Recap: simulation set-up

Data Generating Mechanism	Additive: $f(x) = x_1 + 0.5x_2^2$	Combined: $f(x) = x_1 + 0.5x_2^2 + x_1x_2$
$\rho = 0$, standard normal feature distributions		
$\rho = 0.5$, standard normal feature distributions		
$\rho = 0.9$, standard normal feature distributions		

- 1000 training samples
- SNRs: 10, 5
- 20 repetitions on samples drawn with different random seeds
- additionally: 2 uncorrelated random noise features with same marginals

ML algorithms

- GAM (correctly specified + full)
- XGBoost (interactions correctly specified + full)
- SVM with RBF-kernel

each tuned well for 200 iterations with TPE w.r.t. their 5-CV MSE

Feature effect methods

- PDP (1D)
- ALE (1D)

Performance Measures

- Model Performance: MSE, MAE and R2-Score on holdout test set (10000 samples).
- Feature Effect Error: Average pointwise L2-loss between centered estimated model PD (ALE) and estimated groundtruth PD (ALE) at 100 grid points on training data

$$Err_c(\widehat{PD}_{\hat{f},S}(x_S), \widehat{PD}_{f,S}(x_S))$$

$$Err_c(\widehat{ALE}_{\hat{f},S}(x_S), \widehat{ALE}_{f,S}(x_S))$$

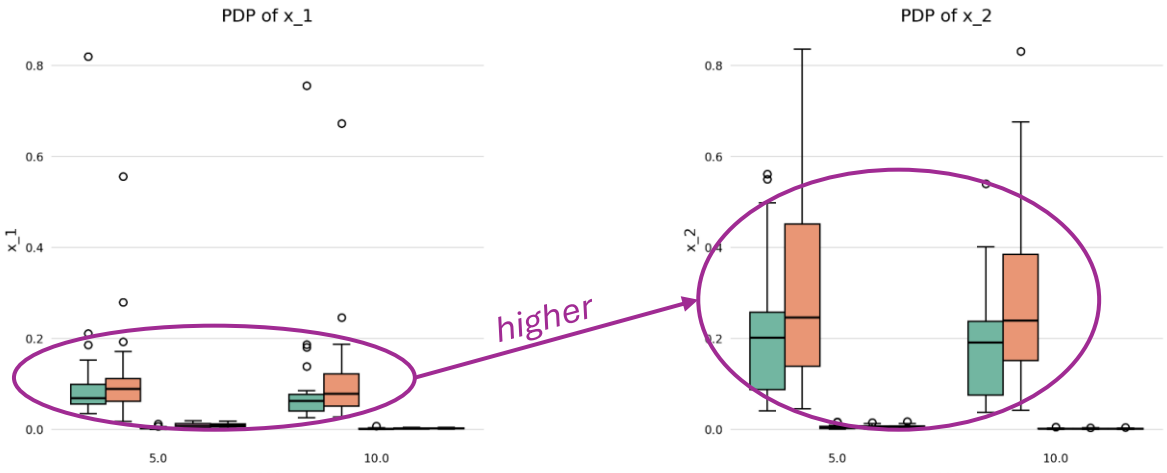
Interesting finding: surprisingly high feature effect error for certain models for PDP of x_1 in the following scenario

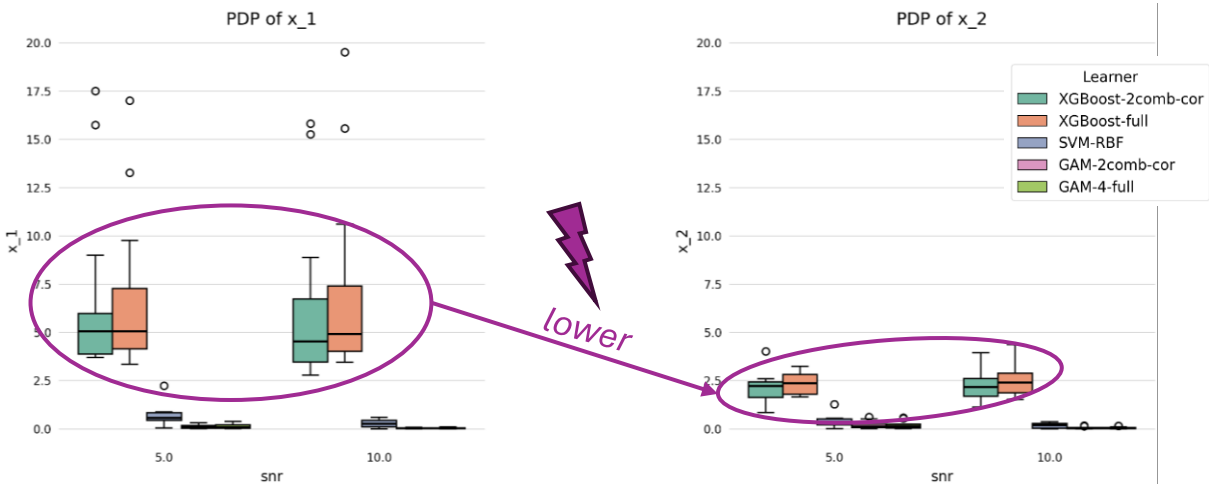
Groundtruth: $x_1 + \frac{1}{2}x_2^2 + x_1x_2$

corr = 0

corr = 0.9

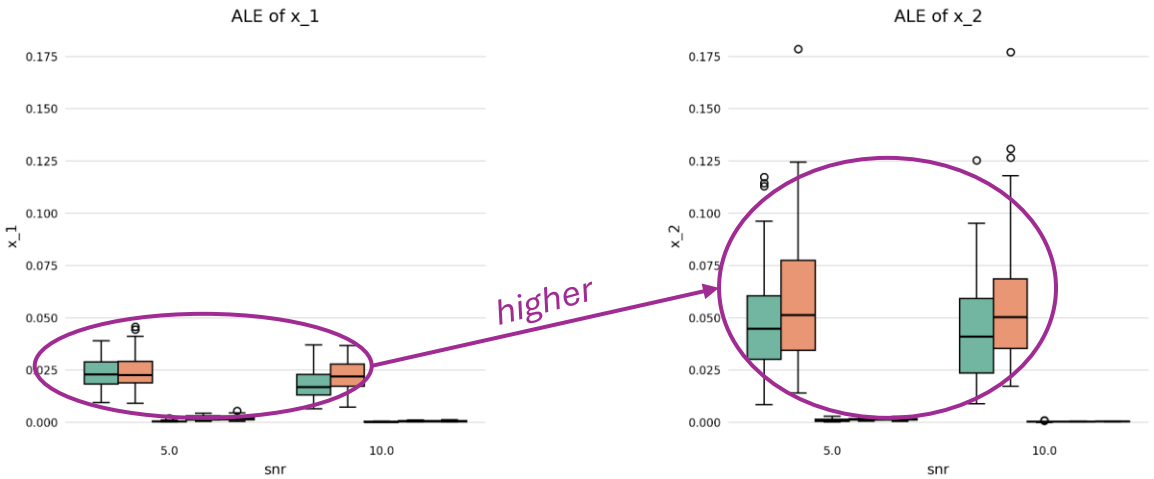
Feature effect error

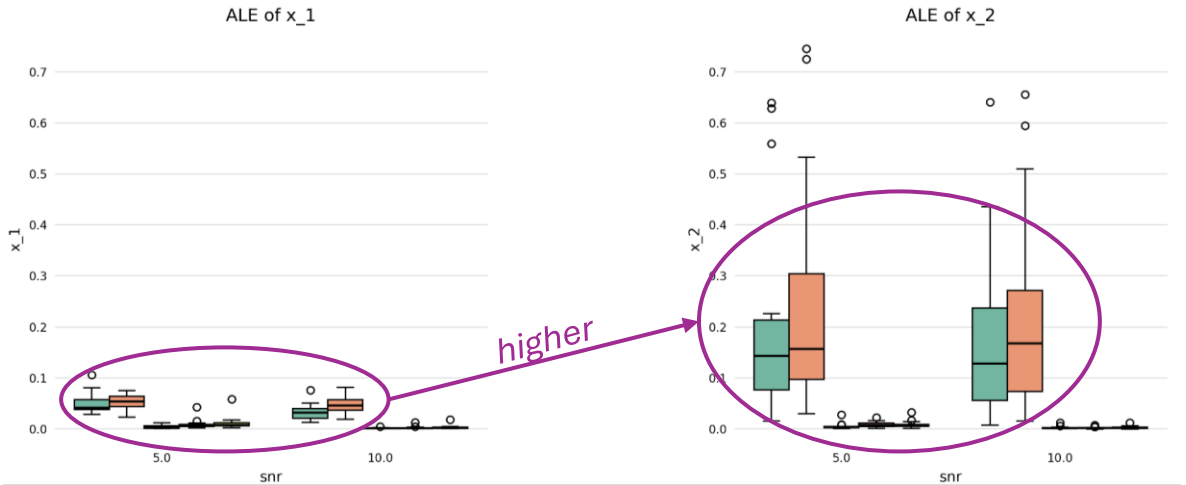




Error for x_2 always higher for both ALE & PDP

For PDP, error for x_1 is now higher (for XGBoost, SVM), while for ALE x_2 error still higher → **WHY?**





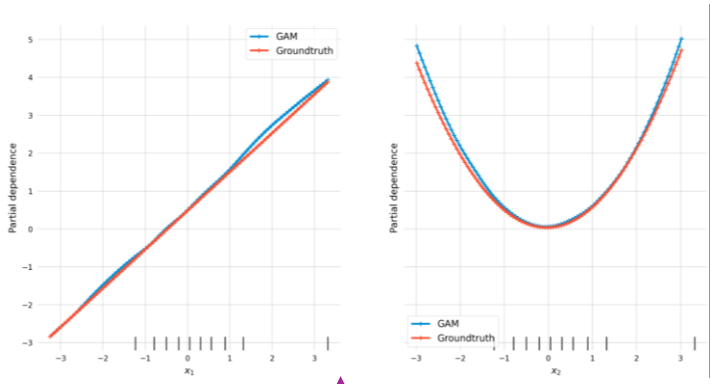
Deeper insights with concrete examples: feature effect curves from 1st simulation with snr=10 for **corr=0.9**

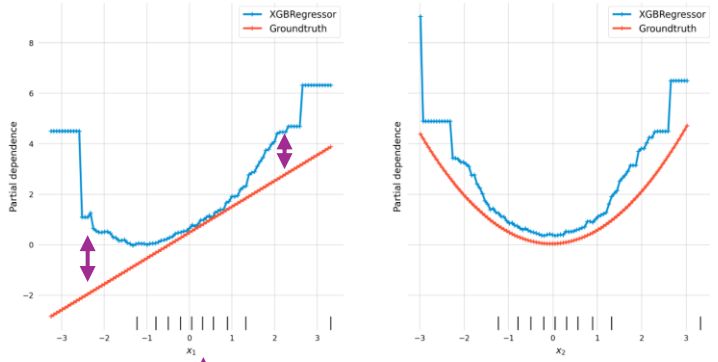
GAM (correctly specified)

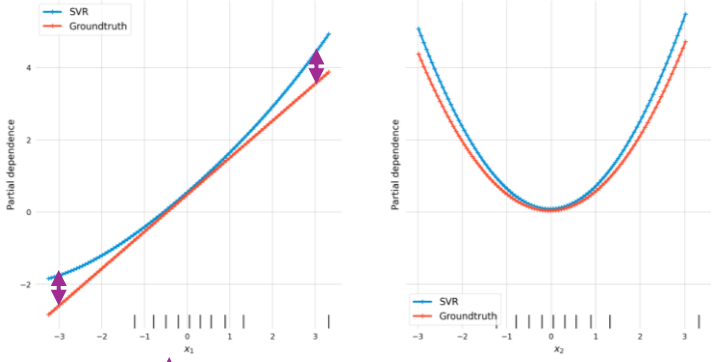
XGBoost (interactions specified)

SVM

PDP





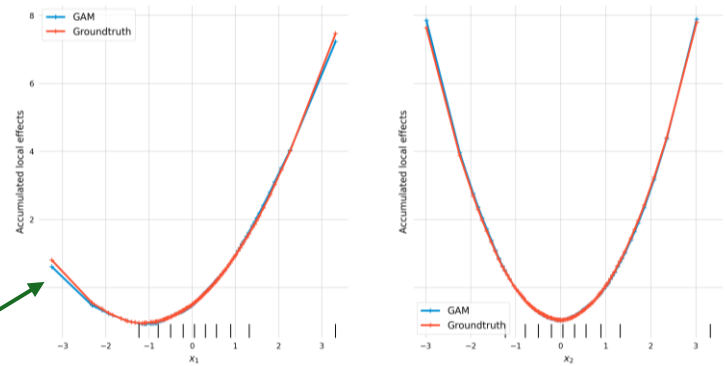


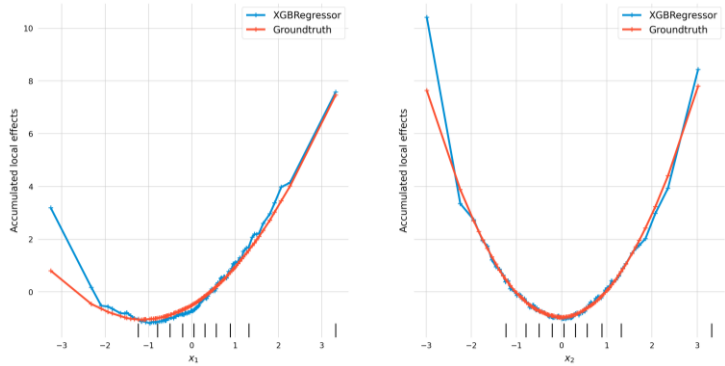
both PDP & ALE of the GAM match with the groundtruths for both x_1 and x_2

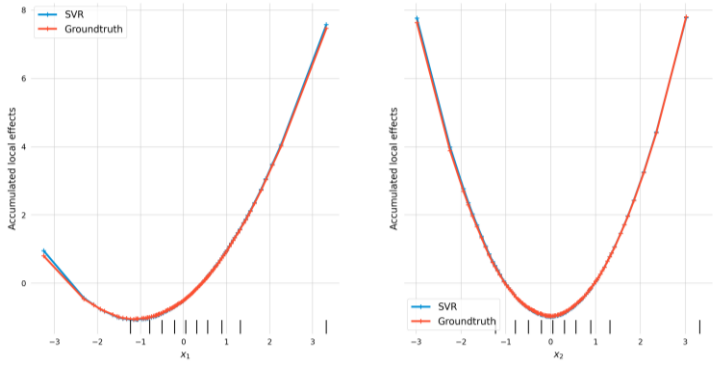
for x_1, XGBoost PDP extremely differs from groundtruth PDP, while ALE similar to groundtruth ALE

for x_1, SVM PDP has a slightly different shape than groundtruth PDP, while similar shapes in case of ALE

ALE







Although not showing linear effect, this is expected behavior of ALE as it accounts for interacting/correlated features

Additional notes:

- The described effect is also visible in weakened form for corr=0.5 (but not for corr=0)
- For purely additive groundtruths (i.e. no interaction), this effect is not visible at all (holds for all correlation strengths between the features)