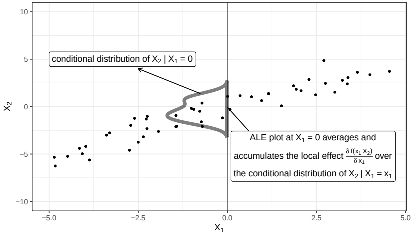
Interpretable Machine Learning

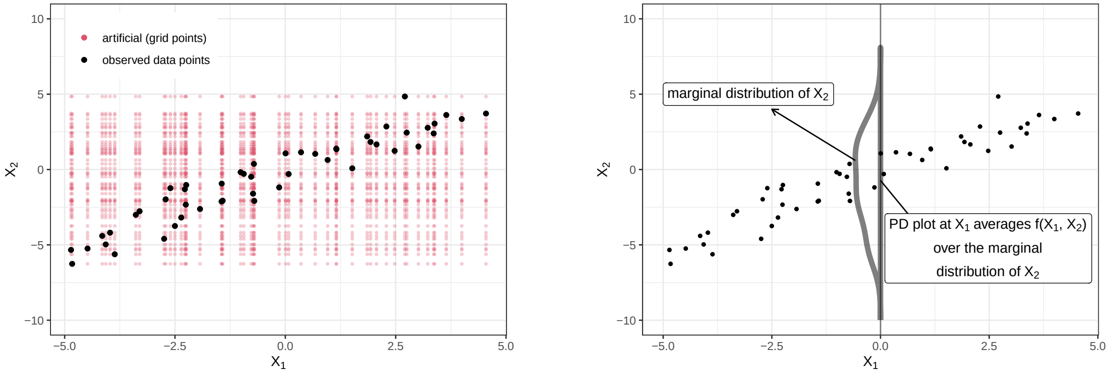
Accumulated Local Effect (ALE) plot



Learning goals

- PD plots and its extrapolation issue
- M plots and its omitted-variable bias
- Understand ALE plots

MOTIVATION - CORRELATED FEATURES

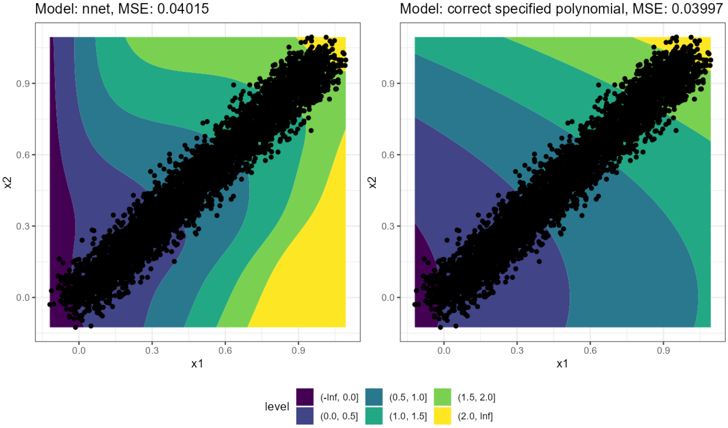


- PD plots **average over predictions** of artificial points that are out of distribution / unlikely (red)
  $\Rightarrow$ Can lead to misleading / biased interpretations, especially if model also contains interactions
- Not wanted if interest is to interpret effects within data distribution

MOTIVATION - CORRELATED FEATURES

Example: Fit a NN to 5000 simulated data points with $x \sim Unif(0,1)$, $\epsilon \sim N(0, 0.2)$ and

$$y = x_1 + x_2^2 + \epsilon, \text{ where } x_1 = x + \epsilon_1,\ x_2 = x + \epsilon_2 \text{ and } \epsilon_1, \epsilon_2 \sim N(0, 0.05).$$

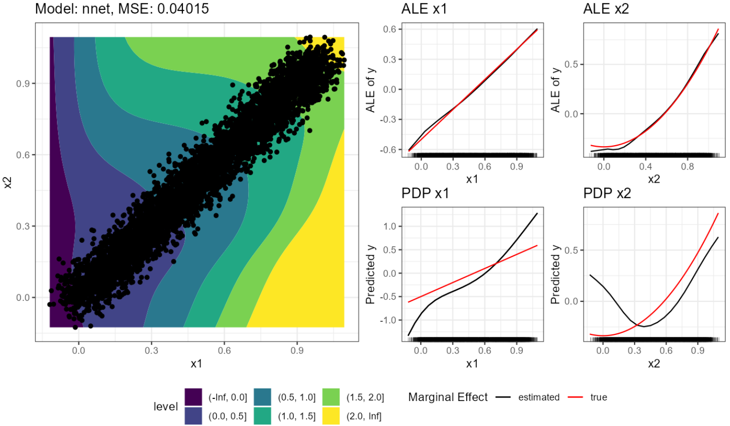


- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)

MOTIVATION - CORRELATED FEATURES

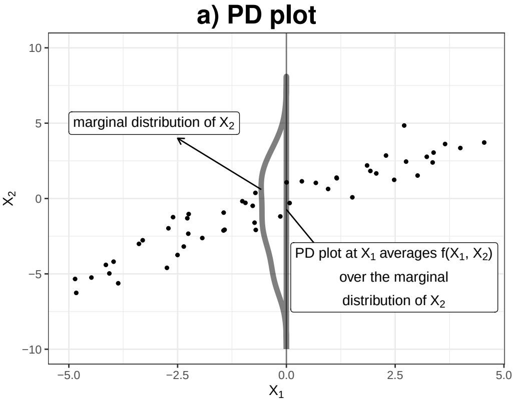
Example: Fit a NN to 5000 simulated data points with $x \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2)$ and

$$y = x_1 + x_2^2 + \epsilon, \text{ where } x_1 = x + \epsilon_1,\ x_2 = x + \epsilon_2 \text{ and } \epsilon_1, \epsilon_2 \sim N(0, 0.05).$$



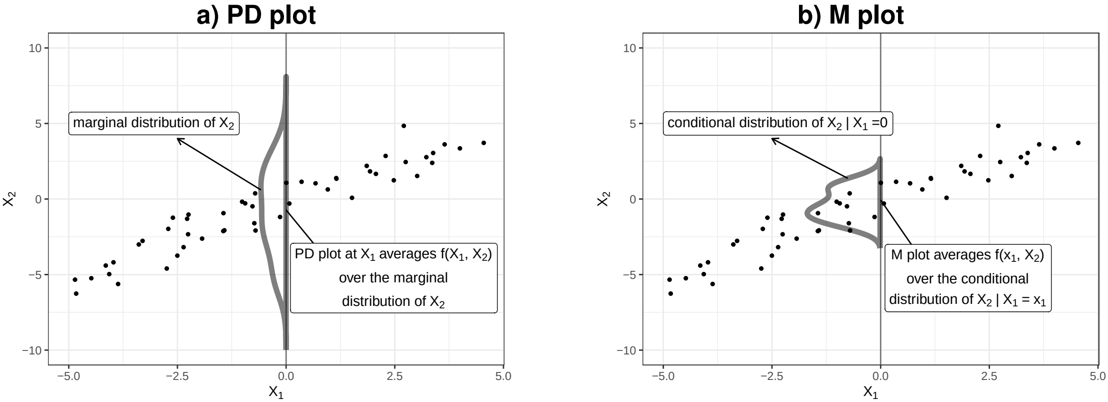
- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)
- ALE in line with ground truth
- PDP does not reflect ground truth effects of DGP well
  $\Rightarrow$ Due to interactions and averaging of points outside data distribution

M PLOT VS. PD PLOT



a) PD plot $\mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(x_1, \mathbf{x}_2)\right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1, \mathbf{x}_2^{(i)})$

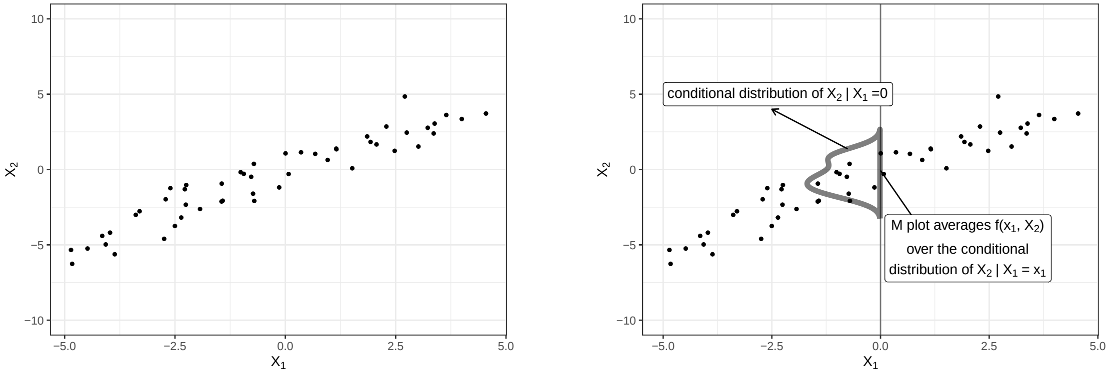
M PLOT VS. PD PLOT



a) PD plot $\mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(x_1, \mathbf{x}_2)\right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_1, \mathbf{x}_2^{(i)})$

b) M plot $\mathbb{E}_{\mathbf{x}_2|\mathbf{x}_1}\left(\hat{f}(x_1, \mathbf{x}_2)\middle|\mathbf{x}_1\right)$ is estimated by $\hat{f}_{1,M}(x_1) = \frac{1}{|N(x_1)|}\sum_{i\in N(x_1)} \hat{f}(x_1, \mathbf{x}_2^{(i)})$, where index set $N(x_1) = \{i : x_1^{(i)} \in [x_1 - \epsilon, x_1 + \epsilon]\}$ refers to observations with feature value close to x_1.

M PLOT VS. PD PLOT



- M plots average predictions over conditional distribution (e.g., $\mathbb{P}(\mathbf{x}_2|x_1)$)
  $\Rightarrow$ Averaging predictions close to data distribution avoid extrapolation issues
- **But:** M plots suffer from omitted-variable bias (OVB)
  - They contain effects of other dependent features
  - Useless in assessing a feature's marginal effect if feature dependencies are present

M PLOT VS. PD PLOT - OVB EXAMPLE

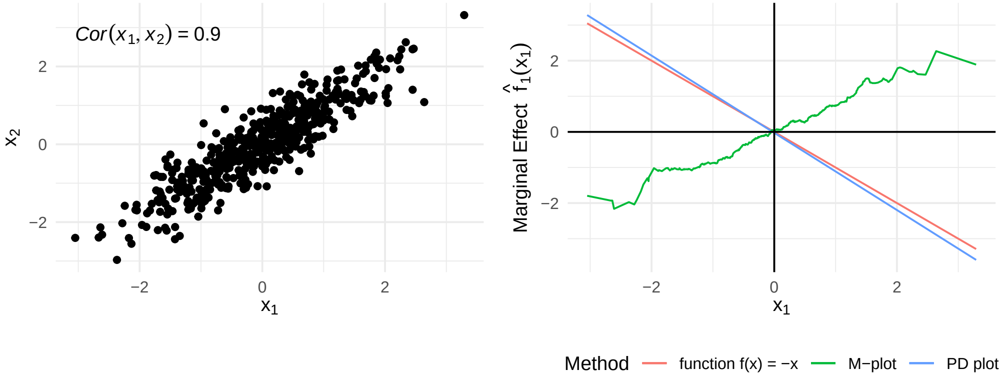


Illustration: Fit LM on 500 i.i.d. observations with features $x_1, x_2 \sim N(0,1)$, $Cor(x_1, x_2) = 0.9$ and

$$y = -x_1 + 2 \cdot x_2 + \epsilon, \ \epsilon \sim N(0,1).$$

Results: M plot of x_1 also includes marginal effect of all other dependent features (here: x_2)

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1 x_2$$

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1 x_2$$

- Partial derivative of $\hat{f}$ w.r.t. x_1: $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- Partial derivative of $\hat{f}$ w.r.t. x_1: $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$
- Integral of partial derivative ($z_0 = \min(x_1)$):

$$\int_{z_0}^{x} \frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} dx_1 = [2x_1 - 4x_1x_2]_{z_0}^{x}$$

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- Partial derivative of $\hat{f}$ w.r.t. x_1: $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$
- Integral of partial derivative ($z_0 = \min(x_1)$):

$$\int_{z_0}^{x} \frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} dx_1 = [2x_1 - 4x_1x_2]_{z_0}^{x}$$

- We removed the main effect of x_2, which was our goal

ACCUMULATED LOCAL EFFECTS (ALE) ▸ Apley, Zhu (2020)

ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$

ACCUMULATED LOCAL EFFECTS (ALE) ‣ Apley, Zhu (2020)

ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$
2. averaging local effects over conditional distribution $\mathbb{P}(\mathbf{x}_{-S} | x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation issue

ACCUMULATED LOCAL EFFECTS (ALE) ▸ Apley, Zhu (2020)

ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$
2. averaging local effects over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation issue
3. integrating averaged local effects up to a specific value $x \sim \mathbb{P}(x_S)$
   $\Rightarrow$ Accumulates local effects to estimate global main effect of x_S
   $\Rightarrow$ Avoids OVB issue as other unwanted main effects were removed in (1)

FIRST ORDER ALE

- Let x_S be feature of interest with $z_0 = \min(x_S)$ and $\mathbf{x}_{-S}$ all other features (complement of S)
- Uncentered first order ALE $\tilde{f}_{S,ALE}(x)$ at feature value $x \sim \mathbb{P}(x_S)$ is defined as:

$$\tilde{f}_{S,ALE}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S}}_{(2)} \left(\underbrace{\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}}_{(1)} \middle| x_S = z_S \right) dz_S$$

FIRST ORDER ALE

- Let x_S be feature of interest with $z_0 = \min(x_S)$ and $\mathbf{x}_{-S}$ all other features (complement of S)
- Uncentered first order ALE $\tilde{f}_{S,ALE}(x)$ at feature value $x \sim \mathbb{P}(x_S)$ is defined as:

$$\tilde{f}_{S,ALE}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S}}_{(2)} \left(\underbrace{\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}}_{(1)} \Bigg| x_S = z_S \right) dz_S$$

- Substract average of uncentered ALE curve (constant) to obtain centered ALE curve $f_{S,ALE}(x)$ with zero mean regarding marginal distribution of feature of interest x_S:

$$f_{S,ALE}(x) = \tilde{f}_{S,ALE}(x) - \underbrace{\int_{-\infty}^{\infty} \tilde{f}_{S,ALE}(x_S)\, d\mathbb{P}(x_S)}_{:= constant}$$

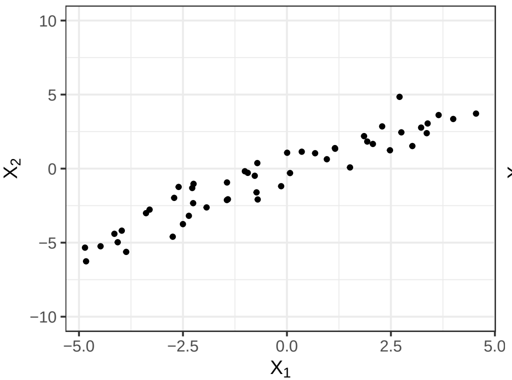
ALE ESTIMATION

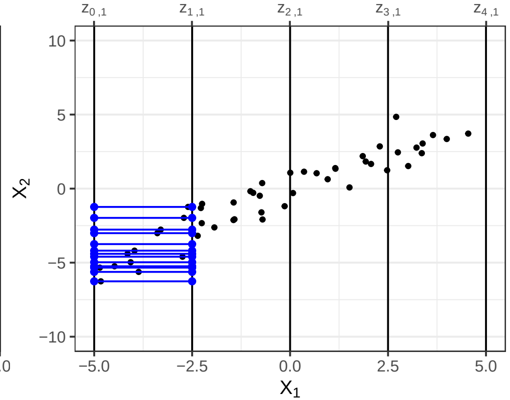
- Partial derivatives not useful for all models (e.g., tree-based methods as random forests)
- Approximate partial derivatives by finite differences of predictions within K intervals for $\mathbf{x}_S$:

$$\begin{aligned} x \in [\min(\mathbf{x}_S), \max(\mathbf{x}_S)] \iff & x \in [z_{0,S}, z_{1,S}] \\ & \vee x \in \,]z_{1,S}, z_{2,S}] \\ & \quad \dots \\ & \vee x \in \,]z_{K-1,S}, z_{K,S}] \end{aligned}$$

- A simple way to create K intervals for feature $\mathbf{x}_S$ is to use its quantile distribution with $K-1$ quantiles as interval bounds $z_{1,S}, \dots, z_{K-1,S}$ (not counting the 0% and 100% quantiles)

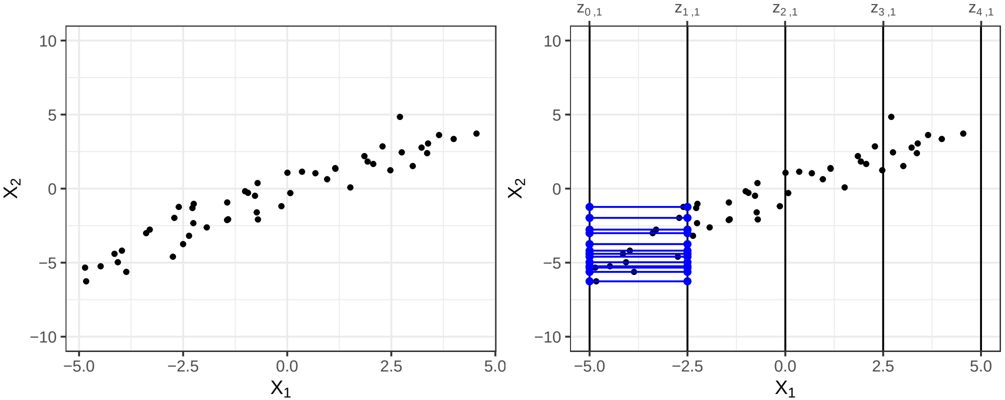
2-D ILLUSTRATION





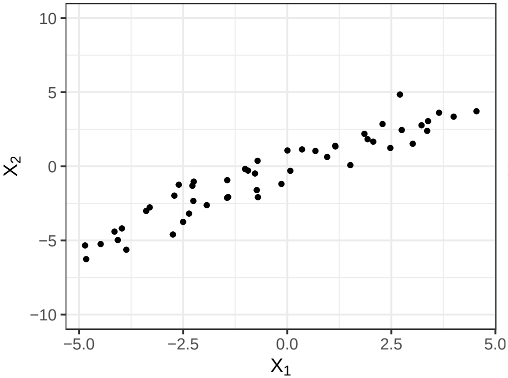
- Divide feature of interest into intervals (vertical lines)
- For all points within an interval, compute **prediction difference** when we replace feature value with upper/lower interval bound (blue points) while keeping other feature values unchanged
- These **finite differences** (approximate local effect) are accumulated & centered $\Rightarrow$ ALE plot

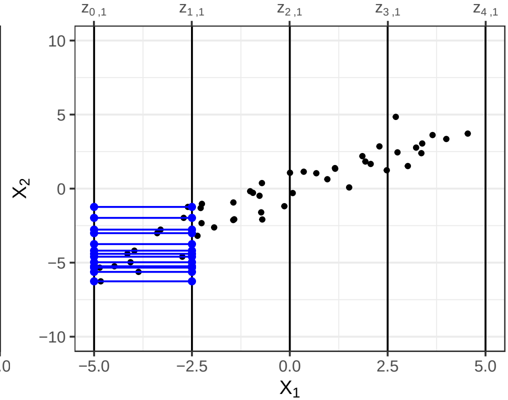
2-D ILLUSTRATION



- For $\mathbf{x}^{(i)} = (x_S^{(i)}, \mathbf{x}_{-S}^{(i)})$, value $x_S^{(i)}$ is located within k-th interval of $\mathbf{x}_S$ ($x_S^{(i)} \in]z_{k-1,S}, z_{k,S}]$)
- Replace $x_S^{(i)}$ by upper/lower interval bound while all other feature values $\mathbf{x}_{-S}^{(i)}$ are kept constant
- Finite differences correspond to $\hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$

2-D ILLUSTRATION





- Estimate local effect of $\mathbf{x}_S$ within each interval by averaging all observation-wise finite differences $\hat{=}$ Approximation of inner integral that integrates over local effects w.r.t. $\mathbb{P}(\mathbf{x}_{-S}|z_S)$.
- Sum up local effects of all intervals up to point of interest $\hat{=}$ Estimates outer integral

ALE ESTIMATION: FORMULA

- Estimated uncentered first order ALE $\hat{\tilde{f}}_{S,ALE}(x)$ at point x:

$$\hat{\tilde{f}}_{S,ALE}(x) = \sum_{k=1}^{k_S(x)} \frac{1}{n_S(k)} \sum_{i:\ x_S^{(i)} \in \,]z_{k-1,S}, z_{k,S}]} \left[\hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})\right]$$

- $k_S(x)$ denotes the interval index a feature value $x \in \mathbf{x}_S$ falls in
- $n_S(k)$ denotes the number of observations inside the k-th interval of $\mathbf{x}_S$
- Substract average of estimated uncentered ALE to obtain centered ALE estimate:

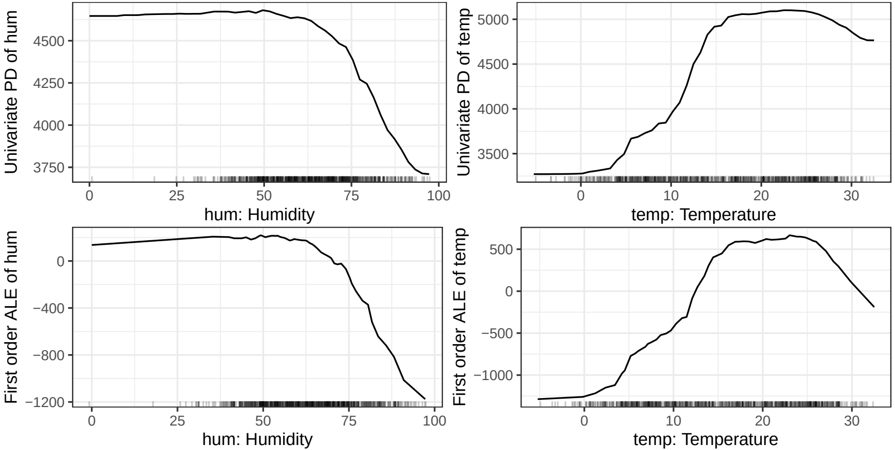
$$\hat{f}_{S,ALE}(x) = \hat{\tilde{f}}_{S,ALE}(x) - \frac{1}{n} \sum_{i=1}^{n} \hat{\tilde{f}}_{S,ALE}(x_S^{(i)})$$

ALE ESTIMATION: ALGORITHM

1. Create K intervals for value range of $\mathbf{x}_S$
2. Repeat for each interval:
   - Replace observation's feature value $x_S^{(i)}$ with upper/lower interval bound for each observation inside k-th interval
   - Compute observation-wise finite difference inside k-th interval and average them to estimate interval-wise local effects
3. Accumulate interval-wise local effects up to value of interest x to estimate uncentered ALE and then center it

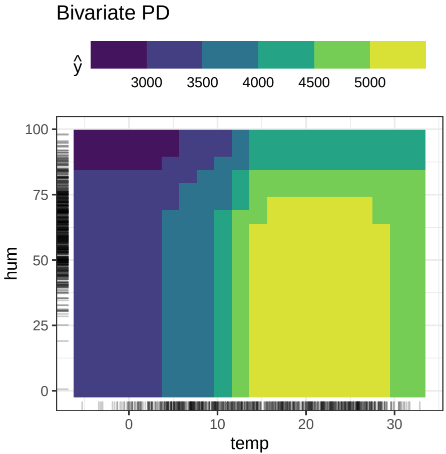
BIKE SHARING DATASET: FIRST ORDER ALE

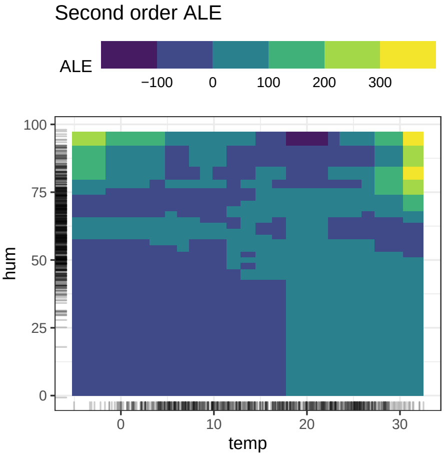
Shape of PD plot (left) often looks similar to (centered) first order ALE plot (right) but on different y-axis scale. In case of correlated features, ALE might be better due to PD's extrapolation issue.



BIKE SHARING DATASET: SECOND ORDER ALE

Unlike bivariate PD plots, 2nd-order ALE plots only estimate pure interaction between two features (1st-order effects are not included).





PD VS. ALE

PD:

$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

$$f_{S,ALE}(x) = \int_{z_0}^{x} \mathbf{E}_{\mathbf{x}_{-S}|x_S}\left(\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}\Big| x_S = z_S\right) dz_S - const$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$

PD VS. ALE

PD:

$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

$$f_{S,ALE}(x) = \int_{z_0}^{x} \mathbf{E}_{\mathbf{x}_{-S}|x_S}\left(\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}\bigg| x_S = z_S\right) dz_S - const$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates partial derivatives of feature S over z_S
  $\rightsquigarrow$ isolates effect of feature S and removes main effect of other dependent features

PD VS. ALE

PD:

$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

$$f_{S,ALE}(x) = \int_{z_0}^{x} \mathbf{E}_{\mathbf{x}_{-S}|x_S}\left(\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}\bigg| x_S = z_S\right) dz_S - const$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates partial derivatives of feature S over z_S
  $\leadsto$ isolates effect of feature S and removes main effect of other dependent features
- Difference 3: ALE is centered so that $\mathbb{E}_{x_S}\left(f_{S,ALE}(x)\right) = 0$