Instructions for Paper Submissions to AISTATS 2023

Anonymous Author
Anonymous Institution

Abstract

The Abstract paragraph should be indented 0.25 inch (1.5 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The **Abstract** heading must be centered, bold, and in point size 12. Two line spaces precede the Abstract. The Abstract must be limited to one paragraph.

1 INTRODUCTION

Recently, ML has flourished in critical domains, such as healthcare and finance. In these areas, we need a combination of accurate predictions along with meaningful explanations to support them. For this reason there is an increased interest in Explainable AI (XAI), the field that provides interpretations about the behavior of complex blackbox models. XAI literature distinguishes between local and global explainability techniques (Molnar et al., 2020a). Local methods explain a specific prediction, whereas global methods explain the entire model behavior. Global methods provide a universal explanation, summarizing the various local explanations into a single interpretable outcome, usually a number or a plot. If a user wants to get a rough overview about which features are significant (feature importance) or whether a particular feature has a positive or negative effect on the output (feature effect), they should opt for a global explainability technique. On the other hand, aggregating the individual explanations for producing a concise global one is vulnerable to misinterpretations. Under strong interactions and correlations between features, the global explanation may obfuscate heterogeneous effects that exist under the hood (Herbinger et al., 2022); a phenomenon called aggregation bias (Mehrabi et al., 2021).

Feature effect (FE) (Grömping, 2020) is a fundamental category of global explainability methods. The objective of FE is to the isolate and visualize the impact of a single feature on the output. [1] FE methods suffer from aggregation bias because, often, the rationale behind the average effect might be unclear. For example, a feature with zero average effect may indicate that the feature has no effect on the output or, contrarily, it has a highly positive effect in some cases and a highly negative in others. There are three widely-used FE methods; Partial Dependence Plots (PDP)(Friedman, 2001), Marignal Plots (MP)(Apley and Zhu, 2020) and Aggregated Local Effects (ALE)(Apley and Zhu, 2020). PDP and MP have been criticized for computing erroneous effects when the input features are (highly) correlated, which is a frequent scenario in many ML problems. Therefore, ALE has been established as the state-of-the-art FE method.

However, ALE faces two crucial limitations, the first concerns ALE definition and the second one the ALE approximation. Regarding the definition, i.e., the formulation of the feature effect, ALE does not inform the user about the level of heterogeneous effects hidden behind the average effect (global explanation). In contrast, in the case of PDP, the heterogeneous effects can be (partially) spotted by exploring the Individual Conditional Expectations (ICE)(Goldstein et al., 2015). Regarding the approximation, e.g. the estimation of ALE from the limited samples of the training set, ALE requires an important additional step, called *bin-splitting*. Bin-splitting consists of partitioning the axis of the feature of interest in a sequence of non-overlapping intervals and estimating a single effect from the population of samples in each interval. Specifying an approriate sequence of intervals is of particular importance, since ALE's interpretation is meaningful only inside each interval (Molnar, 2022). However, this crucial step has not raised the appropriate attention, making the approximation vulnerable to potential misinterpretations.

In this paper, we extend ALE with a probabilistic component for measuring the uncertainty of the explanation. The uncertainty of the global explanation, i.e. how certain we are that the averaged explanation is valid if applied to an instance drawn at random, quantifies the level of heterogeneous effects. Given that ALE's interpretation (expected

Preliminary work. Under review by AISTATS 2023. Do not distribute.

[1] FE methods also isolate the effect of a pair of features to the output. Combinations of more than two features are not usual, because they encounter, among others, visualization difficulties.

effect and uncertainty) is meaningful only inside each interval, we design a principled framework, where we treat the bin-splitting step as a data-driven clustering problem, searching for the optimal splitting given available instances of the training set. We, also, present a computationally-grounded algorithm for finding the optimal solution.

Contributions. The contribution of this paper is NAME, a feature effect method that:

- Extends ALE with a probabilistic component to quantify the heterogeneous effects behind the global explanation.
- Automatically extracts regions with similar effects, improving the estimation and the interpretability of ALE plots.

We provide empirical evaluation of the method in artificial and real datasets. The implementation of our method and the code for reproducing all the experiments is provided in the submission and will become publicly available upon acceptance.

2 BACKGROUND AND RELATED WORK

At Section 2.1, we present the basic Feature Effect (FE) methods and we review the ALE definition(Apley and Zhu, 2020). At Section 2.2, we present the techniques for quantifying the heterogeneous effects and we discuss some of their limitations. Finally, at Section 2.3, we review the ALE approximation(Apley and Zhu, 2020)(add us), describing why it is vulnerable to a poor approximation due to inappropriate bin-splitting.

Notation. We refer to random variables (rv) using uppercase X, to simple variables with plain lowercase x and to vectors with bold $\mathbf{x}$. Often, we partition the input vector $\mathbf{x} \in \mathbb{R}^D$ to the feature of interest $x_s \in \mathbb{R}$ and the rest of the features $\mathbf{x}_c \in \mathbb{R}^{D-1}$. For convenience we denote it as $(x_s, \mathbf{x}_c)$, but we clarify that it implies the vector $(x_1, \cdots, x_s, \cdots, x_D)$. Equivalently, we denote the corresponding rv as $\mathbf{X} = (X_s, \mathbf{X}_c)$. The black-box function is $f : \mathbb{R}^D \to \mathbb{R}$ and the FE of the s-th feature is $f^{<\texttt{method}>}(x_s)$, where $<\texttt{method}>$ is the name of the FE method.[2]

2.1 Feature Effect Methods And ALE Definition.

The three well-known feature effect methods are: PDP, MP and ALE. PDP formulates the FE of the s-th attribute as an expectation over the marginal distribution $\mathbf{X}_c$, i.e., $f^{\texttt{PDP}}(x_s) = \mathbb{E}_{\mathbf{X}_c}[f(x_s, \mathbf{X}_c)]$, whereas MP formulates it as an expectation over the conditional $\mathbf{X}_c|X_s$, i.e., $f^{\texttt{MP}}(x_s) = \mathbb{E}_{\mathbf{X}_c|X_s=x_s}[f(x_s, \mathbf{X}_c)]$. ALE defines the local effect of the s-th feature at point $x_s = z$ as $f^s(z, \mathbf{x}_c) = f^s(z, \mathbf{x}_c)$. All the local explanations at z are, then, weighted by the conditional distribution $p(\mathbf{x}_c|x_s = z)$ and are averaged, to produce the averaged effect $\mu(z)$. Finally, ALE is the accumulation of the averaged local effects:

$$f^{\texttt{ALE}}(x_s) = \int_{x_{s,min}}^{x_s} \underbrace{\mathbb{E}_{\mathbf{X}_c|X_s=z}\left[f^s(z, \mathbf{X}_c)\right]}_{\mu(z)} \partial z \qquad (1)$$

ALE has specific advantages which gain particular value in cases of correlated input features. In these cases, PDP integrates over unrealistic instances due to the use of the marginal distribution $\mathbf{X}_c$, and MP computes aggregated effects, i.e., imputes the combined effect of sets of features to a single feature. ALE manages to resolve both issues, and is therefore the most trustable method in cases of correlated features.

2.2 Quantification Of Heterogeneous Effects.

FE methods pose the question *what is expected to happen to the output (expected effect), if the value of a specific feature is increased/decreased*. Having quantified the expected effect, it comes naturally to also ask *how certain we are about the expected change (uncertainty)*. For this reason, the quantification of the uncertainty, whether there are local explanations that deviate from the expected global effect, has attracted a lot of interest.

ICE and d-ICE(Goldstein et al., 2015) provide a set of curves that are produce a plot plot on top-of PDP. Both methods produce one curve for each instance of the training set, i.e. ICE the instance effect $f_i^{\texttt{ICE}}(x_s) = f(x_s, \mathbf{x}_c^i)$ and d-ICE the derivative effect, i.e. $f_i^{\texttt{d-ICE}}(x_s) = \frac{\partial f}{\partial x_s}(x_s, \mathbf{x}_c^i)$. The user can visually examine the level of homogeneity of the set of curves as an indicator of the level of heterogeneous effects. Some methods try to automate the aforementioned visual exploration, by grouping (d-)ICE into clusters (Molnar et al., 2020b; Herbinger et al., 2022). Some other approaches, like H-Statistic, Greenweal, quantify the level of interaction between the input features. A strong interaction index is an indirect indicator for the existence of heterogeneous effects.

The aforementioned approaches are under two limitations; They either do not quantify the uncertainty of the FE directly or they are based on PDPs, and, therefore, they are subject to the failure modes of PDPs in cases of correlated features, as we will analyze later. To the best of our knowledge, no method so far quantifies the heterogeneous effects on-top of ALE.

[2] An extensive list of all symbols used in the paper is provided in the helping material.

2.3 ALE Approximation.

In real ML scenarios, the FE is estimated from the limited instances of the training set. Therefore, (Apley and Zhu, 2020) proposed dividing the s-th axis in K bins and estimating the local effects in each bin by evaluating the black box-function at the bin limits:

$$\hat{f}^{\texttt{ALE}}(x_s) = \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[f(z_k, \mathbf{x}_c^i) - f(z_{k-1}, \mathbf{x}_c^i) \right] \tag{2}$$

We denote as k_x the index of the bin that x_s belongs to, i.e. $k_x : z_{k_x-1} \leq x_s < z_{k_x}$ and $\mathcal{S}_k$ is the set of training instance that lie in the k-th bin, i.e. $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x_s^i < z_k\}$. Afterwards, (cite) proposed the Differential ALE (DALE) approximation for computing the local effects on the training instances using auto-differentiation:

$$\hat{f}^{\texttt{DALE}}(x_s) = \Delta x \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \frac{\partial f}{\partial x_s}(\mathbf{x}^i) \tag{3}$$

Their method has the advantages of remaining on-distribution even when bins become wider and, most importantly, allows the recomputation of the accumulated effect with different bin-splitting with near-zero computational cost.

Both approximations ask from the user to blindly decide the number of bins K for splitting the axis into K equally-sized bins, an approach with crucial limitations. Firstly, it is vulnerable to poor bin-effect approximations that lead to a misleading ALE plot. On the one hand, setting a K to a small value may hide fine-grain effects due to large bins and setting K to a high value leads to estimations of bin-effect from very limited samples. In general, the user may face contradictory explanations for differnt K without a clue to decide which one to trust. Secondly, as indicated by (Molnar, 2022), in ALE the interpretation of the effect can only be local. Preseting a fixed-size intervals, does not permit the extraction of local regions (intervals) ng of data-driven local regions

3 THE NAME METHOD

The NAME method extends ALE for quantifying the level of heterogeneous effects (uncertainty) and automating the bin-splitting step for robust estimations. At Section 3.1, we define NAME, presenting the component that extends ALE with uncertainty quantification. In Section 3.2, we provide an alternative definition of NAME based on variable-size bins, providing important remarks and proofs for connecting it with the initial definition. In Section 3.3, we define the problem of optimal bin-splitting, i.e. how to optimally approximate METHOD based on limited samples, and we provide an algorithmic solution to the problem. Finally, in Section 3.4, we illustrate the appropriate visualization of NAME for facilitating its interpretation by a non-expert and we discuss important aspects of the method.

3.1 NAME: ALE With Uncertainty Quantification

ALE does not shed light to potential heterogeneous effects that contribute to the average explanation. Therefore, we extend ALE definition of Eq. (1) with a component that quantifies the level of heterogeneous effects. We denote $\mathcal{H}(z)$ the uncertainty of the local effects at a specific point $x_s = z$ and we quantify it as the standard deviation of the local explanations:

$$\mathcal{H}(z) := \sigma(z) = \sqrt{\mathbb{E}_{\mathbf{X}_c|z} \left[\left(f^s(z, X_c) - \mu(z) \right)^2 \right]} \tag{4}$$

The uncertainty emerges from the natural characteristics of the experiment, i.e., the feature correlations existent in the data generating distribution and the implicit interactions of the black-box function. We also define the accumulated uncertainty at x_s, as the accumulation of the standard deviation of the local effects along the axis:

$$f_\sigma^{\texttt{ALE}}(x_s) = \int_{x_{s,min}}^{x_s} \sigma(z) \partial z \tag{5}$$

NAME method formulates the effect at a specific point x_s with a compact entity that consists of the average effect and the uncertainty $\mu(z) \pm \sigma(z)$ and visualizes them as a continuous curve with a confidence region, i.e. $f^{\texttt{METHOD}}(x_s) := f_\mu^{\texttt{ALE}}(x_s) \pm f_\sigma^{\texttt{ALE}}(x_s)$. In Section 3.4, we propose an appropriate visualization for easier interpretation of NAME.

3.2 Interval-Based Formulation of NAME

In real scenarios, the estimations are based on the limited instances of the training set. Estimating $\mu(z), \sigma(z)$ at the granularity of a point is impossible, because the probability of observing a sample inside the interval $[x_s - h, x_s + h]$ tends to zero, when $h \to 0$. For this reason, we firstly define the mean effect (bin-effect) and the uncertainty (bin-uncertainty) inside an interval $[z_1, z_2)$:

$$\mu(z_1, z_2) = \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x}_c|z} \left[f^s(z, \mathbf{x}_c) \right] \partial z \tag{6}$$

$$\sigma(z_1, z_2) = \frac{\int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x}_c|x_s=z} \left[\left(\frac{\partial f}{\partial x_s} - \mu(z_1, z_2) \right)^2 \right] \partial z}{z_2 - z_1} \tag{7}$$

We also define as $\mathcal{Z}_s$ a sequence of $K+1$ points that partitions the s-th feature axis into a sequence K variable-size intervals, i.e. $\mathcal{Z}_s = \{z_0, z_K\}$, where $z_0 = x_{s,min}$ and $z_K = x_{s,max}$. Based on the above, we redefine NAME using an interval-based formulation $\tilde{f}^{\texttt{METHOD}}(x_s) := \tilde{f}_\mu^{\texttt{ALE}}(x_s) \pm \tilde{f}_\sigma^{\texttt{ALE}}(x_s)$, where:

$$\tilde{f}_{\mu}^{\texttt{ALE}}(x_s) = \sum_{k=1}^{k_x} \mu(z_{k-1}, z_k)(z_k - z_{k-1}) \tag{8}$$

$$\tilde{f}_{\sigma}^{\texttt{ALE}}(x_s) = \sum_{k=1}^{k_x} \sigma(z_{k-1}, z_k)(z_k - z_{k-1}) \tag{9}$$

Theorem 1. If we define the residual $\rho(z)$ as the difference between the expected effect at x_s and the mean expected effect at the interval, i.e $\rho(z) = \mu(z) - \mu(z_1, z_2)$, then, the accumulated variance $\sigma^2(z_1, z_2)$ equals to the accumulation of the point-wise variances plus the squared residuals inside the interval:

$$\sigma^2(z_1, z_2) = \int_{z_1}^{z_2} \sigma^2(z) + \rho^2(z)\partial z \tag{10}$$

The proof is in the Appendix.

Theorem 2. The two definitions of NAME are equivalent, i.e., $f_{\mu}^{\texttt{ALE}}(x_s) = \tilde{f}_{\mu}^{\texttt{ALE}}(x_s)$ and $f_{\sigma}^{\texttt{ALE}}(x_s) = \tilde{f}_{\sigma}^{\texttt{ALE}}(x_s)$, if and only if the sequence of limits $\mathcal{Z}$ is such that $\sum_{k=1}^{K} \rho^2(z_{k-1}, z_k) = 0$, where $\rho^2(z_{k-1}, z_k) = \int_{z_{k-1}}^{z_k} \rho^2(z)\partial z$.

Discussion Theorem 1 decouples the aggregated uncertainty inside a bin, i.e., $\mathcal{H}_{bin}(z_1, z_2) := \sigma^2(z_1, z_2)$, into two terms. The first term quantifies the accumulated uncertainty due to the natural characteristics of the experiment, i.e., $\mathcal{H}(z_1, z_2) = \int_{z_1}^{z_2} \mathcal{H}(z)\partial z$, and the second term adds extra nuisance uncertainty due to limiting the resolution $\mathcal{H}_n(z_1, z_2) := \int_{z_1}^{z_2} \rho^2(z)\partial z$. Therefore, it holds that:

$$\mathcal{H}_{bin}(z_1, z_2) = \mathcal{H}(z_1, z_2) + \mathcal{H}_n(z_1, z_2) \tag{11}$$

In Theorem 2 we prove that if we find a sequence of bins, where each bin adds no nuisance uncertainty, then we will get exactly the same NAME explanation as in the original definition.

3.3 Bin-Splitting: Finding Regions With Homogeneous Effects

In this section we formulate bin-splitting as an unsupervised clustering problem. The mean effect (Eq. (6)) and the uncertainty (Eq. (7)) can be directly estimated from the set $\mathcal{S}$ of the dataset instances with the s-th feature lying inside the interval $\mathcal{S} = \{\mathbf{x}^i : z_1 \leq x_s^i < z_2\}$, i.e. $\hat{\mu}(z_1, z_2) = \frac{1}{|\mathcal{S}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}} \left[f^s(\mathbf{x}^i)\right]$ and $\hat{\sigma}^2(z_1, z_2) = \frac{1}{|\mathcal{S}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}} \left(f^s(\mathbf{x}^i) - \hat{\mu}(z_1, z_2)\right)^2$. Both approximations are unbiased only under the assumption that the points are uniformly distributed in $[z_1, z_2]$. Under this assumption, and due to the law of large numbers, the approximations become more robust, as the bins become wider. On the other hand, bins should group local explanations with similar effects and enlarging a bin is under the danger of adding nuisance uncertainty. Therefore, we set-up the following optimization problem:

$$\begin{aligned} \min_{\mathcal{Z}=\{z_0,\ldots,z_K\}} \quad & \mathcal{L} = \sum_{k=1}^{K} \hat{\sigma}^2(z_{k-1}, z_k)\Delta z_k \\ \text{where} \quad & \Delta z_k = z_k - z_{k-1} \\ & \tau_k = 1 - 0.2\frac{|S_k|}{N} \\ \text{s.t.} \quad & |\mathcal{S}_k| \geq N_{\texttt{NPB}} \\ & z_0 = x_{s,min} \\ & z_K = x_{s,max} \end{aligned} \tag{12}$$

We search for the sequence of intervals $z_0, \ldots, z_K$ that minimizes the sum of the cost of each bin. The cost of each bin is the mean variance σ_k^2 scaled by the bin length Δz_k and discounted by the term τ_k. The term $0.8 \leq \tau_K \leq 1$ acts as a tie-braker in cases of two candidate bins with equal mean variance, favoring the selection of the bigger bin. The constraint of at least $N_{\texttt{PPB}}$ points per bin sets the lowest-limit for a *robust* estimation.

According to Eq. (11), the uncertainty of a bin $\mathcal{H}_{bin}$ is the sum of the uncertainty due to the characteristics of the experiment (data and black-box model) $\mathcal{H}$ and the nuisance uncertainty $\mathcal{H}_n$. Since $\mathcal{H}$ is independent of the bin-spliting procedure, by minimizing $\mathcal{L}$ we minimize the nuiscane uncertainty $\mathcal{H}_n$ in each bin.

3.3.1 Solve Bin-Splitting with Dynamic Programming

For achieving a computationally-grounded solution we set a threshold K_{max} on the maximum number of bins which also discretizes the solution space. The width of the bin can take discrete values that are multiple of the minimum step $u = \frac{x_{s,max} - x_{s,min}}{K}$. For defining the solution, we use two indexes. The index $i \in \{0, \ldots, K_{max}\}$ denotes the i-th limit (z_i) and the index $j \in \{0, \ldots, K_{max}\}$ denotes the j-th position along the axis ($x_j = x_{s,min} + j \cdot u$). The recursive cost function $T(i, j)$ is the cost of setting the i-th limit z_i to x_j:

$$\mathcal{T}(i, j) = \min_{l \in \{0,\ldots,K_{max}\}} \left[\mathcal{T}(i-1, l) + \mathcal{B}(x_l, x_j)\right] \tag{13}$$

where $\mathcal{T}(0, j)$ equals zero if $j = 0$ and ∞ in any other case. $\mathcal{B}(x_l, x_j)$ denotes the cost of creating a bin with limits $[x_l, x_j)]$:

$$\mathcal{B}(x_l, x_j) = \begin{cases} \infty, & \text{if } x_j > x_l \text{ or } |\mathcal{S}_{(x_j, x_l)}| < N \\ 0, & \text{if } x_j = x_l \\ \hat{\sigma}^2(x_j, x_l), & \text{if } x_j \leq x_l \end{cases} \tag{14}$$

The optimal solution is given by solving $\mathcal{L} = \mathcal{T}(K_{max}, K_{max})$ and keeping track of the sequence of steps.

- Computational complexity

3.4 Visualization and Interpretation

- Discuss about the meaning of ALE, to find intervals with some effect

4 SIMULATION EXAMPLES

The simulation examples, where the data-generating distribution $p(\mathbf{X})$ and the predictive function $f(\cdot)$ are defined by us, enable the evaluation of competitive approaches against a solid ground-truth. We follow this common XAI practice (Aas et al., 2021; Herbinger et al., 2022) for providing secure empirical evidence about the superiority of NAME method.

We split the simulation examples in two groups. The first group, Section 4.1, aims at showcasing that NAME method is more accurate than PDP-ICE in quantifying the average effect and the level of heterogeneous effects (uncertainty), when input features are correlated. The second group, Section 4.2, illustrates that NAME method achieves a better approximation (average effect and uncertainty) in cases of limited samples, due to automatic bin-splitting. In both groups, we choose to illustrate the most indicative examples; a more extensive evaluation is provided in the Appendix.

4.1 Case 1: Uncertainty Quantification

In this simulation, we will compare NAME method against PDP-ICE in quantifying the main effect f_μ and the uncertainty f_σ^2. Since there is ambiguity about the ground-truth first-order effect in cases of correlated features, e.g. (Apley and Zhu, 2020; Grömping, 2020), we limit ourselves to the following piecewise linear function,

$$f(\mathbf{x}) = \begin{cases} f_{\texttt{lin}} + \alpha f_{\texttt{int}} & \text{if } f_{\texttt{lin}} < 0.5 \\ 0.5 - f_{\texttt{lin}} + \alpha f_{\texttt{int}} & \text{if } 0.5 \leq f_{\texttt{lin}} < 1 \\ \alpha f_{\texttt{int}} & \text{otherwise} \end{cases} \quad (15)$$

where $f_{\texttt{lin}}(\mathbf{x}) = a_1 x_1 + a_2 x_2$ and $f_{int}(\mathbf{x}) = x_1 x_3$. As we observe, the linear term $f_{\texttt{lin}}$ includes the two correlated features and the term $f_{\texttt{int}}$ interacts the two non-correlated variables. The samples that we use in all examples are coming from the following distribution: $p(\mathbf{x}) = p(x_3)p(x_2|x_1)p(x_1)$ where $x_1 \sim \mathcal{U}(0,1)$, $x_2 = x_1$ and $x_3 \sim \mathcal{N}(0, \sigma_3^2)$. We will test the effect computed by NAME and PDP-ICE in three cases; (a) no interaction ($\alpha = 0$) and equal weights ($a_1 = a_2$), (b) no interaction ($\alpha = 0$) and different weigths ($a_1 \neq a_2$) and (c) with interaction ($\alpha \neq 0$) with equal weights ($a_1 = a_2$).

In all cases, we firstly compute the ground-truth average effect and the uncertainty analytically (proofs in the Appendix) and then we compare it against the approximation provided by each method. The approximation is computed after generating $N = 300$ samples. As we will see, despite the model's simplicity, PDP-ICE fail in quantifying the effect correctly, due to correlation between features X_1 and X_2. Finally, besides it is not the focus of the current examples, we will also observe that NAME manages to correctly extract the regions with constant effect. Details for obtaining the ground truth effects and the experimental set-up are in the Appendix.

No Interaction, Equal weights. We test the feature effect when no interaction term is apparent, i.e. $\alpha = 0$, and the weights are equal $a_1 = a_2 = 1$. In this case, the ground truth effect $f_\mu^{\texttt{GT}}(x_1)$ is: x_1 when $0 \leq x_1 < 0.25$, $-x_1$ when $0.25 \leq x_1 < 0.5$ and zero otherwise. In each position, there is total absence of heterogeneous effects; therefore, the uncertainty of the global explanation is zero $f_{\sigma^2}^{\mathcal{GT}}(x_1) = 0$. In Figure **??**, we observe that PDP main effect is wrong and ICE plots imply the existence of heterogeneous effects. In contrast, ALE captures correctly the average effect and the zero uncertainty and it also groups perfectly the regions with constant-effect.

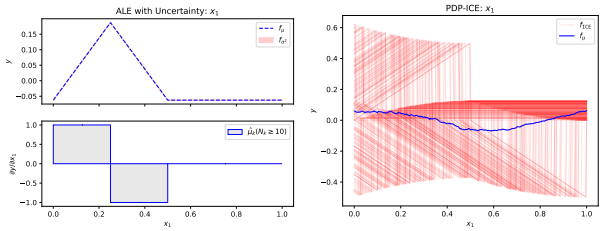

Figure 1: No interaction, Equal weights: Feature effect for x_1; Ground-truth vs (a) NAME method at the left and (b) PDP-ICE at the right.

No Interaction, Different weights. As before, there is no interaction term and, therefore, the ground-truth is zero, i.e. $f_{\sigma^2}^{\mathcal{GT}}(x_1) = 0$. The weigts are different $a_1 = 2, a_2 = 0.5$, therefore, the ground-truth effect is $f_\mu^{\texttt{GT}}(x_1)$ is: $2x_1$ when $0 \leq x_1 < 0.2$, $-2x_1$ when $0.2 \leq x_1 < 0.4$ and zero otherwise. In Figure 2, we observe that PDP estimation is completely opposite to the ground-truth effect, i.e. negative in $[0, 0.2)$ and positive in $[0.2, 0.4)$, and the ICE erroneously implies heterogeneous effects. As before, ALE quantifies perfectly the ground truth effect and the zero-uncertainty, extracting correclty the constant effect regions.

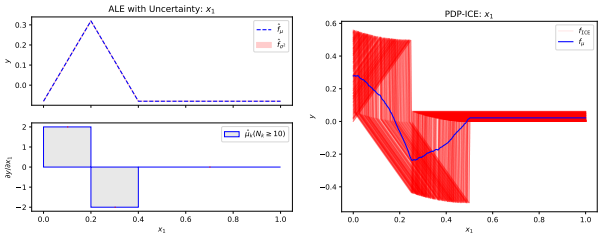

Figure 2: No interaction, Different weights: Feature effect for x_1; Ground-truth vs (a) NAME method at the left and (b) PDP-ICE at the right.

Uncertainty, Equal weights. In this case we activate the interaction term, i.e. $a = 1$, and we set equal weights $a_1 = a_2 = 1$. The interaction term provokes heterogeneous effects in features x_1, x_3, i.e., at any position x_1, the local effects dependend on the unknown value of X_3 and vice-versa. The effect of x_2 continues to have zero-uncertainty, since it does not appear in any interaction term. As we observe in Figure 3, NAME captures perfectly the effect of all features, and the uncertainty in all cases. As expected, PDP-ICE quantify the effect and the uncertainty correctly only in the case of x_3, since X_3 is independent from other features. For the correlated features, i.e. x_1, x_2, the average effect computed by PDP and the uncertainty implied by ICE plots are wrong.

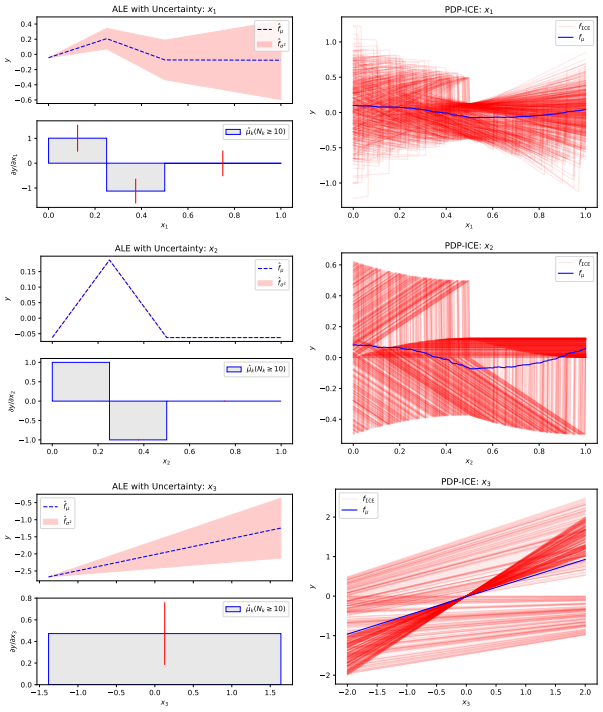

Figure 3: With interaction, equal weights: Feature effect for all features, x_1 to x_3 from top to bottom; Ground-truth vs (a) NAME method at the left columns and (b) PDP-ICE at the right column.

Discussion. In purpose, the examples above do not cover the case where there is an interaction term includes correlated features, i.e. $x_1 x_2$. In this scenario, there is an open debate about the ground-truth effect (Grömping, 2020).

4.2 Case 2: Bin-Splitting

In this simulation, we aim to to quantify the advantages of automatic bin-splitting based on the objective of Eq.(12), through the following validation framework. We generate a big dataset with dense sampling ($N = 10^6$) and we treat the DALE estimation with dense fixed-size bins ($K = 10^3$) as ground-truth. Afterwards, we generate less samples ($N = 500$) and we compare (a) the fixed-size DALE estimation for many different K versus (b) the auto-bin splitting algorithm. In all cases, we use a bivariate black-box function $f(\cdot)$ where the samples are instances of the distribution $p(\mathbf{x}) = p(x_2|x_1)p(x_1)$ where $x_1 \sim \mathcal{U}(0, 1)$ and $x_2 \sim \mathcal{N}(x_1, \sigma_2^2 = 0.5)$.

We denote as $\mathcal{Z}^{\texttt{DP}} = \{z_{k-1}^{\texttt{DP}}, \cdots, z_K^{\texttt{DP}}\}$ the sequence obtained by automatic bin-splitting based on the optimisation problem of Eq. (12) and with $\mathcal{Z}^{\texttt{K}}$ the fixed-size splitting with K bins. The evaluation is done with regard to two metrics; $\mathcal{L}^{\mu}_{\texttt{DP}|\texttt{K}} = \frac{1}{|\mathcal{Z}^{\texttt{DP}|\texttt{K}}|} \sum_{k \in \mathcal{Z}^{\texttt{DP}|\texttt{K}}} |\mu_k - \hat{\mu}_k|$ quantifies the average error in estimating the the expected effect and $\mathcal{L}^{\sigma^2}_{\texttt{DP}|\texttt{K}} = \frac{1}{|\mathcal{Z}^{\texttt{DP}|\texttt{K}}|} \sum_{k \in \mathcal{Z}^{\texttt{DP}|\texttt{K}}} |\sigma_k^2 - \hat{\sigma}_k^2|$ the average error in estimating the variance of the effect. The metric $\mathcal{L}^{\rho^2}_{\texttt{DP}|\texttt{K}} = \frac{1}{|\mathcal{Z}^{\texttt{DP}|\texttt{K}}|} \sum_{k \in \mathcal{Z}^{\texttt{DP}|\texttt{K}}} \rho_k^2$ quantifies the average nuisance uncertainty. The optimal bin-splitting is the one that minimises, at the same time, both $\mathcal{L}^{\mu}$ and $\mathcal{L}^{\sigma^2}$ errors. For consistent results, in all the examples below, we regenerate samples and repeat the computations for $t = 10$ times, providing the mean value for all metrics.

Piecewise-Linear Function. In this example, we define $f(\mathbf{x}) = a_1 x_1 + x_1 x_2$ with 5 piecewise-linear regions of different-size, i.e., a_1 equals to $= \{2, -2, 5, -10, 0.5\}$ in the intervals defined by the sequence $\{0, 0.2, 0.4, 0.45, 0.5, 1\}$. As we observe, in Figure 4, NAME method extracts a sequence of intervals with better $\mathcal{L}^{\mu}_{\texttt{DP}}$ and $\mathcal{L}^{\sigma^2}_{\texttt{DP}}$ error compared to any fixed-size splitting. Analyzing the fixed-size errors helps us understand the importance of variable-size splitting. In Figure 4(b), we observe a positive trend between $\mathcal{L}^{\mu}_{\texttt{K}}$ and K, concluding that bin effect estimation is more incosistent as Δx becomes smaller, due to less points contributing to each bin. The interpretation of variance error is slightly more complex. Given that the smallest interval is $\Delta x = 0.05 \Rightarrow K = 20$ and all intervals are multiples of the smallest interval, any K that is not a multiple of 20 adds nuisance uncertainty $\mathcal{L}^{\rho^2}_{\texttt{K}}$ leading to a high variance error $\mathcal{L}^{\sigma^2}_{\texttt{K}}$. In these cases, the variance error reduces as K grows bigger because the length of the bins that lie in the limit between two piecewise linear regions becomes smaller. For

$K = \{20, 40, 60, 80, 100\}$ where $\mathcal{L}_{\mathrm{K}}^{\rho^2} \approx 0$, we conclude the same as with with the mean effect error, i.e. the estimation becomes more incosistent as K grows larger.

Variable-size extracts correctly the fine-grain bins, e.g., intervals $[0.4, 0.45]$, $[0.45, 0.5]$, and acts as a merging mechanism to create wider bins when the effect remains constant, e.g. interval $[0.5, 1]$, leading to an optimal solution.

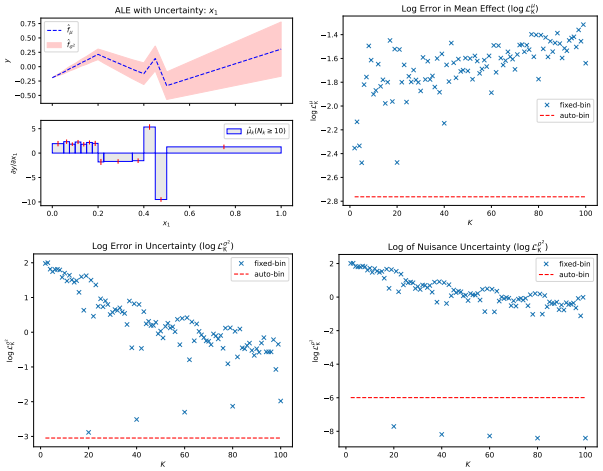

Figure 4: Figure 1

Non-Linear Function. In this example, we define a black-box function $f(\mathbf{x}) = 4x_1^2 + x_2^2 + x_1x_2$, where the effect is non-linear in all the range of x. This case has two specialties. First, there is no obvious advantage of variable-size versus fixed-size splitting, and, second, there is not an apriori optimal bin-size. Widening a bin will increase the resolution error $\mathcal{L}^{\rho^2}$ and narrowing will make less robust. In Figure 5, we observe that automatic bin splitting finds a solution that compromises the conflicting objectives, i.e., it keeps as low as possible both the main effect $\mathcal{L}_{\mathrm{K}}^{\mu}$ and the variance error $\mathcal{L}_{\mathrm{K}}^{\sigma^2}$.

Discussion.

5 REAL-WORLD EXAMPLES

we also apply the methods on some real-world datasets. In these cases, it is infeasible to obtain ground truth information, therefore, we perform evaluation at two levels. First, we compare visually NAME versus PDP with ICE plots. Second, we hold as ground-truth the effects computed on the full training-set and we, then, perform subsampling to evaluate the robustness of the bin-splitting method.

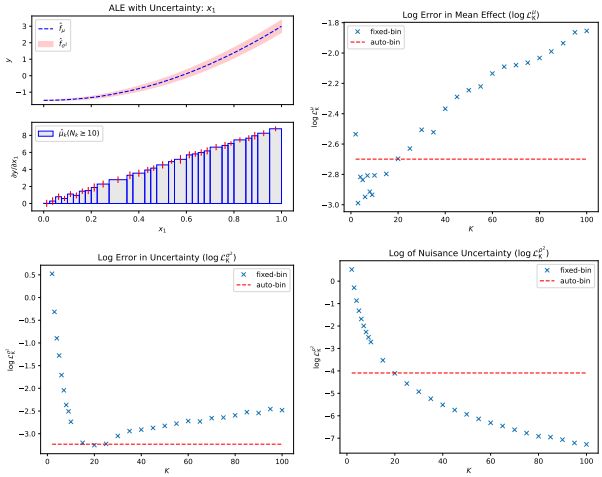

Figure 5: Figure 1

6 CONCLUSION

Acknowledgments

All acknowledgments go at the end of the paper, including thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support. To preserve the anonymity, please include acknowledgments *only* in the camera-ready papers.

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Ulrike Grömping. Model-agnostic effects plots for interpreting machine learning models, 03 2020.

Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. Repid: Regional effect plots with implicit interaction detection. In *International Conference on Artificial Intelligence and Statistics*, pages 10209–10233. PMLR, 2022.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL `https://christophm.github.io/interpretable-ml-book`.

Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer, 2020a.

Christoph Molnar, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio. Model-agnostic feature importance and effects with dependent features–a conditional subgroup approach. *arXiv preprint arXiv:2006.04628*, 2020b.