Conditional Feature Importance revisited: Double Robustness, Efficiency and Inference

Angel Reyero Lobo
Inst. Math. Toulouse ; UMR5219
Université de Toulouse ; CNRS

Pierre Neuvial
Inst. Math. Toulouse ; UMR5219
Université de Toulouse ; CNRS

Bertrand Thirion
Université Paris-Saclay;
Inria; CEA

Abstract

Conditional Feature Importance (CFI) was introduced long ago to account for the relationship between the studied feature and the rest of the input. However, CFI has not yet been studied from a theoretical perspective because the conditional sampling step has generally been overlooked. In this article, we demonstrate that the recent Conditional Permutation Importance (CPI) is indeed a valid implementation of this concept. Under the conditional null hypothesis, we then establish a **double robustness** property that can be leveraged for variable selection: with either a valid model or a valid conditional sampler, the method correctly identifies null coordinates.

Under the alternative hypothesis, we study the theoretical target and link it to the popular Total Sobol Index (TSI). We introduce the **Sobol-CPI**, which generalizes CPI/CFI, prove that it is nonparametrically efficient, and provide a bias correction. Finally, we propose a consistent and valid type-I error test and present numerical experiments that illustrate our findings.

1 Introduction

Modern machine learning models are capable of making accurate predictions, so using these models can help characterize the dependency structure between variables and thus provide valuable insights for further research. For example, if a model can accurately predict a disease based on a set of genes, identifying the important predictive variables can guide specialists on which genes to study further to combat the disease.

There is usually a trade-off between *model transparency* and *model complexity*. Indeed, simple methods such as linear regression are easily interpretable, since each covariate's importance can be directly assessed via its estimated coefficient. However, using a model that poorly fits the underlying distribution may yield misleading results (Molnar et al., 2021).

For this reason, there is a need to rely on *model-agnostic* approaches based on flexible enough models that truly provide information about the data distribution. In this way, one can adapt to complex situations by capturing non-linear dependencies, recovering the truly important variables in highly correlated settings.

The two main model-agnostic approaches to measure the importance of a covariate are *perturbation* and *removal* approaches (Covert et al., 2021). Both of them aim to disable the information given by the covariate.

In perturbation/permutation approaches, such as Permutation Feature Importance (PFI, Mi et al. (2021)) and Conditional Permutation Importance (CPI, Chamma et al. (2024a)), the information is disabled by computing the loss when permuting the j-th coordinate marginally or conditionally.

The most used removal-based approach is Leave One Covariate Out (LOCO), where the importance of a given covariate is estimated by evaluating the performance of a model re-trained without that covariate (Lei et al., 2018; Williamson et al., 2021).

LOCO is essentially a direct plug-in estimate of the predictive power of the model compared to the predictive power of the restricted model (Williamson et al., 2023). As such, it effectively estimates the Total Sobol Index (TSI), an importance index originating from Sensitivity Analysis (Homma and Saltelli, 1996). TSI aims to assess the predictive capacity of a coordinate given the others, making it a *conditional* approach.

Due to extrapolation issues with PFI (Hooker et al., 2021), there has been a shift towards conditional approaches. These include Conditional Variable Importance—also called Conditional Feature Importance (CFI) in Ewald et al. (2024)—and Conditional Model

Reliance (CMR), defined as a ratio rather than a difference in Fisher et al. (2019). However, the conditional sampling step is often *overlooked*; a Gaussian distribution or copula is commonly used but performs poorly in practice due to covariance estimation issues (Blain et al., 2025). Our first contribution is to show that the conditional permutation used in CPI (and CMR) has theoretical grounding, making it a valid CFI approach.

These conditional methods have shown strong empirical performance (see Chamma et al. (2024a,b); Paillard et al. (2025)). We explain this by a double robustness property, typically sought in causal inference, which we introduce here for variable importance.

Variable selection seeks the smallest set of predictive variables—those not independent of the output *given* the others. Even in highly correlated settings, assuming that each covariate is not directly a function of the others, this set is well-defined and unique. This assumption is standard in controlled variable selection (Candès et al., 2018) and feature importance (Verdinelli and Wasserman, 2024).

Permutation-based approaches do not provide a proper quantification of variable importance. In particular, Bénard et al. (2022) have shown that PFI does not converge to TSI. Yet, the theoretical quantity targeted by CPI has not been studied. In this article, we fill this gap and show that, with a simple adjustment, the method can be adapted to estimate TSI. We call this extension *Sobol-CPI*, an unbiased generalization of CPI, for which we establish nonparametric efficiency and study inference to provide statistical guarantees.

In summary, our main contributions are:

- a theoretical framework under which the conditional sampling step of CPI is valid, which shows that CPI is a Conditional Feature Importance.
- a *double robustness* property: to detect a null covariate, the accuracy of only one of the two estimates used in its computation is sufficient. This explains the strong performance of CPI in variable selection. By contrast, LOCO requires both full and restricted models to be accurate. In a simple linear setting, CPI exhibits quadratic bias decay, whereas LOCO decays only linearly.
- *Sobol-CPI*, a new estimator of TSI based on a permutation approach, bridging the gap with removal methods. We prove its asymptotic efficiency and provide a consistent (power tending to one) and valid (type-I error control) test.
- numerical experiments to illustrate these findings.

Proofs and additional experiments are in the appendix.

2 Framework

2.1 Setting

We observe $\{(x_i, y_i)\}_{i=1,\dots,n_{\text{test}},\dots,n_{\text{test}}+n_{\text{train}}}$ a test and train set sampled i.i.d. from P_0, a distribution that belongs to a class of distributions $\mathcal{M}$, where $X \in \mathcal{X} \subset \mathbb{R}^p$ and $y \in \mathcal{Y} \subset \mathbb{R}$. We are interested in assessing the importance of a coordinate $j \in \{1, \dots, p\}$. We denote by X^j the j-th column of X and by X^{-j} the vector X with the j-th coordinate excluded. We consider a space of functions $\mathcal{F}$ with a norm $\|\cdot\|_{\mathcal{F}}$. The notation is summarized in the glossary (Appendix A).

Similarly to the Knockoffs framework (Barber and Candès, 2015; Candès et al., 2018), we want to avoid making strong assumptions about the relationship between inputs and outputs, as it can be complex. However, we assume that the relationship among the input covariates is simple, typically because they originate from the same generative process (e.g. a measurement device). For this reason, it can be much more efficient and accurate to study the relationship between the j-th covariate, X^j, given the rest, X^{-j}, rather than directly regressing y on X^{-j}. This approach is also advantageous in settings where labeling data is expensive, as there are typically more available samples for the input pair (X^{-j}, X^j) than for the pair (X^{-j}, y). However, empirical robustness to this assumption is demonstrated in Paillard et al. (2025).

We denote by $m(X) := \mathbb{E}[y \mid X]$ (resp. $m_{-j}(X^{-j}) := \mathbb{E}[y \mid X^{-j}]$) the conditional expectation of the input y given X (resp. X^{-j}), by $\widehat{m}$ (resp. $\widehat{m}_{-j}$) its estimation and by $\widehat{m}_n$ to make explicit the dependence on the training sample size. Similarly, we denote by $\nu_{-j}(X^{-j}) := \mathbb{E}[X^j \mid X^{-j}]$ and $\widehat{\nu}_{-j}$ its estimate. We denote by $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ the loss used to compare a prediction $\widehat{m}(X) \in \widehat{\mathcal{Y}} \subset \mathbb{R}$ with the true output $y \in \mathcal{Y}$. These estimations are the usual objective of machine learning models. This is obvious for the mean squared error (MSE), but for many other losses (R^2, deviance, classification accuracy, and the area under the ROC curve, see Williamson et al. (2023)) the minimizer is also a function of this conditional expectation.

Definition 2.1 (TSI)**.** Given $(X, y) \sim P_0$ and a loss ℓ we define the Total Sobol Index of $j \in \{1, \dots, p\}$ as

$$\psi_{TSI}(j, P_0) := \mathbb{E}\left[\ell\left(m_{-j}(X^{-j}), y\right)\right] - \mathbb{E}\left[\ell(m(X), y)\right].$$

We define this quantity as the (Generalized) Total Sobol Index because, when the loss is the quadratic loss, it corresponds to the unnormalized TSI (Homma and Saltelli, 1996). It represents the loss decay when the j-th covariate is not used. It has also been referred to as the Generalized ANOVA (Williamson et al.,

2021) and is considered a standard importance measure (Lei et al., 2018; Rinaldo et al., 2019; Hooker et al., 2021; Bénard et al., 2022; Williamson et al., 2023). A large decay indicates that the j-th covariate is useful, whereas a small decay indicates that the covariate lacks predictive power when the remaining covariates are used. Bénard et al. (2022) considered it as the best quantity for support recovery.

2.2 Related work

LOCO consists of a simple plug-in estimate of the TSI:

Definition 2.2 (LOCO)**.** Given j, a loss ℓ, a regressor $\widehat{m}$ of y given X, a regressor $\widehat{m}_{-j}$ of y given X^{-j} and a test set $(X_i, y_i)_{i=1,\dots n_{\text{test}}}$, LOCO is defined as

$$\widehat{\psi}^j_{\text{LOCO}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \ell\left(\widehat{m}_{-j}(x_i^{-j}), y_i\right) - \ell\left(\widehat{m}(x_i), y_i\right).$$

Nonparametric theory (Kennedy, 2023) shows that a simple plug-in estimate does not achieve optimal convergence, typically requiring a one-step correction. However, Williamson et al. (2021) demonstrated that this correction is unnecessary with the quadratic loss, and Williamson et al. (2023) extended this to other losses under certain regularity conditions. This provides valid confidence intervals. Similar results will be established for our method in Section 4.1.

A major practical limitation of LOCO is that it requires retraining a model, $\widehat{m}_{-j}$, for each coordinate j. This is not only computationally intensive, but it also introduces optimization errors that do not compensate as desired, as discussed in Section 3.2.

A first naive *permutation-based* approach consists in comparing the performance of the estimate on a test set where the j-th column is shuffled: the formal definition of PFI is given in Appendix B. Even though it avoids refitting and Mi et al. (2021) report good performance, it suffers from extrapolation bias, as predictions may fall in low-density regions where $\widehat{m}$ was not trained (Hooker et al., 2021). It also fails to control type-I error with highly correlated covariates (Chamma et al., 2024a). Moreover, Bénard et al. (2022) showed that it does not target an interpretable quantity: it decomposes into three components, only the first being desirable (ψ_{TSI}). The second, the unnormalized marginal total Sobol index, can mislead under high correlation, and the third grows with covariate dependence.

To address this issue, Strobl et al. (2008) proposed *conditionally* permuting the j-th coordinate with respect to the others. In this way, only the information exclusively given by X^j is shuffled, while the relationship with the rest of the coordinates is preserved. As a result, one can make predictions based on the new conditional sample, denoted by $\widetilde{x}_i^{\prime(j)}$, where the apostrophe indicates that it was estimated. More formally, Chamma et al. (2024a) proposed:

Definition 2.3 (CPI)**.** Given j, a loss ℓ, a regressor $\widehat{m}$ of y given X and a test set $(X_i, y_i)_{i=1,\dots n_{\text{test}}}$, CPI is defined as

$$\widehat{\psi}^j_{\text{CPI}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \ell\left(\widehat{m}(\widetilde{x}_i^{\prime(j)}), y_i\right) - \ell\left(\widehat{m}(x_i), y_i\right),$$

where the j-th coordinate is *conditionally* permuted.

For the conditional permutation, they proposed regressing the j-th coordinate with respect to X^{-j} to estimate $\nu_{-j}(X^{-j}) = \mathbb{E}[X^j \mid X^{-j}]$, and then adding the permuted residuals of this regression. Thus, $\widetilde{x}^{\prime(j)l} = x^l$ for any $l \neq j$, and $\widetilde{x}^{\prime(j)j} = \widehat{\nu}_{-j}(x^{-j}) + (x^j - \widehat{\nu}_{-j}(x^{-j}))^{\text{perm}}$, where the permutation is across the individuals. We denote by $\widetilde{X}^{(j)} \sim P_j^\star \in \mathcal{M}$ the theoretical random variable based on the true ν_{-j}, and $\widetilde{X}^{\prime(j)} \sim P'_j$ based on the estimated $\widehat{\nu}_{-j}$. We omit the superscript (j) when the coordinate is clear.

CPI requires training a separate regressor $\widehat{\nu}_{-j}$ for each covariate, which may seem counterintuitive since permutation approaches aim to avoid fitting $\widehat{m}_{-j}$. However, as discussed in Section 2.1, while the relationship between y and X may be complex, between X^j and X^{-j} is assumed to be simple. For this reason, unlike with LOCO, we use a simple model to estimate ν_{-j}.

The goal of CPI is to sample $\widetilde{X}^{(j)}$ such that $\widetilde{X}^{(j)} \sim X, \widetilde{X}^{(j)-j} = X^{-j}$ and $\widetilde{X}^{(j)j} \perp\!\!\!\perp y \mid X^{-j}$. However, Chamma et al. (2024a) did not discuss the assumptions under which this method samples from the target distribution; we address this in Section 3.1. Furthermore, CPI does not estimate a known theoretical quantity, so in Section 4 we propose a modification to target TSI.

Finally, in Chamma et al. (2024a), a type-I error control is proposed using asymptotic normality. However, it does not address the fact that, under the conditional null hypothesis, similarly to what happens with LOCO, the influence function vanishes. Consequently, there is a need to correct the estimated variance to ensure type-I error control (Williamson et al., 2023; Dai et al., 2024; Verdinelli and Wasserman, 2024). In Section 4.3, we address this issue.

3 Theoretical properties of CPI/CFI

3.1 Validity of conditional sampling

We observe that there is no need to preserve the same conditional sampling used for CPI in Chamma et al. (2024a). Normalizing flows (Papamakarios et al.,

2021) are theoretically sound, but impractical due to the large number of training samples they require. Other alternatives are domain-specific generative models (Sesia et al., 2020). Since the conditional sampling used in CPI works in practice and makes the procedure computationally efficient thanks to the use of fast and simple models to estimate ν_{-j}, we continue with the same conditional step, which has also been used in Hooker et al. (2021) and Blain et al. (2025), for example. Let us denote the i-th observation in which only the j-th coordinate has been conditionally permuted using the residual from the k-th observation by:

$$\widetilde{x}_{i,k}^{\prime(j),l} = \begin{cases} x_i^l & \text{if } l \neq j \\ \widehat{\nu}_{-j}(x_i^{-j}) + \left[x_k^j - \widehat{\nu}_{-j}(x_k^{-j})\right] & \text{if } l = j \end{cases} \quad (1)$$

First, we assume that each covariate brings an additive independent innovation, similar to a standard assumption in regression (3.7), but for each covariate.

Assumption 3.1 (Additive innovation)**.** For each $j \in \{1, \ldots, p\}$, there exists a function ν_{-j} such that $X^j = \nu_{-j}(X^{-j}) + \epsilon_j$ with $\epsilon_j \perp\!\!\!\perp X^{-j}$ and $\mathbb{E}[\epsilon_j] = 0$.

This ν_{-j} is exactly the conditional expectation because

$$\mathbb{E}\left[X^j \mid X^{-j}\right] = \mathbb{E}\left[\nu_{-j}\left(X^{-j}\right) + \epsilon_j \mid X^{-j}\right] = \nu_{-j}\left(X^{-j}\right).$$

For instance, Gaussian data satisfies this assumption.

Lemma 3.2 (Gaussian additive noise)**.** *For a Gaussian vector X, additive innovation (3.1) is satisfied.*

We also need the estimate $\widehat{\nu}_{-j}$ to be consistent, i.e. $\mathbb{E}\left[\left(\widehat{\nu}_{-j}\left(X^{-j}\right) - \nu_{-j}\left(X^{-j}\right)\right)^2\right] \to 0$.

For instance, the Random Forest is consistent under mild assumptions (Scornet et al. (2015)). If we assume, for instance, that X is Gaussian, a simple linear model is consistent. To deal with high-dimensionality, we generally use a Lasso (Tibshirani (1996)).

Under the previous assumptions, the 2-Wasserstein distance between the estimated conditional distribution and the true distribution converges to 0:

Proposition 3.3 (Empirical conditional sampling)**.** *Under additive innovation (3.1), if the regressor $\widehat{\nu}_{-j}$ is consistent, then $\widetilde{x}^{\prime(j)}$ constructed as in* (1)*, is sampled from P_j', and $\mathcal{W}_2(P_j', P_j^\star) \to 0$.*

Proposition 3.3 implies that the CPI sampler is asymptotically drawing from the desired conditional distribution. Therefore, CPI qualifies as a Conditional Feature Importance (Hooker et al. (2021)). In the sequel, we refer primarily to CPI due to its feasibility. **However, note that the double robustness result from Theorem 3.6, as well as the use of Sobol-CPI to estimate TSI, are completely general and hold for any conditional sampler. Consequently, these results apply to any CFI.**

3.2 Double robustness

We have observed that there is a need to estimate two regressors for both LOCO ($\widehat{m}$ and $\widehat{m}_{-j}$) and CPI ($\widehat{m}$ and $\widehat{\nu}_{-j}$). More generally, for CPI, we could use any conditional sampler $\widetilde{X}^{\prime(j)}$. In this section, we prove a *double robustness* property for detecting conditionally null covariates: to identify a null covariate, it is sufficient that one of the two estimates is consistent. This contrasts with LOCO, where errors in both estimates must compensate. This property explains the good empirical results obtained by CPI for variable selection (Chamma et al. (2024a,b); Paillard et al. (2025)).

We begin by a general result, then illustrate double robustness for the quadratic loss, and finally derive explicit rates in a linear setting: CPI's bias decays quadratically, whereas LOCO's decays only linearly.

We first assume that m depends only on the conditionally dependent features, and vice versa.

Assumption 3.4 (Functional and conditional equivalence)**.** $X^j \perp\!\!\!\perp y \mid X^{-j}$ if and only if $m(X)$ is not a function of X^j.

This assumption limits feature contributions only to higher-order effects—for example, a feature that is conditionally dependent only through its variance and has no predictive power. As shown in Proposition 3.1 of Reyero-Lobo et al. (2025), it holds under general ML settings, such as the additive noise model (3.7) below. For our results, this assumption is unnecessary, since we only require that conditional dependence implies functional dependence, which always holds.

We require that the ML model does not depend asymptotically on the unimportant features:

Assumption 3.5 (Asymptotic relevance)**.** Denote by $g_j(x, s)$ the vector x with the j-th component replaced 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 n_0 such that for $n \geq n_0$,

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

This is a pointwise convergence on the null features, which can be easily verified for any GLM since the estimated null coefficients tend to 0. It is satisfied under the representability assumption for the Lasso (Zhao and Yu, 2006). Under standard assumptions on Random Forests, the splits are performed with high probability only along the important features (Scornet et al., 2015, Proposition 1). For kernel methods, convergence in the RKHS together with the reproducing property implies pointwise convergence, and thus asymptotic relevance. For Gaussian processes, variable selection can be achieved using dimension-specific scalings (Bhattacharya et al., 2014). Finally, there are

also results for 1-norm penalized SVM (Zhu et al., 2003) and neural networks (Dinh and Ho, 2020). In Section D.2 we show numerically that this property holds across a large benchmark of ML models.

In any case, we note that both assumptions are weak and natural. Since our goal is to study conditional independence using a model, we are implicitly summarizing the X–y relationship through the first-order moment ($\mathbb{E}[Y \mid X]$), and implicitly assuming that the model relies only on the important features. Thus, both assumptions are simply formalizations of what is expected from the model and data distribution when using ML models to study conditional independence.

Theorem 3.6 (Double robustness)**.** *Assume that $X^j \perp\!\!\!\perp y \mid X^{-j}$. Given a conditional sampler such that $\widetilde{X}'^{(j)} \xrightarrow{\mathcal{L}} \widetilde{X}^{(j)}$, then for any bounded and continuous loss ℓ and continuous $\widehat{m}$, $\widehat{\psi}_{\text{CPI}} \to 0$ a.s. On the other hand, given Assumption 3.5, for any conditional sampler $\widetilde{X}'^{(j)}$ and continuous loss ℓ, $\widehat{\psi}_{\text{CPI}} \to 0$ a.s.*

The first implication of the theorem extends Theorem 2 from König et al. (2021) by making explicit the estimation of the conditional distribution. The second relies on the model's implicit variable selection: as the loss optimizer, the model will generally not assign importance to irrelevant features, under the standard assumptions ensuring consistency of the ML model. For unbounded losses, the result holds if the loss is clipped with a sufficiently large constant.

Quadratic loss. In the rest of this section, we focus on the bias introduced by the need to estimate the regressors, therefore, having n_{train} fixed. As is usual in regression, we assume an independent additive noise:

Assumption 3.7 (Additive noise)**.** $y = m(X) + \epsilon$ with $\epsilon \perp\!\!\!\perp X$ and $\mathbb{E}[\epsilon] = 0$.

We study the estimation bias of CPI in Proposition 3.8 and of LOCO in Proposition 3.9.

Proposition 3.8. *Assuming $X^j \perp\!\!\!\perp y \mid X^{-j}$ and additive noise (3.7), we have*

$$\mathbb{E}\left[\widehat{\psi}^j_{\text{CPI}} \middle| \mathcal{D}_{\text{train}}\right] = \mathbb{E}\left[\left(\widehat{m}(\widetilde{X}) - \widehat{m}(\widetilde{X}')\right)^2 + 2\left(m(\widetilde{X}) - \widehat{m}(\widetilde{X})\right)\left(\widehat{m}(\widetilde{X}) - \widehat{m}(\widetilde{X}')\right) \middle| \mathcal{D}_{\text{train}}\right].$$

In this expression, we can observe the mentioned double robustness: it is sufficient to have either an accurate $\widehat{m}$ or $\widehat{\nu}_{-j}$. Specifically, the first error term will vanish either because $\widehat{m}$ detected the j-th coordinate as not important (and therefore the function does not change due to it), or because $\widehat{\nu}_{-j}$ is close to ν_{-j}, making $\widetilde{X}'$ close to $\widetilde{X}$. The second term is treated similarly.

Proposition 3.9. *Under additive noise (3.7), we have*

$$\mathbb{E}\left[\widehat{\psi}^j_{\text{LOCO}} \middle| \mathcal{D}_{\text{train}}\right] = \mathbb{E}\left[\left(m_{-j}(X^{-j}) - \widehat{m}_{-j}(X^{-j})\right)^2 - (m(X) - \widehat{m}(X))^2 \middle| \mathcal{D}_{\text{train}}\right].$$

Therefore, there is a bias due to the estimation of both regressors. There are two main reasons why this error does not cancel out. The first one is an *optimization error*, which is more harmful in complex models. Indeed, it corresponds to a difference of errors from different models that have been independently optimized. Due to the variability of the optimization process, there may be many different solutions and multiple local minima, each with different errors, i.e. given $\widehat{m}_1$ and $\widehat{m}_2$ two optimizations, we do not forcely have $\mathbb{E}\left[(m(X) - \widehat{m}_1(X))^2\right] = \mathbb{E}\left[(m(X) - \widehat{m}_2(X))^2\right]$.

The second source of error is an *estimation error*. This arises because the error distributions of both models differ. From statistical learning theory, we know that the distribution typically depends on the dimension of the model. Since the models have different dimensions, they have different error distributions, which prevents the differences from canceling out. For instance, in the linear model example presented below, we show how this estimation error implies that the error remains at the convergence rate of the linear model ($O(1/n)$), rather than achieving a faster rate as for CPI.

In summary, LOCO compares errors of two independently optimized models, which do not cancel because (i) they are optimized separately and (ii) their error distributions differ. In contrast, Proposition 3.8 shows that CPI's main error term compares the effect of estimating $\widetilde{X}$ *within the same optimized model*.

Linear model. In the remainder of this section, we focus on a Gaussian linear setting, which helps build intuition for CPI's advantages over LOCO.

Assumption 3.10 (Linear model)**.** $y = X\beta + \epsilon$ with $\epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ and $\beta \in \mathbb{R}^p$.

We assume that X is Gaussian. Thus, there is an explicit form for the conditional distribution: $X_j \mid X_{-j} = x_{-j} \sim \mathcal{N}(\mu^j_{\text{cond}}, \Sigma^j_{\text{cond}})$ with $\mu^j_{\text{cond}} := \mu_j + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}(x_{-j} - \mu_{-j})$ and $\Sigma^j_{\text{cond}} := \Sigma_{j,j} - \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j}$. Then, $\nu_{-j}(X^{-j}) := \mathbb{E}\left[X^j | X^{-j}\right]$ is a linear model.

As seen in Lemma I.2, $y \mid X^{-j}$ is also a linear model. Thus, $\widehat{m}$, $\widehat{m}_{-j}$, and $\widehat{\nu}_{-j}$ are linear. However, the bias decays linearly for LOCO and quadratically for CPI.

Lemma 3.11. *Assuming $X^j \perp\!\!\!\perp y \mid X^{-j}$, linear model (3.10), Gaussian covariates, and $\widehat{\nu}_{-j}$ and $\widehat{\beta}$ trained in different samples, then $\mathbb{E}\left[\widehat{\psi}_{\text{CPI}}(j)\right] = O(1/n^2_{\text{train}})$.*

We note that assuming two training samples has negligible impact on this convergence result: splitting a single sample reduces the rate by a factor of 4, which remains quadratic in n_{train}. Finally, for LOCO:

Lemma 3.12. *Assuming $X^j \perp\!\!\!\perp y \mid X^{-j}$, linear model (3.10) and Gaussian covariates then $\mathbb{E}\left[\widehat{\psi}_{\text{LOCO}}(j)\right] = O(1/n_{\text{train}})$.*

Therefore, in this linear setting the double robustness translates as a faster convergence rate.

4 Sobol-CPI

We aim to obtain a stable estimate of TSI (2.1). As noted in Section 2.1, the link between y and X^{-j} can be complex, often requiring computationally intensive models. Furthermore, as discussed in Section 3.2, optimization errors do not accumulate efficiently for retraining methods like LOCO. Therefore, we develop a CPI-based approach to estimate TSI. The key idea is to leverage the law of total expectation, because

$$\begin{aligned} m_{-j}(X^{-j}) &= \mathbb{E}\left[y \mid X^{-j}\right] = \mathbb{E}\left[\mathbb{E}\left[y \mid X\right] \mid X^{-j}\right] \\ &= \mathbb{E}\left[m(X) \mid X^{-j}\right] = \mathbb{E}\left[m(\widetilde{X}^{(j)}) \mid X^{-j}\right], \end{aligned}$$

where $\widetilde{X}^{(j)} \sim X \mid X^{-j}$. We propose to compute $1/n_{\text{cal}} \sum_{i=1}^{n_{\text{cal}}} \widehat{m}\left(\widetilde{x}_i^{'(j)}\right)$, where $\{\widetilde{x}_i^{'(j)}\}_{i=1,\dots,n_{\text{cal}}}$ are sampled from the estimated conditional distribution. This can be easily achieved using CPI's conditional sampling, proven valid in Section 3.1. We thus only need to train a regressor $\widehat{\nu}_{-j}$ and add n_{cal} residuals to obtain the estimate.

This idea is not limited to regression settings. Indeed, as discussed in Williamson et al. (2023), most Bayes predictors are functions of the conditional expectation. For example, under the classical 0-1 loss, the Bayes classifier is given by $\mathbb{I}_{\mathbb{E}[y|X]\geq 0.5}$. In this case, for the restricted model, we propose $\mathbb{I}_{1/n_{\text{cal}} \sum_{i=1}^{n_{\text{cal}}} \widehat{m}\left(\widetilde{x}_i^{(j)}\right) \geq 0.5}$, rather than refitting another model for $\widehat{m}_{-j}$. Then, we define the general Sobol-CPI as this TSI plug-in combination of the conditional sampling total's expectation estimate of m_{-j}:

Definition 4.1 (Sobol-CPI)**.** Given a coordinate j, n_{cal}, a regressor $\widehat{m}$ of y given X and a test set $(X_i, y_i)_{i=1,\dots n_{\text{test}}}$, Sobol-CPI is defined as

$$\widehat{\psi}^j_{\text{SCPI}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \ell\left(\frac{1}{n_{\text{cal}}} \sum_{k=1}^{n_{\text{cal}}} \widehat{m}(\widetilde{x}_{i,k}^{'(j)}), y_i\right) - \ell\left(\widehat{m}(x_i), y_i\right),$$

where the j-th coordinate of $\widetilde{x}_{i,k}^{'(j)}$ is conditionally sampled, and the rest fixed to x_i^{-j} (see (1)).

We observe that this idea of using the law of total expectation is related to the marginalization of the conditional SAGE value functions (cSAGEvf, Covert et al. (2020)), while making explicit the fact that neither the conditional density nor the conditional expectation is known and needs to be estimated. Regarding the first issue, we study in particular the sampling step from the CPI and propose an asymptotic efficiency result. We further observe that, in order to extend this result to other cSAGEvf, it is necessary to specify the sampler explicitly. For the second issue, we propose a bias-correction procedure that yields an unbiased estimator without incurring additional cost.

4.1 Asymptotic efficiency

Under the assumptions of Williamson et al. (2023) (discussed in Section E) and an additional Lipschitz condition ensuring local robustness to small input changes, we achieve the same nonparametric efficiency.

Theorem 4.2. *Assuming additive innovation (3.1), that m is Lipschitz and assumptions A1-A4 and B1-B3 in Section E, $\widehat{\psi}^j_{\text{SCPI}}$ is nonparametric efficient.*

Among the assumptions, those on the loss function have been shown by Williamson et al. (2023) to hold for standard losses, such as the quadratic loss or classification accuracy. We also require the usual $O(n^{-1/4})$ convergence rate for $\widehat{m}$ and $\widehat{\nu}_{-j}$, which is standard in semiparametric inference and can be achieved by common ML models under mild conditions.

4.2 Fixing n_{cal} for the quadratic loss ℓ_2

In practice, we need to decide on a finite value for n_{cal}. Doing so introduces a trade-off between *variable selection* and *variable importance*. Indeed, as discussed in Section J.2, with a small n_{cal}, we recover the double robustness of CPI, achieving better results for variable selection but losing nonparametric efficiency, leading to slightly worse results for variable importance. This is because the optimality assumption (Assumption A1 in Section E), which controls the first-order bias due to the need to estimate m, no longer holds. In any case, with the ℓ_2-loss, fixing n_{cal} induces a bias:

Proposition 4.3 (Bias of $\widehat{\psi}_{\text{SCPI}}$)**.** *For $n_{\text{cal}} < \infty$, under additive noise (3.1) and consistency of $\widehat{m}$ and $\widehat{\nu}_{-j}$, then*

$$\widehat{\psi}^j_{\text{SCPI}} \xrightarrow{n_{\text{train}}, n_{\text{test}} \to \infty} \left(1 + \frac{1}{n_{\text{cal}}}\right) \psi_{\text{TSI}}(j, P_0).$$

This bias can be corrected by scaling by $(1+1/n_{\text{cal}})^{-1}$:

Definition 4.4 (Sobol-CPI(n_{cal}))**.** Given a coordinate j, a fixed n_{cal}, a regressor $\widehat{m}$ of y given X and a test

set $(X_i, y_i)_{i=1,\dots n_{\text{test}}}$, Sobol-CPI(n_{cal}) is defined as

$$\widehat{\psi}^j_{\text{SCPI}(n_{\text{cal}})} = \frac{n_{\text{cal}}}{n_{\text{cal}}+1}\frac{1}{n_{\text{test}}} * \sum_{i=1}^{n_{\text{test}}}\left[\left(\frac{1}{n_{\text{cal}}}\sum_{k=1}^{n_{\text{cal}}}\widehat{m}(\widetilde{x}^{'(j)}_{i,k}) - y_i\right)^2 - (\widehat{m}(x_i)-y_i)^2\right],$$

where the j-th coordinate of $\widetilde{x}^{'(j)}_{i,k}$ is conditionally sampled, and the rest fixed to x_i^{-j} (see (1)).

In particular, by taking $n_{\text{cal}} = 1$, we recover the standard CPI, but divided by 2. With this simple correction, the method works effectively for variable selection, as we directly recover double robustness. Moreover, it will converge to TSI, even though it is not efficient (see Section J.2). This result also corrects Theorem 2 from Hooker et al. (2021), which claimed that Dropped and Conditional Variable Importance coincide, highlighting the gap in the literature between the two approaches.

4.3 Confidence intervals

As stated in Theorem 4.2, there is no need for a one-step estimate to correct any first-order bias. The same holds for LOCO, which does not require such corrections when estimated as a plug-in estimate in the difference of generalized ANOVA (Williamson et al. (2021)) or as a difference of predictiveness measures more generally (Williamson et al. (2023)). Whenever the importance is not null, it is possible to construct confidence intervals using the influence functions. Indeed, the variance is given by $0 < \tau_j^2 := \mathbb{E}[\varphi_j^2(z)] < \infty$, where φ_j is the influence function of $\psi_{\text{TSI}}(j, P_0)$. Therefore, it is possible to estimate it using a plug-in estimate (Williamson et al. (2023)).

Moreover, as shown in Appendix H, when taking the MSE as the predictiveness measure, the variance estimated with this plug-in method matches the natural empirical variance exactly. Nevertheless, under the null hypothesis, the influence function vanishes, preventing a Gaussian asymptotic distribution. This poses a challenge, as it hinders variable selection with direct statistical guarantees—an essential component of reliable scientific discovery. This issue is the same as that faced by LOCO and Shapley due to the quadratic functionals (Verdinelli and Wasserman (2024)).

Several approaches have been proposed to address this problem. Williamson et al. (2023) noted that even if the influence function of the variable importance measure vanishes, excluding extreme cases, the influence functions of the predictiveness measures with and without the covariate do not vanish. They attempted to leverage this observation by computing each predictiveness measure on different data splits. However, as observed in numerical experiments in Section 5, this approach does not work well because it does not only increase variability but also significantly increases bias, resulting in a loss of power and poor estimates.

Other alternatives include inflating the confidence interval by an additive term of the order $O_P(\sqrt{n})$ (Dai et al. (2024); Verdinelli and Wasserman (2024)). Validity can then be obtained using Chebyshev's inequality, resulting in a confidence interval of the form $(-\infty, \widehat{\psi}_n + z_\alpha \text{se}_n + c/\sqrt{n}]$, where se_n is the empirical standard error and c is any constant. In practice, c is taken as the standard error of the output, similar to Verdinelli and Wasserman (2024). We use the same correction idea.

Using Theorem 4.2, we establish the consistency of this conditional null hypothesis test with the additive term (see Appendix G). Furthermore, based on Lemma 3.11, we propose applying Markov's inequality to construct a more powerful test—achieving a rate of c/n^γ for $\gamma < 2$ instead of the standard $c/\sqrt{n}$—as demonstrated in Appendix G. The limitations regarding the extent of interval expansion are discussed in Appendix J.4, both for linear and nonlinear settings.

5 Experiments

To compare importance estimators, we work in a regression setting with the standard quadratic loss. We study both linear and more complex settings. In these cases, it is possible to compute explicitly TSI (see Section K). Figures 1 and 2 are run over 50 and 100 repetitions respectively. The code is available at `https://anonymous.4open.science/r/Sobol-CPI-D9CB`.

We compare the proposed method (with varying n_{cal} values) to `LOCO-W`—the data-splitting version from Williamson et al. (2023)—and `LOCO-HD` from Verdinelli and Wasserman (2024), which, to the best of our knowledge, are the only methods aligned with the same interpretability goals. Our primary focus is on the bias in estimating important and non-important covariates (i.e., the double robustness property), and on the ability to perform powerful, type-I-error-controlled variable selection. For a broader discussion of how n_{cal} balances nonparametric efficiency (variable importance) and double robustness (variable selection), we refer to Section J.2.

Complex learners: we study a nonlinear setting similar to the one of Bénard et al. (2022): $y = X_0X_1\mathbb{I}_{X_2>0} + 2X_3X_4\mathbb{I}_{X_2<0}$, with $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma_{i,j} = 0.6^{|i-j|}$, $p = 50$, and $\mu = \mathbf{0}$. We use a simple Lasso for $\widehat{\nu}_{-j}$ and a Super-Learner van der Laan et al. (2007) comprising Random Forests, Lasso, Gradient

Boosting, and SVM for $\widehat{m}$ and $\widehat{m}_{-j}$. For computational reasons, in the data-splitting version (`LOCO-W`), we used only Gradient Boosting. We applied Sobol-CPI with $n_{\text{cal}} = 1$ and $n_{\text{cal}} = 100$.

In Figure 1, we observe that even if we use a complex model to estimate $\widehat{m}_{-j}$, but a simple one to estimate $\widehat{\nu}_{-j}$, we not only achieve a more computationally efficient estimate but also obtain better results with Sobol-CPI than with LOCO. First, for an important covariate (illustrated here by X_0), Sobol-CPI and LOCO-HD yield similar results, but Sobol-CPI without refitting all the super-learners. Furthermore, for null covariates (such as X_6), Sobol-CPI shows no bias. Then, we see that Sobol-CPI performs better in identifying important covariates. Finally, the double robustness property of Sobol-CPI is evident, as with $n_{\text{cal}} = 100$, it still does not assign importance to the null covariates.

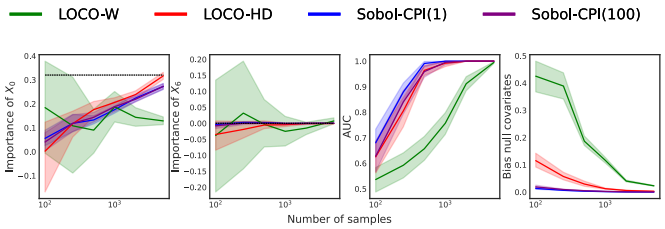


Figure 1: **Double robustness for complex learners:** left to right: TSI estimates for an important covariate (X_0) and a null covariate (X_6); AUC for an importance-based variable selection; bias for null covariates. Sobol-CPI converges at similar rates to TSI and, under the null, it converges faster.

Inference: we study a linear setting with important covariates uniformly sampled with sparsity 0.25, $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma_{i,j} = 0.6^{|i-j|}$, $p = 100$, and $\mu = \mathbf{0}$.

As discussed in Section 4.3, to test the null hypothesis for each covariate, a correction is necessary to address the variance decrease under the null hypothesis. In Figure 2, we present the power of the test using bootstrap variance estimation with a linear correction term. Other approaches, along with analyses of type-I error, computation time, and additional settings (e.g., different correlations and polynomial settings), are discussed in Section J.4.

First, `LOCO-W` lacks power and it requires a large sample size to control the type-I error. We find that Sobol-CPI also reduces the bias for the non-null covariates. Furthermore, it leverages its double robustness property to achieve the largest power. Finally, the type-I error is controlled at the target level 0.05 accross all the procedures but the data-splitting version of Williamson et al. (2023).

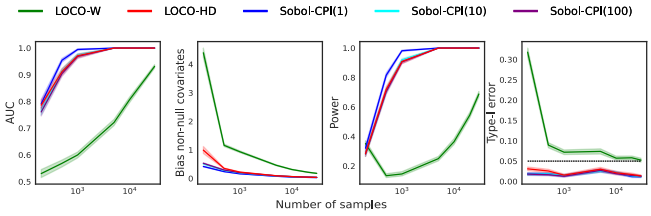


Figure 2: **Statistical Inference on variable importance** in a linear setting with correlation: AUC for variable selection accuracy, bias in non-null TSI estimation, power and type-I error. Sobol-CPI(1) provides the most powerful test. Using the corrected variance, the type-I error is controlled.

6 Conclusion

In this article, we revisited Conditional Feature Importance (CFI). First, we considered the estimation of the conditional distribution, which had been completely ignored in the literature. We then studied both the null and alternative regimes. To explain the good performance of CFI in correctly identifying null features, we proposed a double robustness property. This property means that the method benefits from both the conditional sampler and the model: asymptotically, the sampler draws from the input distribution, so the two sources of error compensate as they share the same distribution, while the model generally induces implicit variable selection during optimization.

For the alternative regime, we modified the CFI and introduced the Sobol-CFI, which we proved to be nonparametrically efficient. We also addressed the bias by proposing a correction that establishes the link between LOCO (refitting) and CFI (perturbation). Finally, we showed how to perform valid inference within the procedure. Overall, we provided a detailed and comprehensive account of the behavior of CFI.

References

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055 – 2085.

Bhattacharya, A., Pati, D., and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth gaussian processes. *The Annals of Statistics*, 42(1).

Blain, A., Lobo, A. R., Linhart, J., Thirion, B., and Neuvial, P. (2025). When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs.

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(4):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(3):551–577.

Chamma, A., Engemann, D. A., and Thirion, B. (2024a). Statistically valid variable importance assessment through conditional permutations. *Advances in Neural Information Processing Systems*, 36.

Chamma, A., Thirion, B., and Engemann, D. (2024b). Variable importance in high-dimensional settings requires grouping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11195–11203. AAAI Press.

Covert, I., Lundberg, S., and Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90.

Covert, I., Lundberg, S. M., and Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223. Curran Associates, Inc.

Dai, B., Shen, X., and Pan, W. (2024). Significance tests of feature relevance for a black-box learner. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1898–1911.

Dinh, V. C. and Ho, L. S. (2020). Consistent feature selection for analytic deep neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2420–2431. Curran Associates, Inc.

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*, page 440–464. Springer Nature Switzerland.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Homma, T. and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):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(82):1–16.

Kennedy, E. H. (2023). Semiparametric doubly robust targeted double machine learning: a review.

König, G., Molnar, C., Bischl, B., and Grosse-Wentrup, M. (2021). Relative feature importance. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9318–9325, Milan, Italy. IEEE.

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(523):1094–1111.

Lobo, A. D. R., Ayme, A., Boyer, C., and Scornet, E. (2024). A primer on linear classification with missing data.

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(1):3008.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B. (2021). General pitfalls of model-agnostic interpretation methods for machine learning models.

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.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Reyero-Lobo, A., Neuvial, P., and Thirion, B. (2025). A principled approach for comparing variable importance.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4).

Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2020). Multi-resolution localization of causal variants across the genome. *Nature Communications*, 11(1):1093.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6:Article25.

Verdinelli, I. and Wasserman, L. (2024). Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4):623–636.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):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(543):1645–1658.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90):2541–2563.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2003). 1-norm support vector machines. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'03, page 49–56, Cambridge, MA, USA. MIT Press.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes,`https://anonymous.4open.science/r/Sobol-CPI-D9CB`]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Notation Glossary

Symbol	Description
$X \in \mathbb{R}^p$	Input
$X^j \in \mathbb{R}$	j-th input covariate
$X^{-j} \in \mathbb{R}^{p-1}$	Input with the j-th covariate excluded
$y \in \mathbb{R}$	Output
x_i	i-th individual
$x_i^{(j)}$	i-th individual with permuted j-th covariate
$m(X)$ (resp. $m_{-j}(X^{-j})$)	$\mathbb{E}[y \mid X]$ (resp. $\mathbb{E}\left[y \mid X^{-j}\right]$)
$\widehat{m}$ (resp. $\widehat{m}_{-j}$)	Estimation of m (resp. of m_{-j})
$\nu_{-j}(X^{-j})$	$\mathbb{E}\left[X^j \mid X^{-j}\right]$
$\widehat{\nu}_{-j}$	Estimation of ν_{-j}
$P_j^\star$	Conditional distribution of X^j given X^{-j}
P_j'	Estimated conditional distribution of X^j given X^{-j}
$\widetilde{X}^{(j)}$	Random variable drawn according to $P_j^\star$
$\widetilde{X}'^{(j)}$	Random variable drawn according to P_j'
$\widetilde{x}_i'^{(j)}$	i-th observation of $\widetilde{X}'^{(j)}$
$\widetilde{x}_{i,k}'^{(j)}$	i-th observation with the added residual from x_k
ℓ	Loss function
$\mathcal{W}_2$	2-Wasserstein distance
$\mathcal{F}$	Generic space of functions
$\xrightarrow{\mathcal{L}}$	Convergence in law
$\mathcal{D}_{\text{train}}$	Train data set
n_{train}	Size of train set
n_{test}	Size of test set
n_{cal}	Size of calibration set
$O(R_n)$	Bounded at a rate of R_n.

The superscript (j) is omitted to avoid index overload when j can be inferred from the context. Also, some notation can be combined; for instance, $\widetilde{x}_{i,k}'^{(j),l}$ denotes the l-th coordinate of $\widetilde{x}_{i,k}'^{(j)}$.

B Permutation Feature Importance (PFI)

We denote by $x_i^{(j)}$ the vector where all coordinates come from the i-th observation, except for the j-th coordinate, which is taken from another observation, as the column has been completely shuffled. Thus, it is given by:

Definition B.1 (PFI)**.** Given a coordinate j, a loss ℓ, a regressor $\widehat{m}$ of y given X and a test set $(X_i, y_i)_{i=1,\dots n_{\text{test}}}$, PFI is defined as

$$\widehat{\psi}_{\text{PFI}}^j = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \ell\left(\widehat{m}(x_i^{(j)}), y_i\right) - \ell\left(\widehat{m}(x_i), y_i\right),$$

where the j-th coordinate is permuted.

C Conditional sampling proofs

C.1 Proof of Lemma 3.2

We begin by decomposing each column as

$$X^j = \mathbb{E}\left[X^j \middle| X^{-j}\right] + \left(X^j - \mathbb{E}\left[X^j \middle| X^{-j}\right]\right).$$

We can denote $\nu_{-j}(X^{-j}) := \mathbb{E}\left[X^j \middle| X^{-j}\right]$ and $\epsilon_j := X^j - \mathbb{E}\left[X^j \middle| X^{-j}\right]$. First, note that they are both Gaussian as $\mathbb{E}\left[X^j | X^{-j}\right] = \mu_j + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}(X^{-j} - \mu_{-j})$, therefore they are linear combinations of coordinates of a Gaussian vector.

Then, note that ϵ_j is centered. Finally, to see that they are independent, as they are both Gaussian variables, we just need to prove that their covariance is null:

$$\begin{aligned}
&\mathbb{E}\left[\left(\nu_{-j}\left(X^{-j}\right) - \mathbb{E}\left[X^j\right]\right)\left(X^j - \mathbb{E}\left[X^j \middle| X^{-j}\right]\right)\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\left(\nu_{-j}\left(X^{-j}\right) - \mathbb{E}\left[X^j\right]\right)\left(X^j - \mathbb{E}\left[X^j \middle| X^{-j}\right]\right) \middle| X^{-j}\right]\right] \\
&\quad = \mathbb{E}\left[\left(\nu_{-j}\left(X^{-j}\right) - \mathbb{E}\left[X^j\right]\right)\mathbb{E}\left[X^j - \mathbb{E}\left[X^j \middle| X^{-j}\right] \middle| X^{-j}\right]\right] = 0.
\end{aligned}$$

C.2 Proof of Proposition 3.3

Proof. In this proof, for simplicity of notation, instead of referring to two random variables X_i and X_k, we will denote them by X and X'. We will first observe that, under Theorem 3.1, $P_j^\star$ can be decomposed as the theoretical counterpart of (1). Then, using the consistency of $\widehat{\nu}_{-j}$, we show that the Wasserstein distance vanishes. We first observe that for $X' \overset{\text{i.i.d}}{\sim} X$ we have that $X'^j - \nu_{-j}\left(X'^{-j}\right) \overset{\text{i.i.d}}{\sim} \epsilon_j$.

Also, we have that $X^j | \left(X^{-j} = x^{-j}\right) = \epsilon_j + \nu_{-j}(x^{-j}) \overset{\text{i.i.d}}{\sim} \left(X'^j - \nu_{-j}\left(X'^{-j}\right)\right) + \nu_{-j}(x^{-j})$, which is exactly the theoretical CPI sampling step, i.e. first we compute the conditional expectation $\nu_{-j}(x^{-j})$ with the regressor ν_{-j} and then we add a permuted residual $\left(X'^j - \nu_{-j}\left(X'^{-j}\right)\right)$. Then, for $\widetilde{X}^{(j)} \sim P_j^\star$, we have that $X^j \overset{\text{i.i.d.}}{\sim} \widetilde{X}^{(j)} | X^{-j}$. We obviously have $X^{-j} = \widetilde{X}^{(j)-j}$. Finally, to ensure that $\widetilde{X}^{(j)} \perp\!\!\!\perp y \mid X^{-j}$, we can rely on the fact that, similar to the construction of knockoffs in Candès et al. (2018), $\widetilde{X}^{(j)}$ is constructed without using y.

To show that the estimated distribution converges to the theoretical one, first recall that the usual 2-Wasserstein distance is given by

$$\left(\inf_{P_\theta \in \Theta(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}P_\theta(\mathrm{d}x, \mathrm{d}y)\right)^{\frac{1}{2}},$$

where $\Theta(\mu, \nu)$ is the set of distributions with marginals μ and ν.

We have that $P_j^\star$ has the same distribution as $\mathbb{E}\left[X^j \middle| X^{-j}\right] + \left(X'^j - \mathbb{E}\left[X'^j | X'^{-j}\right]\right)$ and the empirical couterpart is given by $P'_j := \widehat{\nu}_{-j}\left(X^{-j}\right) + X'^j - \widehat{\nu}_{-j}\left(X'^{-j}\right)$. Now, we bound the distance between them:

$$\begin{aligned}
W_2(P_j^\star, P'_j) &= \left(\inf_{P_\theta \in \Theta(P_j^\star, P'_j)} \int_{\mathbb{R}^{2p} \times \mathbb{R}^{2p}} (x - y)^2 P_\theta(\mathrm{d}x, \mathrm{d}y)\right)^{\frac{1}{2}} \\
&\leq \Bigg(\int_{\mathbb{R}^{2p}} \Big(\mathbb{E}\left[X^j \middle| X^{-j}\right] + \left(X'^j - \mathbb{E}\left[X'^j \middle| X'^{-j}\right]\right) - \widehat{\nu}_{-j}\left(X^{-j}\right) \\
&\qquad - \left(X'^j - \widehat{\nu}_{-j}\left(X'^{-j}\right)\right)\Big)^2 P_X(\mathrm{d}x) P_{X'}(\mathrm{d}x')\Bigg)^{\frac{1}{2}} \\
&= \Bigg(\int_{\mathbb{R}^{2(p-1)}} \Big(\nu_{-j}\left(X^{-j}\right) - \nu_{-j}\left(X'^{-j}\right) \\
&\qquad - \widehat{\nu}_{-j}\left(X^{-j}\right) + \widehat{\nu}_{-j}\left(X'^{-j}\right)\Big)^2 P_{X^{-j}}(\mathrm{d}x^{-j}) P_{X'^{-j}}(\mathrm{d}x'^{-j})\Bigg)^{\frac{1}{2}} \\
&\leq \left(\mathbb{E}\left[\left(\nu_{-j}\left(X^{-j}\right) - \widehat{\nu}_{-j}\left(X^{-j}\right)\right)^2\right]\right)^{\frac{1}{2}} + \left(\mathbb{E}\left[\left(\nu_{-j}\left(X'^{-j}\right) - \widehat{\nu}_{-j}\left(X'^{-j}\right)\right)^2\right]\right)^{\frac{1}{2}}.
\end{aligned}$$

We conclude using that both terms converge to 0 by the consistency of the regressor. □

D Proofs of the double robustness

D.1 Proof of Theorem 3.6

We assume that $X^j \perp\!\!\!\perp y \mid X^{-j}$. We start by assuming that $\widetilde{X}'^{(j)} \xrightarrow{\ell} \widetilde{X}^{(j)}$. We have that for any bounded and continuous ℓ and continuous function f,

$$\begin{aligned}\left|\widehat{\psi}_{\mathrm{CPI}}(j)\right| &= \left|\frac{1}{n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}} \ell\left(f(\widetilde{x}_i'^{(j)}), y_i\right) - \ell\left(f(x_i), y_i\right)\right| \\ &\leq \left|\frac{1}{n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}} \ell\left(f(\widetilde{x}_i'^{(j)}), y_i\right) - \ell\left(f(x_i), y_i\right) - \mathbb{E}\left[\ell\left(f(\widetilde{X}'^{(j)}), y\right) - \ell\left(f(X), y\right)\right]\right| \\ &\quad + \left|\mathbb{E}\left[\ell\left(f(\widetilde{X}'^{(j)}), y\right) - \ell\left(f(X), y\right)\right]\right|.\end{aligned}$$

The first term converges to 0 due to the Law of Large Numbers. The second term converges to 0 because of Portmanteau lemma, because under the null hypothesis $j \in \mathcal{H}_0$, $X^j \perp\!\!\!\perp y \mid X^{-j}$, then $(\widetilde{X}^{(j)}, y) \sim (X, y)$. In particular

$$\ell\left(f(X), y\right) \sim \ell\left(f(\widetilde{X}^{(j)}), y\right).$$

On the other hand, we observe that using Assumption 3.4 we have the equivalence between the conditional independence and functional independence, i.e. as $X^j \perp\!\!\!\perp y \mid X^{-j}$, then $m(X)$ does not depend on X^j. Let $\epsilon > 0$. First, using the continuity of the loss, we have that for any ϵ, there exists $\delta > 0$ such that

$$|\ell(\hat{m}(\tilde{x}_i'), y_i) - \ell(\hat{m}(\tilde{x}_i), y_i)| < \epsilon$$

if $|\hat{m}(\tilde{x}_i') - \hat{m}(\tilde{x}_i)| < \delta$.

Second, using the asymptotic relevance (3.5), for any $\delta > 0$, as $j \in \mathcal{H}_0$, there exists n_0 such that for any $n > n_0$,

$$|\hat{m}_n(\tilde{x}_i') - \hat{m}_n(\tilde{x}_i)| < \delta.$$

Therefore, for any $n > n_0$,

$$\left|\frac{1}{n_{\mathrm{test}}}\sum \ell(\hat{m}(\tilde{x}_i'), y_i) - \ell(\hat{m}(\tilde{x}_i), y_i)\right| \leq \frac{1}{n_{\mathrm{test}}}\sum |\ell(\hat{m}(\tilde{x}_i'), y_i) - \ell(\hat{m}(\tilde{x}_i), y_i)| \leq \epsilon.$$

As this was proven for any $\epsilon > 0$, we have that $\hat{\psi}_{\mathrm{CPI}} \to 0$ a.s.

D.2 On the asymptotic relevance

In this section, we numerically explore the asymptotic relevance assumption (3.5). This assumption can be interpreted as requiring a model that does not extrapolate under the null hypothesis. Thanks to the equivalence between functional and conditional dependence (see Assumption 3.4 and Reyero-Lobo et al. (2025)), the theoretical model does not depend on the j-th coordinate whenever $X^j \perp\!\!\!\perp Y \mid X^{-j}$. This is easily illustrated for the linear model, where the coefficients of the null coordinates converge to zero; hence, imputations on these coordinates are irrelevant, since for sufficiently large n_0 the multiplied terms vanish.

In contrast, most models are not as transparent as the linear model, and one cannot simply verify the absence of dependence on the j-th coordinate by inspecting an estimated coefficient. However, this can be assessed model-agnostically using the Permutation Feature Importance (PFI, see Section B). The theoretical PFI index was proven to be zero under the null hypothesis in Reyero-Lobo et al. (2025), as it satisfies their proposed minimal axiom. In Figure 3, we show that for a large benchmark of ML models, the empirical PFI converges to zero. This indicates that there is no extrapolation under the null hypothesis, as the models indeed do not rely on the null coordinate.

The data are generated according to

$$Y = X_1^2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.2), \quad X \sim \mathcal{N}(0, \Sigma),$$

and we study the effect of varying the correlation between the null coordinate X_2 and the important coordinate X_1. In particular, we consider $\Sigma_{1,2} \in \{0, 0.3, 0.6, 0.9\}$. The sample size is $n = 1000$, with $n_{\text{train}} = 800$ and $n_{\text{test}} = 200$. All models are implemented using the default specifications of `scikit-learn` (Pedregosa et al. (2011)).

We first observe that linear regression and the Lasso perform poorly, as the quadratic relationship with the output is not captured by their linear specifications. Across all other models, the PFI tends to zero, confirming the absence of functional dependence on the null coordinate. The only exceptions arise at very high correlations, where Support Vector Regression (SVR) and k-Nearest Neighbors (KNN) exhibit slight model dependence on the null coordinate.

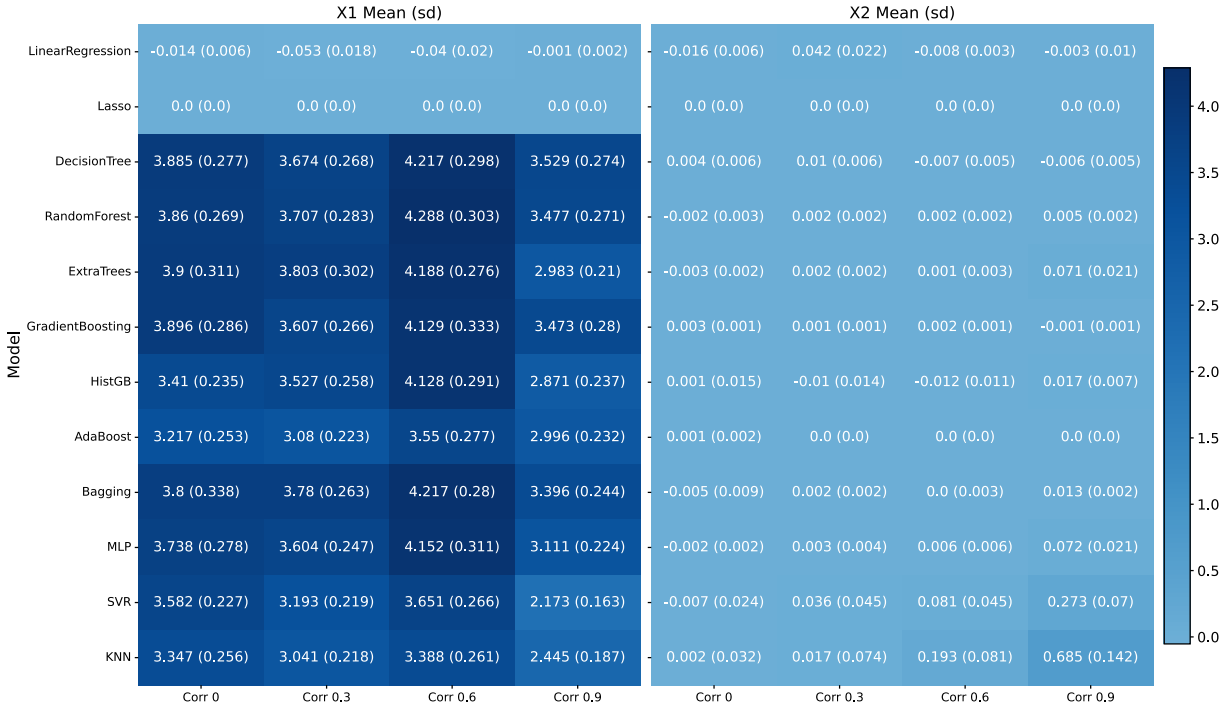


Figure 3: **Asymptotic relevance on standard ML models:** Permutation Feature Importance (PFI) mean and standard deviation in parenthesis for X_1 (left) and X_2 (right) across different correlation levels and models, where $Y = X_1^2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.2), \quad X \sim \mathcal{N}(0, \Sigma)$, with $\Sigma_{1,2} \in \{0, 0.3, 0.6, 0.9\}$. The sample size is $n = 1000$, with 80% used for training and 20% for testing. The experiment was repeated 100 times.

D.3 Proof of Proposition 3.8

Proof. As we are under the conditional null hypothesis, using Theorem I.1 we have that there exists a function $m_{-j} \in \mathcal{F}_{-j}$ such that $m(X) = m_{-j}(X^{-j})$.

To estimate TSI, as seen in Section 4.2, we begin by dividing CPI by 2. We observe that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{2}\widehat{\psi}_{\text{CPI}}(j, P_0)\middle|\mathcal{D}_{\text{train}}\right] &= \mathbb{E}\left[\frac{1}{2n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\left(y_i - \widehat{m}(\widetilde{X}_i')\right)^2 - (y_i - \widehat{m}(X_i))^2\middle|\mathcal{D}_{\text{train}}\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left(y - \widehat{m}(\widetilde{X}')\right)^2 - (y - \widehat{m}(X))^2\middle|\mathcal{D}_{\text{train}}\right]. \end{aligned} \tag{2}$$

We first note that using Theorem 3.7, we have that

$$\mathbb{E}\left[\left(y-\widehat{m}(\widetilde{X}')\right)^2-(y-\widehat{m}(X))^2\middle|\mathcal{D}_{\text{train}}\right]$$
$$=\mathbb{E}\left[\left(m(X)-\widehat{m}(\widetilde{X}')\right)^2-(m(X)-\widehat{m}(X))^2\middle|\mathcal{D}_{\text{train}}\right].$$

We can develop the first term as

$$\mathbb{E}\left[\left(m(X)-\widehat{m}(\widetilde{X}')\right)^2\middle|\mathcal{D}_{\text{train}}\right]$$
$$=\mathbb{E}\left[\left(m(X)-m(\widetilde{X})\right)^2\right]+\mathbb{E}\left[\left(m(\widetilde{X})-\widehat{m}(\widetilde{X}')\right)^2\middle|\mathcal{D}_{\text{train}}\right]$$
$$+2\mathbb{E}\left[(m(X)-m(\widetilde{X}))(m(\widetilde{X})-\widehat{m}(\widetilde{X}')))\middle|\mathcal{D}_{\text{train}}\right].$$

We observe that the last crossed-term is null because using the conditional null hypothesis we have that $m(X)=m(\widetilde{X})$. We note that the first term is exactly twice the Total Sobol Index, which under the conditional null hypothesis it is also null. The second term can be developed as

$$\mathbb{E}\left[\left(m(\widetilde{X})-\widehat{m}(\widetilde{X}')\right)^2\middle|\mathcal{D}_{\text{train}}\right]=\mathbb{E}\left[(m(\widetilde{X})-\widehat{m}(\widetilde{X}))^2\middle|\mathcal{D}_{\text{train}}\right]+\mathbb{E}\left[(\widehat{m}(\widetilde{X})-\widehat{m}(\widetilde{X}'))^2\middle|\mathcal{D}_{\text{train}}\right]$$
$$+2\mathbb{E}\left[(m(\widetilde{X})-\widehat{m}(\widetilde{X})))(\widehat{m}(\widetilde{X})-\widehat{m}(\widetilde{X}'))\middle|\mathcal{D}_{\text{train}}\right].$$

Then, using that $\widetilde{X}\overset{\text{i.i.d.}}{\sim}X$, we have that $\mathbb{E}\left[(m(\widetilde{X})-\widehat{m}(\widetilde{X}))^2\middle|\mathcal{D}_{\text{train}}\right]=\mathbb{E}\left[(m(X)-\widehat{m}(X))^2\middle|\mathcal{D}_{\text{train}}\right]$, so this first term will cancel with the second term in (2).

Finally, combining all the previous we have that

$$\mathbb{E}\left[\frac{1}{2}\widehat{\psi}_{\text{CPI}}(j,P_0)\middle|\mathcal{D}_{\text{train}}\right]$$
$$=\frac{1}{2}\mathbb{E}\left[\left(y-\widehat{m}(\widetilde{X}')\right)^2\middle|\mathcal{D}_{\text{train}}\right]-\mathbb{E}\left[(y-\widehat{m}(X))^2\middle|\mathcal{D}_{\text{train}}\right]$$
$$=\psi_{\text{TSI}}(j,P_0)+\frac{1}{2}\mathbb{E}\left[\left(\widehat{m}(\widetilde{X})-\widehat{m}(\widetilde{X}')\right)^2\middle|\mathcal{D}_{\text{train}}\right]$$
$$+\mathbb{E}\left[\left(m(\widetilde{X})-\widehat{m}(\widetilde{X})))(\widehat{m}(\widetilde{X})-\widehat{m}(\widetilde{X}')\right)\middle|\mathcal{D}_{\text{train}}\right]$$
$$=\frac{1}{2}\mathbb{E}\left[\left(\widehat{m}(\widetilde{X})-\widehat{m}(\widetilde{X}')\right)^2\middle|\mathcal{D}_{\text{train}}\right]+\mathbb{E}\left[\left(m(\widetilde{X})-\widehat{m}(\widetilde{X})))(\widehat{m}(\widetilde{X})-\widehat{m}(\widetilde{X}')\right)\middle|\mathcal{D}_{\text{train}}\right].$$

□

D.4 Proof of Proposition 3.9

Proof.

$$\mathbb{E}\left[\widehat{\psi}_{\text{LOCO}}(j)\middle|\mathcal{D}_{\text{train}}\right]=\mathbb{E}\left[\frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\left(y_i-\widehat{m}_{-j}(X_i^{-j})\right)^2-(y_i-\widehat{m}(X_i))^2\middle|\mathcal{D}_{\text{train}}\right]$$
$$=\mathbb{E}\left[\left(y-\widehat{m}_{-j}(X^{-j})\right)^2-(y-\widehat{m}(X))^2\middle|\mathcal{D}_{\text{train}}\right].$$

On the one hand, we have that

$$\mathbb{E}\left[\left(y-\widehat{m}_{-j}(X^{-j})\right)^2\middle|\mathcal{D}_{\text{train}}\right]=\sigma^2+\mathbb{E}\left[\left(m(X)-\widehat{m}_{-j}(X^{-j})\right)^2\middle|\mathcal{D}_{\text{train}}\right]\text{ using Theorem 3.7}$$
$$=\sigma^2+\mathbb{E}\left[\left(m_{-j}(X^{-j})-\widehat{m}_{-j}(X^{-j})\right)^2\middle|\mathcal{D}_{\text{train}}\right]$$
$$+\mathbb{E}\left[\left(m(X)-m_{-j}(X^{-j})\right)^2\right],\tag{3}$$

where we have used that

$$\begin{aligned}&\mathbb{E}\left[(m_{-j}(X^{-j})-\widehat{m}_{-j}(X^{-j}))(m(X)-m_{-j}(X^{-j}))|\mathcal{D}_{\text{train}}\right]\\&\quad=\mathbb{E}\left[(m_{-j}(X^{-j})-\widehat{m}_{-j}(X^{-j}))\mathbb{E}\left[(m(X)-m_{-j}(X^{-j}))|X^{-j},\mathcal{D}_{\text{train}}\right]|\mathcal{D}_{\text{train}}\right]=0.\end{aligned}$$

We note that $\mathbb{E}\left[\left(m(X)-m_{-j}(X^{-j})\right)^2\right]$ is exactly $\psi_{\text{TSI}}(j,P_0)$. On the other hand, we also have that

$$\mathbb{E}\left[(y-\widehat{m}(X))^2\Big|\mathcal{D}_{\text{train}}\right]=\sigma^2+\mathbb{E}\left[(m(X)-\widehat{m}(X))^2\big|\mathcal{D}_{\text{train}}\right]. \tag{4}$$

Combining (3) and (4), we have that

$$\begin{aligned}&\mathbb{E}\left[\widehat{\psi}_{\text{LOCO}}(j)\Big|\mathcal{D}_{\text{train}}\right]\\&\quad=\psi_{\text{TSI}}(j,P_0)+\mathbb{E}\left[\left(m_{-j}(X^{-j})-\widehat{m}_{-j}(X^{-j})\right)^2\Big|\mathcal{D}_{\text{train}}\right]-\mathbb{E}\left[(m(X)-\widehat{m}(X))^2\big|\mathcal{D}_{\text{train}}\right].\end{aligned}$$

□

D.5 Proofs of double robustness in Gaussian linear setting

As X is Gaussian, as recalled in the main text, there is an explicit form for the distribution of a coordinate given the others: $X_j|X_{-j}=x_{-j}\sim\mathcal{N}(\mu^j_{\text{cond}},\Sigma^j_{\text{cond}})$ with $\mu^j_{\text{cond}}:=\mu_j+\Sigma_{j,-j}\Sigma^{-1}_{-j,-j}(x_{-j}-\mu_{-j})$ and with $\Sigma^j_{\text{cond}}:=\Sigma_{j,j}-\Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j}$. For simplicity, we assume that $\mu=0$. Then, $\mathbb{E}\left[X^j|X^{-j}\right]=\Sigma_{j,-j}\Sigma^{-1}_{-j,-j}X_{-j}$, so it is a linear model. We will then denote it as $\nu_{-j}(X^{-j})=X^{-j}\gamma_{-j}$ for $\gamma_{-j}\in\mathbb{R}^{p-1}$.

We will denote by $\Delta\beta=\beta-\widehat{\beta}$ where $\widehat{\beta}$ was obtained by regression y given X. Similarly, $\Delta\beta'=\beta'-\widehat{\beta}'$, where $\widehat{\beta}'$ was trained only over X^{-j} and $\Delta\gamma_{-j}=\gamma_{-j}-\widehat{\gamma}_{-j}$.

We start by showing that the bias of LOCO decreases linearly with the training sample, and then the quadratic decay for CPI.

D.5.1 Linear decay of LOCO

We decompose this proof into two parts. First, we prove Theorem D.1, which rigorously decomposes the bias into the bias of estimating each linear model $\widehat{m}$ and $\widehat{m}_{-j}$. Then, in Theorem 3.12, we use the underlying distribution to compute the expectation and prove the linear decay.

Proposition D.1 (LM LOCO bias)**.** *Under Theorem 3.10 with X Gaussian, we have that*

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}_{\text{LOCO}}(j,P_0)\Big|\mathcal{D}_{\text{train}}\right]&=\psi_{TSI}(j,P_0)+\Delta\beta'^{\top}\Sigma_{-j}\Delta\beta'-\Delta\beta^{\top}\Sigma\Delta\beta\\&=\psi_{TSI}(j,P_0)+\|\Delta\beta'\|^2_{\Sigma_{-j}}-\|\Delta\beta\|^2_{\Sigma}.\end{aligned}$$

Proof. First we note that

$$y-\widehat{m}(X)=X\beta+\epsilon-X\widehat{\beta}=X\Delta\beta+\epsilon\sim\mathcal{N}(0,\sigma^2+\Delta\beta^{\top}\Sigma\Delta\beta).$$

Similarly, using Theorem I.2, we have that

$$y-\widehat{m}_{-j}(X^{-j})=X^{-j}\Delta\beta'+\epsilon'\sim\mathcal{N}(0,\Delta\beta'^{\top}\Sigma_{-j}\Delta\beta'+\sigma^2+\Sigma_{\text{cond}}{\beta^j}^2).$$

Therefore, we have that

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}_{\text{LOCO}}(j,P_0)\Big|\mathcal{D}_{\text{train}}\right]&=\mathbb{E}\left[(y-\widehat{m}_{-j}(X^{-j}))^2\big|\mathcal{D}_{\text{train}}\right]-\mathbb{E}\left[(y-\widehat{m}(X))^2\big|\mathcal{D}_{\text{train}}\right]\\&=\mathbb{V}\left(y-\widehat{m}_{-j}(X^{-j})\big|\mathcal{D}_{\text{train}}\right)-\mathbb{V}\left(y-\widehat{m}(X)|\mathcal{D}_{\text{train}}\right)\\&=\sigma^2+{\beta^j}^2\Sigma_{\text{cond}}+\Delta\beta'^{\top}\Sigma_{-j}\Delta\beta'-\sigma^2-\Delta\beta^{\top}\Sigma\Delta\beta\\&=\psi_{\text{TSI}}(j,P_0)+\|\Delta\beta'\|^2_{\Sigma_{-j}}-\|\Delta\beta\|^2_{\Sigma},\end{aligned}$$

where we have used that $\psi_{\text{TSI}}(j, P_0) = {\beta^j}^2 \Sigma_{\text{cond}}$.

□

For an initial intuition, let's assume we have independent training samples for each regressor (which is even proposed in Williamson et al. (2023) to control the type-I error).

Therefore, we observe that the bias is distributed as a linear combination of independent chi-squared covariates. Indeed, using that in a linear model $\widehat{\beta} \sim \mathcal{N}(0, \sigma^2 (X_{\text{tr}}^\top X_{\text{tr}})^{-1})$, we have that $\Delta\beta' \sim (\sigma^2 + \beta_j^2 \Sigma_{\text{cond}}) \mathcal{N}(0, (X_{\text{tr},-j}^\top X_{\text{tr},-j})^{-1})$ and $\Delta\beta \sim \mathcal{N}(0, \sigma^2 (X_{\text{tr}}^\top X_{\text{tr}})^{-1})$. Therefore, for independent estimates, we would diagonalize and obtain a linear combination of independent chi-squared variables. This decreases linearly with the number of samples, as the error rate of the linear model. Intuitively, $\Sigma^{-1/2} X_{\text{tr}}^\top X_{\text{tr}} \Sigma^{-1/2}$ is *almost* the identity matrix multiplied by a coefficient decreasing in probability linearly with n_{train}. Similarly for $\Sigma_{-j}^{-1/2} X_{-j}^\top X_{-j} \Sigma_{-j}^{-1/2}$. Therefore, we have that

$$\|\Delta\beta'\|^2_{\Sigma_{-j}} - \|\Delta\beta\|^2_{\Sigma} \approx \Big((\sigma^2 + {\beta^j}^2 \Sigma_{\text{cond}}) \chi^2(p-1) \Big) O_P(1/n_{\text{train}}) - \big(\sigma^2 \chi^2(p) \big) O_P(1/n_{\text{train}}),$$

with $O_P(1/n)$ meaning that it decreases in probability linearly with n. We observe that in this example both errors follow the same distribution. This is because, as shown in Theorem I.2, the restricted model also satisfies the linear model. We observe that this is not generally the case: Lobo et al. (2024) showed that even under the global logistic assumption, the restricted model is not logistic.

Even if both errors are not independent, there is no reason that the error made in a given coordinate would be compensated with the error made in the same coordinate in the global model. Therefore, this is also one difference with the error made by CPI, as will be seen later: in CPI, we will only take into account the error made to estimate the β^j coordinate.

Under the null hypothesis, we have that $\beta^j = 0$.

We show that under the null hypothesis the expectation of the error decreases linearly with the number of samples:

Lemma D.2. *Under the conditional null hypothesis, Assumption 3.10 and Gaussian covariates, we have that*

$$\mathbb{E}\left[\widehat{\psi}_{\text{LOCO}}(j, P_0)\right] = \psi_{\text{TSI}}(j, P_0) + O(1/n_{\text{train}}) = O(1/n_{\text{train}}).$$

Therefore, in any case, assuming two training sets or one training set, the bias is of the order of n_{train}.

Proof. Using Proposition D.1, we only need to compute $\mathbb{E}\left[\|\Delta\beta'\|^2_{\Sigma_{-j}}\right]$ and $\|\Delta\beta\|^2_{\Sigma}$. We compute the second one and the first one is done equivalently.

We first observe that $\Delta\beta | X_{\text{tr}} \sim \mathcal{N}(0, \sigma^2 (X_{\text{tr}}^\top X_{\text{tr}})^{-1})$. We denote by

$$A := \Sigma^{1/2} \Delta\beta / \sigma^2 \sim \mathcal{N}(0, (\Sigma^{-1/2} X_{\text{tr}}^\top X_{\text{tr}} \Sigma^{-1/2})^{-1}).$$

We have that

$$\begin{aligned}
\mathbb{E}\left[\|\Delta\beta\|^2_{\Sigma}\right] &= \sigma^4 \mathbb{E}\left[A^\top A\right] = \sigma^4 \mathbb{E}\left[\text{tr}(A^\top A)\right] = \sigma^4 \mathbb{E}\left[\text{tr}(AA^\top)\right] = \sigma^4 \text{tr}(\mathbb{E}\left[AA^\top\right]) \\
&= \sigma^4 \text{tr}(\mathbb{E}\left[\mathbb{E}\left[AA^\top | X_{\text{tr}}\right]\right]) = \sigma^4 \text{tr}(\mathbb{E}\left[(\Sigma^{-1/2} X_{\text{tr}}^\top X_{\text{tr}} \Sigma^{-1/2})^{-1}\right]).
\end{aligned}$$

We remark that $X_{tr}^i \Sigma^{-1/2} \sim \mathcal{N}(0, I_p)$. Therefore, $(\Sigma^{-1/2} X_{\text{tr}}^\top X_{\text{tr}} \Sigma^{-1/2})^{-1}$ is distributed as a Inverse-Wishart with I_d the scale matrix and n_{train} the degrees of freedom. Therefore, its expectation is $I_d/(n_{\text{train}} - p - 1)$. Therefore, we have that

$$\mathbb{E}\left[\|\Delta\beta\|^2_{\Sigma}\right] = \sigma^4 \frac{1}{n_{\text{train}} - p - 1} \text{tr}(I_p) = \sigma^4 \frac{p}{n_{\text{train}} - p - 1}.$$

Similarly for $\mathbb{E}\left[\|\Delta\beta'\|^2_{\Sigma_{-j}}\right]$, we have that

$$\mathbb{E}\left[\|\Delta\beta'\|^2_{\Sigma_{-j}}\right] = \sigma^4 \frac{p-1}{n_{\text{train}}-p-2}.$$

Therefore, we conclude that

$$\begin{aligned}\mathbb{E}\left[\|\Delta\beta'\|^2_{\Sigma_{-j}}\right] - \mathbb{E}\left[\|\Delta\beta\|^2_{\Sigma}\right] &= \sigma^4 \frac{p-1}{n_{\text{train}}-p-2} - \sigma^4 \frac{p}{n_{\text{train}}-p-1} \\ &= \sigma^4 \frac{-n_{\text{train}}+2p+1}{(n_{\text{train}}-p-1)(n_{\text{train}}-p-2)} = O(1/n_{\text{train}}).\end{aligned}$$

□

D.5.2 Quadratic decay of CPI

In this section, we prove that CPI benefits from double robustness, achieving a faster convergence rate for the conditional null hypothesis. For notational simplicity, we suppress the superscript in $\widetilde{X}^{(j)}$.

Indeed, we first note that in the linear Gaussian setting, the bias will consist of the product of errors.

Proposition D.3 (LM CPI bias)**.** *Under Assumption 3.10 with X Gaussian and $j \in \mathcal{H}_0$, we have that*

$$\mathbb{E}\left[\frac{1}{2}\widehat{\psi}_{\text{CPI}}(j)\middle|\mathcal{D}_{\text{train}}\right] \approx \psi_{\text{TSI}}(j, P_0) + \widehat{\beta}_j^2 \Delta\gamma_{-j}^\top \Sigma_{-j} \Delta\gamma_{-j} \tag{5}$$

$$= \psi_{\text{TSI}}(j, P_0) + \widehat{\beta}_j^2 \|\Delta\gamma_{-j}\|^2_{\Sigma_{-j}}. \tag{6}$$

We first note that $\|\Delta\gamma_{-j}\|^2_{\Sigma_{-j}}$ represents the error resulting from using a finite sample to estimate the function ν_{-j}. Since this function is linear, the error will decrease linearly with the size of the training sample. We also observe that, in this case, contrary to what occurs in the LOCO approach, the only estimation error accounted for in the global model is in the estimation of β_j. In general, this error will converge to the coefficient, leading to the usual linear convergence rate in the product.

Nevertheless, under the conditional null hypothesis, this coefficient is zero. Consequently, it will also decrease linearly with the training sample size. Since both errors are multiplied, we obtain a quadratic convergence rate, meaning that CPI will detect more quickly when a covariate is null, resulting in fewer false positives than when using LOCO for variable selection with statistical guarantees.

Proof. We first note that since X is Gaussian, Assumption 3.1 is satisfied. Therefore, from Proposition 3.8 we have that

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{2}\widehat{\psi}_{\text{CPI}}(j)\middle|\mathcal{D}_{\text{train}}\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left(\widehat{m}(\widetilde{X}) - \widehat{m}(\widetilde{X}')\right)^2\middle|\mathcal{D}_{\text{train}}\right] + \mathbb{E}\left[\left(m(\widetilde{X}) - \widehat{m}(\widetilde{X}))\right)(\widehat{m}(\widetilde{X}) - \widehat{m}(\widetilde{X}')\right)\middle|\mathcal{D}_{\text{train}}\right].\end{aligned}$$

We start by noting that the second bias term is negligible:

$$\begin{aligned}&\mathbb{E}\left[\left(m(\widetilde{X}) - \widehat{m}(\widetilde{X}))\right)(\widehat{m}(\widetilde{X}) - \widehat{m}(\widetilde{X}')\right)\middle|\mathcal{D}_{\text{train}}\right] = \mathbb{E}\left[(\Delta\beta^\top \widetilde{X})(\widehat{\beta}^\top(\widetilde{X} - \widetilde{X}'))\middle|\mathcal{D}_{\text{train}}\right] \\ &= \Delta\beta^\top \mathbb{E}\left[\widetilde{X}(\widehat{\beta}^j(\nu_{-j}(X^{-j}) + (X'^j - \nu_{-j}(X'^{-j})) - \widehat{\nu}_{-j}(X^{-j}) - (X'^j - \widehat{\nu}_{-j}(X'^{-j}))))\middle|\mathcal{D}_{\text{train}}\right] \\ &= \widehat{\beta}^j \Delta\beta^\top \mathbb{E}\left[\widetilde{X}(\nu_{-j}(X^{-j}) - \widehat{\nu}_{-j}(X^{-j})) - (\nu_{-j}(X'^{-j}) - \widehat{\nu}_{-j}(X'^{-j})))\middle|\mathcal{D}_{\text{train}}\right] \\ &= \widehat{\beta}^j \Delta\beta^\top \mathbb{E}\left[\widetilde{X}(\Delta\gamma_{-j}^\top X^{-j} - \Delta\gamma_{-j}^\top X'^{-j})\middle|\mathcal{D}_{\text{train}}\right] \\ &= \widehat{\beta}^j \Delta\beta^\top \mathbb{E}\left[\widetilde{X}(X^{-j} - X'^{-j})^\top\middle|\mathcal{D}_{\text{train}}\right]\Delta\gamma_{-j}.\end{aligned}$$

Then, we observe that it depends both in the estimation error of β and of γ_{-j}, which are both consistent. Moreover, as $j \in \mathcal{H}_0$, then $\widehat{\beta}_j \to 0$. This third-order term is negligible.

For the first bias term we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\widehat{m}(\widetilde{X}) - \widehat{m}(\widetilde{X}')\right)^2 \middle| \mathcal{D}_{\text{train}}\right] \\
&= \widehat{\beta}^2 \mathbb{E}\left[\left(\nu_{-j}(X^{-j}) + (X'^j - \nu_{-j}(X'^{-j}))) - \widehat{\nu}_{-j}(X^{-j}) - (X'^j - \widehat{\nu}_{-j}(X'^{-j}))\right)^2 \middle| \mathcal{D}_{\text{train}}\right] \\
&= \widehat{\beta}^2 \mathbb{E}\left[\left(\nu_{-j}(X^{-j}) - \widehat{\nu}_{-j}(X^{-j}) - (\nu_{-j}(X'^{-j}) - \widehat{\nu}_{-j}(X'^{-j}))\right)^2 \middle| \mathcal{D}_{\text{train}}\right] \\
&= \widehat{\beta}^2 \mathbb{E}\left[\left(\Delta\gamma_{-j}^\top X^{-j} - \Delta\gamma_{-j}^\top X'^{-j}\right)^2 \middle| \mathcal{D}_{\text{train}}\right] \\
&= \widehat{\beta}^2 \Delta\gamma_{-j}^\top \mathbb{E}\left[\left(X^{-j} - X'^{-j}\right)\left(X^{-j} - X'^{-j}\right)^\top \middle| \mathcal{D}_{\text{train}}\right] \Delta\gamma_{-j} \\
&= 2\widehat{\beta}_j^2 \|\Delta\gamma_{-j}\|_{\Sigma_{-j}}^2.
\end{aligned}$$

□

Lemma D.4. *Under the conditional null hypothesis, Assumption 3.10 and Gaussian covariates, assume $\widehat{\nu}_{-j}$ and $\widehat{\beta}$ trained in different samples, we have that*

$$\mathbb{E}\left[\widehat{\psi}_{\text{CPI}}(j)\right] = O(1/n_{\text{train}}^2).$$

Proof. We observe that $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 ({X_{\text{tr}}}^\top X_{\text{tr}})_j^{-1})$ and that $\Delta\gamma_{-j} \sim \mathcal{N}(0, (\sigma^2 + \Sigma_{\text{cond}})({X_{\text{tr}}^{-j}}^\top X_{\text{tr}}^{-j})^{-1})$. Therefore, assuming the independence between both of them because they have been trained in independent training sets, we have that under the conditional null hypothesis the variance of the distribution of the bias will decrease quadratically. To remark so, we can use similar arguments as the ones used in the proof of Lemma 3.12 because $\text{tr}(\Sigma^{1/2}(X_{\text{tr}}^\top X_{\text{tr}})^{-1}\Sigma^{1/2}) \approx p/n_{\text{train}}$. □

E Assumptions and proof of Theorem 4.2

In this section we present the assumptions taken from Williamson et al. (2023) to proof the asymptotic efficiency. To do so, we first introduce the same notations as in this paper.

Denote $\mathcal{R} := \{c(P_1 - P_2) : c \in [0, \infty), P_1, P_2\}$ the linear space of finite signed measures generated by the class of distributions $\mathcal{M}$. In $\mathcal{R}$, the supremum norm $\|\cdot\|_\infty$ is the supremum difference of their distribution functions. The Gâteaux derivative of $P \mapsto \mathbb{E}_P[\ell(f(X), y)]$ at P_0 in the direction $h \in \mathcal{R}$ is denoted by $\dot{V}(f, P_0, h)$. Let's also define the random function $g_n : (X, y) \mapsto \dot{V}(\widehat{m}, P_0; \delta_{(X,y)} - P_0) - \dot{V}(m, P_0; \delta_{(X,y)} - P_0)$, where $\delta_{(X,y)}$ is the degenerate distribution on $\{(X, y)\}$. The assumptions are either deterministic (A) or stochastic (B):

A1 (*optimality*) there exists some constant $C > 0$ such that, for each sequence $\widehat{m}_1, \widehat{m}_2, \ldots \in \mathcal{F}$ such that $\|\widehat{m}_n - m\|_{\mathcal{F}} \to 0$, $|\mathbb{E}[\ell(\widehat{m}_n(X), y)] - \mathbb{E}[\ell(m(X), y)]| \leq C\|\widehat{m}_n - m\|_{\mathcal{F}}^2$ for each n large enough.

A2 (*differentiability*) there exists some constant $\delta > 0$ such that for each sequence $\epsilon_1, \epsilon_2, \ldots \in \mathbb{R}$ and $h, h_1, h_2, \ldots \in \mathcal{R}$ satisfying that $\epsilon_j \to 0$, and $\|h_j - h\| - \infty \to 0$, it holds that

$$\sup_{f \in \mathcal{F}: \|f - m\|_{\mathcal{F}} < \delta} \left| \frac{\mathbb{E}_{P_0 + \epsilon_j h_j}[\ell(f(X), y)] - \mathbb{E}_{P_0}[\ell(f(X), y)]}{\epsilon_j} - \dot{V}(f, P_0; h_j) \right| \to 0.$$

A3 (*continuity of optimization*) $\|m_{P_0 + \epsilon h} - m\|_{\mathcal{F}} = O(\epsilon)$ for each $h \in \mathcal{R}$ where $m_{P_0 + \epsilon h}$ is the minimizer in $\mathcal{F}$ of $f \mapsto \mathbb{E}_{P_0 + \epsilon h}[\ell(f(X), y)]$.

A4 (*continuity of derivative*) $f \mapsto \dot{V}(f, P_0; h)$ is continuous at m relative to $\|\cdot\|_{\mathcal{F}}$ for each $h \in \mathcal{R}$.

B1 (*minimum rate of convergence of* m) $\|\widehat{m}_n - m\|_{\mathcal{F}} = o_P(n^{-1/4})$.

B2 (*weak consistency*) $\int (g_n(X,y))^2 dP_0(X,y) = o_P(1)$.

First, we note that there is no need to include the limited complexity assumption, as we are already using a cross-fitted version. This is because the importance is estimated on a separate test dataset from the training set.

We also recall from Williamson et al. (2023) that Assumption A1 is referred to as *optimality* because we require a development that depends only on the second order, as the derivative tends to zero given that the model under consideration is a minimizer of the loss.

Similarly, as done in Williamson et al. (2023), we will prove the asymptotic efficiency of each empirical predictiveness measure with respect to its theoretical counterpart separately. This means that both the predictiveness measure using the information of the j-th covariate and the one without it will converge efficiently to their respective theoretical predictiveness measures. Therefore, all the previous assumptions need to be fulfilled for the restricted model. Additionally, we will require the minimum convergence rate not only of $\widehat{m}$, which was already required for the complete model convergence, but also for the model $\widehat{\nu}_{-j}$:

B3 (*minimum rate of convergence of* ν_{-j}) $\|\widehat{\nu}_{n,-j} - \nu_{-j}\|_{\mathcal{F}} = o_P(n^{-1/4})$.

We observe that all the above assumptions are exactly the ones required for the asymptotic efficiency of LOCO in Williamson et al. (2023), except for replacing the required minimum rate of convergence of $\widehat{m}_{-j}$ with that of $\widehat{\nu}_{-j}$.

We observe that these convergence rates are achieved by parametric models. They are standard in semiparametric inference. They are also achievable for nonparametric models under additional assumptions on dimensionality and regularity, and for standard machine learning models under mild assumptions.

We recall that, in our setting—the knockoffs framework (Candès et al. (2018))—it is easier to fulfill the latter condition than the former. This is because the relationship between the input and the output is expected to be more complex than the relationship among the inputs.

For a detailed discussion of these assumptions, see Williamson et al. (2023).

For this result, we also need local robustness to small changes in the argument of the function m. This condition is expressed by requiring m to be Lipschitz continuous. This ensures that the error made by estimating the argument does not explode.

Finally, we require Assumption 3.1 to hold for the validity of the conditional sampling step.

Theorem E.1. *Under Assumption 3.1, assuming that m is Lipschitz and assumptions A1-A4 and B1-B3, $\widehat{\psi}_{\mathrm{SCPI}}(j)$ is nonparametric efficient.*

Proof. Similarly as done in Williamson et al. (2023), the asymptotic efficiency will be established by decomposing each predictiveness measure. Indeed, we are going to first prove that

$$\frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \ell(\widehat{m}(x_i), y_i) \to \mathbb{E}\left[\ell(m(X), y)\right].$$

This comes directly from Theorem 2 of Williamson et al. (2023). Then, we need to proof this efficient asymptotic convergence

$$\frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \ell\left(\frac{1}{n_{\mathrm{cal}}} \sum_{l=1}^{n_{\mathrm{cal}}} \widehat{m}(\widetilde{x}_{i,l}^{(j)'}), y_i\right) \to \mathbb{E}\left[\ell(m_{-j}(X^{-j}), y)\right].$$

To achieve this, we will apply the same theorem. For this, we need to prove that the regressor converges

sufficiently fast to m_{-j}, specifically at the rate $o_P(n^{-1/4})$. Observe that

$$\begin{aligned}\left\|\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}\widehat{m}(\widetilde{X}_l^{(j)'})-m_{-j}(X^{-j})\right\| \leq &\left\|\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}\widehat{m}(\widetilde{X}_l^{(j)'})-\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}m(\widetilde{X}_l^{(j)'})\right\| \\ &+\left\|\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}m(\widetilde{X}_l^{(j)'})-\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}m(\widetilde{X}_l^{(j)})\right\| \\ &+\left\|\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}m(\widetilde{X}_l^{(j)})-m_{-j}(X^{-j})\right\|.\end{aligned}$$

We first note that the last quantity exhibits the desired convergence rate by applying the CLT, and because, under Assumption 3.1, $\widetilde{X}^{(j)}$ follows the desired distribution. For the first term, we have that

$$\left\|\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}m\left(\widetilde{X}_l^{(j)'}\right)-\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}m(\widetilde{X}_l^{(j)})\right\| \leq \max_l\|\widehat{m}(\widetilde{X}_l^{(j)'})-m(\widetilde{X}_l^{(j)'})\|.$$

Then, it has the correct convergence rate by using the convergence rate of $\widehat{m}$. Finally, for the second term, using that m is assumed L-Lipschitz, we have that

$$\begin{aligned}&\left\|\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}m(\widetilde{X}_l^{(j)'})-\frac{1}{n_{\text{cal}}}\sum_{l=1}^{n_{\text{cal}}}m(\widetilde{X}_l^{(j)})\right\| \\ &\quad\leq \max_i\|m(\widetilde{X}_l^{(j)'})-m(\widetilde{X}_l^{(j)'})\| \\ &\quad\leq L\max_i\|\widetilde{X}_l^{(j)'}-\widetilde{X}_l^{(j)'}\| \\ &\quad= L\max_i\|\widehat{\nu}_{-j}(X^{-j})+(X_i-\widehat{\nu}_{-j}(X_i^{-j}))-\nu_{-j}(X^{-j})-(X_i-\nu_{-j}(X_i^{-j}))\| \\ &\quad\leq L\left(\|\widehat{\nu}_{-j}(X^{-j})-\nu_{-j}(X^{-j})\|-\max_i\|\widehat{\nu}_{-j}(X_i^{-j})-\nu_{-j}(X_i^{-j})\|\right).\end{aligned}$$

We conclude by using Assumption B3. □

F Fixing n_{cal}: Proof of Proposition 4.3

Proof. The first part of the proof consists on applying the consistency of the estimates, continuous mapping theorem, and finally the Law of Large Numbers. Then, we develop the given expectation as

$$\begin{aligned}&\mathbb{E}\left[\left(y-\frac{1}{n_{\text{cal}}}\sum_{i=1}^{n_{\text{cal}}}m(\widetilde{X}_i^{(j)})\right)^2\right]-\mathbb{E}\left[(y-m(X))^2\right] \\ &=\mathbb{E}\left[\left(m(X)-\frac{1}{n_{\text{cal}}}\sum_{i=1}^{n_{\text{cal}}}m(\widetilde{X}_i^{(j)})\right)^2\right] \\ &=\frac{1}{n_{\text{cal}}^2}\mathbb{E}\left[\left(\sum_{i=1}^{n_{\text{cal}}}(m(X)-m(\widetilde{X}_i^{(j)}))\right)^2\right] \\ &=\frac{1}{n_{\text{cal}}^2}\sum_{i=1}^{n_{\text{cal}}}\mathbb{E}\left[(m(X)-m(\widetilde{X}_i^{(j)}))^2\right]+\frac{2}{n_{\text{cal}}^2}\sum_{i<k}\mathbb{E}\left[(m(X)-m(\widetilde{X}_i^{(j)}))(m(X)-m(\widetilde{X}_k^{(j)}))\right] \\ &=\frac{1}{n_{\text{cal}}}\mathbb{E}\left[(m(X)-m(\widetilde{X}_1^{(j)}))^2\right]+\frac{2}{n_{\text{cal}}^2}\sum_{i<k}\mathbb{E}\left[(m(X)-m(\widetilde{X}_i^{(j)}))(m(X)-m(\widetilde{X}_k^{(j)}))\right].\end{aligned}$$

For the second part, we observe that

$$\begin{aligned}&\mathbb{E}\left[\ (m(X)-m(\widetilde{X}_i^{(j)}))(m(X)-m(\widetilde{X}_k^{(j)}))\right] \\ &\quad=\mathbb{E}\left[m(X)(m(X)-m(\widetilde{X}_k^{(j)}))\right]-\mathbb{E}\left[m(\widetilde{X}_i^{(j)})(m(X)-m(\widetilde{X}_k^{(j)}))\right].\end{aligned}$$

The second term vanishes:

$$\begin{aligned}\mathbb{E}\left[m(\widetilde{X}_i^{(j)})(m(X)-m(\widetilde{X}_k^{(j)}))\right] &= \mathbb{E}\left[\mathbb{E}\left[m(\widetilde{X}_i^{(j)})(m(X)-m(\widetilde{X}_k^{(j)}))|X^{-j}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[m(\widetilde{X}_i^{(j)})|X^{-j}\right]\mathbb{E}\left[(m(X)-m(\widetilde{X}_k^{(j)}))|X^{-j}\right]\right] \\ &= 0.\end{aligned}$$

Now we observe that the first term is exactly the Total Sobol Index:

$$\begin{aligned}\mathbb{E}\left[m(X)(m(X)-m(\widetilde{X}_k^{(j)}))\right] &= \mathbb{E}\left[m(X)^2 - m(X)m(\widetilde{X}_k^{(j)})\right] \\ &= \mathbb{E}\left[m(X)^2\right] - \mathbb{E}\left[\mathbb{E}\left[m(X)m(\widetilde{X}_k^{(j)})|X^{-j}\right]\right] \\ &= \mathbb{E}\left[m(X)^2\right] - \mathbb{E}\left[\mathbb{E}\left[m(X)|X^{-j}\right]\mathbb{E}\left[m(\widetilde{X}_k^{(j)})|X^{-j}\right]\right] \\ &= \mathbb{E}\left[m(X)^2\right] - \mathbb{E}\left[m_{-j}(X^{-j})^2\right],\end{aligned}$$

and

$$\begin{aligned}\psi_{\mathrm{TSI}}(j,P_0) &= \mathbb{E}\left[(m(X)-m_{-j}(X^{-j}))^2\right] \\ &= \mathbb{E}\left[m(X)^2\right] - 2\mathbb{E}\left[m(X)m_{-j}(X^{-j})\right] + \mathbb{E}\left[m_{-j}(X^{-j})^2\right] \\ &= \mathbb{E}\left[m(X)^2\right] - 2\mathbb{E}\left[\mathbb{E}\left[m(X)|X^{-j}\right]m_{-j}(X^{-j})\right] + \mathbb{E}\left[m_{-j}(X^{-j})^2\right] \\ &= \mathbb{E}\left[m(X)^2\right] - \mathbb{E}\left[m_{-j}(X^{-j})^2\right].\end{aligned}$$

Therefore we have

$$\begin{aligned}\widehat{\psi}_{\mathrm{SCPI}}(j) &\xrightarrow{n_{\mathrm{train}},n_{\mathrm{test}}\to\infty} \frac{1}{n_{\mathrm{cal}}}\mathbb{E}\left[(m(X)-m(\widetilde{X}_1^{(j)}))^2\right] \\ &\qquad + \frac{2}{n_{\mathrm{cal}}^2}\sum_{i<k}\mathbb{E}\left[(m(X)-m(\widetilde{X}_i^{(j)}))(m(X)-m(\widetilde{X}_k^{(j)}))\right] \\ &= \frac{1}{n_{\mathrm{cal}}}2\psi_{\mathrm{TSI}}(j,P_0) + \frac{2}{n_{\mathrm{cal}}^2}\sum_{i<k}\psi_{\mathrm{TSI}}(j,P_0) \\ &= \frac{1}{n_{\mathrm{cal}}}2\psi_{\mathrm{TSI}}(j,P_0) + \frac{2}{n_{\mathrm{cal}}^2}\psi_{\mathrm{TSI}}(j,P_0)\frac{n_{\mathrm{cal}}(n_{\mathrm{cal}}-1)}{2} \\ &= \left(1+\frac{1}{n_{\mathrm{cal}}}\right)\psi_{\mathrm{TSI}}(j,P_0).\end{aligned}$$

□

G Inference: type-I error and power

Type-I error: In general, Verdinelli and Wasserman (2024) introduced an additional square root term to ensure control of the type-I error. Indeed, due to the quadratic nature of the statistic, under the null hypothesis, the variance vanishes as $\mathbb{V}(\widehat{\psi}) = O(n^{-\gamma})$ with $\gamma > 1$. Therefore, to maintain a valid type-I error rate, we observe that under the null hypothesis, using Chebyshev's inequality, we have that

$$\mathbb{P}_{\mathcal{H}_0}\left(\widehat{\psi} \geq z_\alpha \mathrm{se}_n + c/\sqrt{n}\right) \leq \frac{\mathbb{V}(\widehat{\psi})}{\left(z_\alpha \mathrm{se}_n + c/\sqrt{n}\right)^2} \to 0.$$

Improving the corrective term in linear settings: We observe that using Markov's inequality we have that

$$\mathbb{P}_{\mathcal{H}_0}\left(\widehat{\psi} \geq z_\alpha \mathrm{se}_n + c/\sqrt{n}\right) \leq \frac{\mathbb{E}(\widehat{\psi})}{(z_\alpha \mathrm{se}_n + c/\sqrt{n})}.$$

From Lemmas 3.11 and 3.12, we observe that in the linear setting, it is possible to use this last inequality to derive a more powerful valid test by changing the additive term $c/\sqrt{n}$ by $c/n^{-\gamma}$, with $\gamma < 1$ for LOCO and $\gamma < 2$ for CPI.

Power: We observe that under the alternative hypothesis, using Theorem 4.2, we have $\widehat{\psi} \xrightarrow{\text{a.s.}} \psi > 0$. In particular, this implies that

$$\mathbb{P}_{\mathcal{H}_1}\left(\widehat{\psi} \geq z_\alpha \mathrm{se}_n + c/\sqrt{n}\right) \underset{n\to\infty}{\to} 1.$$

Thus, the test is *consistent*, i.e. it has power approaching to one.

Standard deviation estimation: For se_n in the numerical experiments, we also explored using empirical variances derived from multiple estimations via bootstrapping on the test set, without refitting the model. Specifically, we computed several means over bootstrap samples on the test set and then estimate the variance among them. Also, the sample variance (which in this case coincides with the variance estimation with influence function as seen in Section H) divided by the sample size for the others, using the relation

$$\begin{aligned}\mathrm{var}(\widehat{\psi}_{\mathrm{SCPI}}) &= \mathrm{var}\left(\frac{n_{\mathrm{cal}}}{n_{\mathrm{cal}}+1}\cdot\frac{1}{n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}}\left[\left(\frac{1}{n_{\mathrm{cal}}}\sum_{k=1}^{n_{\mathrm{cal}}}\widehat{m}(\widetilde{x}_{i,k}^{\prime(j)})-y_i\right)^2-(\widehat{m}(x_i)-y_i)^2\right]\right)\\ &= \frac{1}{n_{\mathrm{test}}}\cdot\mathrm{var}\left(\frac{n_{\mathrm{cal}}}{n_{\mathrm{cal}}+1}\left[\left(\frac{1}{n_{\mathrm{cal}}}\sum_{k=1}^{n_{\mathrm{cal}}}\widehat{m}(\widetilde{x}_{i,k}^{\prime(j)})-y_i\right)^2-(\widehat{m}(x_i)-y_i)^2\right]\right).\end{aligned}$$

H MSE variance estimation with influence functions

From Theorem 4.2, it is possible to estimate the variance using the influence function. Indeed, the variance is given by

$$\tau_j^2 := \mathbb{E}\left[\varphi_j^2(X,y)\right],$$

where φ_j is the influence function of ψ_j, which is given by

$$\varphi_j : (X,y) \mapsto \dot{V}(m_{-j}, P_0, \delta_{(X,y)} - P_0) - \dot{V}(m, P_0, \delta_{(X,y)} - P_0).$$

Here, $\dot{V}(f, P_0; h)$ stands for the Gâteaux derivative of $P \mapsto \mathbb{E}_P\left[\ell(f(X), y)\right]$ in the direction $h \in \mathcal{R}$, with

$$\mathcal{R} := \{c(P_1 - P_2) : c \in [0,\infty), P_1, P_2 \in \mathcal{M}\}$$

being the finite signed measures generated by $\mathcal{M}$. Therefore, it is possible to estimate the variance as a simple plug-in:

$$\widehat{\tau}_j^2 := \frac{1}{n}\sum_{i=1}^{n}\left[\dot{V}(\widehat{m}_{-j}, P_n; \delta_{(X_i^{-j}, y_i)} - P_n) - \dot{V}(\widehat{m}, P_n; \delta_{(X_i,y_i)} - P_n)\right]^2.$$

In Appendix A of Williamson et al. (2023), they propose a simple method to compute the influence function when the predictiveness measure comes from standardized V-measures. We observe that, in general, computing the empirical variance and the variance using this plug-in version do not coincide. Nevertheless, in this section, we easily prove that they do coincide when using the MSE. Indeed, we observe that

$$
\begin{aligned}
\widehat{\tau}_j^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[\dot{V}(\widehat{m}_{-j}, P_n; \delta_{(x_i,y_i)} - P_n) - \dot{V}(\widehat{m}, P_n; \delta_{(x_i,y_i)} - P_n)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - \widehat{m}_{-j}(x_i^{-j})\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{m}_{-j}(x_i^{-j})\right)^2 - (y_i - \widehat{m}(x_i))^2 \right. \\
&\qquad \left. + \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{m}(x_i))^2\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - \widehat{m}_{-j}(x_i^{-j})\right)^2 - (y_i - \widehat{m}(x_i))^2 \right. \\
&\qquad \left. - \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - \widehat{m}_{-j}(x_i^{-j})\right)^2 - (y_i - \widehat{m}(x_i))^2\right]\right]^2 \\
&= \widehat{\mathrm{var}}\left(\ell(y, \widehat{m}_{-j}(x_i^{-j})) - \ell(y, \widehat{m}(x_i)\right).
\end{aligned}
$$

I Some lemmas used in the proofs

Lemma I.1 (Conditional null hypothesis)**.** *Under Assumption 3.7, the j-th covariate is independent of the output y conditionally on the rest of covariates if and only if there exists a measurable function $m_{-j} \in \mathcal{F}_{-j}$ such that $m(X) = m_{-j}(X^{-j})$.*

Proof. Firstly, we assume that $m(X) = m_{-j}(X^{-j})$, or equivalently, that $Y = m(X) + \epsilon = m_{-j}(X^{-j}) + \epsilon$. Therefore, using that ϵ is independent from X and that $m_{-j}(X^{-j})$ is constant conditionally on X^{-j}, then $y \perp\!\!\!\perp X^j | X^{-j}$.

To prove the other way, we first observe that

$$
\mathbb{E}\left[y^2|X^{-j}\right] = \mathbb{E}\left[(m(X)+\epsilon)^2|X^{-j}\right] = \mathbb{E}\left[m(X)^2|X^{-j}\right] + \sigma^2,
$$

using that ϵ is centered and independent of X. On the other hand, we observe that using the conditional independence and also that ϵ is centered and independent of X that

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

Then, we obtained that as both quantities are equivalent that $\mathbb{E}\left[m(X)^2|X^{-j}\right] = \mathbb{E}\left[m(X)|X^{-j}\right]^2$. We observe that Jensen's inequality with an strict convex function is only achieved with degenerate distributions. Therefore, $m(X)$ is $\sigma(X^{-j})$-measurable and therefore there exists a measurable function that we denote m_{-j} such that $m(X) = m_{-j}(X^{-j})$.

□

Lemma I.2 (LM for $y|X^{-j}$)**.** *Under Assumption 3.10 and Gaussian covariate, we have that $y = X^{-j}\beta' + \epsilon'$ with ϵ' independent from X^{-j} and $\epsilon' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 + \Sigma_{\text{cond}}{\beta^j}^2)$.*

Proof. Indeed, we can write

$$
\begin{aligned}
y &= X\beta + \epsilon \\
&= X^{-j}\beta_{-j} + X^j\beta^j + \epsilon \\
&= X^{-j}\beta_{-j} + (X^j - \nu_{-j}(X^{-j}) + \nu_{-j}(X^{-j}))\beta^j + \epsilon \\
&= X^{-j}\beta_{-j} + \nu_{-j}(X^{-j}) + \left((X^j - \nu_{-j}(X^{-j}))\beta^j + \epsilon\right).
\end{aligned}
$$

Using that X is Gaussian we have that $\nu_{-j}(X^{-j}) = X^{-j}\gamma_{-j}$, therefore, we can write $\beta' = \beta_{-j} + \gamma_{-j}$. Finally, we have that $(X^j - \nu_{-j}(X^{-j}))$ is independent of X^{-j} using the Gaussianity and that $(X^j - \nu_{-j}(X^{-j})) \sim \mathcal{N}(0, \Sigma_{\text{cond}})$. Therefore, we have that $\epsilon' = (X^j - \nu_{-j}(X^{-j}))\beta^j + \epsilon \sim \mathcal{N}(0, \sigma^2 + (\beta^j)^2\Sigma_{\text{cond}})$. □

J Additional experiments

J.1 Double Robustness of Sobol-CPI and nonnull bias

In this section, we study a linear Gaussian setting. In this case, both methods need to implement computationally similar procedures because $\widehat{m}$, $\widehat{m}_{-j}$, and $\widehat{\nu}_{-j}$ are all linear. We have $y = X\beta + \epsilon$ with $\beta_0 = 0$, $\beta_{1:p-1} \sim \mathcal{N}(0, 25I_{p-1})$, $\epsilon \sim \mathcal{N}(0,1)$, and $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.6^{|i-j|}$ and $p = 20$, we estimate $\widehat{\psi}_{\text{Sobol-cpi(1)}}$, $\widehat{\psi}_{\text{Sobol-cpi(100)}}$(i.e. with $n_{\text{cal}} = 100$) and $\widehat{\psi}_{\text{LOCO}}$ using linear models trained with $n_{\text{train}} = 50$ and averaged over $n_{\text{test}} = 100000$. On the top of Figure 4, the bias when estimating the importance on the null coordinate, while on the bottom, it corresponds to a nonnull coordinate. On the left, we use the same training set to estimate $\widehat{m}$, $\widehat{m}_{-j}$, and $\widehat{\nu}_{-j}$. Conversely, on the right, we use different training samples while maintaining $n_{\text{train}} = 50$. The histograms are based on 1000 runs.

Figure 4 illustrates that CPI converges faster under the conditional null hypothesis due to double robustness, and that a significant loss is incurred by using the data splitting method proposed in Williamson et al. (2023). However, we also observe that by increasing n_{cal}, the double robustness of CPI can still be preserved. Additionally, the bias distribution for the non-null covariates is similar across methods. Consequently, there is improved variable selection without sacrificing variable importance.

J.2 Effect of n_{cal}

In general, in order to get a valid variable importance that is asymptotically efficient, Theorem 4.2 shows that we need to take a calibration set size that increases with n. Nevertheless, as seen in Proposition 4.3, with $\ell = \ell_2$, it is possible to fix the calibration set size and correct the bias generated. Moreover, when n_{cal} is fixed to 1, as it is just a correction of CPI, it benefits from its double robustness, making it easier to separate null covariates from important ones. In this way, this *hyperparameter* represents a trade-off between *variable selection* and *variable importance*. With a large n_{cal}, we benefit from the asymptotic efficiency, obtaining a more robust estimate for the important covariates; however, for the null covariates, it yields a worse estimate. Indeed, standard CPI does not converge and is not asymptotically efficient when the importance is not null. This can be seen because the optimality assumption (Assumption A1 in Section E) does not hold.

To avoid the first-order contribution of having to estimate the regressor m, which is the maximizer of $f \mapsto \mathbb{E}[\ell(f(X), y)]$ over $\mathcal{F}$, it is reasonable to assume that

$$\frac{d}{d\epsilon}\mathbb{E}[\ell(m_\epsilon(X), y)]\big|_{\epsilon=0} = 0,$$

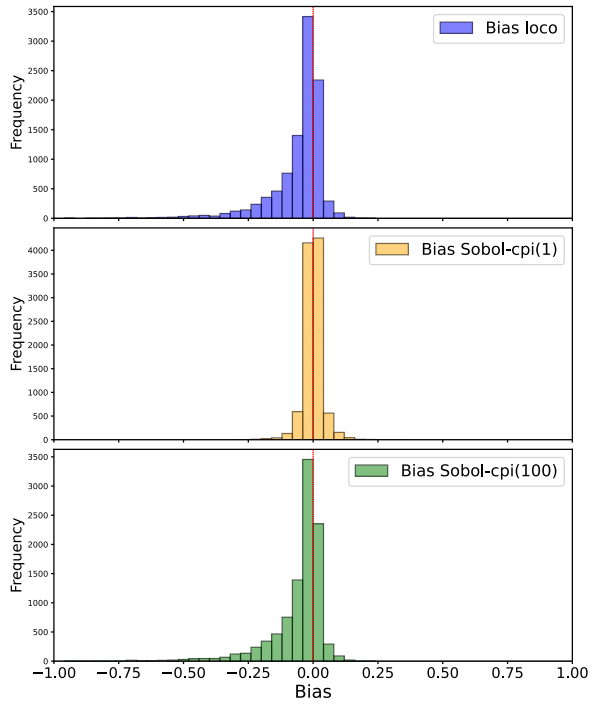
for any smooth path $\{m_\epsilon : -\infty < \epsilon < \infty\} \subset \mathcal{F}$. Nevertheless, this no longer holds for CPI. Indeed, we are not reoptimizing a learner with the empirical conditional distribution; we are only substituting the optimizer of the original distribution on another distribution. Therefore, this first-order term is not expected to vanish.

This trade-off between variable selection and variable importance can be observed, for instance, in Figure 5. In this Figure, we study the estimated importance of two covariates in a modified version of the nonlinear setting from Bénard et al. (2022): $y = X_0X_1\mathbb{I}_{X_2>0} + 2X_3X_4\mathbb{I}_{X_2<0}$, with $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma_{i,j} = \rho^{|i-j|}$, $p = 50$, $n = 10000$ and $\mu = \mathbf{0}$. The x-axis represents ρ, while the y-axis represents the estimated importance. The experiment is repeated 30 times.

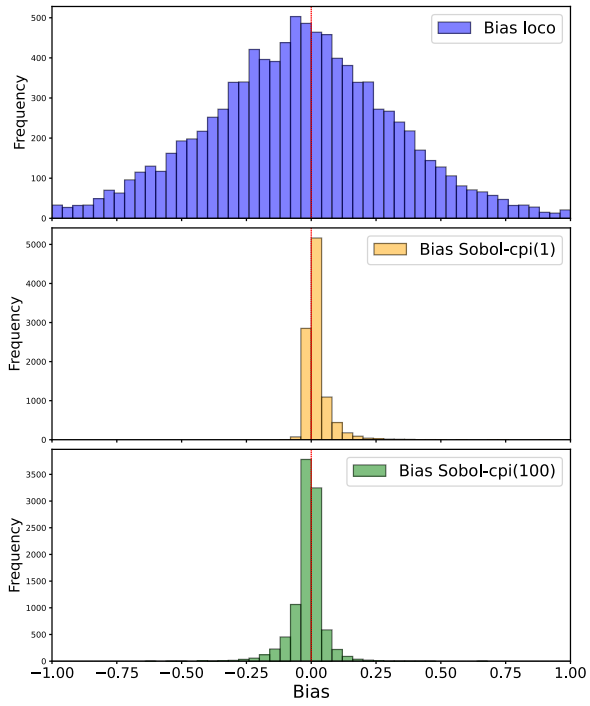
For the important covariate, we observe slightly better results in estimating importance. However, for the null covariates, the results are slightly worse when n_{cal} is larger. This is also remarkable in Figure 7.

J.3 Correlation

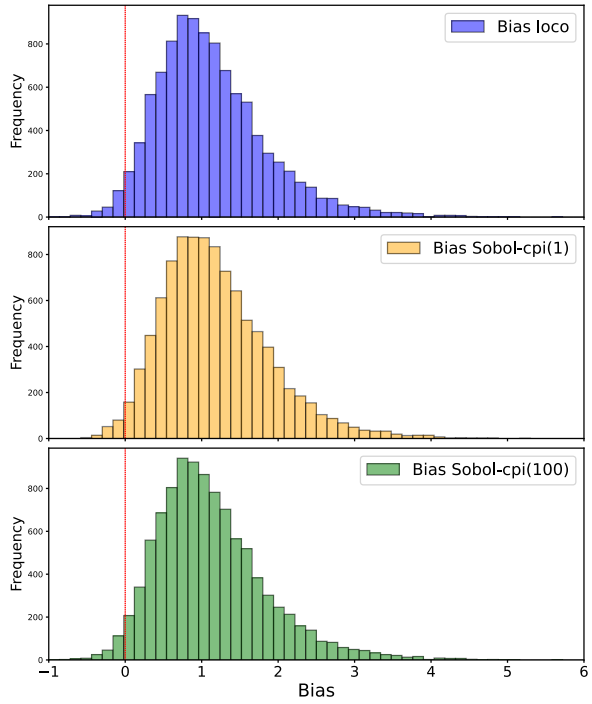
In this section, $\widehat{m}$ and $\widehat{m}_{-j}$ are Gradient Boosting models, while $\widehat{\nu}_{-j}$ is a Lasso model. In this experiments, $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma_{i,j} = \rho^{|i-j|}$, $p = 50$ and $\mu = \mathbf{0}$. The x-axis is for ρ.



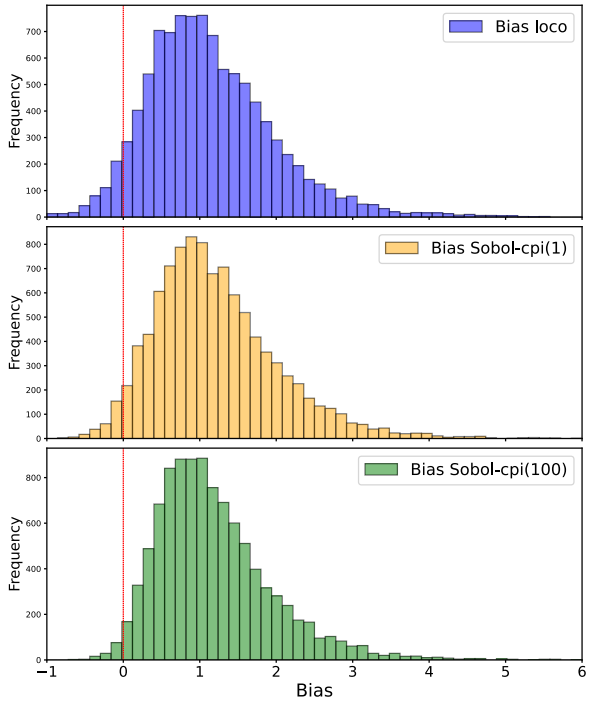
(a) Same training sample, null covariate



(b) Different training sample, null covariate



(c) Same training sample, nonnull covariate



(d) Different training sample, nonnnull covariate

Figure 4: **Double robustness of the Sobol-CPI:** The empirical bias distribution of LOCO, Sobol-CPI(1), and Sobol-CPI(100). From (a) and (b), we observe the benefits of using a CPI-based approach, as its double robustness results in lower bias. In (c) and (d), we see that the estimation error for a non-null covariate is similar. Comparing (a) and (b), we observe the negative effect of data splitting.

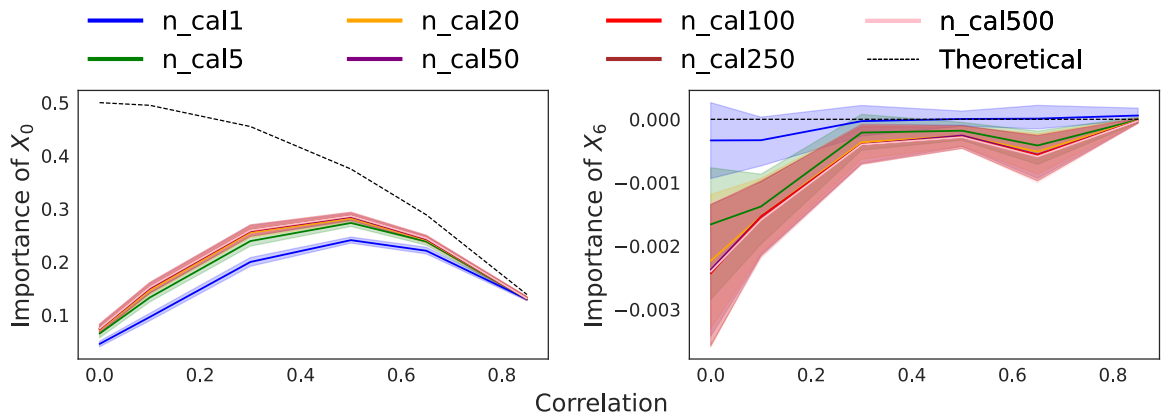


Figure 5: **Calibration set size effect as a trade-off between Variable Importance and Variable Selection:** Total Sobol Index estimation in a nonlinear setting. The first figure represents an important covariate (X_0), while the second represents a non-important covariate (X_6). We observe that with a larger n_{cal}, the importance estimation of the non-null covariate is slightly improved, enhancing variable importance. However, for the null covariate, there is a slightly greater bias, making variable selection less accurate.

In Figure 6, we examine a polynomial setting where the important covariates are randomly sampled with a sparsity of 0.1, with $n = 1000$. On the left the AUC and the right is the mean bias across the null covariates. This experiment is run over 30 repetitions. We observe that the AUC values are similar, but the LOCO method fails to achieve null importance for the null covariates. Additionally, the LOCO method is significantly more computationally expensive.

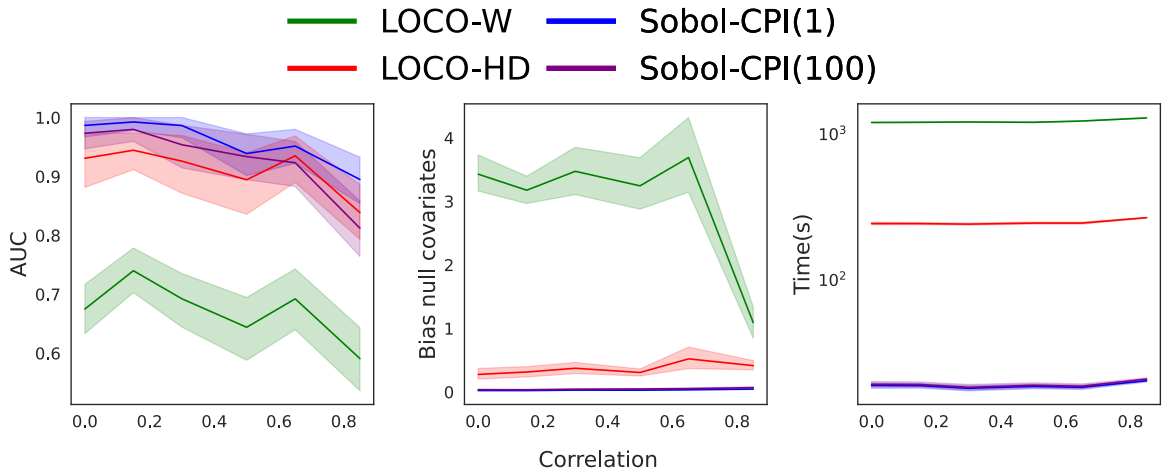


Figure 6: **Correlation Effect in a Polynomial Setting:** For all correlation values, Sobol-CPI achieves better discrimination of important covariates, assigns no importance to null covariates, and is significantly more computationally efficient.

In Figure 7, we use the same standard nonlinear setting: $y = X_0X_1\mathbb{I}_{X_2>0} + 2X_3X_4\mathbb{I}_{X_2<0}$, with $n = 10000$. On the top, the first two figures display the estimated importance of the important covariates, while the third figure represents an unimportant covariate. They are theoretically obtained in Theorem K.2. On the bottom, the left figure shows the AUC, the middle figure presents the mean bias across the null covariates, and the right figure is the computational time. This experiment is conducted over 10 repetitions.

The first two figures show that the importance scores from the Sobol-CPI method with a larger n_{cal} are more accurate, though still comparable to the LOCO method. From the third figure, we observe that Sobol-CPI maintains double robustness with larger n_{cal}, while LOCO exhibits a large bias.

In the bottom row, the first figure demonstrates the superior discriminant power of CPI-based methods. The second figure highlights the bias of the LOCO method for null covariates, and the third figure illustrates that CPI-based methods are much faster than the LOCO methods. This is because CPI-based methods avoid refitting a Gradient Boosting model for each covariate, using instead a Lasso model.

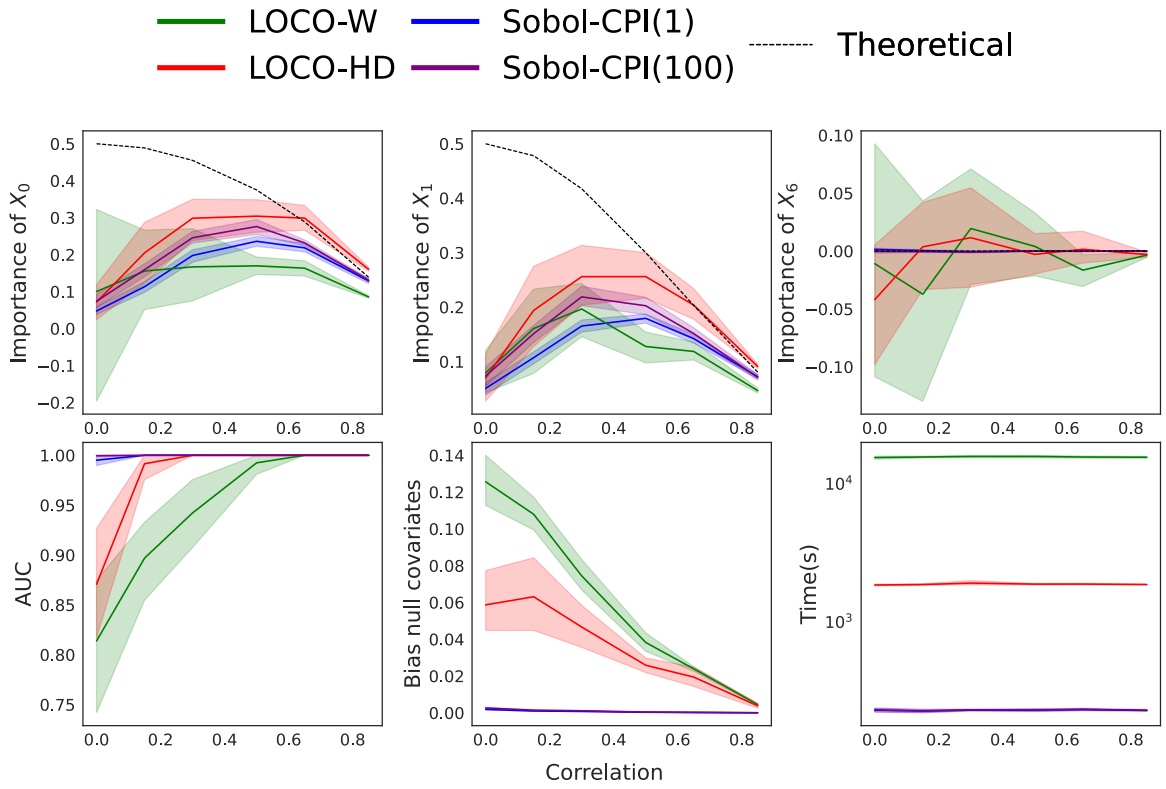


Figure 7: **Correlation effect in non-linear setting:** Sobol-CPI and LOCO achieve comparable results for important covariates while providing double robustness for null covariates in a significantly faster manner.

J.4 Inference

In this section, we compare the power and type-I error of different variance correction methods across LOCO and Sobol-CPI, using various values of n_{cal} in both linear and polynomial correlated settings.

As stated in Appendix G, different variance estimators have been studied: bootstrap-based estimates for the methods with the _bt and _n2 suffixes, and sample variance divided by the sample size for the others for the remaining methods.

When the method has no suffix (denoted by a circle marker), this indicates that no variance correction is applied. This approach was used by Chamma et al. (2024a), who ignored the vanishing variance; therefore, if the variance is zero, the null hypothesis is retained directly, without an statistical test. The suffix _sqrt denotes a square root correction term, _n and _bt indicate a linear correction term, and _n2 represents a quadratic correction term.

The experiments are run 100 times for the linear setting and 30 for the polynomial.

Linear setting: We study a linear setting with varying correlations ($\rho \in \{0.3, 0.6, 0.8\}$) to examine their effect on the power of the methods. However, since the results are similar across different values of ρ, we present only the graphics for $\rho = 0.6$.

In these experiments, $\widehat{m}$, $\widehat{m}_{-j}$, and $\widehat{\nu}_{-j}$ are linear models (see Theorem 3.2). More formally, $y = X\beta + \epsilon$ with β sparse with sparsity 0.25 and value 1 when non-null, $\epsilon \sim \mathcal{N}(0, \|X\beta\|^2/\text{snr})$ with snr $= 2$, and $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.6^{|i-j|}$ and $p = 100$.

From Figure 8, we observe that, despite the computational cost being similar across all methods due to the use of linear models (with the bootstrap-based covariance methods being more computationally intensive), Sobol-CPI demonstrates significantly smaller bias—not only for the null covariates but also for the non-null covariates.

We first note that `LOCO-W` (Williamson et al., 2023) requires a larger sample size to achieve type-I error control. Consequently, while LOCO-W is capable of identifying significant covariates, it also produces a substantial number of false positives. Furthermore, we observe that Sobol-CPI(1) is the most powerful method, benefiting from its double robustness property.

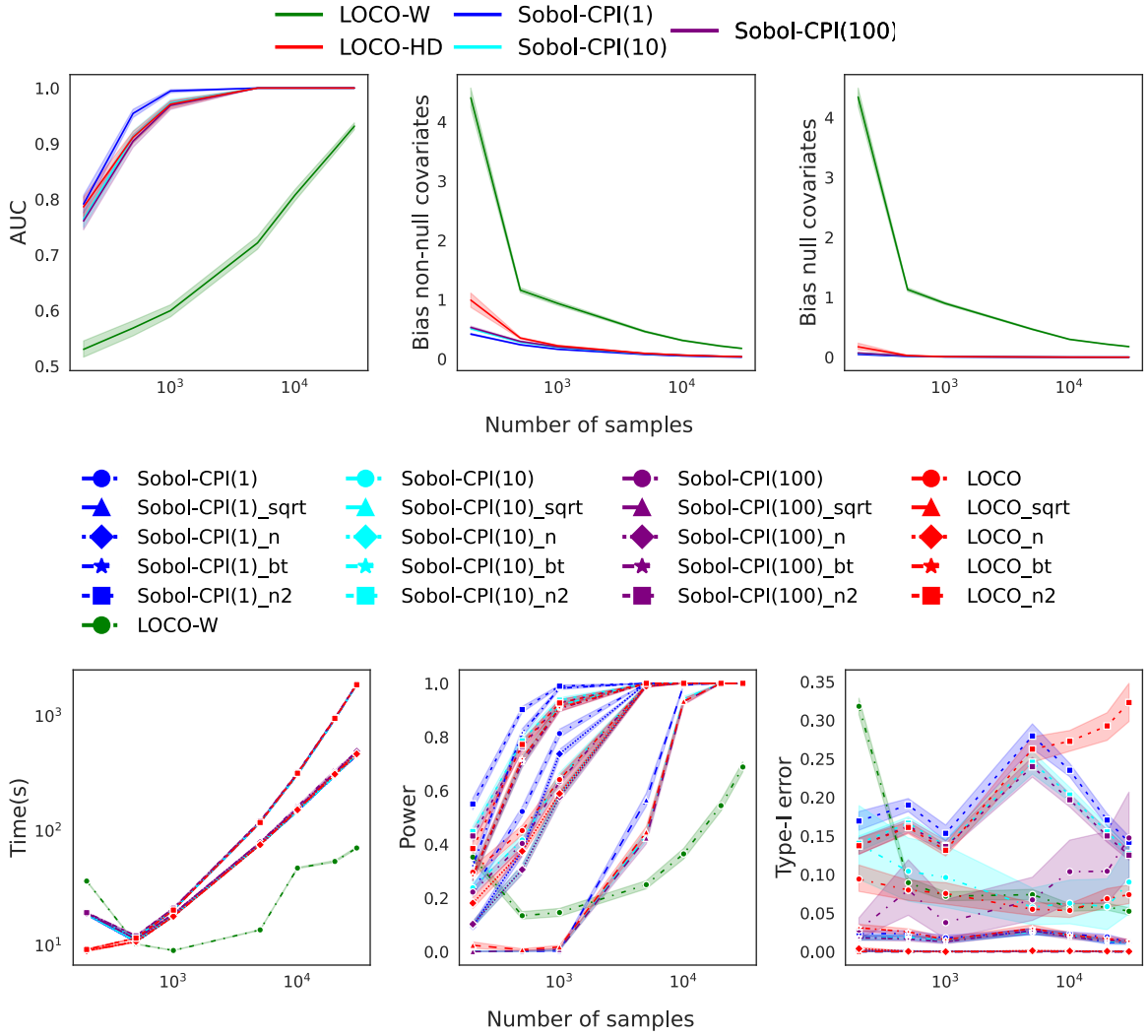


Figure 8: **General comparison of LOCO and Sobol-CPI in linear setting:** From left to right, and top to bottom, we have the AUC to detect the null covariates, the bias in estimating the non-null covariates, the bias in estimating the null covariates, the time to compute the estimates and the tests, the power of the tests, and the type-I error. Sobol-CPI presents lower bias, more power with a valid type-I error with a linear additive correction.

In Figures 9 and 10, we observe a detailed comparison of the type-I error and power, respectively. Among the methods, the most powerful tests involve the quadratic correction. Nevertheless, these do not control the type-I error. Additionally, the uncorrected method (circle marker) also fails to control the type-I error due to vanishing variance. As shown in Section J.4, the linear correction successfully preserves the type-I error.

We observe that the most powerful method is Sobol-CPI(1). When using a larger calibration size n_{cal}—which serves as a trade-off between variable selection and variable importance—the power slightly decreases due to the double robustness property.

Finally, we observe that in this case, there is a substantial gain when computing the variance using the bootstrap compared to using the sample variance in all the procedures. Both methods are valid, as the validity stems from the additive term, not from the specific choice of variance estimator.

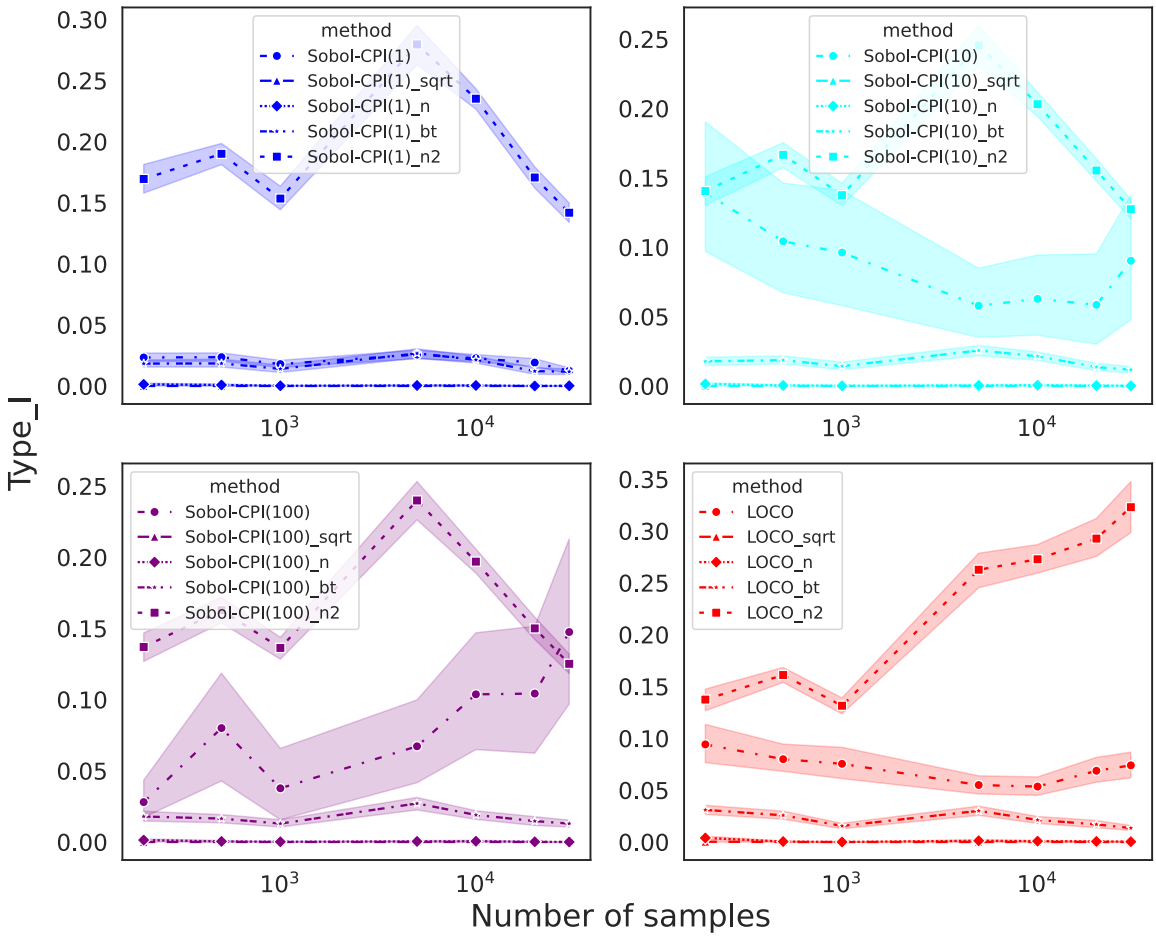


Figure 9: **Type-I error of Sobol-CPI and LOCO with different variance corrections in the Linear Setting:** The first three figures show different corrections applied to Sobol-CPI with varying n_{cal}, followed by the results for LOCO. The quadratic correction is not enough to control the type-I error.

Polynomial setting: We study a polynomial setting with varying correlations($\rho \in \{0.3, 0.6, 0.8\}$) to examine their effect on the power of the methods. Nevertheless, we only report the results for $\rho = 0.6$ as the other correlations give qualitatively similar results. In this experiment, the input matrix is generated as before ($X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = \rho^{|i-j|}$ and $p = 50$), but the output is polynomial in the input with interactions and degree three. The important covariates are randomly sampled with a sparsity of 0.25.

In all these experiments, $\widehat{m}$ and $\widehat{m}_{-j}$ are Gradient Boosting models, and $\widehat{\nu}_{-j}$ is a Lasso. Therefore, in this setting, there is a clear computational benefit to using a permutation-based approach, as seen in the bottom-left panel of Figure 11, where a substantial gain is observed. However, the benefits are not purely computational. At the top of the figure, we observe that the Sobol-CPI demonstrates superior classification performance, as indicated by a higher AUC. Additionally, it does not exhibit any bias on the null covariates, which is significant for the LOCO methods.

Regarding hypothesis testing, the conclusions are not entirely clear from this figure alone; we refer the reader to Figures 12 and 13 for a more detailed comparison. Nevertheless, it is evident that the Sobol-CPI methods are the most powerful, showing a clear separation in power from the LOCO methods. Moreover, the quadratic and linear corrections with bootstrap variance are not strict enough to control the type-I error. The `LOCO-W` method, in particular, fails to control the type-I error altogether—so even though it may yield discoveries, they are not

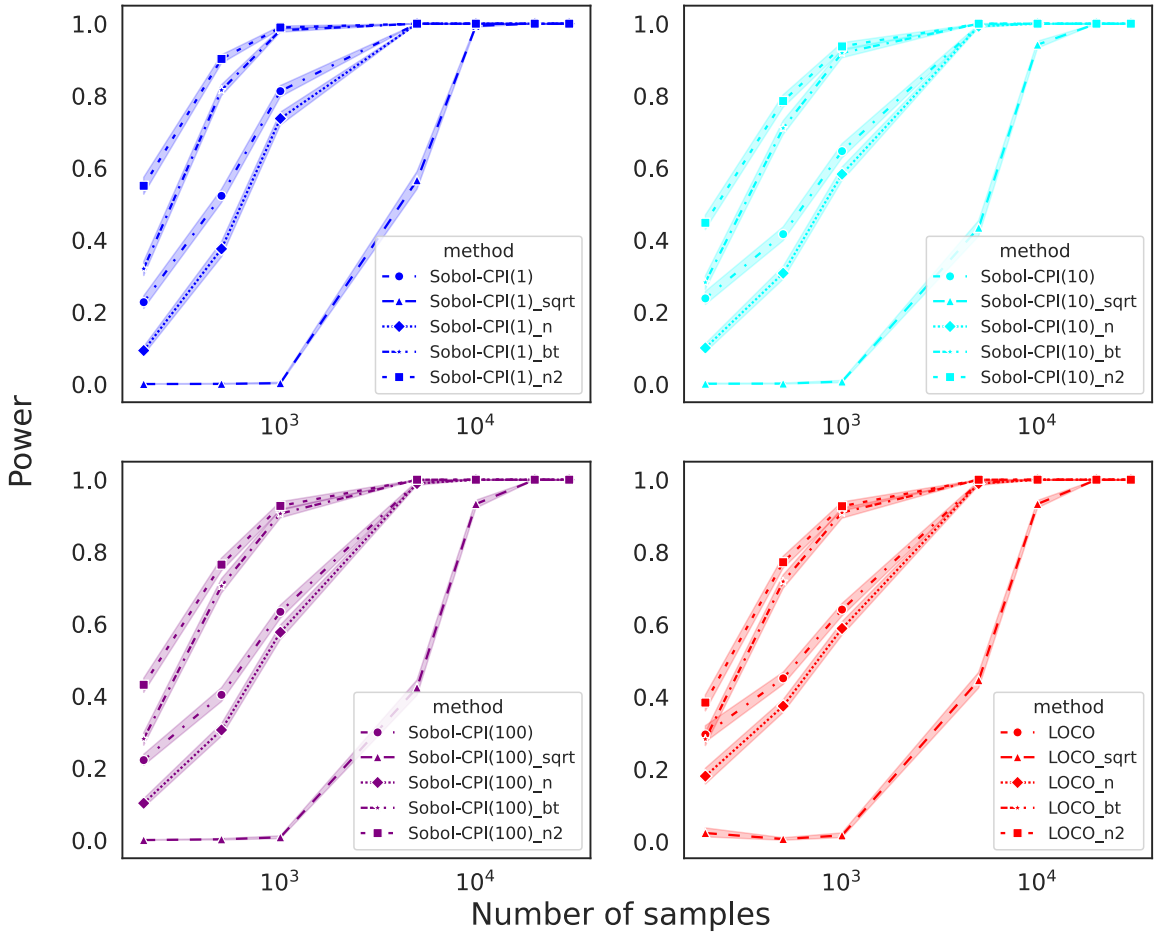


Figure 10: **Power of Sobol-CPI and LOCO with different variance corrections in the Linear Setting:** The first three figures show different corrections applied to Sobol-CPI with varying n_{cal}, followed by the results for LOCO. Sobol-CPI(1) is the most powerful method. Across the corrections, the linear decaying term is the least conservative, with variance estimated via bootstrap.

reliable due to a high number of false positives. This can also be explained by the AUC: even with a large sample size, there is no clear distinction between important and null covariates.

In Figure 12, we observe, similarly to the linear case, that the quadratic correction is not sufficient to control the type-I error. However, we also see that the linear correction with the variance estimated via bootstrap is not sufficient either. This contrasts with the linear setting, and the reason lies in the fact that the convergence rate result (see Theorem 3.11), which allows the use of Markov's inequality (see Section G) to guarantee type-I error control, is only available in the linear setting.

Additionally, we observe that for LOCO, even with the square root correction, the type-I error slightly exceeds the nominal level $\alpha = 0.05$, due to the high variability of the method. Finally, for Sobol-CPI, the square root correction does control the type-I error, but it may be overly conservative, as the observed error is exactly zero.

In Figure 13, we observe that Sobol-CPI(1) is the most powerful procedure, benefiting explicitly from double robustness. Additionally, across all methods, there is a decrease in power associated with more restrictive corrections. The gap observed with the square root correction arises from its overly conservative nature.

K Explicit Total Sobol Index examples

example K.1 (LM with Gaussian covariates). Given $X \sim \mathcal{N}(\mu, \Sigma)$ and $y = \beta X + \epsilon$ we note that

$$\begin{aligned}\psi_{\text{TSI}}(j, P_0) &= \mathbb{E}\left[(m(X) - m_{-j}(X^{-j}))^2\right] \\ &= \beta_j^2 \mathbb{E}\left[(X^j - \mathbb{E}\left[X^j | X^{-j}\right])^2\right] \\ &= \beta_j^2 \mathbb{E}\left[\mathbb{V}(X^j | X^{-j})\right] \\ &= \beta_j^2 \Sigma^j_{\text{cond}},\end{aligned}$$

with $\Sigma^j_{\text{cond}} := \Sigma_{j,j} - \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j}$.

example K.2 (Non-linear setting). In this example we will recover the example of no-linear setting from Bénard et al. (2022) but changing the input covariance matrix to obtain more complex relationships between the covariates. Indeed, we will have $y = \alpha X^0 X^1 \mathbb{I}_{X^2>0} + \beta X^3 X^4 \mathbb{I}_{X^2<0}$, where X is p-dimensional centered Gaussian with a Toeplitz covariance matrix where the i, j-th entry is given by $\rho^{|i-j|}$. In this setting, we are going to compute the $\psi_{\text{TSI}}(j, P_0)$ for the covariate X^0 and X_1.

First, we observe that

$$m_{-0}(X^{-0}) = \mathbb{E}\left[m(X)|X^{-0}\right] = \alpha \mathbb{E}\left[X^0 | X^{-0}\right] X^1 \mathbb{I}_{X^2>0} + \beta X^3 X^4 \mathbb{I}_{X^2<0}.$$

Then, we can develop LOCO as

$$\begin{aligned}\psi_{\text{TSI}}(0, P_0) &= \mathbb{E}\left[(m(X) - m_{-0}(X^{-0}))^2\right] \\ &= \mathbb{E}\left[\left(\alpha X^1 \mathbb{I}_{X^2>0}\left(X^0 - \mathbb{E}\left[X^0 | X^{-0}\right]\right)\right)^2\right] \\ &= \alpha^2 \mathbb{E}\left[(X^1)^2 \mathbb{I}_{X^2>0}\left(X^0 - \mathbb{E}\left[X^0 | X^{-0}\right]\right)^2\right] \\ &= \alpha^2 \mathbb{E}\left[(X^1)^2 \mathbb{I}_{X^2>0}\right] \mathbb{E}\left[\left(X^0 - \mathbb{E}\left[X^0 | X^{-0}\right]\right)^2\right]. \text{ using } X^0 - \mathbb{E}\left[X^0 | X^{-0}\right] \perp\!\!\!\perp X^{-0}\end{aligned}$$

The first term is exactly $\Sigma_{1,1}/2$. To see this, we first observe that as the covariates are centered and symmetrical, then $\mathbb{E}\left[(X^1)^2 \mathbb{I}_{X^2>0}\right] = \mathbb{E}\left[(X^1)^2 \mathbb{I}_{X^2<0}\right]$. Therefore, we have that

$$\Sigma_{1,1} = \mathbb{E}\left[(X^1 - \mathbb{E}\left[X^1\right])^2\right] = \mathbb{E}\left[(X^1)^2\right] = \mathbb{E}\left[(X^1)^2 (\mathbb{I}_{X_2>0} + \mathbb{I}_{X_2<0})\right] = 2\mathbb{E}\left[(X^1)^2 \mathbb{I}_{X_2>0}\right],$$

where we have used that $\mathbb{E}\left[X^1\right] = 0$. We also observe that as it is a Toeplitz matrix, $\Sigma_{1,1} = 1$. Then, $\psi_{\text{TSI}}(0, P_0) = \alpha^2/2\mathbb{E}\left[\left(X^0 - \mathbb{E}\left[X^0 | X^{-0}\right]\right)^2\right] = \alpha^2/2\mathbb{E}\left[\mathbb{V}(X^0 | X^{-0})\right]$. Note that as it is a Gaussian vector, the

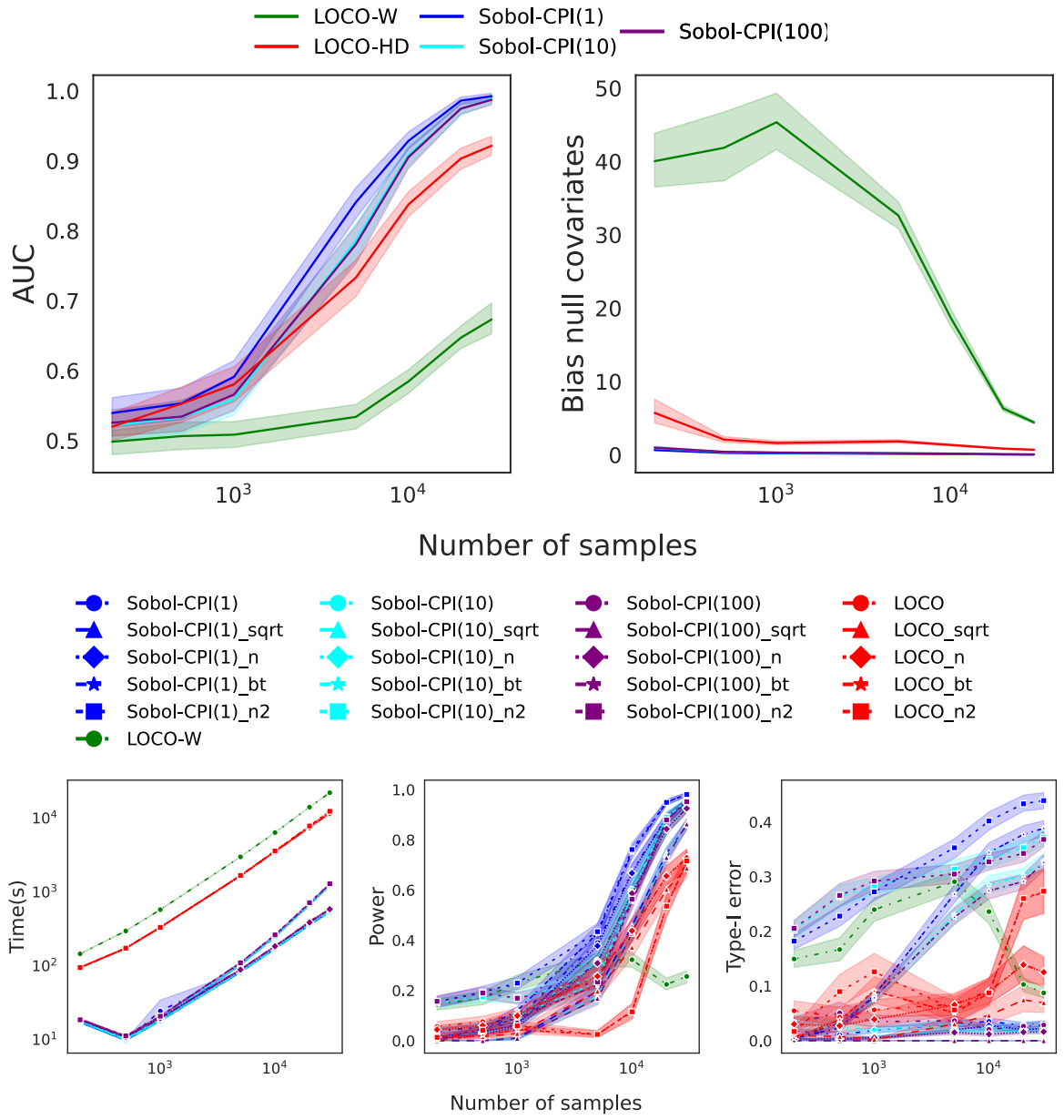


Figure 11: **General comparison of LOCO and Sobol-CPI in polynomial setting:** From left to right, and top to bottom, we have the AUC to detect the null covariates, the bias in estimating the null covariates, the time to compute the estimates and the tests, the power of the tests, and the type-I error. A greater gain is achieved with Sobol-CPI in these more complex settings, as it attains better results on the AUC without introducing bias on null covariates while requiring less computational effort.

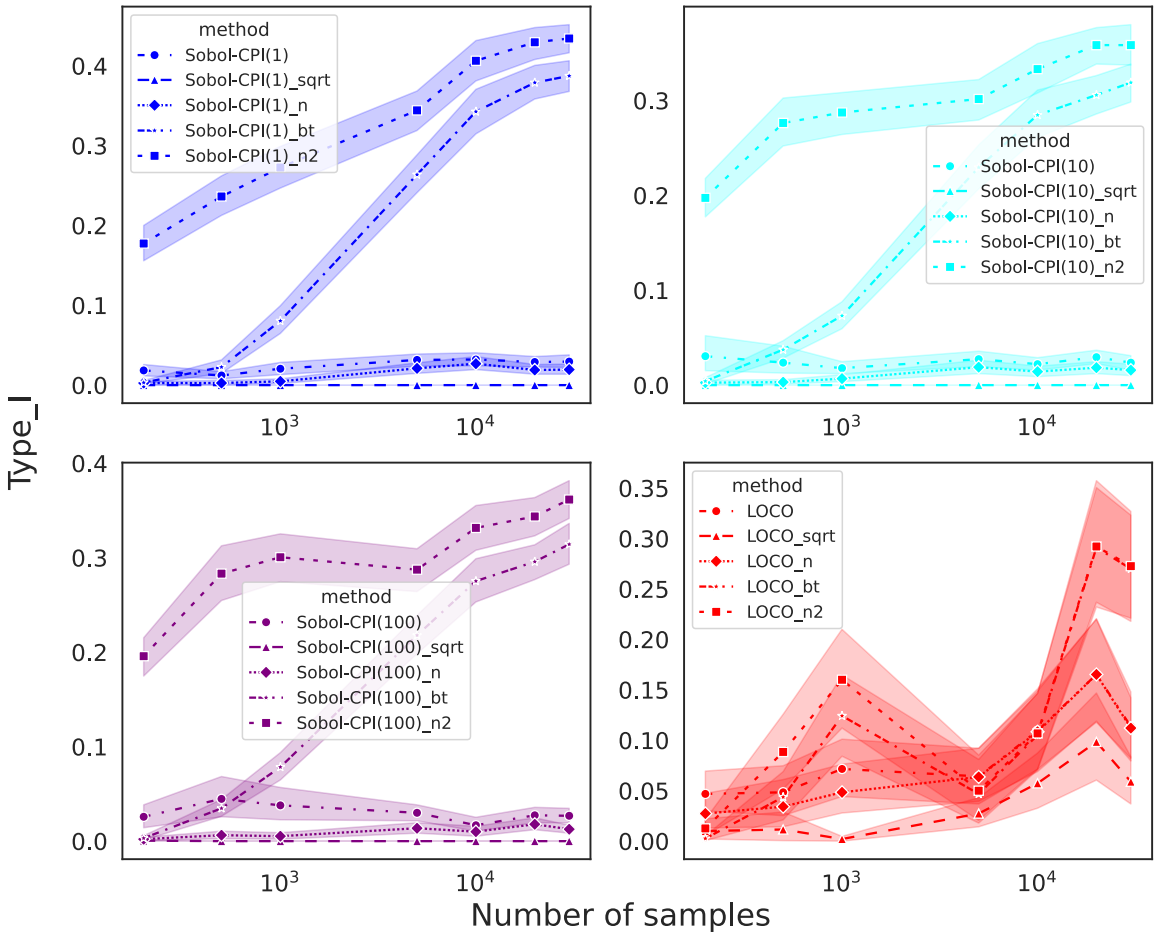


Figure 12: **Type-I error of Sobol-CPI and LOCO with different variance corrections in the Polynomial Setting:** The first three figures show different corrections applied to Sobol-CPI with varying n_{cal}, followed by the results for LOCO. The linear correction with bootstrap variance is not enough to control the type-I error.

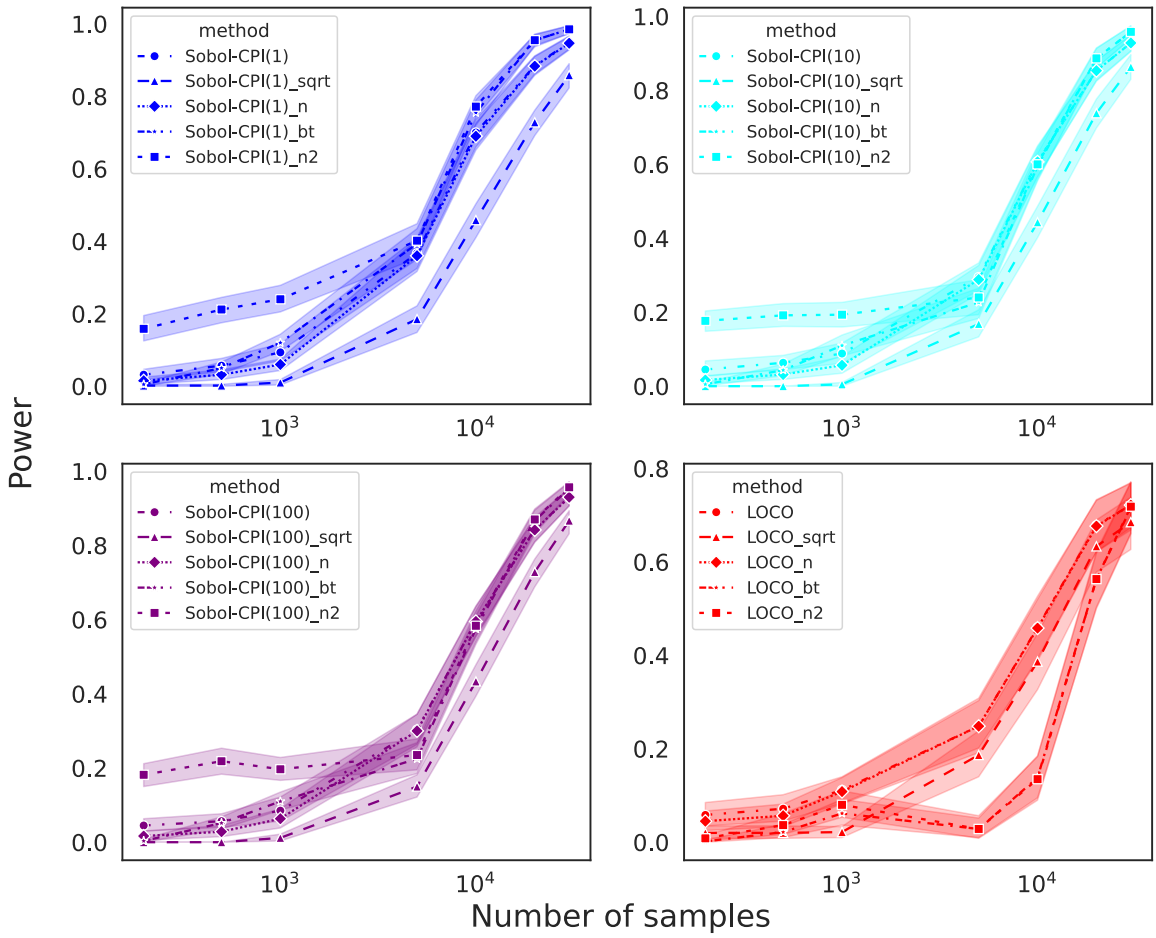


Figure 13: **Power of Sobol-CPI and LOCO with different variance corrections in the Polynomial Setting:** The first three figures show different corrections applied to Sobol-CPI with varying n_{cal}, followed by the results for LOCO. Sobol-CPI(1) is the most powerful method. Across the corrections, the linear decaying term is the least conservative, with variance estimated via bootstrap.

variance is exactly $\Sigma_{0,0} - \Sigma_{0,-0}\Sigma_{-0,-0}^{-1}\Sigma_{-0,0}$. We also observe that as it is a Toeplitz matrix, we have the property that $\Sigma_{-0,0} = \rho\Sigma_{-0,1} = \rho\Sigma_{-0,-0}(\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0})^\top$. Thus, we can develop the last term as

$$\begin{aligned}
\mathbb{E}\left[\mathbb{V}(X^0|X^{-0})\right] &= \Sigma_{0,0} - \Sigma_{0,-0}\Sigma_{-0,-0}^{-1}\Sigma_{-0,0} \\
&= 1 - \rho\Sigma_{0,-0}\Sigma_{-0,-0}^{-1}\Sigma_{-0,-0}(\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0})^\top \\
&= 1 - \rho\Sigma_{0,-0}(\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0})^\top \\
&= 1 - \rho^2.
\end{aligned}$$

Combining the previous, we conclude that, in this setting, $\psi_{\text{TSI}}(0, P_0) = (1 - \rho^2)/2$. Similarly, for the first covariate we obtain $\psi_{\text{TSI}}(1, P_0) = \rho^2/2(1 - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})$.