Uncertainty-aware Accumulated Local Effects (UALE) for quantifying the heterogeneous effects

Anonymous Author
Anonymous Institution

Abstract

Accumulated Local Effects (ALE) is a popular approach to explainable AI that describes how a feature influences the decisions of a model on average, taking into account potential correlations between the features. However, ALE does not provide an explicit component for modeling the uncertainty of the effect, i.e., the level of heterogeneity of the effects behind the average explanation, which is crucial for a complete interpretation when the black-box function contains interactions between the correlated features. In this work, we define Uncertainty-aware ALE (UALE), an extension of ALE for quantifying the heterogeneous effects and we propose a method for estimating it. We formally prove the conditions the bin-splitting, i.e. the partitioning the domain of the feature of interest into consecutive non-overlapping intervals, must respect for an an unbiased estimation of the uncertainty. We finally propose a computationally-grounded algorithm for finding the optimal sequence of bins given the limited instances of the training set. We demonstrate through synthetic and real datasets, that UALE quantifies the average and heterogeneous effects correctly and approximates them robustly by optimizing the bin-splitting procedure.

1 INTRODUCTION

Recently, Machine Learning (ML) has been adopted 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 black-box models. XAI literature distinguishes between local and global explainability techniques (Molnar et al., 2020a). Local methods explain the prediction of a specific instance, whereas global methods provide an explanation that summarizes the entire model behavior. Global methods provide a universal explanation, aggregating the various local explanations into a single interpretable outcome, usually a number or a plot. If users want 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 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 isolate the impact of a single feature on the output. [1] FE methods suffer from aggregation bias because, often, implications of the average effect to the model behavior are unclear. For example, a feature with zero average effect may indicate no effect on the output or, in contrast, highly positive in some cases and 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 cases. 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 ALE approximation. Regarding the definition ALE does not inform the user about the uncer-

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

[1] Often, FE methods also isolate the combined effect of a pair of features to the output. Combinations of more than two features are uncommon, since they are difficult to estimate and visualize.

tainty, i.e., the level of heterogeneous effects hidden behind the average effect due to implicit feature interactions. In contrast, in PDPs the heterogeneous effects can be identified by exploring the Individual Conditional Expectations (ICE)(Goldstein et al., 2015). Regarding the approximation, i.e. the estimation using the (limited) samples of the training set, ALE requires the additional *bin-splitting* step. Bin-splitting consists of partitioning the domain of the feature of interest into consecutive non-overlapping intervals (regions) and estimating a constant effect in each region. The effect is estimated from the population of samples that fall inside the region. Specifying an appropriate set of bins is of particular importance, since ALE's interpretation is meaningful only inside each region (Molnar, 2022). However, this crucial step has not raised the appropriate attention and the approximation is vulnerable to potential misinterpretations.

In this paper, we present Uncertainty-aware Accumulated Local Effects (UALE) an extension of ALE for measuring the uncertainty of the explanation, i.e. how certain we are that the average explanation would predict the feature effect to an instance drawn at random. We show that choosing an appropriate sequence of intervals is particularly important for a robust and unbiased estimation of the uncertainty and we provide an algorithm that automates this step. We also formally prove the bin-splitting conditions for an unbiased estimation of the uncertainty. We apply our method in multiple synthetic and real world datasets for evaluating its performance.

Contributions. The contributions of this paper are:

- UALE, a FE method that extends ALE for quantifying the heterogeneous effects (uncertainty).
- A formal proof of the bin-splitting requirements for an unbiased approximation of the heterogeneous effects.
- A framework for automatically extracting 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

In this section, we describe the basic methods for FE and uncertainty quantification, focusing on ALE definition. We, then, review the ALE approximation(Apley and Zhu, 2020; Gkolemis et al.), describing some of its vulnerabilities.

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 $\mathbf{X} = (X_s, \mathbf{X}_c)$. When we refer to an instance of the training set, we use $\mathbf{x}^i = (\mathbf{x}_c^i, x_s^i)$. The black-box function is $f : \mathbb{R}^D \to \mathbb{R}$ and the FE of the s-th feature is $f^{<\texttt{method}>}(x_s)$, with $<\texttt{method}>$ indicating the particular method in use.[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., $f^{\texttt{MP}}(x_s) = \mathbb{E}_{\mathbf{X}_c|x_s}[f(x_s, \mathbf{X}_c)]$. ALE defines the local effect of the s-th feature at point $x_s = z$ as $f^s(z, \mathbf{x}_c) = \frac{\partial f}{\partial x_s}(z, \mathbf{x}_c)$. 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 $\mu(z)$. ALE is the accumulation of the averaged local effects:

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

Eq. (1) 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 reply to the question *what is the average (global) effect on the output, if the value of a specific feature is increased/decreased*. It comes naturally to also ask *to what extent the local effects deviate from the global explanation*. Quantifying the uncertainty of the global explanation has attracted a lot of interest. ICE and d-ICE(Goldstein et al., 2015) provide a set of curves that are plot on top-of the PDP. Both methods produce one curve for each instance of the training set; $f_i^{\texttt{ICE}}(x_s) = f(x_s, \mathbf{x}_c^i)$ for ICE and $f_i^{\texttt{d-ICE}}(x_s) = \frac{\partial f}{\partial x_s}(x_s, \mathbf{x}_c^i)$ for d-ICE. The user then visually observes if the curves are homogeneous and

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

to what extent they deviate from the PDP. Some methods try to automate the aforementioned visual exploration, by grouping (d-)ICE plots into clusters (Molnar et al., 2020b; Herbinger et al., 2022; Britton, 2019). Some other approaches, like H-Statistic(Friedman and Popescu, 2008), Greenwell's interaction index(Greenwell et al., 2018) or SHAP interaction values(Lundberg et al., 2018), quantify the level of interaction between the input features, with an interaction value. A strong interaction index is an 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(Baniecki et al., 2021), as we will show in-depth in Section 4.1. To the best of our knowledge, no work exist so far that quantifies the heterogeneous effects based on the formulation of ALE.

2.3 ALE Approximation.

In ML scenarios, ALE is estimated from the training set examples. Therefore, (Apley and Zhu, 2020) proposed dividing the feature domain in K equal-sized 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, (Gkolemis et al.) proposed the Differential ALE (DALE) that computes the local effects on the training instances using auto-differentiation:

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

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, denoted with K, for splitting the axis into equally-sized bins. This approach introduces significant limitations. Specifically, setting K to a small value may hide fine-grain effects due to large bins while setting K to a high value leads to poor bin-effect estimations from limited samples. In general, the user may face contradictory explanations for different K without a clue for which one to trust.

3 Uncertainty-Aware ALE (UALE)

UALE extends ALE for quantifying the level of heterogeneous effects (uncertainty) and automates the bin-splitting. At Section 3.1 we define UALE, at Section 3.2 we redefine it using an interval-based formulation, at Section 3.3 we provide an approximation and at Section 3.4 we define and solve the problem of optimal bin-splitting.

3.1 UALE: ALE With Uncertainty Quantification

We extend ALE definition of Eq. (1) with a component for quantifying the uncertainty. We denote as $\mathcal{H}(z)$ 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 $\mathcal{H}(z) = \sigma(z)$, where:

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

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:

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

UALE formulates the effect at a specific point x_s with a compact tuple that consists of the average effect and the uncertainty $(\mu(z), \sigma(z))$ and visualizes them as a continuous curve with a confidence region, i.e. $f^{\texttt{UALE}}(x_s) := f_{\mu}^{\texttt{ALE}}(x_s) \pm f_{\sigma}^{\texttt{ALE}}(x_s)$.

3.2 Interval-Based Formulation and Approximation

In real scenarios, all estimations are based on the instances of the training set. Therefore, we cannot estimate $\mu(z), \sigma(z)$ at the granularity of a point, because the probability of observing a sample inside the region $[x_s - h, x_s + h]$ tends to zero, when $h \to 0$. We define the regional-effect $\mu(z_1, z_2)$ and the regional-uncertainty $\mathcal{H}(z_1, z_2) = \sigma(z_1, z_2)$, as the summary statistics that characterize the effect and the uncertainty inside the interval $[z_1, z_2)$:

$$\mu(z_1, z_2) = \mathbb{E}_{z \sim \mathcal{U}(z_1, z_2)}[\mu(z)] = \frac{\int_{z_1}^{z_2} \mu(z) \partial z}{z_2 - z_1} \tag{6}$$

$$\sigma^2(z_1, z_2) = \mathbb{E}_{z \sim \mathcal{U}(z_1, z_2)}[\sigma^2(z)] = \frac{\int_{z_1}^{z_2} \sigma^2(z) \partial z}{z_2 - z_1} \tag{7}$$

Intuitively, the regional-effect and the regional-uncertainty quantify the expected average effect and the expected uncertainty if we randomly draw a point z^* given a uniform

distribution $z^* \sim \mathcal{U}(z_1, z_2)$. If we also define as $\mathcal{Z}$ the sequence of $K+1$ points that partition the domain of the s-th feature into K variable-size intervals, i.e. $\mathcal{Z} = \{z_0, \ldots, z_K\}$, we can redefine UALE using an interval-based formulation $\tilde{f}^{\texttt{UALE}}_{\mathcal{Z}}(x_s) := \tilde{f}^{\texttt{ALE}}_{\mathcal{Z},\mu}(x_s) \pm \tilde{f}^{\texttt{ALE}}_{\mathcal{Z},\sigma}(x_s)$, where:

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

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

3.3 Interval-Based Approximation

For approximating the mean effect (Eq. (6)) and the uncertainty (Eq. (7)) we use the set $\mathcal{S}_k$ of dataset instances with the s-th feature lying inside the interval $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x^i_s < z_k\}$:

$$\hat{\mu}(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[f^s(\mathbf{x}^i)\right] \tag{10}$$

$$\hat{\sigma}^2(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left(f^s(\mathbf{x}^i) - \hat{\mu}(z_1, z_2)\right)^2 \tag{11}$$

Eq.(10) is an unbiased estimator of Eq. (6) under the assumption that the points are uniformly distributed inside the intervals. The same does not hold in the general case for Eq.(11). In the general case, $\hat{\sigma}^2(z_{k-1}, z_k)$ is an unbiased estimator of the *observable* variance $\sigma^2_{obs}(z_1, z_2) = \frac{\int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x}_c|x_s=z}\left[(f^s(z,\mathbf{x}_c) - \mu(z_1,z_2))^2\right]\partial z}{z_2 - z_1}$, which as we prove in Theorem 1 is an overestimation of $\sigma^2(z_1, z_2)$.

Theorem 1. *If we define the residual $\rho(z)$ as the difference between the expected effect at x_s and the regional effect, i.e $\rho(z) = \mu(z) - \mu(z_1, z_2)$, then the regional variance $\sigma^2(z_1, z_2)$ equals to:*

$$\sigma^2_{obs}(z_1, z_2) = \sigma^2(z_1, z_2) + \frac{\int_{z_1}^{z_2} \rho^2(z)\partial z}{z_2 - z_1} \tag{12}$$

Theorem 1 reveals an important connection between the ground-truth and the observable uncertainty inside a region. For convenience, we denote as $\mathcal{E}^2(z_1, z_2) = \frac{\int_{z_1}^{z_2} \rho^2(z)\partial z}{z_2 - z_1}$ the error term that models the expected square residual inside the interval and as $\mathcal{H}_{obs}(z_1, z_2) := \sigma_{obs}(z_1, z_2)$ the observable uncertainty. It holds that the observable uncertainty is an overestimation of the correct uncertainty

$$\mathcal{H}^2_{obs}(z_1, z_2) = \mathcal{H}^2(z_1, z_2) + \mathcal{E}^2(z_1, z_2) \tag{13}$$

and, therefore, the estimation is unbiased only in case $\mathcal{E}^2(z_1, z_2) = 0$.

3.4 Bin-Splitting: Finding Regions With Homogeneous Effects

In this section we formulate bin-splitting as an unsupervised clustering problem, searching for a solution that compromises two contradictory objectives. On the one hand, we want to minimize the error term $\mathcal{E}$ for approaching an unbiased estimation and on the other hand, we want wide-enough bins for including a fair population of samples for a robust estimation of $\hat{\mu}(z_1, z_2), \hat{\sigma}(z_1, z_2)$. We set-up the following optimization problem:

$$\begin{aligned} \min_{\mathcal{Z}=\{z_0,\ldots,z_K\}} \quad & \mathcal{L} = \sum_{k=1}^{K} \tau_k \hat{\sigma}^2(z_{k-1}, z_k)\Delta z_k \\ \text{where} \quad & \Delta z_k = z_k - z_{k-1} \\ & \tau_k = 1 - \alpha \frac{|S_k|}{N} \\ \text{s.t.} \quad & |\mathcal{S}_k| \geq N_{\texttt{NPB}} \\ & z_0 = x_{s,min} \\ & z_K = x_{s,max} \end{aligned} \tag{14}$$

We search for the sequence of intervals $z_0, \ldots, z_K$ that minimizes the sum of the bin costs. The cost of each bin is the approximated variance $\hat{\sigma}^2_k$ scaled by the bin length Δz_k and discounted by the term τ_k. The term τ_K favors the selection of a bigger bin in case it has similar variance with the aggregate variance of many smaller ones. The constraint of at least $N_{\texttt{PPB}}$ points per bin sets the lowest-limit for a *robust* estimation. The user can choose to what extent they favor the creation of wide bins through the discount parameter by setting the parameter α and where they set the minimum for robust approximation with the parameter $N_{\texttt{PPB}}$. For providing a rough idea, in our experiments we set $\alpha = 0.2$ which means that the discount ranges between $[0\%, 20\%]$ and $N_{\texttt{PPB}} = \frac{N}{20}$. It is also important to clarify that by minimizing $\hat{\sigma}^2_k \approx \mathcal{H}^2_{obs}$, we actually minimize the squared error term $\mathcal{E}^2$, since the term $\mathcal{H}^2_{true}$ is independent of the bin splitting.

3.4.1 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 $u = \frac{x_{s,max} - x_{s,min}}{K_{max}}$. For defining the solution, we use two indexes. The index $i \in \{0, \ldots, K_{max}\}$ denotes the point (z_i) and the index $j \in \{0, \ldots, K_{max}\}$ denotes the position of the j-th multiple of the minimum step, i.e., $x_j = x_{s,min} + j \cdot u$. The recursive cost function $T(i, j)$ is the cost of setting $z_i = x_j$:

$$\mathcal{T}(i, j) = \min_{l \in \{0,\ldots,K_{max}\}} \left[\mathcal{T}(i-1, l) + \mathcal{B}(x_l, x_j)\right] \tag{15}$$

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 $[x_l, x_j)]$:

$$\mathcal{B}(x_l, x_j) = \begin{cases} \infty, & \text{if } x_j > x_l \text{ or } |\mathcal{S}_{(x_j, x_l)}| < N \\ 0, & \text{if } x_j = x_l \\ \hat{\sigma}^2(x_j, x_l), & \text{if } x_j \leq x_l \end{cases} \tag{16}$$

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

- Discuss more aspects (e.g. Computational complexity)

4 SIMULATION EXAMPLES

Simulation examples provide the freedom to select the data-generating distribution $p(\mathbf{X})$ and the predictive function $f(\cdot)$. We follow this common XAI practice (Aas et al., 2021; Herbinger et al., 2022) for empirically evaluating UALE against a solid ground-truth. The first group of synthetic examples (Section 4.1) compares UALE with PDP-ICE plots in quantifying the average and the uncertainy of the effect. The second group (Section 4.2) evaluates the accuracy of the approximation (average effect and uncertainty) using the automatic bin-splitting algorithm. In both groups, we 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 UALE method against PDP-ICE in quantifying the main effect and the uncertainty. 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 create the following piecewise linear function:

$$f(\mathbf{x}) = \begin{cases} f_{\texttt{lin}} + \alpha f_{\texttt{int}} & \text{if } f_{\texttt{lin}} < 0.5 \\ 0.5 - f_{\texttt{lin}} + \alpha f_{\texttt{int}} & \text{if } 0.5 \leq f_{\texttt{lin}} < 1 \\ \alpha f_{\texttt{int}} & \text{otherwise} \end{cases} \tag{17}$$

where $f_{\texttt{lin}}(\mathbf{x}) = a_1 x_1 + a_2 x_2$ and $f_{\texttt{int}}(\mathbf{x}) = x_1 x_3$. The linear term $f_{\texttt{lin}}$ includes the two correlated features and $f_{\texttt{int}}$ interacts the two non-correlated features. We evaluate the effect computed by UALE and PDP-ICE in three cases; (a) without interaction ($\alpha = 0$) and equal weights ($a_1 = a_2$), (b) without 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 provide the ground-truth average effect and uncertainty computed with analytical derivations (proofs at the Appendix) and we compare it against the approximation provided by each method. The approximation is computed after generating $N = 500$ samples.

No Interaction, Equal weights. We test the feature effect when no interaction exits $\alpha = 0$ and the weights are equal $a_1 = a_2 = 1$. 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. There zero heterogeneous effects, therefore $f_{\sigma^2}^{\mathcal{GT}}(x_1) = 0$. In Figure 1, 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 constant-effect regions.

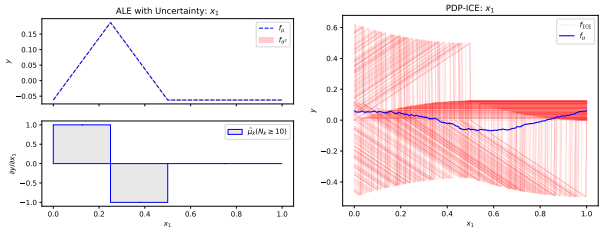

Figure 1: No interaction, Equal weights: Feature effect for x_1; Ground-truth vs (a) UALE method at the left and (b) PDP-ICE at the right.

No Interaction, Different weights. As before, no interaction exits $\alpha = 0$, therefore, the ground-truth uncertainty is zero, i.e. $f_{\sigma^2}^{\mathcal{GT}}(x_1) = 0$. The weigts are $a_1 = 2, a_2 = 0.5$, 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 the region $[0, 0.2)$ and positive in $[0.2, 0.4)$, and the ICE erroneously implies the existence of heterogeneous effects. As before, ALE quantifies correctly the ground truth effect, the zero-uncertainty and extracts the constant effect regions.

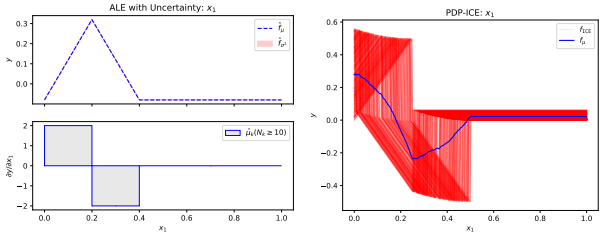

Figure 2: No interaction, Different weights: Feature effect for x_1; Ground-truth vs (a) UALE method at the left and (b) PDP-ICE at the right.

Uncertainty, Equal weights. In this case we activate the interaction term, i.e. $a = 1$, and we set the weights to $a_1 = a_2 = 1$. The interaction term provokes heterogeneous effects for features x_1, x_3, because the local effects at x_1 depend on the unknown value of X_3 and vice-versa. The effect of x_2 has zero-uncertainty since it does not appear in any interaction term. As we observe in Figure 3, UALE correcly models the average effect and the uncertainty in all cases. PDP-ICE quantifies correctly the effect and the uncertainty only in the case of feature x_3, because X_3 is independent from other features. For features x_1, x_2 that are correlated the average effect of PDP and the uncertainty of ICE plots are wrong.

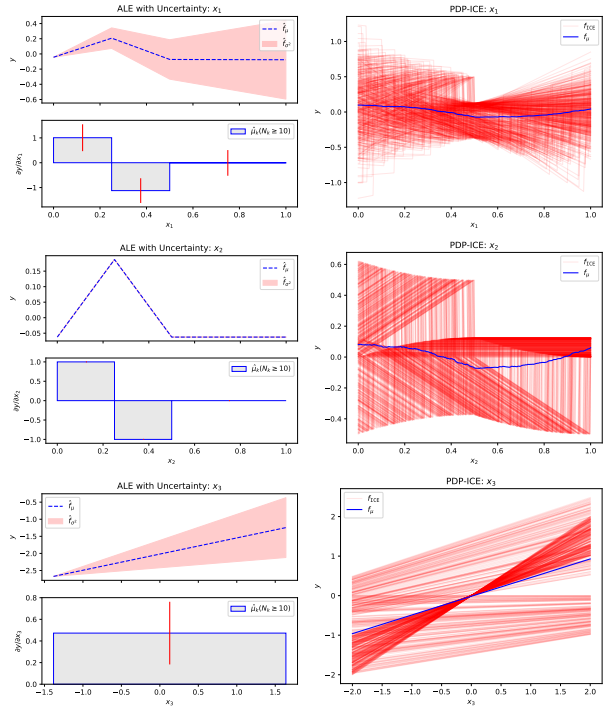

Figure 3: With interaction, equal weights: Feature effect for all features, x_1 to x_3 from top to bottom; Ground-truth vs (a) UALE method at the left columns and (b) PDP-ICE at the right column.

Discussion. Despite the model's simplicity, PDP-ICE fail in quantifying the average effect and the uncertainty due to correlation between features X_1 and X_2. UALE quantifies them correctly and estimates them accurately extracting the regions with constant effect. The examples above do not cover the case of an interaction term between correlated features, e.g. $x_1 x_2$, because there is an open debate about the ground-truth effect in this case (Grömping, 2020).

4.2 Case 2: Bin-Splitting

In this simulation, we aim to to quantify the advantages of automatic bin-splitting. We set-up 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 the fixed-size DALE estimation (for many different K) versus the automatic bin-splitting. 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 $\mathcal{Z}^{\mathtt{DP}} = \{z_{k-1}^{\mathtt{DP}}, \cdots, z_K^{\mathtt{DP}}\}$ the sequence obtained by automatic bin-splitting and with $\mathcal{Z}^{\mathtt{K}}$ the fixed-size splitting with K bins. For consistent results, in all the examples below, we regenerate samples and repeat the computations for $t = 30$ times, providing the mean value for all metrics.

Metrics The evaluation is done with regard to two metrics that count the mean bin error, where bin error is the absolute difference of the approximation from the ground-truth. The metric $\mathcal{L}^{\mu}_{\mathtt{DP|K}} = \frac{1}{|\mathcal{Z}^{\mathtt{DP|K}}|} \sum_{k \in \mathcal{Z}^{\mathtt{DP|K}}} |\mu_k - \hat{\mu}_k|$ counts the mean bin error on the average effect and $\mathcal{L}^{\sigma}_{\mathtt{DP|K}} = \frac{1}{|\mathcal{Z}^{\mathtt{DP|K}}|} \sum_{k \in \mathcal{Z}^{\mathtt{DP|K}}} |\sigma_k - \hat{\sigma}_k|$ the mean bin error on the uncertainty. In both cases the ground truth is computed from the dense bins of ground-truth that overlapse with the wider bin of the approximation. We also provide the mean (per bin) error term $\mathcal{L}^{\rho}_{\mathtt{DP|K}} = \frac{1}{|\mathcal{Z}^{\mathtt{DP|K}}|} \sum_{k \in \mathcal{Z}^{\mathtt{DP|K}}} \rho_k$ for a complete understanding of the uncertainty error.

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, UALE method extracts a sequence of intervals with better $\mathcal{L}^{\mu}_{\mathtt{DP}}$ and $\mathcal{L}^{\sigma^2}_{\mathtt{DP}}$ 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 $\mathcal{L}^{\mu}_{\mathtt{K}}$ 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 $\mathcal{L}^{\rho^2}_{\mathtt{K}}$ leading to a high variance error $\mathcal{L}^{\sigma^2}_{\mathtt{K}}$. 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 $\mathcal{L}^{\rho^2}_{\mathtt{K}} \approx 0$, 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.

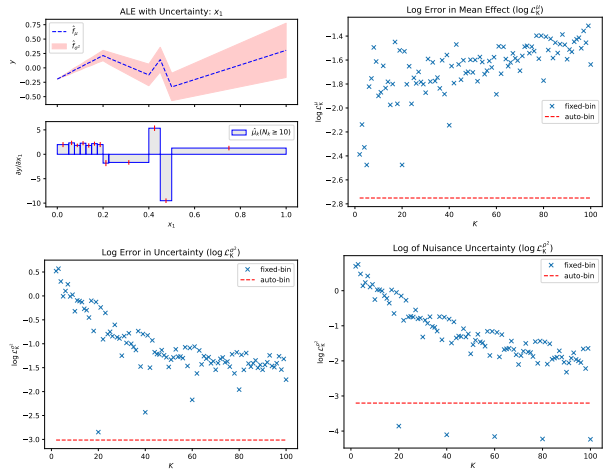

Figure 4: Figure 1

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 $\mathcal{L}_K^{\mu}$ and the variance error $\mathcal{L}_K^{\sigma^2}$.

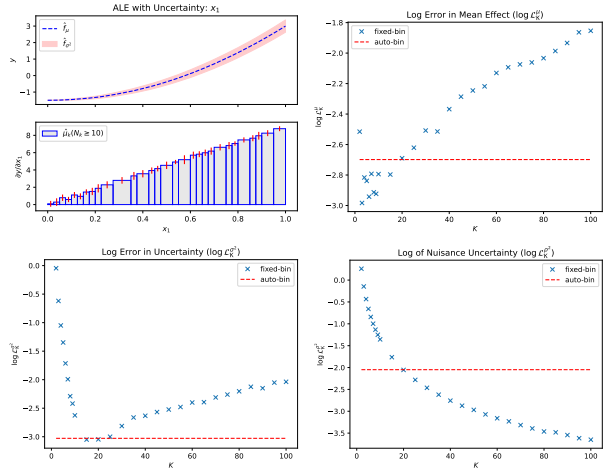

Figure 5: Figure 1

5 REAL-WORLD EXAMPLE

Here, we aim at demonstrating the usefuleness of uncertainty quantification and the advantages of automatic bin-splitting, on the real-world California Housing dataset (Pace and Barry, 1997).

ML setup The California Housing is a largely-studied dataset with approximately 20000 training instances, making it appropriate for robust Monte-Carlo approximations. The dataset contains $D = 8$ numerical features with characteristics about the building blocks of California, e.g. latitude, longitude, population of the block or median age of houses in the block. The target variable is the median value of the houses inside the block in dollars that ranges between $[15, 500] \cdots 10^3$, with a mean value of $\mu_Y \approx 201 \cdot 10^3$ and a standard deviation of $\sigma_Y \approx 110 \cdot 10^3$.

We exclude instances with missing values or with outlier values. As outlier we define the feature values which ar over three standard deviations away from the mean feature value. We also normalize all features to zero-mean and unit standard deviation. We split the dataset into $N_{tr} = 15639$ training and $N_{test} = 3910$ test examples (80/20 split) and we fit a Neural Network with 3 hidden layers of 256, 128 and 36 units respectively. After 15 epochs using the Adam optimizer with learning rate $\eta = 0.02$, the model achieves a normalized mean squared error (R-Squared) of 0.25 (0.75), which corresponds to a MAE of $37 \cdot 10^3$ dollars.

Below, we illustrate the feature effect for three features: latitude x_2, population x_6 and median income x_8. The particular features cover the main FE cases, e.g. positive/negative trend and linear/non-linear curve, and they are appropriate for illustration purposes. The complete results for all features, along with in-depth information about the preprocessing, training and evaluation parts are provided in the Appendix.

Uncertainty Quantification In real-world datasets, it is infeasible to obtain the ground truth FE for seamlessly evaluating the competitive methods. We selected the particular experiment, because, in broad terms, UALE and PDP-ICE plots aggree in the estimation of the average effect and uncertainty. Therefore, we can focus on juding the quality of the information provided by the two methods.

In Figure6, we observe, from top to bottom, the effects for the latitude, population and the median income. The effect of UALE and PDP-ICE are similar for the population and the median income. The population has a negative impact that progressively decreases: from 400 to 1500 people the house value decreases with a rate of $-150(\pm 140)$ dollars per added person, a rate that decreases from $-80(\pm 80)$ to $-60(\pm 60)$ dollars per added person as we move from 1500 to 2800 people. The level of uncertainty indicates significant variance in absolute value of the rate, but in the grant majority of instances the rate is negative. With the same inspection, we observe that the median annual income has a positive impact on the value (all numbers are thousands of dollars): 20 ± 15 per 10 of added median income for incomes in $[8, 15]$, 32 ± 20 per 10 added income in $[15, 60]$ and 40 ± 15 per 10 added income in $[60, 70]$. The uncertainty indicates that there are less heterogeneous effects

about the median income compared to the number of people. In both cases, we can end-up to the same conclusion by inspecting the PDP-ICE plots. For the latitude, there is a small difference in the explanations for the region $[32, 35]$, where UALE estimates a less negative slope with less uncertainty than PDP, while the explanations are similar for the range $[35, 39]$, where both methods reveal an increase in the uncertainty around the feature value 37.5.

In general, we observe that UALE complements ALE in the quantification of the heterogeneous effects, similarly to as ICE complements PDP. The automatic extraction of constant-effect and constant-uncertainty regions provided by UALE is helpful for an easier interpretation. On the other hand, ICE plots sometimes are more descritive on locating the type of heterogeneity, whereas UALE quantifies only the level (not the type) of heterogeneity.

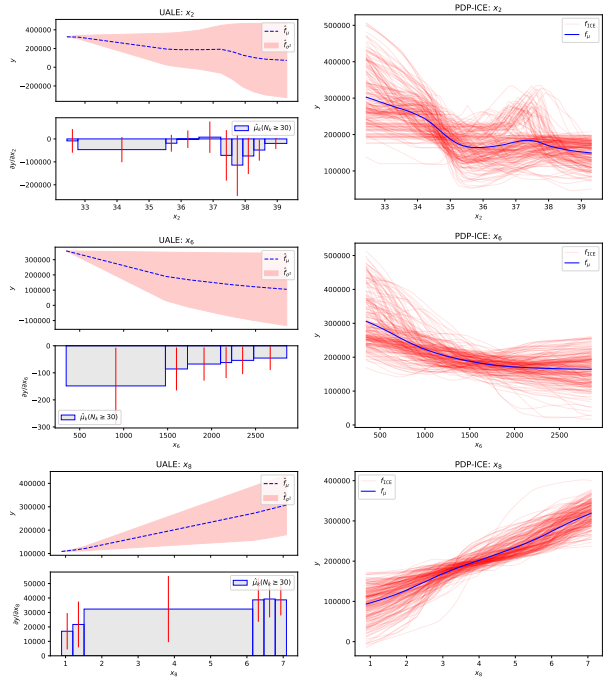

Figure 6: Figure 1

Bin Splitting In this part, we evaluate the robustness of the approximation using automatic bin-splitting. Following the evaluation framework of Section 4.2, we treat as ground-truth the effects computed on the full training-set $N = 20000$ with dense fixed-size bin-splitting ($K = 80$). Given the big number of samples, we make the hypothesis that the approximation with dense binning is close to the ground truth. Afterwards, we randomly select less samples $N = 1000$ and we compare UALE approximation with all possible fixed-size alternatives, repeating this process for 30 times for robust results. In Figure 7, we illustrate the mean values for $\mathcal{L}^{\mu}, \mathcal{L}^{\sigma}$ of the 30 repetitions. We observe that automatic bin-spliting provides (close to the) best approximation in the three features. In the Appendix, we provide the same evaluation for all features.

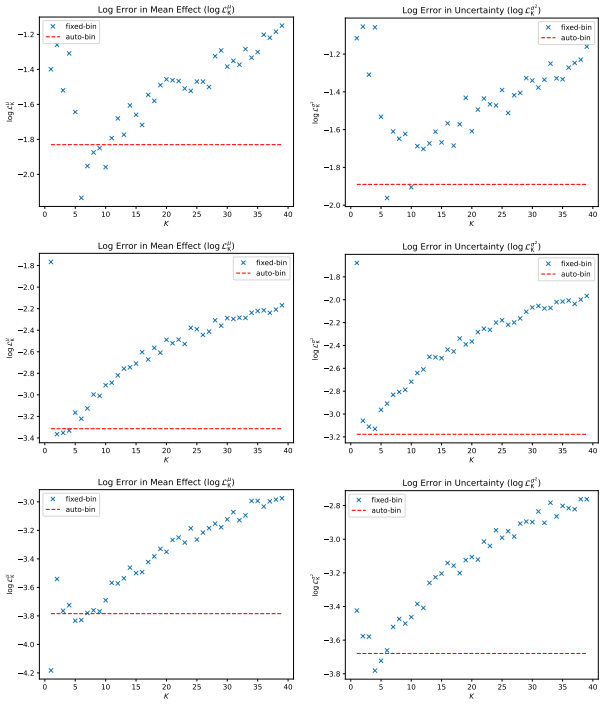

Figure 7: Figure 1

6 DISCUSSION

add limitations.

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.

Hubert Baniecki, Wojciech Kretowicz, and Przemyslaw Biecek. Fooling partial dependence via data poisoning. *arXiv preprint arXiv:2105.12837*, 2021.

Matthew Britton. Vine: visualizing statistical interactions in black box models. *arXiv preprint arXiv:1904.00561*, 2019.

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

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The annals of applied statistics*, pages 916–954, 2008.

Vasilis Gkolemis, Theodore Dalamagas, and Christos Diou. Dale: Differential accumulated local effects for efficient and accurate global explanations.

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.

Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

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.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

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.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.