Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models

Daniel W. Apley and Jingyu Zhu
Northwestern University, USA

Summary. In many supervised learning applications, understanding and visualizing the effects of the predictor variables on the predicted response is of paramount importance. A shortcoming of black box supervised learning models (e.g., complex trees, neural networks, boosted trees, random forests, nearest neighbors, local kernel-weighted methods, support vector regression, etc.) in this regard is their lack of interpretability or transparency. Partial dependence (PD) plots, which are the most popular approach for visualizing the effects of the predictors with black box supervised learning models, can produce erroneous results if the predictors are strongly correlated, because they require extrapolation of the response at predictor values that are far outside the multivariate envelope of the training data. As an alternative to PD plots, we present a new visualization approach that we term accumulated local effects (ALE) plots, which do not require this unreliable extrapolation with correlated predictors. Moreover, ALE plots are far less computationally expensive than PD plots.

Keywords: Functional ANOVA; Marginal plots; Partial dependence plots; Supervised learning; Visualization

1. Introduction

With the proliferation of larger and richer data sets in many predictive modeling application domains, black box supervised learning models (e.g., complex trees, neural networks, boosted trees, random forests, nearest neighbors, local kernel-weighted methods, support vector regression, etc.) are increasingly commonly used in place of more transparent linear and logistic regression models to capture nonlinear phenomena. However, one shortcoming of black box supervised learning models is that they are difficult to interpret in terms of understanding the effects of the predictor variables (aka predictors) on the predicted response. For many applications, understanding the effects of the predictors is critically important. This is obviously the case if the purpose of the predictive modeling is explanatory, such as identifying new disease risk factors from electronic medical record databases. Even if the purpose is purely predictive, understanding the effects of the predictors may still be quite important. If the effect of a predictor violates intuition (e.g., if it appears from the supervised learning model that the risk of experiencing a cardiac event *decreases* as patients age), then this is either an indication that the fitted model is unreliable or that a surprising new phenomenon has been discovered. In addition, predictive models must be transparent in many regulatory environments, e.g., to demonstrate to regulators that consumer credit risk models do not penalize credit applicants based on age, race, etc.

To be more concrete, suppose we have fit a supervised learning model for approximating $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \approx f(\mathbf{x})$, where Y is a scalar response variable, $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ is a vector of d predictors, and $f(\cdot)$ is the fitted model that predicts Y (or the probability that Y falls into a particular class, in the classification setting) as a function of $\mathbf{X}$. To simplify notation, we omit any $\hat{}$ symbol over f, with the understanding that it is fitted from data. The training data to which the model is fit consists of n $(d+1)$-variate observations $\{y_i, \mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d}) :$

†Address for correspondence: Daniel W. Apley, Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208, USA.
E-mail: apley@northwestern.edu

本文用 predictors 即常说的 feature

如果仅直觉的效应出现了，那说明 model 不可靠或者发现了新现象

d 是 feature 数量

f(·) 是 model 方法。

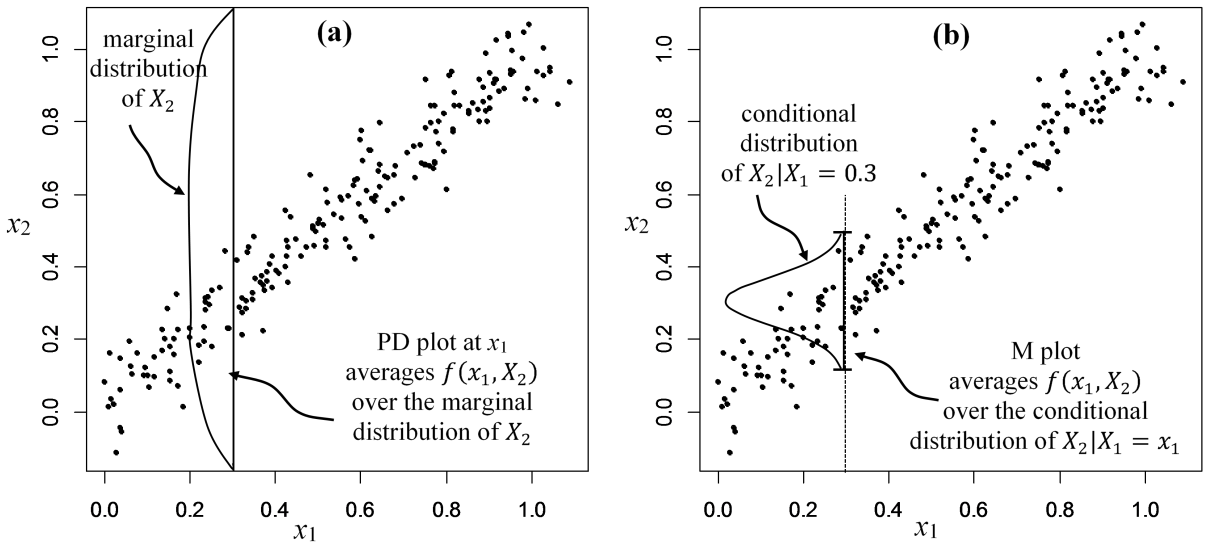

Fig. 1. Illustration of the differences between the computation of (a) $f_{1,PD}(x_1)$ and (b) $f_{1,M}(x_1)$ at $x_1 = 0.3$.

$i = 1, 2, \ldots, n\}$. Throughout, we use upper case to denote a random variable and lower case to denote specific or observed values of the random variable.

Our objective is to visualize and understand the "main effects" dependence of $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_d)$ on each of the individual predictors, as well as the low-order "interaction effects" among pairs of predictors. Throughout the introduction we illustrate concepts for the simple $d = 2$ case. The most popular approach for visualizing the effects of the predictors is partial dependence (PD) plots, introduced by Friedman (2001). To understand the effect of one predictor (say X_1) on the predicted response, a PD plot is a plot of the function

$$f_{1,PD}(x_1) \equiv \mathbb{E}[f(x_1, X_2)] = \int p_2(x_2) f(x_1, x_2) dx_2 \tag{1}$$

versus x_1, where $p_2(\cdot)$ denotes the marginal distribution of X_2. We use $p(\cdot)$ to denote the full joint probability density of $\mathbf{X}$, and use $p_{\cdot}(\cdot)$, $p_{\cdot|\cdot}(\cdot|\cdot)$, and $p_{\cdot,\cdot}(\cdot,\cdot)$ to respectively denote the marginal, conditional, and joint probability density functions of various elements of $\mathbf{X}$, with the subscripts indicating which elements. An estimate of (1), calculated pointwise in x_1 for a range of x_1 values, is

$$\hat{f}_{1,PD}(x_1) \equiv \frac{1}{n} \sum_{i=1}^{n} f(x_1, x_{i,2}). \tag{2}$$

Figure 1(a) illustrates how $f_{1,PD}(x_1)$ is computed at a specific value $x_1 = 0.3$ for a toy example with $n = 200$ observations of (X_1, X_2) following a uniform distribution along the line segment $x_2 = x_1$ but with independent $N(0, 0.05^2)$ variables added to both predictors (see Hooker (2007), for a similar example demonstrating the adverse consequences of extrapolation in PD plots). Although we ignore the response variable for now, we return to this example in Section 4 and fit various models $f(\mathbf{x})$ to these data. The salient point in Figure 1(a), which illustrates the problem with PD plots, is that the integral in (1) is the weighted average of $f(x_1, X_2)$ as X_2 varies over its marginal distribution. This integral is over the entire vertical line segment in Figure 1(a) and requires rather severe extrapolation beyond the envelope of the training data. If one were to fit a simple parametric model of the correct form (e.g., $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$), then this extrapolation might be reliable. However, by nature of its flexibility, a nonparametric supervised learning model like a regression tree cannot be expected to extrapolate reliably. As we demonstrate later (see Figures 5—7), this renders the PD plot an unreliable indicator of the effect of x_1.

main effect：独立的 feature 效应

P(·) 表示 联合（joint）概率密度

$P_{\cdot}(\cdot)$ 表示边缘效应

$P_{\cdot|\cdot}(\cdot|\cdot)$ 表示条件

$P_{\cdot,\cdot}(\cdot,\cdot)$ ：联合

The extrapolation in Figure 1(a) that is required to calculate $f_{1,PD}(x_1)$ occurs because the marginal density $p_2(x_2)$ is much less concentrated around the data than the conditional density $p_{2|1}(x_2|x_1)$, due to the strong dependence between X_2 and X_1. Marginal plots (M plots) are alternatives to PD plots that avoid such extrapolation by using the conditional density in place of the marginal density. As illustrated in Figure 1(b), an M plot of the effect of X_1 is a plot of the function

$$f_{1,M}(x_1) \equiv \mathbb{E}[f(X_1, X_2)|X_1 = x_1] = \int p_{2|1}(x_2|x_1) f(x_1, x_2) dx_2 \tag{3}$$

versus x_1. A crude estimate of $f_{1,M}(x_1)$ is

$$\hat{f}_{1,M}(x_1) \equiv \frac{1}{n(x_1)} \sum_{i \in N(x_1)} f(x_1, x_{i,2}), \tag{4}$$

where $N(x_1) \subset \{1, 2, \ldots, n\}$ is the subset of row indices i for which $x_{i,1}$ falls into some small, appropriately selected neighborhood of x_1, and $n(x_1)$ is the number of observations in the neighborhood. Although more sophisticated kernel smoothing methods are typically used to estimate $f_{1,M}(x_1)$, we do not consider them here, because there is a more serious problem with using $f_{1,M}(x_1)$ to visualize the main effect of X_1 when X_1 and X_2 are dependent. Namely, using $f_{1,M}(x_1)$ is like regressing Y onto X_1 while ignoring (i.e., marginalizing‡ over) the nuisance variable X_2. Consequently, if Y depends on X_1 and X_2, $f_{1,M}(x_1)$ will reflect both of their effects, a consequence of the omitted variable bias phenomenon in regression.

The main objective of this paper is to introduce a new method of assessing the main and interaction effects of the predictors in black box supervised learning models that avoids the foregoing problems with PD plots and M plots. We refer to the approach as accumulated local effects (ALE) plots. For the case that $d = 2$ and $f(\cdot)$ is differentiable (the more general definition is deferred until Section 2), we define the ALE main-effect of X_1 as

$$\begin{aligned} f_{1,ALE}(x_1) &\equiv \int_{x_{\min,1}}^{x_1} \mathbb{E}[f^1(X_1, X_2)|X_1 = z_1] dz_1 - \text{constant} \\ &= \int_{x_{\min,1}}^{x_1} \int p_{2|1}(x_2|z_1) f^1(z_1, x_2) dx_2 dz_1 - \text{constant}, \end{aligned} \tag{5}$$

where $f^1(x_1, x_2) \equiv \frac{\partial f(x_1,x_2)}{\partial x_1}$ represents the *local effect* of x_1 on $f(\cdot)$ at (x_1, x_2), and $x_{\min,1}$ is some value chosen near the lower bound of the effective support of $p_1(\cdot)$, e.g., just below the smallest observation $\min\{x_{i,1} : i = 1, 2, \ldots, n\}$. Choice of $x_{\min,1}$ is not important, as it only affects the vertical translation of the ALE plot of $f_{1,ALE}(x_1)$ versus x_1, and the constant in (5) will be chosen to vertically center the plot.

The function $f_{1,ALE}(x_1)$ can be interpreted as the *accumulated local effects* of x_1 in the following sense. In (5), we calculate the local effect $f^1(x_1, x_2)$ of x_1 at $(x_1 = z_1, x_2)$, then average this local effect across all values of x_2 with weight $p_{2|1}(x_2|z_1)$, and then finally accumulate/integrate this averaged local effect over all values of z_1 up to x_1. As illustrated in Figure 2, when averaging the local effect $f^1(x_1, x_2)$ across x_2, the use of the conditional density $p_{2|1}(x_2|x_1)$, instead of the marginal density $p_2(x_2)$, avoids the extrapolation required in PD plots. The avoidance of extrapolation is similar to M plots, which also use the conditional density $p_{2|1}(x_2|x_1)$. However, by averaging (across x_2) and accumulating (up to x_1) the local effects via (5), as opposed to directly averaging $f(\cdot)$ via (3), ALE plots avoid the omitted nuisance variable bias that renders M plots of little use for assessing the main effects of the predictors. This relates closely to the

‡Regarding the terminology, plots of an estimate of $f_{1,M}(x_1)$ are often referred to as "marginal plots", because ignoring X_2 in this manner is equivalent to working with the joint distribution of (Y, X_1) after marginalizing across X_2. Unfortunately, plots of $\hat{f}_{1,PD}(x_1)$ are also sometimes referred to as "marginal plots" (e.g., in the **gbm** package for fitting boosted trees in R), presumably because the integral in (1) is with respect to the marginal distribution $p_2(x_2)$. In this paper, marginal plots will refer to how we have defined them above.

extrapolation 外推

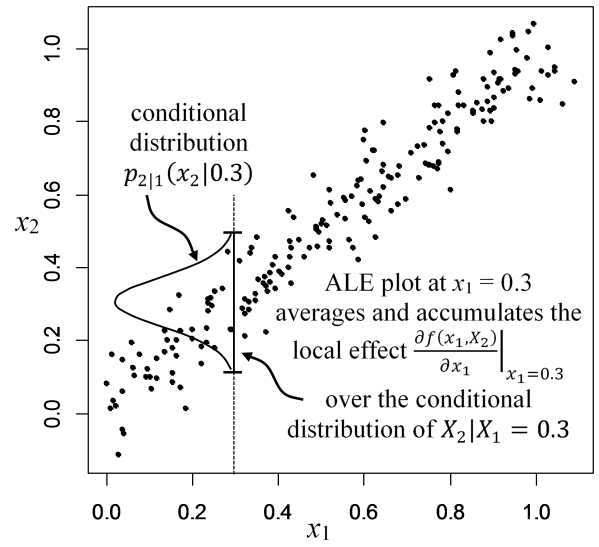

Fig. 2. Illustration of the computation of $f_{1,ALE}(x_1)$ at $x_1 = 0.3$

use of paired differences to block out nuisance factors in more general statistical settings, which we discuss in Section 5.3.

Methods also exist for visualizing the effects of predictors by plotting a collection of curves, rather than a single curve that represents some aggregate effect. Consider the effect of a single predictor X_j, and let $\mathbf{X}_{\setminus j}$ denote the other predictors. Conditioning plots (coplots) (Chambers and Hastie (1992); Cleveland (1993)), conditional response (CORE) plots (Cook (1995)), and individual conditional expectation (ICE) plots (Goldstein and Pitkin (2015)) plot quantities like $f(x_j, \mathbf{x}_{\setminus j})$ vs. x_j for a collection of discrete values of $\mathbf{x}_{\setminus j}$ (CORE and ICE plots), or similarly they plot $\mathbb{E}[f(x_j, \mathbf{X}_{\setminus j})|\mathbf{X}_{\setminus j} \in S_k]$ vs. x_j for each set S_k in some partition $\{S_k : k = 1, 2, \ldots\}$ of the space of $\mathbf{X}_{\setminus j}$. Such a collection of curves have more in common with interaction effect plots (as in Figure 10, later) than with main effect plots, for which one desires, by definition, a single aggregated curve.

The format of the remainder of the paper is as follows. In Section 2, we define the ALE main effects for individual predictors and the ALE second-order interaction effects for pairs of predictors. In Section 3 we present estimators of the ALE main and second-order interaction effects, which are conceptually straightforward and computationally efficient (much more efficient than PD plots), and we prove their consistency. We focus on main and second-order interaction effects and discuss general higher-order effects and their estimation in the appendices. In Section 4 we give examples that illustrate the ALE plots and, in particular, how they can produce correct results when PD plots are corrupted due to their reliance on extrapolation. In Section 5, we discuss interpretation of ALE plots and a number of their desirable properties and computational advantages, and we illustrate with a real data example. We also discuss their relation to functional ANOVA decompositions for dependent variables (e.g., Hooker (2007)) that have been developed to avoid the same extrapolation problem highlighted in Figure 1(a). ALE plots are far more computationally efficient and systematic to compute than functional ANOVA decompositions; and they yield a fundamentally different decomposition of $f(\mathbf{x})$ that is better suited for visualization of the effects of the predictors. Section 6 concludes the paper. We also provide as supplementary material an `R` package **`ALEPlot`** to implement ALE plots.

2. Definition of ALE Main and Second-Order Effects

In this section we define the ALE main effect functions for each predictor (Eq. (5) is a special case for $d = 2$ and differentiable $f(\cdot)$) and the ALE second-order effect functions for each pair of predictors. ALE plots are plots of estimates of these functions, and the estimators are defined in Section 3. We do not envision ALE plots being commonly used to visualize third- and high-order effects, since high-order effects are difficult to interpret and usually not as predominant as main and second-order effects. For this reason, and to simplify notation, we focus on main and second-order effects and relegate the definition of higher-order ALE effects to the appendices.

Throughout this section, we assume that p has compact support $\mathcal{S}$, and the support of p_j is the interval $\mathcal{S}_j = [x_{\min,j}, x_{\max,j}]$ for each $j \in \{1, 2, \ldots, d\}$. For each $K = 1, 2, \ldots,$ and $j \in \{1, 2, \ldots, d\}$, let $\mathcal{P}_j^K \equiv \{z_{k,j}^K : k = 0, 1, \ldots, K\}$ be a partition of $\mathcal{S}_j$ into K intervals with $z_{0,j}^K = x_{\min,j}$ and $z_{K,j}^K = x_{\max,j}$. Define $\delta_{j,K} \equiv \max\{|z_{k,j}^K - z_{k-1,j}^K| : k = 1, 2, \ldots, K\}$, which represents the fineness of the partition. For any $x \in \mathcal{S}_j$, define $k_j^K(x)$ to be the index of the interval of $\mathcal{P}_j^K$ into which x falls, i.e., $x \in (z_{k-1,j}^K, z_{k,j}^K]$ for $k = k_j^K(x)$. Let $\mathbf{X}_{\setminus j}$ denote the subset of $d-1$ predictors excluding X_j, i.e., $\mathbf{X}_{\setminus j} = (X_k : k = 1, 2, \ldots, d; k \neq j)$. The following definition is a generalization of (5) for a function $f(\cdot)$ that is not necessarily differentiable and for any $d \geq 2$. The generalization essentially replaces the derivative and integral in (5) with limiting forms of finite differences and summations, respectively.

Definition 1 (**Uncentered ALE Main Effect**). Consider any $j \in \{1, 2, \ldots, d\}$, and suppose the sequence of partitions $\{\mathcal{P}_j^K : K = 1, 2, \ldots\}$ is such that $\lim_{K\to\infty} \delta_{j,K} = 0$. When $f(\cdot)$ and p are such that the following limit exists and is independent of the particular sequence of partitions $\{\mathcal{P}_j^K : K = 1, 2, \ldots\}$ (see Theorem A.1 in Appendix A for sufficient conditions on the existence and uniqueness of the limit), we define the uncentered ALE main (aka first-order) effect function of X_j as (for $x_j \in \mathcal{S}_j$)

$$g_{j,ALE}(x_j) \equiv \lim_{K\to\infty} \sum_{k=1}^{k_j^K(x_j)} \mathbb{E}[f(z_{k,j}^K, \mathbf{X}_{\setminus j}) - f(z_{k-1,j}^K, \mathbf{X}_{\setminus j}) | X_j \in (z_{k-1,j}^K, z_{k,j}^K]]. \tag{6}$$

The following theorem, the proof of which is in Appendix A, states that for differentiable $f(\cdot)$, the uncentered ALE main effect of X_j in (6) has an equivalent but more revealing form that is analogous to (5).

Theorem 1 (**Uncentered ALE Main Effect for differentiable** $f(\cdot)$). Let $f^j(x_j, \mathbf{x}_{\setminus j}) \equiv \frac{\partial f(x_j, \mathbf{x}_{\setminus j})}{\partial x_j}$ denote the partial derivative of $f(\mathbf{x})$ with respect to x_j when the derivative exists. In Definition 1, suppose

(i) $f(x_j, \mathbf{x}_{\setminus j})$ is differentiable in x_j on $\mathcal{S}$,

(ii) $f^j(x_j, \mathbf{x}_{\setminus j})$ is continuous in $(x_j, \mathbf{x}_{\setminus j})$ on $\mathcal{S}$, and

(iii) $\mathbb{E}[f^j(X_j, \mathbf{X}_{\setminus j}) | X_j = z_j]$ is continuous in z_j on $\mathcal{S}_j$.

Then, for each $x_j \in \mathcal{S}_j$,

$$g_{j,ALE}(x_j) = \int_{x_{\min,j}}^{x_j} \mathbb{E}[f^j(X_j, \mathbf{X}_{\setminus j}) | X_j = z_j] dz_j. \tag{7}$$

(End of Theorem 1)

The (centered) ALE main effect of X_j, denoted by $f_{j,ALE}(x_j)$, is defined the same as $g_{j,ALE}(x_j)$ but centered so that $f_{j,ALE}(X_j)$ has a mean of zero with respect to the marginal distribution of X_j. That is,

$$\begin{aligned} f_{j,ALE}(x_j) &\equiv g_{j,ALE}(x_j) - \mathbb{E}[g_{j,ALE}(X_j)] \\ &= g_{j,ALE}(x_j) - \int p_j(z_j) g_{j,ALE}(z_j) dz_j. \end{aligned} \tag{8}$$

$S_j = [x_{\min,j}, x_{\max,j}]$

紧支撑？

S_j 是 interval，j 是某个 feature

$\mathcal{P}_j^K \equiv \{z_{k,j}^K : k = 0, 1, \ldots, K\}$

是 S_j 的一个分区，有 K 个间隔

$z_{0,j}^K = x_{\min,j}$ $z_{K,j}^K = x_{\max,j}$

S_j K 是分区的细度 (fineness)

对于任何 $x \in S_j$，$k_j^K(x)$

表示 x 所在的间隔 (interval)

Remark 1. The ALE plot function $f_{j,ALE}(x_j)$ attempts to quantify something quite similar to the PD plot function $f_{j,PD}(x_j)$ in (1) and can be interpreted in the same manner. For example, ALE plots and PD plots both have a desirable *additive recovery* property. That is, if $f(\mathbf{x}) = \sum_{j=1}^{d} f_j(x_j)$ is additive, then both $f_{j,ALE}(x_j)$ and $f_{j,PD}(x_j)$ are equal to the desired true effect $f_j(x_j)$, up to an additive constant. Hence, a plot of $f_{j,ALE}(x_j)$ vs. x_j correctly reveals the true effect of X_j on f, no matter how black-box the function f is. If second-order interaction effects are present in f, a similar additive recovery property holds for the ALE second-order interaction effects that we define next (see Section 5.3 for a more general additive recovery property that applies to interactions of any order). In spite of the similarities in the characteristics of f that they are designed to extract, the differences in $f_{j,ALE}(x_j)$ and $f_{j,PD}(x_j)$ lead to very different methods of estimation. As will be demonstrated in the later sections, the ALE plot functions are estimated in a far more computationally efficient manner that also avoids the extrapolation problem that renders PD plots unreliable with highly correlated predictors.

We next define the ALE second-order effects. For each pair of indices $\{j,l\} \subseteq \{1,2,\ldots,d\}$, let $\mathbf{X}_{\setminus\{j,l\}}$ denote the subset of $d-2$ predictors excluding $\{X_j, X_l\}$, i.e., $\mathbf{X}_{\setminus\{j,l\}} = (X_k : k = 1,2,\ldots,d; k \neq j; k \neq l)$. For general $f(\cdot)$, the uncentered ALE second-order effect of (X_j, X_l) is defined similarly to (6), except that we replace the 1-D finite-differences by 2-D second-order finite differences on the 2-D grid that is the Cartesian product of the 1-D partitions of $\mathcal{S}_j$ and $\mathcal{S}_l$, and the summation is over this 2-D grid.

Definition 2 (**Uncentered ALE Second-Order Effect**). Consider any pair $\{j,l\} \subseteq \{1,\ldots,d\}$ and corresponding sequences of partitions $\{\mathcal{P}_j^K : K = 1,2,\ldots\}$ and $\{\mathcal{P}_l^K : K = 1,2,\ldots\}$ such that $\lim_{K\to\infty} \delta_{j,K} = \lim_{K\to\infty} \delta_{l,K} = 0$. When $f(\cdot)$ and p are such that the following limit exists and is independent of the particular sequences of partitions, we define the uncentered ALE second-order effect function of (X_j, X_l) as (for $(x_j, x_l) \in \mathcal{S}_j \times \mathcal{S}_l$)

$$h_{\{j,l\},ALE}(x_j,x_l) \equiv \lim_{K\to\infty} \sum_{k=1}^{k_j^K(x_j)} \sum_{m=1}^{k_l^K(x_l)} \mathbb{E}[\Delta_f^{\{j,l\}}(K,k,m;\mathbf{X}_{\setminus\{j,l\}}) | X_j \in (z_{k-1,j}^K, z_{k,j}^K], X_l \in (z_{m-1,l}^K, z_{m,l}^K]], \tag{9}$$

where

$$\begin{aligned}\Delta_f^{\{j,l\}}(K,k,m;\mathbf{x}_{\setminus\{j,l\}}) = [f(z_{k,j}^K, z_{m,l}^K, \mathbf{x}_{\setminus\{j,l\}}) - f(z_{k-1,j}^K, z_{m,l}^K, \mathbf{x}_{\setminus\{j,l\}})] \\ -[f(z_{k,j}^K, z_{m-1,l}^K, \mathbf{x}_{\setminus\{j,l\}}) - f(z_{k-1,j}^K, z_{m-1,l}^K, \mathbf{x}_{\setminus\{j,l\}})]\end{aligned} \tag{10}$$

is the second-order finite difference of $f(\mathbf{x}) = f(x_j, x_l, \mathbf{x}_{\setminus\{j,l\}})$ with respect to (x_j, x_l) across cell $(z_{k-1,j}^K, z_{k,j}^K] \times (z_{m-1,l}^K, z_{m,l}^K]$ of the 2-D grid that is the Cartesian product of $\mathcal{P}_j^K$ and $\mathcal{P}_l^K$.

Analogous to Theorem 1, Theorem 2 (proved in Appendix A) states that for differentiable $f(\cdot)$, the uncentered ALE second-order effect of (X_j, X_l) in (9) has an equivalent integral form.

Theorem 2 (**Uncentered ALE Second-Order Effect for differentiable** $f(\cdot)$). Let $f^{\{j,l\}}(x_j, x_l, \mathbf{x}_{\setminus\{j,l\}}) \equiv \frac{\partial^2 f(x_j,x_l,\mathbf{x}_{\setminus\{j,l\}})}{\partial x_j \partial x_l}$ denote the second-order partial derivative of $f(\mathbf{x})$ with respect to x_j and x_l when the derivative exists. In Definition 2, suppose

(i) $f(x_j, x_l, \mathbf{x}_{\setminus\{j,l\}})$ is differentiable in (x_j, x_l) on $\mathcal{S}$,

(ii) $f^{\{j,l\}}(x_j, x_l, \mathbf{x}_{\setminus\{j,l\}})$ is continuous in $(x_j, x_l, \mathbf{x}_{\setminus\{j,l\}})$ on $\mathcal{S}$, and

(iii) $\mathbb{E}[f^{\{j,l\}}(X_j, X_l, \mathbf{X}_{\setminus\{j,l\}}) | X_j = z_j, X_l = z_l]$ is continuous in (z_j, z_l) on $\mathcal{S}_j \times \mathcal{S}_l$.

Then, for each $(x_j, x_l) \in \mathcal{S}_j \times \mathcal{S}_l$,

$$h_{\{j,l\},ALE}(x_j,x_l) \equiv \int_{x_{\min,l}}^{x_l} \int_{x_{\min,j}}^{x_j} \mathbb{E}[f^{\{j,l\}}(X_j, X_l, \mathbf{X}_{\setminus\{j,l\}}) | X_j = z_j, X_l = z_l] dz_j dz_l. \tag{11}$$

(End of Theorem 2)

The ALE second-order effect of (X_j, X_l), denoted by $f_{\{j,l\},ALE}(x_j, x_l)$, is defined the same as $h_{\{j,l\},ALE}(x_j, x_l)$ but "doubly-centered" so that $f_{\{j,l\},ALE}(X_j, X_l)$ has a mean of zero with respect to the marginal distribution of (X_j, X_l) *and* so that the ALE main effects of X_j and X_l on $f_{\{j,l\},ALE}(X_j, X_l)$ are both zero. The latter centering is accomplished by subtracting from $h_{\{j,l\},ALE}(x_j, x_l)$ its uncentered ALE main effects via

$$\begin{aligned} g_{\{j,l\},ALE}(x_j, x_l) \equiv\ & h_{\{j,l\},ALE}(x_j, x_l) \\ & - \lim_{K\to\infty} \sum_{k=1}^{k_j^K(x_j)} \mathbb{E}[h_{\{j,l\},ALE}(z_{k,j}^K, X_l) - h_{\{j,l\},ALE}(z_{k-1,j}^K, X_l) | X_j \in (z_{k-1,j}^K, z_{k,j}^K]] \\ & - \lim_{K\to\infty} \sum_{k=1}^{k_l^K(x_l)} \mathbb{E}[h_{\{j,l\},ALE}(X_j, z_{k,l}^K) - h_{\{j,l\},ALE}(X_j, z_{k-1,l}^K) | X_l \in (z_{k-1,l}^K, z_{k,l}^K]]. \end{aligned} \tag{12}$$

By Theorem 1, for differentiable f, (12) is equivalent to

$$\begin{aligned} g_{\{j,l\},ALE}(x_j, x_l) \equiv\ & h_{\{j,l\},ALE}(x_j, x_l) - \int_{x_{\min,j}}^{x_j} \mathbb{E}[\frac{\partial h_{\{j,l\},ALE}(X_j, X_l)}{\partial X_j} | X_j = z_j] dz_j \\ & - \int_{x_{\min,l}}^{x_l} \mathbb{E}[\frac{\partial h_{\{j,l\},ALE}(X_j, X_l)}{\partial X_l} | X_l = z_l] dz_l \\ =\ & h_{\{j,l\},ALE}(x_j, x_l) - \int_{x_{\min,j}}^{x_j} \int p_{l|j}(z_l|z_j) \frac{\partial h_{\{j,l\},ALE}(z_j, z_l)}{\partial z_j} dz_l dz_j \\ & - \int_{x_{\min,l}}^{x_l} \int p_{j|l}(z_j|z_l) \frac{\partial h_{\{j,l\},ALE}(z_j, z_l)}{\partial z_l} dz_j dz_l. \end{aligned} \tag{13}$$

The final centering is accomplished by taking

$$\begin{aligned} f_{\{j,l\},ALE}(x_j, x_l) \equiv\ & g_{\{j,l\},ALE}(x_j, x_l) - \mathbb{E}[g_{\{j,l\},ALE}(X_j, X_l)] \\ =\ & g_{\{j,l\},ALE}(x_j, x_l) - \int\int p_{\{j,l\}}(z_j, z_l) g_{\{j,l\},ALE}(z_j, z_l) dz_j dz_l. \end{aligned} \tag{14}$$

It can be verified that $f_{\{j,l\},ALE}(x_j, x_l)$ is centered in the sense that the ALE main effects of X_j and X_l on $f_{\{j,l\},ALE}(x_j, x_l)$ are both zero (see Appendix C for a formal proof of a related but more general result).

If we define the zero-order effect for any function of $\mathbf{X}$ as its expected value with respect to p, then we can view the ALE first-order effect of X_j as being obtained by first calculating its uncentered first-order effect (6), and then for the resulting function, subtracting its zero-order effect. Likewise, the ALE second-order effect of (X_j, X_l) is obtained by first calculating the uncentered second-order effect (9), then for the resulting function, subtracting both of its first-order effects of X_j and of X_l, and then for this resulting function, subtracting its zero-order effect. The ALE higher-order effects are defined analogously in Appendix B. The uncentered higher-order effect is first calculated, and then all lower-order effects are sequentially calculated and subtracted one order at a time, until the final result has all lower-order effects that are identically zero.

Remark 2. In Appendix B we define ALE higher-order effect functions $f_{J,ALE}(\mathbf{x}_J)$ for $|J| > 2$, where $|J|$ denotes the cardinality of the set of predictor indices J. Appendix C shows that this leads to a functional-ANOVA-like decomposition of f via

$$f(\mathbf{x}) = \sum_{j=1}^{d} f_{j,ALE}(x_j) + \sum_{j=1}^{d} \sum_{l=j+1}^{d} f_{\{j,l\},ALE}(x_j, x_l) + \sum_{J\subseteq\{1,2,\ldots,d\},|J|\geq 3} f_{J,ALE}(\mathbf{x}_J).$$

This ALE decomposition has a certain orthogonality-like property, which we contrast with other functional ANOVA decompositions in Section 5.5.

Remark 3. The ALE function definitions in this section apply to predictor distributions p_j that are continuous numerical with compact support. For discrete p_j, one could consider modifying (6) and (9) by using a fixed finite partition whose interval endpoints coincide with the support of p_j. We do not develop this, however, because our focus is on *estimation* and interpretation of the ALE effects, and the estimators in the following section are meaningful for either continuous or discrete p_j. In the case that X_j is a nominal categorical predictor, one must decide how to order its categories prior to estimating its ALE effect (which requires differencing f across neighboring categories of X_j). In Appendix E, we discuss a strategy for this that we have found to work well in practice.

3. Estimation of $f_{j,ALE}(x_j)$ and $f_{\{j,l\},ALE}(x_j, x_l)$

In Appendix D we briefly describe how to estimate the ALE higher-order effect $f_{J,ALE}(\mathbf{x}_J)$ for a general subset $J \subseteq \{1, 2, \ldots, d\}$ of predictor indices. Our focus in this section is on estimating the first-order ($|J| = 1$) and second-order ($|J| = 2$) effects, since these are the most common and useful (i.e., interpretable).

As an overview, the estimate $\hat{f}_{J,ALE}$ is obtained by computing estimates of the quantities in Eqs. (6)—(14) for $J = j$ (a single index) or for $J = \{j, l\}$ (a pair of indices). For the estimates we (i) replace the sequence of partitions in (6) (or (9)) by some appropriate fixed partition of the sample range of $\{\mathbf{x}_{i,J} : i = 1, \ldots, n\}$ and (ii) replace the conditional expectations in (6) (or (9)) by sample averages across $\{\mathbf{x}_{i,\setminus J} : i = 1, 2, \ldots, n)\}$, conditioned on $\mathbf{x}_{i,J}$ falling into the corresponding interval/cell of the partition. In the preceding, $\mathbf{x}_{i,J} = (x_{i,j} : j \in J)$ and $\mathbf{x}_{i,\setminus J} = (x_{i,j} : j = 1, 2, \ldots, d; j \notin J)$ denote the ith observation of the subsets of predictors $\mathbf{X}_J$ and $\mathbf{X}_{\setminus J}$, respectively.

More specifically, for each $j \in \{1, 2, \ldots, d\}$, let $\{N_j(k) = (z_{k-1,j}, z_{k,j}] : k = 1, 2, \ldots, K\}$ be a sufficiently fine partition of the sample range of $\{x_{i,j} : i = 1, 2, \ldots, n\}$ into K intervals. Since the estimator is computed for a fixed K, we have omitted it as a superscript on the partition, with the understanding that the partition implicitly depends on K. In all of our examples later in the paper, we chose $z_{k,j}$ as the $\frac{k}{K}$ quantile of the empirical distribution of $\{x_{i,j} : i = 1, 2, \ldots, n\}$ with $z_{0,j}$ chosen just below the smallest observation, and $z_{K,j}$ chosen as the largest observation. Figure 3 illustrates the notation and concepts in computing the ALE main effect estimator $\hat{f}_{j,ALE}(x_j)$ for the first predictor $j = 1$ for the case of $d = 2$ predictors. For $k = 1, 2, \ldots, K$, let $n_j(k)$ denote the number of training observations that fall into the kth interval $N_j(k)$, so that $\sum_{k=1}^{K} n_j(k) = n$. For a particular value x of the predictor x_j, let $k_j(x)$ denote the index of the interval into which x falls, i.e., $x \in (z_{k_j(x)-1,j}, z_{k_j(x),j}]$.

For general d, to estimate the main effect function $f_{j,ALE}(\cdot)$ of a predictor X_j, we first compute an estimate of the uncentered effect $g_{j,ALE}(\cdot)$ defined in (6), which is

$$\hat{g}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{\{i:x_{i,j} \in N_j(k)\}} [f(z_{k,j}, \mathbf{x}_{i,\setminus j}) - f(z_{k-1,j}, \mathbf{x}_{i,\setminus j})] \tag{15}$$

for each $x \in (z_{0,j}, z_{K,j}]$. Analogous to (8), the ALE main effect estimator $\hat{f}_{j,ALE}(\cdot)$ is then obtained by subtracting an estimate of $\mathbb{E}[g_{j,ALE}(X_j)]$ from (15), i.e.,

$$\hat{f}_{j,ALE}(x) = \hat{g}_{j,ALE}(x) - \frac{1}{n}\sum_{i=1}^{n} \hat{g}_{j,ALE}(x_{i,j}) = \hat{g}_{j,ALE}(x) - \frac{1}{n}\sum_{k=1}^{K} n_j(k)\hat{g}_{j,ALE}(z_{k,j}). \tag{16}$$

To estimate the ALE second-order effect of a pair of predictors (X_j, X_l), we partition the sample range of $\{(x_{i,j}, x_{i,l}) : i = 1, 2, \ldots, n\}$ into a grid of K^2 rectangular cells obtained as the Cartesian product of the individual one-dimensional partitions. Figure 4 illustrates the

$x_{i,J}$：i是样本下标

式K被省去不再显示

$z_{k,j}$是$\frac{k}{K}$个分位数 quantiles

$n_j(k)$表示 kth个间隔$N_j(k)$中样本数量

对于 feature j的任一个值 x，

$k_j(x)$表示该值落入的间隔的下标

→ 取所有 difference的平均值作为 local effect

Constant：

$\sum_{k=1}^{K}$ 间隔样本数 × 该间隔非中心化ALE / 总样本数

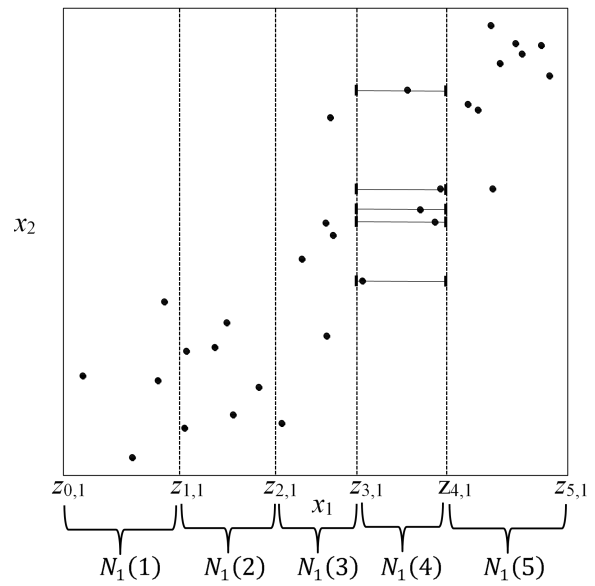

Fig. 3. Illustration of the notation and concepts in computing the ALE main effect estimator $\hat{f}_{j,ALE}(x_j)$ for $j = 1$ with $d = 2$ predictors. The bullets are a scatterplot of $\{(x_{i,1}, x_{i,2}) : i = 1, 2, \ldots, n\}$ for $n = 30$ training observations. The range of $\{x_{i,1} : i = 1, 2, \ldots, n\}$ is partitioned into $K = 5$ intervals $\{N_1(k) = (z_{k-1,1}, z_{k,1}] : k = 1, 2, \ldots, 5\}$ (in practice, K should usually be chosen much larger than 5). The numbers of training observations falling into the 5 intervals are $n_1(1) = 4$, $n_1(2) = 6$, $n_1(3) = 6$, $n_1(4) = 5$, and $n_1(5) = 9$. The horizontal line segments shown in the $N_1(4)$ region are the segments across which the finite differences $f(z_{4,j}, \mathbf{x}_{i,\setminus j}) - f(z_{3,j}, \mathbf{x}_{i,\setminus j})$ are calculated and then averaged in the inner summand of Eq. (15) for $k = 4$ and $j = 1$.

notation and concepts. Let (k, m) (with k and m integers between 1 and K) denote the indices into the grid of rectangular cells with k corresponding to x_j and m corresponding to x_l. In analogy with $N_j(k)$ and $n_j(k)$ defined in the context of estimating $f_{j,ALE}(\cdot)$, let $N_{\{j,l\}}(k, m) = N_j(k) \times N_l(m) = (z_{k-1,j}, z_{k,j}] \times (z_{m-1,l}, z_{m,l}]$ denote the cell associated with indices (k, m), and let $n_{\{j,l\}}(k, m)$ denote the number of training observations that fall into cell $N_{\{j,l\}}(k, m)$, so that $\sum_{k=1}^{K} \sum_{m=1}^{K} n_{\{j,l\}}(k, m) = n$.

To estimate $f_{\{j,l\},ALE}(x_j, x_l)$, we first estimate the uncentered effect $h_{\{j,l\},ALE}(x_j, x_l)$ defined in (9) by

$$\hat{h}_{\{j,l\},ALE}(x_j, x_l) = \sum_{k=1}^{k_j(x_j)} \sum_{m=1}^{k_l(x_l)} \frac{1}{n_{\{j,l\}}(k, m)} \sum_{\{i:\mathbf{x}_{i,\{j,l\}} \in N_{\{j,l\}}(k,m)\}} \Delta_f^{\{j,l\}}(K, k, m; \mathbf{x}_{i,\setminus\{j,l\}}) \tag{17}$$

for each $(x_j, x_l) \in (z_{0,j}, z_{K,j}] \times (z_{0,l}, z_{K,l}]$. In (17), $\Delta_f^{\{j,l\}}(K, k, m; \mathbf{x}_{i,\setminus\{j,l\}})$ is the second-order finite difference defined in (10), evaluated at the ith observation $\mathbf{x}_{i,\setminus\{j,l\}}$, i.e.

$$\begin{aligned} \Delta_f^{\{j,l\}}(K, k, m; \mathbf{x}_{i,\setminus\{j,l\}}) = [f(z_{k,j}, z_{m,l}, \mathbf{x}_{i,\setminus\{j,l\}}) - f(z_{k-1,j}, z_{m,l}, \mathbf{x}_{i,\setminus\{j,l\}})] \\ -[f(z_{k,j}, z_{m-1,l}, \mathbf{x}_{i,\setminus\{j,l\}}) - f(z_{k-1,j}, z_{m-1,l}, \mathbf{x}_{i,\setminus\{j,l\}})] \end{aligned} \tag{18}$$

Analogous to (12), we next compute estimates of the ALE main effects of X_j and X_l for the function $\hat{h}_{\{j,l\},ALE}(x_j, x_l)$ and then subtract these from $\hat{h}_{\{j,l\},ALE}(x_j, x_l)$ to give an estimate of

$g_{\{j,l\},ALE}(x_j, x_l)$:

$$
\begin{aligned}
&\hat{g}_{\{j,l\},ALE}(x_j, x_l) \\
=&\hat{h}_{\{j,l\},ALE}(x_j, x_l) - \sum_{k=1}^{k_j(x_j)} \frac{1}{n_j(k)} \sum_{\{i:x_{i,j} \in N_j(k)\}} [\hat{h}_{\{j,l\},ALE}(z_{k,j}, x_{i,l}) - \hat{h}_{\{j,l\},ALE}(z_{k-1,j}, x_{i,l})] \\
&- \sum_{m=1}^{k_l(x_l)} \frac{1}{n_l(m)} \sum_{\{i:x_{i,l} \in N_l(m)\}} [\hat{h}_{\{j,l\},ALE}(x_{i,j}, z_{m,l}) - \hat{h}_{\{j,l\},ALE}(x_{i,j}, z_{m-1,l})] \\
=&\hat{h}_{\{j,l\},ALE}(x_j, x_l) - \sum_{k=1}^{k_j(x_j)} \frac{1}{n_j(k)} \sum_{m=1}^{K} n_{j,l}(k,m) [\hat{h}_{\{j,l\},ALE}(z_{k,j}, z_{m,l}) - \hat{h}_{\{j,l\},ALE}(z_{k-1,j}, z_{m,l})] \\
&- \sum_{m=1}^{k_l(x_l)} \frac{1}{n_l(m)} \sum_{k=1}^{K} n_{\{j,l\}}(k,m) [\hat{h}_{\{j,l\},ALE}(z_{k,j}, z_{m,l}) - \hat{h}_{\{j,l\},ALE}(z_{k,j}, z_{m-1,l})].
\end{aligned}
\tag{19}
$$

Finally, analogous to (14), we estimate $f_{\{j,l\},ALE}(x_j, x_l)$ by subtracting an estimate of

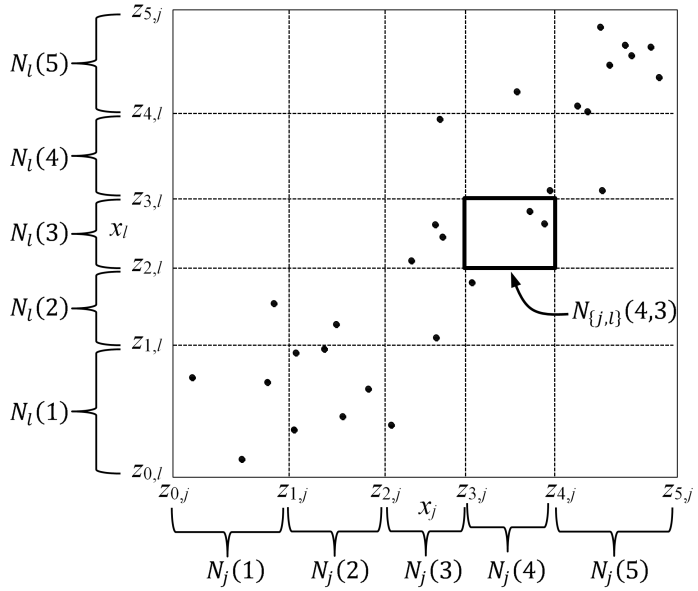

Fig. 4. Illustration of the notation used in computing the ALE second-order effect estimator $\hat{f}_{\{j,l\},ALE}(x_j, x_l)$ for $K = 5$. The ranges of $\{x_{i,j} : i = 1, 2, \ldots, n\}$ and $\{x_{i,l} : i = 1, 2, \ldots, n\}$ are each partitioned into 5 intervals, and their Cartesian product forms the grid of rectangular cells $\{N_{\{j,l\}}(k,m) = N_j(k) \times N_l(m) : k = 1, 2, \ldots, 5; m = 1, 2, \ldots, 5\}$. The cell with bold borders is the region $N_{\{j,l\}}(4,3)$. The second-order finite differences $\Delta_f^{\{j,l\}}(K, k, m; \mathbf{x}_{i,\setminus\{j,l\}})$ in Eq. (18) for $(k,m) = (4,3)$ are calculated across the corners of this cell. In the inner summation of Eq. (17), these differences are then averaged over the $n_{\{j,l\}}(4,3) = 2$ observations in region $N_{\{j,l\}}(4,3)$.

$\mathbb{E}[\hat{g}_{\{j,l\},ALE}(X_j, X_l)]$ from (19), which gives

$$\begin{aligned}\hat{f}_{\{j,l\},ALE}(x_j, x_l) &= \hat{g}_{\{j,l\},ALE}(x_j, x_l) - \frac{1}{n}\sum_{i=1}^{n} \hat{g}_{\{j,l\},ALE}(x_{i,j}, x_{i,l}) \\ &= \hat{g}_{\{j,l\},ALE}(x_j, x_l) - \frac{1}{n}\sum_{k=1}^{K}\sum_{m=1}^{K} n_{\{j,l\}}(k,m)\hat{g}_{\{j,l\},ALE}(z_{k,j}, z_{m,l}).\end{aligned} \tag{20}$$

Theorems 3 and 4 in Appendix A show that, under mild conditions, (16) and (20) are consistent estimators of the ALE main effect (8) of X_j and ALE second-order effect (14) of (X_j, X_l), respectively.

ALE plots are plots of the ALE effect estimates $\hat{f}_{j,ALE}(x_j)$ and $\hat{f}_{\{j,l\},ALE}(x_j, x_l)$ versus the predictors involved. ALE plots have substantial computational advantages over PD plot, which we discuss in Section 5.4. In addition, ALE plots can produce reliable estimates of the main and interaction effects in situations where PD plots break down, which we illustrate with examples in the next section, as well as an example on real data in Section 5.1.

4. Toy Examples Illustrating when ALE Plots are Reliable but PD Plots Break Down

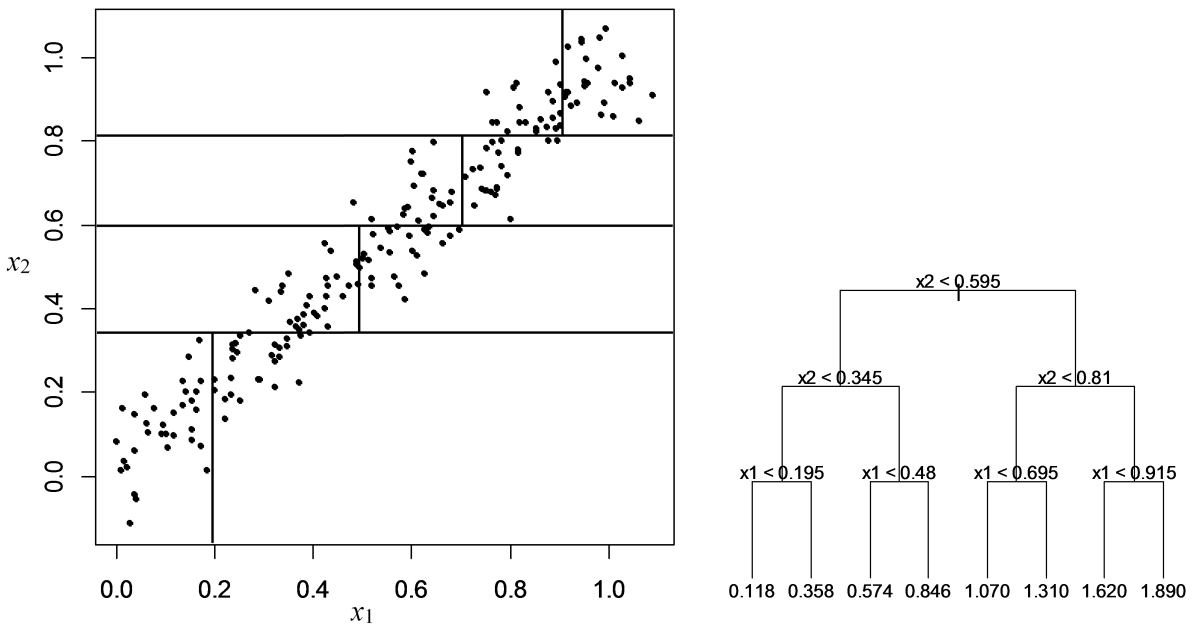

Fig. 5. Depiction of the first eight splits in the tree fitted to the Example 1 data. The left panel is a scatterplot of x_2 vs. x_1 showing splits corresponding to the truncated tree in the right panel.

Example 1. This example was introduced in Section 1. For this example, $d = 2$, $n = 200$, and (X_1, X_2) follows a uniform distribution along a segment of the line $x_2 = x_1$ with independent $N(0, 0.05^2)$ variables added to both predictors. Figure 5 shows a scatter plot of X_2 vs. X_1. The true response was generated according to the noiseless model $Y = X_1 + X_2^2$ for the 200 training observations in Figure 5, to which we fit a tree using the **tree** package of `R` (Ripley (2015)). The tree was overgrown and then pruned back to have 100 leaf nodes, which was approximately the optimal number of leaf nodes according to a cross-validation error sum of squares criterion. Notice that the optimal size tree is relatively large, because the response here is a deterministic function $X_1 + X_2^2$ of the predictors with no response observation error. The first eight splits of the fitted tree $f(\mathbf{x})$ are also depicted in Figure 5. Figure 6 shows main effect PD plots, M plots,