Shapley values and their uncertainties in complex regression models for epidemiological applications

Mark van de Wiel

Joint work with: Matteo Amestoy (Part I), Jeroen Hoogland (Part I); Giorgio Spadaccini (Part II), Marjolein Fokkema (Part II; U Leiden)

Epidemiology & Data Science, Amsterdam University Medical Centers, The Netherlands

Oslo, September 12, 2024

A story in two parts

Part I: Regression models with many two-way interactions

- Linked shrinkage
- Shapleys plus intervals
- Example: canonical epidemiological problem

A story in two parts

Part I: Regression models with many two-way interactions

- Linked shrinkage
- Shapleys plus intervals
- Example: canonical epidemiological problem

Part II: From regression to trees and back

- Translating trees to high-dimensional regression
- Shapleys plus intervals
- Example: academic performance of 1rst yr psychology students

Part I: Setting

- Low-dimensional epidemiological cohort study (Helius)
  - Response: y, say cholesterol
  - Covariates: age, bmi, smoking, ethnicity, etc.
  - Our interest: $n > p$, say $n = 1,000, p = 14$
  - But large N available: $N \approx 21.500$ (allows benchmarking)

Part I: Setting

- Low-dimensional epidemiological cohort study (Helius)
  - Response: y, say cholesterol
  - Covariates: age, bmi, smoking, ethnicity, etc.
  - Our interest: $n > p$, say $n = 1,000, p = 14$
  - But large N available: $N \approx 21.500$ (allows benchmarking)

- Aims
  1. Interpretable model that explains y
  2. Variable importance
  3. Competitive prediction

Our model: linked shrinkage model

$$Y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j,k:j\neq k} \beta_{jk} x_{ij} x_{ik} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

Our model: linked shrinkage model

$$Y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j,k:j\neq k} \beta_{jk} x_{ij} x_{ik} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

$$\beta_j \sim N(0, \sigma^2 \tau_j^2)$$

$$\tau_j \sim C^+(0, 1)$$

$$\alpha \sim N(0, 10^2), \sigma^2 \sim IG(1, 0.001)$$

Our model: linked shrinkage model

$$Y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j,k:j\neq k} \beta_{jk} x_{ij} x_{ik} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

$$\beta_j \sim N(0, \sigma^2 \tau_j^2),\ \beta_{jk} \sim N(0, \sigma^2 \tau_j \tau_k \tau_{int})$$

$$\tau_j \sim C^+(0, 1),\ \tau_{int} \sim U(0.01, 1)$$

$$\alpha \sim N(0, 10^2), \sigma^2 \sim IG(1, 0.001)$$

Our model: linked shrinkage model

$$Y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j,k:j\neq k} \beta_{jk} x_{ij} x_{ik} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

$$\beta_j \sim N(0, \sigma^2 \tau_j^2),\ \beta_{jk} \sim N(0, \sigma^2 \tau_j \tau_k \tau_{int})$$

$$\tau_j \sim C^+(0,1),\ \tau_{int} \sim U(0.01, 1)$$

$$\alpha \sim N(0, 10^2), \sigma^2 \sim IG(1, 0.001)$$

`Bayint`: Coded in R-stan for flexibility

Benefits of the model

Models beats* competitors like spike-and-slab, horseshoe, hierarchical lasso, adaptive lasso, etc on:

- Parameter estimation
- Prediction
- Variable selection

... and also `RF` (prediction)

*See: `https://arxiv.org/abs/2309.13998`; vdW et al, *Epid Meth* 2024

Comparison with RF (prediction)

Set-up

Outcomes: Cholesterol & SBP
$n = 1,000, p = 14$
$n_{\text{test}} \approx 20,000$
25 training-test splits
Competitors:

- `Bayint` (our model)
- `MainEff` (our model with only main effects)
- `OLS`
- `RF` with default parameters
- `RF` with tuned parameters

Comparison with RF (prediction)

Set-up

Outcomes: Cholesterol & SBP

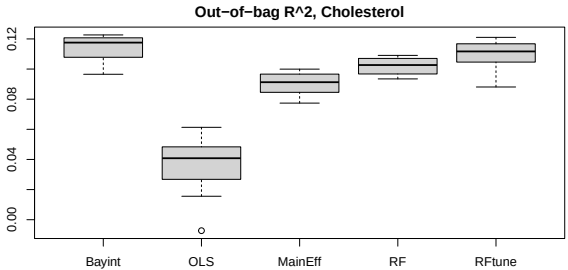
$n = 1,000, p = 14$

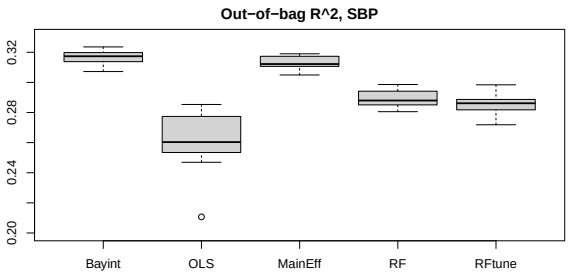
$n_{\text{test}} \approx 20,000$

25 training-test splits

Competitors:

- `Bayint` (our model)
- `MainEff` (our model with only main effects)
- `OLS`
- `RF` with default parameters
- `RF` with tuned parameters





Interpretation: variable importance

Quantifying variable importance not trivial in regression model with interactions (Afshartous and Preston 2011)

Shapley values: metric quantifying variable importance per sample (Aas, Jullum, and Løland 2021)

Interpretation: variable importance

Quantifying variable importance not trivial in regression model with interactions (Afshartous and Preston 2011)

Shapley values: metric quantifying variable importance per sample (Aas, Jullum, and Løland 2021)

- Originates from game theory
- Uniquely combines many desirable properties
- Applies to complex machine learners
- Expensive to compute, usually

Shapley values

Shapley, intuitive explanation (Molnar 2023)

- The variables enter a room in random order. All variables in the room (i.e. the coalition of players) participate in the game (= contribute to the prediction).
- Shapley value of a variable's realization: average change in the prediction when the variable joins the coalition
- Averaged over all possible coalitions (subsets).
- No refitting: predictions are marginalized w.r.t. non-players

Shapley for our model

Interventional Shapley: ignore dependencies between players and non-players

Explicit formula for the interventional Shapley for our model[†]: [e.g. $x_{ij} = 49$, age (j) for individual i]

[†]See: `https://arxiv.org/abs/2309.13998`; vdW et al, *Epid Meth* 2024

Shapley for our model

Interventional Shapley: ignore dependencies between players and non-players

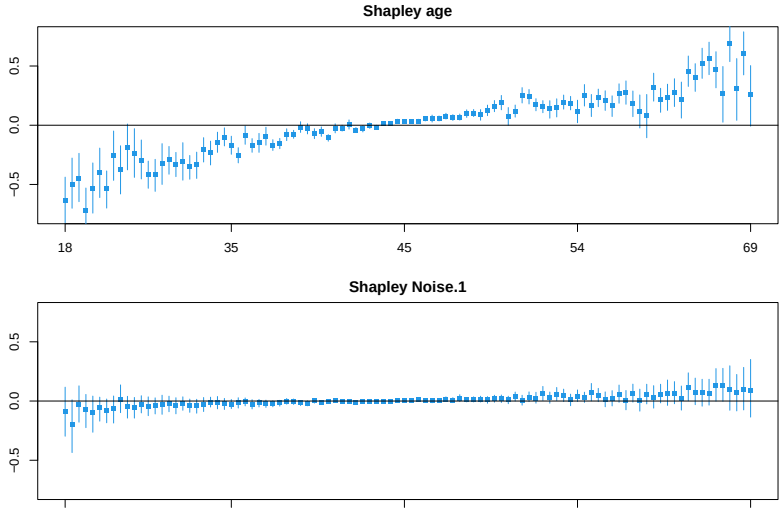
Explicit formula for the interventional Shapley for our model[†]: [e.g. $x_{ij} = 49$, age (j) for individual i]

$$\phi(x_{ij}) = \beta_j x_{ij} + \frac{1}{2}\left(\sum_{k:k\neq j} \beta_{jk} x_{ij} x_{ik} - \sum_{k:k\neq j} \beta_{jk} E[X_{ij} X_{ik}]\right),$$

[†]See: `https://arxiv.org/abs/2309.13998`; vdW et al, *Epid Meth* 2024

Shapleys + credible intervals

MCMC samples of β provides uncertainty estimate of $\phi(x_{ij})$



Intervals have good coverage

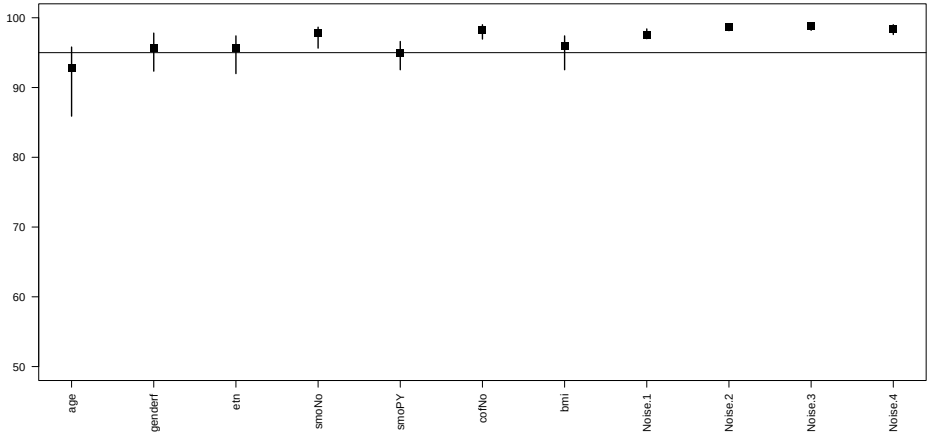
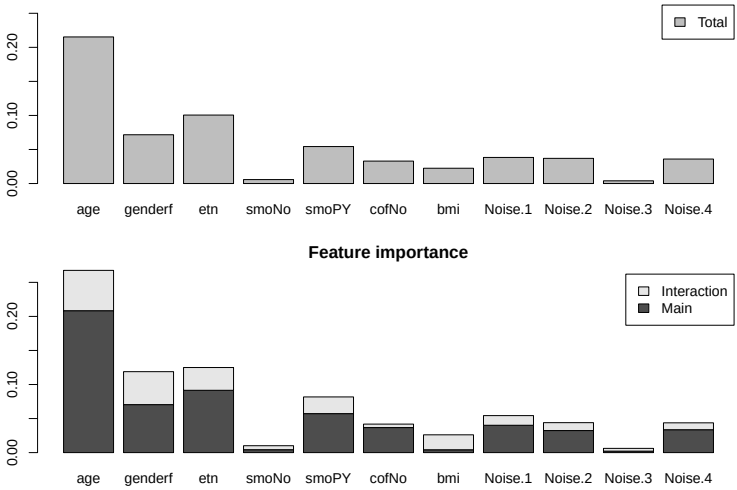


Figure: Coverages of 95% credible intervals for Shapley values of 200 random test individuals (quartiles shown). Estimated Shapley values are obtained from 500 random subsets of size $n = 1,000$.

Distinguish contribution of main and interaction effects

Global importance: $I_j = 1/n \sum_{i=1}^{n} |\phi(x_{ij})|$.



Comparison with hierarchical lasso

Hierarchical lasso (`hlasso`; Bien et al. 2013)

- Stronger link between interactions and main effects (selection)
- Computationally efficient
- Does not provide uncertainty estimates
- Shapleys `hlasso`: simply substitute parameter estimates into formula

Comparison

For variables Age, Etnicity:

1. Estimate Shapley values for 1,000 random test individuals
2. Obtain 'true' Shapley values from OLS estimates on Master set ($N = 21500; p = 14; q = 85$)
3. Plot estimated vs true for 25 nearly non-overlapping training sets

Shapley estimates (1: age)

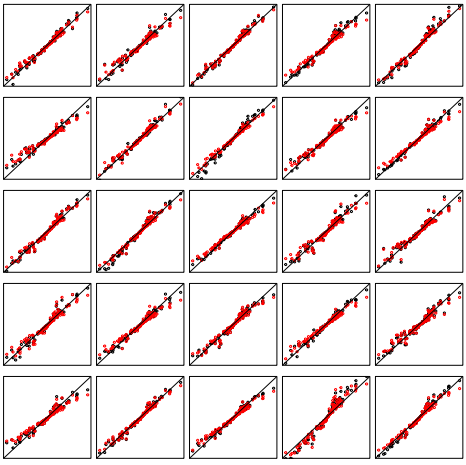


Figure: Shapley values for 'age' over 1,000 random test individuals (dots) for 25 training sets (displays). X-axis: true Shapley values; Y-axis: estimated ones by `Bayint` (black) and `hlasso` (red).

Shapley estimates (2: etn1)

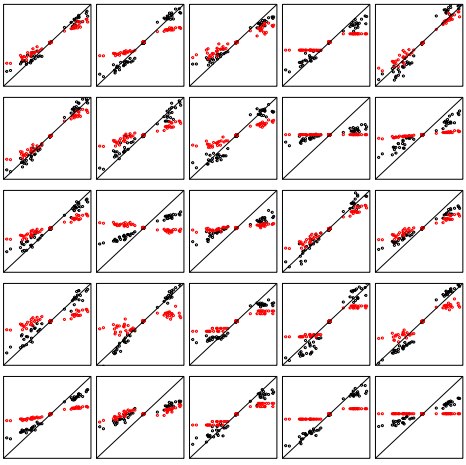


Figure: Shapley values for 'etn1' over 1,000 random test individuals (dots) for 25 training sets (displays). X-axis: true Shapley values; Y-axis: estimated ones by `Bayint` (black) and `hlasso` (red).

Conclusion Part I

1. Linked shrinkage useful concept for model with many interactions

Conclusion Part I

1. Linked shrinkage useful concept for model with many interactions
2. Shapley values are useful for complex regression models

Conclusion Part I

1. Linked shrinkage useful concept for model with many interactions
2. Shapley values are useful for complex regression models
3. Appropriate regularization allows good uncertainty quantification

Part II: From regression to trees and back

Sometimes the world is non-linear and full of complex interactions

Then, tree learners (Random Forest, XGboost, etc) are a better alternative

Part II: From regression to trees and back

Sometimes the world is non-linear and full of complex interactions

Then, tree learners (Random Forest, XGboost, etc) are a better alternative

Current practice in epidemiology:

- Use tree learners and Shap(ley) to select top k features
- Apply a linear (!!!) regression model to those features to perform inference

Part II: From regression to trees and back

Sometimes the world is non-linear and full of complex interactions

Then, tree learners (Random Forest, XGboost, etc) are a better alternative

Current practice in epidemiology:

- Use tree learners and Shap(ley) to select top k features
- Apply a linear (!!!) regression model to those features to perform inference

Main aim: replace this inconsistent approach by one that is true to the tree learner

Main aim

Develop an inferential approach true to the tree learner

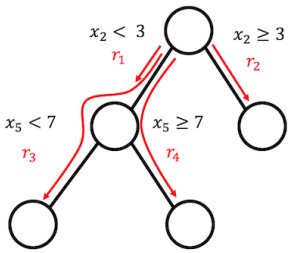
1. Translate tree learners to regression models
2. Efficient computation of Shapley values for such models
3. Inference: credible intervals for Shapley values
4. Decomposition of these into linear contributions and remainder ones

From trees to regression

1. Translate tree learners to regression models
2. Efficient computation of Shapley values for such models
3. Inference on Shapley values: credible intervals
4. Decomposition of these into linear contributions and remainder ones

From trees to regression: RuleFit

RuleFit (Friedman and Popescu 2008; Fokkema 2020): **linear regression** with **rule-based** interactions and non-linear terms.



Rules taken from Random Forest / Boosted Tree Ensemble

Rules + Linear terms $\Longrightarrow$ High interpretability
Rules from Tree Ensemble $\Longrightarrow$ High performance

RuleFit: Example

$$
\begin{aligned}
Y = X\beta = \hat{b}_1 x_1 + \hat{b}_1 x_2 + \hat{b}_3 x_3 + \hat{b}_4 x_4 + \hat{b}_5 x_5 \\
+ \hat{a}_1 I(x_2 < 3) + \hat{a}_2 I(x_2 \geq 3) \\
+ \hat{a}_3 I(x_2 < 3, x_5 < 7) + \hat{a}_4 I(x_2 < 3, x_5 \geq 7) + \cdots
\end{aligned}
$$

$$
X = \begin{bmatrix}
2 & -0.5 & 5 & 3.1 & 7.1 & 1 & 0 & 0 & 1 & \cdots \\
1.5 & 3.7 & -0.1 & 4 & 2.4 & 0 & 1 & 0 & 0 & \cdots \\
0.1 & 1.8 & 1 & 0 & 6 & 1 & 0 & 1 & 0 & \cdots \\
-0.9 & 4.1 & 2.9 & 2.2 & 2.1 & 0 & 1 & 0 & 0 & \cdots \\
-0.4 & 3 & 4.2 & 1.6 & -0.6 & 0 & 1 & 0 & 0 & \cdots
\end{bmatrix}
$$

Very many rules $\Rightarrow$ Sparse model

Back to our aims

1. Translate tree learners to regression models
2. Efficient computation of Shapley values for such models
3. Inference on Shapley values: credible intervals
4. Decomposition of these into linear contributions and rules

Back to our aims

1. Translate tree learners to regression models
2. Efficient computation of Shapley values for such models
3. Inference on Shapley values: credible intervals
4. Decomposition of these into linear contributions and rules

$\rightarrow$ Explicit formula for Shapley values $\phi(x_{ij})$ based on RuleFit

- Computationally competitive to TreeShap (Lundberg et al. 2020)
- Advantage: achieves Aims 3 and 4 as well

Decomposition and inference

1. Translate tree learners to regression models
2. Efficient computation of Shapley values for such models
3. Inference on Shapley values: credible intervals
4. Decomposition of these into linear contributions and rules

Decomposition and inference

1. Translate tree learners to regression models
2. Efficient computation of Shapley values for such models
3. Inference on Shapley values: credible intervals
4. Decomposition of these into linear contributions and rules

Key ingredients for Aims 3 and 4:

- Better inference: Lasso (RuleFit) $\rightarrow$ Horseshoe (Nalenz et al, 2018)
- Decomposition:
  - Rule generation: fit trees on residuals from linear model
  - Joint fitting: many more rules than linear terms $\rightarrow$ differential regularization
  - $\phi(x_{ij}) = \phi^{\text{lin}}(x_{ij}) + \phi^{\text{rule}}(x_{ij})$

Algorithm

1. Fit linear model
2. Fit random forest[‡] on residuals (tweaked to counter over-fitting)
3. Collect rules
4. Fit horseshoe regression with unpenalized linear effects
5. Obtain posteriors from MCMC sample
6. Compute Shapley values and uncertainties with formula

[‡]Could be replaced by one's favorite tree algorithm

Illustration

Academic achievements; 1rst year psychology students

- Response: nr of credit points or mean grade
- Aim: predict and explain
- $n = 638$
- Covariates: gender, age, nationality (4), online_test, test_language, score_Math, score_Engl, score_Psych, program
- Added five independent noise variables as negative controls
- Data available from `pre` package (Fokkema 2020)

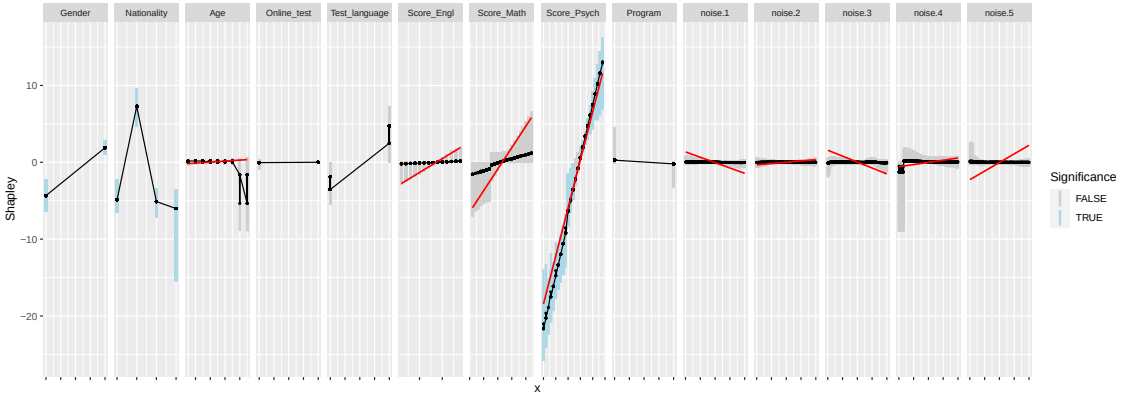
Results: prediction

Evaluation: cross-validated R^2 (the higher, the better)

	Outcome	
Method	Credit Points	Mean Grade
RF	0.327	0.239
RuleFit	0.313	0.218
Tree	0.276	0.189
LassoReg	0.313	0.195
Ours[§]	0.320	0.260

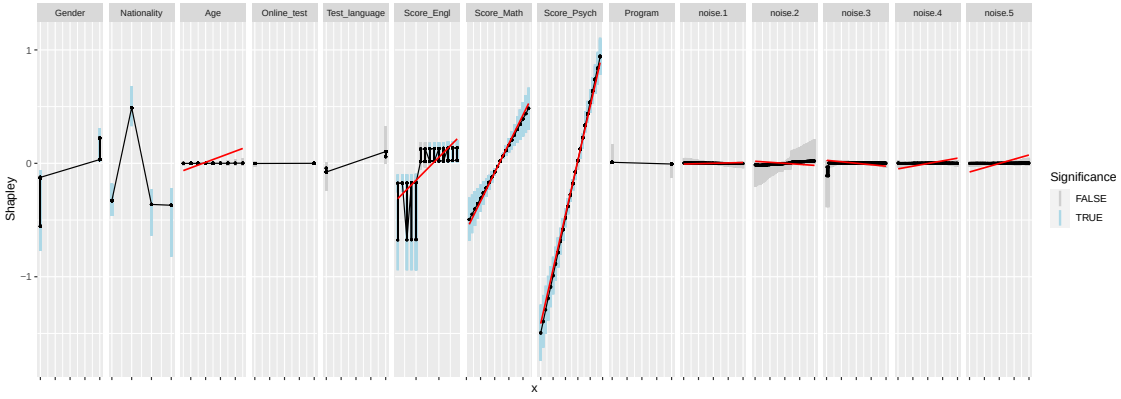
[§]No good name yet!

Predicting credit points: Shapleys + credible intervals



Decomposition $j =$ Score_Engl: $(\overline{|\phi_j^{\mathsf{lin}}|}, \overline{|\phi_j^{\mathsf{rule}}|}) = (0.062, 0.016)$

Predicting mean grade: Shapleys + credible intervals



Decomposition $j =$ Score_Engl: $(\overline{|\phi_j^{\text{lin}}|}, \overline{|\phi_j^{\text{rule}}|}) = (0.003, 0.092)$

Conclusions Part II

1. Translation to regression enables interpretation true to the tree-learner

Conclusions Part II

1. Translation to regression enables interpretation true to the tree-learner

2. Shapley values are useful for complex regression models derived from tree-learners

Conclusions Part II

1. Translation to regression enables interpretation true to the tree-learner
2. Shapley values are useful for complex regression models derived from tree-learners
3. Appropriate regularization allows uncertainty quantification but more work needed to show appropriate coverage of the intervals

References

Aas, Kjersti, Martin Jullum, and Anders Løland (2021). "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values". In: *Artificial Intelligence* 298, p. 103502.

Afshartous, D. and R. A Preston (2011). "Key results of interaction models with centering". In: *Journal of Statistics Education* 19.3.

Bien, Jacob, Jonathan Taylor, and Robert Tibshirani (2013). "A lasso for hierarchical interactions". In: *Annals of statistics* 41.3, p. 1111.

Fokkema, Marjolein (2020). "Fitting Prediction Rule Ensembles with R Package pre". In: *Journal of Statistical Software* 92, pp. 1–30.

Friedman, Jerome H and Bogdan E Popescu (2008). "Predictive Learning via Rule Ensembles". In: *The Annals of Applied Statistics*, pp. 916–954.

Lundberg, Scott M et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1, pp. 56–67.

Nalenz, Malte and Mattias Villani (2018). "Tree ensembles with rule structured horseshoe regularization". In: *The Annals of Applied Statistics* 12.4, pp. 2379–2408.

van de Wiel, Mark A, Matteo Amestoy, and Jeroen Hoogland (2024). "Linked shrinkage to improve estimation of interaction effects in regression models". In: *Epidemiologic Methods* 13.1, p. 20230039.