RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

Vasilis Gkolemis[a, b], **Theodore Dalamagas**[b], **Eirini Ntoutsi**[c] **and Christos Diou**[a]

[a]Harokopio University of Athens
[b]ATHENA RC
[c]Universitat der Bundeswehr Munchen

Abstract. Accumulated Local Effects (ALE) is a widely-used explainability method for isolating the average effect of a feature on the output, because it handles cases with correlated features well. However, it has two limitations. First, it does not quantify the deviation of instance-level (local) effects from the average (global) effect, known as heterogeneity. Second, for estimating the average effect, it partitions the feature domain into user-defined, fixed-sized bins, where different bin sizes may lead to inconsistent ALE estimations. To address these limitations, we propose Robust and Heterogeneity-aware ALE (RHALE). RHALE quantifies the heterogeneity by considering the standard deviation of the local effects and automatically determines an optimal variable-size bin-splitting. In this paper, we prove that to achieve an unbiased approximation of the standard deviation of local effects within each bin, bin splitting must follow a set of sufficient conditions. Based on these conditions, we propose an algorithm that automatically determines the optimal partitioning, balancing the estimation bias and variance. Through evaluations on synthetic and real datasets, we demonstrate the superiority of RHALE compared to other methods, including the advantages of automatic bin splitting, especially in cases with correlated features.

1 Introduction

Recently, Machine Learning (ML) has been adopted across multiple areas of human activity, including mission-critical domains such as healthcare and finance. In such high-stakes areas, it is important to accompany predictions with meaningful explanations [24, 5]. For this reason, there is an increased interest in Explainable AI (XAI) [23, 14]. XAI literature distinguishes between local and global methods [18]. Local methods provide instance-level explanations [4], i.e., explain the prediction for a specific input, whereas global methods explain the entire model behavior [13]. Most of the times, global methods aggregate the instance-level explanations into a single interpretable outcome, usually a number or a plot.

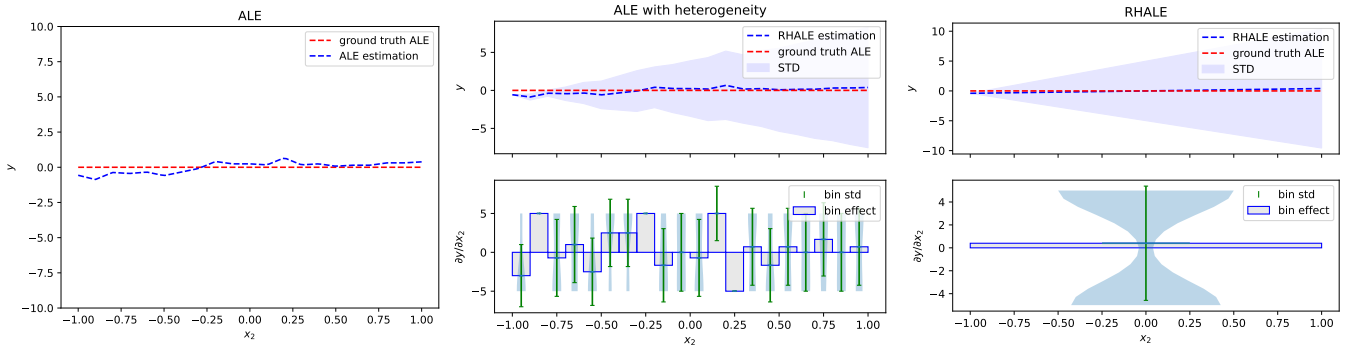
Feature Effect (FE) [11] is a class of global explainability methods that quantify the average (across all instances) partial relationship between one feature and the output. The most popular FE methods are *Partial Dependence Plots* (PDP) [6] and *Accumulated Local Effects* (ALE) [1]. PDPs have been criticized [2, 17, 21] for providing misleading explanations in problems with highly correlated features, making ALE the only reliable solution in such cases. Nevertheless, ALE has two significant limitations. Firstly, the way ALE formulates the FE (*ALE definition* of Eq. (2)), does not take into account the heterogeneity of instance-level effects, a quantity that is crucial for a complete interpretation of the average effect. Secondly, the way ALE estimates the FE from the instances of the training set (*ALE approximation* at Eq. (3)) relies on a user-defined binning process that often results in poor estimations. Therefore, this paper presents RHALE (Robust and Heterogeneity-aware ALE), a FE method build on-top of ALE that overcomes these issues. To better understand the advantages of RHALE over ALE, consider the following example, which was first introduced in [9]:

$$\begin{aligned} &Y = 0.2X_1 - 5X_2 + 10X_2\mathbb{1}_{X_3>0} + \mathcal{E} \\ &\mathcal{E} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad X_1, X_2, X_3 \overset{i.i.d.}{\sim} \mathcal{U}(-1,1) \end{aligned} \tag{1}$$

where we draw $N = 100$ samples, i.e, $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$. Given the knowledge of Eq. (1), the FE of X_3 is zero because the term $10X_2\mathbb{1}_{X_3>0}$, where X_3 appears, is part of the effect of X_2. In contrast, X_2 relates to Y in two opposite ways, as $-5X_2$ when $X_3 < 0$ and as $5X_2$ otherwise. Therefore, the zero average effect of X_2 after aggregating the two opposites effects, should not erroneously imply that X_2 does not affect Y. However, as demonstrated in Figure 1a (for X_2) and Figure 2a (for X_3) *ALE definition* erroneously indicates that both variables are not associated with the output. This phenomenon, known as aggregation bias [16, 12], is a common issue of global XAI methods.

RHALE addresses this issue by quantifying the heterogeneity based on the standard deviation of the underlying instance-level effects. As shown in Figure 1c, although the average X_2 effect is zero, the presence of two opposing groups of instance-level effects, namely, $5X_2$ and $-5X_2$, is revealed by both (a) the shaded area in the top subfigure (the limits of the shaded area are the lines $5X_2$ and $-5X_2$) and (b) the violin plots in the bottom subfigure (the distribution of the instance-level effects has most of its mass at -5 and 5). In contrast, in Figure 2b, the zero heterogeneity states that X_3 is indeed not related to the output.

The second limitation is that *ALE approximation* requires an initial step where the feature axis is partitioned in K non-overlapping fixed-size bins. Afterwards, an average effect (bin-effect) is computed inside each bin, and ALE plot is the aggregation of the bin effects. Since there is no clear indication of an appropriate value for K, users often rely on heuristics, such as ensuring that each bin contains on average at least τ instances on average. In the example above, for $\tau = 5$, then $K = 20$, which, as we show in Figure 1b, results in significant approximation errors. Specifically, the bin-effects and bin-std values deviate significantly from their ground-truth, which are 0 and 5, respectively. To overcome this limitation, RHALE *auto-*

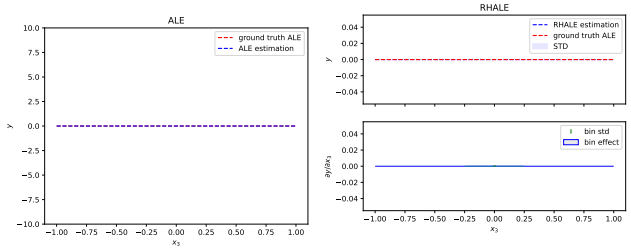


(a) x_2 ALE plot (20 bins)

(b) x_2 ALE + heterogeneity (20 bins)

(c) x_2 RHALE (auto-binning)

Figure 1: Feature effect for x_2 on the simple example of Eq. (1); (a) ALE incorrectly suggests that X_2 does not relate to Y, (b) ALE with heterogeneity using $K = 20$ fixed-size bins leads to significant approximation errors, (c) RHALE accurately estimates both the main effect and the heterogeneity, indicating that the zero average effect comes from opposite groups of instance-level effects.



(a) x_3 ALE plot (20 bins)

(b) x_3 RHALE plot (auto-binning)

Figure 2: Feature effect for x_3 on the simple example of Eq. (1). ALE plot suggests that X_3 does not relate to Y. However, as seen in Figure 1a, this interpretation can be misleading. Only after noticing the zero heterogeneity (STD and bin-std are zero) of RHALE plot in (b), we can confirm this claim.

matically determines the optimal sequence of variable-size bins. For example, in Figure 1c, RHALE automatically finds that it is optimal to create a single bin, which leads to a good approximation of both the average effect and the heterogeneity. The optimal bin splitting depends on the underlying instance-level effects (Section 3.2). Essentially, a wide bin reduces the variance of the estimation by increasing the samples inside the bin (better Monte Carlo approximation) but it may also introduce bias in the estimation. Using this insight, we formulate an optimization problem and propose an algorithm that minimizes the bias while ensuring that each bin has a sufficient number of samples. The main contributions of this paper are:

- A new feature effect method, RHALE, that addresses aggregation bias by providing information on the heterogeneity of instance-level effects.
- A formulation of the selection of variable-sized bins to balance RHALE bias and variance as an optimization problem.
- An algorithm that efficiently solves the optimal bin partitioning problem.
- A thorough experimental evaluation of RHALE on synthetic and real datasets, demonstrating its superiority over other feature effect methods, both in terms of accuracy and efficiency.

The code for reproducing all experiments is provided at https://github.com/givasile/RHALE.

2 Background and related work

Let $\mathcal{X} \in \mathbb{R}^d$ be the d-dimensional feature space, $\mathcal{Y}$ the target space and $f(\cdot) : \mathcal{X} \to \mathcal{Y}$ the black-box function. We use index $s \in \{1, \dots, d\}$ for the feature of interest and $c = \{1, \dots, d\} - s$ for the rest. For convenience, to denote the input vector, we use $(x_s, \mathbf{x}_c)$ instead of $(x_1, \cdots, x_s, \cdots, x_D)$ and, for random variables, (X_s, X_c) instead of $(X_1, \cdots, X_s, \cdots, X_D)$. The training set $\mathcal{D} = \{(\mathbf{x}^i, y^i)\}_{i=1}^N$ is sampled i.i.d. from the distribution $\mathbb{P}_{X,Y}$. Finally, $f^{<\texttt{method}>}(x_s)$ denotes how $<$ `method` $>$ defines the feature effect and $\hat{f}^{<\texttt{method}>}(x_s)$ how it estimates it from the training set.

2.1 *Feature Effect Methods*

The most popular feature effect methods are: *Partial Dependence Plots* (PDP) and *Accumulated Local Effects* (ALE). PDP defines the FE as an expectation over the marginal distribution X_c, i.e., $f^{\texttt{PDP}}(x_s) = \mathbb{E}_{X_c}[f(x_s, X_c)]$. A variation of PDP, known as Marginal Plots (MP), computes the expectation over the conditional distribution $X_c|x_s$, i.e., $f^{\texttt{MP}}(x_s) = \mathbb{E}_{X_c|x_s}[f(x_s, X_c)]$. Both methods suffer from misestimations when features are correlated. PDP integrates over unrealistic instances and MP computes aggregated effects, i.e., attributes the combined effect of sets of features to a single feature [1]. ALE tackles these limitations, using a three-step computation; (a) the local effect at (z, X_c), $f^s(z, X_c) = \frac{\partial f}{\partial x_s}(z, X_c)$, is computed with the derivatives $\frac{\partial f}{\partial x_s}$ to isolate the effect of x_s, (b) the expected effect at z, $\mu(z) = \mathbb{E}_{X_c|z}[f^s(z, X_c)]$, is taken over $X_c|z$, and, (c) the accumulation, $\int \mu(z)dz$, retrieves the main effect. ALE definition is:

$$f^{\texttt{ALE}}(x_s) = \int_{x_{s,\min}}^{x_s} \underbrace{\mathbb{E}_{X_c|X_s=z}\left[f^s(z, X_c)\right]}_{\mu(z)} \partial z \tag{2}$$

where $x_{s,\min}$ is the minimum value of the s-th feature. In real ML problems, $p(\mathbf{X})$ is unknown, so [1] proposed estimating ALE from the training set with:

$$\hat{f}^{\texttt{ALE}}(x_s) = \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^{(i)} \in \mathcal{S}_k} \left[f(z_k, \mathbf{x}_c^{(i)}) - f(z_{k-1}, \mathbf{x}_c^{(i)}) \right] \tag{3}$$

where k_x the index of the bin such that $z_{k_x-1} \leq x_s < z_{k_x}$ and $\mathcal{S}_k$ is the set of the instances of the k-th bin, i.e. $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq$

$x_s^{(i)} < z_k\}$. In Eq. (3) the axis of the s-th feature is split in K equally-sized bins and the average effect in each bin (bin effect) is estimated using synthetic instances, where $x_s^{(i)}$ is set to the right (z_k) and left (z_{k-1}) limits of the bin. Recently, [8] proposed the Differential ALE (DALE) that computes the local effects of differentiable models without modifying the training instances:

$$\hat{f}^{\texttt{DALE}}(x_s) = \Delta x \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^{(i)} \in \mathcal{S}_k} f^s(\mathbf{x}^i) \tag{4}$$

Their method allows formulating large bins without creating out-of-distribution instances and changing the bin size without the need to recalculate the instance-level effects. However, both Eq. (3) and Eq. (4) are limited to equal-width partitioning, which has limitations. As discussed in the Introduction, selecting between narrow and wide bins is challenging and, even more, there are cases (Figure 4) where neither narrow nor wide bins are appropriate. In these cases, it is necessary to use variable bin sizes (Figure 3b).

2.2 *Heterogeneity Of Local Effects*

Relying only on the average effect may provide a misleading interpretation of the model. Thus, there is an increasing interest in quantifying the degree of divergence between local effects and the average effect, which is commonly referred to as heterogeneity of local effects. To measure heterogeneity, PDP has a local equivalent called Individual Conditional Expectation (ICE) [9]. ICE, along with its variations like c-ICE and d-ICE [9], creates one curve per instance, $f_i^{\texttt{ICE}}(x_s) = f(x_s, \mathbf{x}_c^{(i)})$, on top of the average PDP plot, as seen in Figure 3a. In this way, the user assesses the heterogeneity by visually inspecting the similarity between ICE curves. However, as demonstrated in Section 4.1, ICE plots have the same limitations as PDPs in cases of correlated features. Based on the variance of the ICE plots, [19] proposed a method to quantify the standard error around the PDP plot. Some other methods[12, 3, 20] attempt to address PDP-ICE failure in case of correlated features by clustering ICE plots based on their similarity. The focus of these works, however, is on regional effects, i.e., subsets of the input space with homogeneous effects, rather than global effects. Approaches such as the H-Statistic [7], Greenwell's interaction index [10], and SHAP interaction values [15] provide a metric that quantifies the level of interactions between feature pairs but do not provide insight into how interactions influence different parts of the feature effect plot. To the best of our knowledge, no existing method quantifies heterogeneous effects for ALE.

3 RHALE

RHALE visualizes the feature effect with a plot as illustrated in Figure 3b. The plot includes (a) $\hat{f}_\mu^{\texttt{RHALE}}(x_s)$, the robust estimation of ALE that shows the average effect (*RHALE estimation*), (b) $\texttt{STD}(x_s)$, the standard deviation of the ALE effect that shows the heterogeneity of the instance level effects (*STD*), (c) $\hat{\mu}_k \forall k$, the bin effects that show the average change on the output y given a small change in x_s (*bin effect*) and (d) $\hat{\sigma}_k \forall k$ the bin standard deviations that quantify the heterogeneity inside each bin (*bin std*). In each bin, a violin plot on top of the bin effect shows the exact distribution of the local effects. The variable-size partitioning presented in Section 3.2 leads to an accurate estimation of these quantities.

To explain these four interpretable quantities and to highlight the advantages of RHALE compared to PDP-ICE, we will use a running example. We define a generative distribution $p(\mathbf{x}) = p(x_1)p(x_2)p(x_3|x_1)$ where x_3 is highly correlated with x_1, while x_2 is independent from both. Specifically, x_1 lies in $[-0.5, 0.5]$ with most samples inside the first half, i.e. $p(x_1) = \frac{5}{6}\mathcal{U}(x_1; -0.5, 0) + \frac{1}{6}\mathcal{U}(x_1; 0, 0.5)$, x_3 is almost equal to x_1 i.e., $p(x_3|x_1) = \mathcal{N}(x_3; 0, \sigma_3 = 0.01)$ and $p(x_2) = \mathcal{N}(x_2; 0, \sigma_2 = 2)$. In the experiments, we use 60 samples drawn i.i.d. from $p(\mathbf{x})$. The predictive function is:

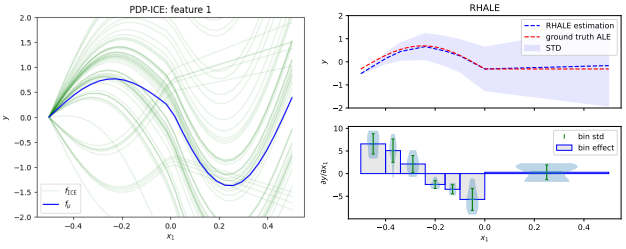
$$f(\mathbf{x}) = \sin(2\pi x_1)(\mathbb{1}_{x_1<0} - 2\mathbb{1}_{x_3<0}) + x_1 x_2 + x_2 \tag{5}$$

The simplicity of the toy example helps us isolate the effect of x_1, which is $f(x_1) \approx -\sin(2\pi x_1)\mathbb{1}_{x_1<0}$. This is because $x_3 \approx x_1$, so $(\mathbb{1}_{x_1<0} - 2\mathbb{1}_{x_3<0} \approx -\mathbb{1}_{x_1<0})$ and the effect of $x_1 x_2$ is $x_1 \mathbb{E}_{x_2}[x_2] = 0$. Furthermore, the only term that introduces heterogeneity is $x_1 x_2$, due to $x_2 \sim \mathcal{N}(0, 2)$ that varies among instances. Detailed derivations are provided at the Appendix B.1.

In Figure 3b we show that a user can interpret these effects from the RHALE plot. Specifically, from the top subplot, a user can interpret that (a) the average effect of x_1 is $\hat{f}_\mu^{\texttt{RHALE}}(x_1) \approx -\sin(2\pi x_1)\mathbb{1}_{x_1<0}$, and is produced after aggregating (b) instance level effects that vary in the region $-\sin(2\pi x_1)\mathbb{1}_{x_1<0} \pm 2x_1$. From the bottom subfigure, the user can interpret the FE at the bin-level. For example, for $x_1 \in [-0.5, -0.4]$ (c) the average change on the output is about $\frac{\partial f}{\partial x_1} \approx 6$ units of y and is produced after aggregating instance-level changes that vary in 6 ± 2. Furthermore, the violin plots show the exact distribution of the instance-level changes.

For estimating the above quantities from the 60 available samples, the optimized partitioning divides the sinusoidal region in six bins (dense enough) and merges the constant region in a single bin (robust estimation), balancing estimation variance and bias (Section 3.2). In contrast, Figure 4 shows that *all* fixed-size splits result in poor estimations; when using sparse bins ($K = 5$) the estimation is biased, as will be explained in Section 3.1, and when using dense bins ($K = 50$) the estimation has high variance.

A natural question that arises is whether we could come to the same interpretation using the PDP-ICE plot. At Figure 3a we observe that PDP with c-ICE, i.e., ICE curves centered to start from zero, lead to a completely misleading interpretation. For example, in $x_1 \in [0, 0.5]$, PDP shows a negative sinusoidal average effect and c-ICE two heterogeneous effects; a negative sinusoidal when $x_3^{(i)} \geq 0$ (for about $\frac{5}{6}$ of the instances), and a linear when $x_3^{(i)} < 0$ (for about $\frac{5}{6}$ of the instances). This is because PDP-based methods ignore the correlation between x_1 and x_3.



(a) PDP-ICE plot (b) RHALE plot

Figure 3: Feature effect for x_1 on the example of Eq. 5. Due to feature correlations, only RHALE provides a robust estimation of the main effect and the heterogeneity.

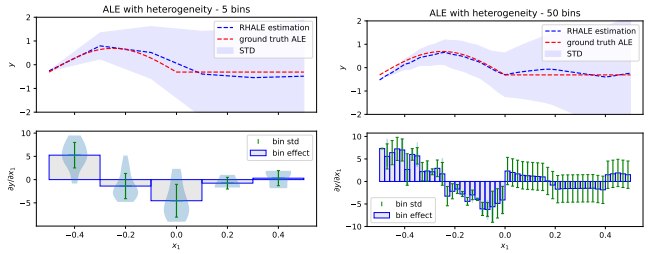


Figure 4: Estimation of the ALE effect, the standard error of ALE, the bin effect and the bin standard deviation using fixed-sized bins, $K = 5$ (left) and $K = 50$ (right).

3.1 Definition

We define the heterogeneity at $x_s = z$ as the standard deviation $\sigma(z)$ of the instance-level effects, where:

$$\sigma^2(z) = \mathbb{E}_{X_c|X_s=z}\left[\left(f^s(z, X_c) - \mu(z)\right)^2\right] \tag{6}$$

The variability is introduced by the implicit feature interactions. If the black-box function does not have any interaction term, i.e., it can be written as $f(\mathbf{x}) = f_s(x_s) + f_c(\mathbf{x}_c)$ then the variability is zero. For the interval-based formulation, we define the bin effect $\mu(z_1, z_2)$ and the bin standard deviation $\sigma(z_1, z_2)$ as:

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

$$\sigma^2(z_1, z_2) = \mathbb{E}_{z\sim\mathcal{U}(z_1,z_2)}[\sigma^2(z)] = \frac{\int_{z_1}^{z_2} \sigma^2(z)\partial z}{z_2 - z_1} \tag{8}$$

The bin effect and the bin standard deviation quantify the average effect and the heterogeneity inside a bin, i.e., for a population $\mathbf{x}^{(i)} = (z^{(i)}, \mathbf{x}_c^{(i)})$, where $z^{(i)}$ is uniformly drawn from $\mathcal{U}(z_1, z_2)$ and $\mathbf{x}_c$ from $X_c|z^{(i)}$. Denoting as $\mathcal{Z}$ the sequence of $K+1$ points that partition the axis of the s-th feature into K variable-size intervals, i.e., $\mathcal{Z} = \{z_0, \ldots, z_K\}$, the interval-based formulation of RHALE is:

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

where k_x is the index of the bin such that $z_{k_x-1} \leq x_s < z_{k_x}$. Eq. (9) is no more than a piece-wise linear approximation of Eq. (2). The approximation of the bin effect and of the bin standard deviation is made from the set $\mathcal{S}_k$ of instances with the s-th feature in the k-th bin, i.e., $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x_s^{(i)} < z_k\}$. The bin effect is estimated with:

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

which is an unbiased estimator of Eq. (7) (Appendix A.1). The estimator of the bin deviation Eq. (8) is:

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

At Appendix A.2, we show that $\hat{\sigma}^2(z_1, z_2)$ is an unbiased estimator of $\sigma_*^2(z_1, z_2) = \frac{\int_{z_1}^{z_2} \mathbb{E}_{X_c|X_s=z}\left[(f^s(z,X_c)-\mu(z_1,z_2))^2\right]\partial z}{z_2-z_1}$ and in Theorem 1 we prove that in the general case, $\sigma_*^2(z_1, z_2) \geq \sigma^2(z_1, z_2)$. Therefore, without a principled bin-splitting strategy, $\hat{\sigma}^2(z_1, z_2)$ leads to an overestimation of the actual bin standard deviation $\sigma^2(z_1, z_2)$.

Theorem 1. *If we define (a) the residual $\rho(z)$ as the difference between the expected effect at z and the bin effect, i.e, $\rho(z) = \mu(z) - \mu(z_1, z_2)$, and (b) $\mathcal{E}(z_1, z_2)$ as the mean squared residual of the bin, i.e., $\mathcal{E}(z_1, z_2) = \frac{\int_{z_1}^{z_2} \rho^2(z)\partial z}{z_2 - z_1}$, then it holds*

$$\sigma_*^2(z_1, z_2) = \sigma^2(z_1, z_2) + \mathcal{E}^2(z_1, z_2) \tag{12}$$

Proof. The proof is at A.3 of the Appendix □

We refer to $\mathcal{E}^2(z_1, z_2)$ as bin error. Based on Theorem1, the estimation is unbiased only when $\mathcal{E}^2(z_1, z_2) = 0$.

3.2 Automatic Bin-Splitting

RHALE approximation is affected by (a) the number of instances (estimation variance) and (b) the error term $\mathcal{E}(z_1, z_2)$ (estimation bias), in each bin. On the one hand, we favor wide bins so that the estimation of $\hat{\mu}(z_1, z_2), \hat{\sigma}(z_1, z_2)$ comes from a sufficient population of samples (low estimation variance). On the other hand, we want to minimize the accumulated bin error, i.e., $\mathcal{E}_{\mathcal{Z}}^2 = \sum_{k=1}^{K} \mathcal{E}^2(z_{k-1}, z_k)\Delta z_k$, where $\mathcal{Z} = \{z_0, \cdots, z_K\}$ and $\Delta z_k = z_k - z_{k-1}$ (low estimation bias). We search for a partitioning that balances this trade-off.

Corollary 2. *If a bin-splitting $\mathcal{Z}$ minimizes the accumulated error $\mathcal{E}_{\mathcal{Z}}^2$, then it also minimizes $\sum_{k=1}^{K} \sigma_*^2(z_1, z_2)\Delta z_k$.*

Proof. The proof is based on the observation that $\sum_{k=1}^{K} \sigma^2(z_{k-1}, z_k)\Delta z_k = \sigma^2(z_0, z_K)(z_K - z_0)$ which is independent of the bin-splitting. A detailed proof is provided at Appendix A4. □

Corollary 2 shows that minimizing $\mathcal{E}_{\mathcal{Z}}^2$ is equivalent to minimizing $\sum_{k=1}^{K} \sigma_*^2(z_{k-1}, z_k)\Delta z_k$, which can be directly estimated from $\sum_{k=1}^{K} \hat{\sigma}^2(z_{k-1}, z_k)\Delta z_k$. Based on that, we set-up the following optimization problem:

$$\begin{aligned} \min_{\mathcal{Z}=\{z_0,\ldots,z_K\}} \quad & \mathcal{L} = \sum_{k=1}^{K} \tau_k \hat{\sigma}^2(z_{k-1}, z_k)\Delta z_k \\ \text{where} \quad & \Delta z_k = z_k - z_{k-1} \\ & \tau_k = 1 - \alpha\frac{|S_k|}{N} \\ \text{s.t.} \quad & |\mathcal{S}_k| \geq N_{\text{PPB}} \\ & z_0 = x_{s,min} \\ & z_K = x_{s,max} \end{aligned} \tag{13}$$

The objective $\mathcal{L}$ searches for a partitioning $\mathcal{Z}^* = \{z_0^*, \ldots, z_K^*\}$ with a low accumulated error $\mathcal{E}_{\mathcal{Z}}^2$ and when many partitionings have similar accumulated errors, the coefficient τ_K favors the one with wider bins (on average, more points per bin). The constraint of at least N_{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 parameter α that controls the discount τ_k and the parameter N_{PPB} that sets the minimum population per bin. A typical choice for α is 0.2, which means a discount between range of $[0\%, 20\%]$ and for N_{PPB} is $\frac{N}{20}$, which means at least $\frac{N}{20}$ points in each bin, where N is the dataset size.

For solving the optimization problem of Eq.13 we discretize the solution space. First, we set a threshold $K_{\max}$ on the maximum number of bins which, in turn, defines the minimum bin width, i.e.

$\Delta x_{\min} = \frac{x_{s,\max} - x_{s,\min}}{K_{\max}}$. Based on that, we restrict the bin limits to the multiples of the minimum width, i.e. $z_k = k \cdot \Delta x_{\min}$, where $k \in \{0, \cdots, K_{\max}\}$. In this discretized solution space, we find the global optimum using Dynamic Programming. To define the solution, we use two indexes; index $i \in \{0, \ldots, K_{\max}\}$ denotes the limit of the i-th bin (z_i) and the index $j \in \{0, \ldots, K_{\max}\}$ denotes the j-th multiple of the minimum step, i.e., $x_j = x_{s,\min} + j \cdot \Delta x_{\min}$. The recursive cost function $T(i, j)$ computes the cost of setting z_i to x_j:

$$\mathcal{T}(i, j) = \min_{l \in \{0, \ldots, K_{max}\}} [\mathcal{T}(i-1, l) + \mathcal{B}(l, j)] \tag{14}$$

The term $\mathcal{B}(l, j)$ is the cost of creating a bin with limits $[x_l, x_j)$. In our case, following Eq. (13), we set it to $\tau_k \hat{\sigma}^2(x_l, x_j)(x_j - x_l)$ if the bin is valid, i.e., $|\mathcal{S}_k| \geq N_{\texttt{PPB}}$, and to ∞ otherwise. The optimal partitioning $\mathcal{Z}^*$ is given by solving $\mathcal{L} = \mathcal{T}(K_{\max}, K_{\max})$ and keeping track of the sequence of steps. Therefore, the main RHALE effect is estimated as in Eq. (15) and its standard deviation as in Eq. (16):

$$\hat{f}^{\texttt{RHALE}}_{\mathcal{Z}^*}(x_s) = \sum_{k=1}^{k_x} \hat{\mu}(z_{k-1}, z_k)(z_k - z_{k-1}) \tag{15}$$

$$\texttt{STD}(x_s) = \sqrt{\sum_{k=1}^{k_x} (z_k - z_{k-1})^2 \hat{\sigma}^2(z_{k-1}, z_k)} \tag{16}$$

The bin effects $\hat{\mu}_k$ are estimated using Eq. 10 and and the heterogeneity by the standard deviation $\hat{\sigma}_k$ in each bin using Eq. 11.

Computational Complexity. The computational complexity of the DP solution is $\mathcal{O}(K^3_{\max})$ because we use the DALE formula of Eq. 4. This allows us to precompute the instance-level effects once in the beginning, and then, the bin-splitting algorithm simply reallocates them to different partitionings without reevaluating f for each partitioning. As a result, for up to roughly $K_{\max} = 100$ bins, our algorithm runs in a couple of seconds, regardless of the dataset size or the cost of evaluating f. On the other hand, a PDP-ICE plot needs to evaluate f on t positions along the x_s axis for all N dataset points, making it a much slower alternative. Additional details and experimental results on the computational aspect can be found in Appendix A5. Finally, it is worth noting that $K_{\max}$ only sets an upper limit and the optimal sequence $\mathcal{Z}^*$ can range from 1 to $K_{\max}$.

4 Simulation examples

To formally evaluate RHALE, we rely on simulation examples as the evaluation requires knowledge of the ground truth generating distribution X and the black-box function f. In contrast, in Section 5, we showcase the applicability of RHALE on a real-world dataset, but it impossible to conduct a formal evaluation in this setting. The evaluation of RHALE on simulation examples is two-fold. First, in Section 4.1 we conduct a formal comparison between RHALE and PDP-ICE to verify that RHALE performs well in cases with correlated features, which PDP-ICE struggles with. Second, in Section 4.2, we compare RHALE's automated bin-splitting approach against the traditional fixed-size approximation. We demonstrate that RHALE's bin-splitting technique produces more accurate estimations across various scenarios.

4.1 RHALE vs PDP-ICE

We consider a data generating distribution $p(\mathbf{x}) = p(x_3)p(x_2|x_1)p(x_1)$, where $x_1 \sim \mathcal{U}(0, 1)$, $x_2 = x_1 + \epsilon$ and $x_3 \sim \mathcal{N}\left(0, \sigma_3^2 = \frac{1}{4}\right)$. Here $\epsilon \sim \mathcal{N}(0, 0.01)$ is a small additive noise. The predictive function is:

$$f(\mathbf{x}) = \alpha f_2(\mathbf{x}) + \underbrace{f_1(\mathbf{x}) \mathbb{1}_{f_1(\mathbf{x}) \leq \frac{1}{2}}}_{g_1(\mathbf{x})} + \underbrace{(1 - f_1(\mathbf{x})) \mathbb{1}_{\frac{1}{2} < f_1(\mathbf{x}) < 1}}_{g_2(\mathbf{x})} \tag{17}$$

where $f_1(\mathbf{x}) = x_1 + x_2$ is the additive term and $f_2(\mathbf{x}) = x_1 x_3$ is the interaction term. We evaluate RHALE and PDP-ICE when (a) there is no heterogeneity ($\alpha = 0$) and (b) there is is heterogeneity implied by the interaction term ($\alpha > 0$). We use this simple example for being able to establish a ground truth for the main effect and the heterogeneity. For the main effect, we use that due to $x_2 \approx x_1$ we can determine the intervals where g_1 or g_2 are active. For the heterogeneity, we use the fact that under no-interactions, $a = 0$, the heterogeneity must be zero and we discuss separately the case with $a > 0$. For detailed derivations see Appendix B1.

Case a: Interaction term disabled. Given that $x_2 \approx x_1$, when $0 \leq x_1 < \frac{1}{4}$, then $f_1(\mathbf{x}) < \frac{1}{2}$, so the effect is x_1 and, similarly, when $\frac{1}{4} \leq x_1 < \frac{1}{2}$ the effect is $-x_1$. Therefore, the ground truth effect is $f^{\texttt{GT}}(x_1) = x_1 \mathbb{1}_{x_1 < \frac{1}{4}} + \left(\frac{1}{4} - x_1\right) \mathbb{1}_{\frac{1}{4} \leq x_1 < \frac{1}{2}}$. Since x_1 does not interact with any other feature, the heterogeneity is zero. In Figure 5, we observe that PDP's main effect is wrong and ICE plots show heterogeneous effects. In contrast, RHALE estimates correctly both the average effect and the heterogeneity. Finally, we observe that RHALE's bin-splitting optimally creates three wide bins, $[0, \frac{1}{4})$, $[\frac{1}{4}, \frac{1}{2})$, $[\frac{1}{2}, 1)$, in the regions with linear effect.

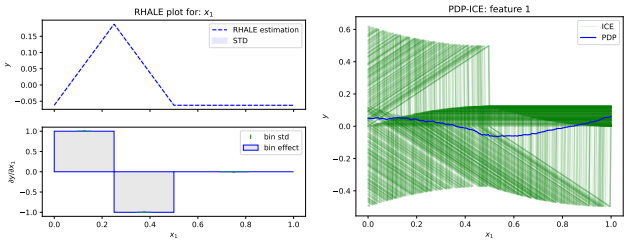


Figure 5: No interaction, Equal weights: Feature effect for x_1 using RHALE (Left) and PDP-ICE (Right).

Case b: Interaction term enabled. The main effects are $f^{\texttt{GT}}(x_j) = x_j \mathbb{1}_{x_j < \frac{1}{4}} + \left(\frac{1}{4} - x_j\right) \mathbb{1}_{\frac{1}{4} \leq x_j < \frac{1}{2}}$ for features $j = 1, 2$ and $f^{\texttt{GT}}(x_3) = \frac{1}{2} x_3$ for feature x_3. The interaction term $x_1 x_2$ induces heterogeneous effects for features x_1 and x_3, and since the two variables are independent, the heterogeneity is $\sigma_3 = \frac{1}{2}$ for x_1 and $\sigma_1 = \frac{1}{4}$ for x_3. In Figure 6, we observe that RHALE correctly estimates the main effect and the heterogeneity of all features. In contrast, PDP-ICE only estimates correctly only the effect and the heterogeneity of x_3. This confirms our previous knowledge that PDP-ICE performs well only when the interaction terms includes non-correlated features, like the term $f_2(\mathbf{x})$. For the correlated features x_1 and x_2, both the average effect and the heterogeneity are erroneously estimated by PDP-ICE.

Discussion. The study shows RHALE's superiority under correlated features, where, PDP and ICE plots can provide highly misleading results. Additionally, RHALE's automatic bin splitting leads to a robust estimation of the average effect and of the heterogeneity, favoring wider bins in regions with (near) constant effects.

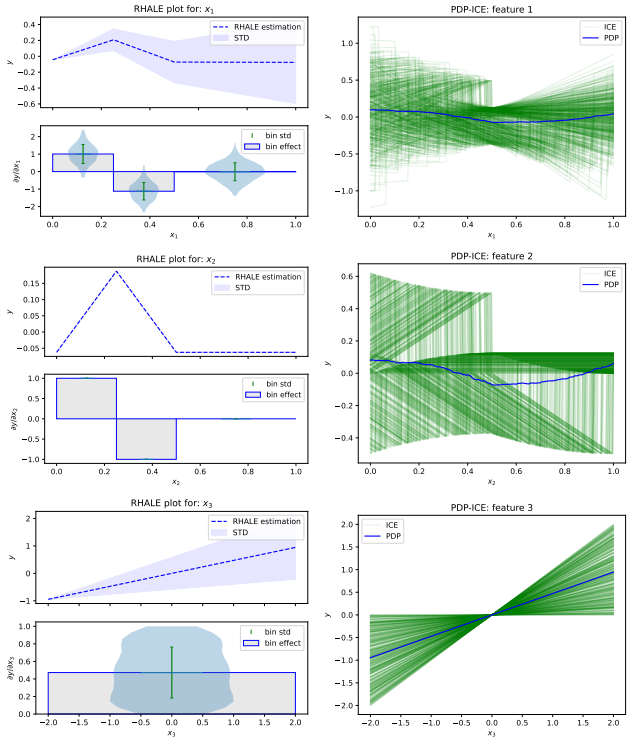


Figure 6: With interaction, equal weights: From top to bottom, feature effect for features $\{x_1, x_2, x_3\}$ using RHALE (left column) and PDP-ICE (right column).

4.2 *RHALE vs ALE*

In this simulation, we compare the performance of RHALE's automatic partitioning with ALE's fixed-size bin-splitting. To assess the accuracy of these approximations, we first estimate the ground truth average effect μ and heterogeneity σ using a large dataset ($N = 10^6$) with dense fixed-size binning ($K = 10^3$). We then generate a smaller dataset ($N = 500$) and compare the estimation of $\hat{\mu}, \hat{\sigma}$, using (a) fixed-size bins for several values of K against (b) RHALE's automatic partitioning. Our objective is to show that RHALE provides better estimates of μ and σ compared to any fixed-size alternative.

The dataset is generated by sampling from $p(\mathbf{x}) = p(x_2|x_1)p(x_1)$ where $x_1 \sim \mathcal{U}(0,1)$ and $x_2 \sim \mathcal{N}(x_1, \sigma_2^2 = 0.5)$. RHALE's approximation is denoted with $\mathcal{Z}^*$ and the fixed-size with K bins as $\mathcal{Z}^K$. The evaluation is based on the Mean Absolute Error (MAE) of the bin effect μ and of the heterogeneity σ across bins, i.e.,

$$\mathcal{L}^{\mu} = \frac{1}{|\mathcal{Z}| - 1} \sum_{k \in \mathcal{Z}} |\mu(z_{k-1}, z_k) - \hat{\mu}(z_{k-1}, z_k)| \tag{18}$$

$$\mathcal{L}^{\sigma} = \frac{1}{|\mathcal{Z}| - 1} \sum_{k \in \mathcal{Z}} |\sigma(z_{k-1}, z_k) - \hat{\sigma}(z_{k-1}, z_k)| \tag{19}$$

The ground truth bin effect, $\mu(z_{k-1}, z_k)$, and heterogeneity, $\sigma(z_{k-1}, z_k)$ are obtained by averaging the dense fixed-size bins within the interval $[z_{k-1}, z_k]$. We also calculate the mean residual error $\mathcal{L}^{\rho} = \frac{1}{|\mathcal{Z}|} \sum_{k \in \mathcal{Z}} \mathcal{E}(z_{k-1}, z_k)$ to interpret cases where the bin standard deviation is biased.

We compare RHALE vs ALE in two different scenarios; when f is piecewise linear and f is non-linear. We execute $t = 30$ independent runs, using each time $N = 500$ different samples, and report the mean values of the metrics.

Piecewise Linear Function. Here, the black-box function is $f(\mathbf{x}) = a_1 x_1 + x_1 x_2$, with 5 piecewise linear regions, i.e., a_1 equals to $\{2, -2, 5, -10, 0.5\}$ in the intervals defined by the sequence $\{0, 0.2, 0.4, 0.45, 0.5, 1\}$. The effect of x_1 is $f^{\text{GT}}(x_1) = a_1 x_1$ and the heterogeneity $\sigma_2 = \sqrt{0.5}$. As we observe in the top left of Figure 7, RHALE splits in fine-grained bins the intervals $[0.4, 0.45]$, $[0.45, 0.5]$ and unites in a single bin most of the constant-effect regions, e.g. region $[0.5, 1]$. Therefore RHALE's estimation is better than any fixed-size binning in terms of both $\mathcal{L}^{\mu}$ and $\mathcal{L}^{\sigma}$.

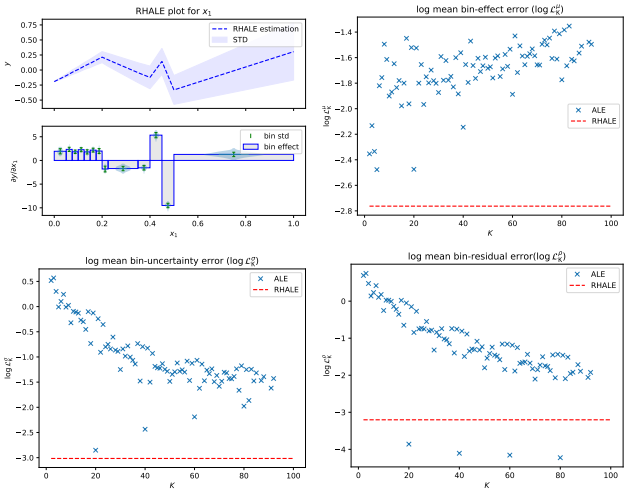


Figure 7: Bin-Splitting, piecewise linear function: RHALE's approximation (Top-Left). RHALE vs fixed-size approximations in terms of: $\mathcal{L}^{\mu}$ (Top-Right), $\mathcal{L}^{\sigma}$ (Bottom-Left), $\mathcal{L}^{\rho}$ (Bottom-Right).

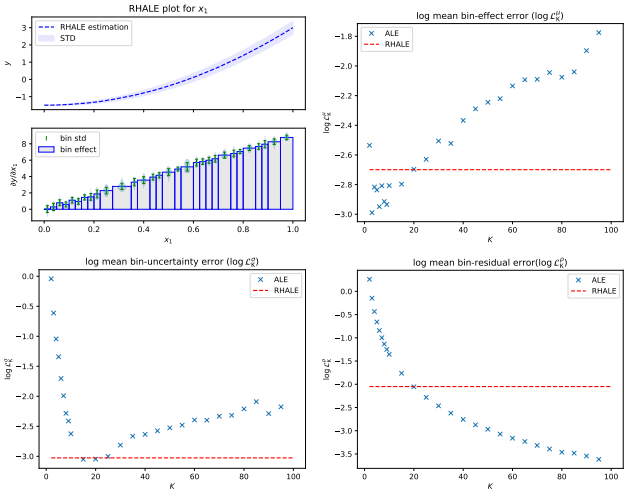


Figure 8: Bin-Splitting, non-linear function: RHALE's approximation (Top-Left). RHALE vs fixed-size approximations in terms of: $\mathcal{L}^{\mu}$ (Top-Right), $\mathcal{L}^{\sigma}$ (Bottom-Left), $\mathcal{L}^{\rho}$ (Bottom-Right).

Non-Linear Function. Here, the black-box function is $f(\mathbf{x}) = 4x_1^2 + x_2^2 + x_1 x_2$, so the effect of x_1 is $f^{\text{GT}}(x_1) = 4x_1^2$ and the heterogeneity is σ_2. When using a wide binning (low K) there is an increase in the mean residual error $\mathcal{L}^{\rho}$ (bottom-right of Figure 8), resulting in a biased approximation of σ. In contrast, narrow bins (high K) lead to a worse approximation due to number of samples

per bin. However, RHALE manages to compromise these competing objectives and achieves an (almost) optimal approximation of both μ (top-right) and σ (bottom-left), as illustrated in Figure 8.

5 Real-world example

Here, since it is infeasible to access the ground-truth FE, we simply demonstrate the usefulness of quantifying the heterogeneity and the advantages of RHALE's approximation, on the real-world California Housing dataset [22].

ML setup. The California Housing is a largely-studied dataset with approximately 20000 training instances, making it appropriate for robust approximation with large K. 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 $[15, 500] \cdot 10^3$, 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 or outlier values and we 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 MAE of $37 \cdot 10^3$ dollars.

Below, we illustrate the feature effect for two features: latitude x_2 and median income x_8. The particular features cover the main FE cases, e.g. positive/negative trend and linear/non-linear curve, and are therefore appropriate for illustration purposes. Results for all features, along with details about the reprocessing, training and evaluation parts are provided in the Appendix B2.

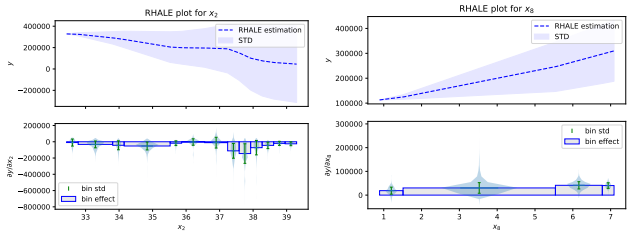


Figure 9: RHALE plot for features x_2 (latitude) and x_6 (median income). Apart from the average effects, i.e., negative for the x_2 and positive for x_8, the heterogeneity ($\pm$STD and BIN-STD) shows that instance-level effects ara more heterogeneous on x_2 case.

Heterogeneity Quantification Figure 9 illustrates the significance of RHALE's heterogeneity quantification for comprehensive interpretation of feature effects. We observe that both features exhibit significant interactions with other features leading to high heterogeneity. However, despite the high heterogeneity, we can confidently infer that the (a) latitude of the house (x_2) negatively impacts the price, and the (b) median income (x_8) has a positive influence on the price, for almost all instances.

Bin Splitting We evaluate the robustness of RHALE approximation, following the same methodology as described in Section 4.2. We consider as ground-truth the effects computed on the entire training set ($N_{tr} = 15639$) with dense fixed-size bin-splitting ($K = 80$). Given the sufficient number of samples, we make the hypothesis that the approximation with dense binning is close enough to the ground truth. Next, we randomly select fewer samples, $N = 1000$, and compare RHALE's approximation against fixed-size approximation (for all K). We repeat this process for $t = 30$ independent runs and we report the mean values of $\mathcal{L}^{\mu}, \mathcal{L}^{\sigma}$. In Figure 10, we observe that RHALE achieves accurate approximations in all cases; $\mathcal{L}^{\mu}$ $\mathcal{L}^{\sigma}$ are close to the best among the fixed-size approximations.

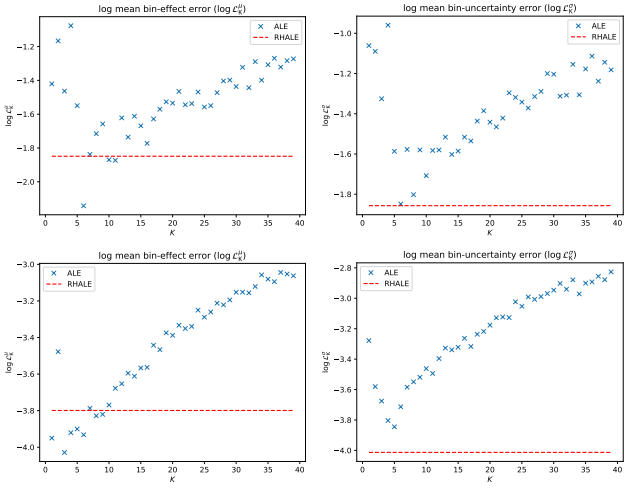


Figure 10: Lower is better. RHALE (red line) vs ALE fixed-size bins (blue crosses) in terms of $\mathcal{L}^{\mu}$ (left column), $\mathcal{L}^{\sigma}$ (right column) for features x_2 (top) x_8 (bottom). We observe that RHALE's estimation is better than (almost) any fixed-size alternative.

6 Conclusion and further work

In this paper, we have introduced Robust and Heterogeneity-aware ALE (RHALE), a global feature effect method that addresses two major limitations of ALE. First, it quantifies the heterogeneity of local effects, which is essential for a complete interpretation of the feature effect. Second, it automates the bin-splitting process to improve the approximation of both the average effect and the heterogeneity. To achieve the latter, we proposed an automatic bin-splitting algorithm that balances estimation bias and variance by creating wider bins only when the underlying local effects are (near) constant. Our experiments on synthetic and real-world examples demonstrate RHALE's superiority over PDP-ICE, which struggles with correlated features, and traditional ALE, as automatic bin-splitting provides more accurate estimates than fixed-size splitting.

Limitations. While the standard deviation of local effects is a good way to express the *level* of heterogeneity, it is challenging to interpret the *type* of heterogeneity. Therefore, we use violin plots to provide the distribution of local effects (type of heterogeneity), but their explanatory power is limited within each bin. At a global level, i.e., between the bins, the user can only determine the magnitude of heterogeneity. Finally, the automatic-binning algorithm comes with three hyperparameters, $K_{\max}, \alpha, N_{\text{PPB}}$. Although, their default values work well in most cases, on exceptional scenarios, such as a very small dataset, may need to be adjusted appropriately for an optimal bin splitting.

Acknowledgements

This work was supported by the XMANAI project (grant agreement No 957362), which has received funding by the European Regional Development Fund of the EU (EU 2020 Programme, ICT-38-2020 - Artificial intelligence for manufacturing).

References

[1] 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).

[2] Hubert Baniecki, Wojciech Kretowicz, and Przemyslaw Biecek, 'Fooling partial dependence via data poisoning', *arXiv preprint arXiv:2105.12837*, (2021).

[3] Matthew Britton, 'Vine: visualizing statistical interactions in black box models', *arXiv preprint arXiv:1904.00561*, (2019).

[4] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl, 'Visualizing the feature importance for black box models', in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pp. 655–670. Springer, (2019).

[5] Timo Freiesleben, Gunnar König, Christoph Molnar, and Alvaro Tejero-Cantero, 'Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena', *arXiv preprint arXiv:2206.05487*, (2022).

[6] Jerome H Friedman, 'Greedy function approximation: a gradient boosting machine', *Annals of statistics*, 1189–1232, (2001).

[7] Jerome H Friedman and Bogdan E Popescu, 'Predictive learning via rule ensembles', *The annals of applied statistics*, 916–954, (2008).

[8] Vasilis Gkolemis, Theodore Dalamagas, and Christos Diou, 'Dale: Differential accumulated local effects for efficient and accurate global explanations', *arXiv preprint arXiv:2210.04542*, (2022).

[9] 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).

[10] 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).

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

[12] Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio, 'Repid: Regional effect plots with implicit interaction detection', in *International Conference on Artificial Intelligence and Statistics*, pp. 10209–10233. PMLR, (2022).

[13] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo, 'Examples are not enough, learn to criticize! criticism for interpretability', *Advances in neural information processing systems*, **29**, (2016).

[14] Pang Wei Koh and Percy Liang, 'Understanding black-box predictions via influence functions', in *International conference on machine learning*, pp. 1885–1894. PMLR, (2017).

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

[16] 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).

[17] Christoph Molnar, *Interpretable Machine Learning*, 2 edn., 2022.

[18] 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*, pp. 417–431. Springer, (2020).

[19] Christoph Molnar, Timo Freiesleben, Gunnar König, Giuseppe Casalicchio, Marvin N Wright, and Bernd Bischl, 'Relating the partial dependence plot and permutation feature importance to the data generating process', *arXiv preprint arXiv:2109.01433*, (2021).

[20] 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*, (2020).

[21] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl, 'General pitfalls of model-agnostic interpretation methods for machine learning models', in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 39–68. Springer, (2022).

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

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, '"Why should i trust you?" Explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, (2016).

[24] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al., 'Do no harm: a roadmap for responsible machine learning for health care', *Nature medicine*, **25**(9), 1337–1340, (2019).