Mathématiques
Orsay

Statistical Inference for Variable Importance in High-Dimensional and Highly Correlated Settings

Angel David REYERO LOBO
Supervisors: Pierre NEUVIAL & Bertrand THIRION

Institut Mathématique d'Orsay(IMO)
Université Paris-Saclay

September 27, 2024

Abstract

Variable importance aims to assess the relevance of each input feature in predicting the output. Traditionally, variable importance has been a heuristic and non-rigorous methodology, often applied at the end of the machine learning pipeline to provide an idea of feature relevance. Furthermore, it has typically been model-dependent, making comparisons of feature importance across different algorithms impossible. Additionally, common variable importance measures struggle in the presence of correlated and high-dimensional settings.

This work aims to establish a more rigorous framework that offers a model-agnostic measure to address these gaps in the literature. First, we present a detailed comparison of the commonly used measures and algorithms. Then, we explore the connection between two key approaches: a permutation-based method and a removal-based method. The permutation-based approach, which includes a conditional sampling step, is computationally more feasible than the removal-based method, and we outline the assumptions under which this conditional step is valid. This provides a more stable method without the inference issues of the removal-based approach.

Finally, we turn to controlled variables selection. In the field of variable importance using global sensitivity analysis, it is common to heuristically suppress some covariates to obtain a minimal set of important variables, often without statistical guarantees. Furthermore, traditional conditional permutation importance only provides type-I error control, which is insufficient for high-dimensional settings. To address this, we propose a new modified version of conditional permutation importance within the knockoff framework that is capable of controlling the False Discovery Rate.

Keywords: variable importance, global sensitivity analysis, statistical inference, controlled variable selection, knockoffs, false discovery rate(FDR).

Notations. We use $\vee$ for the maximum ($a \vee b := \max(a, b)$)and $\wedge$ for the minimum ($a \wedge b := \min(a, b)$).

X^j denotes the j-th component of the random variable X with p covariates. X^{-j} denotes the remaining components. Similarly for $s \subset \{1, \ldots, p\}$ a subset of covariates, X^s stands for the s coordinates of X and X^{-s} for the rest.

We denote $\nu_{-j}(X^{-j}) = \mathbb{E}\left[X^j | X^{-j}\right]$ for the regression of the j-th covariate given X^{-j} and by $m(X) = \mathbb{E}\left[y|X\right]$ and $m_{-j}(X^{-j}) = \mathbb{E}\left[y|X^{-j}\right]$ for the regression of y given X and X^{-j} respectively. We use a hat for their empirical couterpart.

For a given input X and output y, $X^{(j)}$ denotes the random variable where the $-j$-th covariates are preserved and the j-th covariate is independent from the rest of coordinates but marginally preserves the same distribution. Therefore, $X^{-j} = X^{(j)-j}, X^{(j)j} \perp\!\!\!\perp X^{-j}$ and $X^{(j)j} \overset{\text{iid}}{\sim} X^j$. On the other hand, $\widetilde{X}^{(j)}$ is the random variable with conditionally independent j-th covariate to the rest of covariates. This means that all the coordinates are preserved but the j-th covariate, that not only preserves its marginal distribution but also its relationship with the rest of covariates. Nevertheless, conditionally on the rest of coordinates X^{-j}, X^j and $\widetilde{X}^{(j)j}$ are independent and identically distributed ($X^j \overset{\text{iid}}{\sim} \widetilde{X}^j | X^{-j}$). In this way, the link with the output y is broken. Therefore, we observe that in both cases, the coordinates $-j$ are equal, i.e. $X^{-j} = X^{(j)-j} = \widetilde{X}^{(j)-j}$, but is the j-th coordinate that changes.

We use $\overset{d}{=}$ to indicate equality in distribution. For $s \subset \{1, \ldots, p\}$, $(X, \widetilde{X})_{\mathrm{swap}(s)}$ denotes the vector obtained by swapping the coordinates X^j by $\widetilde{X}^j$ for $j \in s$.

Contents

1 Introduction

Due to the growing popularity of machine learning methods, it is increasingly important to decipher these *black-box* models to ensure they behave as expected. This means confirming that the algorithms prioritize relevant variables for prediction. For example, in a medical study of heart disease, it is crucial to verify that the model considers features such as blood pressure or cholesterol levels more significantly than irrelevant factors like eye color. This would indicate that the model is learning the underlying distribution of the data. Similarly, in a candidate application filter, it is important that discriminatory features such as race or gender are not considered to ensure fairness. Thus, there is a clear need for model transparency.

On the other hand, advancements in computational power and algorithm development make it desirable to use these models to understand the underlying data distribution. Modern machine learning models are capable of accurate predictions, and leveraging these models can help interpret data and provide valuable insights for further research. For instance, 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 address the disease.

The distinction between data transparency and model transparency often comes down to the difference between intrinsic variable importance and variable importance, or equivalently, feature importance and variable importance. This distinction is discussed further in Section 1.2.

1.1 Motivation

The importance of a variable is not a formal concept. Indeed, there are many properties that may be considered desirable, and not all of them are compatible. In this section, we will discuss some of the properties that we consider essential for defining a realistic and usable variable importance measure.

Model-agnostic approach: As variable importance has traditionally been a non-rigorous empirical approach, many methods have attempted to determine the importance of each covariate using specific models. For instance, in linear regression, the estimated coefficients for each covariate can be considered as measures of variable importance. Similarly, in tree-based approaches, impurity-based methods aim to assign each covariate an importance score based on the impurity decrease.

Additionally, there is often a trade-off between *model complexity* and *model transparency*, as simpler models tend to be more interpretable, as is the case with linear regression. However, these model-specific approaches do not align with the goal of measuring the predictive capacity of each covariate. In fact, the measures provided by model-specific approaches are not comparable across different models. As seen in Figure 1, a model that is transparent but too simple to capture the underlying data does not offer meaningful insights into the data-generating process. Therefore, we should consider a model-agnostic approach.

Computational feasibility: Some sensitivity analysis measures such as the Total Sobol Index can be considered desirable for quantifying the importance of each covariate. However, they often face inference issues that make their practical use challenging. For instance, these measures may exhibit bias and stability problems because they require refitting the global model while dropping each covariate, and the optimization error does not accumulate as expected. This is the primary limitation of the LOCO approach as it will be seen later.

Other methods, such as Shapley values, require refitting the model a combinatorial number of times, making them computationally infeasible to calculate directly. As a result, we aim to strike a

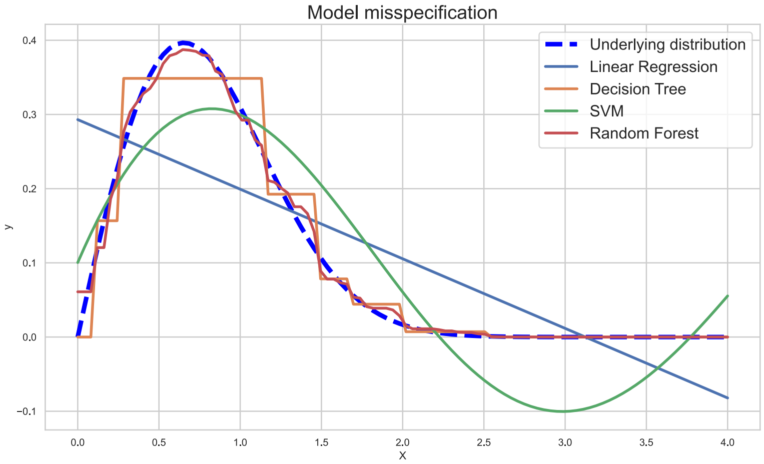

Figure 1: Given $n_{\text{train}} := 30$ of the underlying distribution $y = \sin(X)\exp(-X^2)$ we fit multiple models. Interpreting the underlying distribution with simple and transparent models may be misleading; therefore, we need a **model-agnostic** measure to avoid model misspecification only for the transparency.

balance between the amount of information extracted and the practical feasibility of the algorithm to ensure its usability in real-life applications, particularly in high-dimensional settings.

Statistical control: Sensitivity analysis tools are typically used for *factor screening*, meaning they select a minimal group of covariates that best explain the variance (see Breiman (2001); Bénard et al. (2022b)). However, this process is often done in a heuristic manner and without formal statistical control. Furthermore, when statistical control is provided (see Chamma et al. (2023), Williamson et al. (2021b)), it usually focuses on type-I error control, which becomes ineffective in high-dimensional settings (see Giraud (2021)).

Correlated settings: Not only do inference issues arise when covariates are highly correlated (see Verdinelli and Wasserman (2023)), but interpretation challenges also emerge in such cases (see Molnar et al. (2021)). For instance, a covariate might be mistakenly discovered as important simply because it is correlated with a truly relevant covariate. To address these issues, a conditional approach is required. In future research, grouping covariates may also provide a solution to these challenges.

1.2 Variable importance VS Feature importance

The distinction between variable importance and feature importance can be subtle. When discussing **variable importance**, we assume that we are trying to clarify the behavior of a black-box algorithm that operates on the data, but we lack insight into the covariates considered in predicting the outcome. Therefore, by assessing the importance of a covariate for the algorithm's outcome, we can achieve a more *interpretable AI* by shedding light on its black-box behavior. This approach is *model-agnostic* in that it should function with any model without specifying a particular trained model.

On the other hand, **feature importance** aims to assess a more intrinsic characteristic of the data. It is sometimes referred to as intrinsic variable importance. The goal here is to establish the relationship between the covariates and the outcome. Instead of looking for an interpretable AI we look for *AI to interpret.* Therefore, it is entirely *model-agnostic* in the sense that we do not need to impose any underlying model on the relationship between the outcome and the covariates.

We note that these goals do not always align. For instance, consider an outcome related to two highly correlated covariates. After training an algorithm with a variable selection method, only one of the two covariates might be used in the model, as we typically aim to minimize the number of selected covariates while maximizing predictive accuracy. Therefore, the aim of the first problem is to identify which covariate was selected, while the second problem seeks to recover both covariates. In this regard, we observe that the second goal should be more stable than the first one. The reason is that for some models or even instances of a model, one covariate might be selected while the other is not in the first problem, whereas in the second problem, the recovered solution should be more consistent. We observe that non-sparse models, e.g. deep neural networks, do not involve variable selection.

Another problem outlined in Molnar et al. (2021) is interpreting feature importance with a model that does not adequately fit the underlying distribution. In such cases, we may obtain misleading information about the true relevant covariates because they were not used in the poorly fitted model.

1.3 Marginal vs Conditional approach

Reasoning in marginal and conditional ways poses different questions. For instance, marginal approaches ask for the global importance of a covariate, while the conditional approach focuses on the importance of a covariate *given the others.* As we noted earlier, in the case of dependent covariates, a conditional approach is essential. Otherwise, the importance of covariates inflates, as many unimportant covariates are considered important simply due to their correlation with relevant covariates.

As noted in Candes et al. (2017), this distinction between marginal and conditional reasoning is significant not only with dependent covariates. For instance, in controlled variable selection, which aims to discard irrelevant covariates, even in the independent case where the same hypotheses are being tested, the conditional approach has higher power since the joint model has less residual variance than the marginal one, except in degenerate cases. Moreover, there are simple examples where marginal testing completely fails. For example, with independent covariates $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{Bern}(0.5)$ and output $Y = \mathbb{1}_{\{X_1+X_2=1\}}$, marginal testing would fail, whereas conditional testing would succeed.

There are also technical issues with marginal approaches. For example, as we will see later, when using permutation-based approaches, there are extrapolation issues because we do not preserve the relationship of the covariate with the others. In this case, we are attempting to predict individuals that do not follow the same distribution. Therefore, using a trained predictor for another distribution does not make sense.

For all the aforementioned reasons, a conditional approach will be taken.

1.4 Setting

We assume we are dealing with a regression problem, where $y \in \mathbb{R}$. Future work will aim to generalize this assumption by modifying the considered loss function.

We further assume that the output y can be expressed as a function of the input covariates, plus a centered random noise term that is independent of X.

Assumption 1 (Additive noise)**.** Given $y \in \mathbb{R}$ and $X \in \mathbb{R}^d$, $y = m(X) + \epsilon$ with $\epsilon \perp\!\!\!\perp X$ and $\mathbb{E}[\epsilon] = 0$.

This assumption mainly implies that we can extract all the information about the output through the function. We note that, for instance, in sensitivity analysis, it is often assumed that the output y can be directly explained by the input X (see Song et al. (2016)), which is clearly a stronger assumption.

As is common, this function m is unknown, and therefore we need to estimate it using machine learning algorithms. We assume access to a sample of data $(X_i, y_i) \overset{\text{i.i.d.}}{\sim} P_0$ for $i \in \{1, \ldots, n\}$. To quantify the importance of each covariate, a training-test split is typically applied: one training set is used to estimate the regressors, and a separate test set is used to assess the importance of the covariates based on the accuracy achieved.

1.4.1 On the assumptions of permutation approaches vs removal approaches

As we will see later, there are two main approaches to measure the importance of a covariate j. Both approaches aim to disable the information provided by the covariate (see Covert et al. (2020)). For the permutation approach we will primarily use, known as conditional permutation importance, we need to refit a regressor $\widehat{\nu}_{-j}$ for X^j given X^{-j}. In contrast, for removal approaches, we must refit a model $\widehat{m}_{-j}$ to regress y based on X^{-j}. This might seem paradoxical at first, as it raises questions about why we would do this and whether it's actually more efficient, especially since we claim that the permutation method is faster and more stable.

In reality, we are attempting to shift the *burden of knowledge*, similar to the assumptions made in Model-X knockoffs (see Candes et al. (2017)). Instead of making assumptions about the conditional distribution of y given X, which could be represented by regressing y using X^{-j}, we focus on the distribution of X, assuming that it is easier to regress among the input covariates. By doing so, we can use a computationally intensive, heavily trained model $\widehat{m}$, such as a deep neural network, just once. We can then employ faster and simpler models, such as random forests, to model the conditional regressors $\widehat{\nu}_{-j}$.

As stated in Candes et al. (2017), this approach could be interesting in several scenarios:

- When we have an unlabeled dataset apart from the labeled dataset used to train the general model. This may occur, for instance, when obtaining the data is cheap, but labeling the data is expensive.
- When we know the exact covariate distribution, such as in gene knockout experiments, genetic crossing experiments, sensitivity analysis of numerical models, and admixture mapping.
- When we have more initial information about the distribution of the input covariates, which is used to understand a more complex output. For example, when scientists use simple single nucleotide polymorphisms (SNPs) to understand a complex disease.

Thus, both approaches can be complementary in that we can choose to model the relationship between y and X, rather than just the distribution of the covariates. However, we want to emphasize that making assumptions about the relationship between the output y and the input X does not imply that this relationship is preserved when using the restricted input X^{-j}. For instance, in a logistic regression model, the restricted input does not maintain this structure with the output y (see Lobo et al. (2024)).

1.4.2 Conditional null hypothesis

We first observe that if some input covariates are functions of others, it leads to instability in the selection of important covariates, as there could be multiple valid sets of selected covariates. This occurs because the problem is not well-defined. To address this, we must assume that none of the covariates is a function of the others, ensuring that the set of genuine covariates is uniquely determined. This assumption is standard in controlled variable selection (see Candes et al. (2017)) and variable importance (see Verdinelli and Wasserman (2023)). Importantly, this assumption is not restrictive. In practice, we could assume that a small independent noise exists, making each covariate not an exact function of the others. Under this assumption, which breaks functional dependence, we can properly define the conditional null hypothesis:

Definition 1.1 (Conditional null covariate)**.** A covariate X^j is said to be conditionally null for predicting the output y given the other covariates X^{-j} if and only if y is independent of X^j conditional on X^{-j}. Otherwise, it is a genuine or non-null covariate.

We observe that we can easily reformulate this assumption by stating that the function m, which expresses the relationship between the input and the output, does not depend on the covariate. To formalize this, for a set of functions $\mathcal{F} := \{f : \mathbb{R}^p \to \mathbb{R}\}$, we define the set $\mathcal{F}_{-j} := \{f \in \mathcal{F} : f(u) = f(v), \forall u, v \in \mathbb{R}^p$ satisfying $u_{-j} = v_{-j}\}$.

Lemma 1.2 (Conditional null hypothesis)**.** *Under Assumption 1, 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 $m(X) = m_{-j}(X^{-j})$.

□

2 Related work

In this section, we present the main methods commonly considered for assessing variable importance. To gauge the predictive capacity of a feature, we compare the accuracy of the model when using the information provided by the covariate versus when excluding it. There are several ways to restrict the model from accessing the information provided by the feature. One simple approach is to retrain the model without this information, which leads to refitting-based approaches. However, in high-dimensional settings, this method often encounters problems, as it is computationally intensive and suffers the accumulation of estimation errors.

Other approaches simply disable the information provided by the covariate by breaking its relationship with both the output and input, resulting in permutation-based approaches. However, these methods suffer from extrapolation bias. In the following sections, we first present refitting-based approaches, followed by permutation-based approaches.

2.1 Variance-based global sensitivity analysis literature: LOCO

Sensitivity analysis attempts to study the fluctuations in the output when the input is perturbed. Specifically, local sensitivity analysis focuses on a specific input, while global sensitivity analysis considers all possible input values. Variance-based approaches try to assess the importance of each feature based on the portion of the variance it explains. Homma and Saltelli (1996) introduced two well-known quantities commonly used in this context. The first is the first-order effect, which represents the marginal variance explained by the feature. More formally, it is defined as:

Definition 2.1 (First-order effect)**.** For each $j \in \{1, \ldots, p\}$ and $(X, y) \sim P_0$,

$$\psi_{\text{marg}}(j, P_0) := \mathbb{V}(\mathbb{E}\left[m(X)|X^j\right]) = \mathbb{V}(m(X)) - \mathbb{E}\left[\mathbb{V}(m(X)|X^j)\right].$$

We observe that, being a marginal approach, it does not account for the interactions between covariates. For this reason, a conditional approach is needed to capture these interactions:

Definition 2.2 (Total effect)**.** For each $j \in \{1, \ldots, p\}$ and $(X, y) \sim P_0$,

$$\begin{aligned}
\psi_{\text{LOCO}}(j, P_0) &:= \mathbb{E}\left[(y - m_{-j}(X^{-j}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right] && (1)\\
&= \mathbb{E}\left[(m(X) - m_{-j}(X^{-j}))^2\right] && (2)\\
&= \mathbb{E}\left[\mathbb{E}\left[(m(X) - \mathbb{E}\left[m(X)|X^{-j}\right])^2|X^{-j}\right]\right] && (3)\\
&= \mathbb{E}\left[\mathbb{V}(m(X)|X^{-j})\right]. && (4)
\end{aligned}$$

We observe that to go from (1) to (2) we have just used Assumption 1, and to go from (2) to (3) we have used Tower's property and that $m_{-j}(X^{-j}) = \mathbb{E}\left[y|X^{-j}\right] = \mathbb{E}\left[\mathbb{E}\left[y|X\right]|X^{-j}\right] = \mathbb{E}\left[m(X)|X^{-j}\right]$. We observe that this is a well-studied quantity, known as the unnormalized total Sobol index (see Bénard et al. (2022b)) and as LOCO (Leave One Covariate Out / Leave Out Covariates) (see Williamson et al. (2021a), Verdinelli and Wasserman (2023)). It can also be viewed as a generalization of ANOVA, as it can be rewritten as:

$$\begin{aligned}
\psi_{\text{LOCO}}(j, P_0) &= \mathbb{E}\left[(y - m_{-j}(X^{-j}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right]\\
&= \mathbb{V}(y)\left(\left(1 - \frac{\mathbb{E}\left[(y - m(X))^2\right]}{\mathbb{V}(y)}\right) - \left(1 - \frac{\mathbb{E}\left[(y - m_{-j}(X^{-j}))^2\right]}{\mathbb{V}(y)}\right)\right),
\end{aligned}$$

which is the R^2 difference whenever the covariate is used or not. In the following, we will primarily refer to this as LOCO.

LOCO is particularly popular for feature screening, which aims to select the smallest set of features that explain the output. However, estimating LOCO in practice is challenging. A simple plug-in approach is not effective. On the one hand, it requires a computationally intensive algorithm, as we would need to retrain a model for each covariate, which is costly. On the other hand, using flexible algorithms does not yield good results (see Williamson et al. (2021a)).

First, semi-parametric theory indicates that a one-step correction is necessary to obtain asymptotically optimal estimates when applying the plug-in approach in (2) (see Williamson et al. (2021a)). This one-step correction is not required when applying the plug-in in (1). In practice, sequential regression is also necessary due to estimation noise, meaning that m_{-j} is no longer trained on (X, y) but rather on $(X, \widehat{m}(X))$. Even with this heuristic correction, performance is poor in practice, as shown in Section 3.4, due to the accumulation of estimation noise.

Moreover, valid p-values for testing the null hypothesis do not exist. Since the LOCO statistic is a quadratic functional, its influence function vanishes as it approaches zero (see Williamson et al. (2021a), Verdinelli and Wasserman (2023)). Williamson et al. (2021b) proposed a valid null-testing approach based on cross-fitting, which works because, even though the joint influence function in (1) vanishes, the influence functions of each part do not vanish, except in degenerate cases. However, in practice, this approach does not perform well (see Section 3.4).

2.2 Shapley values

Marginal approaches do not perform well in highly correlated settings because they may suffer from many false positives due to irrelevant covariates being highly correlated with relevant covariates. Conditional approaches do not face this issue. However, they may struggle with making any discoveries. Indeed, when two relevant covariates are highly correlated, conditional methods such as LOCO may incorrectly conclude that neither is important. This happens because, with one covariate, all the information can be recovered, making the other appear irrelevant. To address this, there has been a push to use Shapley values, which were introduced by Owen (2014) from game theory. The goal is to assign each player a fair contribution to the game's outcome. Instead of simply comparing *all covariates vs. all covariates without the one being measured*, Shapley values account for a broader range of relationships between covariates by assessing performance decay across all possible combinations of covariates with the one being measured. More formally:

Definition 2.3 (Shapley values)**.** For each $j \in \{1, \dots, p\}$ and $(X, y) \sim P_0$,

$$\psi_{\text{Shap}}(j, P_0) := \sum_{s \subset j\{j\}} w_s \left(\mathbb{E}\left[(y - m_s(X^s))^2 \right] - \mathbb{E}\left[\left(y - m_{s \cup \{j\}}(X^{s \cup \{j\}}) \right)^2 \right] \right), \tag{5}$$

with $w_s = \frac{1}{p} \binom{p-1}{|s|}^{-1}$.

Therefore, we observe that Shapley values are a weighted sum of LOCO values across the submodels, distributing the signal among correlated features. However, this approach introduces several estimation challenges. Not only do we face the issue of refitting multiple models, but there is also a combinatorial complexity since performance needs to be compared across the submodels. In theory, it is not necessary to compute all combinations, as Williamson and Feng (2020) demonstrated that averaging over a linear number of subsets is sufficient.

To address the combinatorial issue avoiding the brute force algorithms, it is possible a Monte Carlo sampling (see Lundberg and Lee (2017), Williamson and Feng (2020)). For instance, Bénard et al. (2022a) introduced a Monte Carlo approach using importance sampling, where the prior is informed by the most important covariates identified by a Random Forest. To mitigate the challenge of refitting the conditional expectation, some approaches assume a known distribution (see Song et al. (2016)), while others, like SHAFF uses projected random forests (see Bénard et al. (2022a)).

Nevertheless, Shapley values suffer from mathematical issues, and their game-theoretic interpretation may not align with human-centered interpretation (see Kumar et al. (2020)). Moreover, the preference between LOCO and Shapley values is debatable (see Verdinelli and Wasserman (2023)).While it is often claimed that Shapley values address the correlation issues that LOCO does not, this is not entirely accurate. As shown in Verdinelli and Wasserman (2023), a simple linear model was proposed where the covariates are highly correlated, but only the first covariate is relevant. In this case, the importance is spread across all covariates, and as the dimensionality increases, the importance diminishes. One could argue that this example represents a degenerate case since, in the presence of such high correlation, the covariates might also need to be considered important. Furthermore, Shapley values are often defended because they satisfy certain axiomatic properties. However, these axioms may not be as desirable as one might think.

First, the axiom of linearity requires that the sum of all covariates' importances equals the total value of the game. This may not be desirable when covariates are highly correlated, in which the value of two covariates together should not be the sum of each individual importance (see Verdinelli and Wasserman (2023)). Neither it is desirable when using a function m to explain y using X that is not linear (see Kumar et al. (2020)). Additionally, Shapley values are not always easily interpretable. For this reason, Verdinelli and Wasserman (2023) proposed alternative axioms that may be more appealing, such as ensuring that each covariate's importance is assessed independently of its correlation with other covariates. They also introduced another measure that satisfies these axioms, but it suffers from extrapolation bias, as it may make predictions in regions of low data density, where the model is not well-trained, leading to inconsistent results. As we will see later, this is the exact issue that the Permutation Feature Importance (PFI) from Breiman (2001) method encounters.

Moreover, the key limitations of both LOCO and Shapley values discussed in Verdinelli and Wasserman (2023) are in degenerate cases where one covariate is a function of others. In our setting (see Section 1.4), we avoid such cases, which is why we will continue to use theoretical LOCO as a reliable measure of importance in this work.

These earlier approaches relied on refitting multiple models to assess importance. In the following sections, we will explore permutation-based methods, which do not require refitting the output in response to any particular input, but instead study the relationships between inputs.

2.3 Permutation Feature Importance(PFI)

One naive method to assess the importance of covariate j involves disabling the information it provides. To achieve this without refitting the entire model, we can permute the covariate, thereby breaking its relationship with both the other inputs and the output. This allows us to study the resulting shift in accuracy. In practice, this is simply given by

$$\widehat{\psi}_{\mathrm{PFI}}(j, P_0) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left((y_i - \widehat{m}(\boldsymbol{x}_i^{(j)}))^2 - (y_i - \widehat{m}(\boldsymbol{x}_i))^2 \right), \tag{6}$$

where the j-th covariate has been completely permuted.

This quantity is also known as Mean Decrease Accuracy(MDA). Theoretically, it tries to achieve the following quantity:

Definition 2.4 (PFI)**.** For each $j \in \{1, \ldots, p\}$ and $(X, y) \sim P_0$,

$$\psi_{PFI}(j, P_0) := \mathbb{E}\left[(y - m(X^{(j)}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right]. \tag{7}$$

Although it is an efficient algorithm because it avoids refitting, and Mi et al. (2021) claims good performance and statistical validity, this is not the case. Bénard et al. (2022b) studied permutation-based variable importance in the context of Random Forests, comparing different estimation methods used by various software packages. These methods differ, for instance, in whether they use a separate test set to measure importance or rely on the Out-Of-Bag (OOB) sample, and whether the importance is computed across individual trees or the entire forest. They showed that the quantity targeted by PFI (7) is not ideal. It can be decomposed into three components. The first part, which is desirable, is the unnormalized total Sobol index ψ_{LOCO} introduced in Definition 2.2. The second is the unnormalized marginal total Sobol index, which can be misleading in the presence of highly correlated covariates. Lastly, there is a third term that is also problematic, as it increases with covariate dependence. Overall, this theoretical quantity is not what we truly want, and they proposed Sobol-MDA, which aims to directly estimate the total Sobol index using projected forests.

Additionally, we note that when permuting covariate j, both the marginal distributions of X^j and X^{-j} are preserved, but the joint distribution is altered. In assessing the predictive importance of a covariate, we do not necessarily want to change its relationship with the other covariates but rather with the output only. Moreover, using a permuted covariate for prediction can lead to *extrapolation bias*, as it may involve predicting in regions of low density, where the model is not well-trained and where such predictions are nonsensical (see Chamma et al. (2023), Verdinelli and Wasserman (2023)). For instance, we could end up trying to predict the weight of a baby measuring two meters in height. Therefore, the goal of the following method is primarily to break only the relationship with the output while permuting in such a way that the new samples follow the same distribution as the input covariates.

2.4 Conditional Permutation Importance(CPI)

In this section, we discuss the method introduced by Chamma et al. (2023). When assessing the importance of covariate j, this approach aims to preserve its relationship with the other input covariates while breaking only its relationship with the output. To achieve this, they proposed two methods. The first involves regressing X^j on X^{-j} and then adding a permuted residual to this regression. More explicitly, for an input x_i, given another input x_l, the conditional counterpart for covariate j is the vector $\widetilde{x}_i^{(j)}$, where all covariates are preserved except for the j-th covariate, which is modified as $\widetilde{x}_i^{(j)j} = \widehat{\nu}_{-j}(x_i^{-j}) + (x_l^j - \widehat{\nu}_{-j}(x_l^{-j}))$. The second method consists on training a random forest to estimate X^j using X^{-j} and then sample from the leaf. Once this conditional counterpart is computed across the test set, the importance measure is given by

$$\widehat{\psi}_{\text{CPI}}(j, P_0) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(\left(y_i - \widehat{m}(\widetilde{\boldsymbol{x}}_i^{(j)}) \right)^2 - (y_i - \widehat{m}(\boldsymbol{x}_i))^2 \right). \tag{8}$$

The theoretical part that it tries to attain is:

Definition 2.5 (CPI)**.** For each $j \in \{1, \dots, p\}$ and $(X, y) \sim P_0$,

$$\psi_{CPI}(j, P_0) := \mathbb{E}\left[(y - m(\widetilde{X}^{(j)}))^2 \right] - \mathbb{E}\left[(y - m(X))^2 \right], \tag{9}$$

where $\widetilde{X}^{(j)} \perp\!\!\!\perp X | X^{-j}, \widetilde{X}^{(j)-j} = X^{-j}$ and $X \sim X^{(j)}$.

We first observe that (9) introduces a new quantity that is not related to LOCO at a first sight. The link between both quantities will be established in Section 3.2.

PFI vs CPI: Both PFI and CPI are permutation-based approaches. Chamma et al. (2023) described the difference as PFI having an estimation error, which prevents it from achieving the same quantity as CPI due to correlation and the high dimensionality of the problems. However, this is not uniquely due to estimation issues; rather, the theoretical quantities themselves aim to explain different aspects. Indeed, we will demonstrate this in a simple linear setting.

example 2.6 (PFI vs CPI)**.** *Given a linear model $y = \sum_{i=1}^{p} \beta_i X^i + \epsilon$ where ϵ is centered, with variance σ and independent of the covariates, we observe that*

$$\begin{aligned}
\psi_{\mathrm{PFI}}(j, P_0) &= \mathbb{E}\left[(y - m(X^{(j)})^2\right] - \mathbb{E}\left[(y - m(X))^2\right] \\
&= \mathbb{E}\left[(\sum_{i=1}^{p} \beta_i X^i + \epsilon - \sum_{i \neq j} \beta_i X^i - \beta_j X'^j)^2\right] - \sigma^2 \\
&= \mathbb{E}\left[(\beta_j X^j + \epsilon - \beta_j X'^j)^2\right] - \sigma^2 \\
&= \beta_j^2 \mathbb{E}\left[(X^j - X'^j)^2\right] && \text{(using that } X^j \overset{\text{i.i.d.}}{\sim} X'^j) \\
&= 2\beta_j^2 \sigma_j^2. && (10)
\end{aligned}$$

On the other hand, we also have

$$\begin{aligned}
\psi_{\mathrm{CPI}}(j, P_0) &:= \mathbb{E}\left[(y - m(\widetilde{X}^{(j)})^2\right] - \mathbb{E}\left[(y - m(X))^2\right] \\
&= \mathbb{E}\left[(\sum_{i=1}^{p} \beta_i X^i + \epsilon - \sum_{i \neq j} \beta_i X^i - \beta_j \widetilde{X}^{(j)j})^2\right] - \sigma^2 \\
&= \mathbb{E}\left[(\beta_j X^j + \epsilon - \beta_j \widetilde{X}^{(j)j})^2\right] - \sigma^2 \\
&= \mathbb{E}\left[(\beta_j X^j - \beta_j \widetilde{X}^{(j)j})^2\right] \\
&= \beta_j^2 \mathbb{E}\left[\mathbb{E}\left[(X^j - \widetilde{X}^{(j)j})^2 | X^{-j}\right]\right] && \text{(using that } X^j \overset{\text{i.i.d.}}{\sim} \widetilde{X}^{(j)j} | X^{-j}) \\
&= 2\beta_j^2 \mathbb{E}\left[\mathbb{V}(X^j | X^{-j})\right]. && (11)
\end{aligned}$$

We observe that if the j-th column is independent of the rest, then both quantities coincide, and both estimates should converge to the same value. However, this is usually not the case. For instance, in highly-correlated setting, (11) *vanishes while* (10) *remains constant.*

Conditional sampling: We observe that the step of obtaining $\widetilde{X}^{(j)}$ is equivalent to sampling from the conditional distribution $X^j | X^{-j}$. This is not an easy problem. Moreover, we note that no theoretical study has been conducted on this step to provide guarantees that we are sampling from the correct distribution in Chamma et al. (2023). For this reason, in Section 3.1, we establish the assumptions under which permuting the residuals provides a valid sampling method.

Type-I error control: The main contribution of this method is that, by leveraging asymptotic normality, it is possible to provide valid p-values to test the nullity of the importance of the covariates. The proof of this theorem relies primarily on Theorem 1 from Williamson et al. (2021b), applied to both the unmodified predictiveness and the conditional counterpart. Therefore, it is necessary to prove that the conditional sampling indeed samples from the empirical conditional distribution.

Therefore, we observe that there is no obvious link between refitting-based and permutation-based variable importance. Moreover, the only available statistical guarantees offered are for type-I error control, whereas we would prefer stronger control. This is especially problematic in high-dimensional settings, where the number of false positives explodes due to the variability of the covariates (see Giraud (2021)) and the complex dependencies among them.

3 First internship contribution: Theoretical analysis of CPI

As noted earlier, both LOCO and CPI are conditional variable importance measures, meaning they assess the importance of each covariate relative to the others. However, no formal link between them has been established thus far. Establishing such a connection would be important, as it could bridge removal-based and permutation-based approaches. For instance, this would allow us to estimate the LOCO (also known as the unnormalized Total Sobol index), a well-known quantity in sensitivity analysis that assesses the variability explained by each covariate, using only a version of the CPI. This approach would address the challenges of having to refit a model for each covariate, which is not only computationally intensive but also suffers high variability due to the accumulation of estimation errors from the regressors.

To achieve this, we first examine the conditional sampling step of the CPI and establish the assumptions under which it is valid in Section 3.1. Then, in Section 3.2, we establish the link between removal and permutation-based approaches, which provides a LOCO estimate using the CPI. In Section 3.3, we introduce *Robust-CPI*, another permutation-based approach aimed at estimating LOCO, which seeks to improve stability by averaging across permutations to mitigate potential extrapolation issues. Finally, in Section 3.4, we present numerical experiments demonstrating that permutation-based methods exhibit much lower variance compared to refitting approaches.

3.1 On the conditional sampling: the theoretical framework to establish the validity of the sampling step

The conditional permutation importance methods discussed in Chamma et al. (2023)(Section 2.4) and the knockoffs procedure proposed by Blain et al. (2024) rely on the assumption that sampling from the residuals of inter-covariate regressions, combined with the regression on the covariate of interest, is equivalent to sampling from the conditional distribution. However, sampling from the conditional distribution is not a trivial task. Even under Gaussian assumptions, where explicit formulas for the conditional distribution exist, practical challenges arise due to the need to estimate the inverse of a covariance matrix, which is computationally unstable in practice (see Blain et al. (2024)). Approaches such as normalizing flows Papamakarios et al. (2021), while theoretically sound, are also impractical due to the large number of training samples required. Another method, proposed by Chamma et al. (2023), involves sampling from the tree leaves instead of taking the average of the regressed covariate X^j given the rest X^{-j}. This approach, however, proves ineffective in practice because it only considers the covariates involved in the splits, ignoring the rest, trying to make it comparable to a nearest-neighbors method but with only a subset of the *relevant* covariates. Although this idea should work, it does not, as there are no supporting theoretical results. Furthermore, determining hyperparameters, such as the number of samples per leaf to ensure diverse enough conditional sampling, is not trivial. Consequently, this method struggles in high-dimensional settings because, with deep trees, no neighbors are found, leading to a lack of diversity. On the other hand, with shallow trees, the samples from each leaf do not come from the same conditional distribution.

Moreover, the methods that do work are domain-specific generative models (see Sesia et al.

(2020)). On the other hand, sampling from the residuals and adding them to the regressed part works well in practice, as it preserves the information from the other covariates. In this section, we aim to study theoretically the assumptions necessary in this framework. We begin by ignoring the estimation error $\widehat{\nu}_{-j}$ in regressing X^j given X^{-j}. First, in Section 3.1.1, we assume Gaussianity, and then generalize this to assumptions similar to those used in regression models of independent additive noise in Section 3.1.2. Finally, in Section 3.1.3, we examine the convergence of the empirical distribution to the conditional distribution by studying the estimation error when the regressor $\widehat{\nu}_{-j}$ is consistent.

We recall that given X, X' two i.i.d. random variables and $\widehat{\nu}_{-j}$ a regressor of X^j given X^{-j}, $\widetilde{X}^{(j)} \in \mathbb{R}^p$ is given by

$$\widetilde{X}^{(j),l}|X = \begin{cases} X^l & \text{if } l \neq j \\ \widehat{\nu}_{-j}(X^{-j}) + \left[X'^j - \widehat{\nu}_{-j}(X'^{-j})\right] & \text{if } l = j. \end{cases} \tag{12}$$

3.1.1 Assuming Gaussian input

We start studying the validity of the method in the Gaussian setting.

Assumption 2 (Gaussian covariates)**.** $X \sim \mathcal{N}(\mu, \Sigma)$.

First, we note that if X is Gaussian, then the conditional distribution is also Gaussian, i.e. $X_j|X_{-j} = x_{-j} \sim \mathcal{N}(\mu_{\text{cond}}, \Sigma_{\text{cond}})$ with $\mu_{\text{cond}} = \mu_j + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}(x_{-j} - \mu_{-j})$ and with $\Sigma_{\text{cond}} = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j}$. Firstly ignoring the estimation error that will be treated later, we begin by showing that the goal of the CPI sampling step is to attain this distribution.

Lemma 3.1 (CPI sampling step with Gaussian data)**.** *Under the Gaussian covariate assumption (Assumption 2) and ignoring the estimation error of* $\widehat{\nu}_{-j}$, $\widetilde{X}^{(j)j}|X^{-j} = x^{-j}$ *is independently sampled according to the conditional distribution* $X^j|X^{-j} = x^{-j}$.

Proof. First, we start by recalling that in the Gaussian setting, we have that $X^j|X^{-j} = x^{-j} \sim \mathcal{N}(\mu_{\text{cond}}, \Sigma_{\text{cond}})$ with $\mu_{\text{cond}} = \mu_j + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}(x^{-j} - \mu_{-j})$ and with $\Sigma_{\text{cond}} = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j}$.

We also recall that $\widetilde{X}^{(j)j}|X^{-j} = x^{-j}$ is sampled from $\nu_{-j}(x^{-j}) + \left(X^j - \nu_{-j}(X^{-j})\right)^{\text{perm}} = \mathbb{E}\left[X^j|X^{-j} = x^{-j}\right] + (X^j - \mathbb{E}\left[X^j|X^{-j}\right])^{\text{perm}}$.Indeed, we first attempt to estimate the predictable component of covariate j from the remaining covariates for each specific individual, and then we permute the residuals from these predictions within a test set. We observe that this permutation step is equivalent to sampling from the residual of an i.i.d. individual.

We note that the first part corresponds to the mean of the conditional distribution by definition of the regression function $\mathbb{E}\left[X^j|X^{-j} = x^{-j}\right] = \mu_{\text{cond}}$.

Now, we need to prove that the residuals from which we sample are distributed as centered Gaussians with covariance Σ_{cond}, so that, when added to the first part, we obtain the conditional distribution. First, we observe that

$$X^j - \mathbb{E}\left[X^j|X^{-j}\right] = X^j - \mu_j - \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}(X^{-j} - \mu_{-j}).$$

Then, it is a linear combination of coordinates of a Gaussian vector therefore Gaussian. We easily observe that it is centered. Finally, we compute the variance as

$$\begin{aligned}
&\mathbb{V}(X^j - \mu_j - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}(X^{-j} - \mu_{-j}))\\
&\quad= \mathbb{V}(X^j) + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\mathbb{V}(X^{-j})\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} - 2\mathbb{E}\left[(X^j - \mu_j)\Sigma_{j,-j}\Sigma_{-j,-j}^{-1}(X^{-j} - \mu_{-j})\right]\\
&\quad= \Sigma_{j,j} + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} - 2\Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}\\
&\quad= \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}.
\end{aligned}$$

We observe that this is exactly Σ_{cond}. Therefore,

$$\widetilde{X}^{(j)j}|\left(X^{-j} = x^{-j}\right) = \mu_{\text{cond}} + (X^j - \mathbb{E}\left[X^j|X^{-j}\right])^{\text{perm}} \sim \mu_{\text{cond}} + \mathcal{N}(0, \Sigma_{\text{cond}}).$$

□

3.1.2 Towards a more general assumption

We observe that the Gaussian assumption can be generalized. In fact, the only requirement is that the residuals must be identically distributed and independent of the other covariates, ensuring that all the information X^{-j} could provide about X^j has been fully extracted.

Assumption 3. For each $j \in \{1, \dots, p\}$ we have that there exists a function ν_{-j} such that $X^j = \nu_{-j}(X^{-j}) + \epsilon_j$ with $\epsilon_j \perp\!\!\!\perp X^{-j}$ and $\mathbb{E}\left[\epsilon_j\right] = 0$.

We first notice that this assumption is exactly the one given in the regression model (Assumption 1), but applied to each column instead of only to the output y.

We observe that this function ν_{-j} is exactly the conditional expectation because

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

Therefore, the empirical approximation will consist on regressing the column and adding a permuted residual giving the CPI sampling step.

We also notice that the Gaussian assumption (2) is included in this assumption. To see this, we observe that we can decompose each column as

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

We can denote $\nu_{-j}(X^{-j}) := \mathbb{E}\left[X^j|X^{-j}\right]$ and $\epsilon_j := X^j - \mathbb{E}\left[X^j|X^{-j}\right]$. First note that ϵ_j is centered. Then, to see that they are independent, as they are each Gaussian variables, we just need to prove that their covariance is null:

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

Therefore, we have observed that Assumption 2 fulfills Assumption 3.

Finally, we easily observe that under this assumption, the CPI sampling step samples correctly from the conditional distribution:

Proposition 3.2 (CPI sampling step with independent residuals)**.** *Under Assumption 3 and ignoring the estimation error of* $\widehat{\nu}_{-j}$, $\widetilde{X}^{(j)j} \overset{\text{i.i.d.}}{\sim} X^j|X^{-j} = x^{-j}$.

Proof. We first observe that for $X' \overset{\text{i.i.d}}{\sim} X$ we have that $X'^j - \nu_{-j}\left(X'^{-j}\right) \overset{\text{i.i.d}}{\sim} \epsilon_j$.

Finally, we observe 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 $\left(X'^j - \nu_{-j}\left(X'^{-j}\right)\right)$. □

3.1.3 Empirical conditional distribution

We observe that we do not have directly the regressor ν_{-j} but just an estimate of it $\widehat{\nu}_{-j}$. However, the next proposition shows that if we use a consistent estimate, we are going to asymptotically sample from the desired conditional distribution. To do so, we need our estimate $\widehat{\nu}_{-j}$ to be consistent, i.e.

$$\mathbb{E}\left[\left(\widehat{\nu}_{-j}\left(X^{-j}\right)-\nu_{-j}\left(X^{-j}\right)\right)^2\right] \to 0.$$

We observe that for instance, the Random Forest satisfies this property under some mild assumptions (see Scornet et al. (2015)).

Proposition 3.3 (Empirical conditional sampling)**.** *Under Assumption 3, if the regressor is consistent, then the CPI sampler is going to asymptotically sample from the conditional distribution.*

Proof. In this proof we are going to use the usual 2-Wasserstein distance which is given by

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

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

We define P the conditional distribution which using Proposition 3.2 has the same distribution as $\mathbb{E}\left[X^j\middle|X^{-j}\right]+\left(X'^j-\mathbb{E}\left[X'^j|X'^{-j}\right]\right)$. We also define the empirical couterpart by $\widehat{P} := \widehat{\nu}_{-j}\left(X^{-j}\right)+X'^j-\widehat{\nu}_{-j}\left(X'^{-j}\right)$. Now, we bound the distance between them:

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

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

3.2 LOCO vs CPI: Link between removal and permutation based approaches

We observe that both the LOCO and CPI approaches aim to assess the decrease in accuracy when the information provided *uniquely* by the covariate of interest is excluded. In the LOCO approach, this is done by directly training the model without the covariate. In the CPI approach, however, the model is reused while attempting to exclude only the information uniquely provided by the covariate by permuting it conditionnally on the rest. The relationship between the quantities being estimated by these methods is not immediately clear. Specifically, the connection between LOCO, which corresponds to the unnormalized total Sobol index, and Permutation Feature Importance (PFI), also known as Mean Decrease Accuracy (MDA), was established in Bénard et al. (2022b) (see Section 2.3). In this section, we will demonstrate that, unlike PFI, the theoretical estimation in

LOCO does not involve any undesirable terms, but rather only the unnormalized total Sobol index multiplied by a constant.

We recall that ψ_{CPI} was defined in (9) as $\mathbb{E}\left[(Y-m(\widetilde{X}^{(j)}))^2\right]-\mathbb{E}\left[(Y-m(X))^2\right]$, which can be developed as

$$\begin{aligned}\psi_{\mathrm{CPI}}(j,P_0)&=\mathbb{E}\left[(Y-m(\widetilde{X}^{(j)}))^2\right]-\mathbb{E}\left[(Y-m(X))^2\right]\\&=\mathbb{E}\left[(m(X)-m(\widetilde{X}^{(j)}))^2\right]\\&=\mathbb{E}\left[\mathbb{E}\left[(m(X)-m(\widetilde{X}^{(j)}))^2|X^{-j}\right]\right]\\&=\mathbb{E}\left[2\mathbb{E}\left[(m(X)-\mathbb{E}\left[m(X)|X^{-j}\right])^2|X^{-j}\right]\right] &&(\text{using that } X\overset{\text{i.i.d.}}{\sim}\widetilde{X}^{(j)}|X^{-j})\\&=2\mathbb{E}\left[\mathbb{V}(m(X)|X^{-j})\right]\\&=2\psi_{\mathrm{LOCO}}(j,P_0).\end{aligned}$$

Therefore, we could easily correct the CPI to estimate the LOCO by directly dividing it by 2:

Definition 3.4 (0.5CPI)**.** Given a covariate j, a training sample of size n_{train} and a test sample of size n_{test}, we train the regressors $\widehat{m}$ and $\widehat{\nu}_{-j}$ over the train set, we compute the residuals over the test set, and the new LOCO estimate is given by

$$\widehat{\psi}_{0.5\mathrm{CPI}}(j,P_0)=\frac{1}{2n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\left((y_i-\widehat{m}(\widetilde{\boldsymbol{x}}_i^{(j)}))^2-(y_i-\widehat{m}(\boldsymbol{x}_i))^2\right).$$

We observe that this estimate relies primarily on the idea discussed in Section 1.4.1. Specifically, we compute this LOCO estimate without needing to refit a model on y given X^{-j} for each covariate j. Instead, we regress X^j on X^{-j}. This leverages the fact that while the relationship between y and X may be complex and require training long and complicated models, the relationships among the input covariates are simpler, allowing for the use of faster and simpler regressors. Building on this idea, in the next section, we will introduce an alternative LOCO estimate designed to reduce the variance of the method by averaging over multiple conditional samplings. This approach reduces variance while remaining as fast as the original method, as it eliminates the need to refit a model and only requires predictions over other permutations, which is quick and efficient.

3.3 Robust-CPI: a new approach to estimate LOCO

As discussed before, one of the main obstacles in using approaches such as LOCO is that they require retraining the model for each covariate j to test its importance. Such computational burdens can be very costly, making these approaches impractical. Moreover, in practice such approaches are instable due to the optimisation error made by $\widehat{m}$ and $\widehat{m}_{-j}$. One way to possibly overcome this barrier is to use the following Towers property: $m_{-j}(X)=\mathbb{E}\left[Y|X^{-j}\right]=\mathbb{E}\left[\mathbb{E}\left[Y|X\right]|X^{-j}\right]=\mathbb{E}\left[m(X)|X^{-j}\right]$. Therefore, instead of retraining a new model $\widehat{m}_{-j}$ for each covariate, we only need to take the average across different predictions where X^{-j} is fixed and we sample the j-th coordinate independently from the conditional distribution:

$$\frac{1}{n_{\text{cal}}}\sum_{i=1}^{n_{\text{cal}}}\widehat{m}\left(\widetilde{X}_i^{(j)}\right),$$

where for each $i \in \{1, \dots, n_{\text{cal}}\}$, $\widetilde{X}_i^{(j)j} \overset{\text{i.i.d.}}{\sim} X^j | X^{-j}$. Therefore, we will have for each coordinate n_{cal} samples. In particular, if it is 1, we recover the exact CPI, and using the consistency of the regressor and the law of large numbers, as n_{cal} tends to infinity, we recover the restricted regressor. We are trying to reuse the same idea from the removal versus permutation approach, where refitting a complete model is expensive, but making predictions is relatively inexpensive. Moreover, if we apply the conditional sampling step of the CPI, as proven valid in Section 3.1, the only requirement would be to use more permutations from the residuals without needing to retrain any model. This approach would increase stability without incurring additional costs. It should also improve the stability of the method compared to the estimation done by $0.5 * \text{CPI}$. This is because, even with conditional sampling, extrapolation issues can arise when predicting in regions where the regressor was not trained. By taking the mean, we can enhance stability in such cases.

We observe that while trying to estimate m_{-j} using only the already trained regressor $\widehat{m}$ could be directly related to the problem of predicting with missing values. Indeed, in the missing values framework, the Bayes predictor can be decomposed pattern-by-pattern. There are some specific cases in which some shared information across the patterns can be used to efficiently use this decomposition (see Ayme et al. (2022) for regression and Lobo et al. (2024) for classification). However, this is not the general case and usually there is a computational burden to estimate a regressor for each missing pattern, which is similar to the computational burden for the LOCO (indeed it is not the same because for the LOCO we only restrain one coordinate and therefore there would be a need to retrain d models while in the missing data there would be one for each missing pattern, and therefore potentially 2^d. This is what happens with Shapley values. Moreover, with missing data we could only use the data sharing the same missing pattern, and therefore, much less than in the LOCO without missing data). Nevertheless, there have been some studies to predict with missing data using the global model. For instance, ToweranNA (see Matloff and Mohanty (2023)) have tried to predict on a missing input using the global model by predicting on the closest input value across the complete data. However, it relies on the neigborhood concept that is lost with the curse of dimensionality. Therefore, another idea that we could use would be the previous Tower's property with the conditional sampling, giving this estimate:

$$\widehat{\psi}_{\text{LOCO}'}(j, P_0) := \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(y_i - \frac{1}{n_{\text{cal}}} \sum_{k=1}^{n_{\text{cal}}} \widehat{m}(\widetilde{x}_{i,k}^{(j)}) \right)^2 - \left(y_i - \widehat{m}(x_i)\right)^2, \tag{13}$$

where $\widetilde{x}_{i,k}^{(j)}$ the $-j$-th coordinates fixed to the observed coordinates of x_i^{-j}, and the j-th coordinate conditionally sampled on the rest. In practice, under Assumption 3, using Proposition 3.3, we could conditionally sample in the following way:

$$\widetilde{x}_{i,k}^{(j),l} = \begin{cases} x_i^l & \text{if } l \neq j \\ \widehat{\nu}_{-j}(x_i^{-j}) + \left[x_k^j - \widehat{\nu}_{-j}(x_k^{-j}) \right] & \text{if } l = j. \end{cases} \tag{14}$$

We first note that this estimate is consistent:

Proposition 3.5 (Consistency of LOCO'). *Under Assumption 3, assuming the consistency of the regressors $\widehat{m}$ and $\widehat{\nu}_{-j}$, then*

$$\widehat{\psi}_{\text{LOCO}'}(j, P_0) \xrightarrow{n_{\text{train}}, n_{\text{test}}, n_{\text{cal}} \to \infty} \psi_{\text{LOCO}}(j, P_0).$$

Proof. First of all, we note that by Assumption 3 and the consistency of $\widehat{\nu}_{-j}$, we can apply Proposition 3.3 to proof that the sampling step does sample independently from the conditional distribution. Therefore, applying the consistency of $\widehat{m}$ and the LLN, we can obtain that

$$\frac{1}{n_{\text{cal}}}\sum_{k=1}^{n_{\text{cal}}}\widehat{m}(\widetilde{x}_{i,k}^{(j)}) \xrightarrow{n_{\text{train}},n_{\text{cal}}\to\infty} \mathbb{E}\left[m(X)|X^{-j}=x_i^{-j}\right] = m_{-j}(x_i^{-j}).$$

Finally, we conclude using the continuity of the function $g_i(x) := (y_i - x)^2$ and another time the LLN over the test set. For the second part of the sum we use also the consistency of $\widehat{m}$, the continuity and the LLN.

□

In practice: We do not need three separate samples to compute this estimate. In fact, the calibration set is not necessary; the only requirement is that the residuals, from which we sample the conditional distribution, are computed *fairly*, meaning they cannot be derived from the same training sample. Otherwise, the residuals would be artificially small due to overfitting. Therefore, in practice, we can simply obtain the residuals from the test set, as the regressor $\widehat{\nu}_{-j}$ is trained on the training sample. This test set is also used to compute the variable importance measure using the trained regressors. However, to preserve the diversity of the residuals, we cannot let n_{cal} increase indefinitely, as it is constrained by the number of test samples. In practice, we fix the number of calibration samples, n_{cal}, which introduces a small bias. As observed in the previous section Section 3.2, when n_{cal} is set to one, we estimate twice the LOCO. This bias can be easily analyzed and corrected for an arbitrary n_{cal}:

Proposition 3.6 (Bias of LOCO'). *Given $n_{\text{cal}} < \infty$, assuming Assumption 3 and the consistency of the regressors $\widehat{m}$ and $\widehat{\nu}_{-j}$, then*

$$\widehat{\psi}_{\text{LOCO}'}(j, P_0) \xrightarrow{n_{\text{train}},n_{\text{test}}\to\infty} \left(1+\frac{1}{n_{\text{cal}}}\right)\psi_{\text{LOCO}}(j, P_0).$$

Proof. The first part of the proof to obtain the asymptotic quantity is exactly the same as in the proof of Proposition 3.5, but without applying the LLN to n_{cal}. This quantity can be developed 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

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

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 LOCO:

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

and

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

Therefore we have

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

□

Therefore, we can simply correct this bias of this averaged estimate by multiplying by $n_{\text{cal}}/(n_{\text{cal}}+1)$. This gives the Robust-CPI:

Definition 3.7 (Robust-CPI)**.** Given a train, test and calibration set, we define the Robust-CPI estimate as

$$\widehat{\psi}_{\text{Robust-CPI}}(j,P_0):=\frac{n_{\text{cal}}}{n_{\text{cal}}+1}\left(\frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\left(y_i-\frac{1}{n_{\text{cal}}}\sum_{k=1}^{n_{\text{cal}}}\widehat{m}(\widetilde{x}_{i,k}^{(j)})\right)^2-(y_i-\widehat{m}(x_i))^2\right),\qquad(15)$$

where $\widetilde{x}_{i,k}^{(j)}$ is computed as in (14).

We observe that all the consistency of this estimate is obtained using the previous tools for the LOCO'.

3.4 Numerical experiments

In this section, we compare direct-LOCO estimates that attempt to estimate LOCO by refitting with those introduced in this chapter using permutation-based approaches. In the first group, we use either the initial naive estimate of LOCO, implemented in the experiments of Chamma et al. (2023) (denoted as LOCO-AC), or the bias-corrected version from Williamson et al. (2021a), implemented in the Python package VIMPY (denoted as LOCO). For the permutation-based approaches, we first compute $\widehat{\psi}_{0.5\mathrm{CPI}}$ from Definition 3.4, followed by $\widehat{\psi}_{\mathrm{Robust-CPI}}$ from Definition 3.7. We will also compute the PFI, defined in Section 2.3, to demonstrate that this quantity has no relation to the LOCO quantity we aim to estimate and that it indeed has extrapolation issues, which are reflected in the high variance of its results.

In the first two experiments, we consider a simple linear model, and in the following experiments, we increase the complexity by augmenting the dimensions and introducing non-linearity. For the more complex settings, we exclude LOCO-AC due to its computational intensity and its inferior performance compared to the method from Williamson et al. (2021a). Overall, we observe that permutation-based approaches generally have much lower variance than refitting-based approaches, as they do not accumulate the error from repeatedly refitting the model. We will also observe that, although Williamson et al. (2021a) claims to provide valid estimates of LOCO and valid confidence intervals for null hypothesis testing, this approach does not perform as expected in practice.

In practice, n denotes the total number of available samples used to compute each importance measure. Thus, sample splitting will be performed to train the regressors and another to compute the importance measure. To obtain more stable estimates, a cross-fitted version is used, where the final result is the mean of the measures across the folds. We will use 5 folds for the LOCO method of Williamson et al. (2021a) and 2 folds for the other methods, which will still demonstrate greater stability. Moreover, the CPI approaches are also parallelizable. Each experiment is repeated 3 times to get the variance of each method.

3.4.1 Linear setting

In this first experiments we study the effect of correlation and the convergence rates with a simple and low-dimensional linear setting.

Effect of the correlation: We empirically study the estimation of LOCO using previous methods in a simple setting that can be theoretically analyzed (see Example A.1). As seen in Figure 2, even in such a simple case, when the number of samples is insufficient, there is still some bias in achieving the theoretical quantity across all methods. Moreover, the direct-LOCO method estimates exhibit much greater variance, as there are multiple regressors to estimate, and their errors do not couple as expected.

The PFI has nothing to do with the interest quantity and even in this easy setting there is a really high variance.

Convergence rates: Using the exact same setting as before, with the correlation between the covariates fixed at 0.6, we observe the behavior of the various methods as the number of available samples changes. On the left of Figure 3, we see that the importance of the first covariate is greater than that of the second, as its coefficient is higher. We also observe that permutation-based

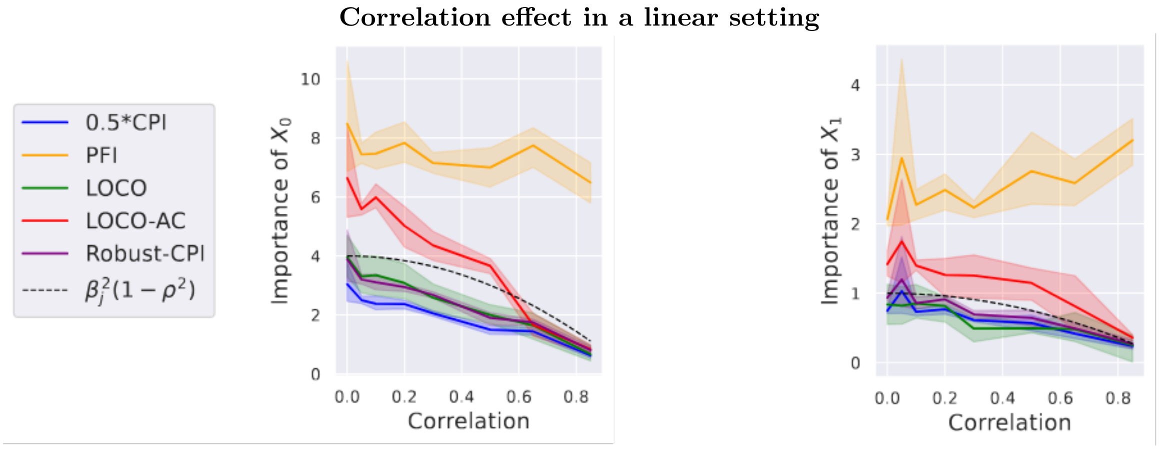

Figure 2: **Setting:** $n = 100, n_{\text{cal}} = 100, y = \beta_0 X_0 + \beta_1 X_1 + \sigma\epsilon$ with $\sigma = \|X\beta\|_2/\text{SNR}$, SNR $= 4$ the signal-to-noise ratio, ϵ a standard centered Gaussian and $[\beta_0, \beta_1] = [2, 1]$. The black dotted line represents the theoretical quantity. The x-axis represents the correlation between X_0 and X_1. On the left is the importance of X_0, and on the right is the importance of X_1.

approaches tend to converge faster and give more accurate results than the other methods, with lower variance. This is even more remarkable with a small number of training samples, as the variance of refitting-based approaches is very large.

3.4.2 Non-linear setting

In the following experiments we will sample the input $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ a Toeplitz matrix, in which the entry i, j is $\rho^{|i-j|}$. Then, the relationship with the output is given by $y = X_0 X_1 \mathbb{1}_{X_2>0} + 2X_3 X_4 \mathbb{1}_{X_2<0}$.

In this setting, an explicit form of the theoretical LOCO is not obvious. Indeed, if $\mu = \mathbf{0}$, we have an explicit form (see Example A.2), but otherwise, we do not. In cases where we do not have the explicit form (in the following two experiments), we compute LOCO with a large number of samples to obtain an asymptotic estimate. For a fair comparison, this is done using the LOCO estimator by Williamson et al. (2021a). Nevertheless, we will show that, in any case, this should not be taken as ground truth, because even with a large number of samples, it is not a reliable quantity. In the last two experiments, where $\mu = \mathbf{0}$, it is possible to compute the theoretical quantity. However, to demonstrate that the previous asymptotic estimate is not reliable, we also compute the asymptotic LOCO from Williamson et al. (2021a) in this case, where the ground truth is known, and observe a significant bias. Additionally, we compute the 0.5CPI and Robust-CPI with the same number of samples and achieve better performance. Future research will study these experiments in more detail.

Convergence rates: In Figure 4, we first observe that the LOCO method of Williamson et al. (2021a) exhibits very high variance. In fact, it assigns importance to covariates that are not important (see the two figures on the right), which even the PFI recognized. On the other hand, the permutation-based methods are much more stable from the beginning, even with a small sample size, producing a line much more parallel to the x-axis and more concentrated across the experiments.

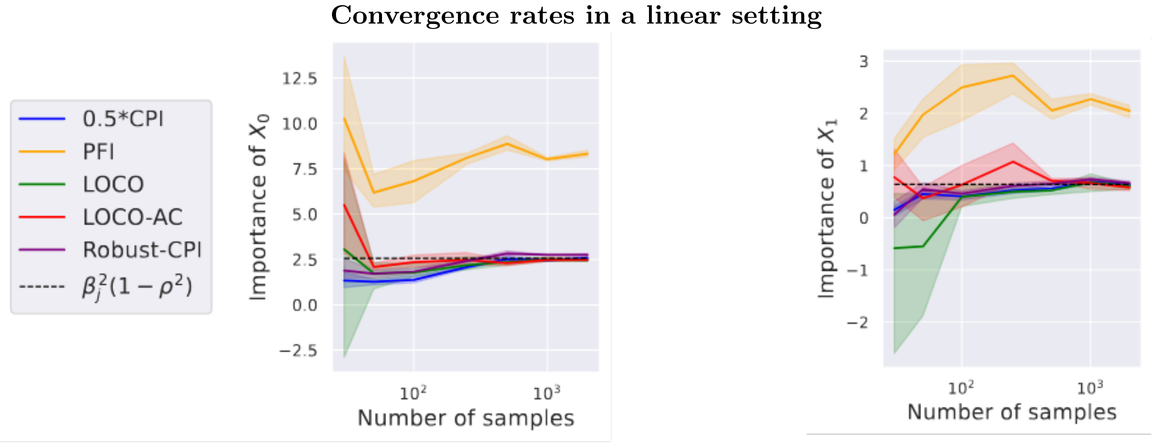

Figure 3: **Setting:** $n_{\text{cal}} = 100, y = \beta_0 X_0 + \beta_1 X_1 + \sigma\epsilon$ with $\sigma = \|X\beta\|_2/\text{SNR}$, SNR $= 4$ the signal-to-noise ratio, ϵ a standard centered Gaussian and $[\beta_0, \beta_1] = [2, 1]$. The black dotted line represents the theoretical quantity. The correlation between X_0 and X_1 is fixed to 0.6. The x-axis represents the number of samples available to compute the importance measure. On the left is the importance of X_0, and on the right is the importance of X_1.

Effect of the correlation: We study the effect of correlation in two different settings. The only change will be the mean of the input covariates X: it will be shifted from $\mu = \mathbf{1}$ in Figure 5 to $\mu = \mathbf{0}$ in Figure 6.

First, in Figure 5, we observe that, as usual, LOCO from Williamson et al. (2021a) exhibits higher variance. Moreover, we observe from the dotted black line that even asymptotically, it does not yield good results. Indeed, even with $n = 100000$, it fails to capture the nullity of covariate 6, as seen in the fourth experiment of Figure 5. Therefore, with a small training set, it suffers from significant variability, making the results unreliable and even with a large number of samples, it fails to discard null covariates.

Next, in Figure 6, we centered X. As a result, covariates X_0 and X_1 lose some importance because covariate X_2 is now centered, causing both parts of the function $y = X_0 X_1 \mathbb{I}_{X_2>0} + 2X_3 X_4 \mathbb{I}_{X_2<0}$ to be used more uniformly. We observe that this reduces the importance measures for the first two covariates, which nearly vanish in Figure 6 for the permutation methods. However, this is only due to the model not being well-trained. When we increase the number of training samples (see the dotted lines) to $n = 10000$, we observe that the permutation-based methods achieve good results in estimating the theoretical quantity. Furthermore, even with a small number of training samples, they were able to reject the null covariates.

On the other hand, LOCO from Williamson et al. (2021a) was unstable. With a small training sample, it exhibited high variability and failed to discard the null covariates. Even with a larger training sample (green dotted line), it was inaccurate in finding the theoretical quantity for the non-null covariates (see the two left images in Figure 6) and was unable to control for the null covariates (see the fourth image in Figure 6).

Effect of dimension: In Figure 7, we study the effect of increasing the dimension of the input covariates X. We observe that the difficulty increases with the dimension. In this case, it is possible to compute an explicit expression for the importance of X_0(see Example A.2). The gray-dotted line stands for this theoretical quantity. We observe that not only do none of the methods estimate this

Convergence rates in a non-linear setting

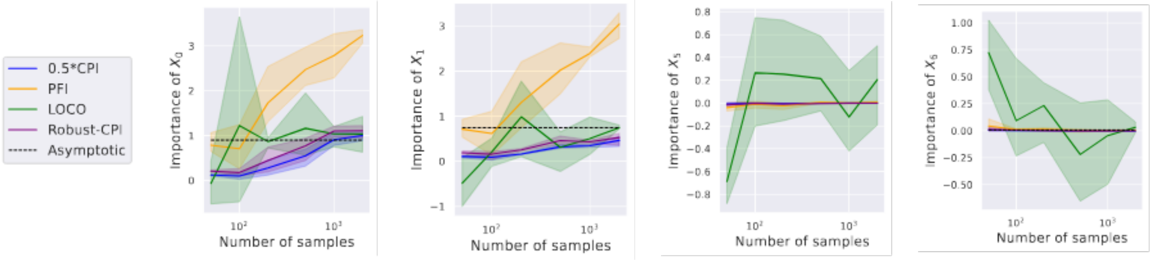

Figure 4: **Setting:** $y = X_0X_1\mathbb{1}_{X_2>0} + 2X_3X_4\mathbb{1}_{X_2<0}$, with $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma_{i,j} = \rho^{|i-j|}$, $p = 50, \rho = 0.6, \mu = \mathbf{1}$ and $n_{\text{cal}} = 100$. The black dotted line represents LOCO of Williamson et al. (2021a) with $n = 100000$. The two figures on the left represent important covariates, while the two on the right represent non-important covariates.

Correlation effect in a non-linear setting with not-centered X

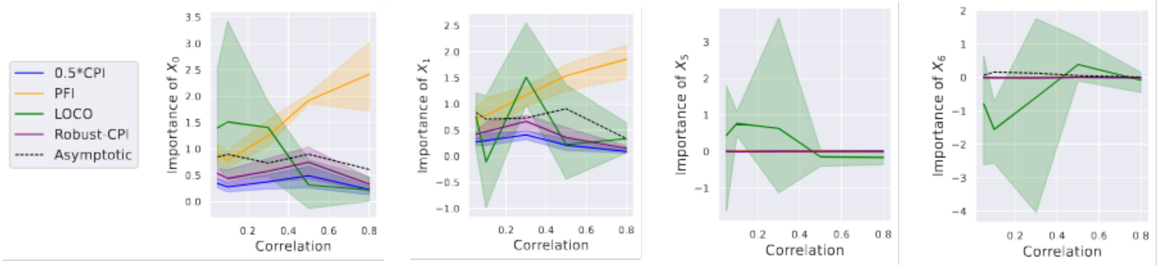

Figure 5: **Setting:** $y = X_0X_1\mathbb{1}_{X_2>0} + 2X_3X_4\mathbb{1}_{X_2<0}$, with $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma_{i,j} = \rho^{|i-j|}$, $p = 50, n = 300, \mu = \mathbf{1}$ and $n_{\text{cal}} = 100$. The black dotted line represents LOCO with $n = 100000$. The two figures on the left represent important covariates, while the two on the right represent non-important covariates. On the x-axis, we vary the correlation ρ between the covariates in the covariance matrix Σ, which is Toeplitz.

quantity accurately with a small number of samples, but even with a large number of samples, the LOCO method of Williamson et al. (2021a) does not correctly estimate it (see the green dotted line) while the methods presented in this work do estimate correctly the quantity. Additionally, it fails to assign zero importance to non-important covariates and exhibits high variance.

4 Second internship contribution: Controlled variable selection using CPI and variants

In the literature, LOCO is commonly used for variable selection, as the goal is to obtain the minimum set of features that can explain the output. For this reason, multiple approaches, such as Sobol-MDA (see Bénard et al. (2022b)), propose algorithms that sequentially remove features based on their total Sobol index. However, this is done without guarantees regarding the selected set. Williamson et al. (2021b) proposed some confidence intervals, which not only fail to provide sufficient statistical control in high-dimensional settings but, as shown in the previous section, also exhibit bias because they are refitting-based methods. Under the null hypothesis, they were unable to discard features in practice.

Correlation effect in a non-linear setting with centered X

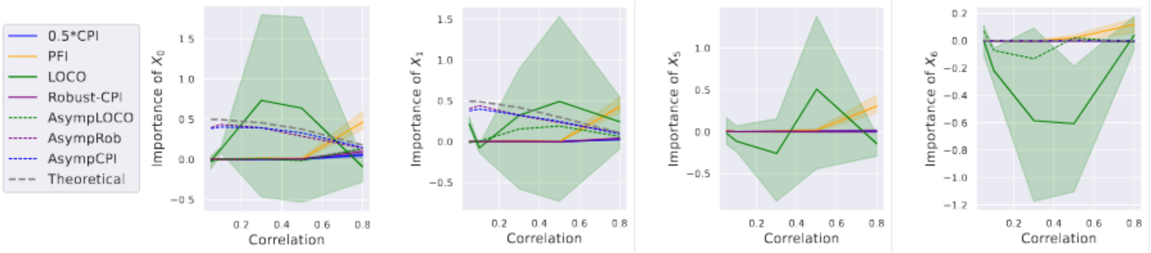

Figure 6: **Setting:** $y = X_0X_1\mathbb{1}_{X_2>0} + 2X_3X_4\mathbb{1}_{X_2<0}$, with $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma_{i,j} = \rho^{|i-j|}$, $p = 50, n = 300, \mu = \mathbf{0}$ and $n_{\text{cal}} = 100$. The two figures on the left represent important covariates, while the two on the right represent non-important covariates. On the x-axis, we vary the correlation ρ between the covariates in the covariance matrix Σ, which is Toeplitz. The dotted lines stand for the estimation of ψ_{LOCO} with $n = 10000$ using the LOCO of Williamson et al. (2021a) (AsympLOCO), the Robust-CPI (AsympRob) and the 0.5CPI (AsympCPI).

Dimension effect in a non-linear setting

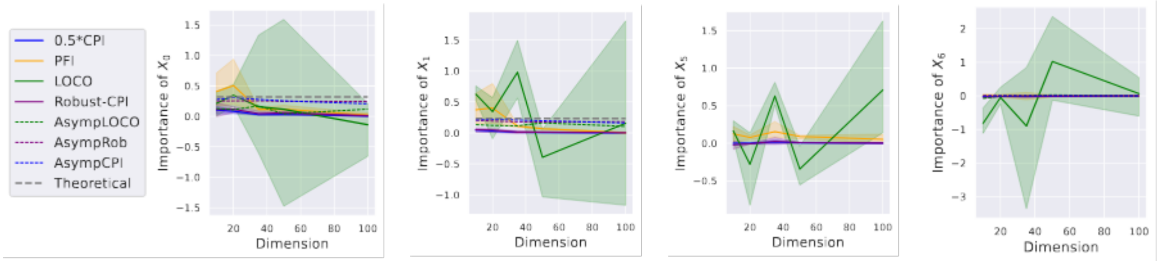

Figure 7: **Setting:** $y = X_0X_1\mathbb{1}_{X_2>0} + 2X_3X_4\mathbb{1}_{X_2<0}$, with $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma_{i,j} = \rho^{|i-j|}$, $\rho = 0.6, n = 300, \mu = \mathbf{0}$ and $n_{\text{cal}} = 100$. The two figures on the left represent important covariates, while the two on the right represent non-important covariates. On the x-axis we vary the dimension by adding null covariates. The dotted lines stand for the estimation of ψ_{LOCO} with $n = 10000$ using the LOCO of Williamson et al. (2021a) (AsympLOCO), the Robust-CPI (AsympRob) and the 0.5CPI (AsympCPI).

For this reason, we have proposed more stable estimation methods based on permutation approaches. In this section, drawing from the knockoffs literature, we will provide robust statistical guarantees that rely on permutation approaches, which are both powerful and stable.

Knockoffs provide a popular framework for conditional independence testing. The goal of conditional independence testing is to test the importance of each covariate X^j given the rest X^{-j} on predicting an output y. Therefore, it tries to understand if the variable remains informative, knowing the other covariates. This is a hard problem (see Shah and Peters (2020)).

Moreover, in high-dimensional settings, as usual with the multiple testing, there is a need for some statistical guarantees on the selected covariates, which are not usually provided by many commonly used methods such as LASSO(Tibshirani (1996)). Indeed, in controlled variable selection, we try to provide these guarantees for the selected variables. Formally, we denote by $\mathcal{H}_0 := \left\{j : y \perp\!\!\!\perp X^j \middle| X^{-j}\right\}$ the unimportant covariates and $\mathcal{H}_1 := \left\{j : y \not\perp\!\!\!\perp X^j \middle| X^{-j}\right\}$ the true unknown support, called the Markov blanket on graphical models. Then, we want to control the False Discovery Rate(FDR) of the selected set $\widehat{S}$ presented by Benjamini and Hochberg (1995), which is the expectation of the False Discovery Proportion (FDP), given by

$$\mathrm{FDP}(\widehat{S}) := \frac{\left|\widehat{S} \cap \mathcal{H}_0\right|}{\left|\widehat{S}\right| \vee 1} \qquad \mathrm{FDR}(\widehat{S}) := \mathbb{E}\left[\mathrm{FDP}(\widehat{S})\right]. \tag{16}$$

We observe that the FDP measures the rate at which non-important covariates are incorrectly discovered as important. Traditionally, there are classical methods that control this quantity, such as Benjamini and Hochberg (1995). However, these methods rely on strong assumptions about the relationship between the p-values, such as weak Positive Regression Dependency (see Giraud (2021) for more details). Moreover, this kind of assumptions does not make sense while considering conditional approaches, as the relationship between p-values across the covariates can be complex. On the other hand, procedures like Benjamini and Yekutieli (2001) do not require these assumptions, but they sacrifice the power of the test. This is because FDR control only addresses one side of the issue: it says nothing about the power of the test. Therefore, the goal is to develop a methodology that is not overly conservative while still controlling the FDR.

In Section 4.1, we provide an introduction to the knockoffs framework. In Section 4.2, we first discuss some pitfalls of a method that works well in practice but does not satisfy the theoretical properties of the knockoffs and then we adapt this method to create valid knockoffs. Then, in Section 4.3, we use this procedure to develop a method that controls the FDR by leveraging Shapley values to create a new statistic. Finally, in Section 4.4, we introduce a new parallelizable method that controls the FDR, with numerical experiments in Section 4.5 demonstrating its effectiveness.

4.1 Mathematical framework of knockoffs

The knockoffs provide a framework to control the FDR (see Candes et al. (2017)). The underlying idea is to sample from a distribution $\widetilde{X}$ that follows the same distribution as X but does not preserve the relationship with the output y. In this way, they can construct a statistic based on the original and the knockoff sample so that if the j-th covariate is important, there is a significant difference between the two.

More formally, there are three main ingredients: the model-X knockoffs, the feature statistics and then choosing a data-dependent threshold.

Definition 4.1 (Model-X knockoffs)**.** For a random variable X, it is a new random variable $\widetilde{X}$ constructed satisfying the following two properties:

1. For any subset $s \subset \{1, \dots, p\}$, the original and the knockoff variables are exchangeable, i.e. $(X, \widetilde{X})_{\mathrm{swap}(s)} \overset{d}{=} (X, \widetilde{X})$.
2. $\widetilde{X} \perp\!\!\!\perp y | X$.

We observe that the second property can be obtained by constructing the knockoffs without using the output.

Next, we need to construct insights on the importance of the coordinates on predicting the output compared with the knockoffs that do not preserve the relationship with the output. This is done by the feature statistics.

Definition 4.2 (Feature statistics from Candes et al. (2017))**.** It is a vector $\mathbf{W} = (W_1, \dots, W_p) \in \mathbb{R}^p$ where each coordinate W_j satisfies

1. It is a function of the input X, the knockoff $\widetilde{X}$ and the output y: $W_j = w_j\left(\left[X, \widetilde{X}\right], y\right)$.
2. It satisfies the flip-sign property:

$$w_j\left(\left[X, \widetilde{X}\right]_{\text{swap}(s)}, y\right) = \begin{cases} w_j\left(\left[X, \widetilde{X}\right], y\right) & \text{if } j \not\in s, \\ -w_j\left(\left[X, \widetilde{X}\right], y\right) & \text{if } j \in s. \end{cases}$$

One popular choice of knockoff statistic is the Lasso Coefficient Difference (LCD) which consists on first training a Lasso estimate by solving the following problem:

$$\min_{b \in \mathbb{R}^{2p}} \frac{1}{2}\|y - [X, \widetilde{X}]b\|_2^2 + \lambda\|b\|_1,$$

where the value of λ does not need to be fixed in advance. Then, we simply compare the difference between the original coefficient estimate with the knockoff one : $w_j = |\widehat{b}_j(\lambda)| - |\widehat{b}_{j+p}(\lambda)|$.

Finally, we select a data-dependent threshold for the knockoffs statistics. To do so, we first observe that

$$\#\left\{j \in \mathcal{H}_0 : W_j \geq t\right\} \overset{\mathrm{d}}{=} \#\left\{j \in \mathcal{H}_0 : W_j \leq -t\right\} \leq \#\left\{j : W_j \leq -t\right\}.$$

Therefore, the FDP that is given by

$$\mathrm{FDP}(t) := \frac{\#\left\{j \in \mathcal{H}_0 : W_j \geq t\right\}}{\#\left\{j : W_j \geq t\right\}},$$

is estimated by

$$\widehat{\mathrm{FDP}}(t) := \frac{\#\left\{j : W_j \leq -t\right\}}{\#\left\{j : W_j \geq t\right\}},$$

as $\mathcal{H}_0$ is unknown.

This estimation should be precise, as there should not be many genuine signals with negative feature statistics, since it is expected to measure their importance. Therefore, by denoting $\mathcal{W} := \{|W_j| : j = 1, \ldots, p\} \backslash \{0\}$, one idea to control the FDR would be to choose the threshold as

$$T_q = \min\left\{t \in \mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\},$$

and $+\infty$ if empty. Nevertheless, this quantity does not control exactly the FDR but a modified one (see Barber and Candès (2015)). To control the original FDR they have introduced a bias corrected version:

$$T_q^\star = \min\left\{t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}. \tag{17}$$

Moreover, Nguyen et al. (2020) proved that this is equivalent to apply the traditional BH procedure (see Benjamini and Hochberg (1995)) on some generated intermediate p-values.

4.2 Non-exchangeability of knockoffs: a review of Blain et al. (2024)

The construction of Model-X knockoffs, as defined in Candes et al. (2017), is not as straightforward as it may seem. In the original work, the authors rely on certain Gaussian assumptions that allow them to sample from a conditional distribution, thereby guaranteeing the exchangeability property (property 1 in Definition 4.1). However, in practice, especially in high-dimensional settings, even under Gaussian assumptions, this approach does not perform well because it requires estimates of the inverse of the covariance matrix, which is often poorly estimated. Therefore, there is a real need to generate knockoffs in practice that satisfy the exchangeability property while differing sufficiently from the original sample in order to maintain a powerful method.

To address this issue, Blain et al. (2024) proposed a method to directly sample knockoffs in a manner similar to the CPI sampling step. To create the j-th coordinate of the knockoff, they first regress the original j-th coordinate on the remaining variables, then add a random residual. While this method works in practice and has the advantage of being parallelizable, the knockoffs described in Blain et al. (2024) do not actually satisfy the desired exchangeability property.

Therefore, in this section, we begin by demonstrating the non-exchangeability of these knockoffs in a standard Gaussian setting. We then propose a sequential alternative that addresses this non-exchangeability issue.

4.2.1 Non-exchangeability in the Gaussian setting

In this section we show that the exchangeability of the proposed Model-X contruction in Blain et al. (2024) does not stand in standard cases such as Gaussian data. To see this, we first observe that if the covariate matrix is Gaussian, we do need the joint distribution to fulfill

$$\left(X^1,\dots,X^p,\widetilde{X}^1,\dots,\widetilde{X}^p\right) \sim \mathcal{N}\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix}\begin{pmatrix}\Sigma & \Sigma-\operatorname{diag}(\mathbf{s})\\ \Sigma-\operatorname{diag}(\mathbf{s}) & \Sigma\end{pmatrix}\right),$$

for some $\mathbf{s}$ such that the covariance matrix is still positive semidefinite. We will observe that the covariate property between the knockoffs with the other knockoffs is not satisfied. We begin by ignoring the optimization error for the conditional sampling due to the estimation of ν_{-j}, which is the regressor of X^j using X^{-j}. We begin by simplifying the knockoff expression using Gaussianity similarly to what was done in Section 3.1.1:

$$\begin{aligned}\widetilde{X}^j &:= \mathbb{E}\left[X^j\middle|X^{-j}\right] + X'^j - \mathbb{E}\left[X'^j\middle|X'^{-j}\right]\\ &= \left(\mu_j + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\left(X^{-j}-\mu_{-j}\right)\right) + X'^j - \left(\mu_j + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\left(X'^{-j}-\mu_{-j}\right)\right)\\ &= X'^j + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\left(X^{-j}-X'^{-j}\right).\end{aligned}$$

We easily observe that the mean is preserved for the knockoffs, and that the variance of each knockoff, as well as the covariance between a knockoff and an original covariate, is also preserved:

- $\mathbb{E}\left[\widetilde{X}^j\right] = \mathbb{E}\left[X'^j + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\left(X^{-j}-X'^{-j}\right)\right] = \mu_j.$

- For any $j \in \{1, \dots, p\}$:

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{X}^j - \mu_j\right)^2\right] &= \mathbb{E}\left[\left(X'^j - \mu_j + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\left(X^{-j} - X'^{-j}\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(X'^j - \mu_j\right)^2\right] + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\mathbb{E}\left[\left(X^{-j} - X'^{-j}\right)\left(X^{-j} - X'^{-j}\right)^\top\right]\Sigma^{-1}_{-j,-j}\Sigma_{-j,j} \\
&\quad + 2\mathbb{E}\left[\left(X'^j - \mu_j\right)\Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\left(X^{-j} - X'^{-j}\right)\right] \\
&= \Sigma_{j,j} + 2\Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j} - 2\Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j} \\
&= \Sigma_{j,j}.
\end{aligned}$$

- Without loss of generality, for readability sake, we choose the first knockoff covariate and the second original covariate (it works for any $j \neq l \in \{1, \dots, p\}$):

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{X}^1 - \mu_1\right)\left(X^2 - \mu_2\right)\right] &= \mathbb{E}\left[\left(X'^1 - \mu_1 + \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\left(X^{-1} - X'^{-1}\right)\right)\left(X^2 - \mu_2\right)\right] \\
&= \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\mathbb{E}\left[\left(X^{-1} - X'^{-1}\right)\left(X^2 - \mu_2\right)\right] \\
&\qquad \text{(using that } X'^1 \perp\!\!\!\perp X^2\text{)} \\
&= \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,2} \\
&= \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,-1}\left(\mathbf{1}, \mathbf{0}, \dots, \mathbf{0}\right) \\
&= \Sigma_{1,2},
\end{aligned}$$

  where in the second-to-last line we have used that $\Sigma_{-1,2}$ is the first column of $\Sigma_{-1,-1}$, and therefore we could rewrite it as $\Sigma_{-1,-1}\left(\mathbf{1}, \mathbf{0}, \dots, \mathbf{0}\right)$.

Without loss of generality, and for readability, we take the first and the second covariates of the knockoff and compute the covariance between them to show that it is not $\Sigma_{1,2}$ as it should be. Indeed, with this procedure, we could either add the residual from the same individual to each coordinate of the knockoff or add different and independent ones. We are going to show that in both cases, the covariance is not the desired one:

Residuals from independent samples: Let $X' \overset{\text{i.i.d.}}{\sim} X''$, then we have that the covariance between the first and second knockoff is given by:

$$\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{X}^1 - \mu_1\right)\left(\widetilde{X}^2 - \mu_2\right)\right] \\
&= \mathbb{E}\left[\left(X'^1 - \mu_1 + \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\left(X^{-1} - X'^{-1}\right)\left(X''^2 - \mu_2 + \Sigma_{2,-2}\Sigma^{-1}_{-2,-2}\left(X^{-2} - X''^{-2}\right)\right)\right)\right] \\
&= \mathbb{E}\left[\Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\left(X^{-1} - X'^{-1}\right)\Sigma_{2,-2}\Sigma^{-1}_{-2,-2}\left(X^{-2} - X''^{-2}\right)\right] && \text{(using } X' \perp\!\!\!\perp X''\text{)} \\
&= \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\mathbb{E}\left[\left(X^{-1} - X'^{-1}\right)\left(X^{-2} - X''^{-2}\right)^\top\right]\Sigma^{-1}_{-2,-2}\Sigma_{-2,2} && \text{(using } a^\top = a \text{ for } a \in \mathbb{R}\text{)} \\
&= \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,-2}\Sigma^{-1}_{-2,-2}\Sigma_{-2,2}.
\end{aligned}$$

To see that this does not coincide with $\Sigma_{1,2}$, we may observe the simple two-dimensional unit variance random variable with covariance ρ. In this case, we have that $\Sigma_{1,-1} = \rho$, $\Sigma^{-1}_{-1,-1} = 1$, $\Sigma_{-1,-2} = \rho$, $\Sigma^{-1}_{-2,-2} = 1$ and $\Sigma_{-2,2} = \rho$, so

$$\operatorname{Cov}\left(\widetilde{X}^1, \widetilde{X}^2\right) = \rho^3.$$

Residuals from the same sample: When the knockoffs are constructed using the residuals from the same sample, the covariance is given by:

$$
\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{X}^1-\mu_1\right)\left(\widetilde{X}^2-\mu_2\right)\right]\\
&\quad=\mathbb{E}\left[\left(X'^1-\mu_1+\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\left(X^{-1}-X'^{-1}\right)\left(X'^2-\mu_2+\Sigma_{2,-2}\Sigma_{-2,-2}^{-1}\left(X^{-2}-X'^{-2}\right)\right)\right)\right]\\
&\quad=\mathbb{E}\left[\left(X'^1-\mu_1\right)\left(X'^2-\mu_2\right)\right]+\mathbb{E}\left[\Sigma_{2,-2}\Sigma_{-2,-2}^{-1}\left(X^{-2}-X'^{-2}\right)\left(X'^1-\mu_1\right)\right]\\
&\qquad+\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\mathbb{E}\left[\left(X^{-1}-X'^{-1}\right)\left(X'^2-\mu_2\right)\right]\\
&\qquad+\mathbb{E}\left[\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\left(X^{-1}-X'^{-1}\right)\Sigma_{2,-2}\Sigma_{-2,-2}^{-1}\left(X^{-2}-X'^{-2}\right)\right].
\end{aligned}
$$

The first term gives $\Sigma_{1,2}$. The second term can be simplified as

$$
\begin{aligned}
\mathbb{E}\left[\Sigma_{2,-2}\Sigma_{-2,-2}^{-1}\left(X^{-2}-X'^{-2}\right)\left(X'^1-\mu_1\right)\right]&=-\Sigma_{2,-2}\Sigma_{-2,-2}^{-1}\Sigma_{-2,1}\\
&=-\Sigma_{2,-2}\Sigma_{-2,-2}^{-1}\Sigma_{-2,-2}\left(\mathbf{1},\mathbf{0},\dots,\mathbf{0}\right)=-\Sigma_{2,1}.
\end{aligned}
$$

Similarly, the third term provides

$$
\begin{aligned}
\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\mathbb{E}\left[\left(X^{-1}-X'^{-1}\right)\left(X'^2-\mu_2\right)\right]&=-\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,2}\\
&=-\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,-1}\left(\mathbf{1},\mathbf{0},\dots,\mathbf{0}\right)=-\Sigma_{1,2}.
\end{aligned}
$$

Finally, the last term can be simplified as

$$
\begin{aligned}
&\mathbb{E}\left[\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\left(X^{-1}-X'^{-1}\right)\Sigma_{2,-2}\Sigma_{-2,-2}^{-1}\left(X^{-2}-X'^{-2}\right)\right]\\
&\quad=\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\mathbb{E}\left[\left(X^{-1}-X'^{-1}\right)\left(X^{-2}-X'^{-2}\right)^{\top}\right]\Sigma_{-2,-2}^{-1}\Sigma_{-2,2}\\
&\quad=2\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,-2}\Sigma_{-2,-2}^{-1}\Sigma_{-2,2}.
\end{aligned}
$$

Therefore, combining the previous, we obtain

$$
\mathbb{E}\left[\left(\widetilde{X}^1-\mu_1\right)\left(\widetilde{X}^2-\mu_2\right)\right]=-\Sigma_{1,2}+2\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,-2}\Sigma_{-2,-2}^{-1}\Sigma_{-2,2}.
$$

Similarly as in the independent residuals setting, we can take the two dimension unit variance random variable to see that this will not coincide in general with the covariance ρ. Indeed, in this case,

$$
\mathrm{Cov}(\widetilde{X}^1,\widetilde{X}^2)=-\rho+2\rho^3.
$$

We notice that when the covariates are independent, the method with either independent or equal individual residual works. However, this is not the general case, therefore, in the next section we propose a methodology to correct this issue.

4.2.2 Sequential Conditional Independent Pairs with CPI sampling

We have observed that these knockoffs preserve all the required properties except the relationship between the knockoffs. This is because they were constructed in parallel and independently, without any guarantee to preserve their covariance. One possible solution could be to construct the knockoffs with this sampling step but in a sequential manner, ensuring that the exchangeability between the knockoffs is maintained. Algorithm 1, known as the Sequential Conditional Independent Pairs (Algorithm 1 of Candes et al. (2017)), presents this idea of sequential construction.

Algorithm 1 Sequential Conditional Independent Pairs

```
Input: An observation X, a conditional sampler
Output: A valid knockoff $\widetilde{X}$
for $j = 1$ to $p$ do
    Sample $\widetilde{X}^j$ from $\mathcal{L}\left(X^j \middle| X^{-j}, \widetilde{X}^{1:j-1}\right)$
end for
```

Algorithm 2 Sequential Conditional Independent Pairs with CPI sampling

```
Input: A matrix $\mathbf{X}$ and a train data set $\mathcal{D}_{\text{train}}$
Output: A valid knockoff $\widetilde{\mathbf{X}}$
for $j = 1$ to $p$ do
    Regress $\mathbf{X}^{\mathbf{j}}_{\text{train}}$ from $\mathbf{X}^{-\mathbf{j}}_{\text{train}}, \widetilde{\mathbf{X}}^{\mathbf{1:j-1}}_{\text{train}}$ to construct $\widehat{\nu}_{-j}$ with $\mathcal{D}_{\text{train}}$
    Permute the residuals and add them to the regression:
        $$\widetilde{\mathbf{X}}^{\mathbf{j}} = \widehat{\nu}_{-j}\left(\mathbf{X}^{-\mathbf{j}}, \widetilde{\mathbf{X}}^{\mathbf{1:j-1}}\right) + \left(\mathbf{X}^{\mathbf{j}} - \widehat{\nu}_{-j}\left(\mathbf{X}^{-\mathbf{j}}, \widetilde{\mathbf{X}}^{\mathbf{1:j-1}}\right)\right)^{\text{perm}}$$
    Construct the *train knockoff* using other permutation of the test residuals:
        $$\widetilde{\mathbf{X}}^{\mathbf{j}}_{\text{train}} = \widehat{\nu}_{-j}\left(\mathbf{X}^{-\mathbf{j}}_{\text{train}}, \widetilde{\mathbf{X}}^{\mathbf{1:j-1}}_{\text{train}}\right) + \left(\mathbf{X}^{\mathbf{j}} - \widehat{\nu}_{-j}\left(\mathbf{X}^{-\mathbf{j}}, \widetilde{\mathbf{X}}^{\mathbf{1:j-1}}\right)\right)^{\text{perm}}$$
end for
```

Moreover, they have proven exchangeability of the knockoffs with this procedure. We could adapt the sampling step to introduce the conditional permutation step to deal with the usual covariance matrix estimation problems in high-dimensional settings. Therefore, in practice, the algorithm could be rewritten as Algorithm 2.

We observe that the validity of this method relies on the ability to sample from the conditional distribution including the already generated knockoffs. To prove that our algorithm works under the Gaussian assumption, we only need to reuse Lemma 3.1 to show that, since it is a Gaussian vector, the conditional distribution can be rewritten as the sum of the regressed part (which is fixed) and a centered Gaussian with the correct covariance matrix, which is precisely given by the residuals distributions. Thus, the only remaining step would be to adapt Proposition 3.3 to this case, demonstrating that using a consistent estimate ensures the empirical distribution converges to the conditional distribution.

We observe that there is a trade-off involved in this algorithm: we sacrifice the computational advantage of parallel computation of each knockoff in order to achieve the theoretical guarantees of the sequential version. In practice, we could just consider the version presented by Blain et al. (2024) as they have demonstrated good empirical results with this approach.

4.3 Shapley-Knockoffs: FDR control using knockoffs

In this section, we present a method that controls the FDR by leveraging the theoretical results on the sampling step of the CPI, as discussed in this document, and the knockoff framework. This approach incorporates concepts from the game-theoretic variable importance measure, Shapley values (see Section 2.2), and addresses issues commonly encountered in high-correlation settings. It does so by considering the multiple relationships between the variable of interest and the other variables, rather than the *every variable vs. every variable excluding the studied one* approach used in LOCO and CPI methods.

First, notice that as proven in the previous section, Algorithm 2 provides valid knockoffs under Gaussian assumption. By defining a vector $\mathbf{W} = (W_1, \dots, W_p)$ where each coordinate j is given by

$$W_j = w_j([X, \widetilde{X}], y) := \left(y - \widehat{m}(\widetilde{X}^{(j)})\right)^2 - (y - \widehat{m}(X))^2 \,,$$

where $\widetilde{X}^{(j)}$ is the vector X with the j-th coordinate is replaced by its knockoff $\widetilde{X}^j$, we observe that it is a function of the input, the knockoff and the output. Then, we observe that it satisfies the flip-sign property:

$$w_j\left(\left[X, \widetilde{X}\right]_{\text{swap}(s)}, y\right) = \begin{cases} w_j\left(\left[X, \widetilde{X}\right], y\right) & \text{if } j \neq s, \\ -w_j\left(\left[X, \widetilde{X}\right], y\right) & \text{if } j = s, \end{cases}$$

for $s \in \{1, \dots, p\}$. However, to fulfill Definition 4.2, this property must hold for any subset s, not just those of size 1. Therefore, to leverage the knockoffs framework for FDR control, we need to modify the statistic to ensure its validity. One idea is to realize that with the current statistic, we are comparing the accuracy of the method by disabling the current coordinate without accounting for its relationship with the other coordinates. This is similar to the difference between LOCO and Shapley measures. Thus, the idea is to disable all covariates except the one being studied in a combinatorial way that satisfies the exchangeability property outlined in Definition 4.2. More formally, we define the Shapley-statistic as:

Definition 4.3 (Shapley-statistic)**.** It is the vector $\mathbf{W}_{\text{Shap}}$, where each coordinate W^j_{Shap} is given by

$$W^j_{\text{Shap}} := \sum_{x^1 \in \{X^1, \widetilde{X}^1\} \times \ldots \times x^p \in \{X^p, \widetilde{X}^p\}} (y - \widehat{m}(x^1, \dots, \widetilde{X}^j, \dots, x^p))^2 - (y - \widehat{m}(x^1, \dots, X^j, \dots, x^p))^2.$$

Therefore, using Shapley-Knockoffs will involve sampling the Knockoffs using the sequential algorithm presented in the previous section and applying the Shapley statistic to determine the importance of each covariate. We can easily recover the FDR-control of the procedure using the standard Knockoffs framework of Candes et al. (2017):

Proposition 4.4 (FDR-control of the Shapley-Knockoffs)**.** *Under Assumption 2 and assuming the consistance of the regressors $\widehat{m}$ and $\widehat{\nu}_{-j}$ for each $j \in \{1, \dots, p\}$, the procedure that consists on computing the Knockoffs following Algorithm 2 and using the standard Knockoffs threshold $T^\star_q$ defined in (17) on the Shapley-statistic from Definition 4.3, controls the FDR at level q.*

Proof. In order to establish FDR control using Theorem 3.4 from Candes et al. (2017), we need two main ingredients: first, the exchangeability of the knockoffs, and second, the validity of the Shapley statistic. The first was established in the previous section using the proof of Sequential Conditional Independent Pairs (Algorithm 1 from Candes et al. (2017)). This relies on the ability to sample from the conditional distribution sequentially. This is achieved through Assumption 2, which allows us to express the conditional distribution as the sum of two independent random variables, specifically the regressed part and the residuals. To address empirical error, we rely on the consistency of the regressors.

To proof the validity of the Shapley statistic we need to proof that the properties of Definition 4.2 are preserved. To see this, we first observe that

$$\begin{aligned} W^j_{\text{Shap}} &= w_j([X, \widetilde{X}], y) \\ &:= \sum_{x^1 \in \{X^1, \widetilde{X}^1\} \times \ldots \times x^p \in \{X^p, \widetilde{X}^p\}} (y - \widehat{m}(x^1, \ldots, \widetilde{X}^j, \ldots, x^p))^2 - (y - \widehat{m}(x^1, \ldots, X^j, \ldots, x^p))^2, \end{aligned}$$

therefore, it is obviously a function of the input X, the knockoff $\widetilde{X}$ and the output y. Then, to prove the flip-sign property, we first observe that for any subset $s \subset \{1, \ldots, p\}$, and for any $l \in s, l \neq j$ we have that

$$w_j([X, \widetilde{X}]_{\text{swap}(s)}, y) = w_j([X, \widetilde{X}]_{\text{swap}(s \setminus l)}, y)$$

because of the commutativity of the sum and because $x^l \in \{X^l, \widetilde{X}^l\}$ is equivalent to $x^l \in \{\widetilde{X}^l, X^l\}$. Therefore, we could proceed in this way to discard all the covariates from s but the j-th covariate. It the j-th covariate is s, then we can easily observe the antisymmetry property because of the antisymmetry of the subtraction. Therefore, we have that

$$w_j\left(\left[X, \widetilde{X}\right]_{\text{swap}(s)}, y\right) = \begin{cases} w_j\left(\left[X, \widetilde{X}\right], y\right) & \text{if } j \not\in s, \\ -w_j\left(\left[X, \widetilde{X}\right], y\right) & \text{if } j \in s. \end{cases}$$

Finally, we conclude that combining all the previous results and using the framework from Candes et al. (2017), FDR control is guaranteed.

□

In practical terms, however, it shares the same combinatorial challenges as the Shapley values. Nonetheless, there are several ways to avoid computing all combinations by leveraging solutions from the Shapley values literature. Some were already discussed in Section 2.2. For example, Bénard et al. (2022a) proposed addressing this issue through Monte-Carlo approach using importance sampling. Future research will focus on developing this method further and studying its performance in correlated settings.

In the following section, we will present another procedure that also controls the FDR, but without directly entering the knockoff framework, unlike the Shapley-knockoffs. Moreover, this procedure will focus uniquely on comparing *every variable vs. every variable excluding the studied one*, rather than considering all the relationships accounted for in the Shapley-knockoffs.

4.4 CPI-Knockoffs: FDR control via *approximate* knockoffs

In this section, we refer to the original CPI framework, where the conditional sampling of each covariate can be done independently and, therefore, in parallel. Given a trained model $\widehat{m}$, for each coordinate j, we compare the performance of the model when predicting the output using the original input, and when using the input where the j-th covariate is conditionally independent of the rest.

At first sight, it may seem similar to the holdout randomization test framework (see Tansey et al. (2021)), which is an specific conditional randomization test (see Candes et al. (2017)) that has the advantage of requiring only a single model refit. In this approach, an importance statistic $T(X, Y, \theta)$ is constructed using both the original test data $T^\star = T(X_{\text{test}}, Y_{\text{test}}, \widehat{\theta})$ and modified test data $T_{j,k} = T(\widetilde{X}^{(j)}_{\text{test},k}, Y_{\text{test}}, \widehat{\theta})$ for $k \in \{1, \ldots, K\}, j \in \{1, \ldots p\}$, where the j-th covariate is replaced by a generated null sample K times. Therefore, this null-generated importance statistic is created multiple times, and a p-value is computed based on the empirical CDF of the null test statistic.

For instance, FLOWSELECT (Hansen et al. (2022)) employs normalizing flows for density estimation and then utilizes an MCMC-based procedure to sample from the conditional distribution, which is used to generate each null statistic. However, after this process, they construct p-values, and to control the FDR, they rely on standard assumptions, such as those needed for the BH methodology(Giraud (2021)). Moreover, as other normalizing flows methods, its accuracy relies on the normalizing flows, which need a lot of training samples to estimate and sample from the density.

In contrast, our methodology diverges from this framework. Although we also sample from the conditional distribution using the CPI sampling methodology and we construct an importance statistic via the squared error, we do not require assumptions on p-values. This is because we leverage the knockoff framework, avoiding the strong PRDS assumptions that may not be met in conditional inference approaches.

Compared to the original CPI framework presented in Chamma et al. (2023), instead of controlling the type I error, we aim to control the FDR. The goal of controlling the FDR (16) is to ensure that, in the selected set, most of the covariates belong to the desired predictive set. This differs from type I error guarantees, which control the probability of incorrectly rejecting the null hypothesis for each covariate individually. This approach does not perform well with high-dimensional data, as the number of false positives increases significantly due to the variability of the random variables (see Giraud (2021)). Moreover, the CPI presented in Chamma et al. (2023) relies on its asymptotic convergence to a Gaussian random variable to provide type I error control, whereas this method could offer finite-sample FDR guarantees.

We recall that $\widetilde{X}^{(j)}$ stands for the vector where all the coordinates are equal to X but the j-th coordinate that is conditionally independent. We also denote by $\widetilde{X}$ the vector in which each coordinate is the conditionally independent sample.

In this context, we begin by defining the CPI statistic W_{CPI}, which is inspired by the knockoff statistics and measures the importance of each coordinate.

Definition 4.5 (CPI-Statistic)**.** Given $\widetilde{X}^{(j)}$ for each coordinate j, we define the CPI-Statistic W_{CPI} as the vector where each coordinate j is given by

$$W_{\text{CPI}}(X, \widetilde{X}, y)^j = \left(y - \widehat{m}(\widetilde{X}^{(j)})\right)^2 - \left(y - \widehat{m}(X)\right)^2 .$$

As discussed in Section 4.2, this approach does not directly fit within either the model-X knockoff framework or the original knockoff framework, as the vector $\widetilde{X}$ is not a knockoff. Neither the flip-sign property of the knockoff statistic is satisfied. However, it is not necessary for controlling the FDR. Indeed, as shown in Barber and Candès (2015), we only need to prove that, conditionally on the magnitude, the sign of the null covariates is i.i.d, or equivalently, that given $\epsilon \in \{\pm 1\}^d$ a sign sequence independent from W_{CPI}, with $\epsilon^j = 1$ for the non-null coordinates and $\epsilon^j \overset{\text{i.i.d}}{\sim} \{\pm 1\}$ independent Rademacher variables for the null coordinates, then

$$(W^1, \dots, W^d) \overset{\text{d}}{=} (\epsilon^1 W^1, \dots, \epsilon^d W^d).$$

To guarantee this, we assume that the difference between the predictions of X and $\widetilde{X}^{(j)}$ satisfies the sign-flip property.

Assumption 4 (Sign-flip difference)**.** Given $\widehat{m}$ a regressor of Y given X, for all the null covariates j, we have

$$\widehat{m}(X) - \widehat{m}(\widetilde{X}^{(j)}) \sim \epsilon^j \left(\widehat{m}(X) - \widehat{m}(\widetilde{X}^{(j)})\right) .$$

with ϵ^j an Rademacher distribution independent from the rest.

We first observe that if we had the theoretical regressor m, the difference would be exactly 0 since it does not depend on the null covariates (see Lemma 1.2 and Definition 1.1). Therefore, this difference can be viewed as an optimization error due to the variability of the covariates. Next, we note that it is a difference between two random variables that, conditionally on X^{-j}, are i.i.d., making the difference between them exchangeable. We are going to prove it in the linear setting.

example 4.6 (Linear model)**.** *For instance, the linear model satisfies this property. To see so, we first note that* $\widehat{m}(X) = \widehat{\beta}^\top X$. *Therefore,* $\widehat{m}(X) - \widehat{m}\left(\widetilde{X}^{(j)}\right) = \widehat{\beta}^j \left(X^j - \widetilde{X}^{(j)j}\right)$. *We note that as stated before, as the regressor improves its accuracy, the* $\widehat{\beta}^j \to 0$ *for the null covariates, so that this difference tends to 0.*

Given $\epsilon \in \{\pm 1\}^d$ *a vector of i.i.d. Rademacher variables, we are going to prove that*

$$\left(\widehat{m}(X) - \widehat{m}\left(\widetilde{X}^{(1)}\right), \dots, \widehat{m}(X) - \widehat{m}\left(\widetilde{X}^{(d)}\right)\right) \overset{\mathrm{d}}{=} \left(\epsilon^1\left(\widehat{m}(X) - \widehat{m}\left(\widetilde{X}^{(1)}\right)\right), \dots, \epsilon^d\left(\widehat{m}(X) - \widehat{m}\left(\widetilde{X}^{(d)}\right)\right)\right),$$

which also proofs the property when restricted to the null covariates.

Using the linear assumption, we only need to prove that

$$\left(X^1 - \widetilde{X}^{(1)1}, \dots, X^d - \widetilde{X}^{(d)d}\right) \overset{\mathrm{d}}{=} \left(\epsilon^1\left(X^1 - \widetilde{X}^{(1)1}\right), \dots, \epsilon^d\left(X^d - \widetilde{X}^{(d)d}\right)\right).$$

We start by noticing that for any j,

$$X^j - \widetilde{X}^{(j)j} = X^j - \left(\mathbb{E}\left[X^j | X^{-j}\right] + X'^j - \mathbb{E}\left[X'^j | X'^{-j}\right]\right) = \left(X^j - \mathbb{E}\left[X^j | X^{-j}\right]\right) - \left(X'^j - \mathbb{E}\left[X'^j | X'^{-j}\right]\right),$$

with X' *independent from* X. *Therefore, each coordinate is independent from the rest. Indeed,* X *is independent from* X' *and by using the regression model assumption (Assumption 3),* $X^j - \mathbb{E}\left[X^j | X^{-j}\right]$ *is independent of* X^{-j}. *To sum up, taking arbitrarily the first coordinate* $X^1 - \widetilde{X}^{(1)1}$, *it is independent from the rest of coordinates and it can be decomposed as* $(X^1 - \mathbb{E}\left[X^1 | X^{-1}\right]) - (X'^1 - \mathbb{E}\left[X'^1 | X'^{-1}\right])$, *so it is a difference of two i.i.d. random variables so we can multiply it with an independent Rademacher variable giving the result.*

Even though this assumption may seem strong, it is easily satisfied in practice. Graphically, we can observe that each coordinate of the vector is symmetrically centered at zero. For example, as shown in Figure 16, we test this assumption in a more complex setting.

We also observe that, even though this property is satisfied for both important and unimportant covariates, if the regressor is consistent, this quantity will vanish for the unimportant covariates, but not for the important ones. This is seen in the histograms 8 and 12 as the scale of the important covariates is larger than the ones of the unimportant ones. This is the property that will be exploited to effectively detect the null covariates.

Finally, we establish control of the FDR for the proposed method by leveraging the knockoffs framework.

Theorem 4.7 (FDR-control on the CPI)**.** *Under Assumption 3 and Assumption 4, using the knockoff threshold (17) on the* W_{CPI}, *the FDR of the CPI procedure is controlled.*

Proof. We only need to prove

$$(W^1_{\mathrm{CPI}}, \dots, W^d_{\mathrm{CPI}}) \overset{\mathrm{d}}{=} (\epsilon^1 W^1_{\mathrm{CPI}}, \dots, \epsilon^d W^d_{\mathrm{CPI}}),$$

with $\epsilon^j = 1$ if j is genuine and an independent Rademacher otherwise. After that, we can conclude using the exact same proof of Theorem 1 and 2 of Barber and Candès (2015). We observe that this is equal to prove that each null coordinate is equally distributed to its opposite, i.e. $W^j_{\text{CPI}} \overset{\text{d}}{=} -W^j_{\text{CPI}}$ for $j \in \mathcal{H}_0$. We begin by decomposing each null coordinate:

$$\begin{aligned}
W^j_{\text{CPI}} &= \left(Y - \widehat{m}(\widetilde{\mathbf{X}}^{(\mathbf{j})})\right)^2 - (Y - \widehat{m}(\mathbf{X}))^2 \\
&= \widehat{m}\left(\widetilde{X}^{(j)}\right)^2 - \widehat{m}(X)^2 - 2Y\widehat{m}\left(\widetilde{X}^{(j)}\right) + 2Y\widehat{m}(X) \\
&= \left(\widehat{m}\left(\widetilde{X}^{(j)}\right) - \widehat{m}(X)\right)\left(\widehat{m}(X) + \widehat{m}\left(\widetilde{X}^{(j)}\right)\right) + 2Y\left(\widehat{m}(X) - \widehat{m}\left(\widetilde{X}^{(j)}\right)\right) \\
&= \left(\widehat{m}\left(\widetilde{X}^{(j)}\right) - \widehat{m}(X)\right)\left(\widehat{m}(X) + \widehat{m}\left(\widetilde{X}^{(j)}\right) - 2Y\right).
\end{aligned}$$

We observe that using Assumption 4 we have

$$\widehat{m}\left(\widetilde{X}^{(j)}\right) - \widehat{m}(X) \overset{\text{d}}{=} \widehat{m}(X) - \widehat{m}\left(\widetilde{X}^{(j)}\right).$$

We conclude by using the commutativity of addition and repeating the same steps in reverse to recover $-W^j_{\text{CPI}}$. □

We observe that under this knockoff framework, rather than simply controlling the asymptotic type-I error provided by the CPI (see Chamma et al. (2023)), we can guarantee the control in finite distance of the FDR.

We observe that this result pertains only to the accuracy of each knockoff draw. There is a need to aggregate the knockoffs to obtain more robust results (Nguyen et al. (2020)). Notably, by reusing the same proof, we can extend the guarantees to the mean, rather than just the accuracy of each individual. This is what we study in practice. Other methods of aggregating individual knockoffs will be considered in future research. For instance, the quantile aggregation Nguyen et al. (2020) of to aggregate the individuals to create a p-value from the empirical CDF distribution as in the Holdout Randomization Test (Tansey et al. (2021)).

4.5 Experiments on the power of the statistic

In this section, we study the performance of the method in identifying important covariates. To do so, we first plot histograms of $\widehat{m}(X) - \widehat{m}(\widetilde{X}^{(j)})$ for each covariate j, verifying that they are centered and symmetric, thus fulfilling Assumption 4. To test it more formally, a Kolmogorov-Smirnov test was also done. We observe that the main difference between null and non-null covariates is reflected in the variance of the histogram. Next, we plot the error for each individual when predicting with the modified input. Typically, the error increases more when important covariates are modified (red crosses). We observe that there is often significant variability, so we average the accuracy across the samples to achieve more robust results. In these figures, we see that the red crosses are clearly separated from the blue dots, demonstrating the power of this method. Finally, we plot histograms of the error increase caused by modifying the null covariates. We observe that these errors are highly concentrated around 0, showing the strength of the method, and that the histograms are symmetric, providing numerical evidence for Theorem 4.7, which provides the FDR control by applying the threshold presented in Candes et al. (2017).

In practice, we use complex and lengthy methods to train the model $\widehat{m}$ to explain y given X, but we employ simpler and faster methods to compute each $\widehat{\nu}_{-j}$ for sampling from the conditional distribution $\widetilde{X}^{(j)}$. For all methods, we use a training sample of $n = 1000$, with 700 used to train the models $\widehat{m}$ and $\widehat{\nu}_{-j}$ for each j, and 300 used to obtain residuals to compute the conditional distributions, and calculate the statistic.

Linear setting: the relationship between y and X is linear and the dimension is fixed to $d = 500$. The correlation between the coordinates of X is a ρ Toeplitz matrix, i.e. the correlation between the coordinates i and j is $\rho^{|i-j|}$. We fix $\rho = 0.6$.

In Figure 8 we can observe that Assumption 4 is satisfied. This was also done more formally using a two-sample test, comparing this data with randomly sign-flipped data, which determined that we could not reject the hypothesis that they came from the same distribution.

We observe that the first row, consisting of the important covariates, exhibits much higher variability than the second row. This property provides the discriminatory power of our method.

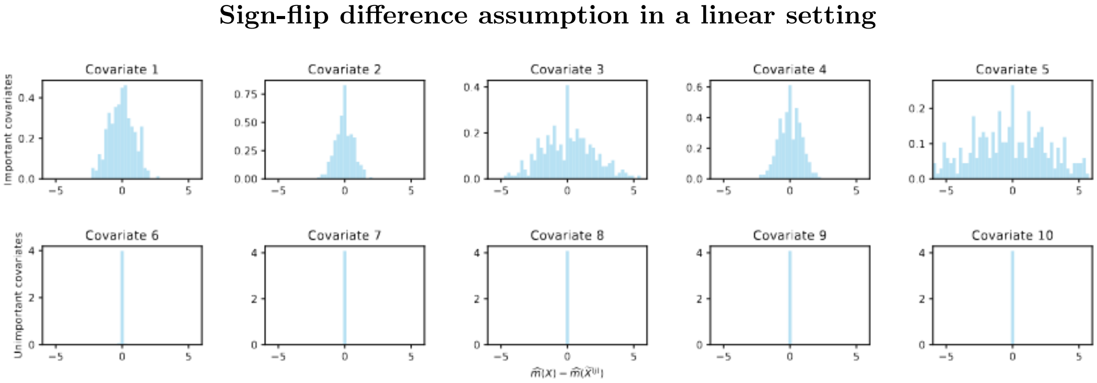

Figure 8: **Setting:** $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. Each plot corresponds to a specific coordinate. Thus, each histogram represents the distribution of the difference between the model's prediction using a test individual and using the same individual with the covariate conditionally independently sampled.

In Figure 9, we observe that the red crosses generally stand out from the blue points, but due to the variability among individuals, there are instances where they overlap. However, in Figure 10, where the test individuals have been aggregated, there is a clear distinction between the red crosses and the blue points.

In Figure 11, we observe that the error difference is centered at 0 and symmetric, providing a numerical example of Theorem 4.7. Additionally, the fact that it is so concentrated around 0 highlights the power of this method, as all important covariates will exhibit a greater difference, enabling the perfect recovery of the relevant covariates.

High-dimensional linear setting: In this experiment we recover the same structure for the dependence between the covariates and the same number of samples, but we take a dimension $d = 5000$. The same models are used to fit the data, therefore the accuracy is worse than in the previous setting. Nevertheless, the important covariates are still separable from the null ones as we will see in Figure 14.

In Figure 12, as before, we can observe that Assumption 4 is satisfied. This was also done more formally using a two-sample test, comparing this data with randomly sign-flipped data, which determined that we could not reject the hypothesis that they came from the same distribution.

In Figure 13, we observe that the red crosses generally stand out from the blue points, but due to the variability among individuals, there are instances where they overlap.However, in Figure 14, where the test individuals have been aggregated, the distinction between the red crosses and the blue

Approximate-knockoff statistic for 10 observations in a linear setting

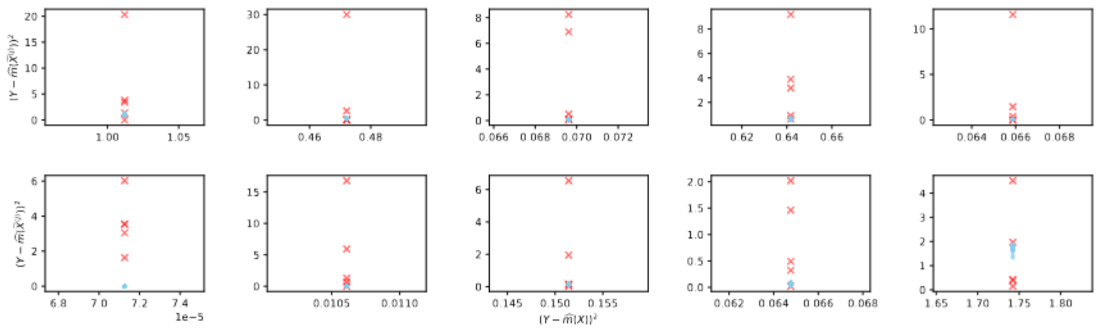

Figure 9: **Setting:** $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. Each plot represents an individual. On the x-axis, we have the prediction error made on the individual, and on the y-axis, the error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

points becomes clearer. Although the model is less accurate than in the previous setting, giving the impression of some overlap, the augmented figure on the right shows that this is not the case, and they remain separable.

In Figure 15, we observe that the error difference is centered at 0 and symmetric, providing a numerical example of Theorem 4.7. Additionally, the fact that it is so concentrated around 0 highlights the power of this method, as all important covariates will exhibit a greater difference, enabling the perfect recovery of the relevant covariates. We also note that as the model is not accurate, the variability is lightly increased compared to the first setting.

Non-linear setting: In this case we take a non-linear relationship between the output y and the input X, given by $y = X_0 X_1 \mathbb{1}_{X_2>0} + 2X_3 X_4 \mathbb{1}_{X_2<0}$. The correlation between the input covariates is preserved as before by a Toeplitz matrix with $\rho = 0.6$. The input covariates are centered in $\mathbf{1}$. The training and test sample is of the same size as before. We used the same models to fit the data as before.

First, we note from Figure 16 that the symmetry required by Assumption 4 is fulfilled. Second, we observe that the histograms of the first and second covariates are much more spread out compared to the other three important covariates. As shown on the left of Figure 18, these covariates are therefore more easily differentiated from the null covariates than the other important covariates. However, all the important covariates remain separable from the null ones. We note that this is because, in the relationship between the input and the output, although the second part of the sum has a higher coefficient, the indicator function will be 0 more often than in the first part, as the covariate X_2 is centered at 1.

In Figure 17, we observe that the red crosses generally stand out from the blue points, but due to the variability among individuals, there are instances where they overlap, mostly for X_2, X_3 and X_4. However, in Figure 14, where the test individuals have been aggregated, there is a clear distinction between the red crosses and the blue points. On the right we have the augmented figure for readability of the result.

In Figure 19, we observe that the error difference is centered at 0 and symmetric, providing

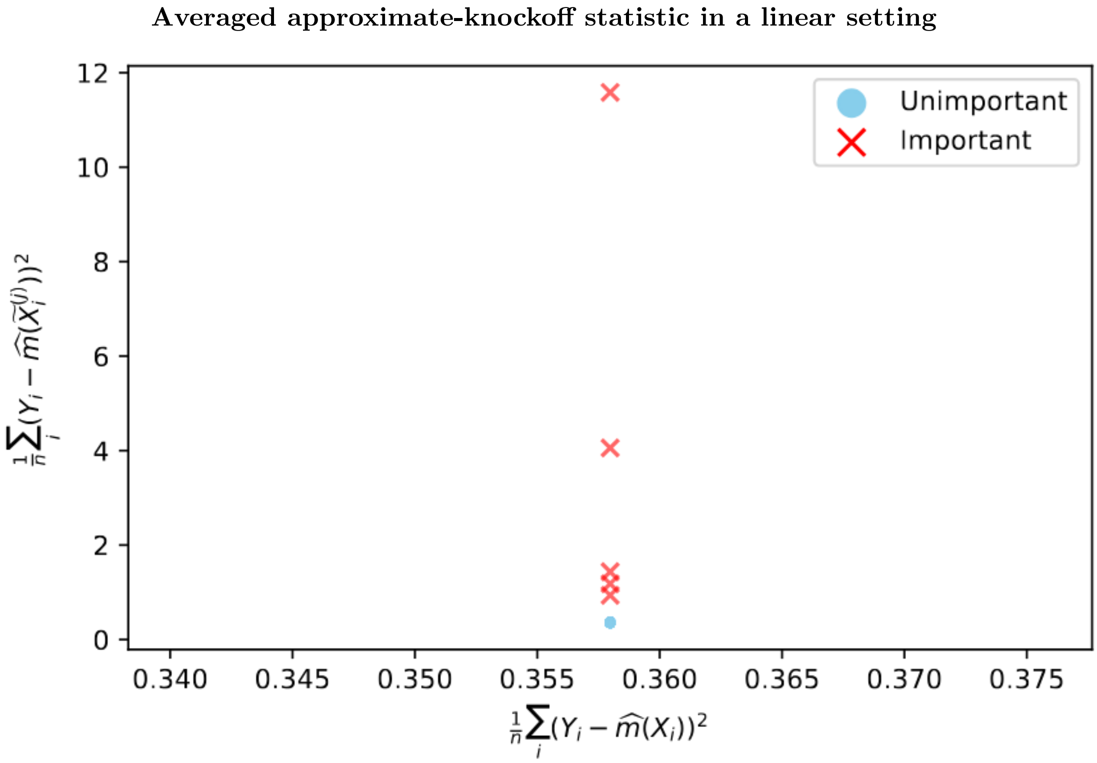

Figure 10: **Setting:** $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. It represents the mean of the errors made across the individuals. On the x-axis, we have the mean prediction error, and on the y-axis, the mean error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

a numerical example of Theorem 4.7. Additionally, the fact that it is so concentrated around 0 highlights the power of this method, as all important covariates will exhibit a greater difference, enabling the perfect recovery of the relevant covariates. We observe that, since this data setting is more challenging, the histogram is more spread out compared to the other two cases.

5 Conclusion/ Perspectives

5.1 Conclusions

This work has reviewed the most widely used methods for measuring variable importance. First, we observe that although removal-based approaches, such as LOCO and Shapley values, may provide theoretically desirable quantities, they suffer from practical limitations that make them difficult to implement. On the other hand, while permutation-based approaches offer good stability and computational feasibility, they do not directly compute any theoretically desired quantities. To address this, we have established a connection between conditional permutation importance and

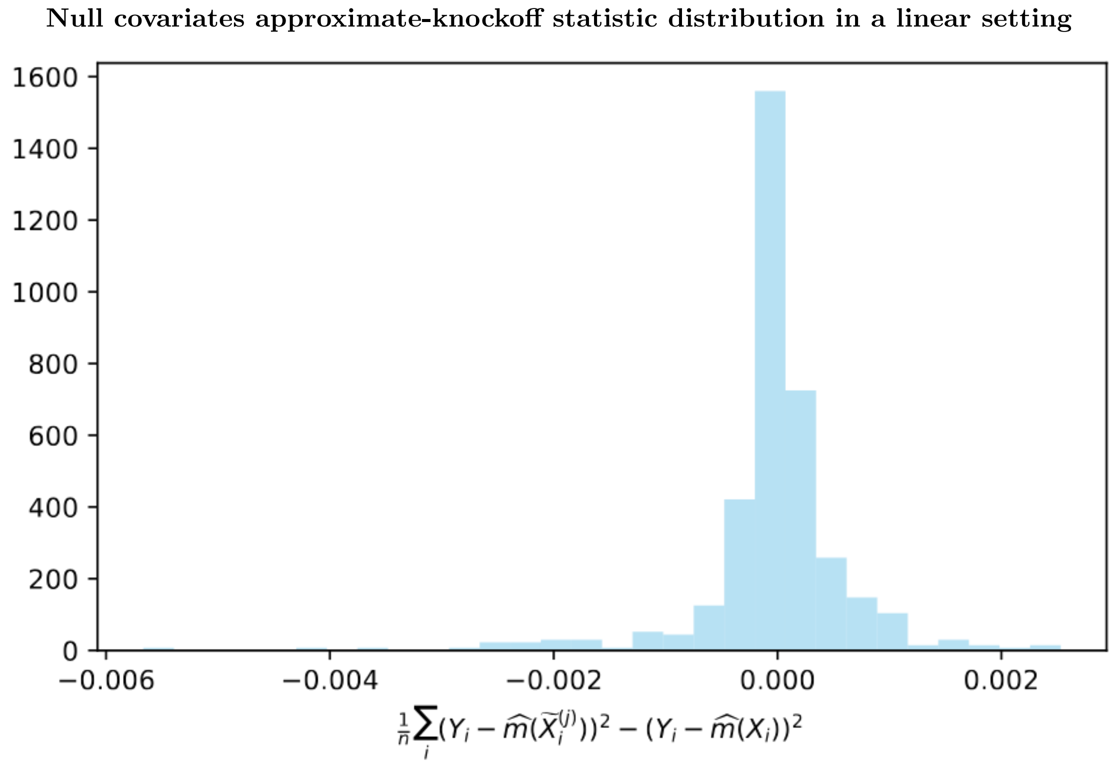

Figure 11: **Setting:**$y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. It represents the histogram of the mean of difference between the errors made across the individuals by using the original and a conditionally independent sample on the null covariates.

LOCO, resulting in a stable estimate of LOCO. This was achieved by identifying a theoretical framework where the conditional sampling step is valid. Additionally, we introduced an aggregated version of this estimate, which adds no computational complexity but may help mitigate extrapolation issues.

In the context of controlled variable selection, we also presented an algorithm that leverages the previously established theoretical framework for conditional sampling to produce valid knockoffs. We then introduced a Shapley-based statistic that accounts for all predictive relationships across subsets, integrating directly into the knockoff framework and providing False Discovery Rate control for the selected subset. The main drawback of this knockoff algorithm is its sequential nature, which looses the parallelizability of the CPI. To overcome this limitation, we introduced a CPI-based method that, while not directly part of the knockoff framework, also provides FDR control and is parallelizable. This method enhances the statistical control of CPI under mild assumptions, and in practice, its discriminatory power depends on the accuracy of the trained model. By using consistent estimates, this approach offers a powerful methodology that separates the true signal from optimization errors. Moreover, it does not directly fit within the Holdout Randomization Test framework, but still provides the same statistical guarantees without requiring the strong dependency assumptions between p-values, which do not align well with conditional approaches.

Sign-flip difference assumption in a high-dimensional linear setting

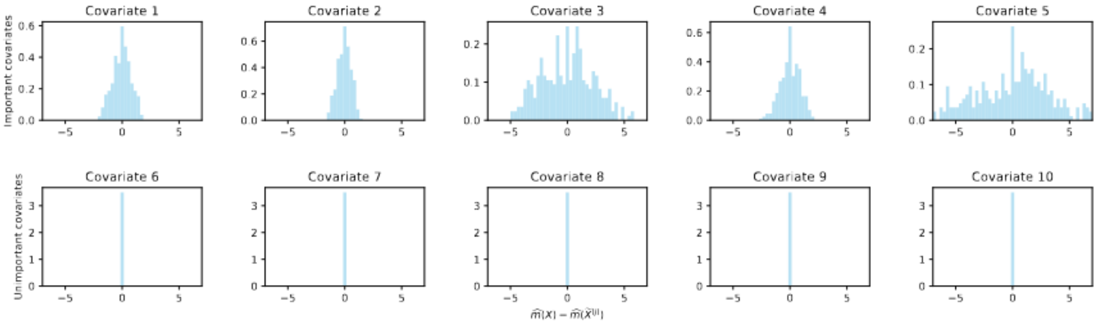

Figure 12: **Setting:** $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. Each plot corresponds to a specific coordinate. Each histogram represents the distribution of the difference between the model's prediction using a test individual and using the same individual with the covariate conditionally independently sampled.

5.2 Perspectives

Much work remains to build on the presented advances, particularly in practical applications to demonstrate the benefits of the proposed methods. For instance, future work will compare the CPI-knockoff with the current state-of-the-art method, as presented in Candes et al. (2017), which uses the difference between the LASSO coefficients of the original covariate and its knockoff counterpart as the test statistic. This approach primarily works in linear settings. However, our proposed method should also work in non-linear contexts, as it relies on the performance of the model $\widehat{m}$. With a sufficiently flexible and well-trained model, we should be able to efficiently recover the true signal.

Additionally, following the work of Blain et al. (2023), instead of focusing on controlling the False Discovery Rate, we aim to control the False Discovery Proportion, which offers stronger statistical guarantees. After all, we are less concerned with controlling the average False Discovery Proportion across multiple datasets and more interested in having theoretical guarantees for the selected set in our specific dataset.

Moreover, while we have aggregated the CPI-knockoffs by simply averaging, other possible methods of aggregation are available (see Nguyen et al. (2020)). Future research will explore these alternatives to potentially improve the performance of the method.

Finally, it is also worth noting that when dealing with highly correlated covariates, it may be beneficial to consider groups of covariates rather than individual ones. For example, in brain imaging, instead of analyzing each voxel independently, it may be more appropriate to consider regions of voxels. This approach aligns with the issue of having highly correlated covariates, where none is truly important because the other one is available, but they are jointly important. This also relates to the critic on the axioms of the Shapley values, which assume linearity across subsets, implying that the importance of two covariates should equal the sum of their individual importances.

Approximate-knockoff statistic for 10 observations in a high-dimensional linear setting

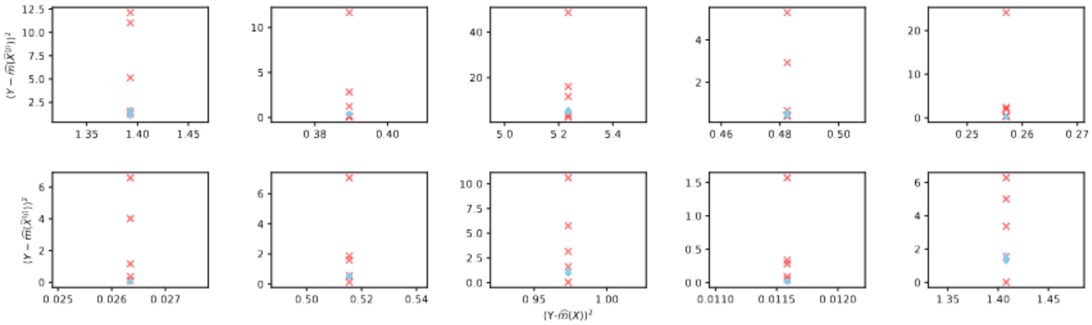

Figure 13: **Setting:** $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. Each plot represents an individual. On the x-axis, we have the prediction error made on the individual, and on the y-axis, the error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

References

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. (2022). Minimax rate of consistency for linear models with missing values.

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society series b-methodological*, 57:289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165 – 1188.

Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2023). False discovery proportion control for aggregated knockoffs.

Blain, A., Thirion, B., Linhart, J., and Neuvial, P. (2024). When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bénard, C., Biau, G., da Veiga, S., and Scornet, E. (2022a). Shaff: Fast and consistent shapley effect estimates via random forests.

Bénard, C., da Veiga, S., and Scornet, E. (2022b). MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2017). Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection.

Averaged approximate-knockoff statistic in a high-dimensional linear setting

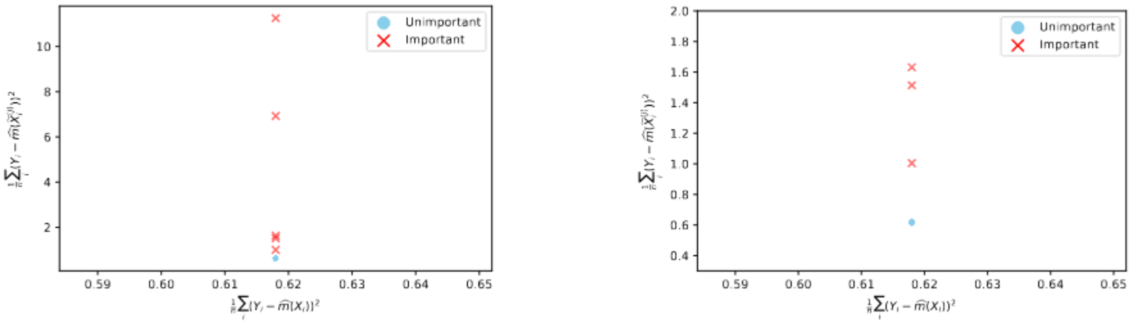

Figure 14: **Setting:** $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. The figure represents the mean of the errors made across the individuals. On the x-axis, we have the mean prediction error, and on the y-axis, the mean error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

Chamma, A., Engemann, D. A., and Thirion, B. (2023). Statistically valid variable importance assessment through conditional permutations.

Covert, I., Lundberg, S. M., and Lee, S. (2020). Explaining by removing: A unified framework for model explanation. *CoRR*, abs/2011.14878.

Giraud, C. (2021). *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 2nd edition.

Hansen, D., Manzo, B., and Regier, J. (2022). Normalizing flows for knockoff-free controlled feature selection.

Homma, T. and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. (2020). Problems with shapley-value-based explanations as feature importance measures.

Lobo, A. D. R., Ayme, A., Boyer, C., and Scornet, E. (2024). Harnessing pattern-by-pattern linear classifiers for prediction with missing data.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Matloff, N. and Mohanty, P. (2023). *A Method for Handling Missing Values in Prediction Applications*. R package version 0.1.0.

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. Published: 21 May 2021.

Null covariates approximate-knockoff statistic distribution in a high-dimensional linear setting

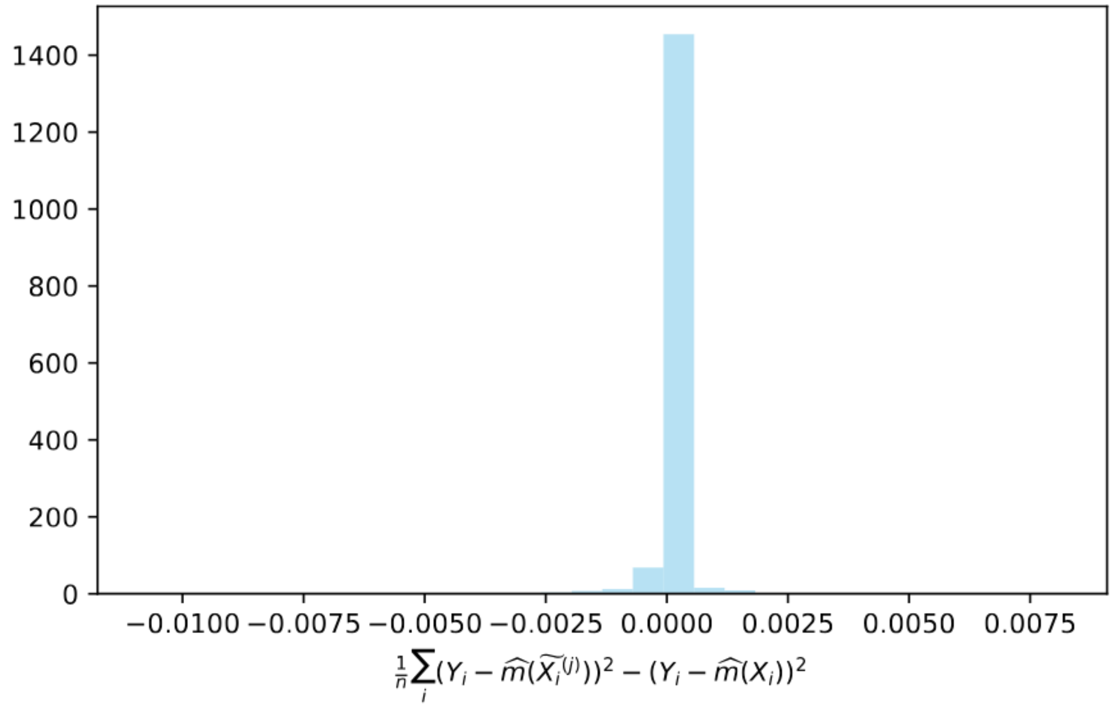

Figure 15: **Setting:**$y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. It represents the histogram of the mean of difference between the errors made across the individuals by using the original and a conditionally independent sample on the null covariates.

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.

Nguyen, T.-B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2020). Aggregation of Multiple Knockoffs. In *ICML 2020 - 37th International Conference on Machine Learning*, number 119 in Proceedings of the ICML 37th International Conference on Machine Learning,, Vienne / Virtual, Austria.

Owen, A. B. (2014). Sobol' indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251.

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

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.

Sign-flip difference assumption in a non-linear setting

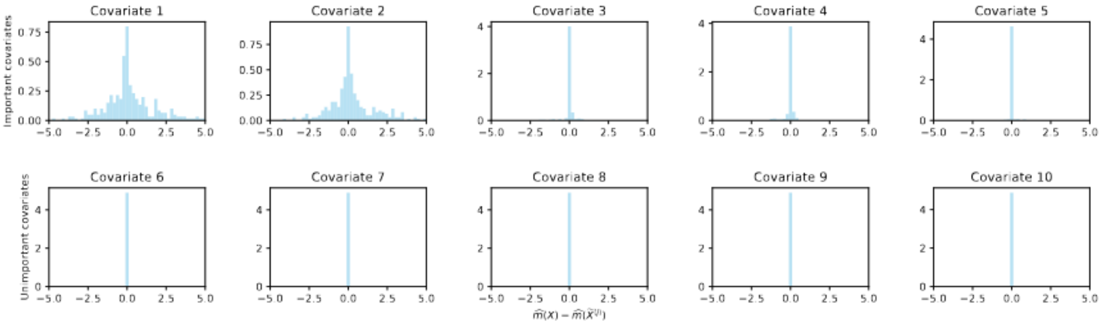

Figure 16: **Setting:** $y = X_1X_2\mathbb{1}_{X_3>0} + 2X_4X_5\mathbb{1}_{X_3<0}$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. Each plot corresponds to a specific coordinate. Thus, each histogram represents the distribution of the difference between the model's prediction using a test individual and using the same individual with the covariate conditionally independently sampled.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3).

Song, E., Nelson, B. L., and Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. (2021). The holdout randomization test for feature selection in black box models.

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

Verdinelli, I. and Wasserman, L. (2023). Feature importance: A closer look at shapley values and loco.

Williamson, B. and Feng, J. (2020). Efficient nonparametric statistical inference on population feature importance using shapley values. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10282–10291. PMLR.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021a). 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. (2021b). A general framework for inference on algorithm-agnostic variable importance.

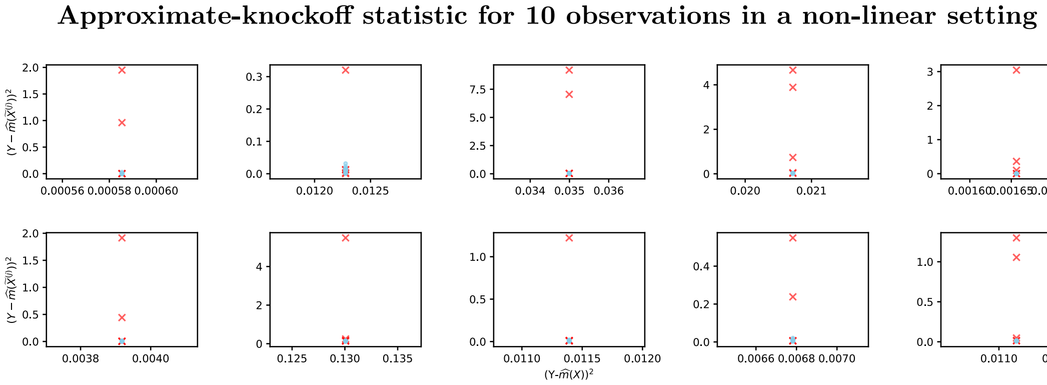
Approximate-knockoff statistic for 10 observations in a non-linear setting

Figure 17: **Setting:** $y = X_1X_2\mathbb{1}_{X_3>0} + 2X_4X_5\mathbb{1}_{X_3<0}$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. Each plot represents an individual. On the x-axis, we have the prediction error made on the individual, and on the y-axis, the error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

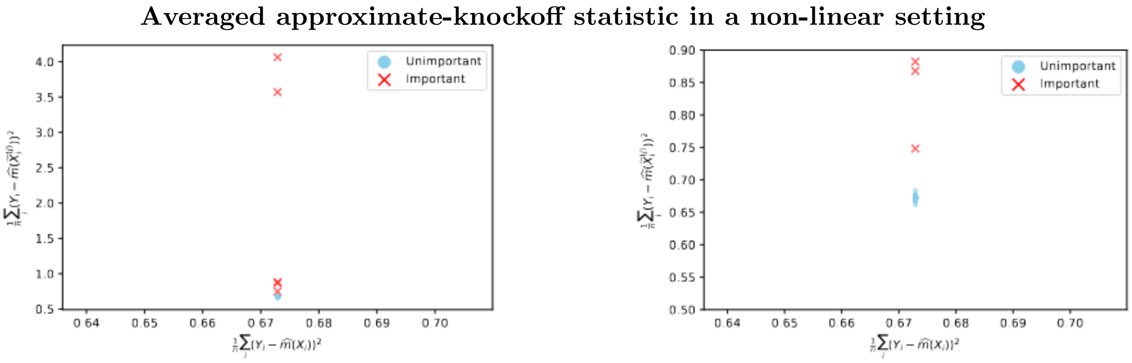
Averaged approximate-knockoff statistic in a non-linear setting

Figure 18: **Setting:** $y = X_1X_2\mathbb{1}_{X_3>0} + 2X_4X_5\mathbb{1}_{X_3<0}$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. It represents the mean of the errors made across the individuals. On the x-axis, we have the mean prediction error, and on the y-axis, the mean error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

Null covariates approximate-knockoff statistic distribution in a non-linear setting

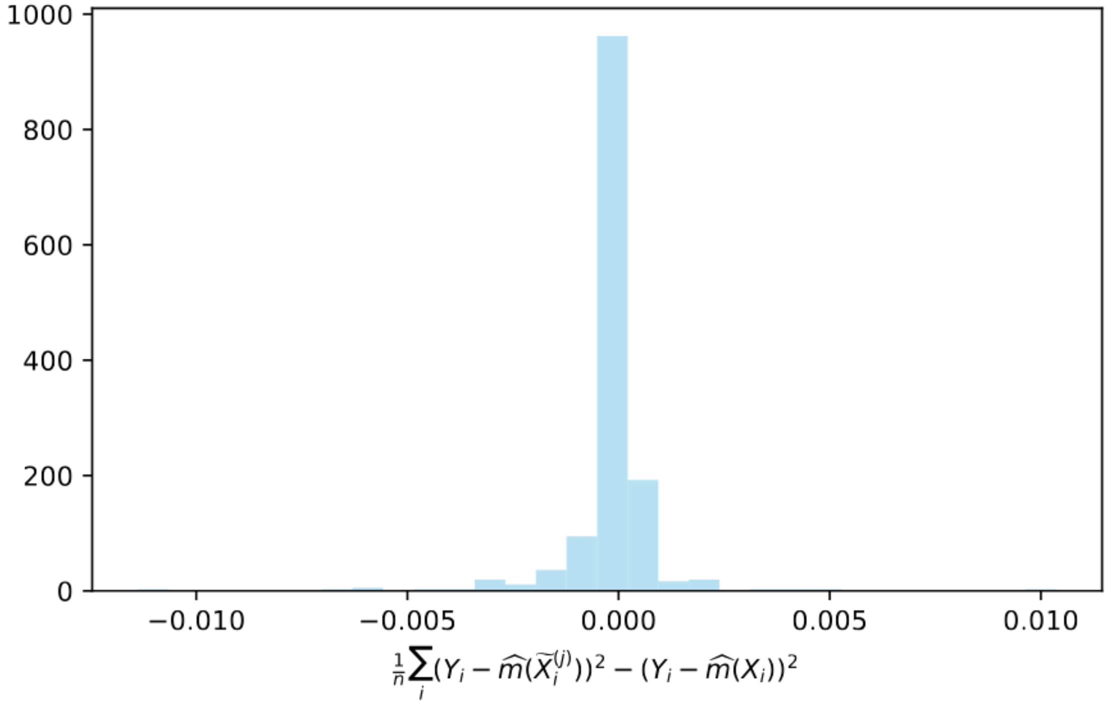

Figure 19: **Setting:**$y = X_1 X_2 \mathbb{1}_{X_3>0} + 2X_4 X_5 \mathbb{1}_{X_3<0}$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. It represents the histogram of the mean of difference between the errors made across the individuals by using the original and a conditionally independent sample on the null covariates.

A Some explicit LOCO examples

example A.1 (LM with two Gaussian covariates)**.** *Given two Gaussian coviates X_0 and X_1 with a correlation ρ we note that under the linear model setting $Y = \beta_0 X_0 + \beta_1 X_1 + \epsilon$, LOCO can be expressed as*

$$\begin{aligned}
\psi_{\text{LOCO}}(j, P_0) &= \mathbb{E}\left[(m(X) - m_{-j}(X^{-j}))^2\right] \\
&= \beta_j^2 \mathbb{E}\left[(X^j - \mathbb{E}\left[X^j | X^{-j}\right])^2\right] \\
&= \beta_j^2 \mathbb{E}\left[(X^j - \mathbb{E}\left[X^j\right] - \frac{\text{Cov}(X^j, X^{-j})}{\mathbb{V}(X^{-j})}(X^{-j} - \mathbb{E}\left[X^{-j}\right]))^2\right] \\
&= \beta_j^2 \left(\mathbb{V}(X^j) + \frac{\text{Cov}(X^j, X^{-j})^2}{\mathbb{V}(X^{-j})}\mathbb{V}(X^{-j}) - 2\frac{\text{Cov}(X^j, X^{-j})^2}{\mathbb{V}(X^{-j})}\right) \\
&= \beta_j^2 \left(\mathbb{V}(X^j) - \frac{\rho^2}{\mathbb{V}(X^{-j})}\right).
\end{aligned}$$

example A.2 (Explicit LOCO in a non-linear setting)**.** *In this example we will recover the example of no-linear setting from Bénard et al. (2022b) but changing the input covariance matrix to obtain more complex relationships between the covariates. Indeed, we will have $y = \alpha X^0 X^1 \mathbb{1}_{X^2>0} + \beta X^3 X^4 \mathbb{1}_{X^2<0}$, 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 LOCO for the covariate X^0.*

First, we observe that

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

Then, we can develop LOCO as

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

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

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

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, $\psi_{\text{LOCO}}(0, P_0) = \alpha^2/2\mathbb{E}\left[\left(X^0 - \mathbb{E}\left[X^0 | X^{-0}\right]\right)^2\right] = \alpha^2/2\mathbb{E}\left[\mathbb{V}(X^0 | X^{-0})\right]$. Note that as it a Gaussian vector, the 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, $\psi_{\mathrm{LOCO}}(0, P_0) = (1-\rho^2)/2$. *Similarly, for the first covariate we obtain* $\psi_{\mathrm{LOCO}}(1, P_0) = \rho^2/2(1 - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})$.