Feature Effect Empirical Analysis

Preliminary Results 2024/06/18

Overview

Aim: Quantification of the error between groundtruth 1D feature effect and estimated feature effects with different ML models and Feature Effect methods (PDP, ALE) for simple groundtruth functions

Data Generating Mechanisms

	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

Overview

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 equidistant grid points

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

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

Definitions

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

where f is the groundtruth function, x_S are the features for which the partial dependence function is computed and $x_C^{(i)}$ are actual feature values from the training sample for the features in which we are not interested, n is the number of instances in the sample.

$$\widetilde{\widehat{ALE}}_{f,S}(x_S) = \sum_{k=1}^{k_S(x_S)} \frac{1}{n_S(k)} \sum_{\{i:x_S^{(i)} \in N_S(k)\}} [f(z_{k,S}, x_C^{(i)}) - f(z_{k-1,S}, x_C^{(i)})]$$

This effect is centered so that the mean effect is zero:

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

Again S is the feature for which the feature effect is computed (with observations x_S, $x_S^{(i)}$ for the ith observation), C the remaining features.
For each feature, $\{N_S(k) = (z_{k-1,S}, z_{k,S}] : k = 1, 2, \ldots, K\}$ describes a sufficiently fine partition of the sample range of $\{x_S^{(i)} : i = 1, 2, \ldots, n\}$ into K intervals.
For $k = 1, 2, \ldots, K$, $n_S(k)$ denotes the number of training observations that fall into the kth interval $N_S(k)$. For a particular value x of the predictor x_S, $k_S(x)$ denotes the index of the interval into which x falls.

Definitions

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

where $\hat{f}$ is the trained model (also estimated on the training data).

$$\widetilde{\widehat{ALE}}_{\hat{f},S}(x_S) = \sum_{k=1}^{k_S(x_S)}\frac{1}{n_S(k)}\sum_{\{i:x_S^{(i)}\in N_S(k)\}}[\hat{f}(z_{k,S}, x_C^{(i)}) - \hat{f}(z_{k-1,S}, x_C^{(i)})]$$

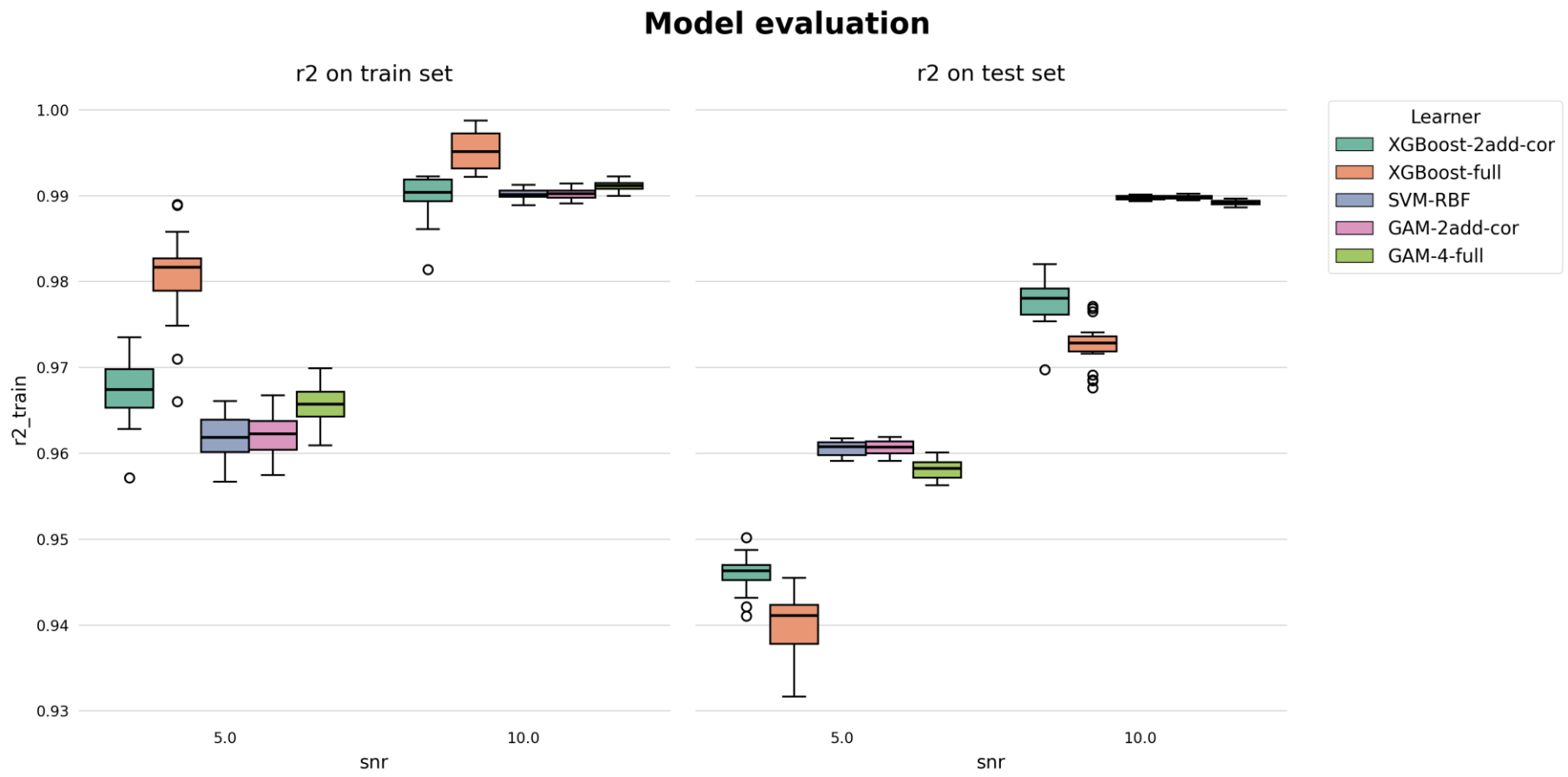
for the uncentered effect, where $\hat{f}$ is the estimated model. This effect is again centered so that the mean effect is zero:

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

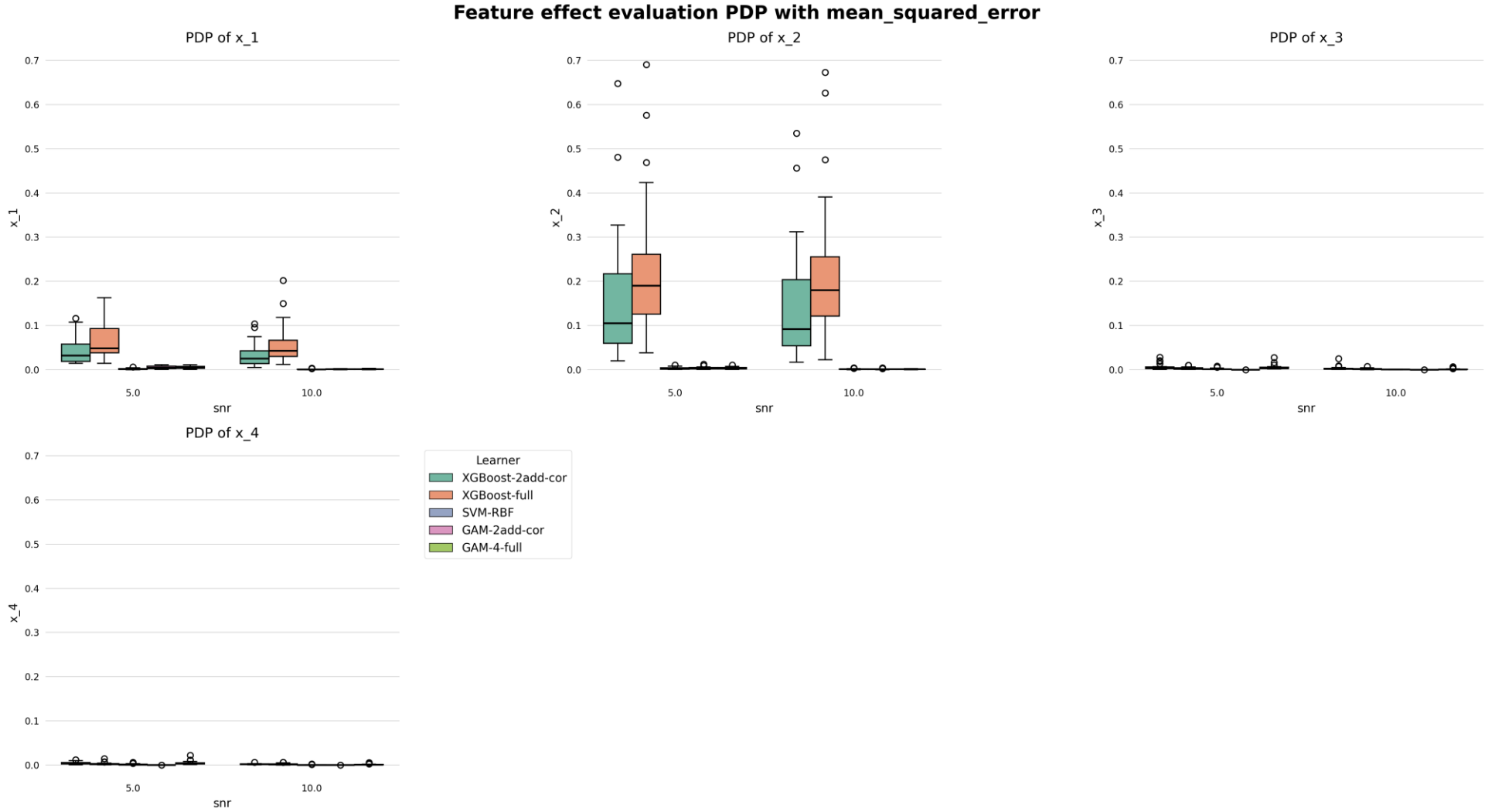
Results Additive Scenario

Model Performance [MSE]

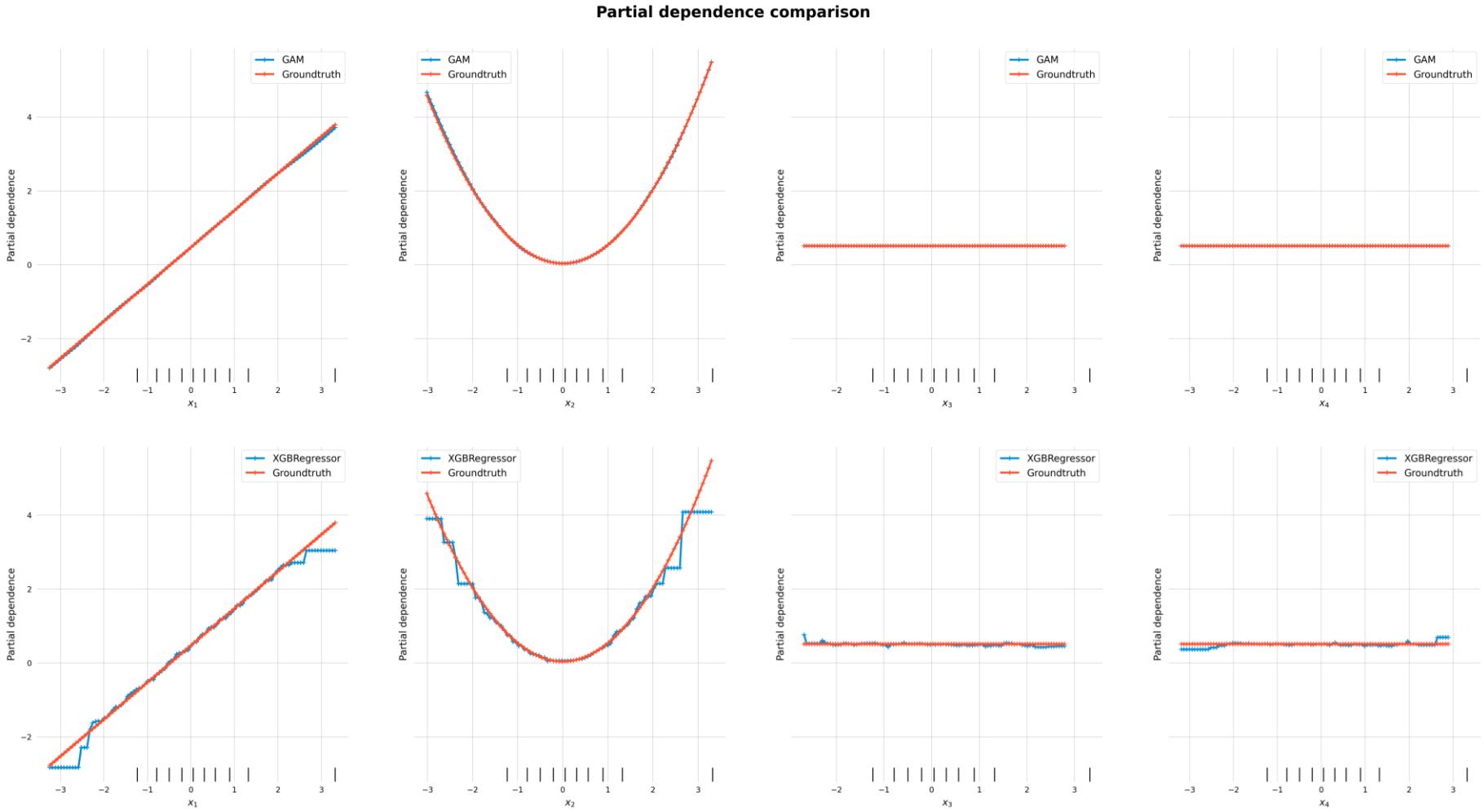
Model Performance [R2]



Feature Effect Error [PDP]



Feature Effect Examples [PDP]

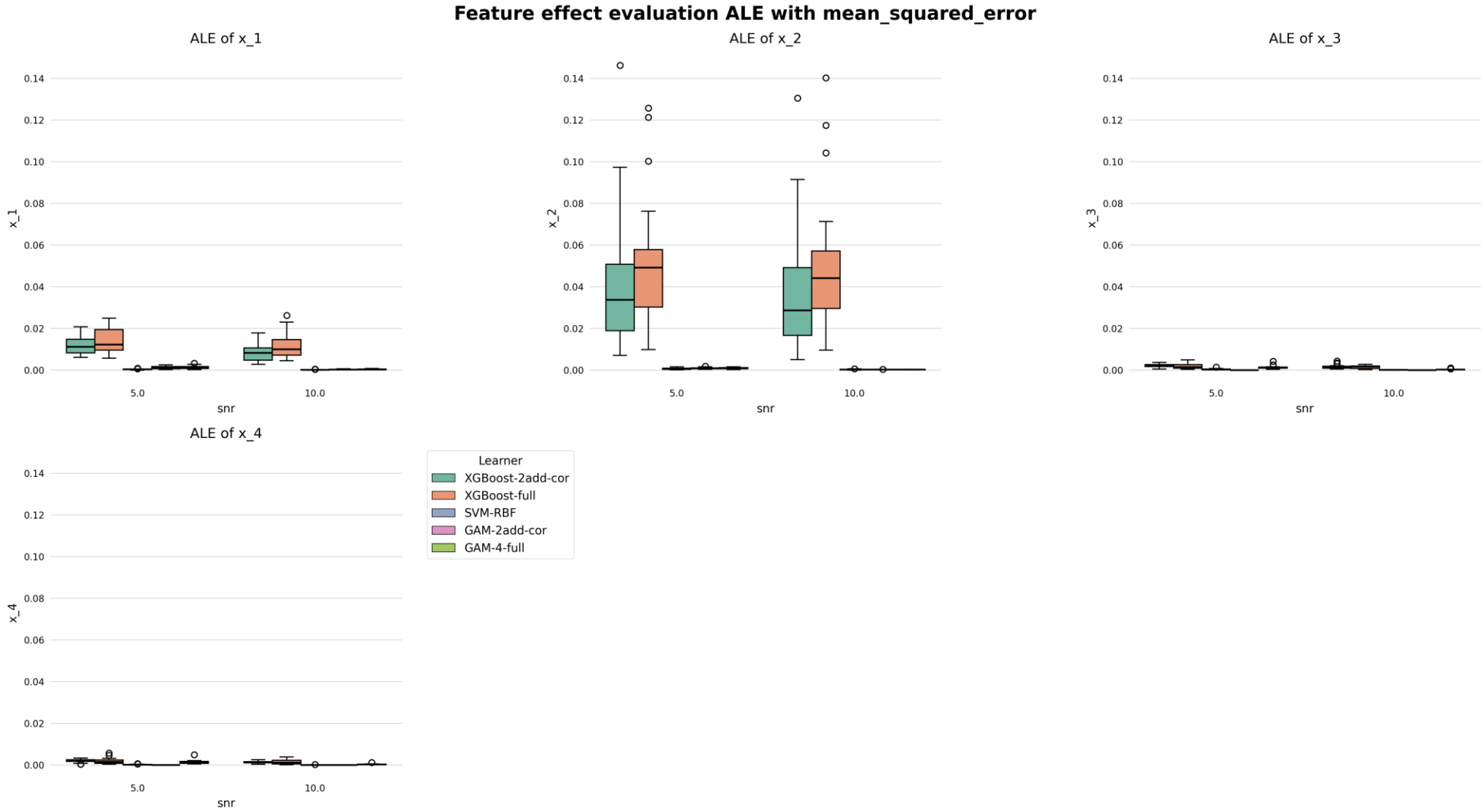


Note: Feature effects are not centered in the plots

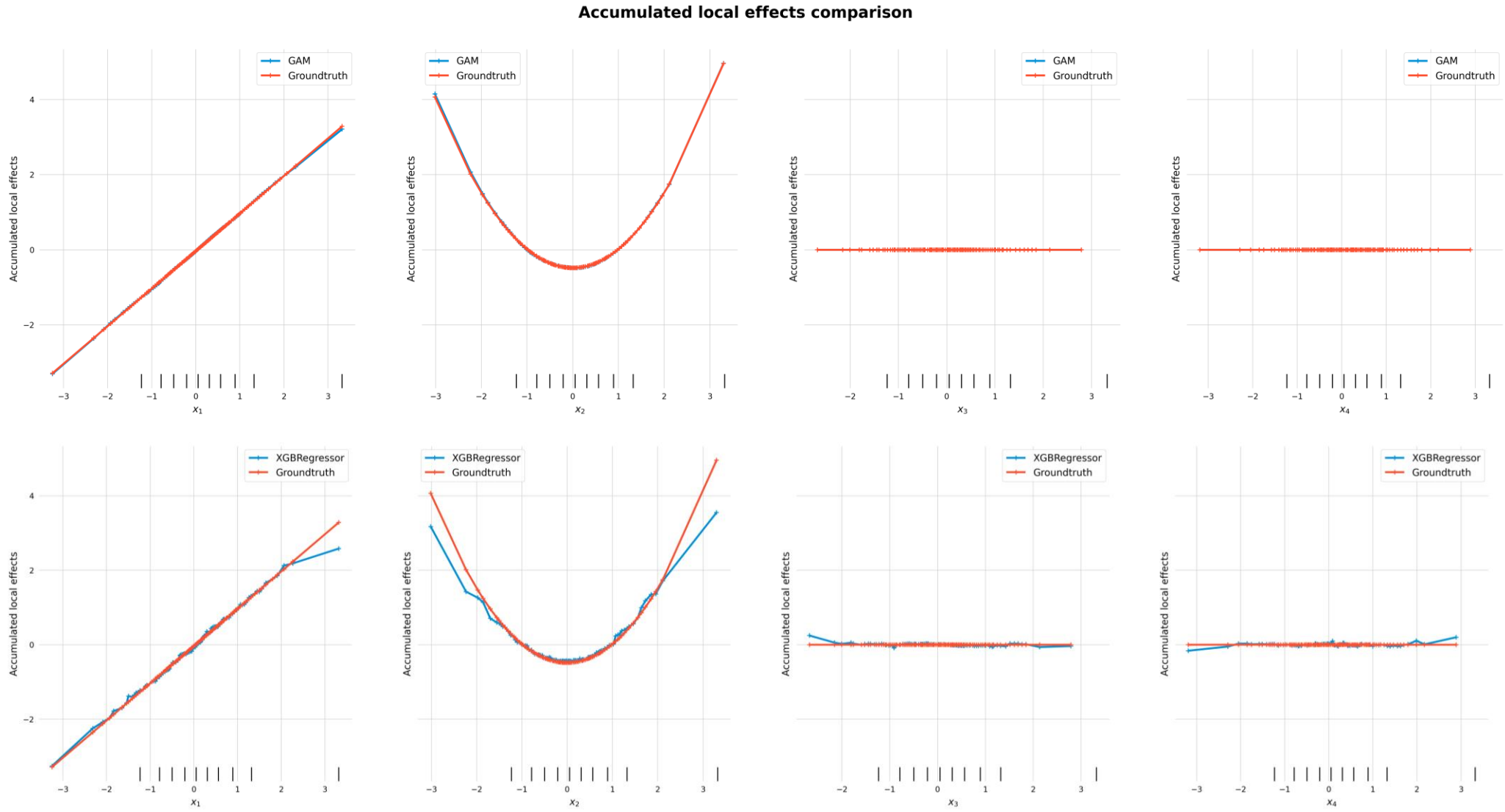
Correlation Analysis [PDP]

feature	snr	model	correlation
x_1	5.0	Overall	0.820078
		XGBoost-2add-cor	-0.109774
		XGBoost-full	0.288722
		SVM-RBF	0.497744
		GAM-2add-cor	0.269173
		GAM-4-full	0.042105
	10.0	Overall	0.796376
		XGBoost-2add-cor	0.063158
		XGBoost-full	0.228571
		SVM-RBF	0.366917
		GAM-2add-cor	0.181955
		GAM-4-full	0.046617
x_2	5.0	Overall	0.768449
		XGBoost-2add-cor	0.239098
		XGBoost-full	-0.276692
		SVM-RBF	0.209023
		GAM-2add-cor	0.090226
		GAM-4-full	0.305263
	10.0	Overall	0.749511
		XGBoost-2add-cor	0.082707
		XGBoost-full	-0.030075
		SVM-RBF	0.009023
		GAM-2add-cor	0.114286
		GAM-4-full	0.398496
x_3	5.0	Overall	0.526613
		XGBoost-2add-cor	-0.163910
		XGBoost-full	0.057143
		SVM-RBF	0.323308
		GAM-2add-cor	0.248120
		GAM-4-full	0.222556
	10.0	Overall	0.720600
		XGBoost-2add-cor	-0.081203
		XGBoost-full	-0.111278
		SVM-RBF	0.175940
		GAM-2add-cor	0.260150
		GAM-4-full	0.177444
x_4	5.0	Overall	0.513075
		XGBoost-2add-cor	-0.172932
		XGBoost-full	0.093233
		SVM-RBF	-0.320301
		GAM-2add-cor	0.248120
		GAM-4-full	0.303759
	10.0	Overall	0.702562
		XGBoost-2add-cor	-0.347368
		XGBoost-full	0.070677
		SVM-RBF	0.004511
		GAM-2add-cor	0.260150
		GAM-4-full	0.255639

Feature Effect Error [ALE]



Feature Effect Examples [ALE]



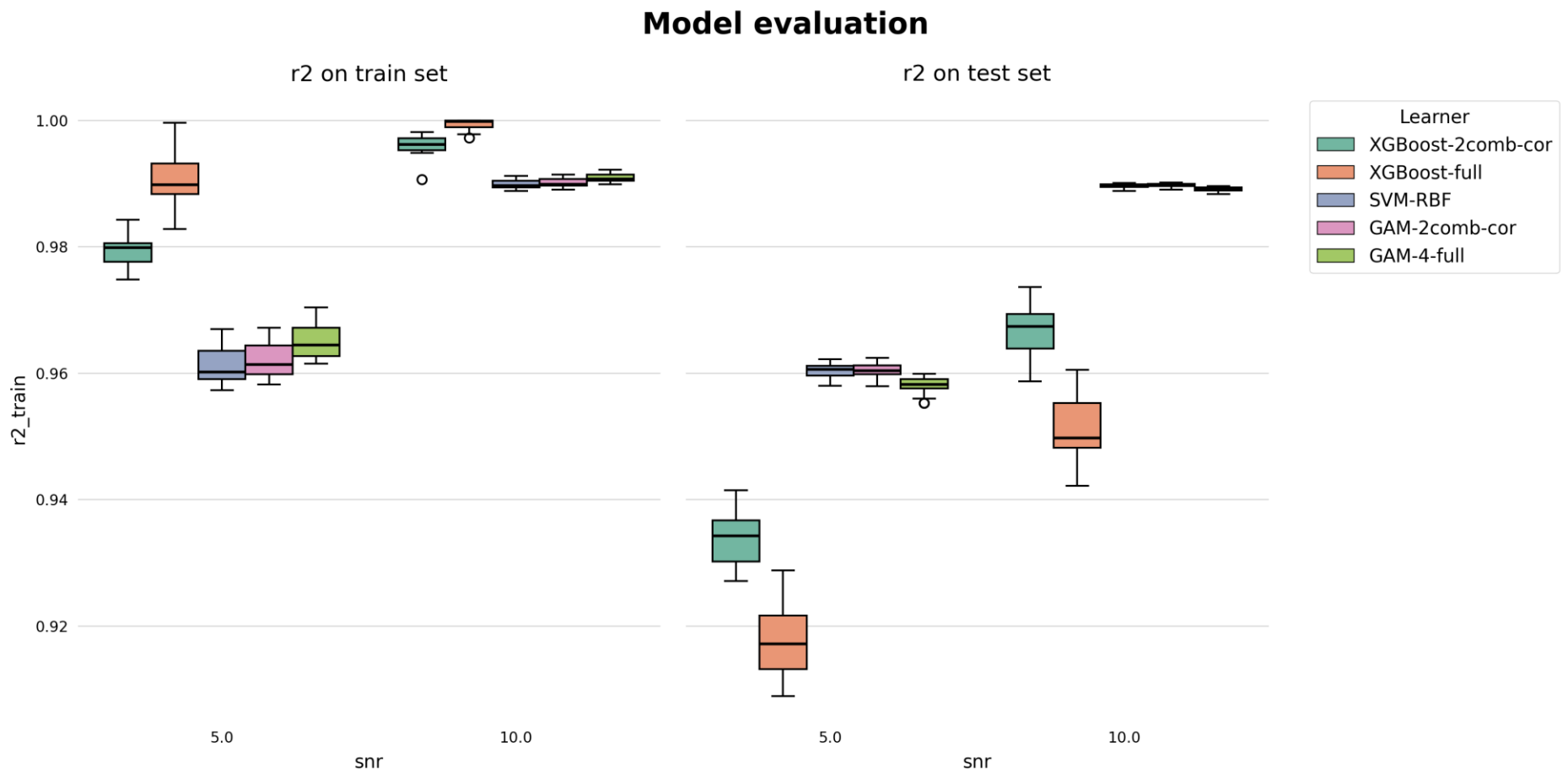
Limited comparability to PDP: intervals not equidistant

Correlation Analysis [ALE]

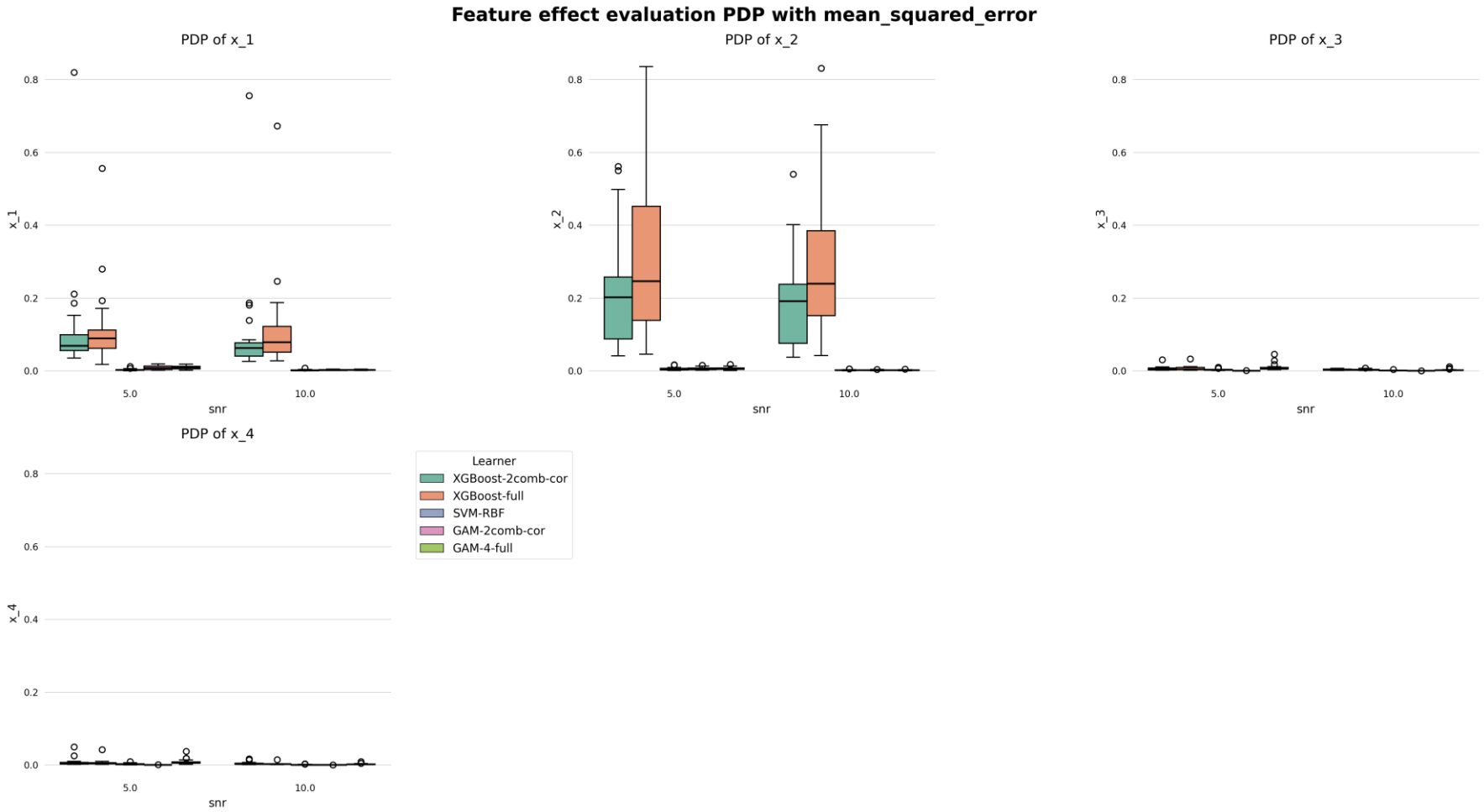
feature	snr	model	correlation
x_1	5.0	Overall	0.811029
		XGBoost-2add-cor	-0.004511
		XGBoost-full	0.261654
		SVM-RBF	0.434586
		GAM-2add-cor	-0.052632
		GAM-4-full	-0.245113
	10.0	Overall	0.781674
		XGBoost-2add-cor	0.135338
		XGBoost-full	0.215038
		SVM-RBF	0.285714
		GAM-2add-cor	-0.069173
		GAM-4-full	-0.230075
x_2	5.0	Overall	0.794059
		XGBoost-2add-cor	0.236090
		XGBoost-full	-0.276692
		SVM-RBF	0.204511
		GAM-2add-cor	0.239098
		GAM-4-full	0.344361
	10.0	Overall	0.763900
		XGBoost-2add-cor	0.078195
		XGBoost-full	-0.043609
		SVM-RBF	0.034586
		GAM-2add-cor	0.401504
		GAM-4-full	0.378947
x_3	5.0	Overall	0.750134
		XGBoost-2add-cor	-0.189474
		XGBoost-full	0.457143
		SVM-RBF	0.287218
		GAM-2add-cor	NaN
		GAM-4-full	0.124812
	10.0	Overall	0.867266
		XGBoost-2add-cor	-0.091729
		XGBoost-full	0.254135
		SVM-RBF	0.255639
		GAM-2add-cor	NaN
		GAM-4-full	0.069173
x_4	5.0	Overall	0.747893
		XGBoost-2add-cor	-0.132331
		XGBoost-full	0.311278
		SVM-RBF	-0.464662
		GAM-2add-cor	NaN
		GAM-4-full	0.315789
	10.0	Overall	0.864591
		XGBoost-2add-cor	-0.073684
		XGBoost-full	0.320301
		SVM-RBF	0.043609
		GAM-2add-cor	NaN
		GAM-4-full	0.221053

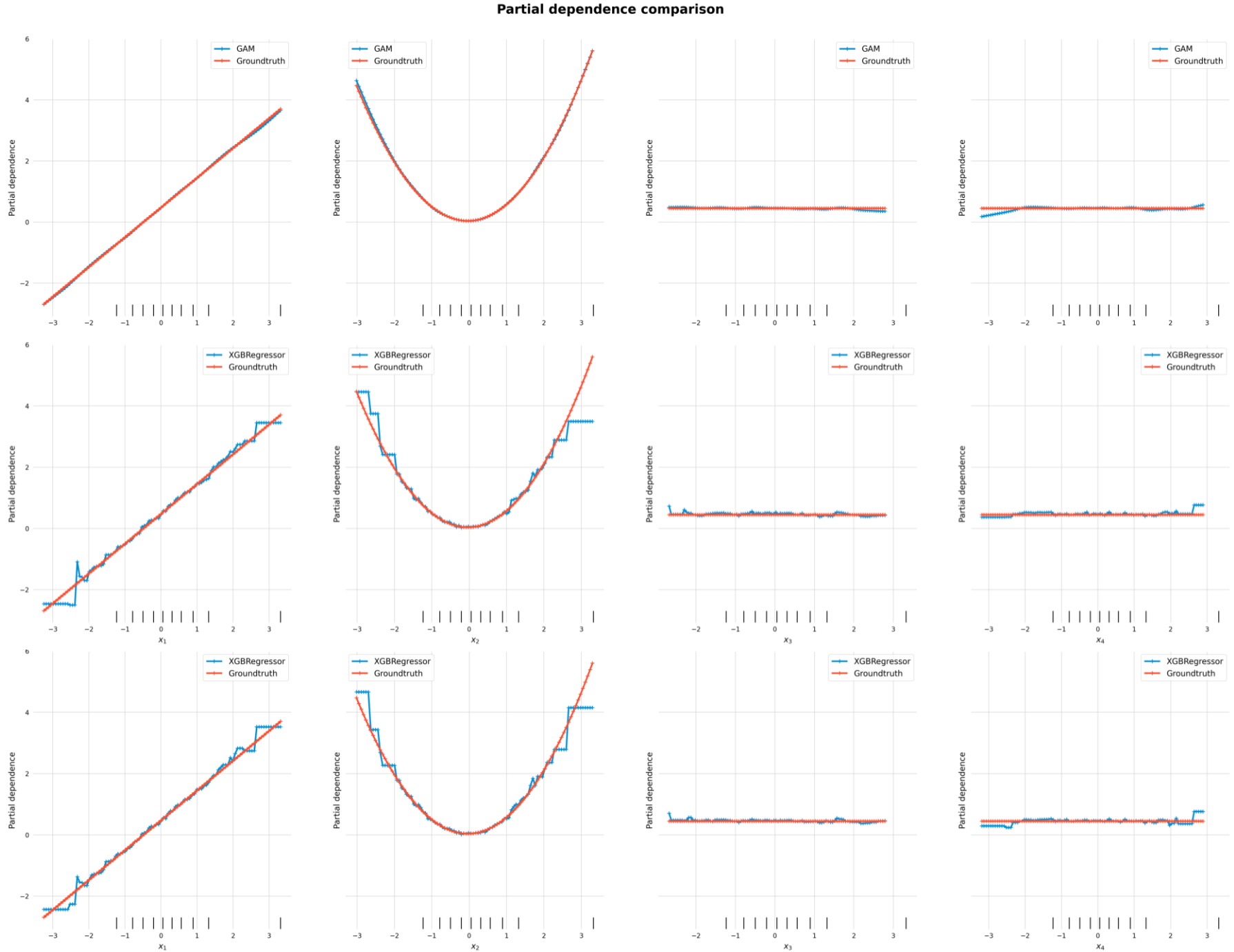
Results Combined Scenario

Model Performance Combined [R2]



Feature Effect Error Combined [PDP]





full

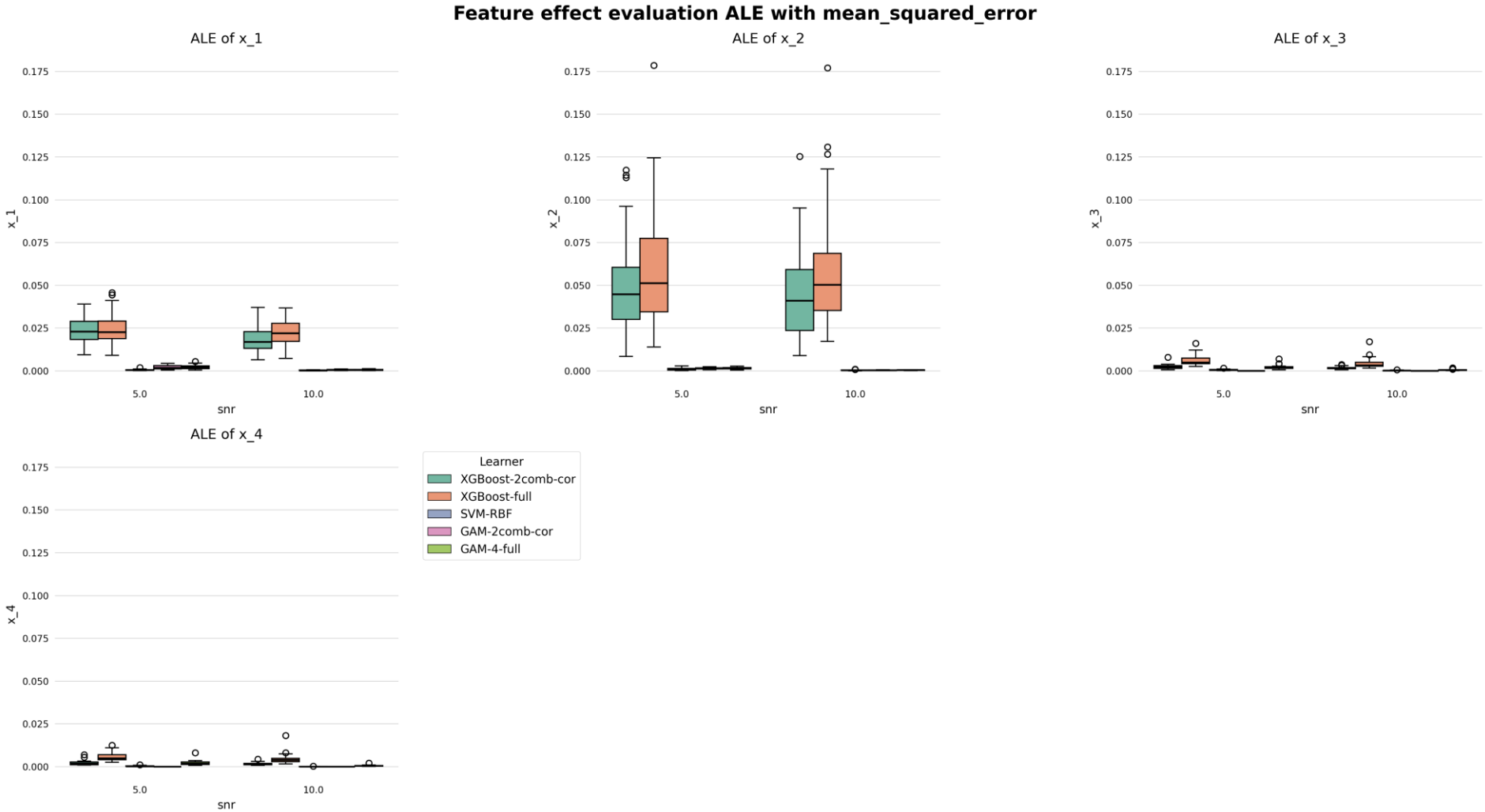
full

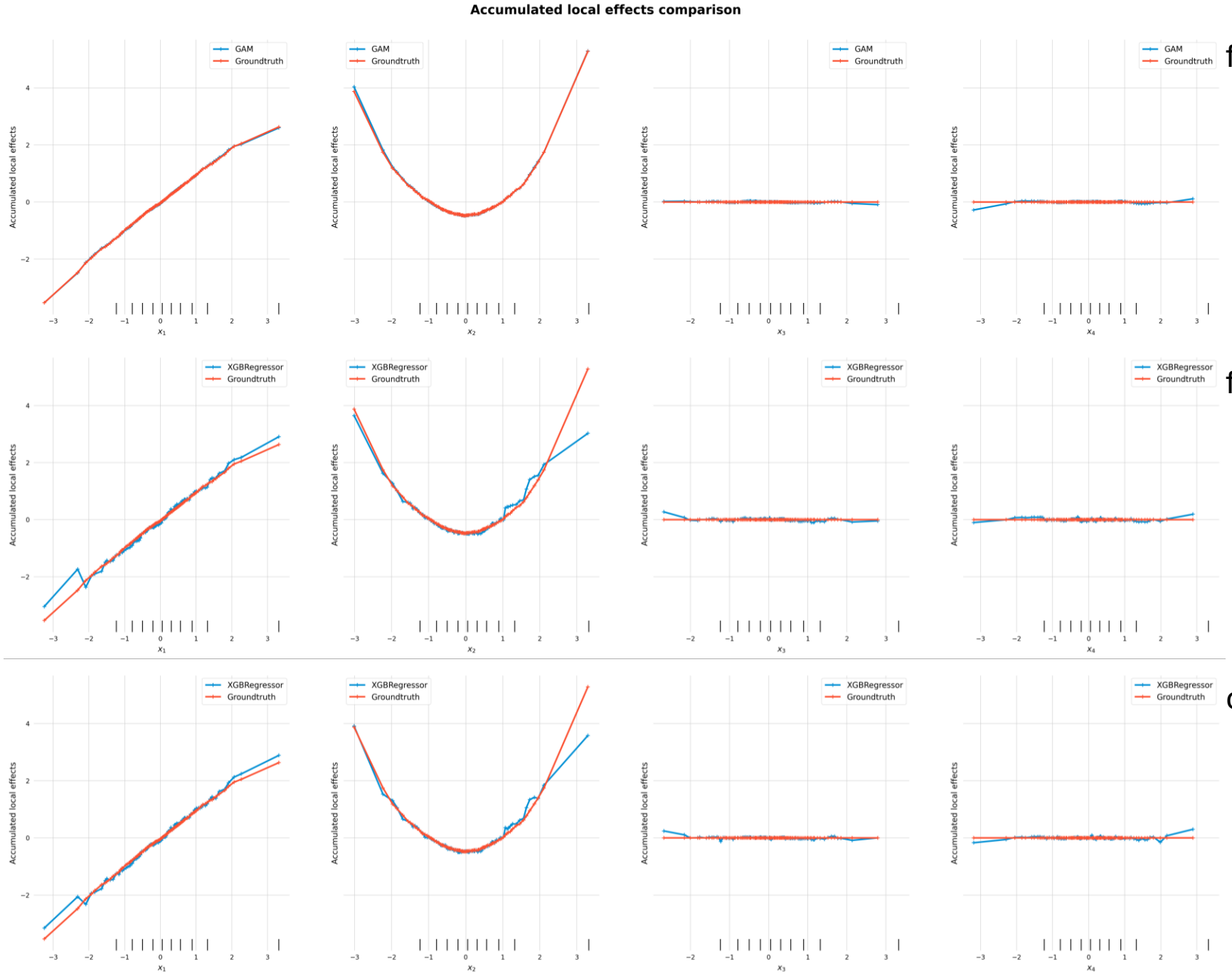
correctly specified

Correlation Analysis Combined [PDP]

feature	snr	model	correlation
x_1	5.0	Overall	0.802952
		XGBoost-2comb-cor	0.502256
		XGBoost-full	0.375940
		SVM-RBF	0.428571
		GAM-2comb-cor	0.009023
		GAM-4-full	0.024060
	10.0	Overall	0.791071
		XGBoost-2comb-cor	0.440602
		XGBoost-full	0.454135
		SVM-RBF	0.440602
		GAM-2comb-cor	0.001504
		GAM-4-full	0.009023
x_2	5.0	Overall	0.787375
		XGBoost-2comb-cor	0.027068
		XGBoost-full	-0.063158
		SVM-RBF	0.254135
		GAM-2comb-cor	0.303759
		GAM-4-full	0.278195
	10.0	Overall	0.782022
		XGBoost-2comb-cor	-0.108271
		XGBoost-full	-0.218045
		SVM-RBF	0.308271
		GAM-2comb-cor	0.485714
		GAM-4-full	0.353383
x_3	5.0	Overall	0.501818
		XGBoost-2comb-cor	-0.049624
		XGBoost-full	-0.311278
		SVM-RBF	0.033083
		GAM-2comb-cor	0.282707
		GAM-4-full	0.178947
	10.0	Overall	0.697930
		XGBoost-2comb-cor	0.105263
		XGBoost-full	-0.568421
		SVM-RBF	-0.144361
		GAM-2comb-cor	0.320301
		GAM-4-full	0.169925
x_4	5.0	Overall	0.539994
		XGBoost-2comb-cor	-0.195489
		XGBoost-full	-0.109774
		SVM-RBF	-0.210526
		GAM-2comb-cor	0.282707
		GAM-4-full	0.306767
	10.0	Overall	0.688833
		XGBoost-2comb-cor	0.060150
		XGBoost-full	-0.096241
		SVM-RBF	-0.192481
		GAM-2comb-cor	0.320301
		GAM-4-full	0.201504

Feature Effect Error Combined [ALE]





full

full

correctly specified

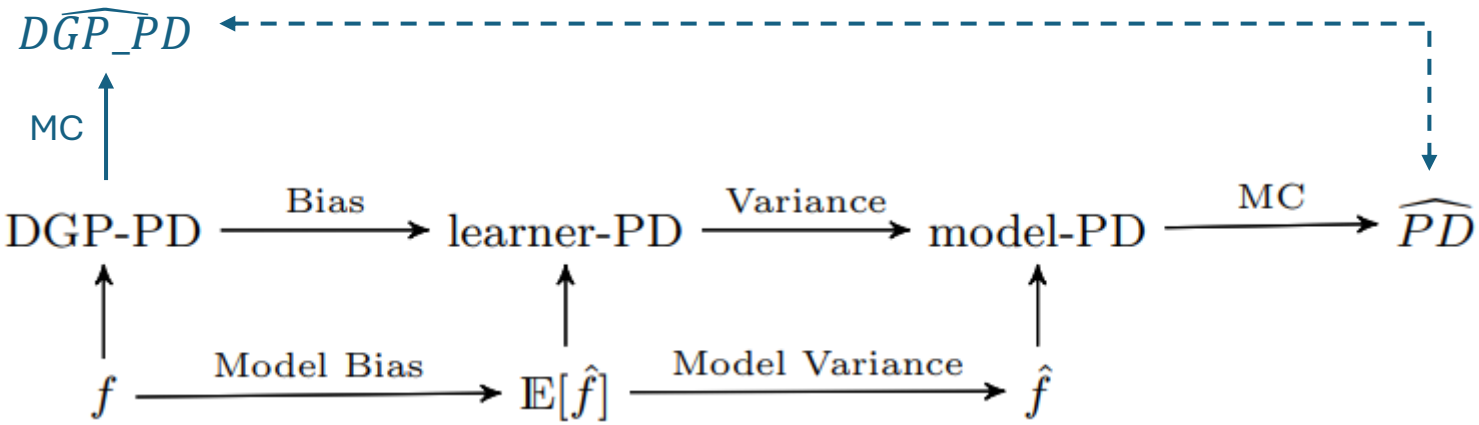
Correlation Analysis Combined [ALE]

feature	snr	model	correlation
x_1	5.0	Overall	0.791431
		XGBoost-2comb-cor	0.395489
		XGBoost-full	0.255639
		SVM-RBF	0.243609
		GAM-2comb-cor	-0.240602
		GAM-4-full	-0.270677
	10.0	Overall	0.770285
		XGBoost-2comb-cor	0.171429
		XGBoost-full	0.117293
		SVM-RBF	0.324812
		GAM-2comb-cor	-0.225564
		GAM-4-full	-0.309774
x_2	5.0	Overall	0.806613
		XGBoost-2comb-cor	-0.216541
		XGBoost-full	-0.327820
		SVM-RBF	0.311278
		GAM-2comb-cor	0.493233
		GAM-4-full	0.347368
	10.0	Overall	0.794479
		XGBoost-2comb-cor	-0.260150
		XGBoost-full	-0.390977
		SVM-RBF	0.203008
		GAM-2comb-cor	0.645113
		GAM-4-full	0.390977
x_3	5.0	Overall	0.826262
		XGBoost-2comb-cor	-0.064662
		XGBoost-full	0.272180
		SVM-RBF	-0.093233
		GAM-2comb-cor	NaN
		GAM-4-full	0.070677
	10.0	Overall	0.898757
		XGBoost-2comb-cor	0.103759
		XGBoost-full	0.102256
		SVM-RBF	-0.057143
		GAM-2comb-cor	NaN
		GAM-4-full	0.034586
x_4	5.0	Overall	0.830720
		XGBoost-2comb-cor	-0.001504
		XGBoost-full	-0.019549
		SVM-RBF	-0.335338
		GAM-2comb-cor	NaN
		GAM-4-full	0.178947
	10.0	Overall	0.901781
		XGBoost-2comb-cor	0.067669
		XGBoost-full	0.037594
		SVM-RBF	-0.324812
		GAM-2comb-cor	NaN
		GAM-4-full	0.218045

Next Steps

Next Steps

- Fix grid calculation for ALE
- Run remaining simulations
- Create error-graph: What do we want to find out?
- …


$\widehat{DGP\_PD}$
MC
DGP-PD
Bias
learner-PD
Variance
model-PD
MC
$\widehat{PD}$
f
Model Bias
$\mathbb{E}[\hat{f}]$
Model Variance
$\hat{f}$

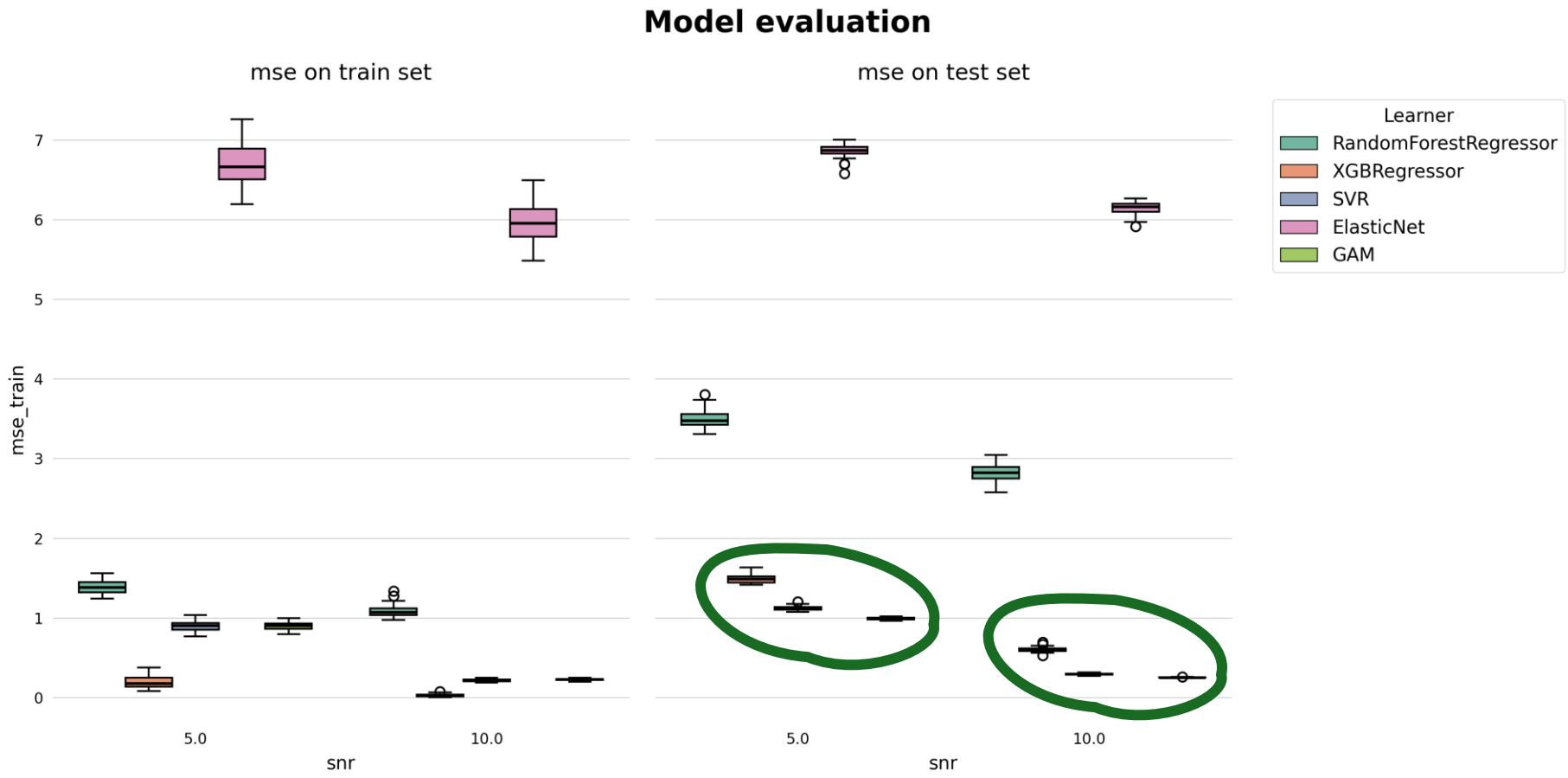
Back-Up: Results Friedman1

Simulation Setting

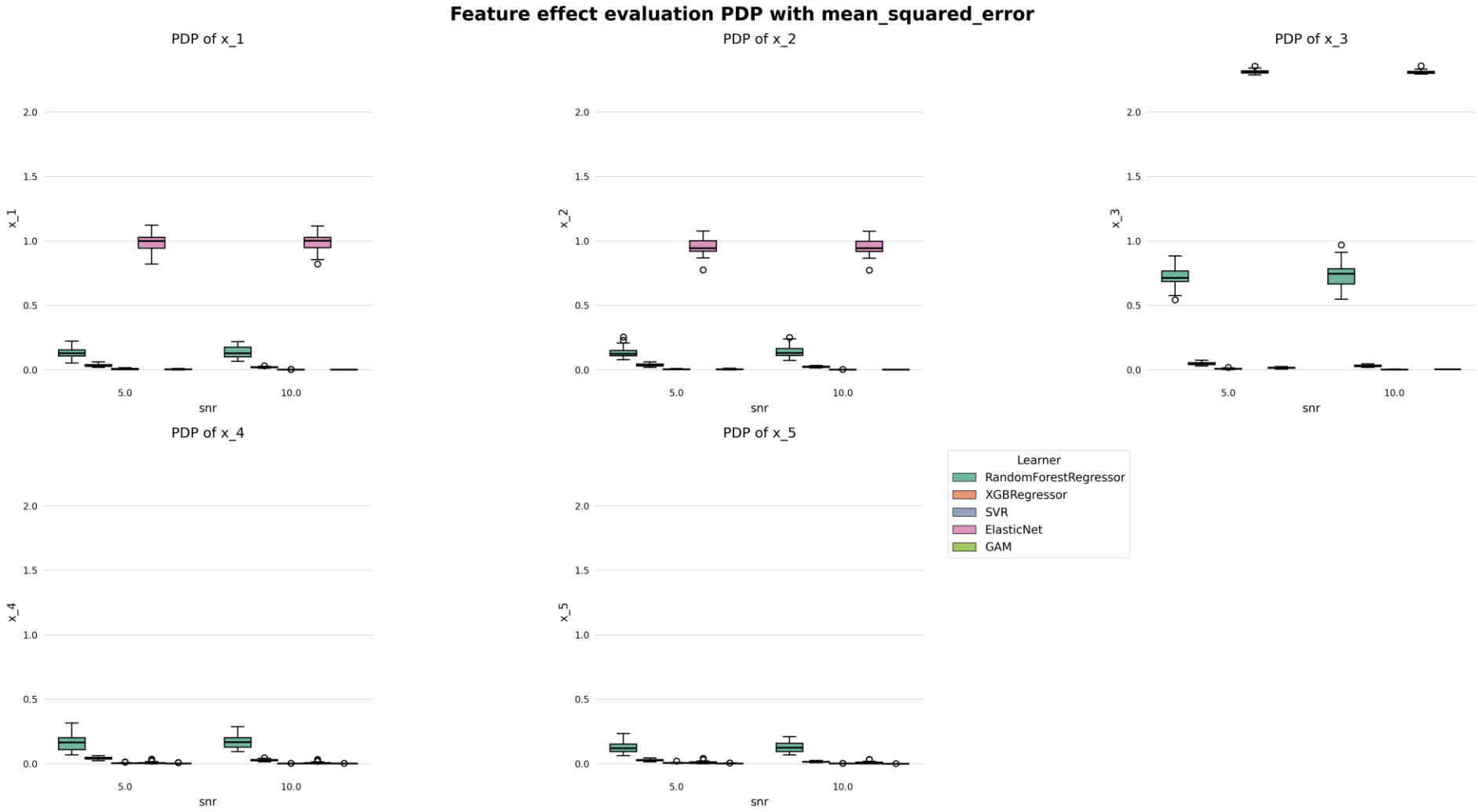
Same setting as before, with the exception of the following points:

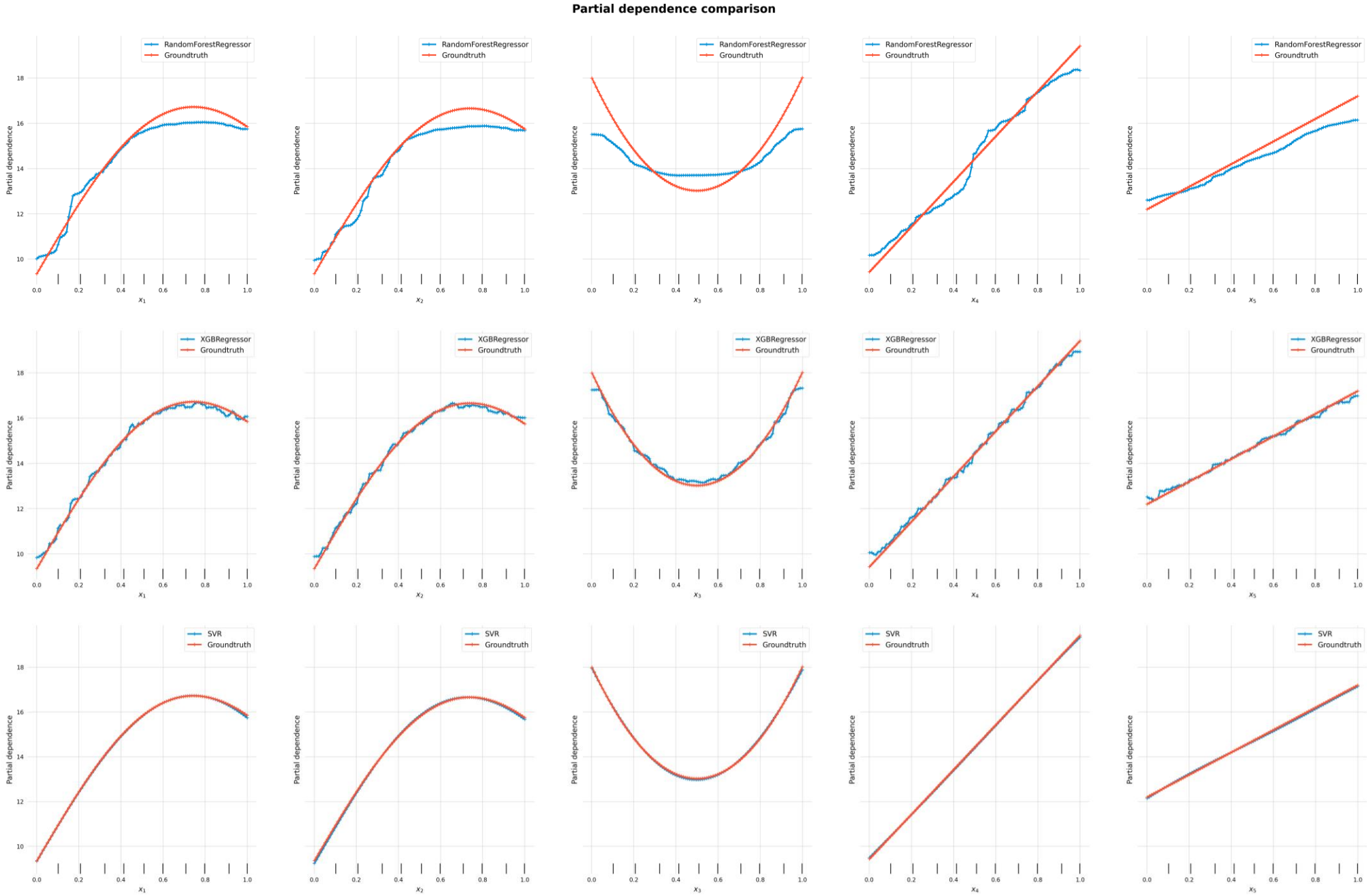
- uses Frieman1 dataset as groundtruth:
  - 5 standard-uniformly distributed, uncorrelated features (no additional noise features)
  - Friedman1 function: $f(x) = 10 * \sin(\pi * x_1 * x_2) + 20 * (x_3 - 0.5)^2 + 10 * x_4 + 5 * x_5 + e * N(0, 1)$
- Different models: RandomForest, XGBoost (full), SVM-RBF, GAM (correctly specified), ElasticNet

Model Results Friedman1

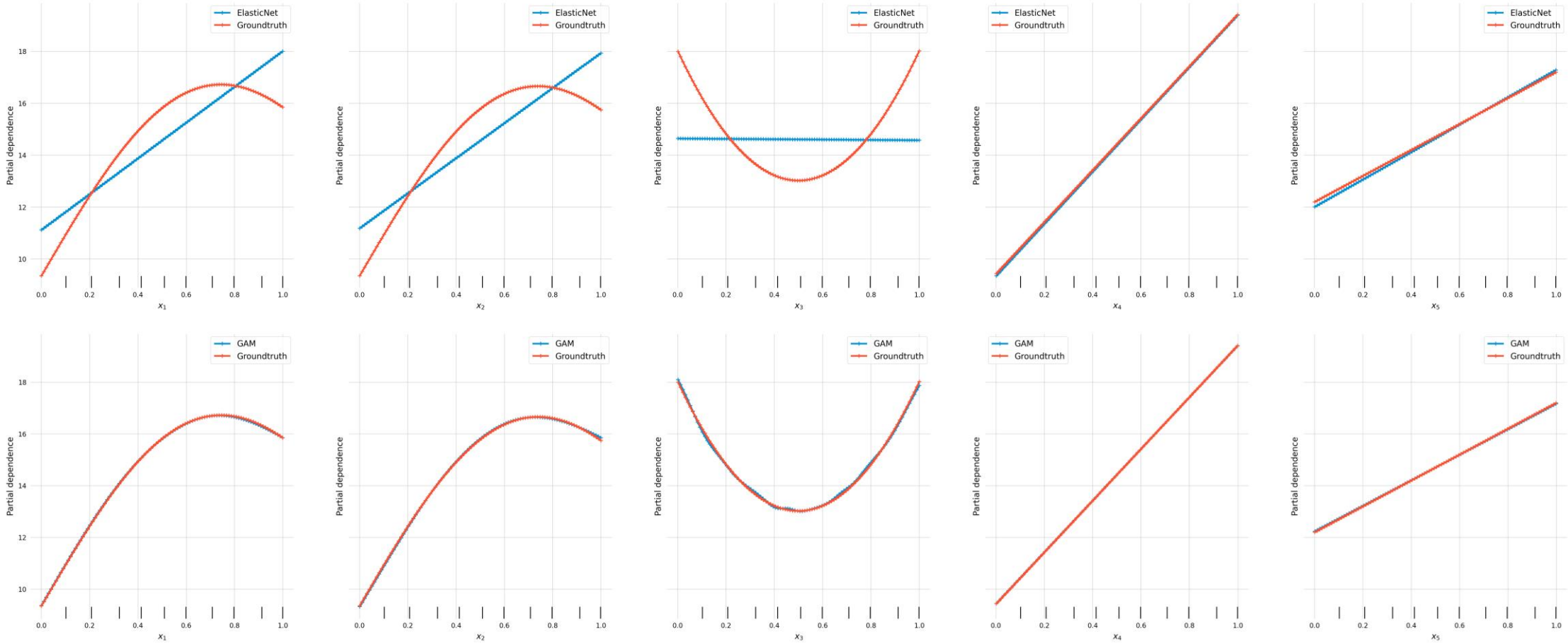


Feature Effect Error Friedman1 [PDP]

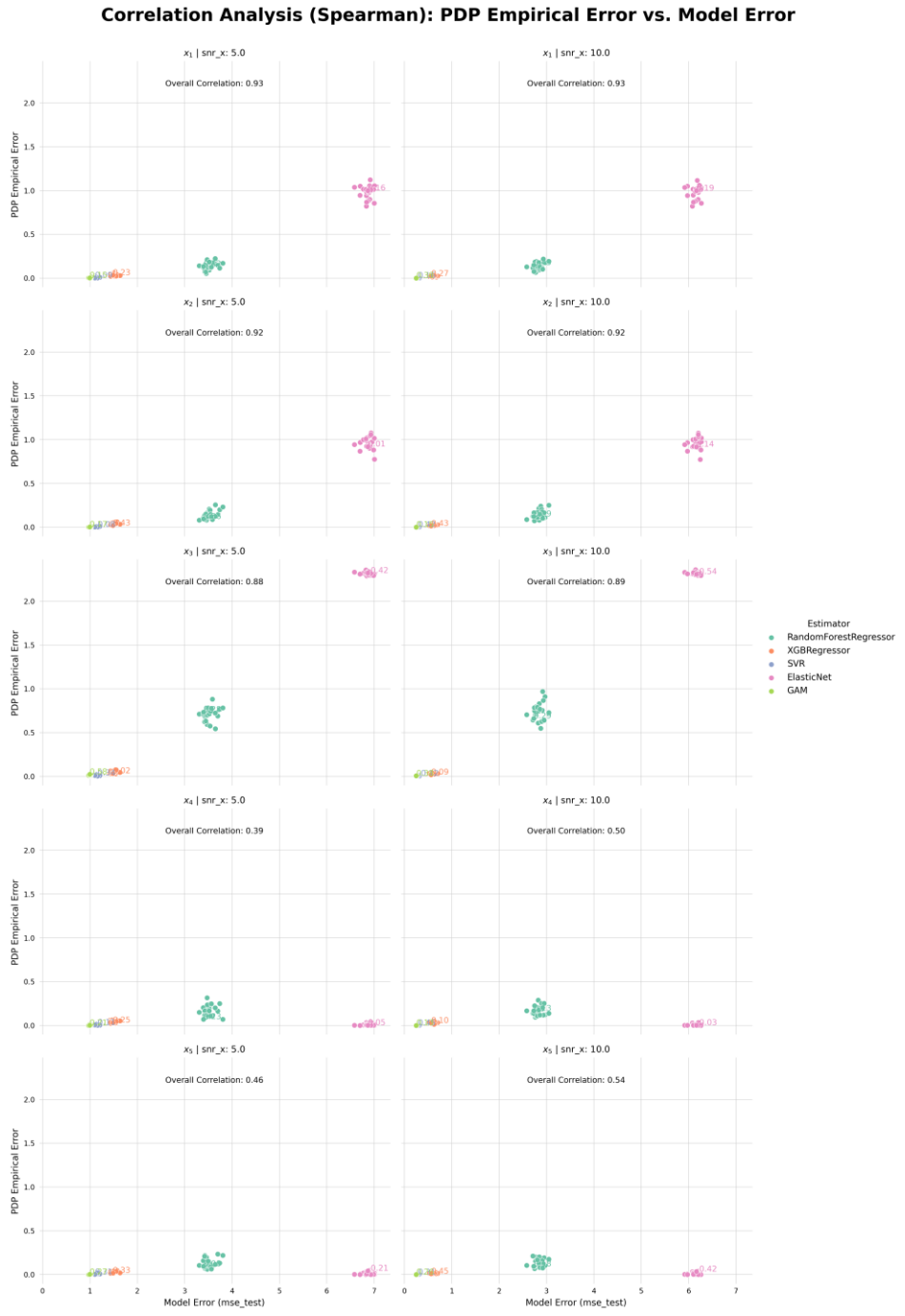



Partial dependence comparison
RandomForestRegressor
Groundtruth
XGBRegressor
Groundtruth
SVR
Groundtruth
Partial dependence
x_1
x_2
x_3
x_4
x_5

Partial dependence comparison

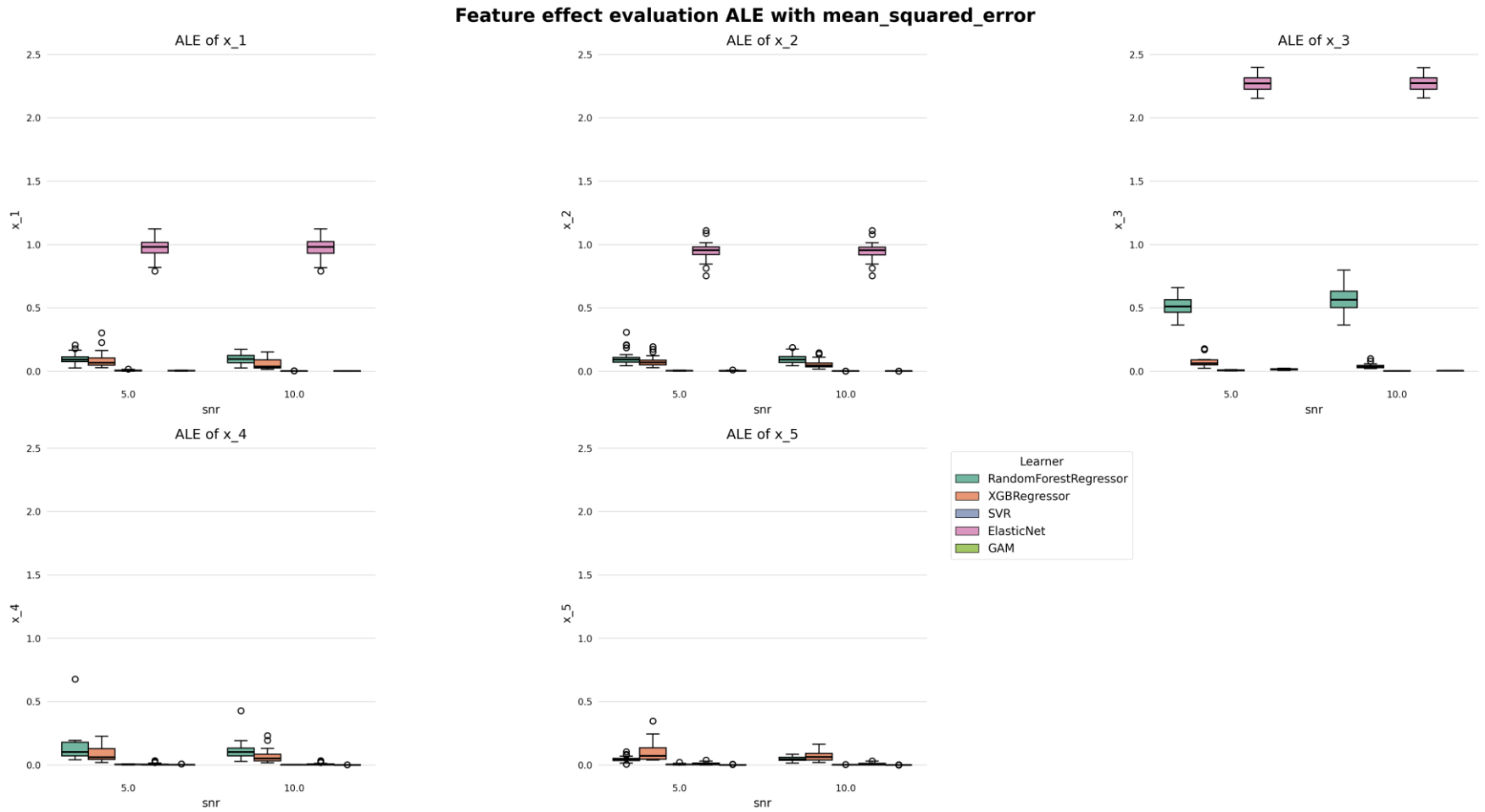


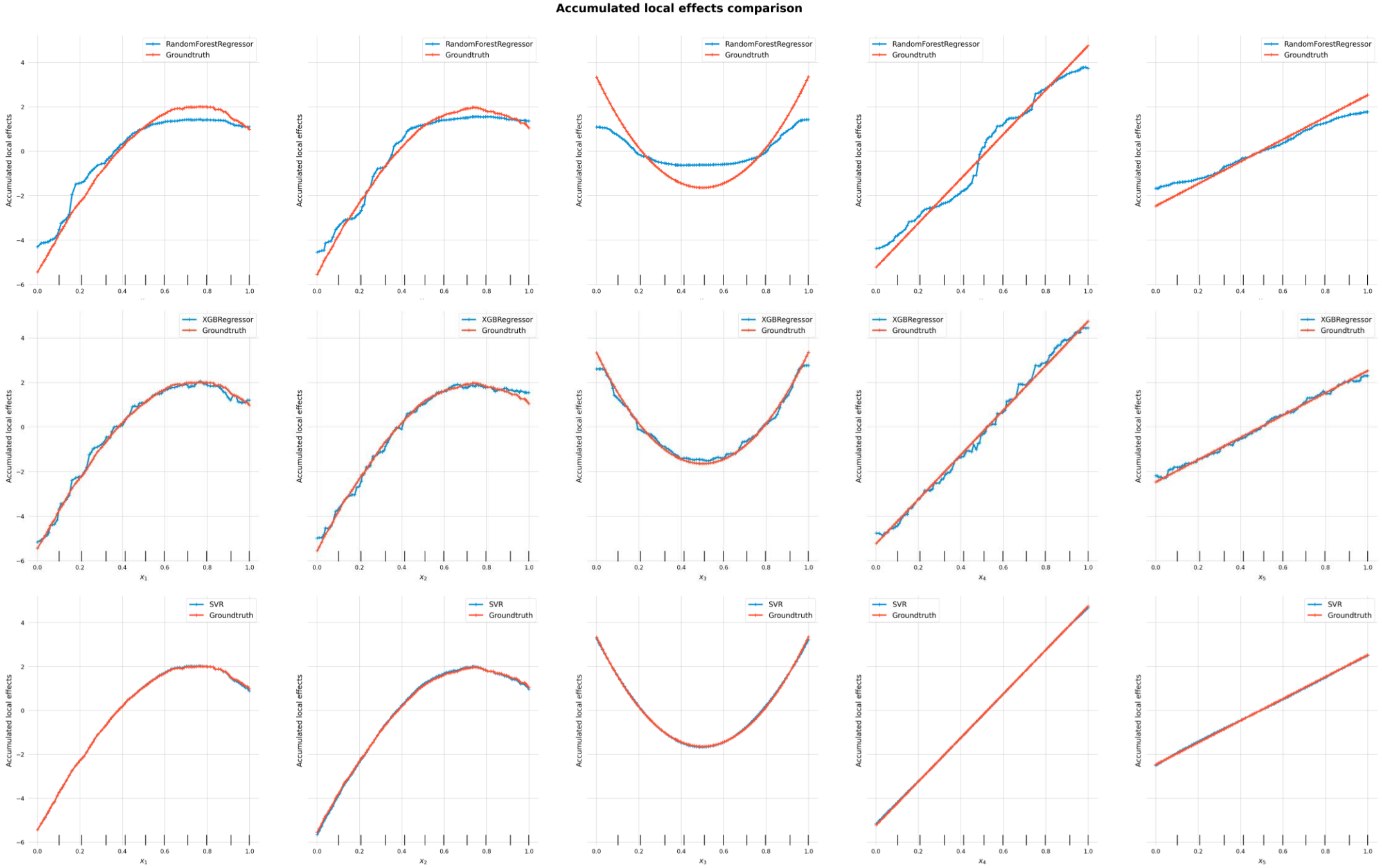
Correlation Analysis Friedman1 [PDP]

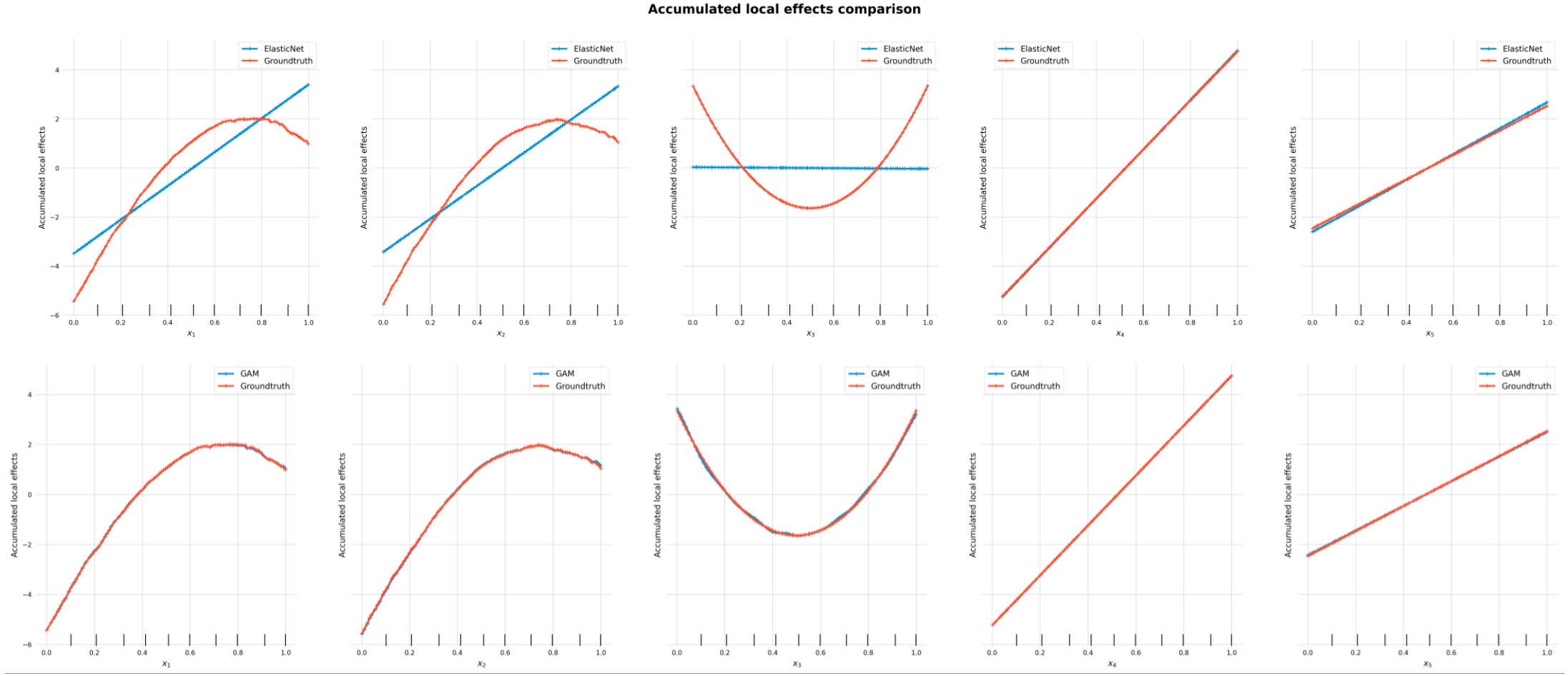


feature	snr	model	correlation
x_1	5.0	Overall	0.930390
		RandomForestRegressor	0.420769
		XGBRegressor	0.225385
		SVR	0.088462
		ElasticNet	0.155385
		GAM	0.154615
	10.0	Overall	0.933155
		RandomForestRegressor	0.425385
		XGBRegressor	0.273077
		SVR	0.159231
		ElasticNet	0.186923
		GAM	0.376923
x_2	5.0	Overall	0.924381
		RandomForestRegressor	0.557692
		XGBRegressor	0.427692
		SVR	-0.081538
		ElasticNet	0.013077
		GAM	0.173846
	10.0	Overall	0.922495
		RandomForestRegressor	0.290769
		XGBRegressor	0.433846
		SVR	0.420769
		ElasticNet	0.135385
		GAM	0.152308
x_3	5.0	Overall	0.877217
		RandomForestRegressor	0.277692
		XGBRegressor	0.024615
		SVR	-0.340769
		ElasticNet	-0.421538
		GAM	0.581538
	10.0	Overall	0.887152
		RandomForestRegressor	0.286923
		XGBRegressor	0.090000
		SVR	-0.115385
		ElasticNet	-0.536923
		GAM	0.383846
x_4	5.0	Overall	0.394482
		RandomForestRegressor	0.229231
		XGBRegressor	0.252308
		SVR	-0.112308
		ElasticNet	0.045385
		GAM	0.210769
	10.0	Overall	0.501512
		RandomForestRegressor	0.128462
		XGBRegressor	0.101538
		SVR	0.448462
		ElasticNet	-0.028462
		GAM	0.128462
x_5	5.0	Overall	0.455920
		RandomForestRegressor	0.080769
		XGBRegressor	0.330000
		SVR	0.313077
		ElasticNet	-0.208462
		GAM	0.269231
	10.0	Overall	0.543318
		RandomForestRegressor	0.176923
		XGBRegressor	0.446923
		SVR	0.306923
		ElasticNet	-0.418462
		GAM	0.260769

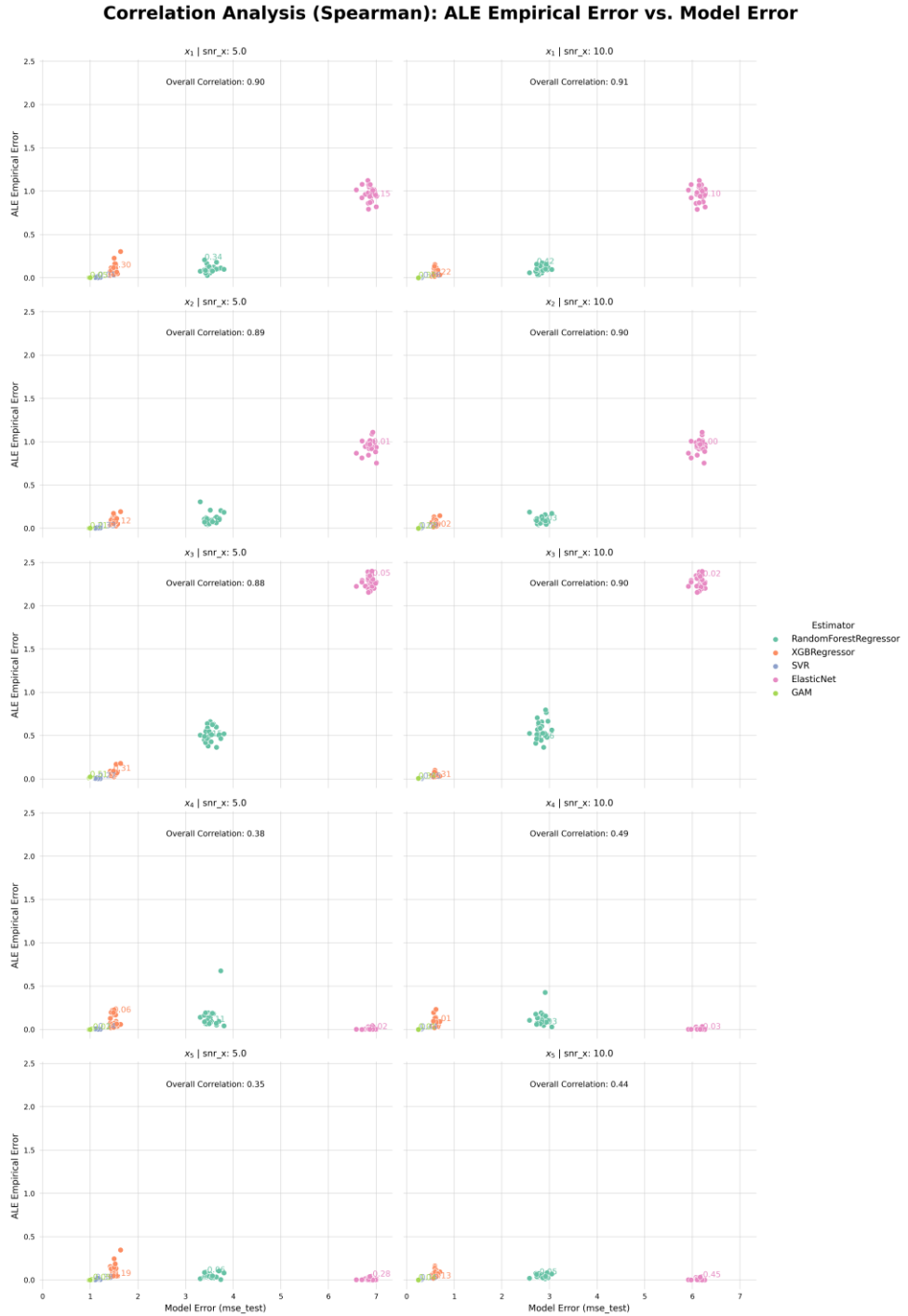
Feature Effect Error Friedman1 [ALE]




Accumulated local effects comparison
RandomForestRegressor
Groundtruth
XGBRegressor
Groundtruth
SVR
Groundtruth
Accumulated local effects
x_1
x_2
x_3
x_4
x_5


Accumulated local effects comparison
ElasticNet
Groundtruth
GAM
Groundtruth
Accumulated local effects
x_1
x_2
x_3
x_4
x_5

Correlation Analysis Friedman1 [ALE]



feature	snr	model	correlation
x_1	5.0	Overall	0.895201
		RandomForestRegressor	0.339231
		XGBRegressor	0.297692
		SVR	-0.013077
		ElasticNet	-0.153846
		GAM	0.048462
	10.0	Overall	0.905493
		RandomForestRegressor	0.423846
		XGBRegressor	0.216154
		SVR	-0.056154
		ElasticNet	-0.097692
		GAM	0.348462
x_2	5.0	Overall	0.887736
		RandomForestRegressor	0.090769
		XGBRegressor	0.116923
		SVR	-0.330000
		ElasticNet	-0.013077
		GAM	0.207692
	10.0	Overall	0.897585
		RandomForestRegressor	-0.026923
		XGBRegressor	0.015385
		SVR	0.088462
		ElasticNet	0.002308
		GAM	0.218462
x_3	5.0	Overall	0.879515
		RandomForestRegressor	0.150769
		XGBRegressor	0.310769
		SVR	-0.274615
		ElasticNet	-0.049231
		GAM	0.510769
	10.0	Overall	0.895644
		RandomForestRegressor	0.258462
		XGBRegressor	0.308462
		SVR	-0.050000
		ElasticNet	-0.023846
		GAM	0.342308
x_4	5.0	Overall	0.380664
		RandomForestRegressor	-0.110000
		XGBRegressor	0.055385
		SVR	-0.136923
		ElasticNet	0.021538
		GAM	-0.022308
	10.0	Overall	0.488762
		RandomForestRegressor	-0.032308
		XGBRegressor	0.011538
		SVR	0.443846
		ElasticNet	-0.029231
		GAM	0.014615
x_5	5.0	Overall	0.353942
		RandomForestRegressor	-0.058462
		XGBRegressor	0.186154
		SVR	0.376923
		ElasticNet	-0.283077
		GAM	-0.048462
	10.0	Overall	0.441886
		RandomForestRegressor	-0.054615
		XGBRegressor	0.131538
		SVR	0.147692
		ElasticNet	-0.448462
		GAM	0.036923