Challenges in interpretability of additive models

Xinyu Zhang[1] , **Julien Martinelli**[2] , **S.T. John**[1]

[1]Department of Computer Science, Aalto University, Espoo, Finland

[2]Inserm Bordeaux Population Health, Vaccine Research Institute, Université de Bordeaux, Inria Bordeaux Sud-ouest, France

Abstract

We review generalized additive models as a type of 'transparent' model that has recently seen renewed interest in the deep learning community as *neural additive models*. We highlight multiple types of nonidentifiability in this model class and discuss challenges in interpretability, arguing for restraint when claiming 'interpretability' or 'suitability for safety-critical applications' of such models.

1 Introduction

Machine-learning models provide predictions just based on existing training data, without any requirement for understanding. However, understanding how a model arrives at its predictions may be required, for example by law, in high-stakes domains such as healthcare, criminal justice, or finance. This has given rise to "interpretable" or "explainable" machine learning. However, as pointed out by [Lipton, 2017], model interpretability "is not a monolithic concept" and requires careful consideration of what exactly we mean by it.

Instead of relying on post-hoc explanation methods that only approximate a model's behavior (e.g. LIME [Ribeiro *et al.*, 2016] or SHAP [Lundberg and Lee, 2017]), it has become popular to develop "transparent" or "glass-box" models that are "readily understandable" [Rudin, 2019; Rai, 2020].

Here we focus on (generalized) additive models (GAMs) as a subclass of "transparent" models. Additive models have a long history in statistical modeling [Hastie and Tibshirani, 1986], but received a resurgence in the deep learning community with "neural additive models" [Agarwal *et al.*, 2021] that purport to "combine some of the expressivity of DNNs with the inherent intelligibility of generalized additive models".

In the following, we provide an overview of recent developments in additive models and discuss how, despite their simplistic nature, additive models may not be as interpretable as one might think.

2 Background on additive models

2.1 General definition

For a D-dimensional input $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, an additive model takes the following form:

$$f(\mathbf{x}) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_D(x_D), \quad (1)$$

where the individual components $\{f_d(\cdot)\}_{1 \leq d \leq D}$ are referred to as *shape functions*. We can model the response variable y in different settings, like classification or regression, as $g(\mathbb{E}[y|\mathbf{x}]) = f(\mathbf{x})$, where $g(\cdot)$ is a link function.

The shape functions can take various forms such as smoothing splines, polynomials, or neural networks. As such they can model highly nonlinear behaviors of each feature while maintaining a relatively simple and 'interpretable' structure since each feature is only involved in one term independently of the others. We will return to this in section 2.3.

To train the additive model of eq. (1), one needs to specify (i) the functional form of the shape functions and (ii) how they are optimized. We cover both points in the following section.

2.2 Literature review: modeling and learning shape functions

In the first GAMs, the shape functions were constructed with smoothing splines and iteratively trained through a backfitting algorithm [Hastie and Tibshirani, 1986]. Within each iteration of backfitting, features are cycled through and shaped from scratch on residuals obtained by fixing other shape functions. More recently, [Lou *et al.*, 2012] proposed shallow boosted bagged trees as the shape functions and gradient boosting [Friedman, 2000] as an alternative fitting method, which gradually expands shape functions instead of learning an entirely new set of shape functions every iteration. [Chang *et al.*, 2021] showed that the tree-based models achieved a better balance of feature sparsity, data fidelity and accuracy compared to the spline-based counterparts.

The Neural Additive Model (NAM) of [Agarwal *et al.*, 2021], a linear combination of feed-forward networks that are jointly fitted using standard backpropagation, revived additive models in the deep learning community. Since then, numerous extensions of NAMs have been proposed, bringing in ideas like feature selection, feature interactions, and uncertainty estimates. Notably, the recently proposed Bayesian variant of NAMs [Bouchiat *et al.*, 2024], LA-NAMs, leverages linearized Laplace-approximated Bayesian inference over subnetworks, enabling principled uncertainty estimation as well as feature selection for both individual features and pairwise interactions.

[Radenovic *et al.*, 2022] proposed the Neural Basis Model (NBM) as a more parameter-efficient alternative to the NAM. Instead of separate independent neural networks for each in-

put, NBM learns a basis expansion of each input, where the basis functions are shared among shape functions. With a theoretical bound of $B = O(\log D)$ for the number of bases, NBM effectively reduces the model size while guaranteeing comparable performance to the NAM.

As an intrinsically probabilistic model, [Duvenaud *et al.*, 2011] proposed additive Gaussian Processes (GPs) for modeling additive functions, leveraging a novel kernel encoding additive structure of interactions of all orders. The standard method of maximizing marginal likelihood on the training set is used for fitting hyperparameters. [Lu *et al.*, 2022] extended this idea to the orthogonal additive kernel (OAK), thus enabling a more identifiable and low-dimensional interaction selection by imposing orthogonality between shape functions of different orders [Durrande *et al.*, 2013].

2.3 How we interpret additive models

One major perk of additive models is the ability to visualize each feature's contribution to the response across the domain, by plotting the corresponding shape function. Moreover, classical feature importance metrics can also be used. For instance, [Agarwal *et al.*, 2021] suggested to define the importance score for feature x_i as $\frac{1}{N}\sum_{n=1}^{N}|f_i(x_{n,i}) - \overline{f_i}|$, where $\overline{f_i} = \frac{1}{N}\sum_{n=1}^{N} f_i(x_{n,i})$, averaged over the entire training set. Alternatively, GAMI-Net [Yang *et al.*, 2021] utilizes the empirical variance of shape functions, $\frac{1}{N-1}\sum_{n=1}^{N} f_i^2(x_{n,i})$ [1], which can be connected to the variance-based sensitivity analysis techniques widely used for measuring relative importance of (sets of) inputs to a function $f(\mathbf{x})$. One of the most prevalent importance measures are the Sobol indices, used for assessing feature importance in additive models with orthogonality constraints [Lu *et al.*, 2022].

2.4 Extensions of additive models

To ensure intelligibility, additive models often integrate inductive biases, such as the absence of feature interactions and smoothness assumption for shape functions. However, strong inductive biases could compromise both accuracy and fidelity, particularly when data exhibits patterns that cannot be solely explained by individual effects, or sharp changes (jumps). Related extensions have been proposed to address these scenarios.

Incorporating higher-order interactions

The performance gap between tree-based additive model and full-complexity models such as random forests suggests a need for incorporating high-order interaction terms [Lou *et al.*, 2012] by generalizing eq. (1) to

$$f(\mathbf{x}) = \sum_{p \in \mathcal{P}(\{1,\dots,D\})} f_p(\mathbf{x}_p). \tag{2}$$

As the number of additive components grows exponentially with interaction orders, this decomposition is commonly restricted to only first- and second-order interactions both for tractability and interpretability, commonly referred to as GA^{2}M variants [Lou *et al.*, 2013]. However, even with this restriction, the search space of $\binom{D}{2}$ feature pairs can still be computationally challenging when dealing with high-dimensional data.

To effectively identify a subset of pairwise interactions, [Lou *et al.*, 2013] introduced a greedy forward selection strategy called FAST, which first fits the first-order components, then adds the second-order interactions one by one based on a greedy minimization of the residual error. The sparse interaction additive network (SIAN) of [Enouen and Liu, 2022] gathers interactions up to a given order using a secant approximation of the Hessian for second-order interactions and a similar approximation of higher-order derivatives for interactions of order more than two. On the other hand, LA-NAM [Bouchiat *et al.*, 2024] computes mutual information between the last-layer parameters θ_d and θ'_d of subnetworks f_d and f'_d and selects the top-k highest pairs (d, d'). Finally, rather than adaptively selecting interactions, a global sparsity constraint can be imposed on the full interaction set using classic methods such as LASSO [Lin and Zhang, 2006; Liu *et al.*, 2020].

Modeling "jagged" data

[Chang *et al.*, 2021] observed that additive models with overly smooth shape functions, such as the spline-based models or fused LASSO additive models (FLAM) [Ashley Petersen and Simon, 2016], fail to capture true patterns when there are sharp jumps in data compared to tree-based variants. The vanilla NAM [Agarwal *et al.*, 2021] attempted to model more jagged functions by incorporating "exponential units" (ExU) activation functions, which can fit even Bernoulli noise. However, follow-up work suggests that, in practice, standard ReLU activation functions appear to perform better [Bouchiat *et al.*, 2024].

Uncertainty estimation

Most NAM variants lack an inherent framework for epistemic uncertainty, necessitating model-agnostic uncertainty estimation techniques such as ensembles of models. [Bouchiat *et al.*, 2024] extended NAM with Laplace approximation to Bayesian inference to enable a more principled uncertainty estimate. Gaussian process models such as OAK [Lu *et al.*, 2022] naturally provide uncertainty estimates; however, their use of normalizing flow preprocessing makes their uncertainty estimates questionable in practice, as a stationary Gaussian process is fitted in a highly nonlinearly transformed space.

3 Nonidentifiability

Papers often make fairly strong claims about having identified shape functions and postulating that this is how it is in the real world. However, there are multiple types of non-identifiability in additive models, possibly causing multiple models that yield the same predictive performance under very different *explanations* of the data. Various solutions have been proposed for obtaining stable parameter estimates with good predictive performance, often claiming to improve interpretability as well. However, it is crucial to question what

[1] In GAMI-Net, all shape functions are centered ($\overline{f_i} = 0$) and the overall offset is modeled using an explicit bias term, thus $\frac{1}{N-1}\sum_{n=1}^{N}\left[f_i(x_{n,i}) - \overline{f_i}\right]^2 = \frac{1}{N-1}\sum_{n=1}^{N} f_i^2(x_{n,i})$.

interpretability means in each use case and whether identifiability truly satisfies this requirement.

3.1 Nonidentifiability due to only observing sum of terms

The most salient type of nonidentifiability stems from only observing the sum of additive components rather than each individual component on its own. As an illustration, let us consider a two-dimensional additive function:

$$f(x_1, x_2) = f_1(x_1) + f_2(x_2). \tag{3}$$

If we only observe the overall function value $f(x_1, x_2)$, then for any constant Δ, we get a different, valid decomposition $g_1(x_1) + g_2(x_2) = (f_1(x_1) + \Delta) + (f_2(x_2) - \Delta)$.

Such nonidentifiability was shown to produce unstable parameter estimates causing disparate interpretations with different initializations as well as unnecessarily complicated models when considering interactions, since higher-order terms can absorb the effect from low-order terms.

Orthogonality constraints were suggested as a solution [Märtens *et al.*, 2019; Lu *et al.*, 2022]. [Märtens *et al.*, 2019] considered additional covariate information $\mathbf{x}$ and its interactions with the latent variables $\mathbf{z}$ in the context of probabilistic dimensionality reduction, by decomposing the mapping into a linear combination of the marginal and interaction effects, while [Lu *et al.*, 2022] focused on encouraging a more parsimonious interaction set for additive GPs. In both works, the proposed kernels enforce the integral of each additive component to be zero with respect to the input measure:

$$\int f_p(\mathbf{x}_p) \mathrm{d}\mu(x_i) = 0 \;\; \forall i \in p. \tag{4}$$

Similarly, LA-NAM ensures orthogonality in individual effects through the block-diagonal covariance matrix for the approximate posterior over feature subnetworks.

3.2 Nonidentifiability due to concurvity/correlated covariates

Even with orthogonality constraints, nonidentifiability appears when covariates are not independent of each other. This can be either in the form of multicollinearity or of concurvity. While multicollinearity describes a linear relationship between the input features x_i, concurvity extends this notion by considering the shape functions f_i. This concept effectively captures nonlinear dependencies among features, indicating redundancy in the feature set. Formally, following the definition in [Ramsay *et al.*, 2003], let $\mathcal{H} \in \{(f_1, \ldots, f_D) | f_d : \mathbb{R} \to \mathbb{R}\}$ be a family of functions. We say that concurvity occurs between feature x_i and other features if there exists $(h_1, \ldots, h_D) \in \mathcal{H}$ such that $h_i(x_i)$ can be approximated by a combination of other functions, $h_i(x_i) \approx \sum_{j \neq i} h_j(x_j)$. In this way, the model becomes nonidentifiable, as for any constant Δ, it holds that:

$$\sum_{d=1}^{D} f_d(x_d) = f_i(x_i) + \Delta h_i(x_i) + \sum_{j \neq i} [f_j(x_j) - \Delta h_j(x_j)] \tag{5}$$

Concerns regarding concurvity were already raised in the 1990s when smoothing spline-based additive models were introduced [Buja *et al.*, 1989]. Using a more flexible function class such as NAMs only worsens the issue.

In order to address this type of nonidentifiability, several feature selection algorithms controlling for concurvity have been proposed: both the mRMR [De Jay *et al.*, 2013] and HSIC-Lasso [Climente-González *et al.*, 2019] methods adopt the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2005] to measure the dependence between two variables. This enables the formulation of scoring criteria in mRMR and the optimization objective in HSIC-Lasso. Moreover, the recent work by [Siems *et al.*, 2023] introduced a concurvity regularizer for differentiable additive models, enforcing decorrelation between transformed features by penalizing the average Pearson correlation coefficient computed on training batches. By carefully choosing the regularization strength from the measured concurvity–accuracy trade-off curve, the authors claim that their method can effectively stabilize model parameter estimates without significantly compromising performance.

Single identifiable model does not guarantee full insight into data under concurvity

These algorithms return a non-redundant, more parsimonious feature subset with comparable predictive performance to the full model. However, when strong concurvity exists, multiple subsets with equal size may support similar predictive performance, making it difficult to define the "right" answer. Let us consider the extreme case with two-dimensional inputs discussed in [Siems *et al.*, 2023], where x_1 and x_2 are identical, i.e. perfectly correlated: the regularized model will predict solely based on either one of them. However, this selection is arbitrary as there is no natural preference: either $\{x_1\}$ or $\{x_2\}$ could be a "right" subset for the model.

A more involved example can be found in [Kovács, 2022]:

$$\begin{aligned}
X_1 &\sim X_2 \sim X_3 \sim U(0, 1), \\
X_4 &= X_2^3 + X_3^2 + N(0, \sigma_1), \\
X_5 &= X_3^2 + N(0, \sigma_1), \\
X_6 &= X_2^2 + X_4^2 + N(0, \sigma_1), \\
X_7 &= X_1 \cdot X_2 + N(0, \sigma_1), \\
Y &= 2X_1^2 + X_5^3 + 2\sin(X_6) + N(0, \sigma_2),
\end{aligned} \tag{6}$$

with $\sigma_1 = 0.05, \sigma_2 = 0.5$. Obtaining nonzero shape functions for $\{X_1, X_5, X_6\}$ would be regarded as the optimal response as only these features show "direct" effects on the target Y. However, severe concurvity occurs, for example, on feature X_5, as its effect on the target can be easily approximated with the polynomial mapping $h_3 : X_3 \mapsto X_3^6$ (fig. 1). This makes $\{X_1, X_3, X_6\}$ another reasonable solution, as can be observed from the similar RMSE shown in table 1.

This scenario has been described as the "Rashomon" effect [Breiman, 2001], highlighting the possibility of multiple models giving near-optimal accuracy but different interpretations of data. It necessitates a careful specification of the dimension of model interpretability most relevant to each task as emphasized by [Lipton, 2017], such that an appropriate

Input subset	test-set RMSE
Full model	0.5205 ± 0.0096
$\{X_1, X_5, X_6\}$	0.5281 ± 0.0123
$\{X_1, X_3, X_6\}$	0.5387 ± 0.0092

Table 1: Results of fitting NAMs with different feature sets to data generated from eq. (6). Mean and standard deviation are computed over 5 random seeds.

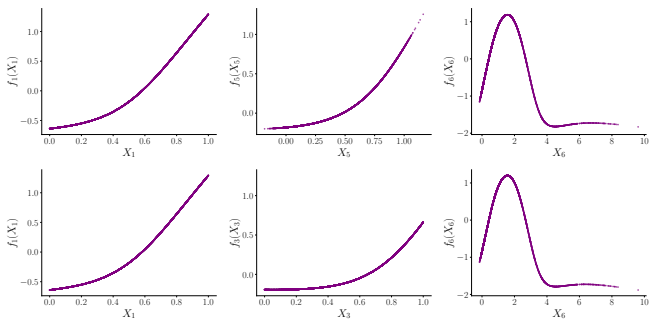

Figure 1: Shape functions learned by fitting NAMs on the subsets $\{X_1, X_5, X_6\}$ (top) and $\{X_1, X_3, X_6\}$ (bottom) to data generated from eq. (6).

criterion can be determined for meaningful model evaluation. For example, an identifiable algorithm might be sufficient when the main task is to automate the decision-making processes, as predictable model behavior is the primary demand; instead, if the goal is to gain comprehensive insight into the data, relying on one single feature subset supporting the minimal predictive error might not be a safe choice. [Fisher *et al.*, 2019] suggested to analyze variable importance on the *entire Rashomon set*, rather than on a *single best* model. Another important line to interpret feature effects is causal inference, which is beyond the scope of our paper.

Uncorrelated shape functions do not guarantee understandable marginal effects

Let us consider the two-dimensional example

$$f(\mathbf{x}) = \min(x_1, x_2) - 0.1x_1 - 0.1x_2, \tag{7}$$

where the individual effects $f_1(x_1) = -0.1x_1$ and $f_2(x_2) = -0.1x_2$ both slope downward, but the interaction effect $f_{12}(x_1, x_2)$ increases. Variables x_1 and x_2 follow a Gaussian distribution, $(x_1, x_2)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$. The covariance matrix Σ is represented as $\left[\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right]$, with an adjustable covariance term. We set $\rho = 0.9$ to induce a strong correlation between x_1 and x_2. Next, we consider the concurvity regularizer introduced by [Siems *et al.*, 2023] as an additional penalizing term to the optimization objective of NAMs:

$$\underset{f_i}{\text{argmin}}\; L(\hat{y}, f_i(x_i)) + \lambda \cdot R_\perp\left(\{f_i\}, \{x_i\}\right). \tag{8}$$

The first term is the task-specific loss function, and $R_\perp\left(\{f_i\}, \{x_i\}\right)$ is computed by averaging the Pearson correlation coefficient between shape functions:

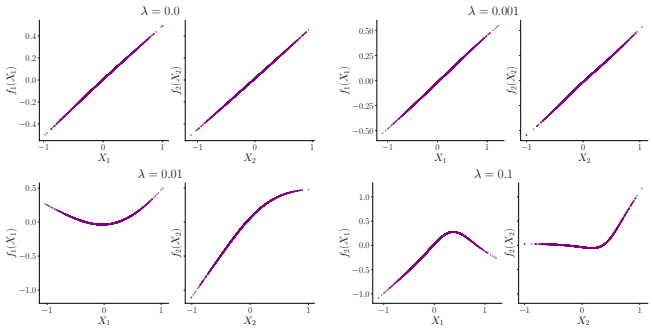

Figure 2: Shape functions learned by fitting NAMs with different concurvity regularization strength to data generated from eq. (7).

$$R_\perp(\{f_i\}, \{x_i\}) = \frac{2}{D(D+1)} \sum_{i=1}^{D} \sum_{j=i+1}^{D} |\text{Corr}(f_i(x_i), f_j(x_j)|. \tag{9}$$

The hyperparameter λ governs the tradeoff between goodness-of-fit and the uncorrelated nature of transformed features. While [Siems *et al.*, 2023] claim that achieving the latter can effectively be seen as a proxy for interpretability, fig. 2 shows a different story. Indeed, NAMs fitted using the objective eq. (9) give rise to gradually less correlated shape functions as λ increases. However, they remain dependent in a nonlinear manner. This hinders interpretability, as deciphering the unique contribution of either x_1 or x_2 on the response remains challenging, and suggests employing other penalties like HSIC to handle nonlinear dependencies.

4 Conclusions

The recent introduction of GAM's deep learning counterpart, NAMs, offered the promise of flexible, easy-to-train and transparent models. While the additive nature of these models can be seen as a "seal of interpretability" through their apparent transparency, specifically given the effect of each feature can be visualized, interpretability still remains a challenge. As we illustrated, this is due to several nonidentifiability issues of the shape functions. Though orthogonality constraints effectively allow one to distinguish between $f_1(\cdot) + f_2(\cdot)$ and $(f_1(\cdot) - \Delta) + (f_2(\cdot) - \Delta)$, they fall short as far as concurvity is concerned. As such, linear dependencies can still be found among shape functions. This blurs the interpretation of the actual effect of a feature on the response. Ultimately, since GAMs cannot select optimal feature sets, the selection remains to be done by domain experts. Domain experts may also be able to attribute a *causal* meaning to each feature, thus easing the choice of an optimal feature subset. For this reason, providing experts with an ensemble of solutions rather than a unique candidate set, as proposed by Rashomon-set-based approaches, currently seems like the best tradeoff [Zhong *et al.*, 2023].

Acknowledgements

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial In-

telligence FCAI and decision 341763) and EU Horizon 2020 (European Network of AI Excellence Centres ELISE, grant agreement 951847). We also acknowledge the computational resources provided by the Aalto Science-IT Project from Computer Science IT. XZ acknowledges the use of OpenAI's GPT-3.5 language model for assistance in polishing the language of this manuscript. The AI tool was utilized to enhance the clarity and readability of the text without influencing the substantive content of the research.

References

[Agarwal *et al.*, 2021] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural Additive Models: Interpretable Machine Learning with Neural Nets. In *Advances in Neural Information Processing Systems*, volume 34, pages 4699–4711. Curran Associates, Inc., 2021.

[Ashley Petersen and Simon, 2016] Daniela Witten Ashley Petersen and Noah Simon. Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25(4):1005–1025, 2016. PMID: 28239246.

[Bouchiat *et al.*, 2024] Kouroche Bouchiat, Alexander Immer, Hugo Yèche, Gunnar Rätsch, and Vincent Fortuin. Improving neural additive models with Bayesian principles, 2024.

[Breiman, 2001] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231, 2001.

[Buja *et al.*, 1989] Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989.

[Chang *et al.*, 2021] Chun-Hao Chang, Sarah Tan, Ben Lengerich, Anna Goldenberg, and Rich Caruana. How Interpretable and Trustworthy are GAMs? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 95–105, New York, NY, USA, August 2021. Association for Computing Machinery.

[Climente-González *et al.*, 2019] Héctor Climente-González, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14):i427–i435, 07 2019.

[De Jay *et al.*, 2013] Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*, 29(18):2365–2368, 2013.

[Durrande *et al.*, 2013] N. Durrande, D. Ginsbourger, O. Roustant, and L. Carraro. ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, 115:57–67, 2013.

[Duvenaud *et al.*, 2011] David K Duvenaud, Hannes Nickisch, and Carl Rasmussen. Additive Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[Enouen and Liu, 2022] James Enouen and Yan Liu. Sparse Interaction Additive Networks via Feature Interaction Detection and Sparse Selection. In *Advances in Neural Information Processing Systems*, volume 35, pages 13908–13920. Curran Associates, Inc., 2022.

[Fisher *et al.*, 2019] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[Friedman, 2000] Jerome Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 11 2000.

[Gretton *et al.*, 2005] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[Hastie and Tibshirani, 1986] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986.

[Kovács, 2022] László Kovács. Feature selection algorithms in generalized additive models under concurvity. *Computational Statistics*, 39:1–33, 11 2022.

[Lin and Zhang, 2006] Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5), October 2006.

[Lipton, 2017] Zachary C. Lipton. The mythos of model interpretability, 2017.

[Liu *et al.*, 2020] Guodong Liu, Hong Chen, and Heng Huang. Sparse shrunk additive models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6194–6204. PMLR, 13–18 Jul 2020.

[Lou *et al.*, 2012] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 150–158, New York, NY, USA, 2012. Association for Computing Machinery.

[Lou *et al.*, 2013] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 623–631, New York, NY, USA, 2013. Association for Computing Machinery.

[Lu *et al.*, 2022] Xiaoyu Lu, Alexis Boukouvalas, and James Hensman. Additive Gaussian processes revisited. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine*

Learning Research, pages 14358–14383. PMLR, 17–23 Jul 2022.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Märtens *et al.*, 2019] Kaspar Märtens, Kieran Campbell, and Christopher Yau. Decomposing feature-level variation with covariate Gaussian process latent variable models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4372–4381. PMLR, 09–15 Jun 2019.

[Radenovic *et al.*, 2022] Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural Basis Models for Interpretability. In *Advances in Neural Information Processing Systems*, volume 35, pages 8414–8426. Curran Associates, Inc., 2022.

[Rai, 2020] Arun Rai. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141, January 2020.

[Ramsay *et al.*, 2003] Timothy O. Ramsay, Richard T. Burnett, and Daniel Krewski. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, 14(1):18–23, 2003.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 05 2019.

[Siems *et al.*, 2023] Julien Siems, Konstantin Ditschuneit, Winfried Ripken, Alma Lindborg, Maximilian Schambach, Johannes Otterbach, and Martin Genzel. Curve Your Enthusiasm: Concurvity Regularization in Differentiable Generalized Additive Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 19029–19057. Curran Associates, Inc., 2023.

[Yang *et al.*, 2021] Zebin Yang, Aijun Zhang, and Agus Sudjianto. GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192, 2021.

[Zhong *et al.*, 2023] Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models, 2023.