Towards Trustworthy Feature Importance by Avoiding Unrestricted Permutations

Emanuele Borgonovo
Department of Decision Sciences and Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy, emanuele.borgonovo@unibocconi.it

Francesco Cappelli
Department of Decision Science, Bocconi University, Milan, Italy, francesco.cappelli@phd.unibocconi.it

Xuefei Lu
SKEMA Business School, Université Côte d'Azur, France, xuefei.lu@skema.edu

Elmar Plischke
Clausthal University of Technology, Clausthal-Zellerfeld, Germany, elmar.plischke@tu-clausthal.de

Cynthia Rudin
Department of Computer Science, Duke University, Durham NC, 27708 USA, cynthia@cs.duke.edu

The determination of feature importance is essential for explainability. However, methods based on unrestricted permutations deliver incorrect insights due to extrapolation errors; this is a problem that affects essentially all non-trivial variable importance metrics to date, leading them to make untrustworthy estimates of the importance of variables. We introduce and compare alternative approaches to address the problem. First, we present a design that combines the calculation of conditional model reliance with a Gaussian transformation of the features to obtain new data points that lie close to the original ones, but that exclude the unique information from the feature of interest. Second, we propose a strategy that combines Knockoffs for the generation of the new data and a Gaussian transformation. Third, we consider the design of accumulated local effect (ALE) plots to define variable importance indices. We establish the theoretical connections of the new indices with total effects under a quadratic loss. We examine the performance of the new indices through extensive numerical experiments, using analytical test cases in which the ground truth importance of the variables is known, as well as two well-known datasets: Boston Housing and the recently introduced Voter-Name-Ethnicity Prediction dataset. Results reveal that the proposed permutation strategies successfully decrease extrapolation and are much more trustworthy for estimating the importance of variables.

Key words: Machine Learning; Variable Importance; Explainability; Explainable AI; XAI; Permutation Importance; Variance-Based Sensitivity Indices; ALE plots

sec:intro

1. Introduction.

Management scientists increasingly rely on complex machine learning (ML) tools to address problems across diverse domains, from advertising (Souyris et al. 2023) to finance (Chen et al. 2024). However,

the successful utilization of ML models can be undermined by a lack of transparency. Providing reliable explanations becomes critical in increasing trust in the model results, troubleshooting, assessing fairness, and in defending the overall modeling approach (Murdoch et al. 2019).

There are many explainability methods that estimate feature importance, including Breiman's variable importance measures (Breiman 2001), Shapley values (Lundberg and Lee 2017, Sundararajan and Najmi 2020), and graphical indicators such as partial dependence plots (Friedman 2001). All of these approaches use *permutations*, where parts of data points are rotated and the change in performance is assessed. Even variable importance measures that are total indices (Homma and Saltelli 1996, Bénard et al. 2022, Verdinelli and Wasserman 2023) are equivalent to using permutations.

A major disadvantage of these types of approaches is that all the permutations are *unrestricted*. Randomly permuting the entries of a dataset may create new points far from the original data or even "impossible data." In a permuted dataset, we may find children who are retired or individuals who graduated from high school before they were born (Mase et al. 2022, p. 1). Forcing ML models to make predictions at these points causes them to extrapolate, making explanations unreliable (Hooker et al. 2021).

There have been a few recent attempts to deal with this problem. Apley and Zhu (2020) aimed to repair Friedman's partial dependence plots to remove unrestricted permutations by introducing accumulated local effect (ALE) plots. ALE plots limit extrapolations by constraining features to assume values on a predetermined grid, and movement is limited to the nearest grid point. Fisher et al. (2019) introduce a method that inadvertently reduces extrapolation: their goal was to distinguish the *overall* and the *unique* importance of a variable. They introduced the notion of conditional model reliance (CMR) that allows us to find the unique information of the variable that cannot be gleaned from other variables or their combinations. Hooker et al. (2021) found that CMR is also exposed to the risk of calculating the model on impossible values, though less so because the only quantity being permuted is the variable's unique information (regression residuals from imputing the function values). Hooker et al. (2021) introduced the idea of using Knockoffs with permutation importance to have a similar effect of reducing impossible values. Ideally, however, a variable importance metric would not use impossible values at all. Our hypothesis is that such metrics would yield more reliable variable importance estimates.

We propose three strategies that substantially reduce impossible data. The first combines the calculation of conditional model reliance with a Gaussian transformation. The intuition is to map the quantiles from our data to the quantiles of a Gaussian distribution before computing conditional model reliance, and map back to the original space afterwards. This way, only the quantiles of the point values are adjusted, which dramatically reduces extrapolation. We prove that under a Gaussian copula assumption for the feature distribution, the new data points follow the same probability

distribution as the original data (Proposition 1). The method is advantageous on several accounts: it allows analysts to compute conditional variable importance indices that contain only the unique information of the variables, it substantially reduces extrapolation in variable importance calculation, and it is computationally frugal.

The second strategy is based on combining the Gaussian transformation with the Knockoffs method. As discussed, Hooker et al. (2021) used Knockoffs to restrict permutations in Breiman's variable importance measures. We improve over that approach by preceding the Knockoffs generation by a Gaussian transformation and mapping the new points back into the original space after permuting. This reduces extrapolation in variable importance.

The third strategy combines Apley and Zhu's ALE plot design with Jansen's estimator (Jansen 1999) of total indices, obtaining a total-effect-like index called τ^{ALE}. Because this index produces numerical noise when the ALE grid becomes more and more refined, we introduce a second index, κ^{ALE}, built as an average of squared Newton ratios.

We derive the theoretical relationship between permutation-based importance measures and total indices, with and without permutation restrictions.

We carry out a detailed examination of the numerical behavior of all these feature importance measures to determine which are able to reliably pinpoint important variables. We start with experiments that have analytical benchmarks, including Hooker et al. (2021)'s test case and the Ishigami function (Ishigami and Homma 1990). We find that variable importance measures (specifically SHAP and Knockoffs) that use unrestricted permutations often create impossible data, relying on predictions that are orders of magnitude away from being realistic. Their results are unstable, unreliable, and untrustworthy. Variable importance measures ν_j^{GCMR} and ν_j^{GKnock} that use restricted permutations do not have these problems. Even though they have completely different derivations and calculations, their variable importance values tend to be stable and agree with each other. We then examine the well-known Boston housing dataset, where we compare both traditional and novel importance measures after fitting three ML models with similar performance. Here again, we see variable importance measures that use extrapolation yield unrealistic predictions and unstable and non-intuitive results. We finally consider the real-world problem of determining which variables could be important for predicting the ethnicity of a person from only their first and last names. This task of *name-ethnicity classification* is a key type of analysis in assessing the fairness of policies when ethnicity has not been reported in a database, but where names are present. Specifically, we perform a variable importance analysis on a recently published dataset for name-based ethnicity identification of registered U.S. voters (Jain et al. 2022). This analysis shows the benefit of using stable variable importance metrics that agree with each other – since there is no ground truth, we would be less confident in a conclusion

drawn from variable importance measures that disagree. Our results using ν_j^{GCMR} and ν_j^{GKnock} are more stable and intuitive.

The remainder of the paper is organized as follows. Section 2 provides a review of the related literature. Section 3 presents the three strategies and concludes with a set of methodological results that link feature importance measures and total indices. Sections 4 and 5 are devoted to numerical experiments. Section 6 offers conclusions and further research perspectives.

sec:LitRev

2. Background Literature

This section concisely reviews feature importance measures and graphical visualization tools. Section 2.1 presents permutation importance measures and ALE plots. Section 2.2 presents relevant indices from sensitivity analysis.

ec_TIandALE

2.1. Breiman's Variable Importance Measures

In the reference framework of Hastie et al. (1994), analysts have a dataset of feature and target realizations and aim to determine the relationship

$$\mathbf{Y} = g(\mathbf{X}, \mathcal{E}), \tag{1}$$

eq:ygxE

where $\mathbf{X}$, $\mathbf{Y}$ and $\mathcal{E}$ are regarded as random variables on a probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$, with $X \in \mathcal{X}$, $\mathcal{X} \subseteq \mathbb{R}^d$, $g : \mathcal{X} \to \mathbb{R}^m$, $\mathcal{E} : \mathcal{X} \times \Omega \to \mathbb{R}$. Throughout the work, we suppose that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \cdots \times \mathcal{X}_n$, where $\mathcal{X}_j$ is the support of X_j, $j = 1, 2, \ldots, d$. We denote the vector of features excluding X_j and the corresponding support by $\mathbf{X}_{-j}$ and $\mathcal{X}_{-j} = \mathcal{X} \setminus \mathcal{X}_j$, respectively.

The map $g(\mathbf{X}, \mathcal{E})$ is assumed to be unknown and is approximated by a model $\widehat{g}(\mathbf{x}; \theta)$, $\widehat{g} : \mathcal{X} \times \Theta \to \mathbb{R}^m$, where $\theta \in \Theta$ is a set (vector) of (hyper) parameters or rules. The parameters determine $\widehat{g}$ via the solution of the (population version of the) optimization problem:

$$\min_{\theta \in \Theta} \mathbb{E}[\mathcal{L}(Y, \widehat{g}(\mathbf{X}; \theta))], \tag{2}$$

eq:MLOpt

where $\mathcal{L} : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty)$ is a loss function, with $\mathcal{L}(a, a') = 0$ if $a = a'$ for all $a, a' \in \mathbb{R}^m$. In practice, for a dataset $D = \{(\mathbf{x}^n, \mathbf{y}^n); n = 1, 2, \ldots, N\}$ containing N realizations of $(\mathbf{X}, \mathbf{Y})$, the sample version of Problem (2) requires us to minimize the empirical expected value of the loss function, namely to find $\theta^* = \arg\min\{\frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(\mathbf{y}^n, \widehat{g}(\mathbf{x}^n; \theta))\}$ with $(\mathbf{x}^n, \mathbf{y}^n) \in D$. Then, the model $\widehat{g}(\cdot; \theta^*)$ is used for further analysis. We refer to Gambella et al. (2020) for a review of the optimization problems that emerge for alternative ML models.

The (empirical) permutation feature importance of X_j under model $\widehat{g}(\cdot; \theta^*)$ is defined by

$$\widehat{\nu}_j = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}\left(\mathbf{y}^n, \widehat{g}(\mathbf{x}^n_{j,\text{perm}}; \theta^*)\right) - \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}\left(\mathbf{y}^n, \widehat{g}(\mathbf{x}^n; \theta^*)\right), \tag{3}$$

eq:nuhat1

where $\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(\mathbf{y}^n, \widehat{g}(\mathbf{x}^n;\theta^*))$ is the expected minimal loss for the sample and $\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(\mathbf{y}^n, \widehat{g}(\mathbf{x}^n_{j,\text{perm}};\theta^*))$ is the average loss when we randomly permute the entries of X_j. The method is model-agnostic. However, as there may be multiple optimal (or almost optimal) $\hat{g}$, given a fixed dataset, the same feature X_j may be assigned different values of $\widehat{\nu}_j$, depending on which model $\widehat{g}$ has been (arbitrarily) chosen. To handle this arbitrariness, Fisher et al. (2019) introduce the notion of model class reliance that takes into account the values of $\widehat{\nu}_j$ across the set of all predictive models that provide near-optimal accuracy. This set is called the Rashomon set (see Fisher et al. 2019, Dong and Rudin 2020, Semenova et al. 2022, for greater details on this notion). Because the selection of the best model is not a central part of this work, we shall consider representative ML models, noting that our technique could be applied across models in the Rashomon set (e.g., using Donnelly et al. 2023). We also use the simplified notation $\widehat{g}(\mathbf{X})$ instead of $\widehat{g}(\mathbf{X};\theta^*)$, when the context is clear.

We focus on the observation that Equation (3) is based on evaluation of the model on points $(\mathbf{x}^n_{j,\text{perm}};\theta^*)$ obtained after free (or unrestricted) permutations of the values in $\mathbf{x}_j$. Without control over the permutations, the new points may fall far from the data or even be impossible, leading the models to unreliable predictions. In the next sections, we review a few tools which are necessary to the strategies we are to propose.

sensitivity

2.2. Total Indices

Total indices are feature importance measures that originate in the simulation literature (Homma and Saltelli 1996, Saltelli and Tarantola 2002) and have attracted recent attention for explainability (Hart and Gremaud 2018, Bénard et al. 2022, Huang and Joseph 2024). One defines the total index of feature X_j as the expected portion of the variance of Y that remains after all features are fixed but X_j:

$$\tau_j = \mathbb{V}_{-j}[\mathbb{E}_j[Y|\mathbf{X}_{-j}]] = \mathbb{V}[Y] - \mathbb{E}_{-j}[\mathbb{V}_j[Y|\mathbf{X}_{-j}]], \tag{4}$$

eq:totalHar

We denote by T_j the corresponding normalized total sensitivity index, $T_j = \dfrac{\tau_j}{\sigma_y^2}$. Hart and Gremaud (2018) show that τ_j is the L^2 error that we incur for approximating g with a function that does not depend on X_j. Under feature independence, the total effects τ_j enjoy notable properties: they coincide with the portion of $\mathbb{V}[Y]$ accounted for by its individual contribution and by its interactions with the remaining features (Homma and Saltelli 1996); they are null if and only if Y is independent of X_j. If the independence assumption does not hold, total indices lose their interpretation as the overall fraction of the output variance associated with X_j, and they may be equal to zero even if $g(\mathbf{X})$ depends on X_j (Kucherenko et al. 2012). Notably, the works of Jansen (1999) and Kucherenko

et al. (2012), Mara and Tarantola (2012), Mara et al. (2015) show that τ_j can be written both under feature dependence and independence as

$$\tau_j = \frac{1}{2}\left(\mathbb{E}\left[\left(g(X'_j, \mathbf{X}_{-j}) - g\left(\mathbf{X}\right)\right)^2\right]\right), \tag{5}$$ eq:total_in

where $\mathbf{X}'$ is an independent replicate of $\mathbf{X}$. The empirical version of Equation (5) is know as Jansen's estimator. When features are correlated, Equation (5) is equivalent to:

$$\tau_j = \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}_j}\left(g\left(x'_j, \mathbf{x}_{-j}\right) - g\left(\mathbf{x}\right)\right)^2 dF_{X_j|\mathbf{X}_{-j}}(x'_j|\mathbf{x}_{-j})dF_{\mathbf{X}}(\mathbf{x}), \tag{6}$$ eq:tauicorr

where $F_{X_j|\mathbf{X}_{-j}}(x'_j|\mathbf{x}_{-j})$ is the conditional distribution of X_j given $\mathbf{X}_{-j}$, and $F_{\mathbf{X}}(\mathbf{x})$ the joint distribution of $\mathbf{X}$.

When X'_j is obtained from a free permutation, i.e., sampled from the marginal distribution of X_j, Equation (5) becomes

$$\tau'_j = \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}_j}\left(g\left(x'_j, \mathbf{x}_{-j}\right) - g\left(\mathbf{x}\right)\right)^2 dF_{X_j}(x'_j)dF_{\mathbf{X}}(\mathbf{x}). \tag{7}$$ eq:tauiprim

Notably, τ'_j in Equation (7) is the decorrelated importance index Ψ_{Dloco} introduced by Verdinelli and Wasserman (2023, Equation 8, p. 6) as a modification of the leave-out covariance index in Lei et al. (2018). Clearly, under feature independence, τ'_j is the total index of X_j for $\widehat{g}(\mathbf{X})$. It can be proven that $\tau'_j = 0$ if and only if $\widehat{g}(\mathbf{X})$ does not depend on X_j (Verdinelli and Wasserman 2023).

Lastly, for Random Forests, Bénard et al. (2022) address the computation of Breiman's permutation feature importance measures and relates it to the calculation of total indices. The mean decrease in impurity (MDI) is a tailored feature importance measure for Random Forests that quantifies the decrease in variance associated with each split in feature X_j. Agarwal et al. (2023) recently modifies the MDI to yield an indicator close to total indices (see Equation 4 on page 7 of Agarwal et al. 2023). A question that remains to be explored is then the connection between Breiman's feature importance measures and total indices when permutations are restricted.

2.3. ALE Plots

Apley and Zhu (2020) propose accumulated local effects (ALE) plots. Assuming that $\widehat{g}(\cdot)$ is differentiable, the ALE function of X_j is defined as

$$ALE_j(x_j) = \int_{x_{j,\min}}^{x_j} \mathbb{E}_{\mathbf{X}_{-j}|X_j}[\widehat{g}'_j(\mathbf{X})|X_j = t_j]dt_j, \tag{8}$$ MR=ale

where $\widehat{g}'_j(\mathbf{X})$ is the estimated partial derivative of $\widehat{g}$ with respect to X_j and $x_{j,\min}$ is a chosen value close to the lower bound of the support of the distribution of X_j. (Apley and Zhu 2020 also introduce a centered version of ALE plots by subtracting a constant in Equation (8)). ALE functions have interesting properties, as highlighted by Apley and Zhu (2020). For instance, if the ML model is

additive in X_j, that is $\widehat{g}(\mathbf{x}) = \widehat{g}_j(x_j) + \widehat{g}_{-j}(\mathbf{x}_{-j})$, then $ALE_j(x_j)$ differs from $\widehat{g}_j(x_j)$ only by an additive constant. A similar behavior is recorded if the model is multiplicatively separable in X_j and X_j is independent of the remaining features.

We note a *null conditional expectation effect* associated with ALE plots. Suppose that $\widehat{g}(\mathbf{x})$ is differentiable and that it can be written as the product of a function of x_j and a function of all other features, $\widehat{g}(\mathbf{x}) = h_j(x_j)h_{-j}(\mathbf{x}_{-j})$. Also, suppose that X_j is independent of $\mathbf{X}_{-j}$. Then, if $\mathbb{E}[h_{-j}(\mathbf{X}_{-j})] = 0$, we have $ALE(x_j) = 0$. In fact,

$$ALE(x_j) = \int_{x_j,\min}^{x_j} h_j'(t)\mathbb{E}[h_{-j}(\mathbf{X}_{-j})|X_j = t]dt = \int_{x_j,\min}^{x_j} h_j'(t)\mathbb{E}[h_{-j}(\mathbf{X}_{-j})]dt = \int_{x_j,\min}^{x_j} h_j'(t_j)dt_j \cdot \mathbb{E}[h_{-j}(\mathbf{X}_{-j})] = [h_j(x_j) - h_j(x_{j,\min})]\mathbb{E}[h_{-j}(\mathbf{X}_{-j})] = 0. \quad (9)$$

and the ALE plot of X_j is flat, no matter how Y depends on X_j. We show that this false negative can be avoided by computing feature importance measures that involve squared ALE effects (see Appendix B).

Sec:Knock

2.4. Model-X Knockoffs

Model-X Knockoffs are a feature selection method introduced in the works of Barber and Candés (2015) and Candès et al. (2018). The intuition behind this method is that of generating artificial variables, the knockoffs, for controlling the false discovery rate. The knockoffs are replicates of the original variables and they are used to create test statistics to assess the relevance of the original variables. The Knockoffs procedure has been intensively studied and applied (Sesia et al. 2019, Romano et al. 2020, Ren and Candès 2023). Recently, Barber et al. (2023) introduce a methodology to derive the power of Knockoffs test statistics, combining asymptotic estimation theory with advances in the theory of approximate message-passing. Relevant to our work is that, under appropriate conditions, the Knockoffs variables follow the same distribution as the original variables. While this intuition is used in one of the experiments of Hooker et al. (2021), we aim to explore it further. Our first experiment shows a failure of this method, where new data falls outside the original domain for the case of uniformly distributed features, correlated via a Gaussian Copula (the theory guarantees the same distribution when the data are multivariate normal). We then explore the option of generating Knockoffs after a Gaussian transformation, where theoretical guarantees hold, and study the efficacy of this new proposal.

Sec:SHAPs

2.5. Shapley Additive Explanations

Shapley values are popular explanations in computer science and artificial intelligence (Sundararajan and Najmi 2020, Chen et al. 2023). Among the alternative implementations, Lundberg and Lee

(2017) provide a unified perspective and a computational framework aimed at efficiently estimating SHapley Additive exPlanations (SHAPs). As underlined by Mase et al. (2022), calculation of the SHAPs is exposed to the same extrapolation issues as partial dependence functions. However, to our knowledge, a systematic investigation of the intensity with which this effect occurs has not previously been performed. We address this point in our experiments.

sec:Impute

3. Three New Strategies

This section presents the novel permutation restriction strategies discussed in this work. Section 3.1 considers a permutation importance measure from a design that obtains new values of X_j through a conditional model reliance calculation performed after a Gaussian transformation of the data. Section 3.2 performs a Knockoffs procedure after a Gaussian transformation. Section 3.3 illustrates feature importance measures based on the ALE design.

sec:GCMR

3.1. Conditional Model Reliance after Gaussian Transformation

Our first approach extends the residual imputation strategy and specifically, the notion of conditional model reliance of Fisher et al. (2019). The conditional model reliance calculation from Fisher et al. (2019) uses the following steps:

1. First, the feature of interest (X_j) is regressed on the remaining features ($\mathbf{X}_{-j}$), via $g_{impute}(\mathbf{X}_{-j}) = \mathbb{E}[X_j|\mathbf{X}_{-j}]$, where $g_{impute}(\mathbf{X}_{-j})$ is a generic non-parametric regression model.
2. The residuals are then defined as usual via $\overline{X}_j := X_j - \mathbb{E}[X_j|\mathbf{X}_{-j}]$.
3. A random permutation of the residuals is performed to obtain $\overline{X}_j^{\pi}$ and a new point X_j' is defined as as $X_j' := \mathbb{E}[X_j|\mathbf{X}_{-j}] + \overline{X}_j^{\pi}$.
4. The original points $\mathbf{X} = (X_j, \mathbf{X}_{-j})$ are mapped into the new points $\mathbf{X}' = (X_j', \mathbf{X}_{-j})$. Thus, the intuition is to randomly permute the residuals instead of the original points.
5. The total index, as the difference between (or ratio of) the imputed data's loss and the original data's loss, as in (5), is used to measure importance of the unique information carried by variable j (information that is not included in other variables).

The problem with this approach is that even with permuting only residuals, it constructs points that are out of distribution. For instance, consider a nonnegative variable. A residual of -1 given to a point with value 0.1 would yield value 0.1-1 = -.9, which is an impossible value.

Our fix for this problem is to perform residual imputation *after transforming the data into normal scores*. Our approach is as follows:

i) Let each Z_j be a normally distributed random variable and let each h_j be a real one-to-one function with $Z_j = h_j(X_j)$ transforming each marginal distribution into a 1D Gaussian by applying the normal quantile function to the (empirical) marginal cumulative distribution function. We can now work with the newly transformed dataset (Z, Y).

ii) Implement the residual imputation procedure on the normal scores Z (using labels Y) and permute, following Steps 1, 2 and 3 above, to obtain $Z'_j := \mathbb{E}[Z_j|\mathbf{Z}_{-j}] + \overline{Z}^{\pi}_j$. Importantly, because this step edits only z-scores, it is identical to adjusting the quantile of X_j within its marginal distribution, which does not go out of range. For instance, even if Z'_j is set to the lowest quantile, it would still correspond to the lowest values of X_j that are realized in the dataset.

iii) Apply the inverse transformation mapping Z'_j to X'_j (i.e., $X'_j = h_j^{-1}(Z'_j)$).

iv) The change in the loss between using X'_j and X^j is our measure of variable importance.

This approach is summarized in Algorithm 1.

Algorithm 1 Steps of the Gaussian-transformation and conditional model reliance strategy (GCMR).

algoGauss

Input: feature dataset $\widehat{X}$.

Output: A permuted dataset $\widehat{X}'$ that avoids unrestricted permutations, and a measure of the importance of variable j.

for $j = 1, 2, \ldots, d$ **do**

Transform $\widehat{X}_j$ into a standard uniform random variable U_j using the (empirical) marginal cdf

Map U_j into a standard normal random variable Z_j, using the Gaussian quantile function.

end for

for $j = 1, 2, \ldots, d$ **do**

Calculate the non-parametric regression curve $\mathbb{E}[Z_j|\mathbf{Z}_{-j}]$ using any regression method.

Calculate the residual $\overline{Z}_j = Z_j - \mathbb{E}[Z_j|\mathbf{Z}_{-j}]$;

Permute $\overline{Z}_j$ to obtain $\overline{Z}^{\pi}_j$;

Add back $\overline{Z}^{\pi}_j$ to $\mathbb{E}[Z_j|\mathbf{Z}_{-j}]$ to obtain $Z'_j = \mathbb{E}[Z_j|\mathbf{Z}_{-j}] + \overline{Z}^{\pi}_j$;

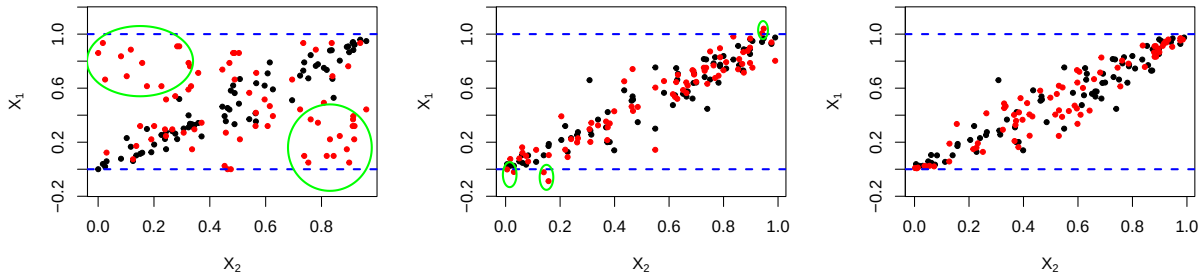
Map Z'_j into X'_j.

end for

Calculate the difference (or ratio) between the loss using X'_j and the loss using X_j. This is our measure of importance of the unique information within X_j.

e:testcase3

EXAMPLE 1. We consider two uniformly distributed random variables (X_1 and X_2) on the interval $[0,1]^2$. We assume that X_1 and X_2 are strongly correlated ($\rho = 0.95$). We model the correlation through a Gaussian copula. In Figure 1, we report original data (black) and new data (red) obtained, respectively, after an unrestricted permutation of X_1 (panel (a)), a residual imputation (panel (b)) and a residual imputation with Gaussian transformation (panel (c)). In Figure 1a, we apply an unrestricted permutation: Several new data fall far from the original ones (we evidence them by a green oval). This implies that an ML model will be forced to extrapolate. In Figure 1b, we apply a



_permutation

plot_gimpute_NO_Norm

plot_gimpute_Norm

Figure 1 **Original (black) and permuted data (red): (a) with unrestricted permutations; (b) with residual imputation without Gaussian transformation; (c) with residual imputation with Gaussian transformation.**

g_impute

permutation restricted through residual imputation: The new points (red) lie closer to the original points (black). However, some new points (red) fall outside the support of X_1. In Figure 1c, we apply a restricted permutation through residual imputation mediated by a Gaussian transformation: New points (red) not only lie within the cloud of the original points but also never leave the support of X_1 and X_2. Note that the data dependence structure is preserved as the new (red) points have the same distribution as the original data. It is important to note here that Figure 1(a) considers the variable's overall importance, whereas Figures 1(b) and (c) consider only the information about the variable that is not contained within other variables.

The previous example involved two continuous features. In the next example, we consider a case in which one of the features is discrete.

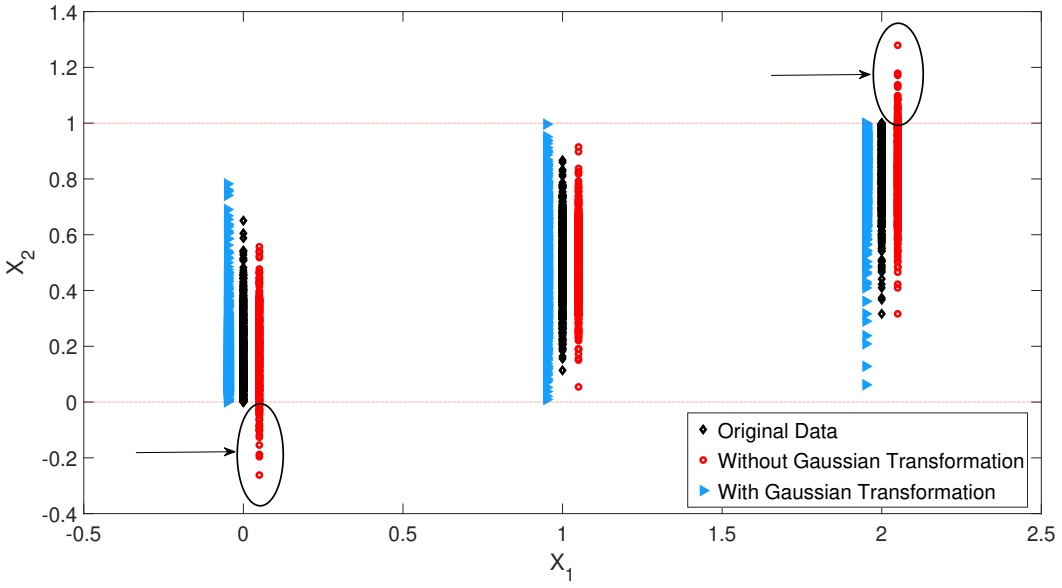


Figure 2 **Vertical axis: X_2. Horizontal axis X_1. Original data ($\diamond$ black) and permuted data ($\circ$ red, right of the original data) permuted data after Gaussian Transformation ($\triangleright$ blue, left of the original data).**

mpute_discr

EXAMPLE 2. Consider X_1 and X_2 strongly correlated ($\rho = 0.95$), with X_1 discrete with support $\{0, 1, 2\}$ and X_2 uniformly distributed on $[0, 1]$. Because X_1 is discrete, the scatterplot assumes the shape of three vertical bins (Figure 2), centered, respectively, at $X_1 = 0, X_1 = 1$ and $X_1 = 2$. In each bin, the central column represents the original data. The right column (red) shows the new data obtained after a restricted permutation without a Gaussian transformation. Several points follow outside the support of X_2. Performing the residual imputation in the Gaussian space yields new points (blue) that lie close to the cloud of the original data and inside the support of X_2.

It is easy to recognize that, from a theoretical viewpoint, it might not be possible to guarantee that the newly generated points follow the same distribution as the original ones in all cases. However, we provide some theoretical results in the proposition below.

GaussCopula

PROPOSITION 1. *If the joint distribution of the features* $\mathbf{X}$ *can be modeled by a Gaussian copula, then* $\mathbf{X}'$ *returned by Algorithm 1 is an independent copy of* $\mathbf{X}$.

In other words, Proposition 1 guarantees that the new permuted points follow the same distribution as the original points when the feature-generating process can be modeled by a Gaussian copula. While this assumption may not hold in all practical situations, from a methodological viewpoint it still encompasses a broad family of statistical distributions.

The new points can then be used to calculate feature importance measures based on permutations, without going outside of the variables' marginal distributions. Similarly to Fisher et al. (2019), calculations from these new points would yield conclusions only about the importance of the unique information about the feature, i.e., the information this feature contains that cannot be gleaned from other variables. We denote the indices by ν_j^{GCMR}, to distinguish them from the unrestricted ν_j.

sec:GKnock

3.2. The Gaussian Knockoffs Strategy

As an alternative to GCMR, we explore the use of the Knockoffs strategy with our Gaussian transformation. In this case, the procedure is as follows. We map the original data into the corresponding Gaussian transformed random variables, as per the first for loop in Algorithm 1. The second for loop is replaced by generating new values of Z variables from the Knockoffs procedure. The knockoffs Z' are then mapped back to the original space, to form the permuted sample X'. We call this strategy GKnock. Similarly to the GCMR strategy, the following proposition holds.

prop:GKnock

PROPOSITION 2. *If the joint distribution of the features* $\mathbf{X}$ *can be modeled by a Gaussian copula, then* $\mathbf{X}'$ *returned by the Gknock procedure is a Knockoff copy of* $\mathbf{X}$.

x:GknockOut

EXAMPLE 3 (EXAMPLE 1 CONTINUED). Applying GKnock to Example 1 yields results similar to those in Figure 1. The central panel in Figure 3 shows that without the Gaussian transformation, the Knockoffs sample falls outside the [0,1] boundary of the marginal distributions of X_1 and

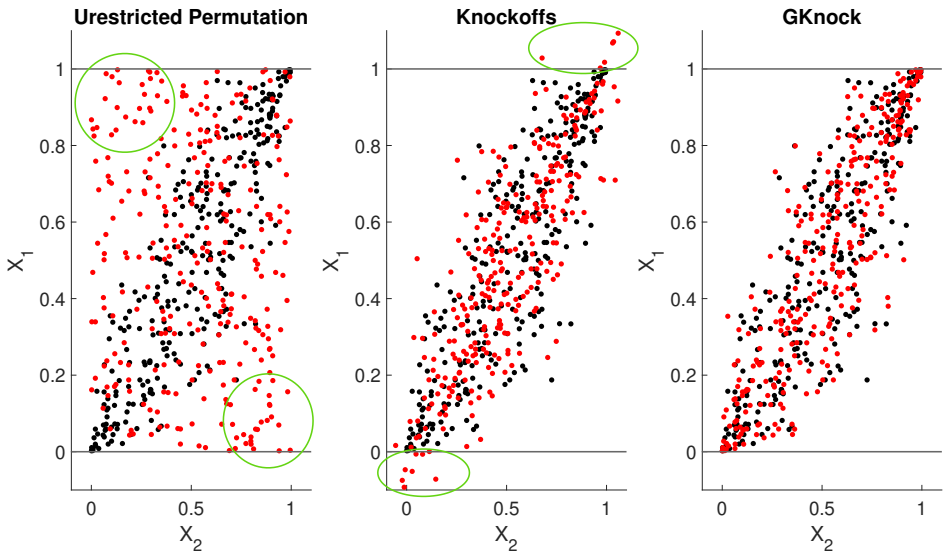


Figure 3 **Original and new points generated with unrestricted (left panel), Knockoffs-restricted (central panel and GKnock-restricted permutations.**

e5knockoffs

X_2. The third panel, instead, shows that if the Knockoffs operation is performed after a Gaussian transformation of the data, the new points lie back within the original support. Thus, applying a Gaussian transformation before implementing the knockoffs improves the data generation also in a case of non-normal variables.

Similarly to the case of GCMR, the new points can be used to compute permutation-based importance measures with data that follow the variable marginal distributions. We use the symbol ν_j^{GKnock} to denote Breiman's importance measures when the new points are Knockoffs of the original variables.

sec:tauALE

3.3. The ALE Plots Feature Importance Measures

The intuition of this strategy is to associate feature importance measures with the design of ALE plots. As in Apley and Zhu (2020), assume that $\mathcal{X}_j = [x_{j,\min}, x_{j,\max}]$ is an interval on the real line (or the union of a possibly disjoint set of intervals). We can partition $\mathcal{X}_j$ into K subintervals $\mathcal{X}_j^k = [z_j^{k-1}, z_j^k)$, with $z_j^0 = x_{j,\min}$ and $z_j^K = x_{j,\max}$, such that $\bigcup_{k=1}^{K} \mathcal{X}_j^k = \mathcal{X}_j$ and $\mathcal{X}_j^k \bigcap \mathcal{X}_j^m = \emptyset$, $k, m = 1, 2, \ldots, K$, $k \neq m$ (Figure 4). Then, as in Apley and Zhu (2020), we denote with $n_j^K(k)$ the number of realizations of X_j that belong to $\mathcal{X}_j^k$, and with $k_j^K(x)$ the index of the subinterval that contains x. An estimate of $ALE_j(x_j)$ is given by

$$\widehat{ALE}_j(x_j) = \sum_{k=1}^{k_j^K(x_j)} \frac{1}{n_j^K(k)} \sum_{n:\mathbf{x}_j^n \in \mathcal{X}_j^k} \left(\widehat{g}\left(z_j^k, \mathbf{x}_{-j}^n\right) - \widehat{g}\left(z_j^{k-1}, \mathbf{x}_{-j}^n\right) \right). \tag{10}$$

MR=ale3

The calculation of $\widehat{ALE}_j(x_j)$ requires averaging differences in predictions over the conditional distribution of the feature of interest and also for points close to a given realization $\mathbf{x}^n$. Note that the

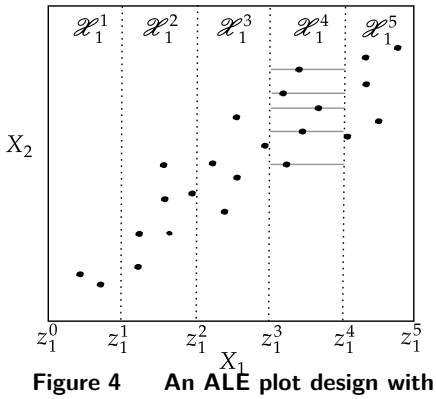


fig:phi_ALE

Figure 4 An ALE plot design with $K = 5$.

L^1 distance between $\left(z_j^k, \mathbf{x}_{-j}^n\right)$ and $\left(z_j^{k-1}, \mathbf{x}_{-j}^n\right)$ is $|z_j^k - z_j^{k-1}|$. Supposing equally spaced partitions, we have $|z_j^k - z_j^{k-1}| = 1/K$, so that the difference is bounded. Thus, the new evaluation points are forced to lie close to the original points and unrestricted permutations are avoided.

We rewrite the ALE effects as random variables

$$\Phi_j^{\mathrm{ALE}}(\mathbf{X}_{-j}; K) = \widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^{k-1}, \mathbf{X}_{-j}), \tag{11}$$

eq:phiALE

where K indicates that we are constraining the variations of X_j to adjacent points on the grid. Inserting these effects into the Jansen estimator in Equation (5), we find

$$\tau_j^{\mathrm{ALE}}(K) = \frac{1}{2}\mathbb{E}\left[\left(\Phi_j^{\mathrm{ALE}}(\mathbf{X}_{-j}; K)\right)^2\right]. \tag{12}$$

eq:tauALEK

By construction of the ALE plots and by the total probability theorem, Equation (12) is equivalent to:

$$\tau_j^{\mathrm{ALE}}(K) = \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}[\left(\widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^{k-1}, \mathbf{X}_{-j})\right)^2 | X_j \in \mathcal{X}_j^k] \cdot \mathbb{P}(X_j \in \mathcal{X}_j^k), \tag{13}$$

eq:tauALEco

where $\mathbb{P}(X_j \in \mathcal{X}_j^k)$ is the probability that the j^{th} component of $\mathbf{X}$, X_j, belongs to the sub-interval $\mathcal{X}_j^k$. Equation 13 makes explicit the dependence of $\tau_j^{\mathrm{ALE}}(K)$ on the grid. Equation (13) shows that the numerical value of the total ALE indices $\tau_j^{\mathrm{ALE}}(K)$ are, in general, different from the classical ones in Equation (5). They are built in the same spirit of total indices as expectations of first order finite differences, however they differ with respect to the points where the model is evaluated.

To explain the differences between $\tau_j^{\mathrm{ALE}}(K)$ and τ_j, consider that the ALE main effects $\Phi_j^{\mathrm{ALE}}(X_j', \mathbf{X}_{-j}; K)$ are calculated for values of X_j fixed at z_k and z_{k-1} for all realizations of $\mathbf{X}_{-j}$. Conversely, the classical effects $\Phi_j'(X_j', \mathbf{X})$ are calculated with the new value X_j' sampled independently from $\mathbf{X}_{-j}$. Figure 5 offers a visualization of the points implied by the two designs when X_j' in the total index is obtained from an unrestricted permutation.

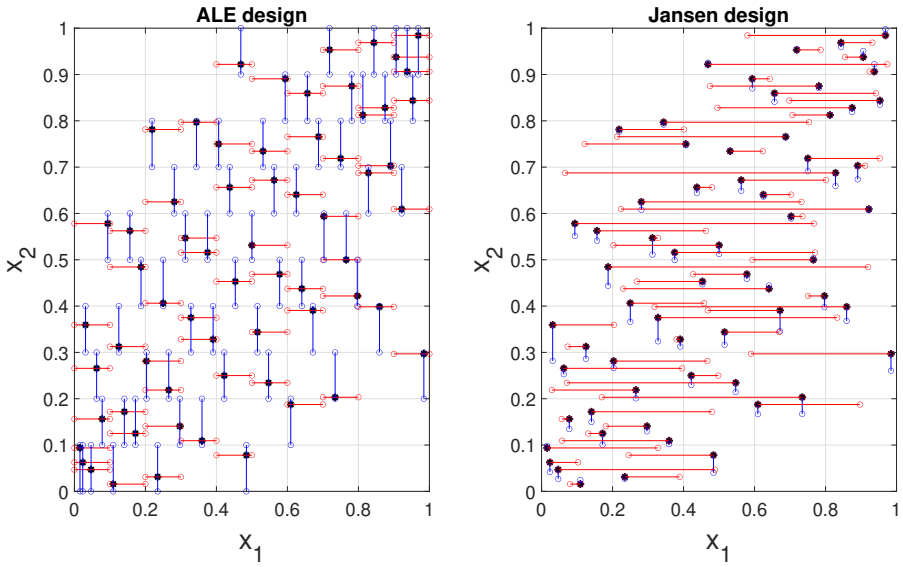


Figure 5 **ALE design (left graph) and Jansen design (right graph) for a correlated case. Legend: • original points; red (∘), new points after varying X_1; blue (∘), new points after varying X_2.**

g:AleDesign

The left graph of Figure 5 displays the points visited by an ALE algorithm. The new points $(z_j^k, \mathbf{x}_{-j}^k)$ (red and blue dots) are always close to the original data $\mathbf{x}^k$ (black dots) and this reduces extrapolation issues. The right panel in Figure 5 shows points visited by an algorithm that follows Equation (5), in which X_j' is sampled independently of $\mathbf{X}_{-j}$ (this is called the Jansen estimator). New points $(x_j^{'}, \mathbf{x}_{-j}^k)$ now can fall far away from the original points $\mathbf{x}^k$, with potential extrapolation problems. The next example illustrates Equations (12) and (13).

xample:Ishi

EXAMPLE 4. Consider the input-output mapping proposed in Ishigami and Homma (1990),

$$g(\mathbf{X}, \mathcal{E}) = \sin(X_1) + 7\sin^2(X_2) + 0.1X_3^4 \sin(X_1), \tag{14}$$

eq:Ishigami

with $\mathcal{E} = 0$ and the features uniformly and independently distributed on the intervals $[-\pi, +\pi]$. Figure 6 provides a visual representation of the input-output mapping, evidencing its non-monotonic and non-convex behavior.

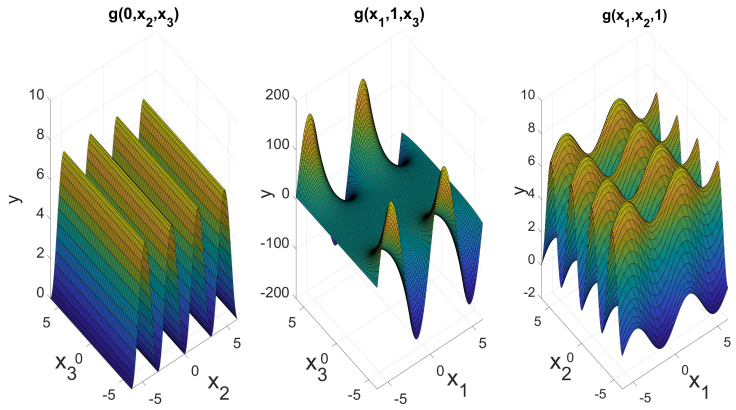


Figure 6 **Three bi-dimensional projections of the 3-variate Ishigami model. Left plot:** $g(0, x_2, x_3)$**. Middle plot:** $g(x_1, 1, x_3)$**. Left plot** $g(x_1, x_2, 1)$**.**

higamiplots

We can obtain the expression of the τ_j^{ALE} indices analytically as a function of K combining Equations (14) and (13). For X_1, setting $z_1^0 = -\pi$, $z_1^K = \pi$, and $z_1^k - z_j^{k-1} = \dfrac{2\pi}{K}$, Equation (12) becomes

$$\tau_j^{\text{ALE}}(K) = \frac{1}{2K(2\pi)^2} \sum_{k=1}^{K} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} (g(z_1^k, x_2, x_3) - g(z_1^{k-1}, x_2, x_3))^2 dx_2 dx_3. \tag{15}$$

eq:tauALE1

Substituting the expression of the Ishighami function into Equation (15) we obtain

$$\tau_1^{\text{ALE}}(K) = \frac{1}{2K(2\pi)} \sum_{k=1}^{K} \int_{-\pi}^{\pi} \left(\sin\left(\frac{2k\pi}{K}\right) - \sin\left(\frac{2(k-1)\pi}{K}\right) + 0.1 X_3^4 \left(\sin\left(\frac{2k\pi}{K}\right) - \sin\left(\frac{2(k-1)\pi}{K}\right) \right) \right)^2 dX_3 = \frac{97.00}{4K\pi} \sum_{k=1}^{K} \left(\sin\left(\frac{2k\pi}{K}\right) - \sin\left(\frac{2(k-1)\pi}{K}\right) \right)^2. \tag{16}$$

Similar expressions are obtained for $\tau_2^{\text{ALE}}(K)$ and $\tau_3^{\text{ALE}}(K)$.

	$\tau_1^{\text{ALE}}(K)$	$\tau_2^{\text{ALE}}(K)$	$\tau_3^{\text{ALE}}(K)$	$\kappa_1^{\text{ALE}}(K)$	$\kappa_2^{\text{ALE}}(K)$	$\kappa^{\text{ALE}}(K)_3$
$K = 10$	1.47	4.32	2.08	1.77	5.10	2.50
$K = 50$	0.061	0.192	0.087	1.83	5.79	2.6
$K = 100$	0.015	0.048	0.022	1.83	5.81	2.61
$K = 200$	0.00	0.012	0.00	1.83	5.82	2.61

Table 1 **Values of** $\tau^{\text{ALE}}(K)$ **and** $\kappa^{\text{ALE}}(K)$ **for the Ishigami function with the partition cardinality** K **varying from** $K = 10$ **to** $K = 200$**.**

b:tkALEIshi

The values of $\tau_j^{\text{ALE}}(K)$ can then be easily computed. Table 1 reports the results for partition sizes $K = 10, 50, 100, 200$. The results show that $\tau_j^{\text{ALE}}(K)$ ranks X_2 as most important for all the selected values of K. It is interesting to compare the values in Table 1 with the values of the classical total indices τ_j in Equation (5), which are analytically known for this model, with $\tau_1 = 7.7169$, $\tau_2 = 6.1248$,

and $\tau_3 = 3.3725$. The comparison shows that the indices do not agree and in contrast with $\tau_j^{\mathrm{ALE}}(K)$, the classical τ_j ranks X_1 as the most important feature. Table 1 also shows that the values of $\tau_j^{\mathrm{ALE}}(K)$ are highly sensitive to the cardinality of the grid, and that they tend to zero as K increases.

For this example, the inputs are not correlated and we know the true input-output mapping. Thus, no extrapolation issues are present. The difference underlines that $\tau_j^{\mathrm{ALE}}(K)$ and total indices are different importance measures. Table also shows that the value of $\tau_j^{\mathrm{ALE}}(K)$ strongly depends on the choice of K. Moreover, as K grows, $\tau_j^{\mathrm{ALE}}(K)$ tends to zero. The reason is that as K increases the size of the partition decreases and the difference $z_j^k - z_j^{k-1}$ becomes infinitesimal. If the input-output mapping is smooth, we obtain small values of the ALE main effects in Equation (11), because $\widehat{g}\left(z_j^k, \mathbf{x}_{-j}^n\right) - \widehat{g}\left(z_j^{k-1}, \mathbf{x}_{-j}^n\right) \approx 0$ if $z_j^k \approx z_j^{k-1}$. Correspondingly, their square will also be numerically small. To illustrate, for the Ishigami function, at $K = 200$, the squared differences $\mathbb{E}\left[(\widehat{g}\left(z_1^k, \mathbf{X}_{-1}\right) - \widehat{g}\left(z_1^{k-1}, \mathbf{X}_{-1}\right))^2\right]$ can be calculated analytically and range from a minimum of $3.76 \cdot 10^{-6}$, close to numerical noise, to a maximum of $1.5 \cdot 10^{-2}$, yielding $\widehat{\tau}_1^{\mathrm{ALE}} = 4.0 \cdot 10^{-3} \approx 0$. To remedy this situation, assume that $\hat{g}(\mathbf{X})$ is differentiable and define

$$\kappa_j^{\mathrm{ALE}}(K) = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left(\frac{\widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^{k-1}; \mathbf{X}_{-j})}{z_j^k - z_j^{k-1}}\right)^2\right] \frac{\sigma_j^2}{\sigma_y^2}. \tag{17}$$

eq:kappa

The index $\kappa_j^{\mathrm{ALE}}(K)$ is a normalized expectation of Newton ratios computed at randomized locations in the feature space in the spirit of derivative-based sensitivity measures of Sobol and Kucherenko (2009) (see also Kucherenko and Iooss 2017 and Song et al. 2019). The normalization in Equation (17) makes $\kappa_j^{\mathrm{ALE}}(K)$ the ALE-plot-based extension of early sensitivity indices defined by Bier (1983) and Helton (1993). Observe that if the model is differentiable, Newton ratios remain finite as K increases.

le:Ishicont

Example 5 (Example 4 continued). For the same setting as in Example 4, we obtain the values of $\widehat{\kappa}_j^{\mathrm{ALE}}(K)$ in Table 1 which are different from zero also as the grid is progressively refined.

We then study whether a zero value of τ_j^{ALE} and κ_j^{ALE} implies independence between Y and X_j.

LEzeroindep

Proposition 3. *If $\widehat{g}(\cdot)$ does not depend on X_j then $\tau_j^{ALE}(K) = 0$ and $\kappa_j^{ALE}(K) = 0$. Conversely, if $\tau_j^{ALE}(K) = 0$ or $\kappa_j^{ALE}(K) = 0$ for every choice of the partition of $\mathcal{X}_j$ then $\widehat{g}(\cdot)$ does not depend on X_j.*

For a fixed a partition of $\mathcal{X}_j$, a null value of $\tau_j^{\mathrm{ALE}}(K)$ or of $\kappa_j^{\mathrm{ALE}}(K)$ does not necessarily reassure the analyst that $\widehat{g}$ is independent of X_j. For instance, consider that $\widehat{g}(z_j^k, \mathbf{X}_{-j}^k)$ is periodic of period $\frac{1}{K}$. If we select $z_j^k - z_j^{k-1} = \frac{1}{K}$ then $\tau_j^{\mathrm{ALE}}(K) = 0$ and $\kappa_j^{\mathrm{ALE}}(K) = 0$. Note that the corresponding ALE plot would be a flat line. However, an easy fix is represented by repeating the calculation with alternative choices of the grid $z_j^1, \ldots, z_j^K$. If there is a dependence on X_j, varying the grid will allow us to detect it.

The indices introduced in this section are close to two well-known global sensitivity measures: $\tau_j^{\text{ALE}}(K)$ are in the spirit of total indices, and $\kappa_j^{\text{ALE}}(K)$ in the spirit of derivative-based sensitivity measures. Sobol and Kucherenko (2009) show that the two indices are related, but not equivalent. In particular, if the derivative-based index of X_j is zero then the corresponding total index must be zero. The converse might not be true. Then, for our indices, $\kappa_j^{\text{ALE}}(K) = 0$ implies $\tau_j^{\text{ALE}}(K) = 0$. The fact that the converse does not hold is shown by the values in Table 1.

While we generally do not expect an identity between ALE-based and total indices, we have the following for linear models.

cor:KALETau

COROLLARY 1. *Let $y : \mathcal{X} \to \mathbb{R}$, with $\mathcal{X} \subseteq \mathbb{R}^d$ be a linear mapping $y = \mathbf{a}\mathbf{x}^T$, where $\mathbf{a} \in \mathbb{R}^d$ is a vector of coefficients. Then, we have:*

1. *For any feature distribution and any grid cardinality K,*

$$\kappa_j^{ALE}(K) = \frac{a_j^2 \sigma_j^2}{\sigma_y^2}. \tag{18}$$

eq:KaleLine

2. *If the features are independent, then $\kappa_j^{ALE}(K) = \dfrac{\tau_j}{\sigma_y^2}$.*

Corollary 1 states that $\kappa_j^{\text{ALE}}(K)$ and total indices τ_j essentially coincide for linear models with independent inputs.

Until now, we regarded X_j as a continuous random variable. If X_j is discrete, its support $\mathcal{X}_j$ is the countable number of realizations $\mathcal{X}_j = \{x_j^1, x_j^2, \ldots, x_j^K\}$. Then, the above equations need to be interpreted in a finite difference sense. The grid is naturally determined by the support of X_j, with the obvious assignment $z_j^k = x_j^k$ and $z_j^{k+1} = x_j^{k+1}$. For a binary random variable, then we cannot but set $z_j^1 = x_j^1$ and $z_j^2 = x_j^2$. Notably, we cannot increase the partition cardinality, and we are not reassured that new data will fall close to the original points (see Appendix D for experimental results).

rmandtotals

3.4. Relating Permutation-based Importance Measures and Total Indices

Let us consider the population definition of the variable importance measures in Equation (3):

$$\nu_j = \mathbb{E}\left[\mathcal{L}\left(Y, \widehat{g}(X_j', \mathbf{X}_{-j}; \theta^*)\right)\right] - \mathbb{E}\left[\mathcal{L}\left(Y, \widehat{g}(\mathbf{X}; \theta^*)\right)\right]. \tag{19}$$

eq:nujLossp

prop:nutau

PROPOSITION 4. *If $\mathcal{L}(\cdot, \cdot)$ is the quadratic loss and if the model is a perfect predictor, then*

1)

$$\nu_j = \mathbb{E}\left[\left(\widehat{g}(\mathbf{X}; \theta^*) - \widehat{g}(X_j', \mathbf{X}_{-j}; \theta^*)\right)^2\right]. \tag{20}$$

eq:nusquare

2) If X_j' is sampled independently of $\mathbf{X}_{-j}$, then

$$\nu_j = 2\tau_j', \tag{21}$$

eq:nu2tauLo

where τ_j' is the Ψ_{DLoco} total index in Equation (7).
3) If $X_j^{'}$ is sampled conditionally on $\mathbf{X}_{-j}$, then

$$\nu_j = 2\tau_j, \tag{22}$$ eq:nu2tau

where τ_j is the classical total index in Equation (6).

Proposition 4 sheds light on the meaning of Breiman's variable importance measures with and without permutation restrictions. Without permutation restrictions, Breiman's variable importance measures are equal to twice the Ψ_{DLoco} importance of Verdinelli and Wassermann. With permutation restrictions, they are equal to twice the classical (Sobol') total indices. For the empirical estimates $\widehat{\nu}_j$ in Equation (3), we have the following.

r:estimates COROLLARY 2. *Under the assumptions of Proposition 4,*
1) $\widehat{\nu}_j$ in Equation (3) *is an asymptotically unbiased estimator of ν_j in Equation* (20).
2) If the new points $X_j^{'}$ are obtained from a free permutation of X_j, then

$$\widehat{\nu}_j = 2\widehat{\tau}_j', \tag{23}$$

where τ_j' is an estimate of the Ψ_{DLoco} total index in Equation (7).
3) If, in addition, under the assumptions, respectively of Propositions 1 or 2, if the new points $X_j^{'}$ are sampled through Algorithms 1 *or* 2, *then*

$$\widehat{\nu}_j^{GCMR} = \widehat{\nu}_j^{GKnock} = 2\widehat{\tau}_j, \tag{24}$$

where τ_j is the classical total index in Equation (6).

mericalExps 4. Numerical Experiments with Analytical Test Cases

We performed a series of numerical experiments to test the behavior of all indices discussed in the paper. A first set of experiments were performed on the Ishigami function. Results show that we can avoid the null conditional expectation effect associated with ALE plots by computing the τ_j^{ALE} and κ_j^{ALE} indices. Moreover, all estimates of ALE-indices and permutation-based indices tend to the ground truth values, in accordance with Equation (13) and Proposition 4. We refer to Appendix B for full details. We focus the remainder of this section on the important test case introduced by Hooker et al. (2021) to study the impact of extrapolation errors on Breiman's feature importance measures. Our goal is to test the efficacy of the proposed permutation restriction strategies and the behavior of ALE-based indices.

The test case. Hooker et al. (2021) select, as input-output mapping, the linear model

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 + 0 \cdot X_6 + 0.5 \cdot X_7 + 0.8 \cdot X_8 + 1.2 \cdot X_9 + 1.5 \cdot X_{10} + \epsilon, \tag{25}$$

Hooker:mode

with $X_j \sim \mathcal{U}[0,1]$ and $\epsilon \sim \mathcal{N}(0, 0.1^2)$. With this assignment, the feature ranking is determined by the magnitudes of their coefficients. The most important feature is then X_{10}, followed by X_9, followed by the first five features, then by X_8, X_7 and by X_6, which is inactive. To test the effect of correlations, Hooker et al. (2021) make X_1 and X_2 statistically dependent, generating their values via a Gaussian copula with correlation coefficient ρ_{X_1,X_2}. They generate synthetic input-output data via Monte Carlo simulation. They then fit a linear model (LM), a Random Forest (RF), an artificial neural network (NN) and compute ν_j for alternative values of ρ_{X_1,X_2}. Hooker et al. (2021) then perform a series of experiments and calculate Breiman's variable importance measures for increasing values of the correlation between X_1 and X_2. Their results show that the variable importance of X_1 and X_2 increases as the correlation increases.

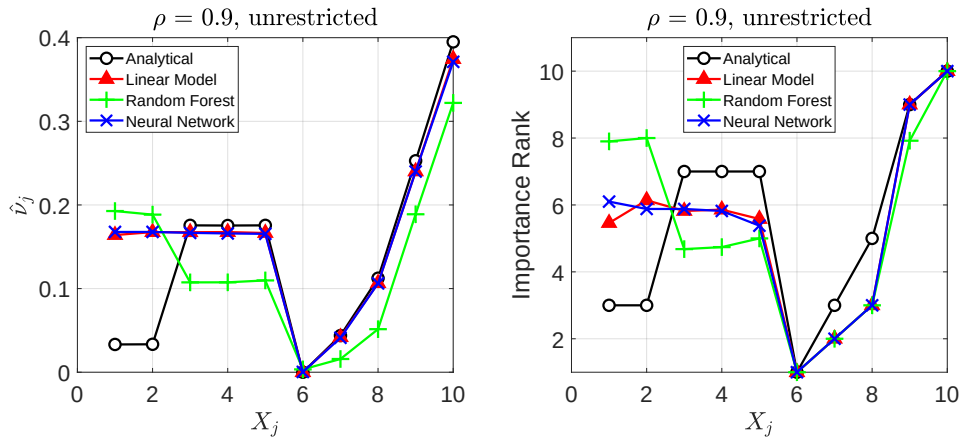
:HookerTrue

Table 2 Ground Truth values for $\nu_j = 2\tau_j$ in the Hooker test case.

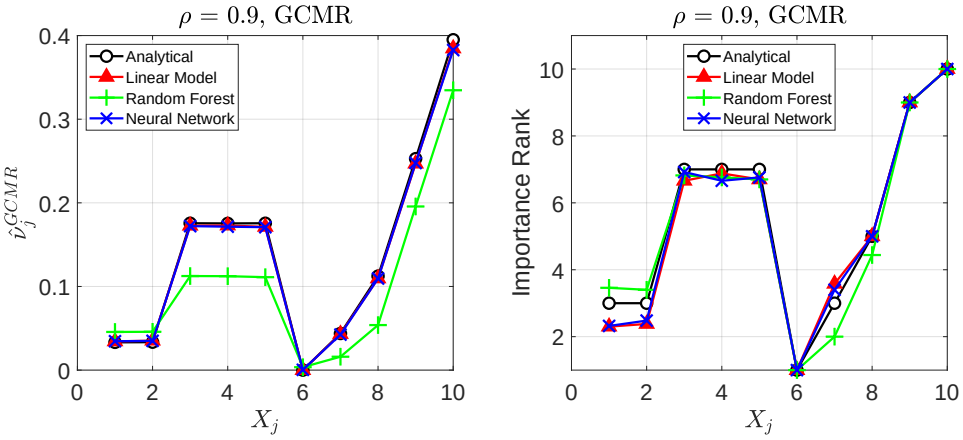
Feature	X_1, X_2	X_3, X_4, X_5	X_6	X_7	X_8	X_9	X_{10}
$\rho_{X_1,X_2} = 0.00$	0.1667	0.1667	0	0.0417	0.1067	0.2400	0.3750
$\rho_{X_1,X_2} = 0.90$	0.0332	0.1756	0	0.0438	0.1124	0.2528	0.3950

The ground truth values. Having the true model g at our disposal, we can exploit Proposition 4 and determine the ground truth values of the variable importance measures. In particular, we know that $\nu_j = 2\tau_j$ in this case [Equation (22)]. Then, we can employ unbiased and consistent estimators for τ_j to calculate the ground truth values. We report the resulting values in Table 2 for $\rho_{X_1,X_2} = 0.00$ and $\rho_{X_1,X_2} = 0.9$. For the uncorrelated case, the values of ν_1 and ν_2 coincide with those of ν_3, ν_4 and ν_5. When X_1 and X_2 are correlated, $\nu_1 = \nu_2$ decreases to 0.0332, while the importance of X_3, X_4 and X_5 slightly increases.

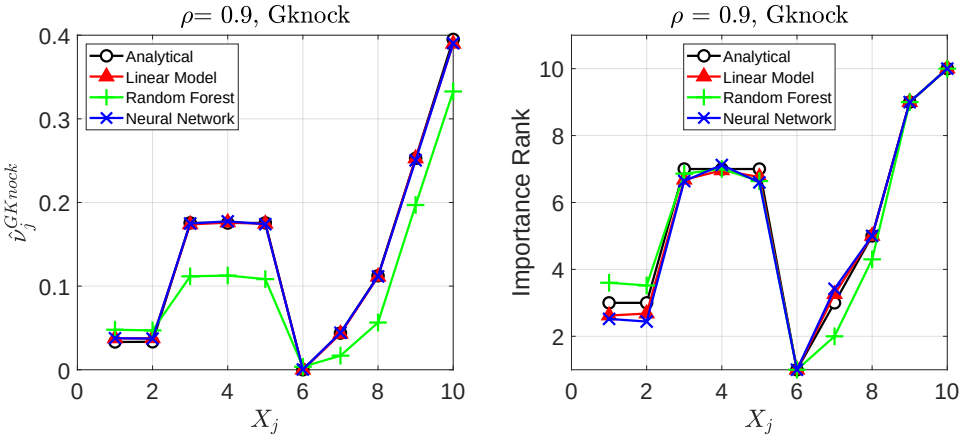
Setup. We proceed as in Hooker et al. (2021). We sample a feature datasets of size $N = 2000$ from the variable distributions for each correlation assignment. We then use the model in Equation (25) to synthetically generate the corresponding values of Y. On each input-output dataset, we train a linear model, a neural network, and a Random Forest as in Hooker et al. (2021), using the MATLAB subroutines FITLM, FITNET, and FITRENSEMBLE. The linear model and the neural network show systematically similar performance, with an average mean squared error MSE ≈ 0.01 and coefficient of determination $R^2 \approx 0.99$. The Random Forest has a slightly lower performance, with average MSE ≈ 0.12 and $R^2 \approx 0.85$. For the two input-output datasets and the ML models, we calculate $\widehat{\nu}_j$ and the new feature importance measures discussed in this work, $\widehat{\nu}_j^{\mathrm{GCMR}}$, $\widehat{\tau}'_j$, $\widehat{\tau}'^{,\mathrm{GCMR}}_j$, $\widehat{\tau}_j^{\mathrm{ALE}}$ and $\widehat{\kappa}_j^{\mathrm{ALE}}$. The local effects needed to compute $\widehat{\tau}_j^{\mathrm{ALE}}$ and $\widehat{\kappa}_j^{\mathrm{ALE}}$ are extracted from a MATLAB implementation of the ALE plots in the package ALEPLOT by Apley and Apley (2018). Similarly to Hooker et al. (2021), we perform 50 replicates of the experiments and report the average values.



(a) Estimates (left) and ranking (right) based on $\widehat{\nu}_j$ (unrestricted permutations).



(b) Estimates (left) and ranking (right) based on $\widehat{\nu}_j^{\mathrm{GCMR}}$ (restricted permutations).



(c) Estimates (left) and ranking (right) based on $\widehat{\nu}_j^{\mathrm{GKnock}}$ (restricted permutations).

Figure 7 **Left panels: $\widehat{\nu}_j$, $\widehat{\nu}_j^{\mathbf{GCMR}}$, $\widehat{\nu}_j^{\mathbf{GCMR}}$. Right panels: corresponding importance-ranking as defined in Hooker et al. (2021). The symbols $\circ$, $\triangle$, $+$, and $\times$ denote the feature rankings for the original model (i.e., the analytical ranking), the linear model, the Random Forest, and the artificial neural network. Results are averages over 50 replicates.**

Results on Breiman's variable importance measures. We find two main insights. First, by restricting permutations, we avoid the importance overestimation occurring in the unrestricted case. Second, the conditional model reliance strategy as well as the Knockoffs strategy yield results in line with Proposition 4: estimates are close to the ground truth values of the total indices, confirming that the generated data remain within distribution.

Let us discuss the results in more detail. In the uncorrelated case, $\rho_{X_1,X_2} = 0$, as expected, the variable importance estimates are equal to twice the total indices with the ground truth values in the second row of Table 2, that is $\widehat{\nu}_j = \widehat{\nu}_j^{\text{GCMR}} = \widehat{\nu}_j^{\text{GKnock}} = 2\widehat{\tau}_j$.

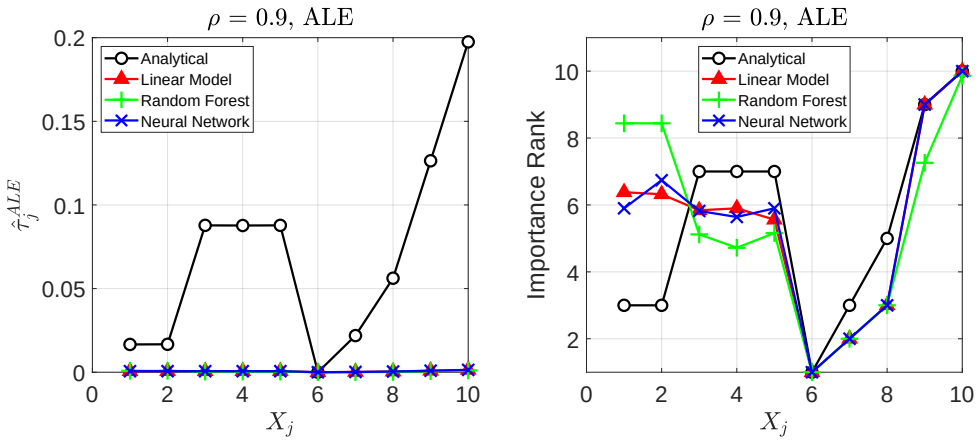
Switching to the $\rho_{X_1,X_2} = 0.9$ case, Figure 7 displays the variable importance measure estimates in the left panels, and the corresponding ranking in the right panels for $\rho_{X_1,X_2} = 0.9$.

The top row reports estimates of ν_j, without permutation restrictions. The values of $\widehat{\nu}_1$ and $\widehat{\nu}_2$ differ from the ground truth for all models. Items 2 in Proposition 4 and Corollary 2 help explain the findings. The estimates of ν_j when features are uncorrelated tend to $2\tau_j'$ rather than $2\tau_j$. When we break the dependence between X_1 and X_2 when freely permuting them, all features become uncorrelated, so that τ_1' and τ_2' coincide with the total indices we would obtain in the absence of correlation, yielding the Breiman's importance indices reported in the first row of Table 2.

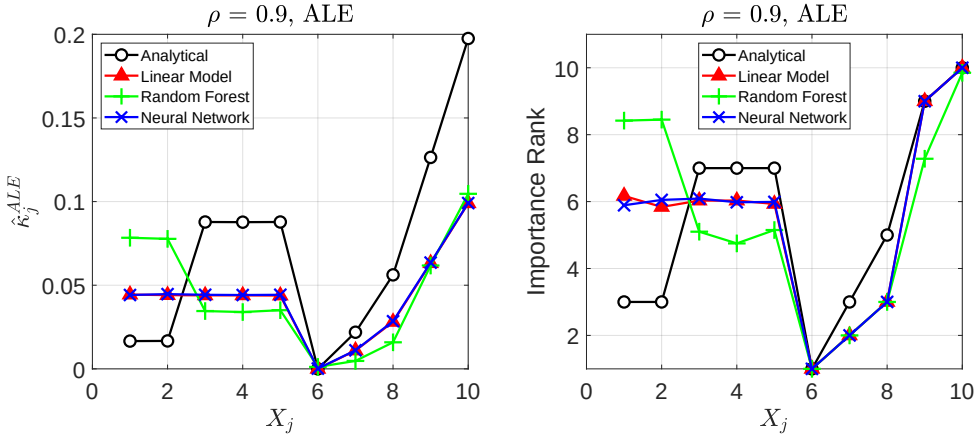
The left panel in the central row of Figure 7 displays the estimates of $\widehat{\nu}_j^{\text{GCMR}}$ obtained using Algorithm 1. For the linear regression and the neural network, the values of $\widehat{\nu}_j^{\text{GCMR}}$ practically coincide with the ground-truth values $2\tau_j$. For the Random Forest, we record greater variability, due to the lower accuracy of this model. The left panel in the bottom row of Figure 7 shows that restricting permutations with the Gaussian-Knockoffs strategy yields similar results, with the importance measure estimates matching the ground truth values for the neural network and the linear model, and some greater variability of the Random Forest. The results are in line with items 3 in Proposition 4 and Corollary 2. We expect that with accurate model predictions, estimates obtained with these algorithms tend to twice the value of $\widehat{\tau}_j$, as the generated data remain within distribution. However, we have less control over their convergence if the model is inaccurate.

Disregarding, for the moment, our knowledge of the ground truth values, the results show that calculating variable importance measures with permutation restrictions yields lower values than without permutation restrictions for this model. The reason is that the input-output dependence is linear and permutation restrictions yield new points $\widehat{g}\left(x'_{n,j}, \mathbf{x}^n_{-j}\right)$ closer to the original ones $\widehat{g}(\mathbf{x}^n)$. For a linear model, this implies smaller squared differences in Equation (4), and therefore, lower importance indices.

Results on ALE-based indices. Figure 8 reports results for the ALE-plot-based indices. For $\widehat{\tau}_j^{\text{ALE}}$, the estimates produce unreliable feature ranking. However, with $\widehat{\kappa}_j^{\text{ALE}}$, we find the same ranking obtained with total indices. In fact, for a linear model, the identity in Equation (18) holds, and



(a) Estimates (left) and ranking (right) based on $\widehat{\tau}_j^{\mathrm{ALE}}$.



(b) Estimates (left) and ranking (right) based on $\widehat{\kappa}_j^{\mathrm{ALE}}$.

Figure 8 **Estimates (left panels) and ranking (right panels) of $\widehat{\tau}_j^{\mathrm{ALE}}$ (upper row) and $\widehat{\kappa}_j^{\mathrm{ALE}}$ (lower row) for the Hooker test case. Results are averages over 50 replicates.**

this identity yields the same value as the total indices for the independent case in agreement with Corollary 1 (Item 2).

5. Applications

This section discusses applications of the proposed strategies for variable importance in the context of the Boston Housing (Section 5.1), and Name Ethnicity (Section 5.2) datasets.

5.1. Boston Housing

Our goal with this dataset is to test the efficacy of variable importance measures with and without permutation restrictions and to compare them with SHAP explanations. We fit several machine learning models and select the three best-performing ones. Some insights are anticipated here. As we have learned, unrestricted variable importance measures overestimate the feature importance. Extrapolation errors are particularly severe for one of the models, and for this model, the extrapolations in computing the SHAPs are extreme, resulting in unreliable variable importance estimates.

The Gaussian imputation and restriction on Knockoffs strategies effectively reduce the extrapolation effect.

Description of the Boston Housing dataset. The data has been recorded in 1978 by the U.S Census Service (Harrison and Rubinfeld 1978), with 13 features listed in Table 3, and 506 observations, each of which represents a house in the Boston area. The features are characteristics of houses (e.g., CHAS

Description

Table 3 Features and target in the Boston Housing dataset

Symbol	Acronym	Description
X_1	CRIM	per capita crime rate by town
X_2	ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
X_3	INDUS	proportion of non-retail business acres per town
X_4	CHAS	Charles River vicinity categorical Variable (1 or 0)
X_5	NOX	nitric oxides concentration (parts per 10 million)
X_6	RM	average number of rooms per dwelling
X_7	AGE	proportion of owner-occupied units built prior to 1940
X_8	DIS	weighted distances to five Boston employment centres
X_9	RAD	index of accessibility to radial highways
X_{10}	TAX	full-value property-tax rate per $10,000[/10k]$
X_{11}	PTRATIO	pupil-teacher ratio by town
X_{12}	B	The result of the equation $B = 1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
X_{13}	LSTAT	% lower status of the population
Y	MEDV	Median value of owner-occupied homes in $1000's[k]$

is the vicinity to the Charles river) and the target (Y) is their sales value. The mean house price $\mathbb{E}[Y]$ and variance $\mathbb{V}[Y]$ are approximately equal to \$22.53K and \$84.30K dollars^2, respectively (We omit the units in the remainder of this section).

Setup. We trained several ML models on this dataset, with the data split into 80% training and 20% testing (All subroutines are available on a dedicated github not revealed for anonymity?). At the end of these preliminary experiments, we select the three best-performing models for our experiments: a linear regression with pairwise interaction terms, an artificial neural network with two seven-neuron layers, and a Random Forest. The models exhibit the following testing performance: model coefficients of determination (R^2) of about 0.91, 0.86, and 0.80, respectively, root mean squared errors of 2.7, 3.5, and 4.5, respectively, and mean absolute deviations of about 1.82, 2.70 and 2.4, respectively.

Variable importance measures for the artificial neural network. We start with results regarding Breiman's feature importance measures with unrestricted permutations, ν_j, and with restricted permutations ν_j^{GCMR}, τ_j^{GKnock}, when the model is the artificial neural network. The bars in Figure 9 show that the variable importance measures agree in ranking RAD (X_{10}) and TAX (X_{11}) as most important features, followed by RM (X_6) and LSTAT (X_{13}). The remaining features have lower

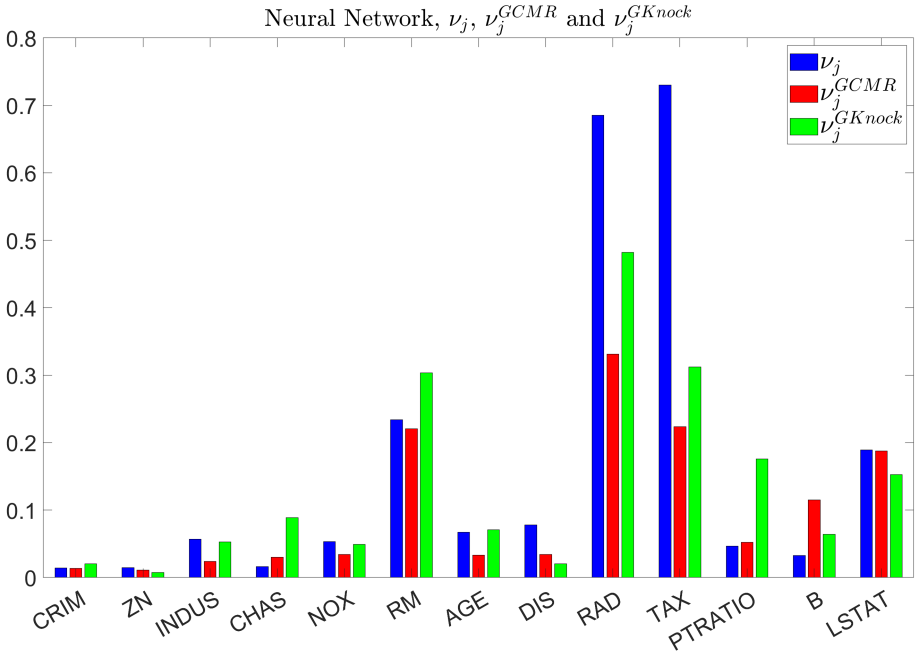


Figure 9 **Permutation based importance measures ν_j (blue), and ν_j^{GCMR} (red), and ν_j^{GKnock} (green) for the neural network.**

sBreiman_NN

relevance. However, the importance values of RAD and TAX are inflated by unrestricted permutations.Restricting permutations decreases ν_{TAX} by about 300% when compared to $\nu_{\text{TAX}}^{\text{GCMR}}$ and by about 200% when compared to $\nu_{\text{TAX}}^{\text{GKnock}}$.

We analyze this aspect further by comparing three distributions: (i) model predictions using the original data; (ii) model predictions when feature X_j has been permuted; and (iii) values of Y in the original data. We use the entire sample in this case. We juxtapose the corresponding empirical distributions in Figure 10. The left panel in Figure 10 showcases a *fat right tail in predictions made by the model after unrestricted permutations; the tail of the dashed red line extends from beyond 55, which is about the maximum of the original feature values, up to values of 100.* The tail is curtailed at around 65 when permutations are restricted (right panel).

Let us now compare these results with those of the SHAPs (Figure 11). A calculation involving all feature subsets repeated for all model predictions would make the computational time infeasible. Following the steps of Lundberg and Lee (2017), we rely on the linearity assumption (Equation 12 in Lundberg and Lee 2017) and employ the model predictions obtained by fixing the remaining variables $\mathbf{x}_{-u}$ at their mean value. In spite of this, the SHAP analysis takes 153.5s, about 20 times higher than the estimation time of the permutation-based importance measures. We visualize the SHAPs in the violin-and-box plot of Figure 11. Figure 11 shows that there is an overall agreement about the first five most important features, when compared to permutation-based variable importance measures. However, the SHAPs would attribute higher relevance to RM and LSTAT for the neural

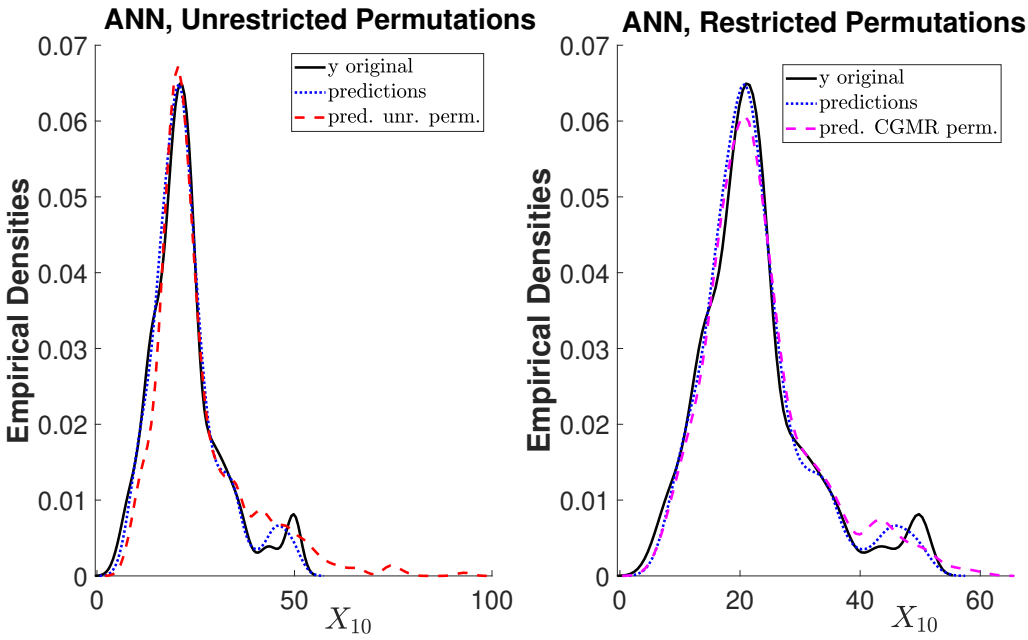


Figure 10 **Boston Housing: Artificial Neural network predictions with free permutations of X_{10} (left) and with permutations restricted by GCMR (right).**

auRestrX2X9

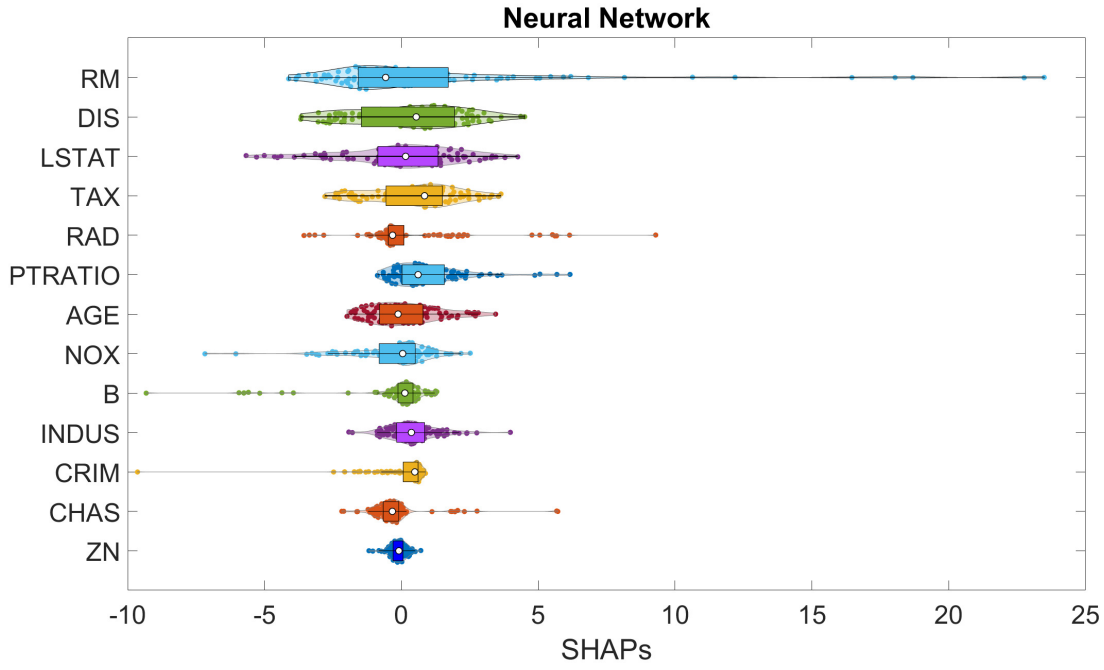


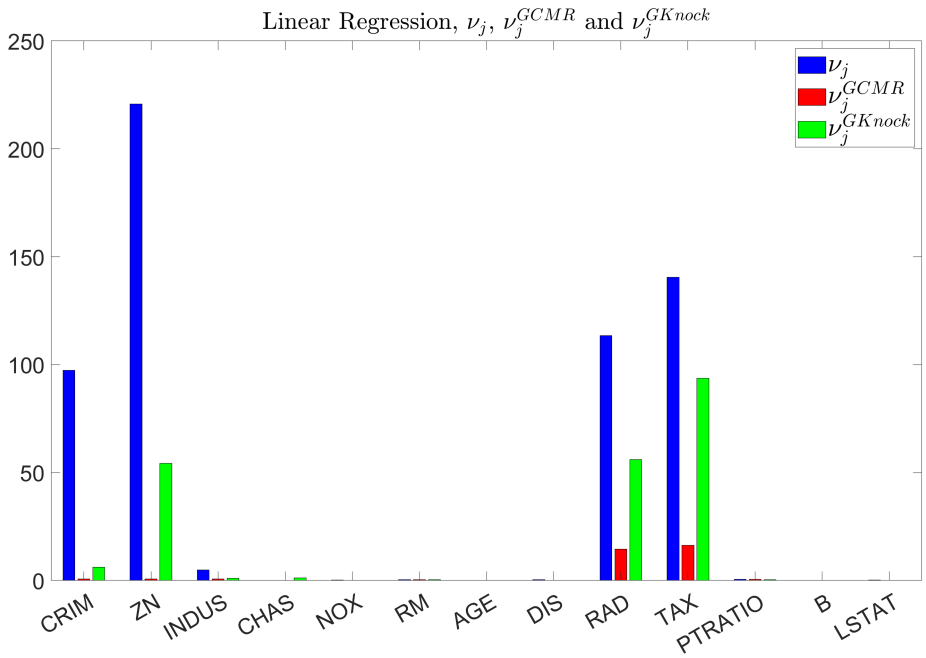
Figure 11 **Features SHAPs when the model is the neural network. Vertical axis: features. Horizontal axis: SHAP values. The features are then listed in descending order after sorting based on their mean absolute value.**

psNNTornado

network. There are two potential reasons for the differences. The SHAPs produce results prediction by prediction and inferring overall importance exposes us to some degree of arbitrariness regarding the aggregation method. As in Senoner et al. (2022), we aggregated them by the mean absolute value,

but different aggregation criteria may yield different rankings. Moreover, the SHAPs do not take any provision against free permutations and the model may be forced to extrapolate when computing them. We illustrate this point further in the next paragraph.

Variable importance measures for the linear model. Figure 12 reports the values of the permutation-based importance measures, for the linear model. There is a notable difference between the importance



primeLinReg **Figure 12** **Feature importance measures for the Linear Regression. Left panel: ν_j, ν^{GCMR}. Right panel: τ_j', ν_j^{GCMR}**

measures in Figures 12 and 9. Without permutation restriction, for the linear regression, the most important feature would be ZN, followed by TAX, RAD and CRIM, with the remaining features playing a very minor role. However, this importance ranking is affected by extrapolation errors, which is revealed by restricting permutations. Consider features ZN and CRIM. *Without restrictions, we record* $\nu_{ZN} \approx 220$ *and* $\nu_{ZN} \approx 95$ *(first two blue bars in Figure 12). After restrictions with GCMR, these values plummet to* $\nu_{ZN}^{GCMR} \approx 0.72$ *and* $\nu_{CRIM}^{GCMR} \approx 0.65$*; that is, three orders of magnitude smaller.* After permutation restrictions, RAD and TAX become the most important variables, which is a similar ranking result to the neural network. The green bars in Figure 12 reveal that the GKnock strategy deflates the feature importance measures, but to a lesser extent than GCMR. (For instance, ν_{ZN} remains at a high $\nu_{ZN}^{\text{GKnock}} \approx 54$ even after permutation restriction with GKnock).

We investigate this aspect further by considering the model predictions with and without permutations of ZN in Figure 13. Comparing the density of the predictions with and without permutation restrictions (dashed red line and dotted blue line, respectively, in Figure 13) shows that the long tail

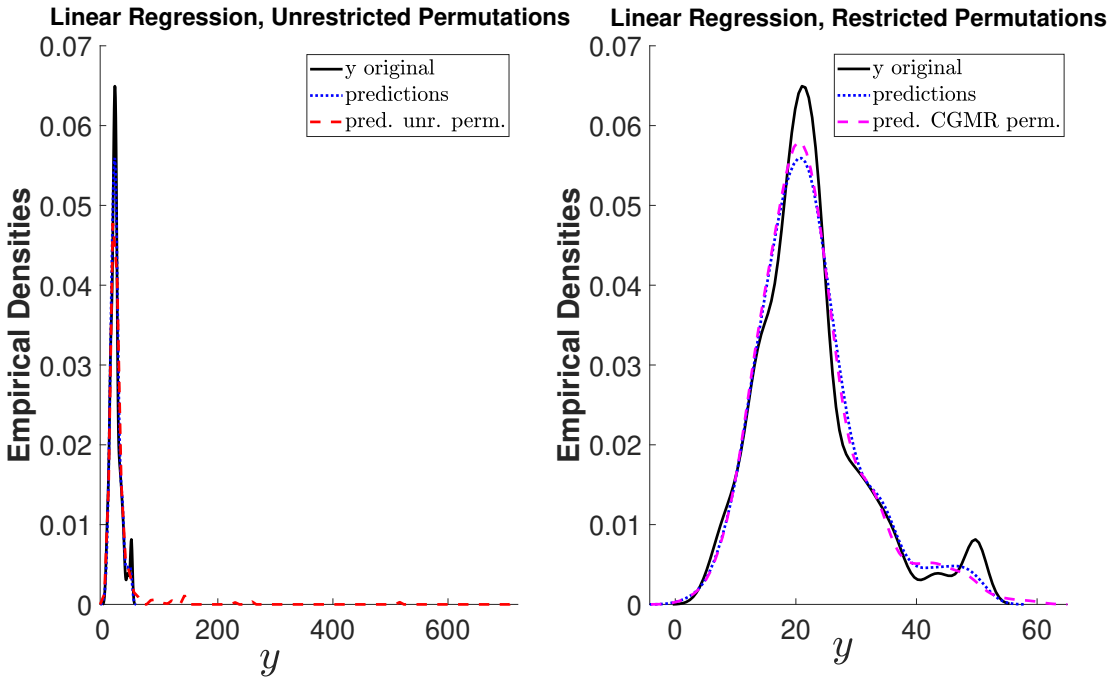


Figure 13 **Linear Regression: prediction densities with free permutations of X_2 (left) and with permutations restricted by GCMR (right). Note the difference in horizontal axes between the plots.**

inRegX2hist

associated with unrestricted permutations is curtailed when permutations are restricted by GCMR. We then compare these results with an analysis of the SHAPs. Figure 14 shows that features TAX

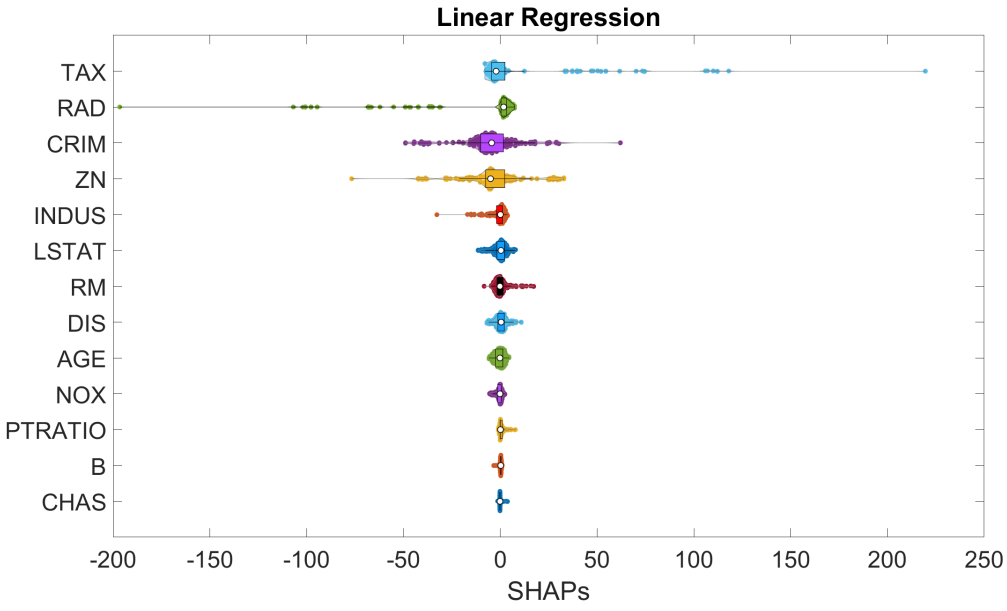
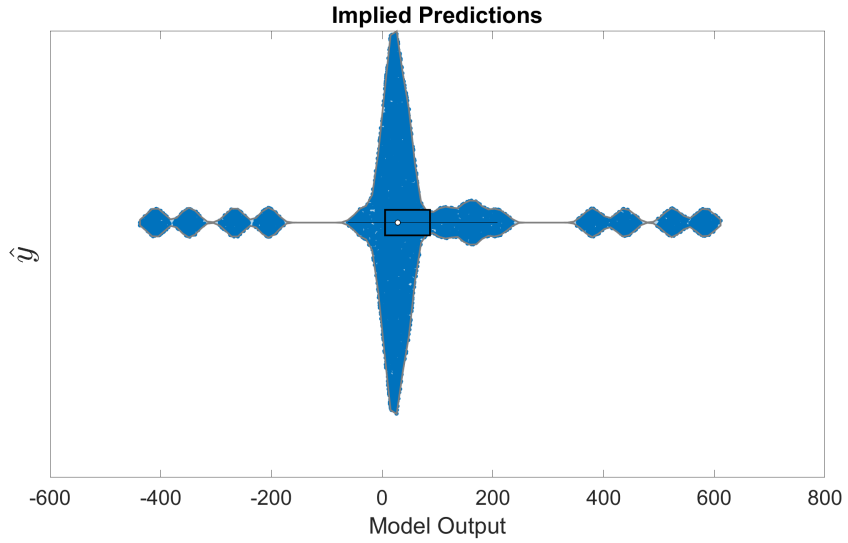


Figure 14 **Feature SHAPs when the model is a linear regression.**

nRegTornado

and RAD are ranked as most important, followed by CRIM, ZN, INDUS, with NOX, B, and CHAS



ountedhouse

Figure 15 Violin-and-box plot of predictions created by the calculation of the SHAPs of TAX.

playing a minor role. *Keeping in mind that SHAPs are in the same units as the predictions, notably, the SHAPs associated with TAX range from a few slightly negative values up to 220, and the SHAPs associated with RAD range from about -200 up to a few positive values just lower than 10. The values of these explanations are much larger than the values of the predictions to be explained.* As a reference, the value SHAP=220 for TAX is recorded for house #35 (in the test subset, which corresponds to realization 414). This house has a market value of 16.3 and the linear regression predicts a value of 12.4; an explanation of 220 makes little sense. To help us understand the values of these SHAPs, we report all the predictions implied by the calculation of the SHAP for TAX in Figure 15. *The graph shows that several predictions (on the horizontal axis) are orders of magnitudes different from the original scale.* The boxplots are barely noticeable, showing that the analysis is driven by outliers. *It is difficult to place confidence in predictions of the value of a house when this figure is approximately -$420K.* Such prediction would imply the that the buyer *receives* a very high sum of money to reside in the house, a suspicious insight. A similar argument accompanies the interpretation of the results for RAD. These results show that the design of the SHAPs forces the model to extrapolate, making the SHAPs unreliable.

Variable importance measures for the Random Forest. The analysis reveals the absence of extrapolation. There is an overall consistency in the ranking of the alternative variable importance measures (ν_j, ν_j^{GCMR}, ν_j^{GKnock}) and also with the SHAPs, which agree in indicating LSTAT and RM as the most important variables, and lRAD, ZN, CHAS as the least important variables for the Random Forest. Further details are presented in Appendix C.

sec:Ethn

5.2. The Name Ethnicity Dataset

This section aims to evaluate the performance of variable importance measures for a dataset containing millions of entries, a size that poses significant computational challenges. Our findings indicate that while calculating the SHAP values becomes impractical, permutation-based importance measures remain computationally efficient. We then compare indices with and without permutation restrictions. Results show that permutation-restricted indices provide more convincing explanations regarding the model's dependence on the features than without permutation restrictions.

The *name-ethnicity* dataset is studied in Jain et al. (2022). It is a primary dataset for name-based ethnicity classification studies, which in turn are key tools in determining algorithmic fairness. The dataset divides the population into four ethnicities, named Asian, Black, Hispanic, and White. The distribution of data in these categories is uneven, with 1.9% of the individuals in the Asian, 14.2% in the Black, 16.7% in the Hispanic, and 67.1% in the White category, respectively. In Jain et al. (2022), the dataset is used to create an interpretable machine learning model that overcomes the problems related to the use of deep learning models for the same task. While capable of unveiling subtle patterns in character names, deep learning models are difficult to interpret. The approach of Jain et al. (2022) consists of extracting an ensemble of features that can be efficient predictors of ethnicity. After extracting these features, Jain et al. (2022) train a multivariate logistic regression (MLR) model.

The dataset at our disposal contains $13,043,270$ names of voters in selected US States. We train two one-hot-encoding machine learning models: a Random Forest and a neural network, using a randomly generated sample of 50% of the data, amounting to about six million entries. The resulting dataset is split into 80% training and 20% testing, with about 5 million entries for training and 1 million entries for testing. We utilize a personal PC with Intel processor 11th Gen Intel(R) Core(TM) i7-11800H, 2.30GHz, 80GB RAM, with MATLAB 2023a-based subroutines. The performance metrics of the two models are reported in Table 4.

ab:PerfEthn

Table 4 Model Performance for the Name Ethnicity Dataset.

Avg performance	Neural Networks		Random Forest	
over 4 classes	Train	Test	Train	Test
Precision	0.85	0.85	0.70	0.70
Recall	0.79	0.79	0.85	0.84
Accuracy	0.88	0.88	0.82	0.82
F1score	0.82	0.82	0.75	0.75
Area under the ROC Curve	0.94	0.94	0.96	0.96

Table 4 shows that the neural network exhibits a slightly better performance than the Random Forest. We then restrict attention to this model for simplicity of exposition.

For variable importance, we started calculating four indices: the SHAPs, ν_j, ν_j^{GCMR} and ν_j^{GKnock}. However, computing the SHAP values turned out to be impractical (after 8 hours SHAPs had processed 2,500 data out of 6 million). Therefore we restrict attention to permutation-based importance measures with and without permutation restrictions: $\nu_j, \nu_j^{\text{GCMR}}$ and ν_j^{GKnock}. Figures 16 shows that,

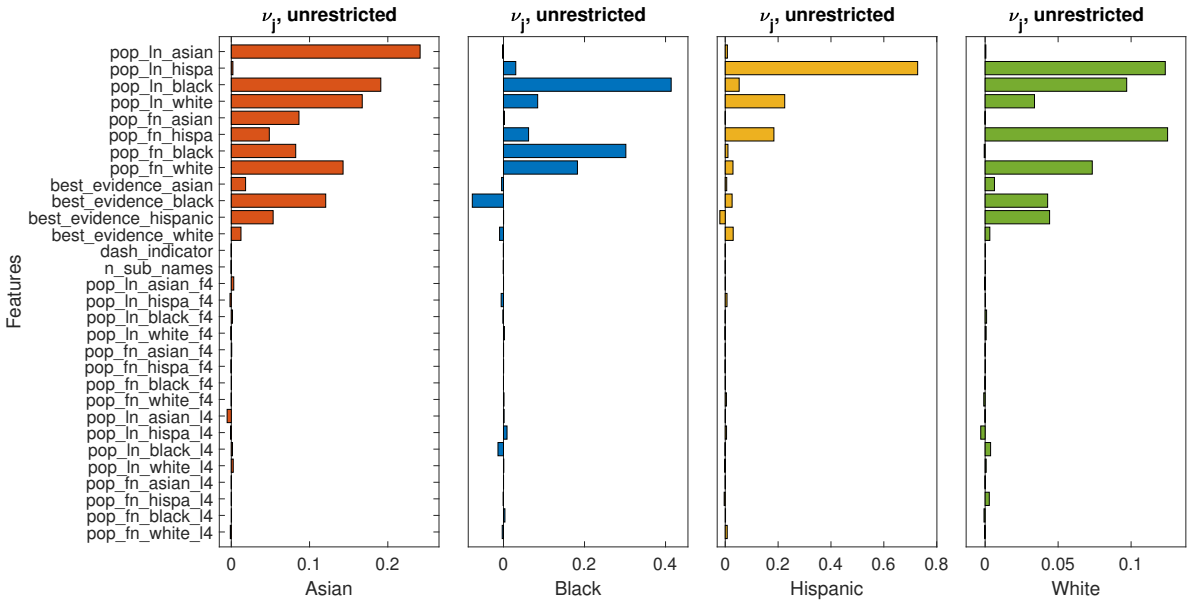


Figure 16 Breiman's importance with unrestricted permutations (ν_j) for the Name Ethnicity dataset. Each panel represents one ethnicity, in alphabetical order, Asian, Black, Hispanic, and White. The vertical axis reports the features, and the horizontal axis reports the values of $\widehat{\nu}_j$.

Nethnicity1

with unrestricted permutations, for Asian, Black, and Hispanic, the most important feature is the probability that a last name is of the same category. For instance, the probability that the last name is Asian (p_ln_asian) is the most important feature for the Asian category. The White category makes an exception, with the probability of the last name being White ranking seventh, and with the probability of the first name being Hispanic as the most important feature, followed by the probability of the last name being Hispanic and by the probability of last name Black; this makes sense because it is often easier to identify names of other categories than names in the White category. Figure 17 shows the results obtained with permutations restricted by the conditional imputation after Gaussian transformation, ν_j^{GCMR}. For the Asian, Black and Hispanic categories, the most important feature is the probability of the corresponding last name. For the White category, instead, we record a tie between the probability of a Hispanic last name and that of a Black last name, followed by the probability of a White first name. The probability of a White last name ranks seventh. Figure 18 displays the results obtained with ν_j^{GKnock}. The results agree with the results of ν_j^{GCMR}, with a substantially high concordance on the most important features. The ranking agreement between ν_j

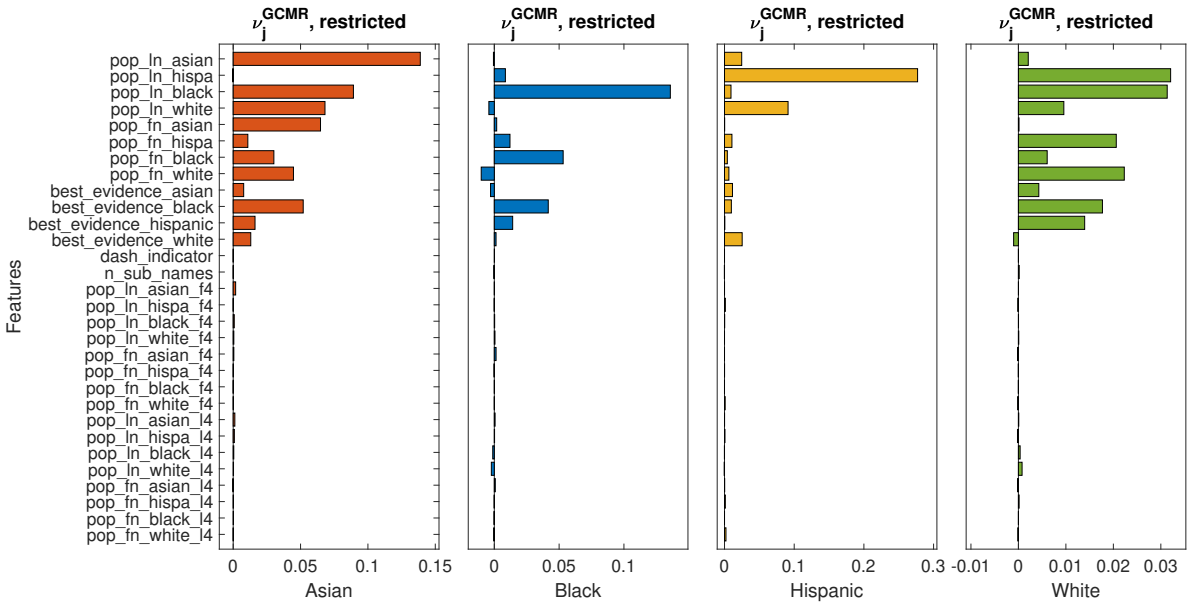


Figure 17 **Breiman's importance with restricted permutations (ν_j^{GCMR}) for the Name Ethnicity dataset. Each panel represents an ethnicity, in alphabetical order, Asian, Black, Hispanic, and White. The vertical axis reports the features, the horizontal axis reports the values of $\widehat{\nu_j^{GCMR}}$.**

Nethnicity2

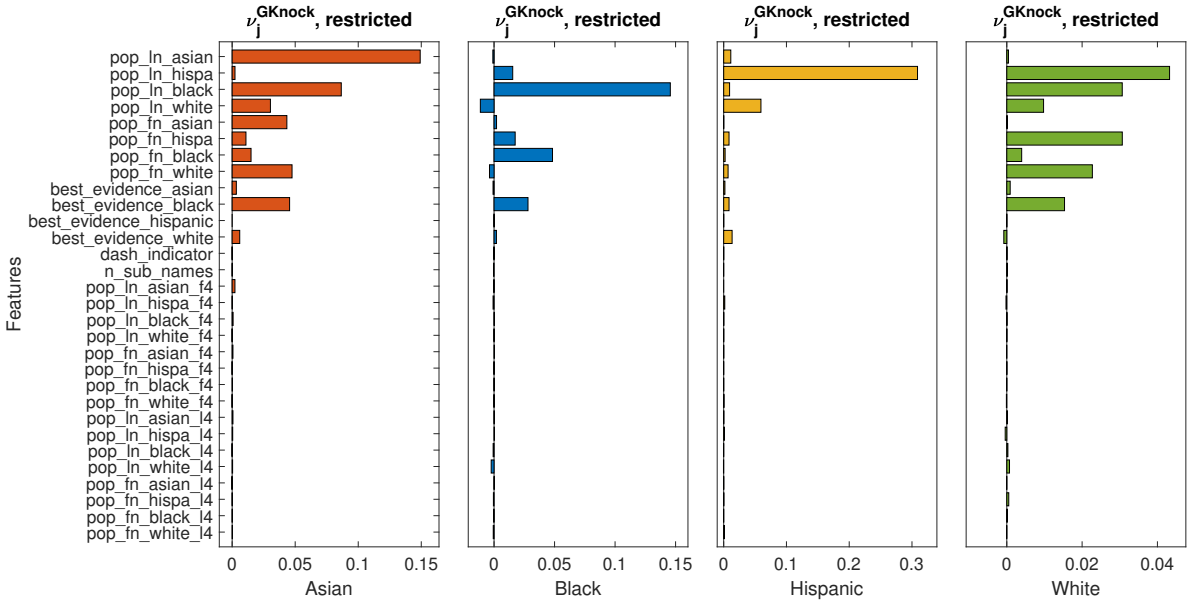


Figure 18 **Breiman's importance with restricted permutations (ν_j^{GKnock}) for the Name Ethnicity dataset. Each panel represents an ethnicity, in alphabetical order, Asian, Black, Hispanic, and White. The vertical axis reports the features, the horizontal axis reports the values of $\widehat{\nu_j^{GKnock}}$.**

Nethnicity3

(without permutation restrictions), and ν_j^{GCMR} and ν_j^{GKnock} (with permutation restrictions) is high for the Asian, Hispanic, and White categories.

The interesting part of this analysis is in where these variable importance analyses disagree. There is a notable disagreement between ν_j on the one side and ν_j^{GCMR} and ν_j^{GKnock} on the other side, for the Black category. Without restricting permutations we obtain a significantly negative value for the feature best_evidence_black (Figure 16, third panel). This is avoided using ν_j^{GCMR} or ν_j^{GKnock}, according to which this variable becomes relevant and with a positive importance. This is in greater agreement with intuition, as having best_evidence_black important in the Black category seems natural.

Overall, the variable importance measures indicate that the neural network, for the White category (the most frequent in the dataset), tends to rely on features outside this category, whereas, for the Asian, Black, and Hispanic categories tends to rely on features within these respective categories.

As a last step, we compare the feature ranking yielded by the restricted permutation indices ν_j^{GKnock} and ν_j^{GCMR} with the ranking induced by the magnitudes of the coefficients of the multinomial regression model of Jain et al. (2022). The magnitudes of the coefficients are taken from Figure 2 on p. 17 of Jain et al. (2022). To compare the rankings obtained with the permutation-restricted importance measures, we consider the heatmap in Figure 19. The figure shows some insights. First, the MLR and the Neural Network do not rely on the same features. To illustrate, best_evidence_asian ranks first or second for the MLR in all categories, whereas it ranks in intermediate positions for the Neural Network. However, both models tend to rely more on the first 12 than on the remaining features. This is true for the neural network, while there are two exceptions for the MLR, with probability_asian_last_name_l4 (the of the last four letters being of an Asian name) ranking fifth for the Hispanic category and probability_black_last_name_l4 ranking fifth for the Asian category. This result also shows that the MLR tends to rely less on "within-the-category" variables than the Neural Network for all categories.

conclusion

6. Final Remarks

Extrapolation is known to cause misleading and unstable variable importance calculations (Hooker et al. 2021). Our work proposes new solutions to variable importance that reduce or eliminate extrapolation risk. We have developed and studied four new indices, two based on new designs for Breiman's variable importance measures (ν_j^{GCMR} and ν_j^{GKnock}) and two based on the ALE plot designs (τ_j^{ALE} and κ_j^{ALE}) and compared them to existing alternatives, finding that the permutation-restricted ones (ν_j^{GCMR} and ν_j^{GKnock}) have substantial benefits, both theoretically and empirically, over the other approaches.

The SHAPs have been heavily criticized as variable importance measures for their lack of logical consistency in classification problems by Marques-Silva and Huang (2024). Our experiments show that the underlying algorithm is exposed to strong extrapolation errors. Our variable importance

	MLR				Neural Network			
Feature	Asian	Black	Hispanic	White	Asian	Black	Hispanic	White
probability_asian_last_name	1	3	3	2	1	26	4	10
probability_hispanic_last_name	7	6	8	11	30	5	1	1
probability_black_last_name	4	2	4	3	2	1	5	2
probability_white_last_ name	12	24	12	4	3	29	2	6
probability_asian_first_ name	3	4	2	6	4	6	19	16
probability_hispanic_first_name	16	15	10	17	10	4	5	3
probability_black_first_name	6	7	14	10	7	2	10	7
probability_white_first_name	17	20	28	7	6	29	9	4
best-evidence-asian	2	1	1	1	11	27	8	8
best-evidence-black	8	13	6	13	5	3	5	5
best_evidence_hispanic	18	5	7	5	8	9	20	14
best_evidence_white	26	23	26	15	9	7	3	30
dash_indicator	10	10	15	9	25	16	23	19
n_sub_names	19	16	21	24	23	22	29	13
probability_asian_last_name_f4	14	11	9	22	12	17	22	21
probability_hispanic_last_name_f4	25	28	30	30	26	22	11	26
probability_black_last_name_f4	11	12	13	14	15	13	16	17
probability_white_last_name_f4	29	19	24	23	16	14	21	19
probability_asian_first_name_f4	28	22	17	21	17	8	24	28
probability_hispanic_first_name_f4	30	29	25	26	24	19	24	22
probability_black_first_name_f4	9	9	16	8	27	15	27	23
probability_white_first_name_f4	23	14	22	29	22	17	14	26
probability_asian_last_name_l4	13	18	5	20	13	10	18	15
probability_hispanic_last_name_ l4	24	21	23	28	14	10	13	29
probability_black_last_name_l4	5	8	11	16	18	25	28	11
probability_white_last_name_l4	22	27	27	25	19	28	30	9
probability_asian_first_name_l4	21	25	19	18	29	10	16	25
probability_hispanic_first_name_l4	27	30	29	27	21	20	14	11
probability_black_first_name_l4	15	17	20	12	20	20	26	17
probability_white_first_name_l4	20	26	18	19	28	24	12	24

Figure 19 Ranking heatmap. The four columns under MLR report the rankings induced by the magnitudes of the coefficients of the multinomial regression. The four columns under Neural Network report the average ranking between ν_j^{GKnock} and ν_j^{GCMR}. A white background indicates a rank between 1 and 5, light grey between 6 and 15, and dark grey a ranking below 15. To illustrate Figure 19 consider the row of the first feature, probability_asian_last_name. The value 1 under the Asian ethnicity column indicates indicates that this feature is ranked first for this category by the MLR coefficient.

nHeatMmapNN

measures ν_j^{GCMR} and ν_j^{GKnock} do not have this problem, and we have theoretical guarantees under which the new data used for computing Breiman's indices *remain within distribution*.

The most important aspect of our work is that the new permutation-restricted variable importance measures ν_j^{GCMR} and ν_j^{GKnock} – despite arising from completely different calculations – yield almost identical results in practice. Unlike other variable importance measures (e.g., SHAP, ALE indices), they are *stable*, with results that make sense intuitively. All other methods we studied showed instability, often relying on predictions that were orders of magnitude too large than any reasonable outcome could possibly be. Those variable importance calculations clearly are not trustworthy.

The variable importance measures ν_j^{GCMR} and ν_j^{GKnock} are easy to calculate and do not require retraining the model. This allows them to be used with large datasets, like the Name Ethnicity

dataset, which has millions of observations. For new – even complicated – models like neural networks on this dataset, we provided the first stable and reliable variable importance calculations.

References

AgarKenn23 Agarwal A, Kenney AM, Tan YS, Tang T, Yu B (2023) Mdi+: A flexible random forest-based feature importance framework. *arXiv:* 2307:1–83.

2018package Apley D, Apley MD (2018) Package 'ALEPlot'.

Apley2020 Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82:1059–1086.

BarbCand23 Barber RF, Candès EJ, Ramdas A, Tibshirani RJ (2023) Conformal prediction beyond exchangeability. *Annals of Statistics* 51:816–845.

BarbCand15 Barber RF, Candés EJ (2015) Controlling the false discovery rate via knockoffs. *Annals of Statistics* 43:2055–2085, ISSN 00905364.

Bier83 Bier V (1983) A measure of uncertainty importance for components in fault trees. *Transactions of the American Nuclear Society* 45:384–5.

Brei02 Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32.

BenaDaVe22 Bénard C, Veiga SD, Scornet E (2022) Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika* 109:881–900, ISSN 1464-3510, URL `http://dx.doi.org/10.1093/biomet/asac017`.

Candes2018 Candès E, Fan Y, Janson L, Lv J (2018) Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 80:551–577.

LundSHaps23 Chen H, Covert IC, Lundberg SM, Lee SI (2023) Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence* 5:590–601, ISSN 2522-5839, URL `http://dx.doi.org/10.1038/s42256-023-00657-x`.

ChenPelg24 Chen L, Pelger M, Zhu J (2024) Deep learning in asset pricing. *Management Science* 70:714 – 750.

ong2020810a Dong J, Rudin C (2020) Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence* 2(12):810–824.

llyEtAl2023 Donnelly J, Katta S, Rudin C, Browne EP (2023) The rashomon importance distribution: Getting rid of unstable, single model-based variable importance. *Proceedings of Neural Information Processing Systems (NeurIPS)*.

FishRudi19 Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20:1–81.

riedman2001 Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 29:1189–1232.

GambGhad20 Gambella C, Ghaddar B, Naoum-Sawaya J (2020) Optimization problems for machine learning: A survey. *European Journal of Operational Research* Forthcomin:1–83.

arrison1978 Harrison D, Rubinfeld D (1978) Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 5:81–102.

HartGrem18 Hart JL, Gremaud PA (2018) An approximation theoretic perspective of sobol indices with dependent variables. *International Journal for Uncertainty Quantification* 8:483–493, arXiv:1801.01359v3.

tie19941255 Hastie T, Tibshirani R, Buja A (1994) Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* 89:1255–1270.

Helton1993 Helton J (1993) Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. *Reliability Engineering and System Safety* 42:327–367, cited By (since 1996): 156 Export Date: 29 June 2007 Source: Scopus.

Homma1996 Homma T, Saltelli A (1996) Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety* 52:1–17.

Hooker2021 Hooker G, Mentch L, Zhou S (2021) Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing* 31.

HuanRosh24 Huang C, Joseph RV (2024) Factor importance ranking and selection using total indices. *ArXiv* 2401:1–43, URL `https://arxiv.org/abs/2401.00800`.

igami1990aa Ishigami T, Homma T (1990) An Importance Quantification Technique in Uncertainty Analysis for Computer Models. *Proceedings. First International Symposium on Uncertainty Modeling and Analysis*, 398–403, URL `http://dx.doi.org/doi:10.1109/ISUMA.1990.151285`.

ainEnRu2022 Jain V, Enamorado T, Rudin C (2022) The Importance of Being Ernest, Ekundayo, or Eswari: An Interpretable Machine Learning Approach to Name-Based Ethnicity Classification. *Harvard Data Science Review* 4(3), https://hdsr.mitpress.mit.edu/pub/wgss79vu.

en_cpc_1999 Jansen MJW (1999) Analysis of variance designs for model output. *Computer Physics Communications* 117:35–43.

erenko2017a Kucherenko S, Iooss B (2017) Derivative-Based Global Sensitivity Measures. Ghanem R, Higdon D, Owhadi H, eds., *Handbook of Uncertainty Quantification*, 1241–1263 (Springer International Publishing).

KuchTara12 Kucherenko S, Tarantola S, Annoni P (2012) Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications* 183:937–946.

LeiGsel18 Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018) Distribution free predictive inference for regression. *Journal of the American Statistical Association* 113:1094–1111.

LundLee17 Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, volume 2017-Decem, 4766–4775.

MaraTara12a Mara T, Tarantola S (2012) Variance-based Sensitivity Indices for Models with Dependent Inputs. *Reliability Engineering & System Safety* 107:115–121.

MaraTara15a Mara TA, Tarantola S, Annoni P (2015) Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environmental Modelling & Software* 72:173–183.

MarqHuan24 Marques-Silva J, Huang X (2024) Explainability is not a game. *Communications of the ACM* 67:66–75, URL `http://dx.doi.org/10.1145/3635301`.

MaseOwen22 Mase M, Owen AB, Selier BB (2022) Variable importance without impossible data. *ArXiv* 2205.:1–36.

dSing19PNAS Murdoch WJ, Signh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods and applications in interpretabile machine learning. *Proceedings of the National Academy of Sciences* 116:22071–22080.

RenCand23 Ren Z, Candès E (2023) Knockoffs with side information. *Annals of Applied Statistics* 17:1152–1174.

RomaSesi20 Romano Y, Sesia M, Candès E (2020) Deep knockoffs. *Journal of the American Statistical Association* 115:1861–1872.

i_jasa_2002 Saltelli A, Tarantola S (2002) On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. *Journal of the American Statistical Association* 97:702–709.

SemeRudi22 Semenova L, Rundin C, Parr R (2022) On the existence of simpler machine learning models. *ArXiv* :1908:1–46.

etl22MSShap Senoner J, Netland T, Feuerriegel S (2022) Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science* 68:5704–5723.

SesiSaba19 Sesia M, Sabatti C, Candès EJ (2019) Gene hunting with hidden markov model knockoffs. *Biometrika* 106:1–18.

ol_mcs_2009 Sobol IM, Kucherenko S (2009) Derivative Based Global Sensitivity Measures and their Links with Global Sensitivity Indices. *Mathematics and Computers in Simulation* 79:3009–3017.

Song2019142 Song S, Zhou T, Wang L, Kucherenko S, Lu Z (2019) Derivative-based new upper bound of sobol' sensitivity measure. *Reliability Engineering and System Safety* 187:142–148.

ouySesh23OR Souyris S, Seshadri S, Subramanian S (2023) Scheduling advertising on cable television. *Operations Research* 71:2217 – 2231.

SundNajim20 Sundararajan M, Najmi A (2020) The many shapley values for model explanation. *Proceedings of 37th International Conference on Machine Learning* 119:1–11.

VerdWass23 Verdinelli I, Wasserman L (2023) Feature importance: A closer look at shapley values and loco. *ArXiv* 2303.:1–14.

Appendix A: Proofs

[Proof of Proposition 1] If the joint distribution of $\mathbf{X}$ can be mapped via a Gaussian copula, the random vector $\mathbf{Z}$ in line 3 of Algorithm 1 is multivariate normal. Then, the residuals $\overline{\mathbf{Z}}$ in line 7 are normally distributed with a variance that is independent of the actual value $\mathbf{z}_j$, and rearranging them to produce the new residuals $\overline{\mathbf{Z}}^{\pi}$ yields $\mathbf{Z}'$ to have the same distribution as $\mathbf{Z}$. Then, mapping back $\mathbf{Z}'$ through the marginal transformation returns a vector $\mathbf{X}'$ which has the same distribution as $\mathbf{X}$.

[Proof of Proposition 2] The proof is similar to the previous one, with the additional observation that, by the results in Candès et al. (2018), if Z is a multivariate normal random vector, then the Knockoffs Z' has the same probability distribution of Z. The values Z' are then transformed back into X' by a deterministic relationship, which guarantees that X' is a Knockoff of X.

[Proof of Proposition 3] The "if"part is trivial. For the *only if* part, let us start with $\tau_j^{\mathrm{ALE}}(K)$ and suppose that $\tau_j^{\mathrm{ALE}}(K) = 0$ for all choices of z_j^k and z_j^{k-1}. By Equation (15), $\tau_j^{\mathrm{ALE}}(K)$ is the weighted sum of K positive conditional expectations. Then, $\tau_j^{\mathrm{ALE}}(K) = 0$ implies that

$$\mathbb{E}[(\widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^k, \mathbf{X}_{-j}))^2 | X_j \in \mathcal{X}_j^k],$$

for all $k = 1, 2, ..., K$. Then, for a generic k, the corresponding conditional expectation is null if $\widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^k, \mathbf{X}_{-j}) = 0$ almost everywhere for . Then, this last condition is equivalent to say that $\widehat{g}$ does not depend on X_j on z_j^k, $k = 1, 2, \ldots, K$. The assumption that this occurs for all selections of z_j^k completes the proof. For the $\widehat{\kappa}_j^{\mathrm{ALE}}$ indices, suppose that $\widehat{\kappa}_j^{\mathrm{ALE}} = 0$ for all of z_j^k and z_j^{k-1}. By definition, $\widehat{\kappa}_j^{\mathrm{ALE}}$ is the weighted sum of K positive ratios $\mathbb{E}\left[\left(\frac{\widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^k, \mathbf{X}_{-j})}{z_j^k - z_j^{k-1}}\right)^2\right]$. Then, $\widehat{\kappa}_j^{\mathrm{ALE}} = 0$ implies that $\mathbb{E}\left[\left(\frac{\widehat{g}(z_j^k, \mathbf{X}_{-j}^k) - \widehat{g}(z_j^{k-1}; \mathbf{X}_{-j}^k)}{z_j^k - z_j^{k-1}}\right)^2\right] = 0$ for all k. Then, observe that, in kALE, the generic summand in the k^{th} partition contains the expectation of a quantity of the type $\left(\frac{\widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^k, \mathbf{X}_{-j})}{z_j^k - z_j^{k-1}}\right)^2$. For this quantity to be null we need to have $\left(\widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^{k-1}, \mathbf{X}_{-j})\right)^2 = 0$ almost everywhere, because by construction $z_j^k - z_j^{k-1} \neq 0$. Then, this last condition is equivalent to say that, fixed $\mathbf{X}_{-j}$, $\widehat{g}(z_j^k, \mathbf{X}_{-j}) = \widehat{g}(z_j^{k-1}, \mathbf{X}_{-j})$. However, becayse the values z_j^k can be arbitrarily chosen, this implies that $\widehat{g}(z_j^k, \mathbf{X}_{-j})$ does not depend on X_j.

[Proof of Corollary 1] For Item 1, by Equation (17) for a linear model we have

$$\begin{aligned}
\widehat{\kappa}_j^{\mathrm{ALE}} &= \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left(\frac{\widehat{g}(z_j^k, \mathbf{X}_{-j}) - \widehat{g}(z_j^{k-1}; \mathbf{X}_{-j})}{z_j^k - z_j^{k-1}}\right)^2\right] \frac{\sigma_j^2}{\sigma_y^2} = \\
&\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left(\frac{a_j z_j^k - a_j z_j^{k-1} + \mathbf{a}_{-j}\mathbf{X}_{-j} - \mathbf{a}_{-j}\mathbf{X}_{-j}}{z_j^k - z_j^{k-1}}\right)^2\right] \frac{\sigma_j^2}{\sigma_y^2} \\
&= \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left(\frac{a_j(z_j^k - z_j^{k-1})}{z_j^k - z_j^{k-1}}\right)^2\right] \frac{\sigma_j^2}{\sigma_y^2} = \frac{a_j^2\sigma_j^2}{\sigma_y^2}.
\end{aligned}$$

For item 2, we recall that for a linear model without interactions and independent inputs, it is $\tau_j = \frac{a_j^2\sigma_j^2}{\sigma_y^2}$, and the thesis follows.

[Proof of Proposition 4] Under a quadratic loss, Equation (19) becomes

$$\nu_j = \mathbb{E}[\Big(Y - \widehat{g}(X_j^{'}, \mathbf{X}_{-j}; \theta^*)\Big)^2] - \mathbb{E}[(Y - \widehat{g}(\mathbf{X}; \theta^*))^2]. \tag{26}$$

eq:nusquare

Then, if the model is a perfect predictor, it is $Y = \widehat{g}(\mathbf{X}; \theta^*)$, so that the rightmost summand in Equation (26) becomes $\mathbb{E}[(Y - \widehat{g}(\mathbf{X}; \theta^*))^2] = 0$. Replace Y in the first summand with $\widehat{g}(\mathbf{X}; \theta^*)$ we obtain

$$\nu_j = \mathbb{E}[\Big(Y - \widehat{g}(X_j^{'}, \mathbf{X}_{-j}; \theta^*)\Big)^2 - (Y - \widehat{g}(\mathbf{X}; \theta^*))^2]. \tag{27}$$

For item 2, if $X_j^{'}$ is sampled independently of $\mathbf{X}_{-j}$, then we have

$$\nu_j = \int_{\mathcal{X}} \int_{\mathcal{X}_j} \Big(\widehat{g}(\mathbf{x}; \theta^*) - \widehat{g}(x_j^{'}, \mathbf{x}_{-j}; \theta^*)\Big)^2 dF_{X_j}(x_j^{'}) dF_{\mathbf{X}}(\mathbf{x}), \tag{28}$$

which implies Equation (21) by Equation (7).

For item 3, if $X_j^{'}$ is sampled conditionally on $\mathbf{X}_{-j}$, we have

$$\nu_j = \int_{\mathcal{X}} \int_{\mathcal{X}_j} \Big(\widehat{g}(\mathbf{x}; \theta^*) - \widehat{g}(x_j^{'}, \mathbf{x}_{-j}; \theta^*)\Big)^2 dF_{X_j|\mathbf{X}_{-j}}(x_j^{'}|\mathbf{x}_{-j}) dF_{\mathbf{X}}(\mathbf{x}), \tag{29}$$

so that $\nu_j = 2\tau_j$ by Equation (6).

[Proof of Corollary 2] Item 1. To prove that $\widehat{\nu}_j$ is an asymptotically unbiased estimator of $\mathbb{E}[(\widehat{g}(X_j^{'}; \mathbf{X}_{-j}) - \widehat{g}(\mathbf{X}))^2]$, we need to prove that under the assumptions of Proposition 4,

$$\lim_{N \to \infty} \widehat{\nu}_j(N) = \mathbb{E}[(\widehat{g}(X_j^{'}; \mathbf{X}_{-j}; \theta^*) - \widehat{g}(\mathbf{X}; \theta^*))^2] \tag{30}$$

and that $\mathbb{E}[\widehat{\nu}_j] = \mathbb{E}[(\widehat{g}(X_j^{'}; \mathbf{X}_{-j}; \theta^*) - \widehat{g}(\mathbf{X}; \theta^*))^2]$. For the limit, under a quadratic loss, we can write

$$\widehat{\nu}_j = \frac{1}{N} \sum_{n=1}^{N} (y^n - \widehat{g}(\mathbf{x}^n_{j,\text{perm}}; \theta^*))^2 - \frac{1}{N} \sum_{n=1}^{N} (y^n - \widehat{g}(\mathbf{x}^n; \theta^*))^2. \tag{31}$$

eq:nuhataux

Noting that, under the assumption of a perfect predictor, it is $y^n = \widehat{g}(\mathbf{x}^n; \theta^*)$, the right summand in Equation (31) vanishes and we have:

$$\widehat{\nu}_j = \frac{1}{N} \sum_{n=1}^{N} (\widehat{g}(\mathbf{x}^n; \theta^*) - \widehat{g}(\mathbf{x}^n_{j,\text{perm}}; \theta^*))^2, \tag{32}$$

which tends to $\mathbb{E}[(\widehat{g}(X_j^{'}; \mathbf{X}_{-j}; \theta^*) - \widehat{g}(\mathbf{X}; \theta^*))^2]$ by the law of large numbers. Also, taking the expectation we find

$$\begin{gathered} \mathbb{E}[\widehat{\nu}_j] = \mathbb{E}[\frac{1}{N} \textstyle\sum_{n=1}^{N} (\widehat{g}(X_j^{'}; \mathbf{X}_{-j}; \theta^*) - \widehat{g}(\mathbf{X}; \theta^*))^2] = \\ \frac{1}{N} \textstyle\sum_{n=1}^{N} \mathbb{E}[(\widehat{g}(X_j^{'}; \mathbf{X}_{-j}; \theta^*) - \widehat{g}(\mathbf{X}; \theta^*))^2] = \nu_j. \end{gathered} \tag{33}$$

Then, items 2 and 3 follow from how the new points are obtained in the alternative empirical procedures implied by Ψ_{DLoco} (τ_j'), $\widehat{\nu}_j^{\text{GCMR}}$ and $\widehat{\nu}_j^{\text{GKnock}}$. Specifically, in the calculations of Ψ_{DLoco} (τ_j') the points are freely permuted and therefore sampled from the marginal distribution of X_j, so that $\widehat{\nu}_j$ becomes an estimate of twice τ_j', by Proposition 4. Conversely, Propositions 1 and 2 imply a conditional resampling, which yields the classical total indices, again by Proposition 4.

sec:Ishi

Appendix B: Ishigami function

This section presents results of numerical experiments aimed at testing convergence of the estimates of the new indices, τ_j^{ALE} and κ_j^{ALE}. As a benchmark for the calculations, we use the values of the indices displayed in Table 1. The results are performed first using the true input-output mapping and then an ML tool fitted on a synthetic sample simulated from the original input-output mapping.

We start with a first batch of experiments in which we use directly the original input-output mapping in Equation (14). Figure 20 reports the graphs obtained by setting $N = 10^4$ and selecting a grid size of $K = 100$ to obtain smooth graphical representations. The graph of the ALE plot curve $ALE_2(x_2)$ (second panel in Figure 20) reports the marginal dependence of Y on X_2 exactly, thanks to the additive recovery property of ALE plots — the Ishigami function is additive in X_2. However, the ALE plot of X_3 is a flat line. This is

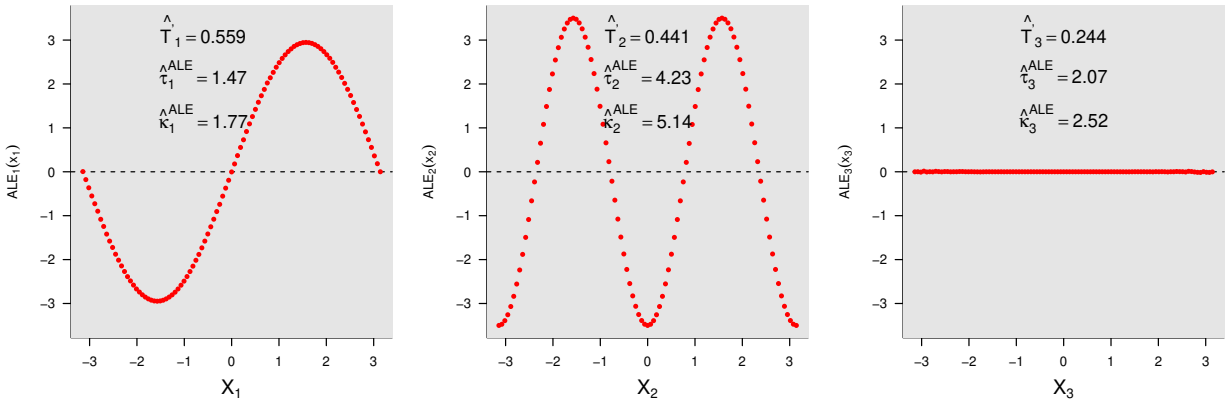
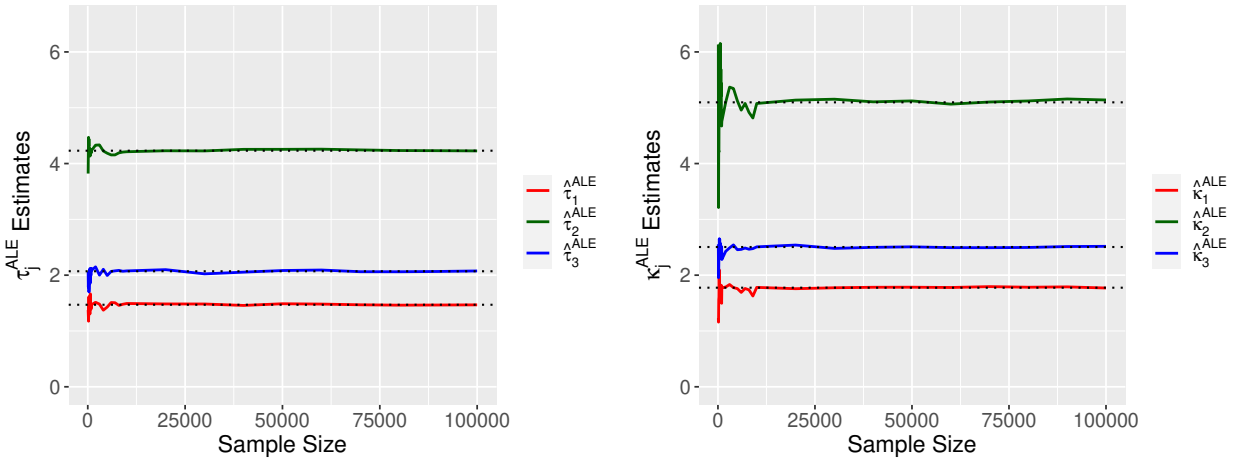


Figure 20 **Ishigami function: ALE plots, total indices τ_j (value normalized between 0 and 1) and ALE indices, τ_j^{ALE} and κ_j^{ALE}.**

ig_g_ALE_TI

a false negative, because we know that g is functionally dependent on X_3. However, calculating the values of τ_j^{ALE} and κ_j^{ALE} helps us understanding whether this feature is active in the model. We set $K = 10$ and report results for the estimates for τ_j^{ALE} and κ_j^{ALE} as the sample size increases. The estimates are expected to tend to the analytical values in the second row of Table 1.

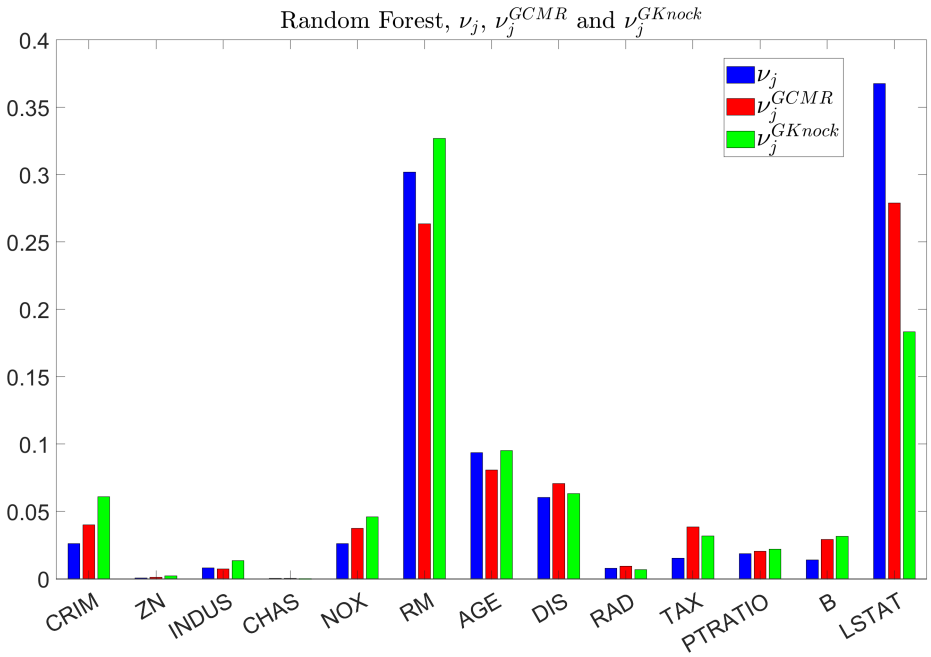


(a) Estimates of τ_j^{ALE} (b) Estimates of κ_j^{ALE}.

Figure 21 **Ishigami function: behavior of estimates of τ_j^{ALE} and κ_j^{ALE} as the sample size N increases, for $K = 10$. Dotted lines correspond to analytical values.**

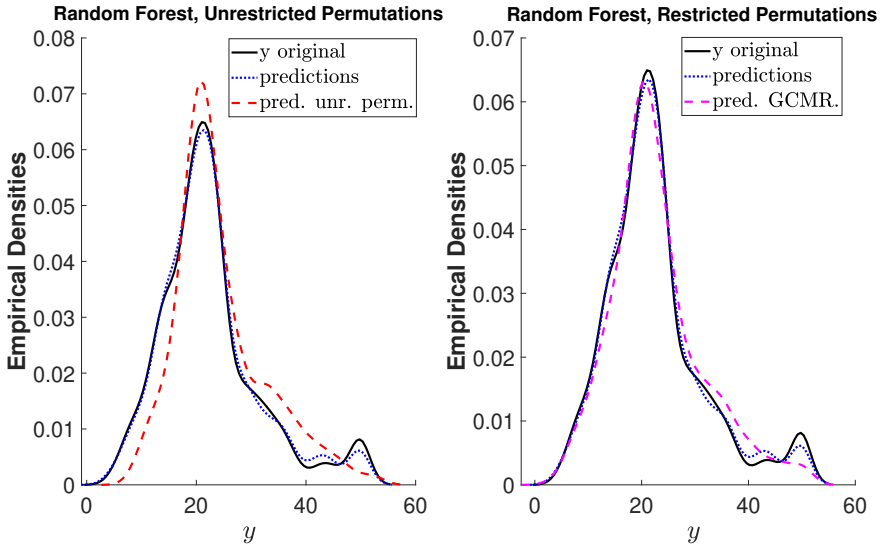
Figure 21 shows that $\widehat{\tau}_j^{\mathrm{ALE}}$ and $\widehat{\kappa}_j^{\mathrm{ALE}}$ converge rapidly toward the corresponding analytical values as the sample size increases. Their non-null values clarify that X_3 is an active feature. Thus, estimating the ALE importance measures simultaneously with an ALE plot avoids the null-conditional expectation false negative.

Regarding permutation variable importance measures, we perform a first series of experiments to test Proposition 4. We use the data generated with the true model to train a single hidden-layer neural network ($\widehat{g}$), using a sample of size $N = 10^5$. The data is divided into 80% training and 20% testing. The R^2 value for testing is 0.97. We then calculate the permutation feature importance measures for the three features, obtaining $\widehat{\nu}_1 = 14.6$, $\widehat{\nu}_2 = 12.1$, and $\widehat{\nu}_3 = 5.96$. These values are equal to twice the corresponding total indices, by Proposition 4. The neural network approximates the true feature-output mapping g with great accuracy, and there are no extrapolation problems because the features are independent. Finally, for τ_j^{ALE} and κ_j^{ALE} computed using $K = 10$ calculated on $\widehat{g}$, we obtain $\widehat{\tau}_1^{\mathrm{ALE}} = 1.09$, $\widehat{\tau}_2^{\mathrm{ALE}} = 4.61$, $\widehat{\tau}_3^{\mathrm{ALE}} = 0.99$ and $\widehat{\kappa}_1^{\mathrm{ALE}} = 1.30$, $\widehat{\kappa}_2^{\mathrm{ALE}} = 5.48$, $\widehat{\kappa}_3^{\mathrm{ALE}} = 1.17$. These values are close to the analytical ones in Table 1.



rimeRegTree

Figure 22 **Boston Housing: Feature importance measures ν_j, ν^{GCMR} and ν_j^{GKnock} for the Random Forest.**



g:RFX13hist

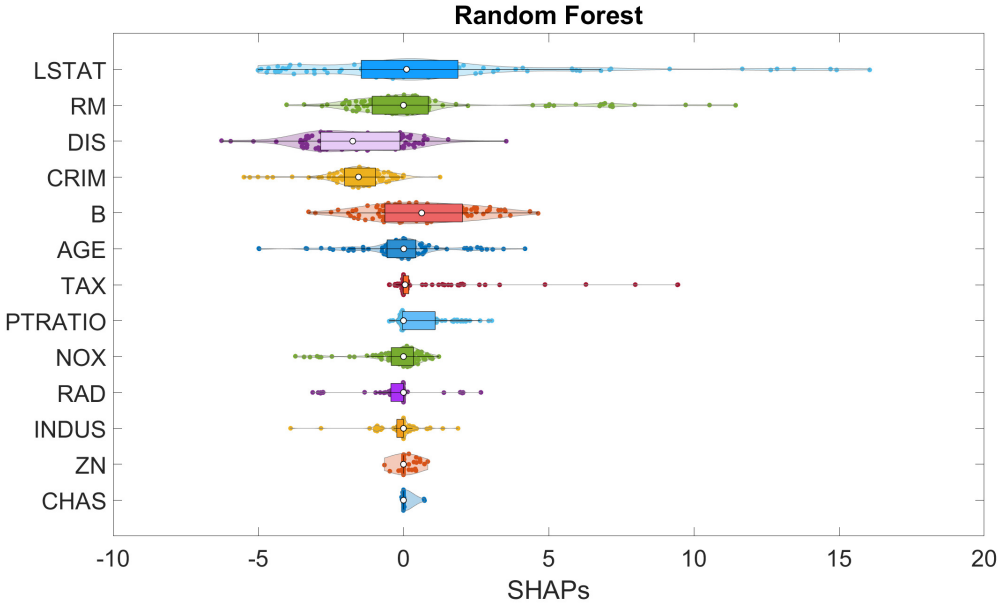
Figure 23 **Boston Housing: Random forest predictions when permuting LSTAT, without restriction (left panel), with CGMR restrictions (right panel).**

restresults

Appendix C: Boston Housing: Results for the Random Forest

Figure 22 displays the values of the feature importance measures. The bar charts in Figure 22 show that, for the Random Forest, the values of the feature importance measures are much less affected by extrapolation errors than the ones of the other models. This conclusion is also obtained by inspecting the resulting prediction plots (Figure 23). Figure 23 shows only small differences between the cases of unrestricted and restricted permutations. For the Random Forest, the most important variables are LSTAT and RM, followed

by AGE, CRIM, DIS and TAX. The remaining features play a minor role. As a comparison, we consider the SHAPs for this model.



g:RFtornado

Figure 24 Boston Housing: Features SHAPs for the Random Forest. LSTAT is ranked first, CHAS least.

Figure 24 shows that the SHAPs rank LSTAT and RM as the two most important features, and that they rank RAD, ZN, and CHAS as the three least relevant features, with the remaining variables with an intermediate importance, for an overall ranking correlation of 84%. Some differences are in the ranking of B, which is third with the SHAPs, but 8th with ν_j^{GCMR}, and INDUS, which is 7th with the SHAPS, but 11th with ν_j^{GCMR}. However, the lack of evident extrapolation and this agreement in the ranking allows us to conclude that inference about the most (LSTAT and RM) and least (RAD, ZN, CHAS) important features is robust for the Random Forest.

tionalthird

Appendix D: Boston Housing Results for the ALE indices

We discuss results for the calculation of ALE-based indices τ_j^{ALE} for the Boston Housing dataset. Figure 25

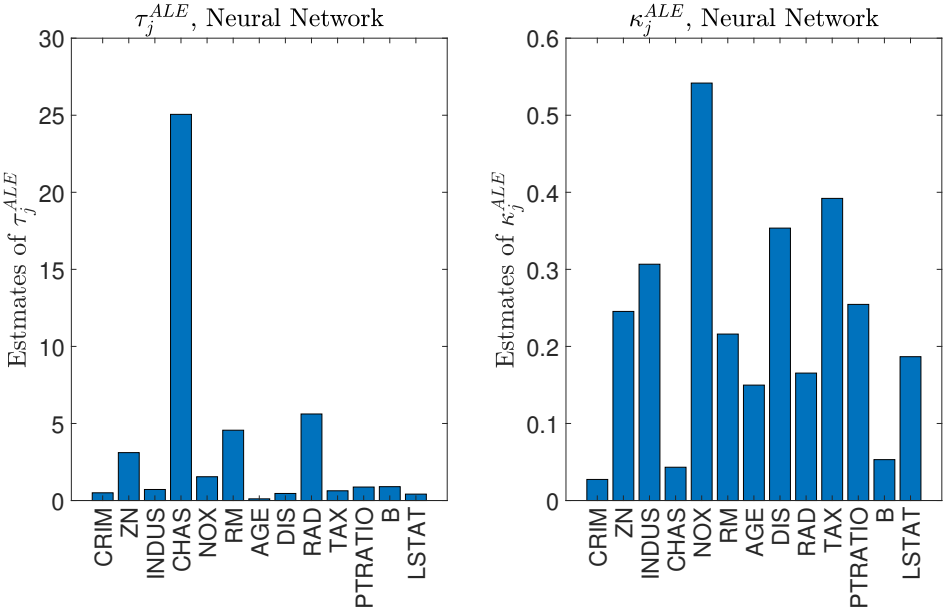


Figure 25 Boston Housng:$\widehat{\tau}_j^{\text{ALE}}$ and $\widehat{kappa}_j^{\text{ALE}}$ for the artificial Neural Network. The two indices tend to disagree on the variable importance.

ig:TauALENN

shows notable differences in the features identified by $\widehat{\tau}_j^{\text{ALE}}$ and $\widehat{\kappa}_j^{\text{ALE}}$. For instance, feature CHAS (X_4) is by far the most important variable for the neural network. In contrast, it is ranked among the least important variables by κ_j^{ALE}, which ranks NOX (X_6) as most important. Because the two indices are based on the same design, the difference stems from the renormalization occorring in Equation (17).

For this model, let us examine the predictions involved in the feature importance calculations. For CHAS, we report the empirical densities of the original data and of the predictions in Figure 26.

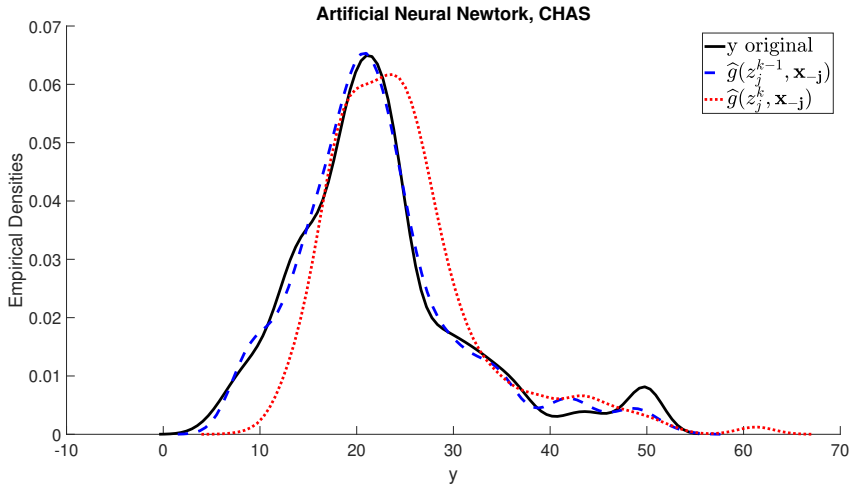
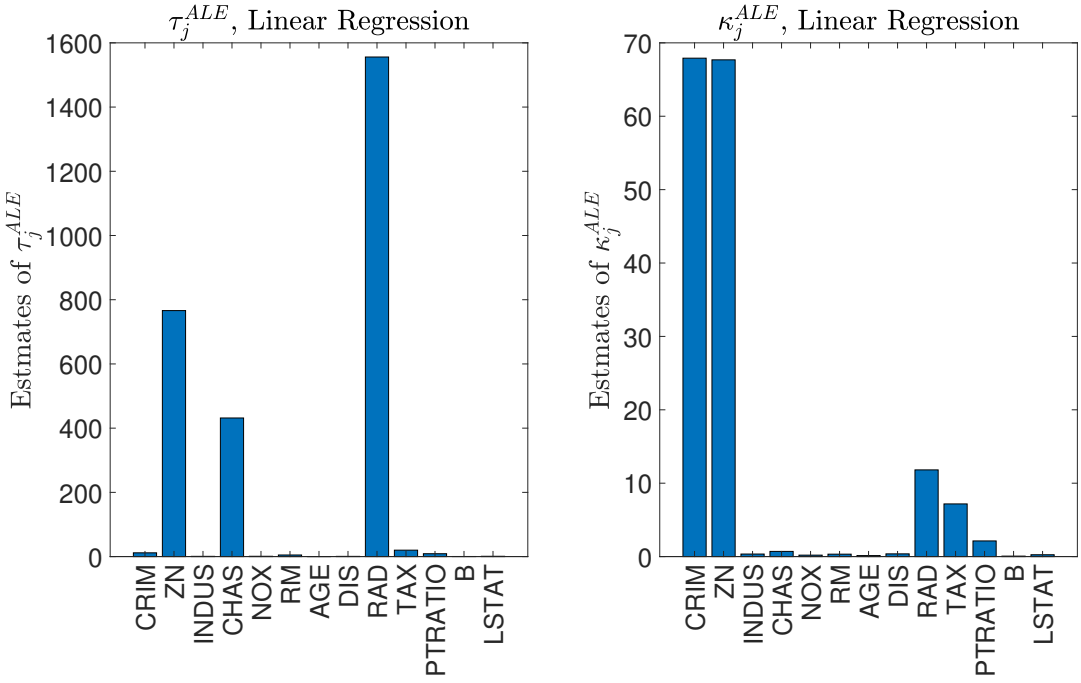


Figure 26 **Empirical densities for the model predictions of the neural network when computing ALE indices for CHAS, X_4. Continuous (black): empirical density of the original data. Dashed (blue): empirical density of predictions at points of the type $(z_j^{k-1}, \mathbf{x}_{-j})$. Dotted (red): empirical densities of predictions at points of the type $(z_j^{k}, \mathbf{x}_{-j})$.**

ies_CHAS_X4

Figure 26 shows that the empirical density of the predictions at points of the type $\hat{g}(z_j^k m\mathbf{x} - j$ is right-shifted when compared to the density of the original data. Also, the distribution has a longer tail, with predictions falling in the 60-80 region, which is far above the mean of 22.53. This signals a tendency to overestimation, although the extrapolation is mild.

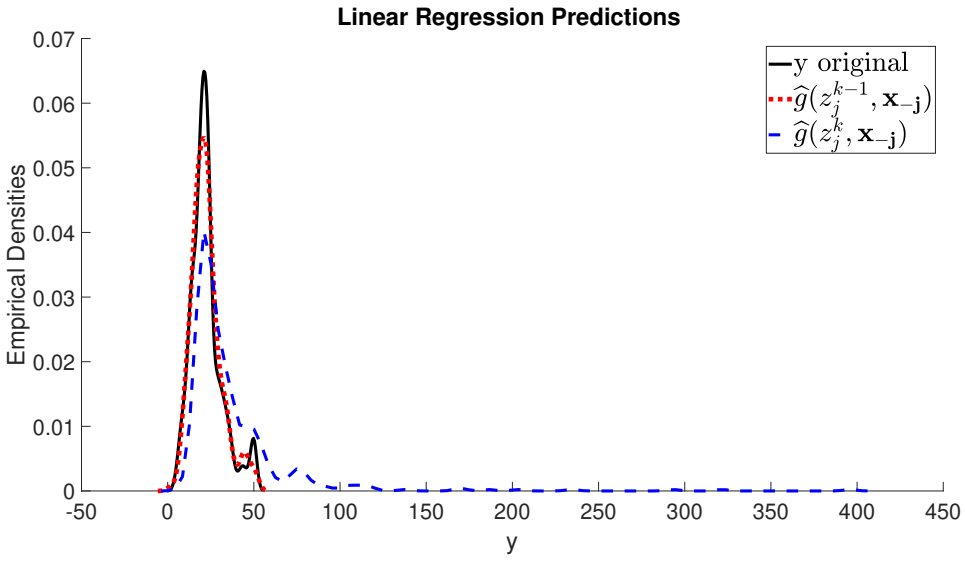
The second model is the linear regression. Figure 27 reports the ALE importance measures.



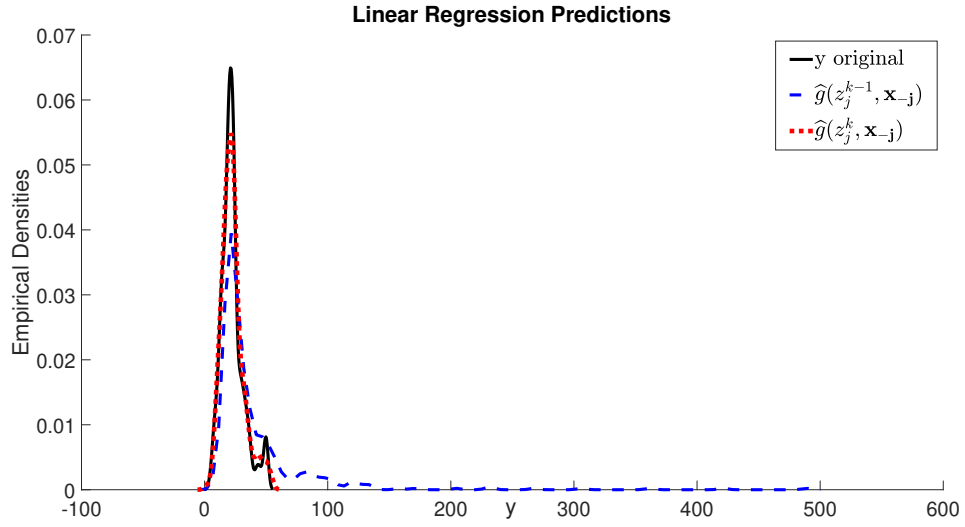
stonHousing

Figure 27 **Boston Housing: ALE importance indices for the Linear Regression.**

The panels signal notable disagreement between the two indices. Moreover, the left panel of Figure 27 shows importance values of $\widehat{\tau}_{ZN}^{\text{ALE}} \approx 800$ and $\widehat{\tau}_{RAD}^{\text{ALE}} \approx 1500$, much higher than the variance of Y, which is 84.60 (This value should be a benchmark for the values of $\widehat{\tau}_{j}^{\text{ALE}}$). These values are clearly untrustworthy and we argue they are the results of extrapolations. Examining the model predictions implied in the calculations of $\widehat{\tau}_{ZN}^{\text{ALE}}$ and $\widehat{\tau}_{RAD}^{\text{ALE}}$ provides a clear answer (Figures 28 and 29).



(a) Densities of Linear Regression Predictions when computing ALE indices for X_2.



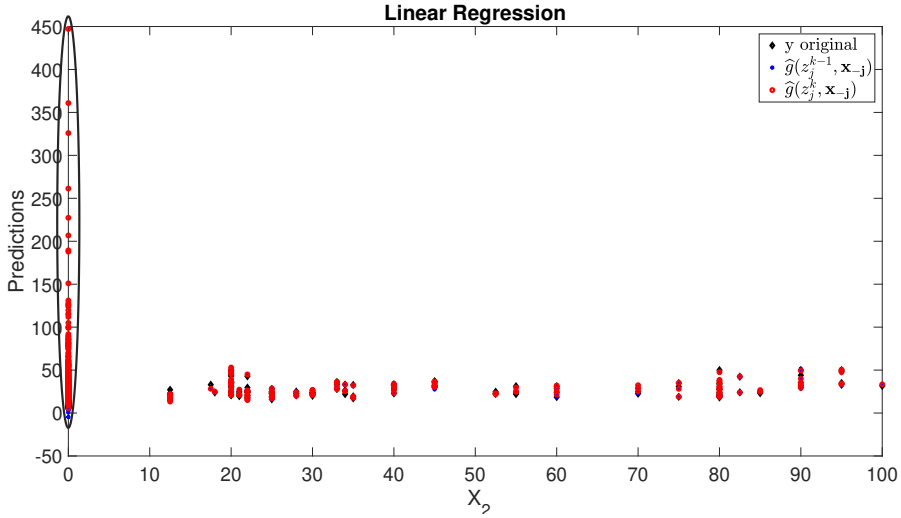
(b) Densities of Linear Regression Predictions when computing ALE indices for X_9.

Figure 28 Boston Housing: Densities of Linear Regression Predictions induced by the ALE indices design for X_2 and X_9. Continuous (black): empirical density of the original data. Dashed (blue): empirical density of predictions at points of the type $(z_j^{k-1}, \mathbf{x}_{-j})$. Dotted (red): empirical densities of predictions at points of the type $(z_j^{k}, \mathbf{x}_{-j})$.

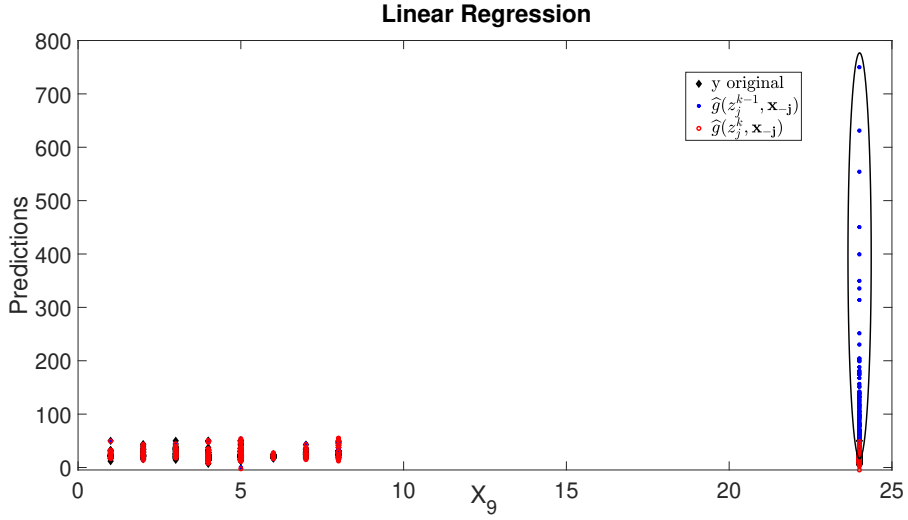
Figure 28 overlays the empirical densities of the original data (continuous, black), the model predictions at the left (dashed, blue) and right (dotted, red) points involved in calculating the ALE indices of the two most

important features. These graphs show that the implied model predictions are right-skewed, with extremely long right-tails, and unrealistic predictions falling in the 100-500 range (5 to 20 times the mean value of the data) for both ZN (X_2) and CHAS (X_9). Thus, estimates of $\widehat{\tau}_j^{\text{ALE}}$ are based on extrapolations and calculations are not reliable. For κ_j^{ALE}, the resulting indications are also untrustworthy, as they are based on the same design and model predictions.

To further investigate the source of these problems (in principle the ALE design should guard against extrapolation), we visualized the data creating a correspondence between the feature values and the predictions (Figure 29).



(a) Linear Regression: ALE plot induced predictions $\widehat{g}(x_j;\mathbf{x}_{-j})$ ($\diamond$), $\widehat{g}(z_j^k;\mathbf{x}_{-j})$ ($*$), and $\widehat{g}(z_j^{k+1};\mathbf{x}_{-j})$ ($\circ$). Left graph: X_2 (ZN). Right graph: X_9 (RAD).



(b) Linear Regression: Values of ALE plot predictions $\widehat{g}(x_j;\mathbf{x}_{-j})$ ($\diamond$), $\widehat{g}(z_j^k;\mathbf{x}_{-j})$ ($*$), and $\widehat{g}(z_j^{k+1};\mathbf{x}_{-j})$ ($\circ$). Left graph: X_2 (ZN). Right graph: X_9 (RAD).

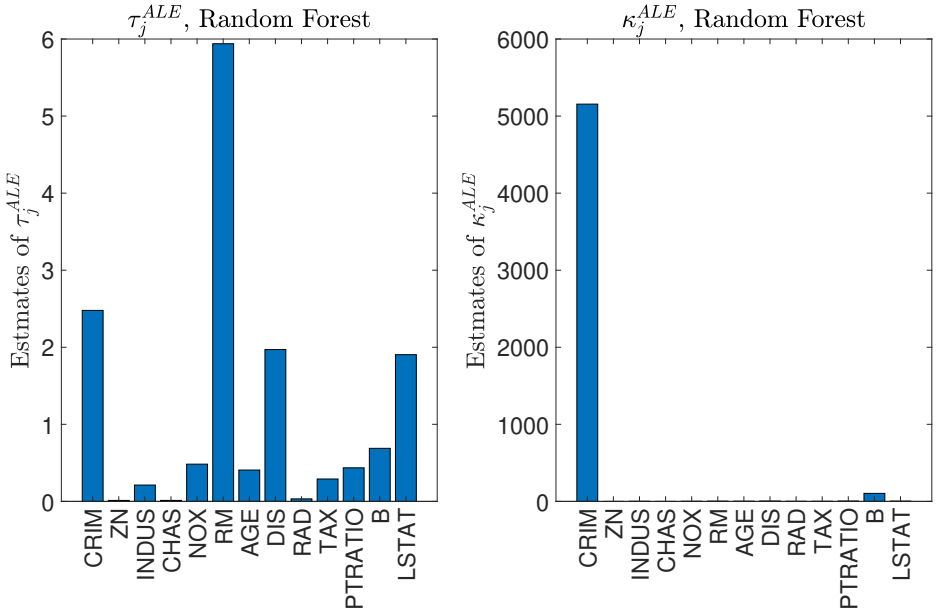
:LinRegX2X9ALEhist

Figure 29 **Boston Housing: linear regression, original data vs predictions in the calculations of ALE plots. The ellipses evidence predictions far from the original data.**

tHousLinReg

Figure 29 shows in a scatterplot the original data, the values of predictions of the type $\widehat{g}(z_2^k, \mathbf{x}_{-2})$ (red dots) and $\widehat{g}(z_9^{k-1}, \mathbf{x}_{-9})$ (blue dots). In the upper panel of Figure 29, we find points of the type $(z_2 = 0, \mathbf{x}_{-2}^k)$, for which the linear regression predicts house values that are up to 30 times higher than the mean of Y. These predictions do not make sense. The lower panel shows a similar strong extrapolation risk for predictions at points of the type $(z_9 = 24, \mathbf{x}_{-9}^k)$, with values up to 40 times the mean of Y (right panel, Figure 29). Then, because ALE effects are given by differences of the type $\Psi_j^{ALE}(\mathbf{X}_{-j}^k; K) = \widehat{g}(z_9 = 24, \mathbf{x}_{-9}^k) - \widehat{g}(z_9 = 8, \mathbf{x}_{-9}^k)$, the very high values spuriously inflate the importance of this variable. The same occurs for X_2. We note the discrete nature of the support of both X_2 and X_9. In particular, X_9 has support $\mathcal{X}_9 = \{1, 2, \ldots, 8, 24\}$, which naturally defines a grid. Because X_9 is discrete, we cannot refine it, and we cannot control whether new points of the type $(z_9 = 24, \mathbf{x}_{-9}^k)$ fall "far"from the original data, causing unreliable predictions. Thus, we believe the discrete nature of these inputs has a role (of course not the sole one) in the extrapolation effects, as it makes the ALE design stategy less effective against extrapolation.

We then examine the results for the Random Forest.



stonHousing

Figure 30 Boston Housing: ALE importance indices for the Random Forest. Left panel: τ_j^{ALE}. Right panel: κ_j^{ALE}.

The left panel in Figure 30 shows that according to τ_j^{ALE} feature RM (X_6) is the most important, following by CRIM (X_1), DIS (X_8) and LSTAT (X_{13}). Instead, the right panel shows that κ_j^{ALE} ranks CRIM as the most important variable, with all others playing a minor role. We checked and the Random Forest predictions are not affected by extrapolation issues. Thus, the disagreement between the two indices cannot be attributed to unreliable predictions. The absence of a ground truth for the variable importance calls for employing additional variable importance measures to understand what factors drive the model response (we have reported the results for ν_j^{GCMR} and ν_j^{GCMR} in the main text).