Applied R - Spring Edition 2019

Sampling, Intervention, Prediction, Aggregation (SIPA): A Generalized Framework for Model Agnostic Interpretation Techniques - Theory and Demonstration

---

Christian Alexander Scholbeck

About me

Christian Alexander Scholbeck

- Graduated from Ludwigs-Maximilians-University Munich in statistics and economics
- Since March 2019: doctoral candidate at Department of Statistics at Ludwigs-Maximilians-University Munich
- Research focus on interpretable / explainable machine learning
- Working groups: computational statistics + methods for missing data, model selection and model averaging

Motivation

- Various model agnostic techniques to interpret predictive black box models have been developed, e.g. individual conditional expectation (ICE), partial dependence (PD), accumulated local effects (ALE), marginal effects, permutation feature importance (PFI), Shapley values from game theory
- They use different terminology and notations $\rightarrow$ difficult to see how they are related
- When deconstructing these methods into sequential work stages, one discovers that they operate according to the exact same principle $\rightarrow$ SIPA

Today's topic: the SIPA framework

- My first research paper: available on Arxiv.org (https://arxiv.org/abs/1904.03959)
- Model agnostic methods are based on 4 work stages: sampling, intervention, prediction, aggregation (SIPA)
- The SIPA work stages inspired the development of the R package `iml` (creator and maintainer: Christoph Molnar)
- We hope to work towards a unified view on model agnostic interpretations in machine learning and inspire the development of novel methods

Talk outline

- What's interpretable machine learning?
- Overview on several prominent techniques: ICE, PD, ALE, PFI
- They really are based on the SIPA framework
- Demonstration in R

Interpretable machine learning (IML)

- Everybody likes to interpret predictive models
- Trade off: predictive performance - interpretability
- IML serves as an umbrella term: not yet a field with standardized vocabulary

Several distinctions:

- Feature effects - feature importance
- Model specific - model agnostic
- Intrinsic interpretability - post hoc interpretability
- Local - global

Notation

- We assume an unknown functional relationship between the input space and the target space plus a random error:

$$\mathcal{Y} = f(\mathcal{X}) + \epsilon$$

- We select one or several features S. The complement is denoted by C.
- The corresponding random variables are denoted by X_S and X_C.
- We collect a sample of data from the population and train a supervised learning model $\hat{f}$.
- The realized vectors of feature values are denoted by x_S and x_C.

Individual conditional expectation (ICE)

- Local feature effect estimate (for each observation)
- How does the variation of x_S affect the prediction for a single observation?
- Prediction dependent on x_S, conditional on $x_C^{(i)}$
- One curve for each observation (ICE curve)

Computations for ICE curves

data:

x_1	x_2	x_3
1	5	6
2	8	9
3	10	11

permutations for ICE curve 1:

x_1	x_2	x_3
1	5	6
2	5	6
3	5	6

permutations for ICE curve 2:

x_1	x_2	x_3
1	8	9
2	8	9
3	8	9

permutations for ICE curve 3:

x_1	x_2	x_3
1	10	11
2	10	11
3	10	11

estimated PD curve:

x_1	$\widehat{PD}(x_1)$
1	$1/3\left[\hat{f}(1,5,7)+\hat{f}(1,8,9)+\hat{f}(1,10,11)\right]$
2	$1/3\left[\hat{f}(2,5,7)+\hat{f}(2,8,9)+\hat{f}(2,10,11)\right]$
3	$1/3\left[\hat{f}(3,5,7)+\hat{f}(3,8,9)+\hat{f}(3,10,11\right]$

ICE curves - example

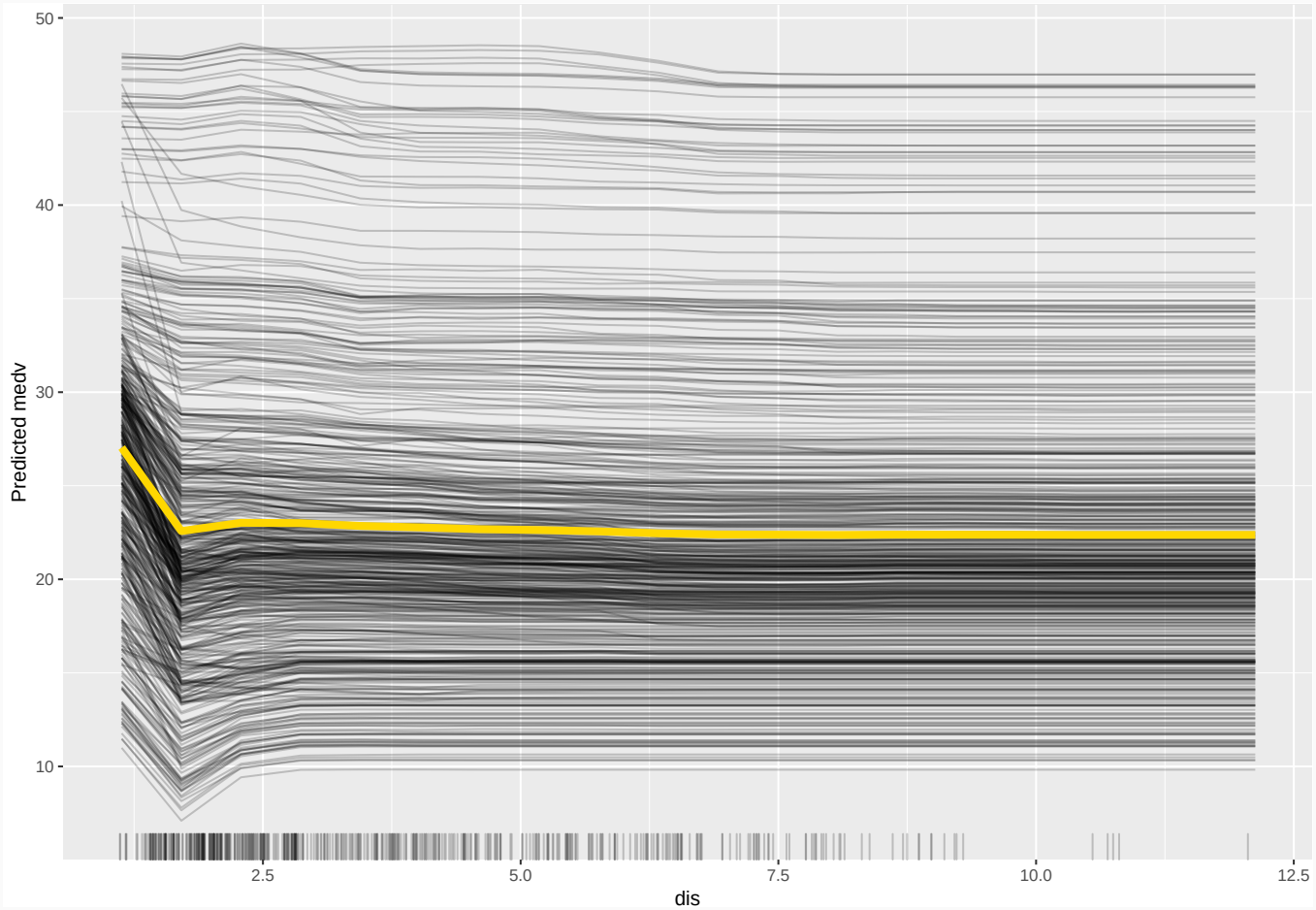

Figure 1: ICE plot of feature ‘dis‘ (weighted mean of distances to five Boston employment centres) in the Boston Housing data (with a trained random forest)

Partial dependence (PD)

- The pointwise average of ICE curves is an *estimate* of the PD and corresponds to a Monte Carlo integration:

$$\widehat{PD}(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

- PD: Dependence of the true model f on one or several features: we marginalize out other features by integrating over their marginal distribution:

$$PD(X_S) = \mathbb{E}_{X_c} \left[f(X_S, X_C) \right] = \int f(X_S, X_C) \, dP(X_C)$$

- Two uncertainties: we neither know the true model f nor the marginal distribution of unselected feature values.

Partial dependence (PD)

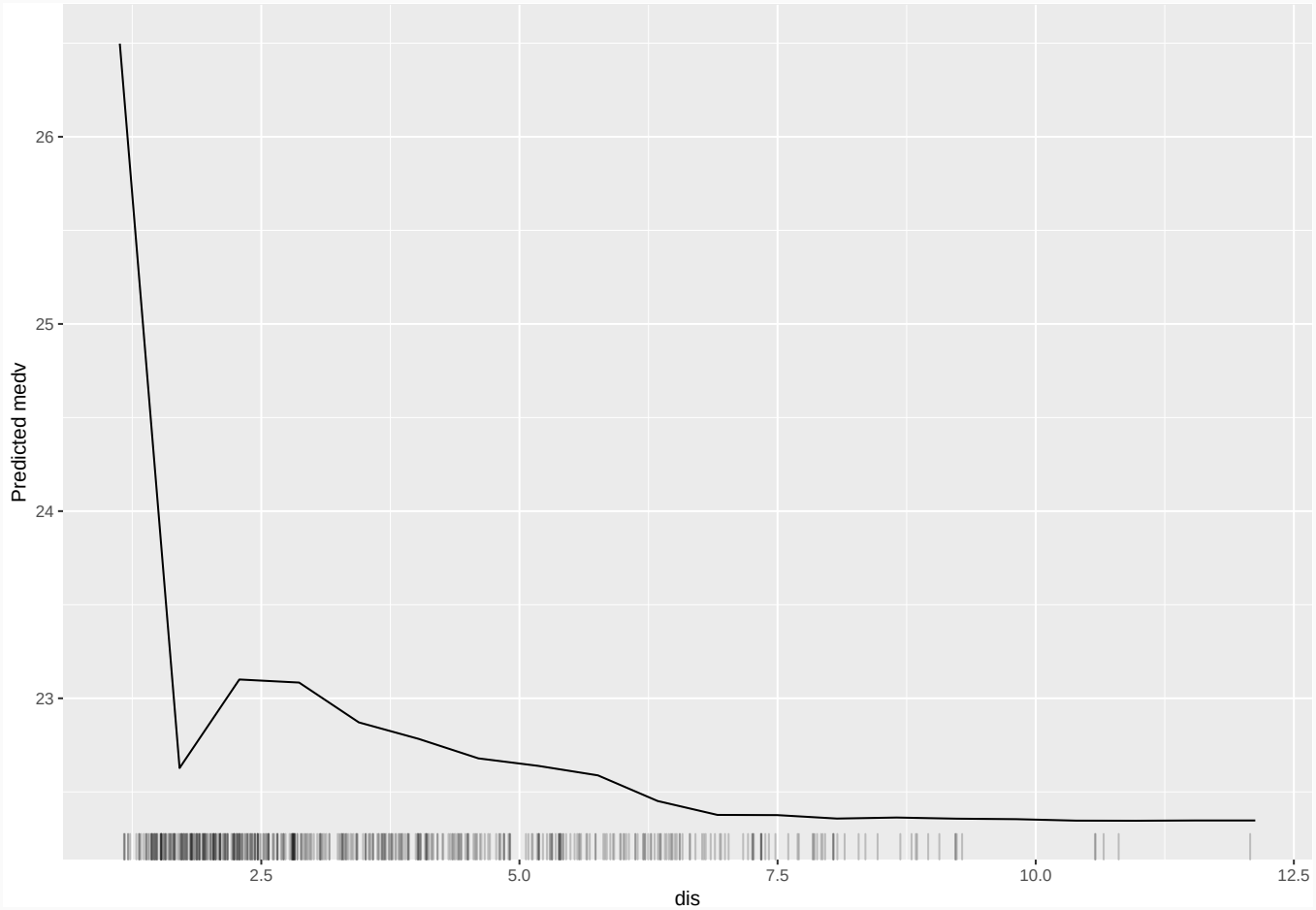

Figure 2: PD of the predicted median housing value per district on feature ‘dis‘ in the Boston Housing data (with a trained random forest)

Flaws of the PD

The PD has two major flaws:

- Flaw no. 1: It hides possible interaction effects of features
- Flaw no. 2: Its estimate is biased when features are correlated (the predictive model extrapolates because we predict in regions without sufficient training data)

The ICE is the disaggregation of the PD. It solves flaw no. 1

Flaw no. 1: interactions and the PD

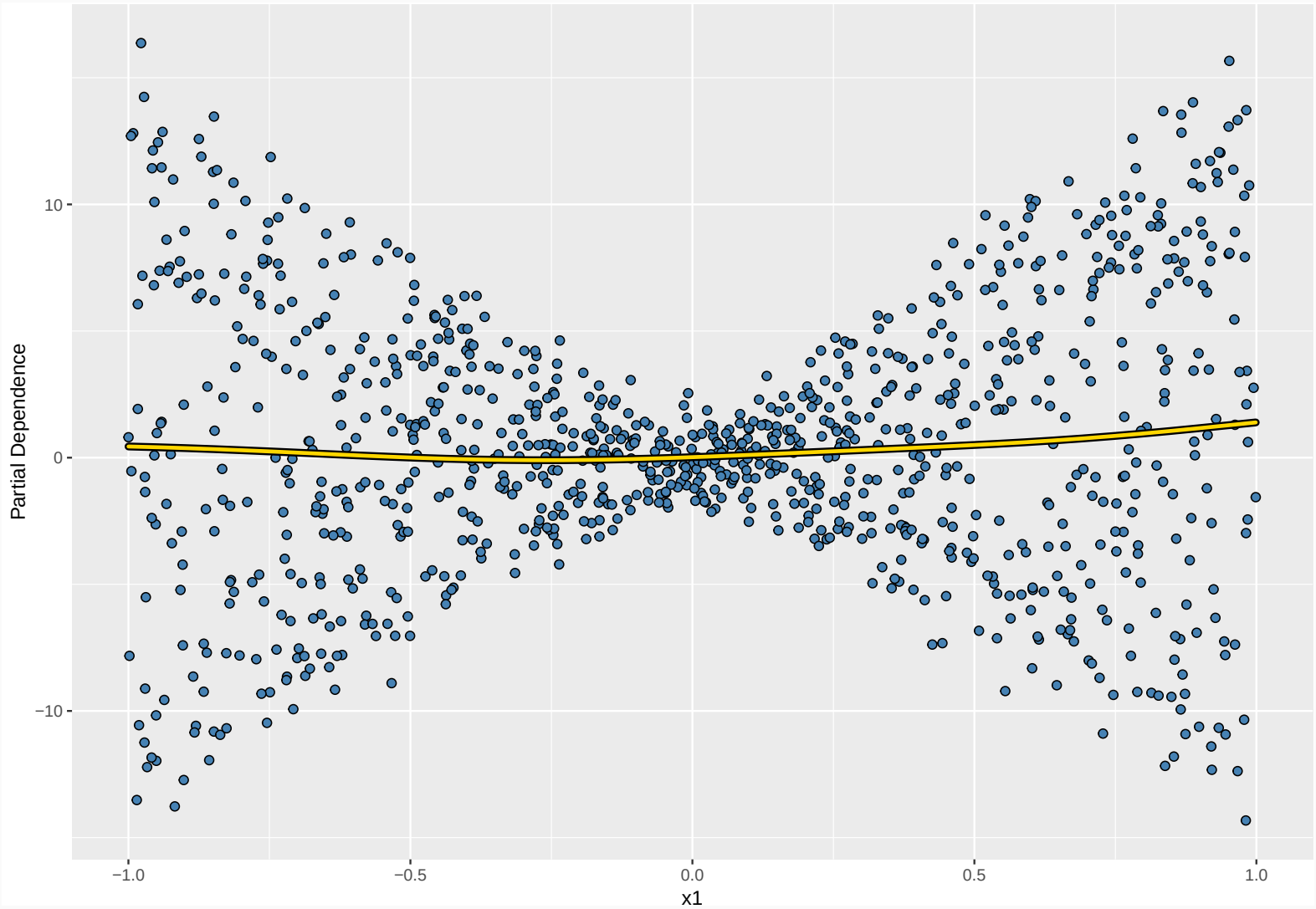

Figure 3: Scholbeck (2018)

ICE curves can be used to discover interactions

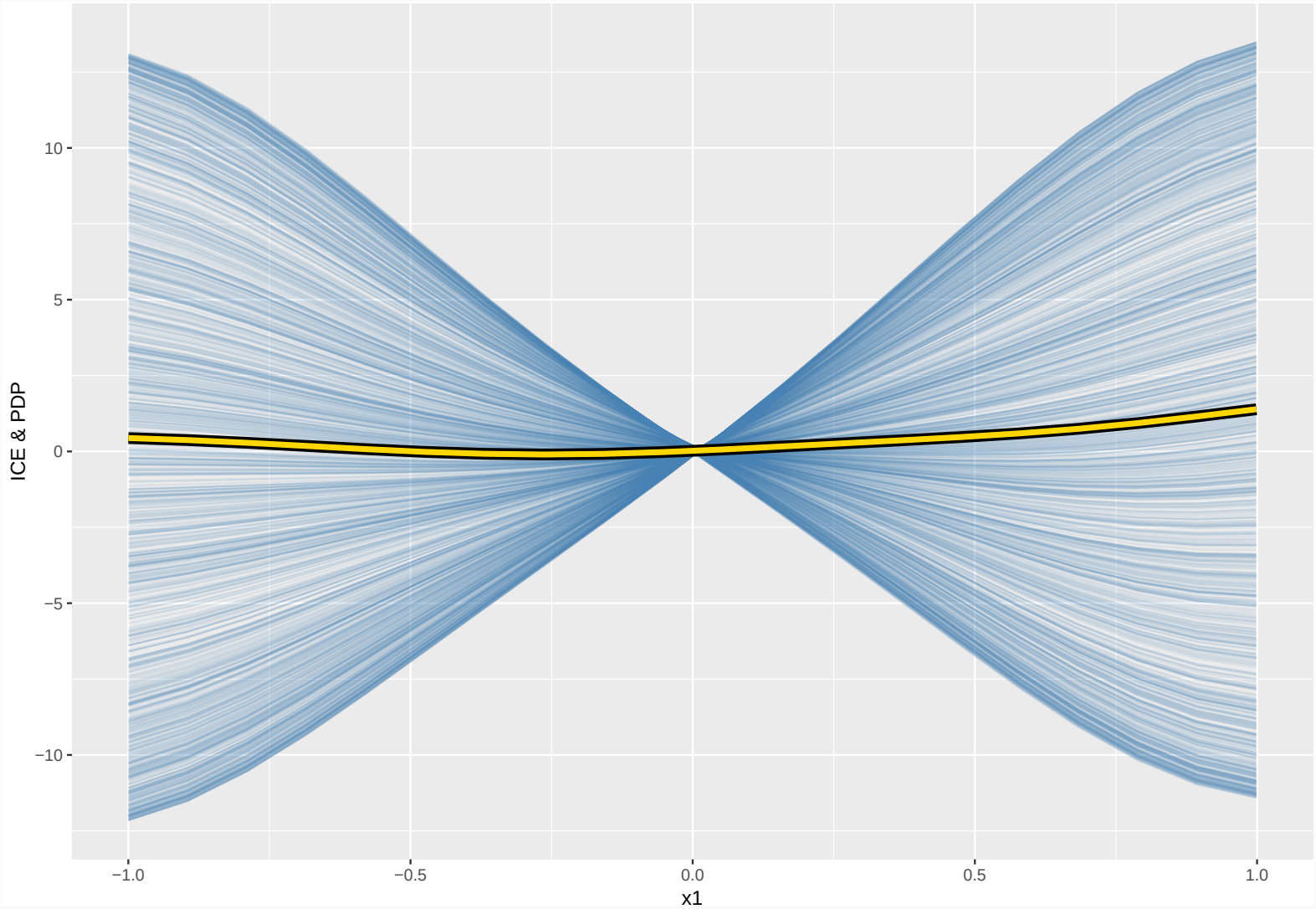

Figure 4: Scholbeck (2018)

ICE curves can be used to discover interactions

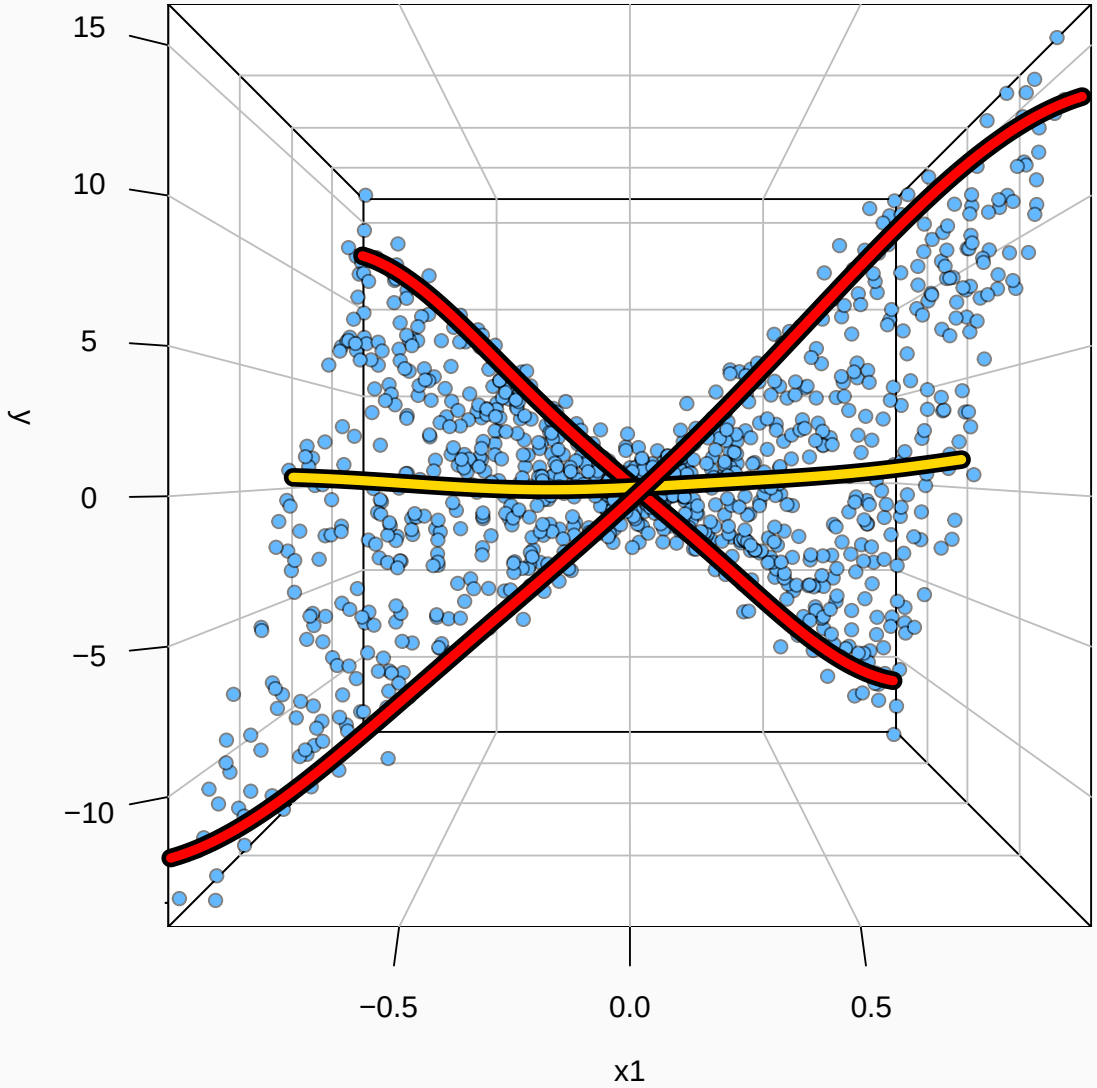

Figure 5: Scholbeck (2018)

ICE curves can be used to discover interactions

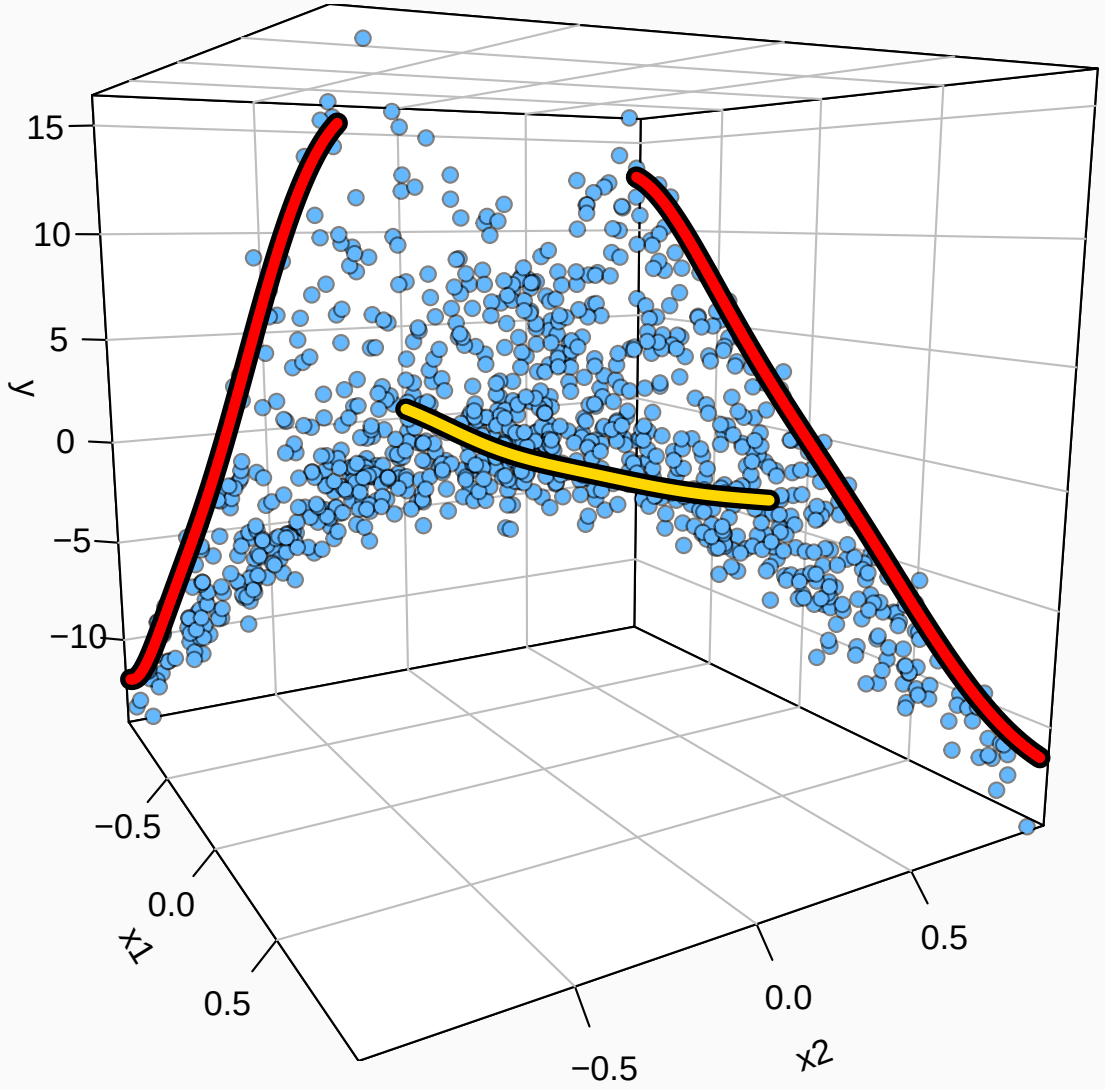

Figure 6: Scholbeck (2018)

Centered ICE (c-ICE)

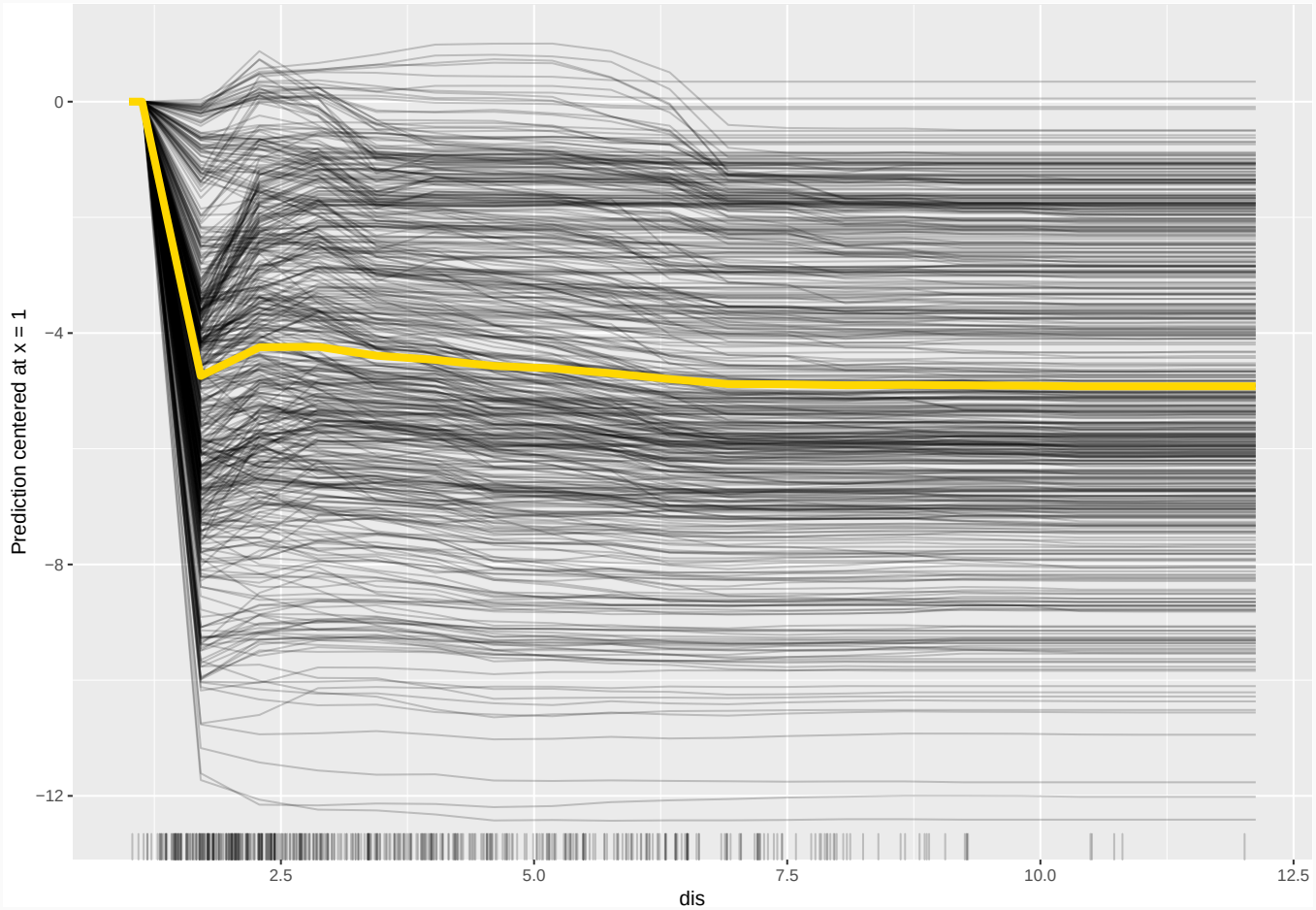

Figure 7: We can center the predictions in order to assess the divergence of curves (i.e. interaction effects).

Flaw no. 2: extrapolation when features are correlated

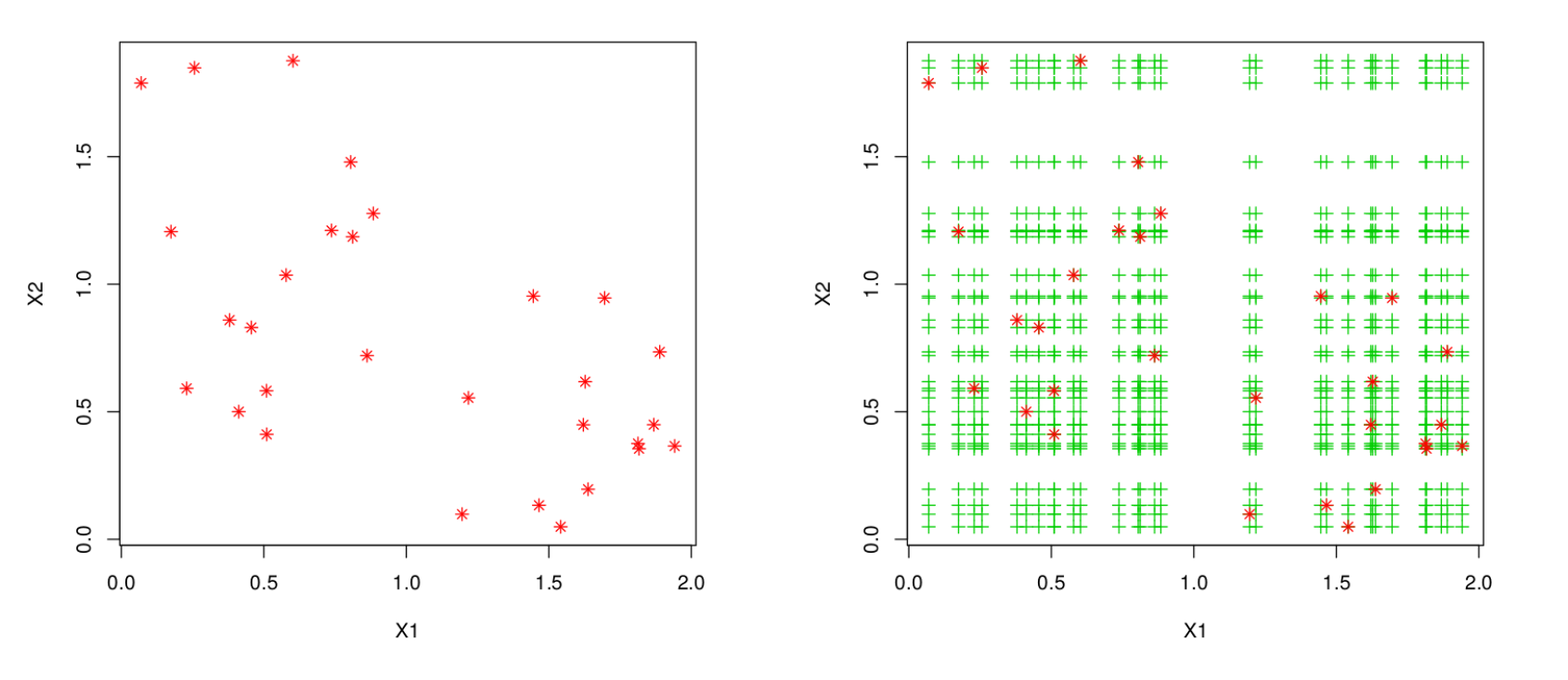

Figure 8: Hooker (2004)

Accumulated local effects (ALE)

- Solve the extrapolation problem by integrating with respect to the conditional distribution of unselected features (where the training data is located) instead of the marginal distribution
- Central idea: integrate with respect to x_S the partial derivative of the prediction function with respect to x_S
- Consider the additive prediction function: $\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$
- Partial derivative with respect to x_1: $\frac{\partial \hat{f}(x_1,x_2)}{\partial x_1} = 2 - 4x_2$
- Integral of the partial derivative: $\int \frac{\partial \hat{f}(x_1,x_2)}{\partial x_1} dx_1 = 2x_1 - 4x_1x_2$
- We removed the main effect of x_2 (which is what we wanted)

Accumulated local effects (ALE)

$$
\begin{aligned}
ALE(X_S) &= \int_{z_0}^{x_S} \mathbb{E}_{X_C|X_S} \left[\frac{\partial \hat{f}(X_S, X_C)}{\partial X_S} \middle| X_S = z_S \right] dz_S - constant \\
&= \int_{z_0}^{x_S} \left[\int \mathcal{P}(X_C|z_S) \frac{\partial \hat{f}(z_S, X_C)}{\partial z_S} dX_C \right] dz_S - constant
\end{aligned}
$$

Estimation:

- Approximate the partial derivative interval-wise by computing average interval-wise finite differences
- Integrate by summing up average interval-wise finite differences

Computing interval-wise finite differences

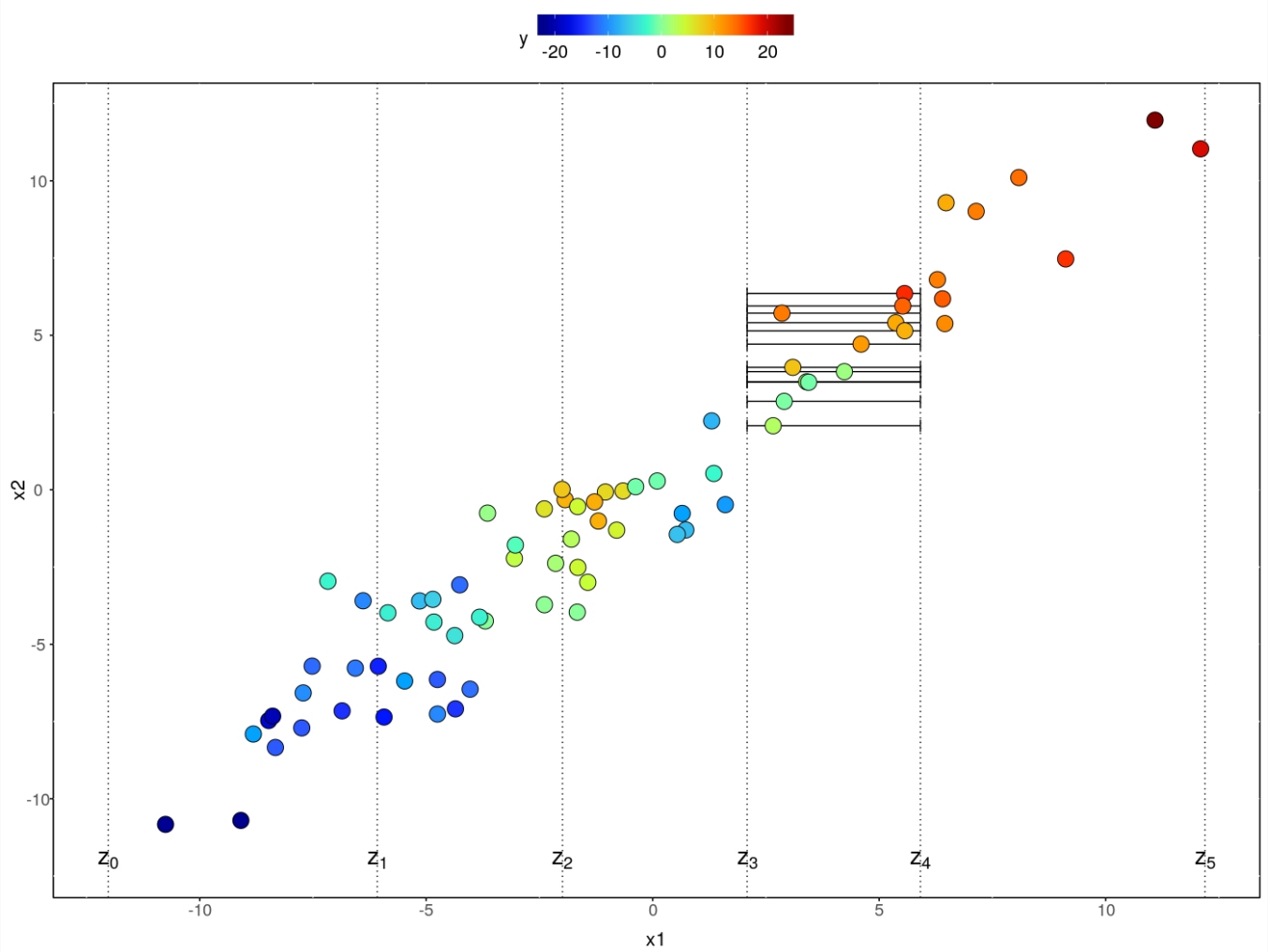

Figure 9: Scholbeck (2018)

Accumulated local effects: illustration of estimation

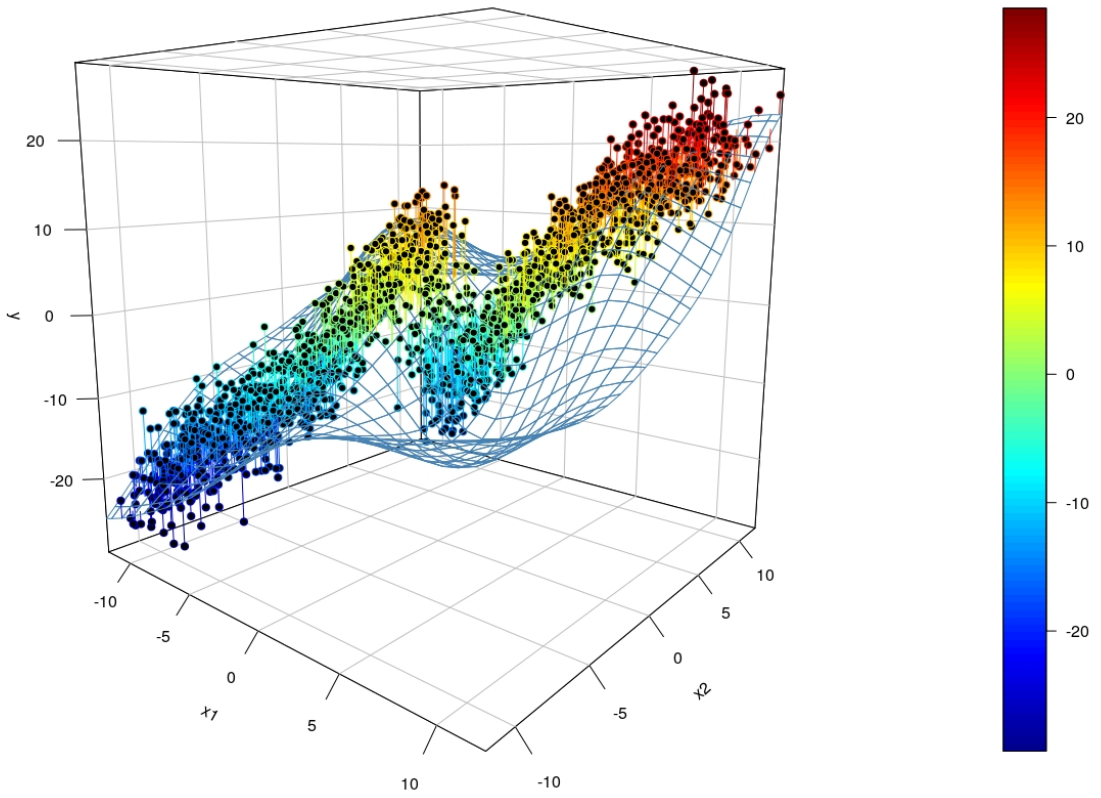

Figure 10: Scholbeck (2018)

Accumulated local effects: illustration of estimation

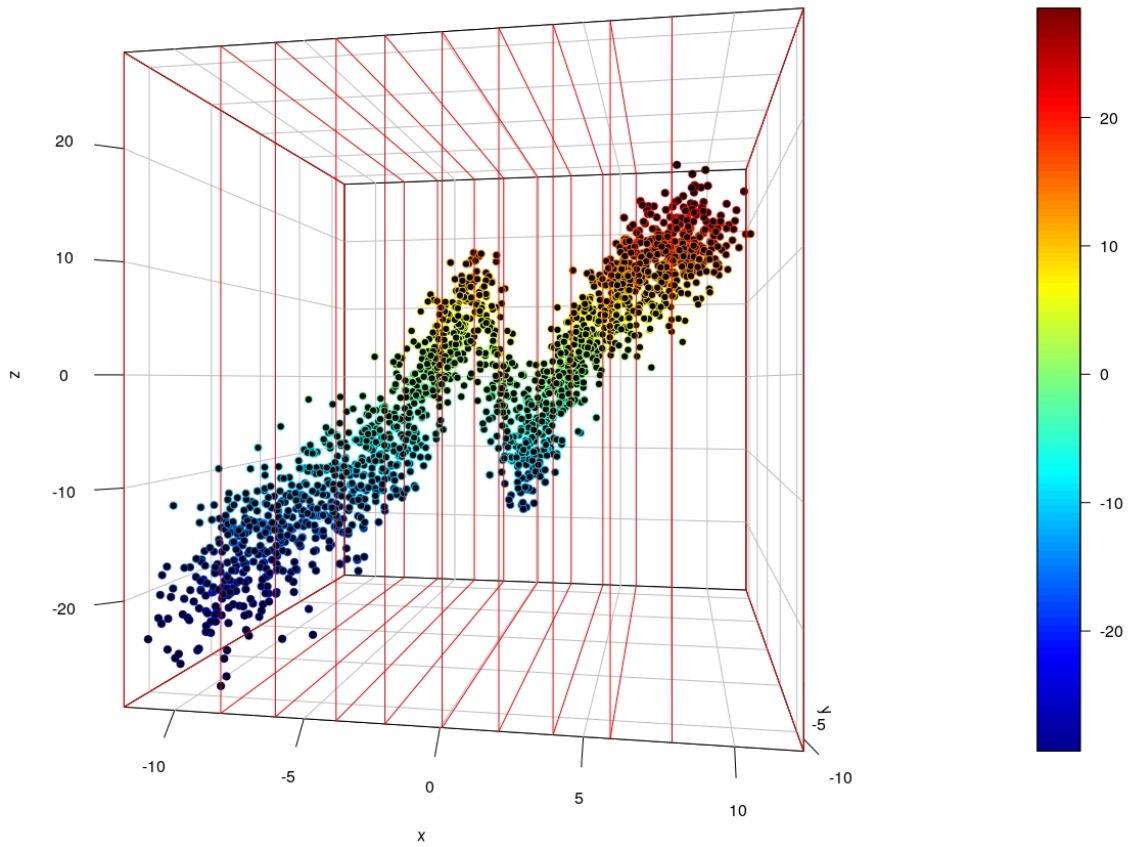

Figure 11: Scholbeck (2018)

Accumulated local effects: illustration of estimation procedure

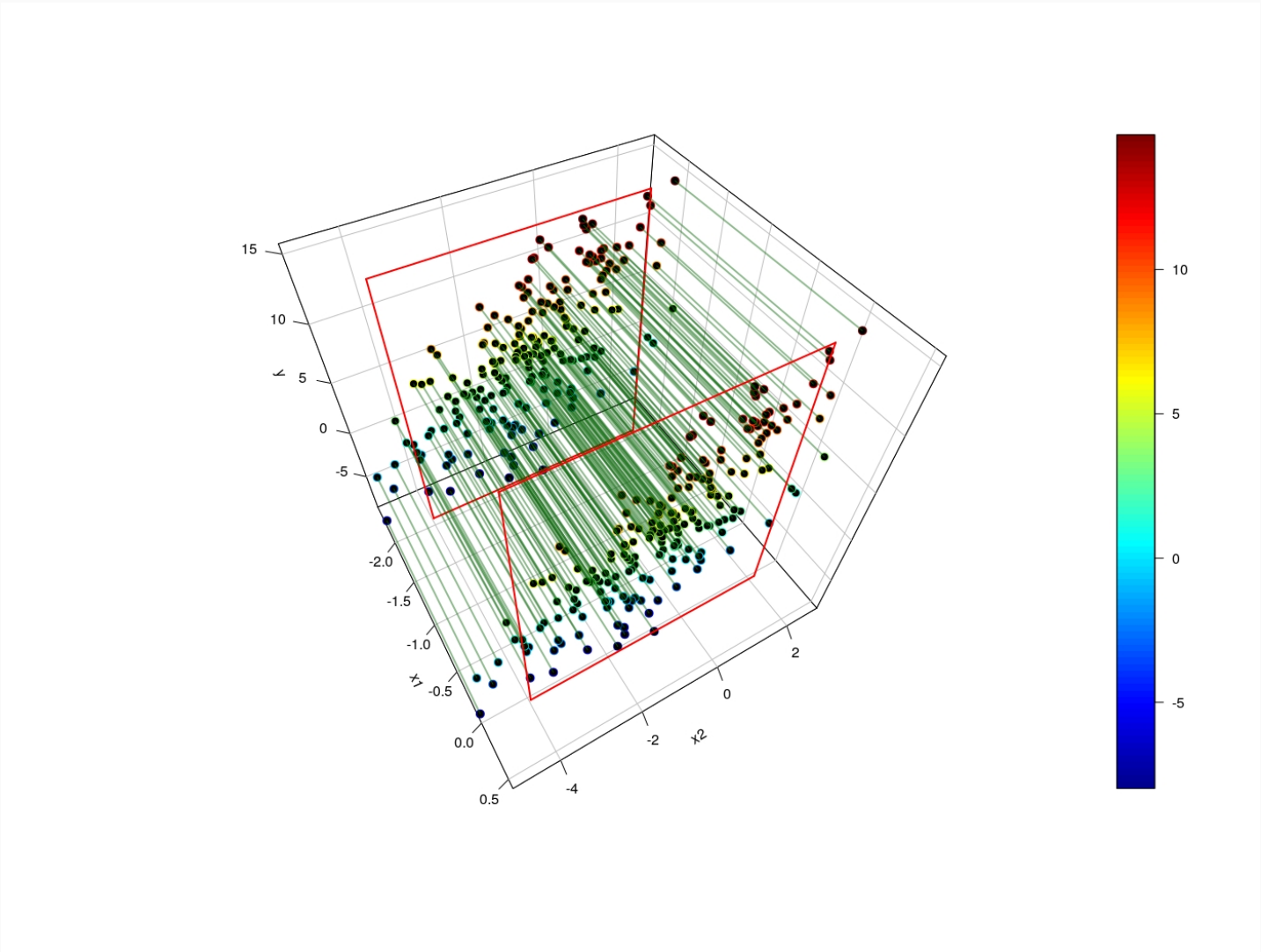

Figure 12: Scholbeck (2018)

Accumulated local effects: illustration of estimation procedure

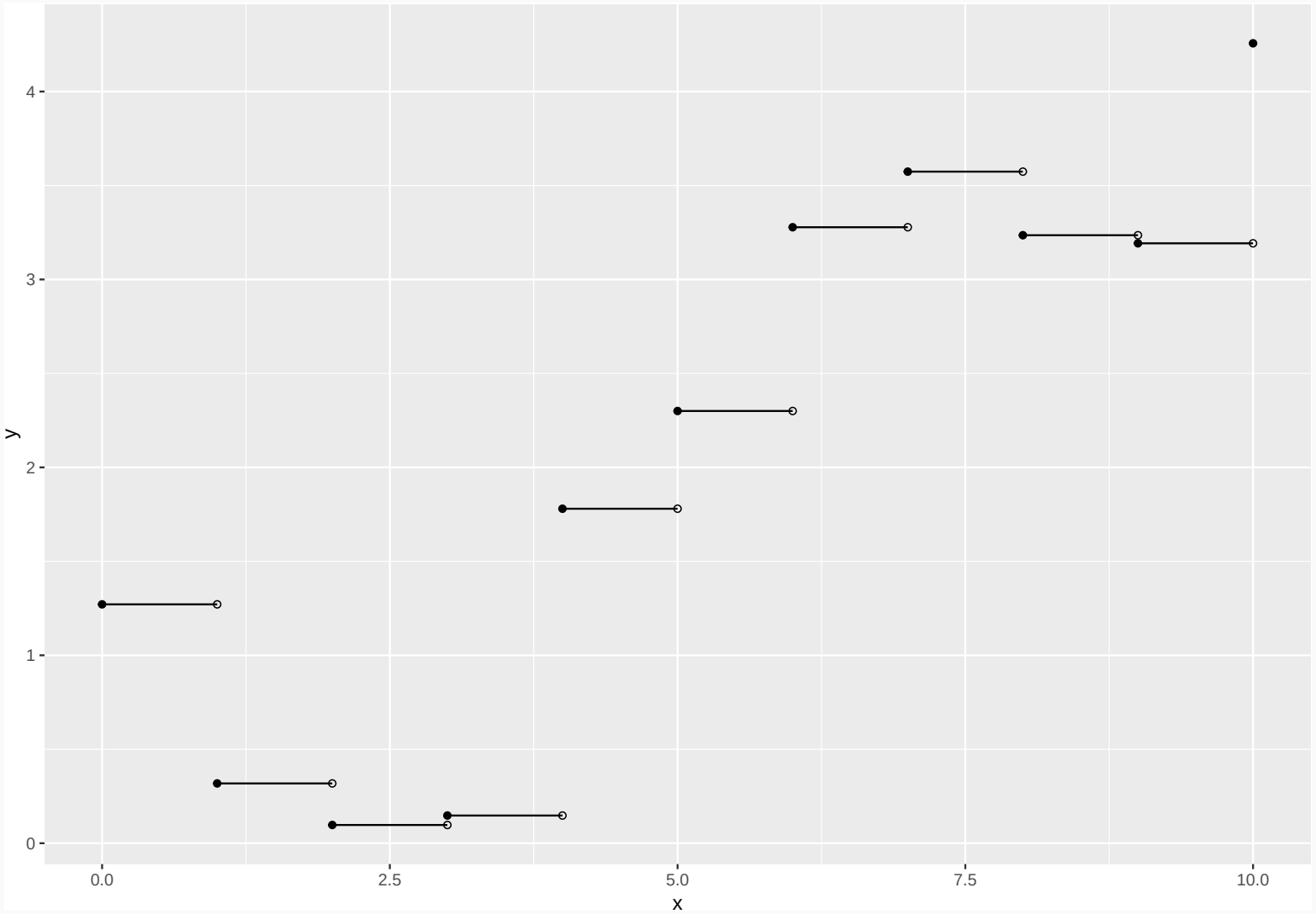

Figure 13: The approximated first derivative of the prediction function is a step function.

Accumulated local effects: comparison with PD

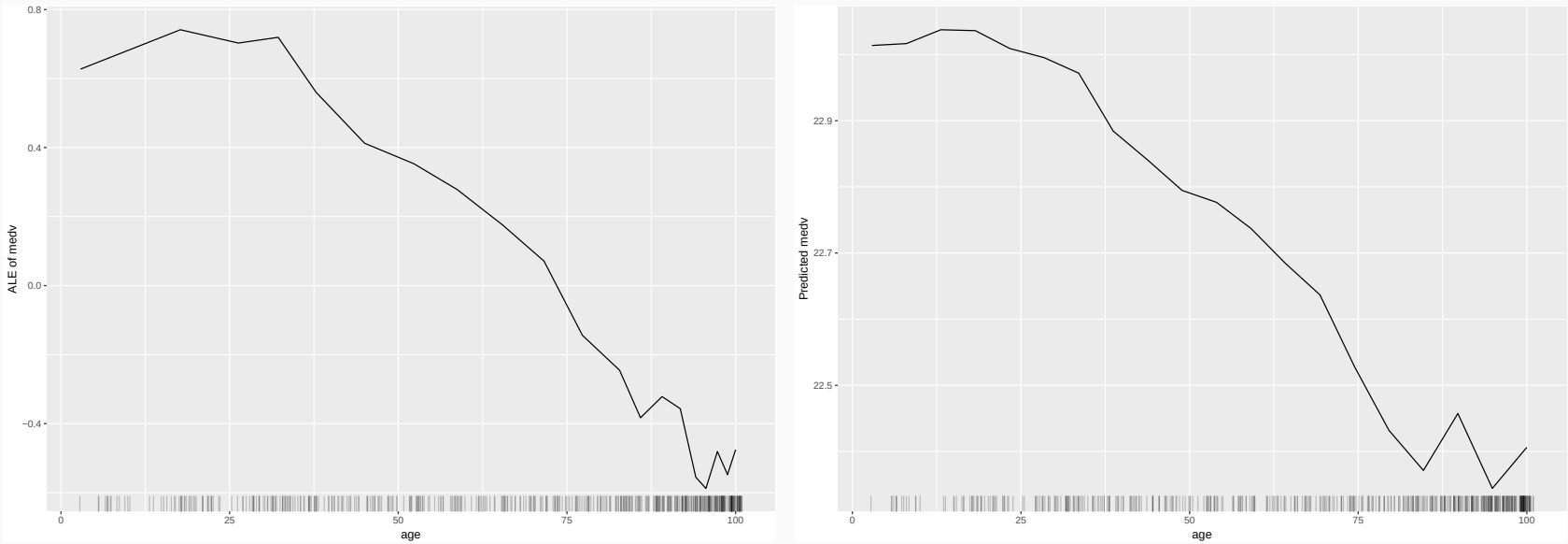

Figure 14: Comparison of ALE (left) to PD (right).

The SIPA framework

Model agnostic techniques work according to the same principle $\rightarrow$ we evaluate the predictions of the black box model when changing inputs:

- **Sampling**: In order to reduce computational costs, we first sample a subset of observations, e.g. to evaluate as ICE curves
- **Intervention**: We have to intervene in our dataset in order to change the predictions. Either set values to ones from the observed marginal distributions (ICE & PD), or unobserved values (finite difference based such as ALE).
- **Prediction**: Predict on intervened data
- **Aggregation**: Aggregate local predictions (ICE, observation-wise FDs) to global ones (PD, ALE)
- Optional: visualization

What about the feature importance?

- **Variance-based** feature importance: estimated the variance of effect estimates such as the PD curve $\rightarrow$ If a feature doesn't have an effect on the outcome it might also be considered unimportant.
- **Performance-based** feature importance: consider changes in loss $\rightarrow$ model agnostic permutation feature importance

Permutation feature importance (PFI)

- Assume there is a relationship between feature X_S and the outcome Y. If a feature is shuffled in isolation, the relationship between the feature and target is broken up. If the feature was important, this should result in an increased loss.
- With the shuffled feature $\tilde{X}_S$, the PFI is denoted by:

$$PFI_S = \mathbb{E}\left[\mathcal{L}(\hat{f}(\tilde{X}_S, X_C), Y)\right] - \mathbb{E}\left[\mathcal{L}(\hat{f}(X), Y)\right]$$

- We estimate the PFI with the generalization error (GE) on data with shuffled columns $\mathcal{D}_S$ and unshuffled columns $\mathcal{D}$:

$$\widehat{PFI}_S = \widehat{GE}(\hat{f}, \mathcal{D}_S) - \widehat{GE}(\hat{f}, \mathcal{D})$$

Extending the SIPA framework to the feature importance

Variance-based feature importance: we already demonstrated effect estimates to be based on the SIPA work stages

Performance-based feature importance $\rightarrow$ 2 modifications:

- Predict both on intervened and non-intervened (observed) data during the prediction stage
- Consider the loss function instead of the prediction and take the difference between predictions with intervened and non-intervened data during the aggregation stage

Novel visualization method that takes advantage of this modification: individual conditional importance (ICI) and partial importance (PI) curves (Casalicchio et al., 2018)

Methods we demonstrated to be based on the SIPA framework

Feature effects:

- individual conditional expectation
- partial dependence
- accumulated local effects
- marginal effects (AME, MER, MEM)
- Shapley values

Feature importance:

- permutation feature importance
- individual conditional importance
- partial importance
- Shapley feature importance

Demonstration in R