UiO : Department of Mathematics
University of Oslo

Explaining individual predictions when features are dependent

More accurate approximations to Shapley values

Lars H. B. Olsen



Table of contents

1. **Authors**
2. **Explainable artificial intelligence (XAI)**
3. **Shapley values**
   - Example: *julebord*
4. **Shapley values as a model-agnostic explanation framework**
5. **Modeling feature dependencies**
   - Gaussian
   - Gaussian Copula
   - Empirical
6. **Simulation studies**
7. **Real data example**
8. **Conclusion with self-advertisement**
9. **References**

Authors





- Kjersti Aas (Dr.Philos): Asst. Research Director SAMBA at NR, big in copulas.
- Martin Jullum (Ph.D): Senior Research Scientist at NR, Nils' Ph.D. candidate.
- Anders Løland (Dr.Philos): Asst. Research Director SAMBA at NR.

Part of BigInsight's XAI group.

Classical machine learning scenario and XAI

Supervised learning task:

- We have a training set $\{y^{(i)}, \boldsymbol{x}^{(i)}\}_{i=1,\ldots,n_{\text{train}}}$.
- Fit a ML predictive model $f(\boldsymbol{x})$ which resembles the response y.
  - Boosted models, random forest, NN, *et cetera*.
  - They often outperform statistical models, but at a cost of interpretability (black-box).
- Goal: make black-box models transparent.
  - Why? Understand the underlying decision process, debugging the model, EU's *right to an explanation* (GDPR): *[the data subject should have] the right ... to obtain an explanation of the decision reached.* Breiman: make culture B more into culture A.
- Many types of explanations:
  - Model-specific (assume model structure, e.g. layers in NNs) *vs.* model-agnostic (change input, measures change in predictive output).
  - Global (model behaviour across all data) *vs.* local (individual prediction).
- Shapley values: model-agnostic explanation framework with local explanations.
  - Explain the prediction $f(\boldsymbol{x}^*)$ for a specific feature vector $\boldsymbol{x} = \boldsymbol{x}^*$.

Shapley values

- Originated in cooperative game theory (Shapley, 1953). (Nobel economics, 2012)
- Solution concept of how to fairly divide a payout $v(\mathcal{M})$ of a game among the players?

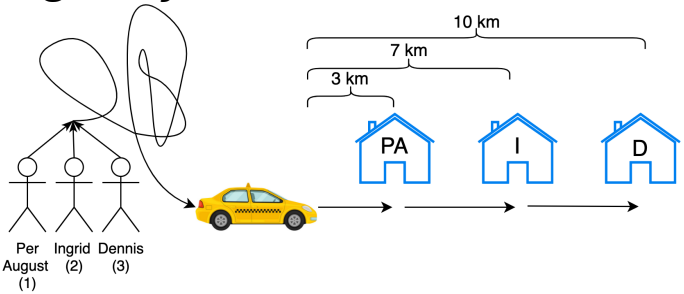
Shapley values

$$\varphi_j = \varphi_j(v) = \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \frac{|\mathcal{S}|!(M - |\mathcal{S}| - 1)!}{M!} \left(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S}) \right)$$

- Where v is the contribution function, i.e. $v(\mathcal{S})$ is the worth of coalition $\mathcal{S}$.
- A weighted average of marginal contributions over all coalitions.
- *Shapley values* are the unique solution to the following axioms:
  - *Efficiency*: $\sum_{j=0}^{M} \varphi_j = v(\mathcal{M})$.
  - *Symmetry*: two players i and j contribute equally $\Rightarrow \varphi_i = \varphi_j$.
  - *Dummy player*: player j no contribution $\Rightarrow \varphi_j = 0$.
  - *Linearity*: $\varphi_j(av + bw) = a\varphi_j(v) + b\varphi_j(w)$.

Example: fairly splitting the *julebord* taxi fare

- $v(\emptyset) = 0$,
- $v(\{1\}) = 3$, $v(\{2\}) = 7$, $v(\{3\}) = 10$,
- $v(\{1,2\}) = 7$, $v(\{1,3\}) = 10$, $v(\{2,3\}) = 10$,
- $v(\{1,2,3\}) = 10$



$$\phi_1 = \frac{1}{3}\left(v(\{1,2,3\} - v(\{2,3\})\right) + \frac{1}{6}\left(v(\{1,2\}) - v(\{2\})\right) + \frac{1}{6}\left(v(\{1,3\}) - v(\{3\})\right) + \frac{1}{3}\left(v(\{1\}) - v(\emptyset)\right) = 1$$

$$\phi_2 = \frac{1}{3}\left(v(\{1,2,3\} - v(\{1,3\})\right) + \frac{1}{6}\left(v(\{1,2\}) - v(\{1\})\right) + \frac{1}{6}\left(v(\{2,3\}) - v(\{3\})\right) + \frac{1}{3}\left(v(\{2\}) - v(\emptyset)\right) = 3$$

$$\phi_3 = \frac{1}{3}\left(v(\{1,2,3\} - v(\{1,2\})\right) + \frac{1}{6}\left(v(\{1,3\}) - v(\{1\})\right) + \frac{1}{6}\left(v(\{2,3\}) - v(\{2\})\right) + \frac{1}{3}\left(v(\{3\}) - v(\emptyset)\right) = 6$$

Shapley values as a model explanation framework

$$\varphi_j = \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \frac{|\mathcal{S}|!(M - |\mathcal{S}| - 1)!}{M!} \left(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S}) \right)$$

- The coalition game is replaced by the prediction model f, the players are the features, and the individual prediction $f(\boldsymbol{x}^*)$ is the payout to be distributed.
- Efficiency axiom: $\sum_{j=1}^{M} \varphi_j = f(\boldsymbol{x}^*) - \varphi_0$, where $\varphi_0 = v(\emptyset) = \mathrm{E}[f(\boldsymbol{x})]$ quantifies how much of a prediction is not due to any of the features.
  - If $|\varphi_0| \gg |\varphi_j|$, then feature j is said to be *unimportant* for that specific prediction.
- Problems in this setup:
  - Computationally expensive:
    the number of terms grows exponentially with number of features: $\mathcal{O}(2^M)$.
  - What should the contribution function be?
    Should say something about the importance of the features $\boldsymbol{x}^*_{\mathcal{S}}$ for the prediction $f(\boldsymbol{x}^*)$.

Lundberg and Lee (2017)

They "solved" both problems with their `KernelSHAP` methodology:

- Computational tractable Shapley values approximations:
  - Solution of a weighted least squares problem: $\varphi = (Z^T W Z)^{-1} Z^T W v = Rv$.
    - Z is the binary matrix ($2^M \times (M+1)$) with all possible coalitions.
    - W is the diagonal matrix ($2^M \times 2^M$) with the Shapley kernel weights.
    - v is the vector (2^M) of contribution function values.
    - R only needs to be computed once, while v change with individual.
  - Importance sampling w.r.t. Shapley kernel weights: evaluate a subset $\mathcal{D} \overset{\text{replace}}{\subseteq} \mathcal{M}$.
    - $\varphi \approx (Z_{\mathcal{D}}^T W_{\mathcal{D}} Z_{\mathcal{D}})^{-1} Z_{\mathcal{D}}^T W_{\mathcal{D}} v_{\mathcal{D}} = R_{\mathcal{D}} v_{\mathcal{D}}$.
- Defined the contribution function to be:

$$v(\mathcal{S}) = \mathsf{E}\left[f(\boldsymbol{x})|\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}^*\right] = \mathsf{E}\left[f(\boldsymbol{x}_{\bar{\mathcal{S}}}, \boldsymbol{x}_{\mathcal{S}})|\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}^*\right] = \int f(\boldsymbol{x}_{\bar{\mathcal{S}}}, \boldsymbol{x}_{\mathcal{S}}^*) p(\boldsymbol{x}_{\bar{\mathcal{S}}}|\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}^*) \, \mathrm{d}\boldsymbol{x}_{\bar{\mathcal{S}}}$$

$$\approx \frac{1}{K} \sum_{k=1}^{K} f(\boldsymbol{x}_{\bar{\mathcal{S}}}^k, \boldsymbol{x}_{\mathcal{S}}^*), \text{ where } \boldsymbol{x}_{\bar{\mathcal{S}}}^k \sim p(\boldsymbol{x}_{\bar{\mathcal{S}}}|\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}^*).$$

- Lundberg & Lee assumed feature independence $p(\boldsymbol{x}_{\bar{\mathcal{S}}}|\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}^*) = p(\boldsymbol{x}_{\bar{\mathcal{S}}})$ for simplicity.

Aas, Jullum and Løland (2021) I

Generate on-distribution values $\boldsymbol{x}_{\bar{\mathcal{S}}}^k \sim p(\boldsymbol{x}_{\bar{\mathcal{S}}} | \boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}^*)$.
Otherwise, cannot trust the Shapley values as f is evaluated on data off-distribution.

- Gaussian:
  1. Assume $\boldsymbol{x} \sim \mathcal{N}_M(\mu, \Sigma)$, where $\mu = (\mu_{\mathcal{S}}, \mu_{\bar{\mathcal{S}}})^T$ and $\Sigma = \begin{bmatrix} \Sigma_{\mathcal{S}\mathcal{S}} & \Sigma_{\mathcal{S}\bar{\mathcal{S}}} \\ \Sigma_{\bar{\mathcal{S}}\mathcal{S}} & \Sigma_{\bar{\mathcal{S}}\bar{\mathcal{S}}} \end{bmatrix}$.
  2. The conditional distributions are then known: $p(\boldsymbol{x}_{\bar{\mathcal{S}}} | \boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}^*) = \mathcal{N}_{|\bar{\mathcal{S}}|}(\mu_{\bar{\mathcal{S}}|\mathcal{S}}, \Sigma_{\bar{\mathcal{S}}|\mathcal{S}})$, where $\mu_{\bar{\mathcal{S}}|\mathcal{S}} = \mu_{\bar{\mathcal{S}}} + \Sigma_{\bar{\mathcal{S}}\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}(\boldsymbol{x}_{\mathcal{S}}^* - \mu_{\mathcal{S}})$ and $\Sigma_{\bar{\mathcal{S}}|\mathcal{S}} = \Sigma_{\bar{\mathcal{S}}\bar{\mathcal{S}}} - \Sigma_{\bar{\mathcal{S}}\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\bar{\mathcal{S}}}$.

- Gaussian copula:
  1. Convert each empirical marginal X_j to Gaussian feature $V_j = \Phi^{-1}(\hat{F}_j(X_j))$.
  2. Assume $V \sim \mathcal{N}_M(\mu, \Sigma)$ and sample from $p(\boldsymbol{v}_{\bar{\mathcal{S}}} | \boldsymbol{v}_{\mathcal{S}} = \boldsymbol{v}_{\mathcal{S}}^*)$ using method above.
  3. Convert margins V_j in the conditional dist. to orginal dist. by $\hat{X}_j = \hat{F}_j^{-1}(\Phi(V_j))$.

Aas, Jullum and Løland (2021) II

Empirical:

- Best when the dependence structure or marginal dist. of $\boldsymbol{x}$ are far from Gaussian.
- Idea is that samples $(\boldsymbol{x}_{\bar{\mathcal{S}}}, \boldsymbol{x}_{\mathcal{S}})$ with $\boldsymbol{x}_{\mathcal{S}}$ close $\boldsymbol{x}_{\mathcal{S}}^*$ are informative about $p(\boldsymbol{x}_{\bar{\mathcal{S}}}|\boldsymbol{x}_{\mathcal{S}}^*)$.
  1. Compute (scaled) Mahalanobis distances: $\mathrm{D}_{\mathcal{S}}(\boldsymbol{x}^*, \boldsymbol{x}^{(i)}) = \sqrt{\frac{(\boldsymbol{x}_{\mathcal{S}}^* - \boldsymbol{x}_{\mathcal{S}}^{(i)})^T \hat{\Sigma}_{\mathcal{S}}^{-1} (\boldsymbol{x}_{\mathcal{S}}^* - \boldsymbol{x}_{\mathcal{S}}^{(i)})}{|\mathcal{S}|}}$, where $\hat{\Sigma}_{\mathcal{S}}$ is the sample covariance matrix for the n_{train} instances of $\boldsymbol{x}_{\mathcal{S}}$.
  2. Compute weights using the Gaussian distribution kernel: $w_{\mathcal{S}}(\boldsymbol{x}^*, \boldsymbol{x}^{(i)}) = \exp\left(-\frac{\mathrm{D}_{\mathcal{S}}(\boldsymbol{x}^*, \boldsymbol{x}^{(i)})^2}{2\sigma^2}\right)$, where σ is a smoothing/bandwidth parameter.
  3. Sort $w_{\mathcal{S}}(\boldsymbol{x}^*, \boldsymbol{x}^{(i)})$ in increasing order and let $\boldsymbol{x}^{[k]}$ be the instance with kth largest weight.
  4. Estimate the contribution function as $v_{\text{cond}}(\mathcal{S}) = \frac{\sum_{k=1}^{K} w_{\mathcal{S}}(\boldsymbol{x}^*, \boldsymbol{x}^{[k]}) f(\boldsymbol{x}_{\mathcal{S}}^{[k]}, \boldsymbol{x}_{\mathcal{S}}^*)}{\sum_{k=1}^{K} w_{\mathcal{S}}(\boldsymbol{x}^*, \boldsymbol{x}^{[k]})}$, where $K = \min_{L \in \mathbb{N}} \left\{ \frac{\sum_{k=1}^{L} w_{\mathcal{S}}(\boldsymbol{x}^*, \boldsymbol{x}^{[k]})}{\sum_{i=1}^{n_{\text{train}}} w_{\mathcal{S}}(\boldsymbol{x}^*, \boldsymbol{x}^{i})} > \eta \right\}$ and η can be e.g. 0.9.
- Small σ (dep. data): low bias, high var. Large σ (ind. data): high bias, low var.
- Regression problem, Nadaraya-Watson estimator, corrected AIC to select σ. Three approaches; exact AICc, approximate AICc, and fixed $\sigma = 0.1$.

Aas, Jullum and Løland (2021) III

Combined:

- Empirical works well for $\dim(\boldsymbol{x}_S) \leq D^*$.
- Gaussian and Gaussian copula are better when we condition on many features.
- Combine empirical with Gaussian or Gaussian copula.
- Set D^* manually based on experience.
  - 3 in the 10-dimensional simulation study.

Simulation studies: dimension 3

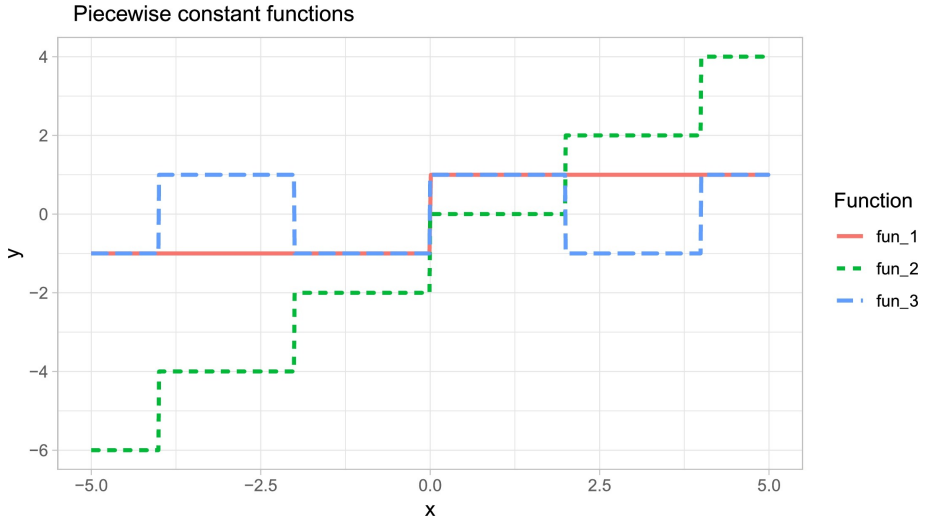
Evaluate the precision of the Shapley value estimates of the different methods.

- Evaluation measures:
  - MAE(method q) $= \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \frac{1}{M} \sum_{j=1}^{M} |\varphi_{j,\text{true}}(\boldsymbol{x}_i) - \varphi_{j,q}(\boldsymbol{x}_i)|$.
  - Skill(MAE, method q) $= 1 - \text{MAE(method } q) \big/ \text{MAE(method original)}$.

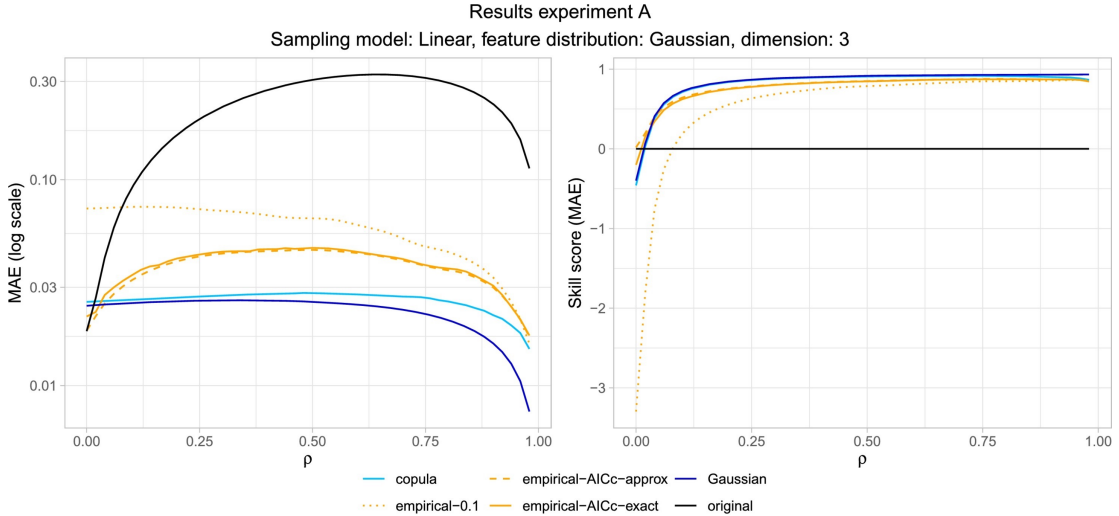
- General: $n_{\text{train}} = 2000$, $n_{\text{test}} = 10 \cdot 100 = 1000$, and $K = 1000$.

- Dimension 3: (Experiments A-F)
  - Multivariate data distribution:
    - Gaussian: $p(\boldsymbol{x}) = \mathcal{N}_3(\mathbf{0}, \Sigma(\rho))$, where $\Sigma(\rho)$ is one on diagonal and ρ off-diagonal.
    - Generalized Hyperbolic: more advanced, but with mean 0, skewness κ varies.
    - Gaussian mixture: $p(\boldsymbol{x}) = \sum_{k=1}^{2} \pi_k \mathcal{N}_3(\mu_k(\gamma), \Sigma(0.2))$, with $\pi_1 = \pi_2 = 0.5$ and $\mu_1(\gamma) = -\mu_2(\gamma) = \gamma \cdot (1, -0.5, 1)^T$.
  - Sampling model:
    - Linear: $y = g(\boldsymbol{x}) = \sum_{j=1}^{3} x_j + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1^2)$. Use Ordinary least squares.
    - Piecewise constant: $y = g(\boldsymbol{x}) = \sum_{j=1}^{3} \text{func}_j(x_j) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1^2)$. Use `XGBoost`.

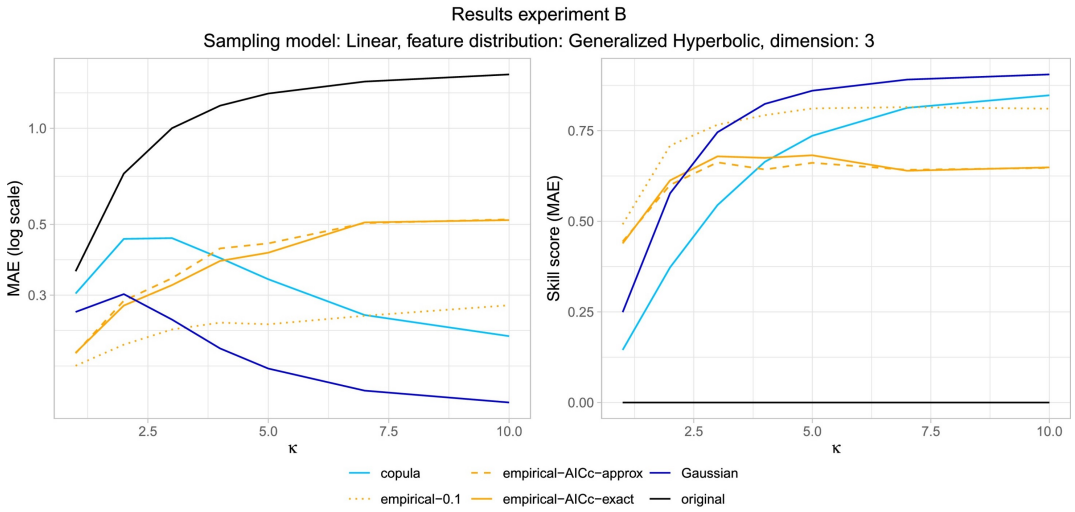
Illustration of piecewise constant functions



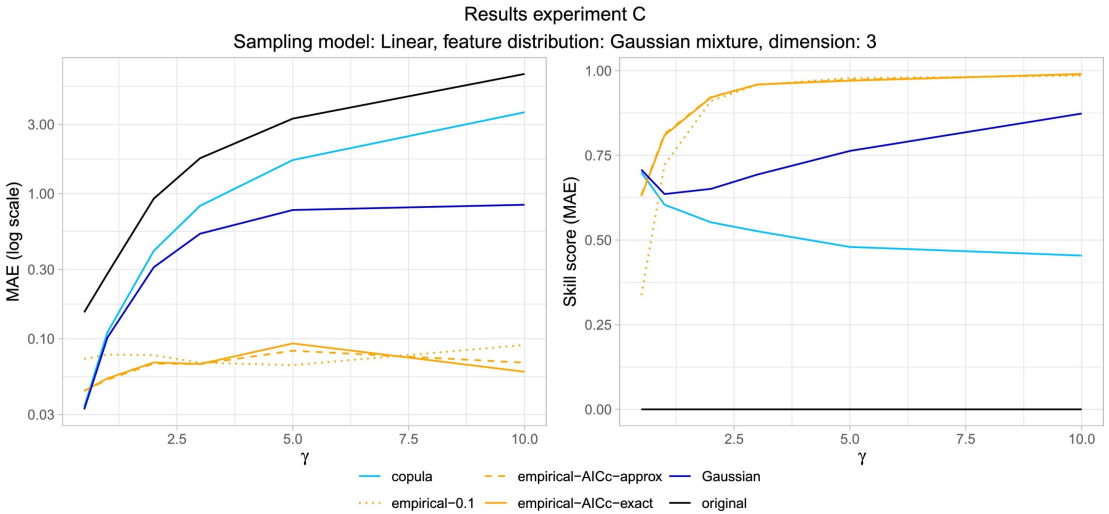
Simulation study: experiment A



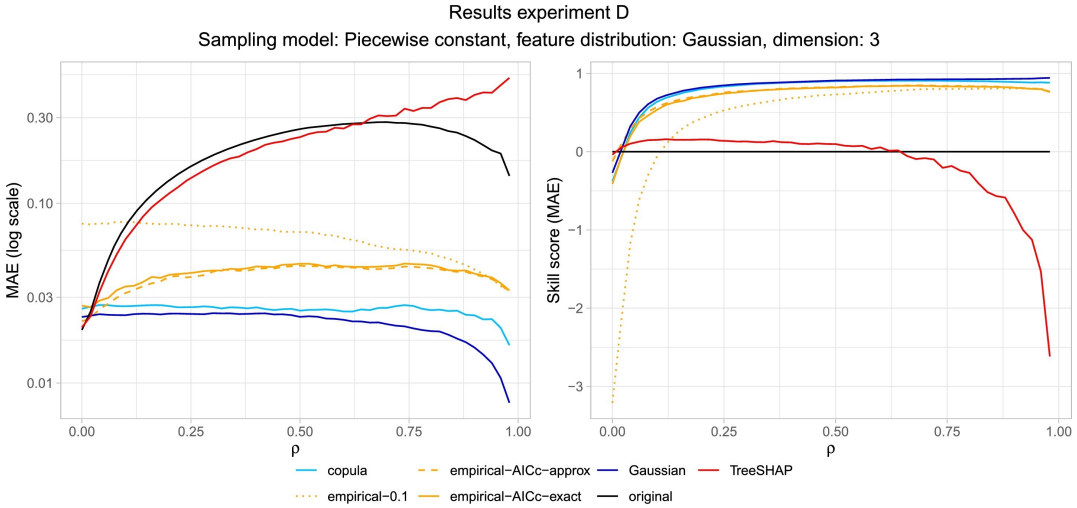
Simulation study: experiment B



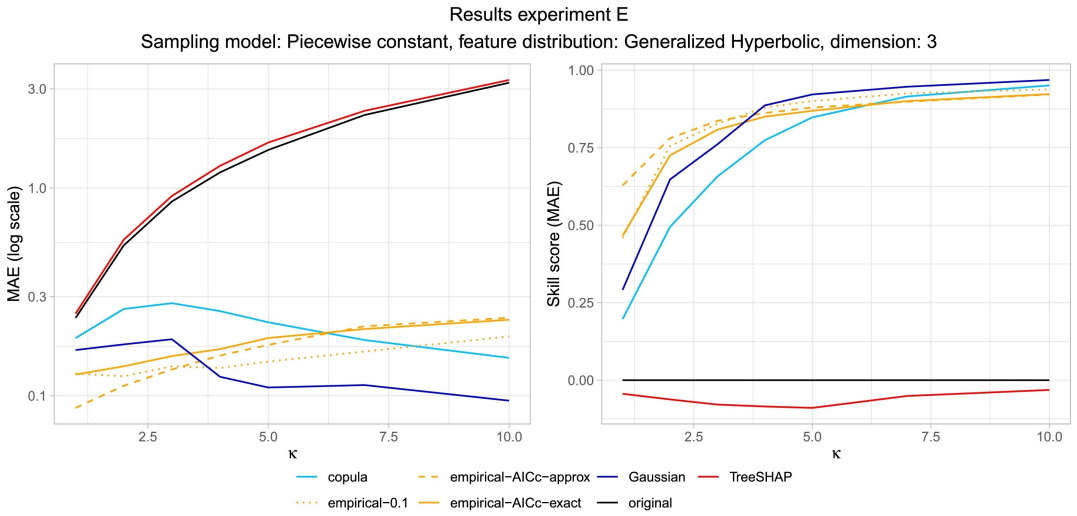
Simulation study: experiment C



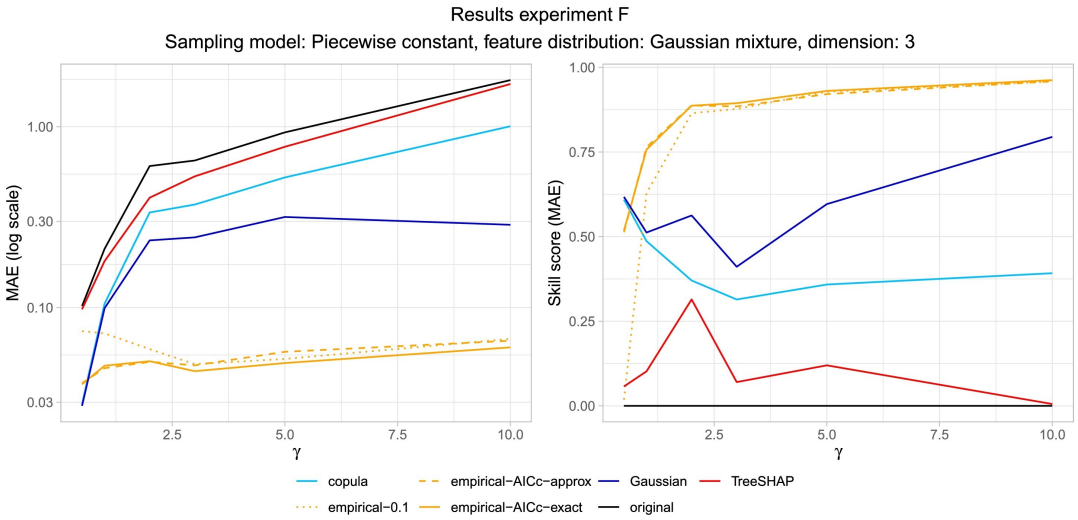
Simulation study: experiment D



Simulation study: experiment E



Simulation study: experiment F



Simulation studies: dimension 10

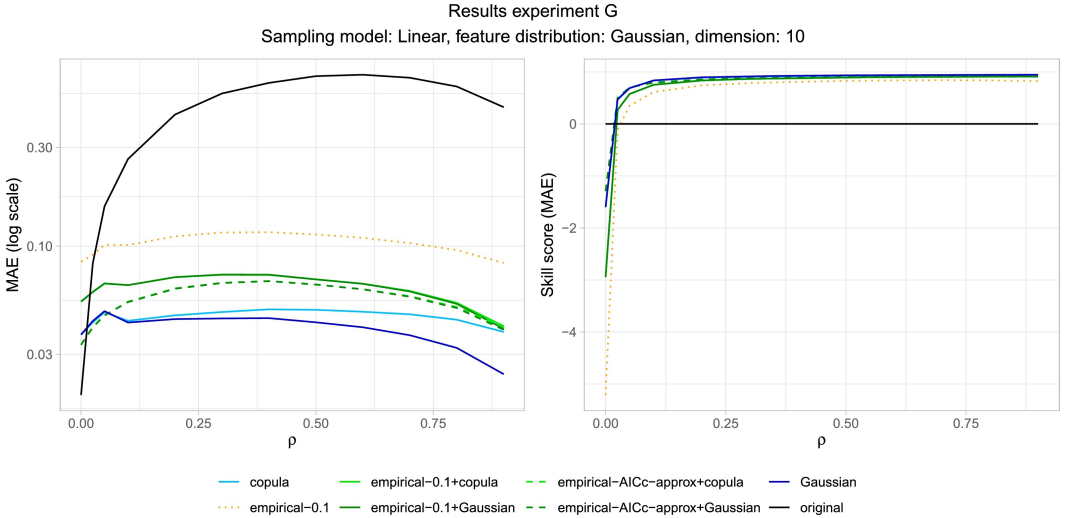
Dimension 10: (Experiments G-I)

1. $p(\boldsymbol{x}) = \mathcal{N}_{10}(\mathbf{0}, \Sigma(\rho))$, $y = g(\boldsymbol{x}) = \sum_{j=1}^{9} x_j + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1^2)$.
   Use ordinary least squares. Feature 10 no direct influence on y. Vary ρ.

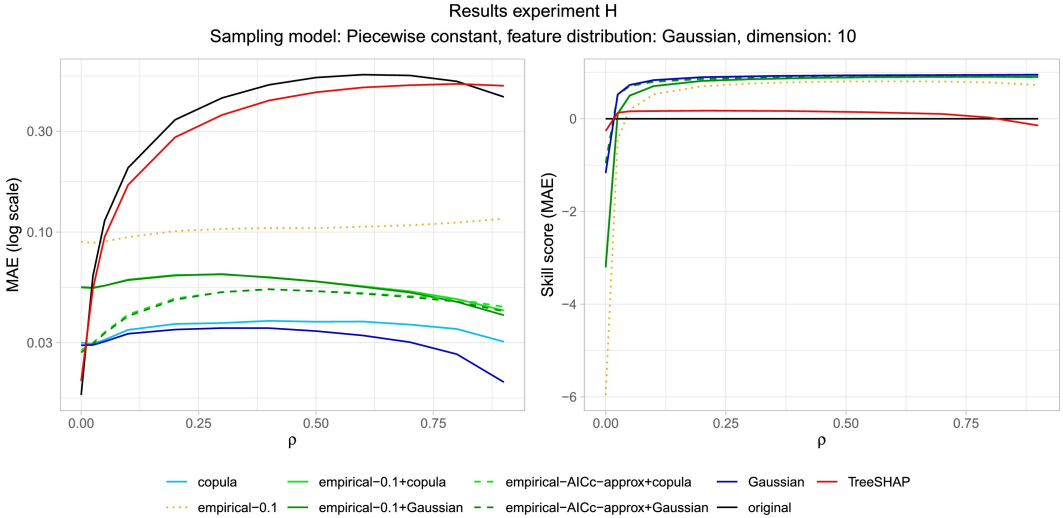
2. $p(\boldsymbol{x}) = \mathcal{N}_{10}(\mathbf{0}, \Sigma(\rho))$,
   $y = g(\boldsymbol{x}) = \sum_{j \in \{1,2,3\}} \text{func}_1(x_j) + \sum_{j \in \{4,5,6\}} \text{func}_2(x_j) + \sum_{j \in \{7,8,9\}} \text{func}_3(x_j) + \epsilon$,
   where $\epsilon \sim \mathcal{N}(0, 0.1^2)$. Use `XGBoost`. Feature 10 no direct influence on y. Vary ρ.

3. GH-distributed data (see paper), same piecewise constant function as in *2*. Only one value for κ. Use Monte Carlo integration to calculate true Shapley values.

Simulation study: experiment G



Simulation study: experiment H



Simulation study: experiment I

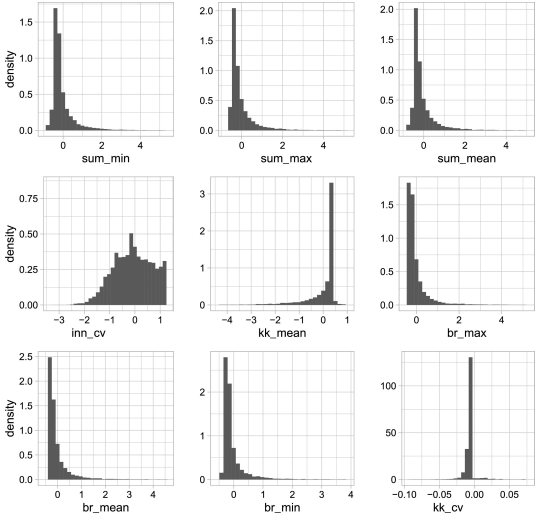
Table 1. Dimension 10: MAE and skill score for piecewise constant model with GH-distributed features.

Approach	MAE	Skill score
original	1.182	0.000
Gaussian	0.377	0.633
copula	0.526	0.504
empirical-0.1	0.307	0.737
empirical-0.1+Gaussian	0.199	0.821
empirical-0.1+Copula	0.236	0.791
TreeSHAP	1.181	0.014

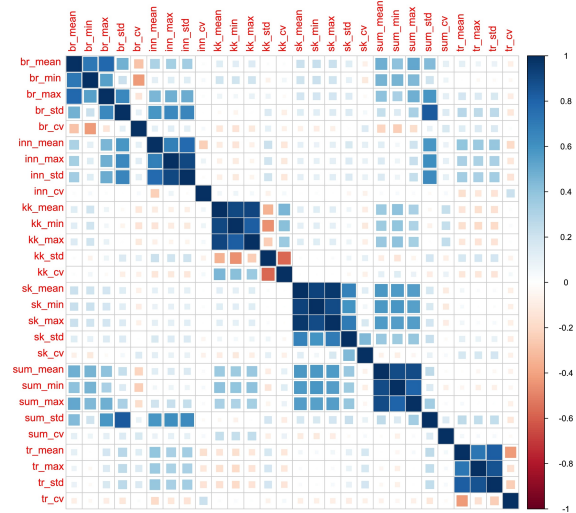
Real data example: transaction data

- 28 features from 6 transaction time series. $n_{\text{train}} = 12,696$, $n_{\text{test}} = 1921$, `XGBoost` model with 50 trees, $AUC = 0.88$
- Features are scaled to have mean 0 and standard deviation 1.
- Feature distributions are skewed and heavy-tailed.
- Pairwise rank correlations measured by Kendall's τ.
  - As features are far from linear dependent.
  - Many are close to 0, but groups of features with high mutual correlations.
  - Hence, new methods are likely more accurate than the original approach.
- The Shapley values of the different methods differ.
  - Primarily different for correlated features.
- No true Shapley values to compare with, but can compare how likely the generated data are.
  - Considers three conditional distributions.
  - Look at mean Mahalanobis distance between each generated sample and its 10 nearest training samples. See that new methods obtain lower means, i.e. more realistic.

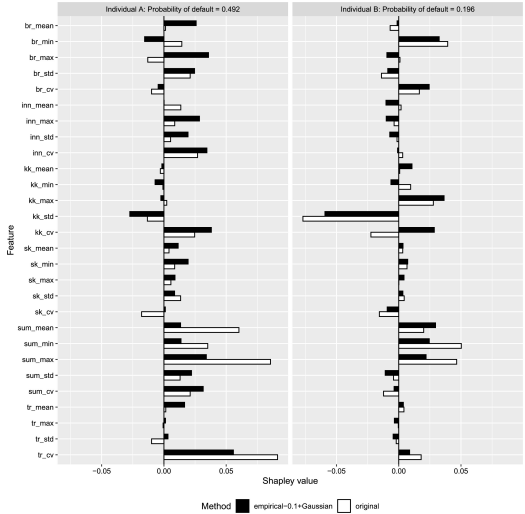
Transaction data: histograms



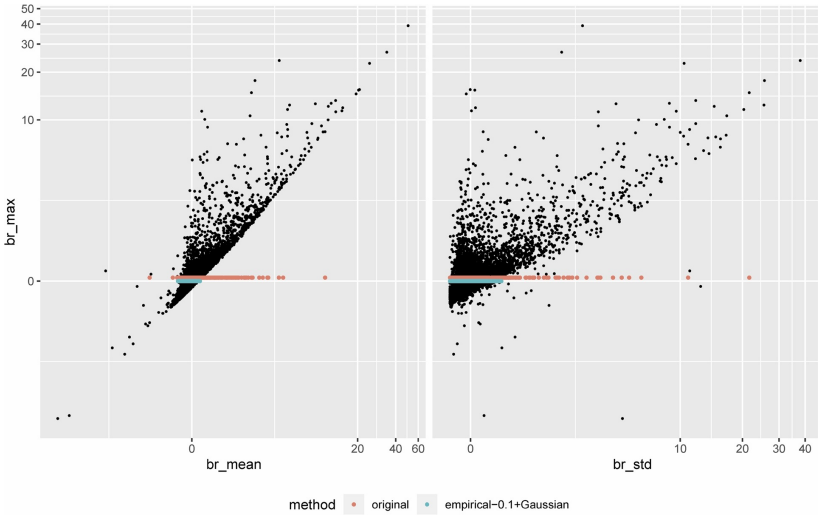
Transaction data: pairwise Kendall's τ



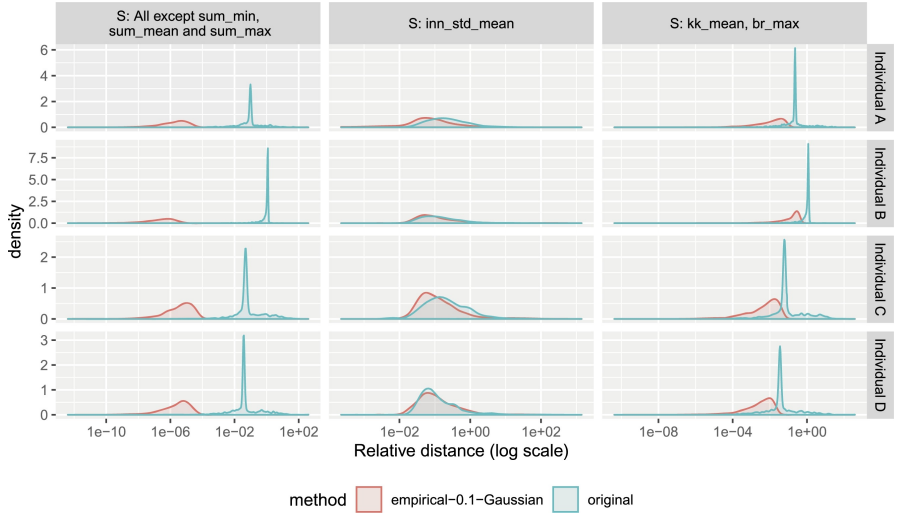
Transaction data: explanations



Transaction data: generated data I



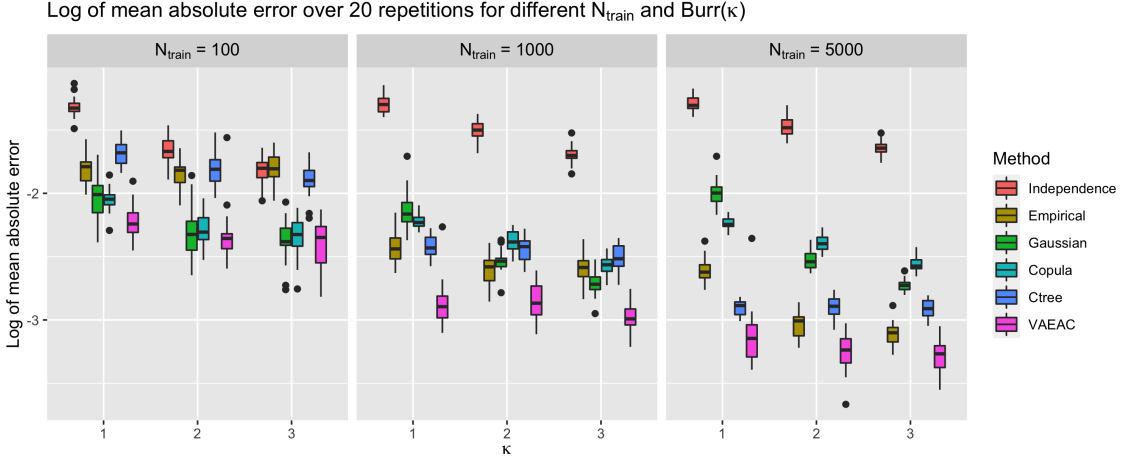
Transaction data: generated data II



Conclusion with self-advertisement I

- Aas et al. introduced methods for modelling feature dependencies in the `Kernel SHAP` methodology.
- The simulation studies demonstrated the necessity of modeling the feature dependencies for accurate Shapley values when the features were dependent.
- Drawback/limitations of new methodology:
  - Slower than the original method.
  - Cannot handle mixed-data, which is common in tabular data.
    - Can use feature encoding, but further increase the computational time.
- Aas et al. have implemented their methods in the R-package `shapr`.
- Self-advertisement:
  - Inspired by Emanuel Parzen from *Statistical Modeling: The Two Cultures*.
  - I use variational autoencoders to model $p(\boldsymbol{x}_{\bar{\mathcal{S}}}|\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}^*_{\mathcal{S}})$ for all $\mathcal{S}$ simultaneously.
  - Can handle mixed data and is more accurate than these methods.
  - Will talk about it on the BigInsight lunch Wednesday the 29th of September.

Conclusion with self-advertisement II



References I



Shapley, L. S.
‘A value for n-person games’.
Contributions to the Theory of Games, 28(2):307–317, 1953.



Lundberg, S. M. and Lee, S.
‘A unified approach to interpreting model predictions’.
Advances in neural information processing systems, 4765–4774, 2017.



Aas, K., Jullum, M. and Løland, A.
‘Explaining individual predictions when features are dependent: More accurate approximations to Shapley values’.
Artificial Intelligence, 298:103502, 2021.

UiO : Department of Mathematics
University of Oslo



Lars H. B. Olsen



Explaining individual predictions when features are dependent

More accurate approximations to Shapley values

