Additive Distributional Forests for Interpretable Modeling with Applications in Insurance

Hong Li [*,1], Qifan Song [†,2], Jianxi Su [‡,2,3], and Ziyi Wang [§,2]

[1]University of Guelph, Guelph, ON N1G 2W1, Canada

[2]Purdue University, West Lafayette, Indiana 47907, United States

[3]RISC Foundation, Toronto, Ontario M2N 7E9, Canada

Abstract

Distributional forests are an ensemble tree learning method for estimating the conditional distribution of the response given covariates. The method allows analysts to examine various distributional features, such as dispersion, skewness, and tail behavior of the response distribution, beyond conditional mean. These distributional features often play an important role in risk assessment and decision-making. Despite the ease of implementation and nonparametric flexibility, distributional forests inherit the limited interpretability of general ensemble tree methods, as the estimated distribution function is expressed as implicit aggregations of joint covariate effects. To address this limitation, we propose an additive distributional regression framework that explicitly decomposes the conditional distribution function of the response into a linear aggregation of single-covariate-specific component functions. When the component functions are modeled using distributional forests, we further develop an expectation–smoothing iterative algorithm to estimate the model. Real data are used to demonstrate the practical usefulness of the proposed additive distributional forest method.

Key words and phrases: Distributional regression, nonlinear effects, heteroscedasticity, healthcare, medical expenditures

[*] e-mail: lihong@uoguelph.ca

[†] e-mail: qfsong@purdue.edu

[‡] Corresponding author; e-mail: jianxi@purdue.edu

[§] e-mail: wang4538@purdue.edu

1 Introduction

Mean regression has long been adopted by actuaries to identify key risk drivers of actuarial outcomes (e.g., claim frequency, claim severity, aggregate losses, and healthcare expenditures) and to examine their impacts on the average level of the response variable. To conduct mean regression, relevant paradigms range from classical linear regression and generalized linear models (De Jong and Heller, 2008; Frees, 2009; McCullagh and Nelder, 1989), to more flexible nonparametric methods such as spline-based regression (Hastie and Tibshirani, 1990), Bayesian nonparametric approaches (Huang and Meng, 2020; Richardson and Hartman, 2018), clustering-based segmentation methods (Gan and Valdez, 2020; Jamotton et al., 2024), and tree-based methods (James et al., 2023). Among these, ensemble tree methods have shown superior statistical performance, comparable to or sometimes even exceeding that of more complex deep learning methods. Combined with their conceptual simplicity and ease of implementation, ensemble tree methods have gained increasing popularity among actuarial risk analysts.

Despite the widespread use of mean regression, many actuarial and risk management problems require consideration beyond the conditional mean of the response. For instance, in claim management, it is often necessary to assess how a policyholder's profile influences not only the central tendency of future claims but also the variability and tail behavior of the claim distribution. These additional distributional characteristics can have direct implications for operational and financial decisions, such as underwriting, claim triage, fraud detection, and reserve adequacy assessment. Similar considerations arise in several other contexts. In ratemaking, risk loading is inherently linked to dispersion and tail risk, so two policies with similar expected loss may warrant different premiums when their loss volatility or upper-tail exposure differs substantially. In reinsurance pricing, the study of tail behavior is of central importance because contract structures such as excess-of-loss are explicitly triggered by extreme outcomes. In healthcare analytics, a key objective is to identify risk drivers of high-cost claimants, where the upper tail of the cost distribution carries the most immediate managerial relevance for early intervention and resource allocation. For related discussions from the practitioners' perspective, we refer to Reed (2015).

To understand how covariates affect the shape of the response distribution, ensemble tree methods, which have been successfully used in mean regression problems, can be naturally extended to model the conditional distribution of the response. In particular, distributional forests, which are also known as quantile regression forests (Meinshausen, 2006), estimate the conditional cumulative distribution function (CDF) of the response by adopting the same splitting strategy as random forests to partition the feature space, while replacing the local empirical mean estimator within each terminal node by an empirical CDF estimator. Owing to this conceptual and algorithmic similarity to random forests, distributional forests can be adopted for distributional regression with minimal additional modeling and implementation effort. Further, it is known that under mild regularity conditions, the distributional forest estimator is consistent (Meinshausen, 2006). Therefore, in principle, with sufficiently large data sets, distributional forests can accurately recover the target conditional CDF of the response.

That said, the aforementioned ease of implementation and nonparametric flexibility come at the cost of interpretability, a trade-off encountered in many machine learning methods. Specifically, the conditional CDF estimated by a distributional forest is constructed through an implicit aggregation of joint covariate effects across a large collection of tree-based partitions. As a result, it is typically unclear how to disentangle and interpret the marginal distributional impact of an individual covariate. This lack of structural interpretability may undermine the usefulness of distributional forests in applications where assessing risk driver impact, communicating model behavior, or supporting decision-making is as important as predictive performance. From a regulatory and model risk governance perspective, we refer to Henckaerts et al. (2021); Wüthrich and Merz (2019), which emphasize that machine learning models for insurance usage should remain transparent and possess a reasonable level of interpretability.

This paper aims to address the limited interpretability inherent in distributional forests. Motivated by the success of additive mean regression models in balancing interpretability and modeling flexibility in capturing nonlinear covariate effects, we extend the underlying additive decomposition principle to distributional forests. Specifically, we propose to decompose the conditional CDF of the response into a linear aggregation of component functions, each depending on a single covariate. We show that each component function is equivalent to a valid marginal conditional CDF, and we estimate these component functions using distributional forests. The resulting additive distributional forests can provide tractable interpretations of covariate-specific marginal distributional effects while retaining the flexibility of tree-based learners in capturing nonlinear relationships.

We note that imposing an additive structure on distributional forests may reduce their statistical fidelity, as interactions among covariates cannot be captured explicitly. The same limitation is also recognized in the context of additive mean regression models. When the primary objective of a modeling task is solely predictive accuracy, users should be cautious in applying the proposed additive framework and may instead favor more general nonparametric approaches that capture the joint covariate effects without imposing structural constraints.

Several alternative approaches to distributional regression are available (Kneib et al., 2023). For instance, generalized linear models can be viewed as a classical parametric framework for distributional regression, though the regression structure is typically imposed on parameters that govern the mean of a specified family of distributions. Quantile regression embodies another route to distributional regression via estimating the conditional quantile function of the response, which can be implemented nonparametrically (Koenker, 2005); see Baione and Biancalana (2019); Heras et al. (2018); Kudryavtsev (2009); Li et al. (2021) for its applications in insurance. However, standard quantile regression is typically fitted at one probability level at a time. As a result, quantile estimates obtained separately across different probability levels may exhibit crossing unless additional non-crossing constraints or post-processing strategies are imposed; see (Dai et al., 2023) for a comprehensive discussion. From this perspective, distributional forests represent a more convenient approach that directly returns a valid conditional CDF by construction, and thus multiple levels of quantiles can be computed simultaneously.

The rest of this paper is organized as follows. In Section 2, we provide a concise review of distributional forests. In Section 3, we present the proposed interpretable distributional regression framework, which decomposes the conditional CDF of the response as a linear aggregation of single-covariate-dependent component functions. In Section 4, we propose to estimate these component functions using distributional forests and develop an iterative algorithm for estimating the resulting additive distributional forests. The performance of the proposed estimation procedure is evaluated in Section 5 through simulation studies. In Section 6, we illustrate the practical usefulness of the proposed additive distributional forest framework using real data. Finally, Section 7 concludes the paper.

2 Background on distributional forests

Consider a set of $n \in \mathbb{N}$ training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $y_i \in \mathbb{R}$ denotes a scalar response variable and $\boldsymbol{x}_i \in \Omega \subset \mathbb{R}^p$ is a vector of $p \in \mathbb{N}$ explanatory variables associated with the i-th sample, $i = 1, \ldots, n$. Here and throughout, Ω denotes the feature space containing all plausible values that the explanatory variable vector $\boldsymbol{x}$ can take.

Distributional forests can be viewed as a natural extension of regression forests. We therefore begin by reviewing the regression tree framework and its ensemble generalization via random forests to establish the necessary background and notation, and then proceed to discussing distributional forests.

2.1 Random forests

To construct a regression tree, the algorithm recursively partitions the feature space Ω into a finite collection of disjoint regions by repeatedly splitting either Ω itself or one of its previously formed subregions along a single explanatory variable. At each internal node, the splitting variable and split point are selected to optimize a local impurity criterion. One commonly used criterion is to minimize the within-node sum of squared errors, or equivalently, to maximize the reduction in the conditional variance of the response variable. Based on the selected impurity criterion, the optimal split is identified via a discrete, greedy search over candidate splits. This recursive binary splitting strategy keeps splitting the feature space Ω into a partition:

$$\Omega = \bigcup_{l=1}^{L} R_l,$$

where $\{R_l\}_{l=1}^L$ denote the terminal regions, or leaves, of the tree, and $L \in \mathbb{N}$ denotes the total number of leaves determined by the user-specified stopping parameters such as minimum node size, maximum tree depth, maximum number of leaves, and so forth.

After a regression tree is fitted to the training data, it predicts the conditional mean of the response variable Y given a query point $\boldsymbol{x} \in \Omega$ by averaging the observed responses within the terminal region that contains $\boldsymbol{x}$. More formally, the tree-based estimator can be written as

$$\widehat{\mathbb{E}}(Y \mid \boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i(\boldsymbol{x})\, y_i, \qquad \text{with} \quad \alpha_i(\boldsymbol{x}) = \frac{\mathbb{1}\{\boldsymbol{x}_i \in R(\boldsymbol{x})\}}{\sum_{j=1}^{n} \mathbb{1}\{\boldsymbol{x}_j \in R(\boldsymbol{x})\}}, \tag{1}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, and $R(\boldsymbol{x})$ represents the leaf to which $\boldsymbol{x}$ is assigned.

Random forests are an extension of regression trees designed for improving their stability and predictive performance by aggregating a large number of decorrelated trees through ensemble learning (Breiman, 2001). A random forest consists of $K \in \mathbb{N}$ regression trees, each trained on a bootstrap sample of the original data and/or grown using randomized feature selection, techniques known as bootstrap aggregating (Breiman, 1996) and feature bagging (Ho, 1998), respectively. For $k = 1, \ldots, K$, the k-th tree partitions the feature space Ω into $L_k > 0$ disjoint terminal regions, denoted by $\{R_{k,l}\}_{l=1}^{L_k}$. Given a query point $\boldsymbol{x} \in \Omega$, the k-th tree assigns $\boldsymbol{x}$ to a unique terminal region, denoted by $R_k(\boldsymbol{x})$, and predicts the response using the sample average of the training observations falling in that region. Accordingly, the random forest predictor is obtained by averaging the predictions from all k trees. The random forest estimator can be expressed in a form analogous to the regression tree estimator in Equation (1) as a locally weighted regression estimator, but with a different set of weights given by

$$\widehat{\mathbb{E}}(Y \mid \boldsymbol{x}) = \sum_{i=1}^{n} \omega_i(\boldsymbol{x})\, y_i, \qquad \omega_i(\boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} \alpha_{i,k}(\boldsymbol{x}), \qquad \alpha_{i,k}(\boldsymbol{x}) = \frac{\mathbb{1}\{\boldsymbol{x}_i \in R_k(\boldsymbol{x})\}}{\sum_{j=1}^{n} \mathbb{1}\{\boldsymbol{x}_j \in R_k(\boldsymbol{x})\}}, \tag{2}$$

where $\alpha_{i,k}(\boldsymbol{x})$ denotes the contribution weight assigned to the i-th training observation by the k-th tree in the forest, and $\omega_i(\boldsymbol{x})$ averages these weights across all trees and reflects the overall forest-induced similarity between $\boldsymbol{x}$ and $\boldsymbol{x}_i$ based on how frequently $\boldsymbol{x}_i$ falls into the same terminal region as $\boldsymbol{x}$. By construction, for any $\boldsymbol{x} \in \Omega$, the weights $\{\omega_i(\boldsymbol{x})\}_{i=1}^{n}$ are nonnegative and sum to one.

2.2 Distributional forests

The conditional mean $\mathbb{E}(Y \mid \boldsymbol{x})$ only captures one aspect of the conditional distribution of Y given $\boldsymbol{x}$. In contrast, the conditional CDF

$$F(y \mid \boldsymbol{x}) := \mathbb{P}(Y \leq y \mid \boldsymbol{x}),$$

provides a complete characterization of the conditional distribution of the response, from which the conditional mean and other distributional properties can be derived. Distributional forests extend random forests from conditional mean estimation to conditional CDF estimation (Meinshausen, 2006). Specifically, a distributional forest predicts the conditional CDF $F(y \mid \boldsymbol{x})$ via

$$\widehat{F}(y \mid \boldsymbol{x}) = \sum_{i=1}^{n} \omega_i(\boldsymbol{x}) \, \mathbb{1}\{y_i \leq y\}, \tag{3}$$

where the weights $\{\omega_i(\boldsymbol{x})\}_{i=1}^{n}$ are defined in exactly the same manner as those used in the random forest estimator in Equation (2).

Algorithmically, the formulation in Equation (3) implies that distributional forests adopt exactly the same partitioning scheme as random forests. The distinction between the two methods lies solely in the leaf-level aggregation. Namely, random forests compute the sample mean of responses within each terminal region, whereas distributional forests compute the corresponding empirical CDF. At the forest level, both estimators in Equations (2) and (3) aggregate the leaf-level estimates in the same way, and therefore share the same set of locally adaptive weights $\{\omega_i(\boldsymbol{x})\}_{i=1}^{n}$. This shared splitting and weighting structure makes the implementation of distributional forests particularly convenient, since efficient algorithms and software packages for constructing random forests are already widely available. To implement distributional forests, the user only needs to extract the tree-induced partitions and forest weights from a fitted random forest model and replace the empirical mean estimator at each leaf with the corresponding empirical CDF estimator.

3 Additive distributional regression

The distributional forest framework reviewed in Section 2 enables flexible nonparametric modeling of the entire distribution of Y conditional on $\boldsymbol{x}$. However, because the resulting conditional CDF is driven by joint covariate effects, the marginal distributional contribution of each covariate is not explicitly identifiable. To address this limitation, in this section, we introduce an additive structure for distributional regression. This framework forms the methodological foundation, upon which the interpretable distributional forests are built.

For ease of presentation, we begin with the two-covariate case and examine its structural properties. Extending to a multi-covariate setting is conceptually straightforward, but the corresponding theoretical analyses become notationally cumbersome. Thereby, we confine the rigorous technical analysis to the two-covariate setting.

3.1 The two-covariate case

Throughout this subsection, we focus on the two-covariate setting with $\boldsymbol{x} = (x_1, x_2)^\top \in \Omega$. Recall that to enhance interpretability, the additive mean regression method models the conditional mean of Y given $\boldsymbol{x}$ via

$$\mathbb{E}(Y \,|\, \boldsymbol{x}) = \Psi_1(x_1) + \Psi_2(x_2), \tag{4}$$

where each component function Ψ_i represents the marginal contribution of the covariate x_i to the conditional mean of Y. In practice, the component functions are typically modeled using flexible smoothers, such as splines, to accommodate nonlinear effects.

Borrowing this idea, to enable a similarly interpretable characterization of how each covariate impacts the entire conditional distribution of the response, we impose an additive structure on the conditional CDF:

$$F(y \,|\, \boldsymbol{x}) = \Psi_1(y, x_1) + \Psi_2(y, x_2). \tag{5}$$

The additive specification in Equation (5) is not intended as a universal approximation to arbitrary conditional distributions, but rather as a deliberately structured modeling choice that prioritizes interpretability of marginal distributional effects, in direct analogy to additive mean regression models. The component functions Ψ_1 and Ψ_2 may be viewed as main effects on the conditional distribution of Y, describing how each covariate individually shapes the response distribution while holding the other covariate fixed.

It is noteworthy that a key distinction between additive mean regression and additive distributional regression lies in the nature of the modeling target, which naturally leads to several implications. First, unlike the additive mean regression model in Equation (4), where for a given $\boldsymbol{x}$, the modeling target is a scalar quantity, the object we aim to model in Equation (5) is the conditional CDF, which is a function of y. Consequently, each component function Ψ_i depends not only on the corresponding covariate x_i, but also on the response level y, which reflects how x_i influences the distribution of Y across different quantiles or regions of the support.

Moreover, imposing an additive structure at the distributional level raises a pivotal yet nontrivial question regarding the admissible behavior of the component functions. Clearly, not any form of component functions can accommodate the additive structure in Equation (5). For example, if both component functions Ψ_1 and Ψ_2 are decreasing in y, then their sum is also decreasing, which would yield an invalid conditional CDF. The following assertion characterizes conditions under which the additive decomposition in Equation (5) leads to a legitimate conditional CDF. In particular, we show that there always exists an additive decomposition such that the component functions can be chosen to be scaled CDFs.

Proposition 1. *If the functional additive structure in (5) holds for a conditional CDF $F(y \mid \boldsymbol{x})$, then there exist a pair of component functions $\Psi_1(y, x_1)$ and $\Psi_2(y, x_2)$ which are non-decreasing, right-continuous in y, and satisfy $\Psi_1(-\infty, x_1) = \Psi_2(-\infty, x_2) = 0$ for all x_1 and x_2, such that (5) holds.*

Proof. See Appendix A. □

Henceforth, we shall assume that the component functions Ψ_1 and Ψ_2 satisfy the conditions outlined in Proposition 1, which ensures the validity of the additive decomposition in (5). Since $F(\infty \mid x_1, x_2) =$

1 for all pairs (x_1, x_2), it is elementary to see that $\alpha_1 := \Psi_1(\infty, x_1)$ does not depend on x_1, nor does $\alpha_2 := \Psi_2(\infty, x_2)$ depend on x_2. Moreover, it must hold that $\alpha_1 + \alpha_2 = 1$. Consequently, these component functions can be represented as

$$\Psi_1(y, x_1) = \alpha_1 \, F_1(y \,|\, x_1), \qquad \Psi_2(y, x_2) = \alpha_2 \, F_2(y \,|\, x_2),$$

where F_1 and F_2 are valid conditional CDFs. Thereby, the additive decomposition (5) can be equivalently written as

$$F(y \,|\, x_1, x_2) = \alpha_1 \, F_1(y \,|\, x_1) + \alpha_2 \, F_2(y \,|\, x_2), \qquad \alpha_1 + \alpha_2 = 1.$$

This representation corresponds to a mixture model. Specifically, conditional on (x_1, x_2), the response variable Y is generated in a way such that, with probability α_1, $(Y \,|\, x_1, x_2)$ follows the distribution $F_1(\cdot \,|\, x_1)$, and with probability α_2, it follows $F_2(\cdot \,|\, x_2)$.

This mixture representation leads to several complementary interpretations. From a causal perspective, the representation suggests a latent regime mechanism. Namely, with probability α_1, the outcome Y is driven primarily by covariate x_1, while x_2 plays no role; with probability α_2, the reverse holds. The representation can also be viewed through a population heterogeneity lens, where the population consists of two latent sub-populations, each governed by a different data-generating mechanism. In one sub-population, the conditional distribution of Y depends mainly on x_1, whereas in the other it depends mainly on x_2. The mixing weights α_1 and α_2 represent the relative prevalence of these sub-populations. Further, from a risk-decomposition perspective, the mixture representation reflects the idea that, for any given realization, one of two competing risk drivers, either x_1 or x_2, dominates the outcome, while the other is effectively inactive. Collectively, these complementary interpretive perspectives support the practical usefulness of the proposed additive distributional regression framework for deriving analytical insights into the marginal impact of individual covariates on the distributional behavior of the response variable.

We openly acknowledge that the additive decomposition in (5) deliberately abstracts away from interaction effects among covariates in exchange for enhanced interpretability and analytical transparency. Similar compromises are routinely adopted in linear models and other additive modeling frameworks, where the primary objective is to obtain interpretable marginal effects rather than to exhaustively capture all possible dependencies.

3.2 The multi-covariate case

In this subsection, we deal with a p-dimensional covariate vector $\boldsymbol{x} = (x_1, \dots, x_p)^\top \in \Omega$. In direct analogy with the two-covariate formulation in (5), we consider the additive structure

$$F(y \,|\, \boldsymbol{x}) = \sum_{j=1}^{p} \Psi_j(y, x_j), \tag{6}$$

where each component function $\Psi_j(y, x_j)$ captures the marginal contribution of covariate x_j to the conditional CDF of the response variable Y.

As in the two-covariate case, imposing an additive structure at the distributional level raises similar questions regarding the admissible behavior of the component functions in (6). To ensure that the additive decomposition defines a valid conditional CDF, the main ideas underlying Proposition 1 can be extended to the multi-covariate setting via a series of two-block decomposition iterations. To sketch out the iterative process, fix an arbitrary ordering of the covariates and write $\boldsymbol{x} = (x_1, \boldsymbol{x}_{-1})^\top$, where $\boldsymbol{x}_{-1} = (x_2, \dots, x_p)^\top$. Using the same argument as in Proposition 1, one can show that there exist non-decreasing and right-continuous component functions $\Psi_1(y, x_1)$ and $\widetilde{\Psi}_{-1}(y, \boldsymbol{x}_{-1})$ satisfying $\Psi_1(-\infty, x_1) = \widetilde{\Psi}_{-1}(-\infty, \boldsymbol{x}_{-1}) = 0$, such that the decomposition

$$F(y \,|\, \boldsymbol{x}) = \Psi_1(y, x_1) + \widetilde{\Psi}_{-1}(y, \boldsymbol{x}_{-1})$$

yields a valid conditional CDF of Y given $(x_1, \boldsymbol{x}_{-1})$.

Next, treat $\widetilde{\Psi}_{-1}(y, \boldsymbol{x}_{-1})$ as another scaled conditional CDF to be decomposed, and further split $\boldsymbol{x}_{-1} = (x_2, \boldsymbol{x}_{-2})^\top$ with $\boldsymbol{x}_{-2} = (x_3, \dots, x_p)^\top$. A repeated application of the same argument as in Proposition 1 implies that there exist component functions $\Psi_2(y, x_2)$ and $\widetilde{\Psi}_{-2}(y, \boldsymbol{x}_{-2})$ that are non-decreasing and right-continuous in y and satisfy $\Psi_2(-\infty, x_2) = \widetilde{\Psi}_{-2}(-\infty, \boldsymbol{x}_{-2}) = 0$, such that the decomposition

$$\widetilde{\Psi}_{-1}(y, \boldsymbol{x}_{-1}) = \Psi_2(y, x_2) + \widetilde{\Psi}_{-2}(y, \boldsymbol{x}_{-2})$$

yields a valid scaled conditional CDF for $\widetilde{\Psi}_{-1}(y, \boldsymbol{x}_{-1})$. Repeating the above argument to sequentially isolate the marginal effects of all covariates leads to an analogous conclusion to Proposition 1 for the multi-covariate case. Accordingly, throughout the multi-covariate setting we assume that the component functions $\Psi_j(y, x_j)$ are non-decreasing and right-continuous in y and satisfy $\Psi_j(-\infty, x_j) = 0$ for all x_j, $j = 1, \ldots, p$.

Moreover, under these conditions, each component function $\Psi_j(y, x_j)$ can be treated as a scaled conditional CDF:

$$\Psi_j(y, x_j) = \alpha_j \, F_j(y \,|\, x_j), \qquad j = 1, \ldots, p,$$

where $\alpha_j \geq 0$, $\sum_{j=1}^{p} \alpha_j = 1$, and $F_j(\cdot \,|\, x_j)$ is a valid conditional CDF. Consequently, the multi-covariate additive distributional regression model admits the following representation:

$$F(y \,|\, \boldsymbol{x}) = \sum_{j=1}^{p} \alpha_j \, F_j(y \,|\, x_j), \tag{7}$$

which may be interpreted as a finite mixture of single-covariate-specific distributional components.

4 Additive distributional forests and their estimation

In practice, the component conditional CDFs $F_j(y \,|\, x_j)$ involved in Equation (7) may be modeled using nonparametric tools that accommodate the distributional relationship between x_i and y. In particular, when each component CDF is estimated using distributional forests, we refer to the resulting model as an *additive distributional forest* (ADF). This construction combines the interpretability of additive distributional regression with the flexibility of tree-based distributional learners.

To implement ADF to estimate the component CDFs F_j, the mixture representation in Equation (7) motivates us to develop an iterative algorithm analogous to the classical expectation–maximization (EM) algorithm, which has been widely adopted for estimating parametric mixture models. Following the terminology discussed in Li et al. (2021), we term the iterative algorithm for fitting ADF the expectation–smoothing (ES) algorithm, reflecting the nonparametric nature of the learner involved.

To describe the ES algorithm, it is convenient to introduce a latent indicator $Z_i \in \{1, \ldots, p\}$ for each training observation $(x_{i1}, \ldots, x_{ip}, y_i)$, where x_{ij} denotes the value of the j-th covariate in the i-th sample. Then the mixture representation in Equation (7) admits the following representation:

$$\mathbb{P}(Z_i = j) = \alpha_j, \qquad (Y_i \mid Z_i = j,\ x_{i1}, \ldots, x_{ip}) \sim F_j(\,\cdot \mid x_{ij}\,), \qquad i = 1, \ldots, n.$$

We emphasize that the latent indicator Z_i is introduced primarily as an algorithmic device for constructing observation-specific weights that guide the iterative fitting procedure, and should not be over-interpreted as representing a literal subpopulation membership. Throughout, let us define posterior responsibilities:

$$r_{ij} = \mathbb{P}(Z_i = j \mid \boldsymbol{x}_i, y_i), \qquad i = 1, \ldots, n, \quad j = 1, \ldots, p.$$

Suppose that at the t-th iteration of the ES algorithm, we have obtained the additive weight estimates $\widehat{\alpha}_j^{(t)}$, and the posterior responsibility estimates $\widehat{r}_{ij}^{(t)}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. To update the estimates in the $(t+1)$-th iteration, the S-step refits each component's conditional CDF $F_j(\cdot \mid x_j)$ using a weighted distributional forest. Specifically, we apply the tree-based procedure described in Section 2 to the weighted sample $\{(x_{ij}, y_i)\}_{i=1}^n$, where observation i is assigned posterior weight $\widehat{r}_{ij}^{(t)}$. In this way, observations with larger posterior weights $\widehat{r}_{ij}^{(t)}$ play a more significant role in determining the splitting of the forests for the j-th covariate. Based on the refitted forest, for any query point x_j, the ensemble induces weights $\{\omega_{ij}^{(t+1)}(x_j)\}_{i=1}^n$ satisfying $\omega_{ij}^{(t+1)}(x_j) \geq 0$ and $\sum_{i=1}^n \omega_{ij}^{(t+1)}(x_j) = 1$, and yields the forest-based conditional CDF estimator

$$\widehat{F}_j^{(t+1)}(y \mid x_j) = \sum_{i=1}^n \omega_{ij}^{(t+1)}(x_j)\, \mathbb{1}(y_i \leq y).$$

In the E-step, we define a local scoring quantity $s_{ij}^{(t+1)}$ that measures how well the j-th component distribution explains the observed response y_i at covariate value x_{ij}. Since the distributional forests grown in the S-step already yield weighted local CDF estimates, we naturally adopt the same weighting schemes to construct $s_{ij}^{(t+1)}$ using a kernel-smoothed local density estimator. To be specific, we define

$$s_{ij}^{(t+1)} := \widehat{f}_j^{(t+1)}(y_i \,|\, x_{ij}) = \sum_{k=1}^{n} \omega_{kj}^{(t+1)}(x_{ij})\, K_h(y_i - y_k), \qquad K_h(u) = h^{-1}\, K(u/h),$$

where $K(\cdot)$ is a kernel function (e.g., Gaussian) and $h > 0$ is a bandwidth parameter. Using these local scores, the posterior responsibilities are updated according to

$$\widehat{r}_{ij}^{(t+1)} = \frac{\widehat{\alpha}_j^{(t)}\ s_{ij}^{(t+1)}}{\sum_{k=1}^{p} \widehat{\alpha}_k^{(t)}\ s_{ik}^{(t+1)}}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, p.$$

Moreover, the additivity weights are updated as

$$\widehat{\alpha}_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{r}_{ij}^{(t+1)}, \qquad j = 1, \ldots, p.$$

The ES algorithm proceeds by iterating between the E-step and the S-step described above until convergence. In practice, convergence can be monitored using either the stabilization of the responsibilities or the mixture weights. For instance, a simple stopping criterion can be

$$\max_{1 \le i \le n} \sum_{j=1}^{p} \left| \widehat{r}_{ij}^{(t+1)} - \widehat{r}_{ij}^{(t)} \right| \le \varepsilon,$$

or alternatively,

$$\max_{1 \le j \le p} \left| \widehat{\alpha}_j^{(t+1)} - \widehat{\alpha}_j^{(t)} \right| \le \varepsilon,$$

for a prescribed tolerance level $\varepsilon > 0$. To avoid unnecessary iterations, we also impose a maximum number of iterations $T_{\max}$. After convergence, the final output of the ES algorithm consists of the fitted component conditional CDFs $\{\widehat{F}_j(\cdot \,|\, \cdot)\}_{j=1}^{p}$ as well as the estimated additive weights $\{\widehat{\alpha}_j\}_{j=1}^{p}$. Together, they yield the ADF estimator evaluated at a query point $(x_1, \ldots, x_p)$:

$$\widehat{F}(y \,|\, x_1, \ldots, x_p) = \sum_{j=1}^{p} \widehat{\alpha}_j\, \widehat{F}_j(y \,|\, x_j).$$

The ES algorithm discussed in this section is summarized in Algorithm 1.

Algorithm 1: Expectation–Smoothing algorithm for additive distributional forests

```
Input: Data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$; distributional-forest hyperparameters;
       kernel $K$ and bandwidth $h$ ; max epochs $T_{\max}$; tolerance $\varepsilon$.
Output: Estimated component forests $\{\widehat{F}_j\}_{j=1}^p$ and additive weights $(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)$.

Init. Set $\widehat{\alpha}_j^{(0)} \leftarrow 1/p$ for $j = 1, \ldots, p$; set $\widehat{r}_{ij}^{(0)} \leftarrow 1/p$ for all $i, j$.
for $t = 0, 1, \ldots, T_{\max} - 1$ do
    // S-step: weighted refit of component distributional forests
    for $j = 1$ to $p$ do
        Fit $\widehat{F}_j^{(t+1)}$ on $\{(x_{ij}, y_i)\}_{i=1}^n$ using posterior responsibilities $\{\widehat{r}_{ij}^{(t)}\}_{i=1}^n$.
    // E-step: local scoring and posterior update
    for $i = 1$ to $n$ do
        for $j = 1$ to $p$ do
            Compute forest weights $\{\omega_{kj}^{(t+1)}(x_{ij})\}_{k=1}^n$ induced by $\widehat{F}_j^{(t+1)}$ at the query point $x_{ij}$.
            Compute local score $s_{ij}^{(t+1)} \leftarrow \sum_{k=1}^n \omega_{kj}^{(t+1)}(x_{ij})\, K_h(y_i - y_k)$.
        for $j = 1$ to $p$ do
            $\widehat{r}_{ij}^{(t+1)} \leftarrow \widehat{\alpha}_j^{(t)} s_{ij}^{(t+1)} \Big/ \sum_{k=1}^p \widehat{\alpha}_k^{(t)} s_{ik}^{(t+1)}$.
    // Update additive weights
    for $j = 1$ to $p$ do
        $\widehat{\alpha}_j^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^n \widehat{r}_{ij}^{(t+1)}$.
    // Stopping criterion
    if $\max_{1 \le i \le n} \sum_{j=1}^p \big| \widehat{r}_{ij}^{(t+1)} - \widehat{r}_{ij}^{(t)} \big| \le \varepsilon$ then
        break
for $j = 1$ to $p$ do
    Refit $\widehat{F}_j$ on $\{(x_{ij}, y_i)\}_{i=1}^n$ using final weights $\{\widehat{r}_{ij}^{(\infty)}\}_{i=1}^n$.
    Set $\widehat{\alpha}_j \leftarrow \widehat{\alpha}_j^{(\infty)}$.
```

5 Simulation analysis

This section evaluates the finite-sample performance of the proposed ADF method through controlled simulation studies. We first consider an idealized setting in which the data-generating process satisfies the assumed additive structure, and then examine the robustness of the method when this structure is violated.

5.1 Estimation performance

Consider two features $\boldsymbol{x} = (x_1, x_2)^\top \in [0,1] \times [0,1]$, and assume that the conditional CDF of Y given $\boldsymbol{x}$ is given by

$$F(y \,|\, \boldsymbol{x}) = \alpha \times F_1(y \,|\, x_1) + (1-\alpha) \times F_2(y \,|\, x_2), \tag{8}$$

where $F_j(y \,|\, x_j) = \Phi\big(y; \mu_j(x_j), \sigma(x_j)\big)$, $j = 1, 2$, are the Gaussian CDF with means $\mu_j(\cdot) \in \mathbb{R}$ and standard deviations $\sigma(\cdot) \in \mathbb{R}_+$. Furthermore, we set $\alpha = 0.3$, and for $x \in [0,1]$,

- $\mu_1(x) = x^3 + 4x^2 + 2x + 8$;
- $\mu_2(x) = 3\sin(\pi\, x) + 5$;
- $\sigma(x) = 0.25 + 0.5\,(1 - \cos(\pi\, x))$.

Figure 1 illustrates the mean and standard deviation functions used in the data-generating mechanism, along with a simulated dataset of size $n = 3\,000$. The generated data reveal clear nonlinear dependencies of both covariates x_1 and x_2 on Y, as well as pronounced heteroskedasticity, reflecting how the conditional variance of Y varies with the input features.

Throughout this analysis, we assume that the data-generating process is unknown and apply the ES algorithm proposed in Section 4 to estimate the two components' conditional CDFs F_1 and F_2. To this end, we train the model using a simulated dataset of size $n = 3\,000$ generated from Equation (8), and evaluate its statistical performance on an independent testing set of size $m = 1\,000$.

To assess the model's ability to recover the conditional distribution of Y given (x_1, x_2), we compare the estimated CDF values $\widehat{F}(y_i^{\text{test}} \,|\, x_{1,i}^{\text{test}}, x_{2,i}^{\text{test}})$ with the true CDF values $F(y_i^{\text{test}} \,|\, x_{1,i}^{\text{test}}, x_{2,i}^{\text{test}})$ over the

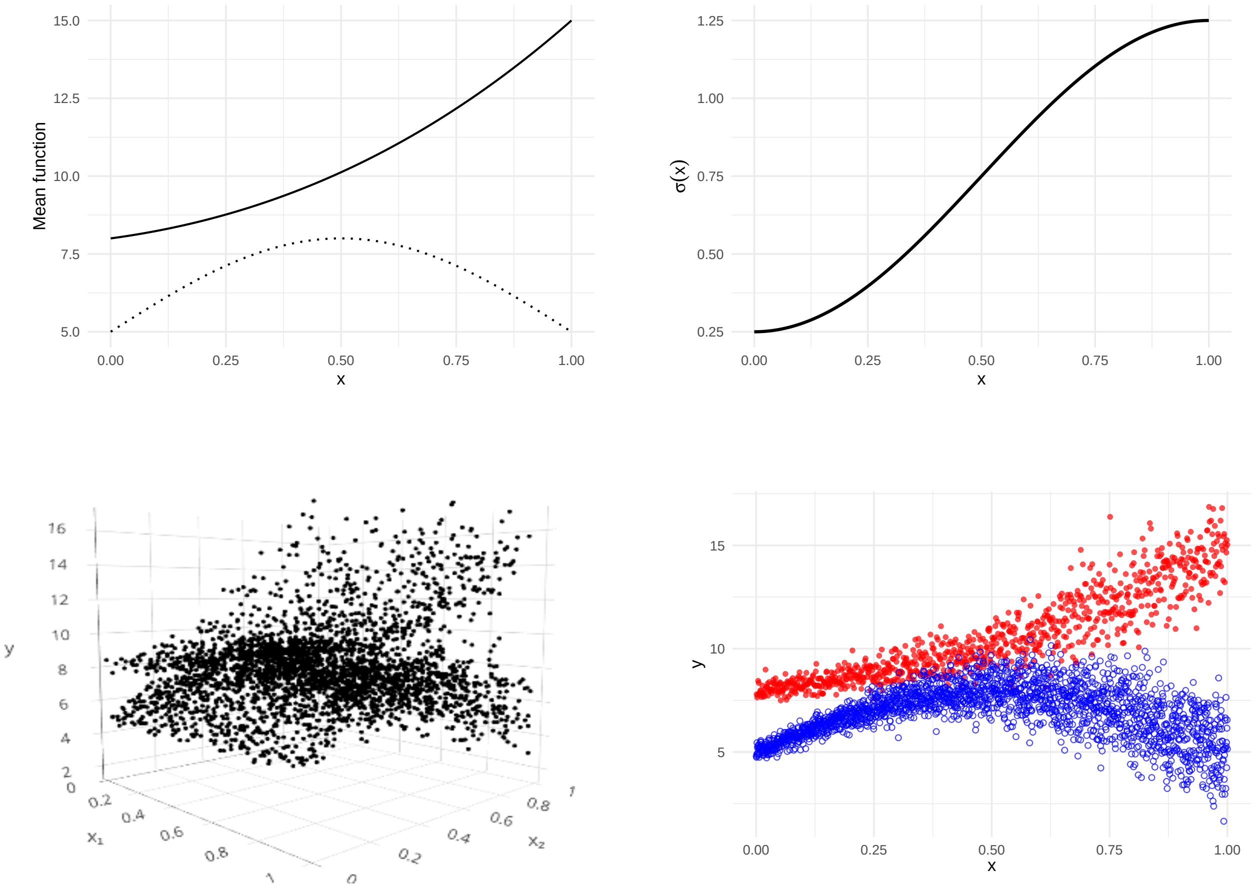

Figure 1: Illustration of the data-generating process. The top-left panel shows the conditional mean functions of the two additive components, with the solid and dotted curves corresponding to $x_1 \mapsto \mu_1(x_1)$ and $x_2 \mapsto \mu_2(x_2)$, respectively. The top-right panel displays the conditional standard deviation function $x \mapsto \sigma(x)$. The bottom-left panel presents the scatter plot of the simulated data. The bottom-right panel visualizes the simulated data projected onto the (x, y) plane, where red solid dots represent the F_1 component (varying x_1 while fixing x_2), and blue hollow circles represent the F_2 component (varying x_2 while fixing x_1).

testing dataset $\{(x_{1,i}^{\text{test}}, x_{2,i}^{\text{test}}, y_i^{\text{test}})\}_{i=1}^m$. If the proposed ES algorithm accurately recovers the underlying model, the comparison points should lie close to the 45-degree diagonal line, which is indeed what we observe in Figure 2.

Additionally, we compare our proposed ADF with the general distributional forest, which does not contain an additive structure. This comparison is also shown in Figure 2. We find that both methods satisfactorily capture the conditional distribution of Y. However, under the current scenario where the data-generating process adheres to an additive structure, the ADF exhibits smaller dispersion around the diagonal. This behavior is consistent with the imposed additive structure, which constrains the estimation space and thereby reduces variability in the fitted model.

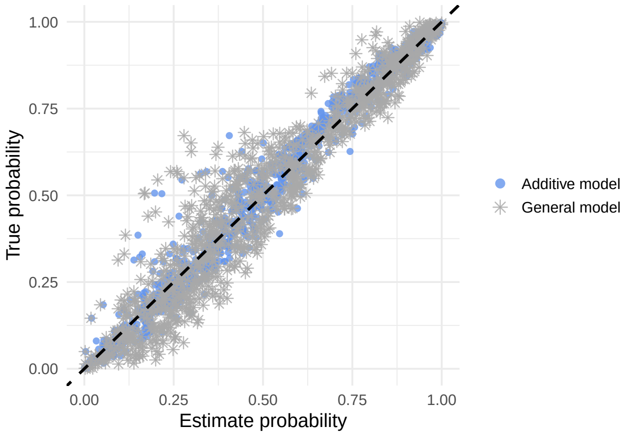

Figure 2: Comparison between the estimated and true conditional CDF values on the testing set. Each point represents an estimated probability $\widehat{F}(y_i^{\text{test}} \mid x_{1,i}^{\text{test}}, x_{2,i}^{\text{test}})$ plotted against the corresponding true value $F(y_i^{\text{test}} \mid x_{1,i}^{\text{test}}, x_{2,i}^{\text{test}})$. The blue dots correspond to the proposed ADF model, while the gray asterisks represent the general (non-additive) distributional forest.

Next, we examine how the proposed ES algorithm recovers the marginal conditional CDFs F_1 and F_2. Figure 3 displays heat maps of the estimated conditional CDFs $F_1(\cdot \mid x_1)$ and $F_2(\cdot \mid x_2)$ across different values of y and the corresponding covariate. The overlaid red curves represent the true conditional mean functions for reference. As shown, the estimated conditional CDFs closely reflect both the nonlinear relationships and the heteroskedastic variance structure imposed by the data-generating process. Unlike traditional mean regression, which can only recover the conditional expectation of Y given (x_1, x_2), the distributional forest approach simultaneously captures the full distributional behavior of Y, and how it may change as covariates vary. The estimated conditional CDF can be inverted to obtain the corresponding conditional quantile function, as illustrated in Figure 4. This transformation offers an alternative visualization of the estimated component conditional CDFs.

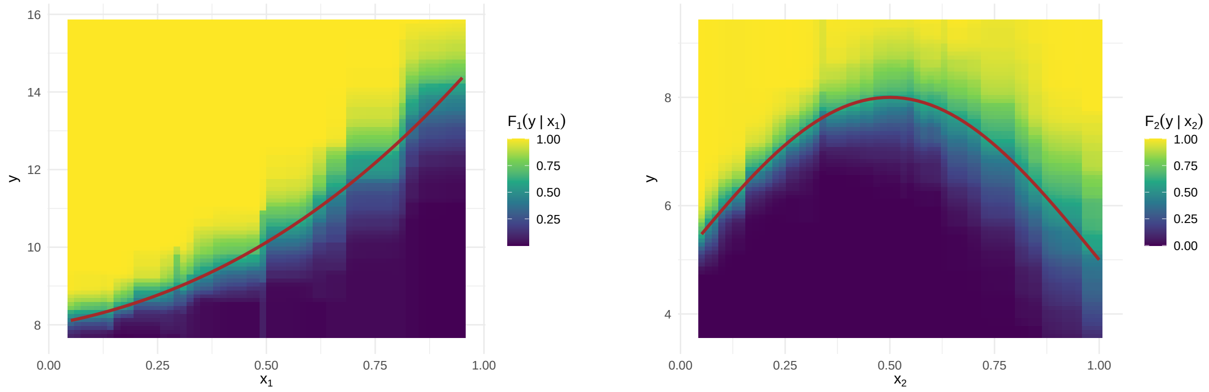

Figure 3: Heat maps of the estimated component conditional CDFs $F_1(y \,|\, x_1)$ (left) and $F_2(y \,|\, x_2)$ (right) under the proposed ADF model. The red curves overlaid on each panel represent the true conditional mean functions for reference.

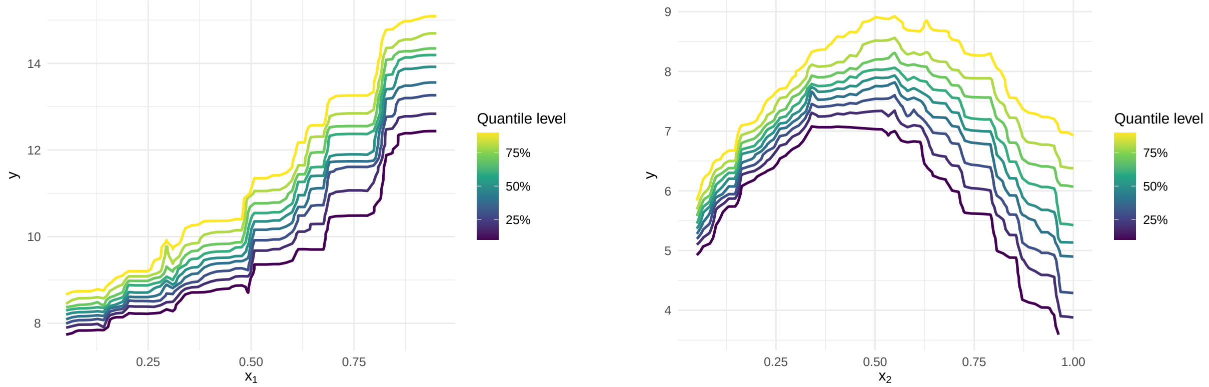

Figure 4: Estimated marginal conditional quantile functions derived from the fitted ADF model. The left panel shows conditional quantiles of Y at a certain quantile level as a function of x_1, and the right panel shows the corresponding quantiles as a function of x_2.

5.2 Robustness check

The simulation study in the previous section was conducted under an idealized setting where the data-generating process precisely adheres to the additive structure assumed by the proposed ADF model. In that case, the result demonstrates that the method is capable of accurately recovering the conditional distribution of the response variable Y.

In this subsection, we investigate how the proposed ADF performs when the true data-generating process deviates from the assumed additive structure. Specifically, we consider the following more general model:

$$F(y \mid \boldsymbol{x}) = (1-\beta) \times \left[\alpha\, F_1(y \mid x_1) + (1-\alpha)\, F_2(y \mid x_2)\right] + \beta\, F_3(y \mid x_1, x_2), \tag{9}$$

where $\beta \in [0,1]$, and $F_3(y \mid \boldsymbol{x}) = \Phi(y; \mu_3(\boldsymbol{x}), 1)$ with $\mu_3(\boldsymbol{x}) = 5\, x_1\, x_2 + 5$, $x_1, x_2 \in (0,1)$. Compared to the data distribution in Equation (8), the augmented model in Equation (9) introduces an interaction term F_3 that cannot be decomposed into additive components. The parameter β governs the extent of non-additivity. Namely, when $\beta = 0$, the model reduces to the fully additive case, and as β increases, the data-generating process becomes increasingly non-additive.

We continue to fit the proposed ADF to data simulated from this more general model, although fully aware that the assumed structure is now misspecified. To assess estimation performance, similar to the setting in the previous subsection, we simulate a training dataset of size $n = 3\,000$ and evaluate the estimation on an independently simulated test set of size $m = 1\,000$.

To account for variability in the simulation process, we repeat each experiment 100 times for fixed values of $\beta \in \{0,\ 10\%,\ 20\%,\ 30\%,\ 40\%\}$. As a benchmark, we compare the ADF with the general (non-additive) distributional forest. Performance is evaluated using two metrics: (a) the Pearson correlation and (b) the mean absolute difference (MAD) between the true conditional CDF values $F(y_i^{\text{test}} \mid x_{1,i}^{\text{test}}, x_{2,i}^{\text{test}})$ and their estimates $\widehat{F}(y_i^{\text{test}} \mid x_{1,i}^{\text{test}}, x_{2,i}^{\text{test}})$. Higher correlation and lower MAD indicate better approximation of the conditional distribution.

Figure 5 summarizes the comparison between the two methods across different levels of non-additivity. As expected, when β is small and the data-generating process is nearly additive, the ADF noticeably outperforms the general forest in both accuracy and stability. However, as β increases, the performance gap narrows. Once β exceeds 30%, the general forest begins to outperform ADF, though the two methods remain roughly comparable around this range. These findings underscore a fundamental trade-off that the additive structure introduces a modeling bias when violated, but provides gains in estimation efficiency and interpretability when valid. The performance degradation under moderate misspecification is relatively mild, suggesting that the ADF is perhaps an acceptable interpretable alternative in practical settings.

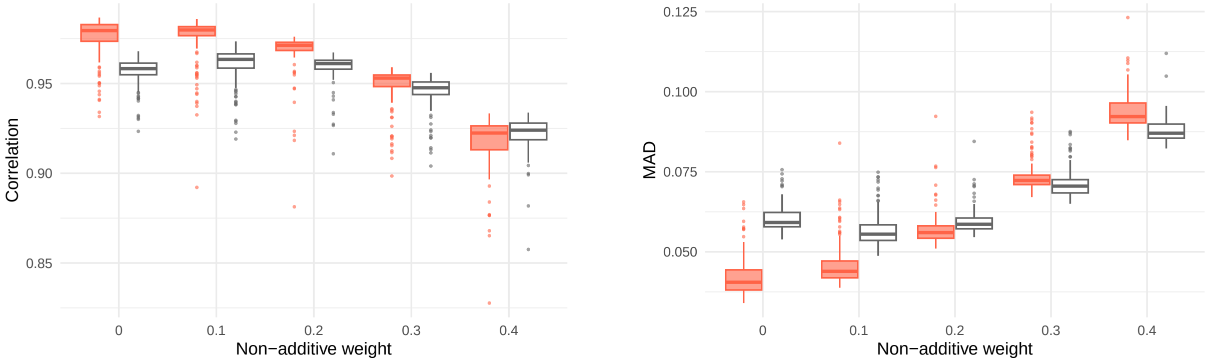

Figure 5: Comparison of estimation performance between the proposed ADF (solid, red) and general distributional forests (empty, white) across varying degrees of non-additivity in the data-generating process. The left panel shows the distribution of Pearson correlation values between the estimated and true conditional CDF's across 100 simulation replicates. The right panel displays the distribution of the corresponding MAD of the estimated and true conditional CDF's.

6 Real data illustration

We demonstrate the practical usefulness of the proposed ADF by applying it to examine the distributional relationship between the body mass index (BMI) and medical expenditure. Understanding this relationship is of particular interest to insurers, as it can provide insights for practical purposes, including interpreting predictive model results, identifying key drivers of high-cost claimants, and supporting underwriting decisions.

According to the U.S. Centers for Disease Control and Prevention (CDC), for adults aged 20 and older, a BMI below 18.5 is underweight, between 18.5–24.9 is healthy, and a BMI $\geq$ 25 is above the healthy range.[1] Naturally, the following conjecture may be proposed:

> **Conjecture C1.** Individuals whose BMI falls within the healthy weight range generally exhibit more favorable health outcomes, which in turn correspond to lower overall medical expenditures. In contrast, both underweight and overweight individuals face elevated health risks that can lead to higher healthcare costs. However, the nature of these risks differs between groups. Underweight status may arise from diverse underlying conditions, ranging from chronic illness to temporary nutritional or metabolic issues, resulting in a

[1] Source: https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

population with highly heterogeneous health profiles. As a consequence, medical expenditures among underweight individuals tend to display greater volatility compared to those in the overweight group, where the associated health burdens are more consistently linked to excess body mass and related chronic diseases.

To examine the conjecture made above, we consider healthcare outcome data extracted from the 2023 wave of the Medical Expenditure Panel Survey (MEPS).[2] MEPS is a nationally representative survey conducted by the U.S. Agency for Healthcare Research and Quality, which provides detailed information on individuals' health status, healthcare utilization, and medical expenditures. The MEPS dataset has been widely employed in the actuarial literature as a comprehensive source for illustrating the practical applications of newly developed methodologies and for conducting health- and economics-related research (Akakpo et al., 2019; Frees et al., 2013; Li et al., 2021; Xia et al., 2018).

We extract the person-level total medical expenditure from the MEPS data and use its shifted-logarithmic transformation (i.e., $\log(1 + \text{Expenditure})$ for accommodating zero expenditures) as the response variable, denoted by y in our analysis. The associated BMI is denoted by x_1. Figure 6 presents the scatter plot between x_1 and y. As shown, without controlling for other relevant variables, no clear relationship between BMI and medical expenditures can be directly observed from the scatter plot. Conjecture C1, although intuitive, is not apparent when ignoring the influence of confounding factors. The reason may be that medical spending is jointly determined by a broad set of demographic, socioeconomic, and clinical characteristics, many of which are simultaneously correlated with BMI. For instance, age and gender both exert significant influences on medical spending while also driving the distribution of BMI across the population. When these key drivers are not accounted for, their effects may mask the conditional expenditure patterns associated with different BMI levels, preventing the intuitive relationship noted in conjecture C1 from being observed in an unconditional analysis.

Next, we examine whether the relationship described in conjecture C1 may emerge after controlling for some key confounding factors. Note that conjecture C1 concerns not only potential nonlinear effects of BMI on the conditional mean of medical expenditures, but also changes in the dispersion and overall shape of the conditional distribution as BMI varies. This probabilistic perspective naturally motivates

[2] Source: https://meps.ahrq.gov/mepsweb/

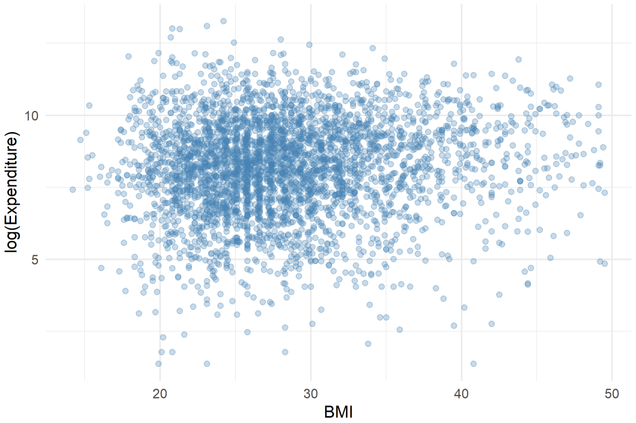

Figure 6: Scatter plot of log medical expenditure versus BMI, extracted from the MEPS dataset.

the use of distributional forests. However, a general distributional forest does not readily yield an interpretable marginal distributional relationship with respect to BMI when additional covariates are present. In this setting, the proposed ADF becomes particularly useful.

In the following analysis, we still focus on the marginal distributional relationship between medical expenditure y and BMI x_1, while incorporating age and gender, denoted by x_2 and x_3, respectively, as additional covariates. It is not our intention to claim that these are the only significant confounding factors in the relationship between medical expenditure and BMI. However, controlling for age and gender is sufficient to allow the relationship described in conjecture C1 to emerge. Therefore, we intentionally keep this illustrative example simple and include only these two additional covariates. The objective of this empirical illustration is not to optimize predictive performance, but to isolate an interpretable BMI-specific distributional component of medical expenditure while flexibly controlling for key confounding variables.

We adopt the following ADF:

$$F(y \,|\, \boldsymbol{x}) = \alpha_1 \, F_1(y \,|\, x_1) + \alpha_2 \, F_2(y \,|\, x_2) + \alpha_3 \, F_3(y \,|\, x_3), \qquad \boldsymbol{x} = (x_1, x_2, x_3), \tag{10}$$

where the additive coefficients $\alpha_1, \alpha_2, \alpha_3 \in (0, 1)$ satisfy $\alpha_1 + \alpha_2 + \alpha_3 = 1$. The ES algorithm discussed in Section 4 is used to estimate the additive weights as well as the component conditional CDFs F_1,

F_2, and F_3. Figure 7 presents the estimated conditional quantile curves associated with the BMI component $F_1(\cdot \,|\, x_1)$ in Equation (10), across a range of probability levels. In real-data applications, estimated conditional quantile functions may exhibit spiked behavior due to data noise and the inherent instability of nonparametric estimators in regions with sparse data. For improved visualization, we apply cubic smoothing splines to smooth the estimated quantile functions.

As shown in Figure 7, once age and gender are accounted for through ADF, the relationship between medical expenditure and BMI becomes considerably more intuitive and meaningful. Specifically, the central tendency of log medical expenditure exhibits a clear "V"-shaped pattern, with its minimum occurring when BMI is approximately between 20 and 25. The estimated conditional quantile curves are uniformly higher for both excessively low and excessively high BMI values. Compared with the low-BMI region, high BMI is associated with substantially higher medical expenditures across all quantile levels, especially among the higher quantiles. Moreover, the spread between quantile curves is noticeably larger in the low-BMI region. In contrast, expenditures in the high-BMI region exhibit a generally increasing trend with BMI, which suggests a more systematic relationship driven by obesity-related health conditions. We note that there is a local dip observed in the two lowest quantile curves around BMI values slightly above 40. This is likely due to data noise, as there is no clear clinical rationale for unusually low medical expenditures at that specific BMI level. Taken together, the distributional relationship between medical expenditure and BMI discussed above aligns with that conjectured in C1, thus lending support to the conjecture.

So far, in order to isolate an interpretable BMI-specific distributional component while incorporating age and gender, the model in Equation (10) imposes an additive structure on the conditional CDF $F(y \,|\, \boldsymbol{x})$. In what follows, we examine the impact of this additive structure by assessing the degree of agreement between distributional forests with and without the additive specification in estimating the conditional CDF $F(y \,|\, \boldsymbol{x})$. To evaluate out-of-sample agreement, we randomly hold out 20% of the observations from the training process and treat them as a testing dataset. Since the true conditional distribution is unknown, we treat the general non-additive distributional forest as the reference benchmark for comparison.

Figure 8 summarizes the comparison discussed above. As can be seen, most comparison points

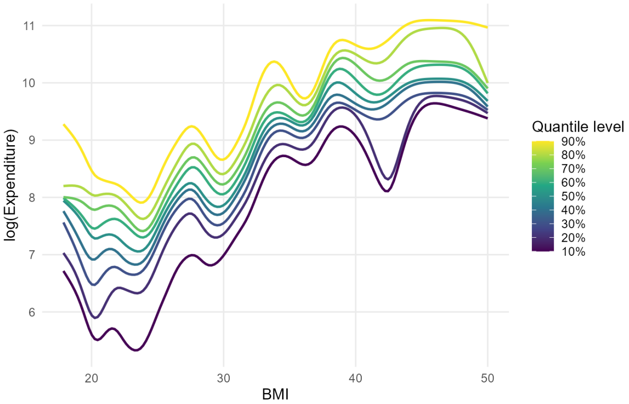

Figure 7: The estimated marginal conditional quantile functions of log medical expenditure y given BMI x_1, evaluated across a range of probability levels after accounting for age and gender through the ADF model in Equation (10).

are concentrated around the diagonal reference line. The extent of deviation of these comparison points from the reference line is comparable to that observed in the analogous comparison plot in Figure 2 from the simulation study in Section 5, where the underlying data-generating process truly admits an additive structure. This suggests that the estimated conditional CDF $F(y \mid \boldsymbol{x})$ obtained from the ADF model in Equation (10) closely tracks the estimates produced by the unconstrained general distributional forests. Therefore, we conclude that in this application, imposing the additive structure yields a substantial gain in interpretability with only a modest loss in estimation fidelity relative to the unconstrained model.

It is worth emphasizing that we do not compare the predictive performance of the proposed ADF with other predictive modeling approaches, such as generalized linear models, gradient boosting, or neural networks, because predictive accuracy is not the primary motivation for the proposed additive framework. Instead, the motivation for adopting the proposed ADF is to obtain distributional insights into the shape of the relationship between an actuarial outcome (e.g., medical expenditure) and a key covariate (e.g., BMI), while flexibly accounting for the nonlinear effects of additional confounding factors. Such insights are valuable for deepening the understanding of the presence and nature of important risk drivers and, in turn, for supporting a wide range of business decisions from an integrated quantitative and qualitative perspective. When the sole objective is to achieve an "optimal" prediction performance, users are instead advised to employ unconstrained general distributional forests or other

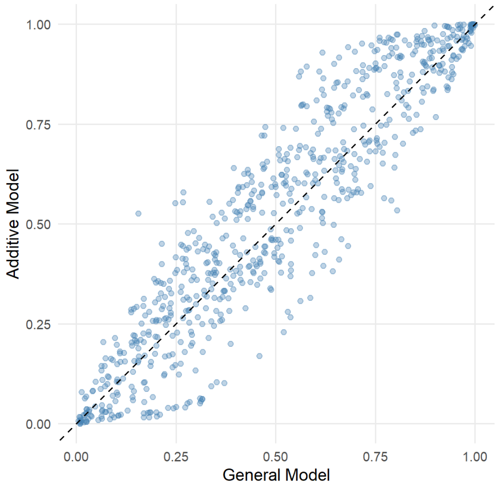

Figure 8: Comparison of the estimated conditional probabilities $\widehat{F}(y \mid \boldsymbol{x})$ obtained from the unconstrained general distributional forest and the proposed ADF model.

state-of-the-art nonparametric predictive methods.

7 Conclusions

This paper proposes an additive distributional forest (ADF) framework for modeling the conditional distribution of the response variable given covariates. Motivated by the need for interpretable distributional learning in insurance applications, ADF combines the flexibility of distributional forests with an additive structure that enables tractable explanations of marginal distributional effects of covariates. To estimate the proposed model, we develop an Expectation–Smoothing (ES) algorithm that alternates between weighted distributional-forest refits and posterior updates based on local scoring. Simulation studies demonstrate that the ES algorithm can effectively recover the additive distributional structure and provide satisfactory conditional distribution estimates under a range of data-generating scenarios. The real-data application further illustrates the practical usefulness of ADF in supporting interpretable assessment of risk drivers.

Beyond the setting considered in this paper, the proposed framework admits several natural extensions, including incorporating structured interaction terms, grouping covariates into interpretable blocks, or tailoring additive decompositions to policy- or decision-relevant risk drivers. Exploring these directions may further broaden the applicability of additive distributional forests while preserving their interpretability advantages.

References

Akakpo, R. M., Xia, M., and Polansky, A. M. (2019). Frequentist inference in insurance ratemaking models adjusting for misrepresentation. *ASTIN Bulletin: The Journal of the IAA*, 49(1):117–146.

Baione, F. and Biancalana, D. (2019). An individual risk model for premium calculation based on quantile: A comparison between generalized linear models and quantile regression. *North American Actuarial Journal*, 23(4):573–590.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Dai, S., Kuosmanen, T., and Zhou, X. (2023). Non-crossing convex quantile regression. *Economics Letters*, 233:111396.

De Jong, P. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data.* Cambridge University Press, Cambridge.

Frees, E. W. (2009). *Regression Modeling with Actuarial and Financial Applications.* Cambridge University Press, Cambridge.

Frees, E. W., Jin, X., and Lin, X. (2013). Actuarial applications of multivariate two-part regression models. *Annals of Actuarial Science*, 7(2):258–287.

Gan, G. and Valdez, E. A. (2020). Data clustering with actuarial applications. *North American Actuarial Journal*, 24(2):168–186.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall/CRC, London.

Henckaerts, R., Côté, M.-P., Antonio, K., and Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2):255–285.

Heras, A., Moreno, I., and Vilar-Zanón, J. L. (2018). An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal*, 2018(9):753–769.

Ho, T.-K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell*, 20(8):832–844.

Huang, Y. and Meng, S. (2020). A Bayesian nonparametric model and its application in insurance loss prediction. *Insurance: Mathematics and Economics*, 93:84–94.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R, 2nd Edtition.* Springer, New York.

Jamotton, C., Hainaut, D., and Hames, T. (2024). Insurance analytics with clustering techniques. *Risks*, 12(9):141.

Kneib, T., Silbersdorff, A., and Säfken, B. (2023). Rage against the mean—A review of distributional regression approaches. *Econometrics and Statistics*, 26:99–123.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, Cambridge.

Kudryavtsev, A. A. (2009). Using quantile regression for rate-making. *Insurance: Mathematics and Economics*, 45(2):296–304.

Li, H., Song, Q., and Su, J. (2021). Robust estimates of insurance misrepresentation through kernel quantile regression mixtures. *Journal of Risk and Insurance*, 88(3):625–663.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Edition.* Chapman and Hall/CRC, London.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Reed, J. (2015). Quantile regression—A new actuarial approach to claims estimation. Technical Report, Society of Actuaries.

Richardson, R. and Hartman, B. (2018). Bayesian nonparametric regression models for modeling and predicting healthcare claims. *Insurance: Mathematics and Economics*, 83:1–8.

Wüthrich, M. V. and Merz, M. (2019). Yes, we cann! *ASTIN Bulletin: The Journal of the IAA*, 49(1):1–3.

Xia, M., Hua, L., and Vadnais, G. (2018). Embedded predictive analysis of misrepresentation risk in GLM ratemaking models. *Variance*, 12(1):39–58.

Appendix A Technical proofs

Proof of Proposition 1. Suppose that the functional additive structure in (5) holds. Without loss of generality, assume that the feature space contains a set of anchor points, namely $(x_1, 0)$, $(0, x_2)$, and $(0, 0)$, for all x_1 and x_2. Let us define the following shorthand notation $\Psi_i^\circ(y) = \Psi_i(y, 0)$ for $i = 1, 2$. On the one hand, by construction, we have

$$F(y \,|\, 0, 0) = \Psi_1^\circ(y) + \Psi_2^\circ(y). \tag{11}$$

On the other hand, we know $F(y \,|\, x_1, 0) = \Psi_1(y, x_1) + \Psi_2^\circ(y)$. Collectively, the two relationships noted above yield

$$\Psi_1(y, x_1) = F(y \,|\, x_1, 0) - F(y \,|\, 0, 0) + \Psi_1^\circ(y). \tag{12}$$

By a repeated application of the same argument to Ψ_2, we can also get

$$\Psi_2(y, x_2) = F(y \,|\, 0, x_2) - F(y \,|\, 0, 0) + \Psi_2^\circ(y). \tag{13}$$

Thereby, to establish the assertion in this proposition, it is sufficient for us to construct Ψ_1° and Ψ_2° such that the following criteria are satisfied: (i) $\Psi_1^\circ(-\infty) = \Psi_2^\circ(-\infty) = 0$, and (ii) the resulting $\Psi_1(y, \cdot)$ in (12) and $\Psi_2(y, \cdot)$ in (13) are non-decreasing and right-continuous in y.

Now, suppose that the conditional CDF $F(\cdot \,|\, \boldsymbol{x})$ is absolutely continuous, and its corresponding probability density function (PDF) is given by $f(\cdot \,|\, \boldsymbol{x})$. Choose Ψ_1° and Ψ_2° such that they are differentiable and satisfy

$$\Psi_i^\circ(y) = \int_{-\infty}^{y} \frac{\mathrm{d}}{\mathrm{d}u} \Psi_i^\circ(u) \, \mathrm{d}u, \qquad i = 1, 2. \tag{14}$$

In addition, specify the derivative of Ψ_1° to be

$$\frac{\mathrm{d}}{\mathrm{d}y} \Psi_1^\circ(y) = f(y \,|\, 0, 0) - \min_{x_1} f(y \,|\, x_1, 0). \tag{15}$$

Based on the above specification of $\mathrm{d}\Psi_1^\circ(y)/\mathrm{d}y$, by (11), we have

$$\frac{\partial}{\partial y} \Psi_2^\circ(y) = f(y \,|\, 0, 0) - \frac{\partial}{\partial y} \Psi_1^\circ(y) = \min_{x_1} f(y \,|\, x_1, 0). \tag{16}$$

By such a construction, for $i = 1, 2$, we have $\mathrm{d}\Psi_i^\circ(y)/\mathrm{d}y \geq 0$, so $\Psi_i^\circ(y)$ is non-decreasing in y. Moreover, defining $\Psi_i^\circ(y)$ via (14) ensures $\Psi_i^\circ(-\infty) = 0$ for $i = 1, 2$, which ensures (i).

To verify that the above construction also ensures criterion (ii), we note that the following relationships hold:

$$\begin{aligned}
& F(y \,|\, x_1, 0) + F(y \,|\, 0, x_2) - F(y \,|\, 0, 0) \\
=& \big[\Psi_1(y, x_1) + \Psi_2(y, 0)\big] + \big[\Psi_1(y, 0) + \Psi_2(y, x_2)\big] - \big[\Psi_1(y, 0) + \Psi_2(y, 0)\big] \\
=& \Psi_1(y, x_1) + \Psi_2(y, x_2) \\
=& F(y \,|\, x_1, x_2).
\end{aligned}$$

Consequently, we have $f(y \,|\, x_1, x_2) = f(y \,|\, x_1, 0) + f(y \,|\, 0, x_2) - f(y \,|\, 0, 0) > 0$, or equivalently $f(y \,|\, x_1, 0) \geq f(y \,|\, 0, 0) - f(y \,|\, 0, x_2)$ for any pairs of x_1 and x_2 and a given y. This further implies that

$$\min_{x_1} f(y \,|\, x_1, 0) \geq f(y \,|\, 0, 0) - \min_{x_2} f(y \,|\, x_2, 0). \tag{17}$$

Recall the construction of derivatives in (15) and (16). Together with the relationship in (15), we can conclude that

$$\frac{\mathrm{d}}{\mathrm{d}y}\Psi_2^{\circ}(y) \geq f(y \,|\, 0, 0) - \min_{x_2} f(y \,|\, 0, x_2).$$

Combining (12)–(13) with the derivative bounds above, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}y}\Psi_1(y, x_1) &= f(y \,|\, x_1, 0) - f(y \,|\, 0, 0) + \frac{\mathrm{d}}{\mathrm{d}y}\Psi_1^{\circ}(y) \\ &= f(y \,|\, x_1, 0) - \min_{x_1} f(y \,|\, x_1, 0) \geq 0.\end{aligned}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}y}\Psi_2(y, x_2) &= f(y \,|\, 0, x_2) - f(y \,|\, 0, 0) + \frac{\mathrm{d}}{\mathrm{d}y}\Psi_2^{\circ}(y) \\ &\geq f(y \,|\, 0, x_2) - \min_{x_2} f(y \,|\, 0, x_2) \geq 0.\end{aligned}$$

Collectively, we obtain that both $\Psi_1(y, \cdot)$ and $\Psi_2(y, \cdot)$ are non-decreasing. We also know that $\Psi_1(y, x_1)$ and $\Psi_2(y, x_2)$ are right-continuous in y, because they are defined through differences of conditional distribution functions, as specified in (12) and (13). Together, this verifies (ii), completing the proof for the absolutely continuous case.

Next, we proceed to studying the discrete case where $F(y \,|\, \boldsymbol{x})$ is a right-continuous step function. For a given function g, define its jump at y by $\Delta g(y) = g(y) - g(y-)$. In this case, we can similarly construct discrete counterparts of Ψ_1° and Ψ_2°, analogous to those in the previous absolutely continuous case. Specifically, we can set

$$\Delta\Psi_1^{\circ}(y) = \Delta F(y \,|\, x_1, 0) - \min_{x_1} \Delta F(y \,|\, x_1, 0)$$

and

$$\Delta\Psi_2^{\circ}(y) = \min_{x_1} \Delta F(y \,|\, x_1, 0).$$

Note that the above construction can indeed yield valid $\Delta\Psi_1^\circ$ and $\Delta\Psi_2^\circ$ in the sense that they define a pair of legitimate (sub-)distributional jump measures. To see this, by construction, it is elementary to check that $\Delta\Psi_i^\circ \geq 0$, $i = 1, 2$. Further, note that $\Delta\Psi_2^\circ$ has at most countably many nonzero jump points, and its total mass satisfies

$$\sum_y \Delta\Psi_2^\circ(y) = \sum_y \min_{x_1} \Delta F(y \,|\, x_1, 0) \leq \sum_y \Delta F(y \,|\, 0, 0) = 1.$$

Thereby, $\Delta\Psi_2$ is valid, and so is $\Delta\Psi_1$ due to their inherent relationships induced by the additive decomposition of interest. Using the discrete analogue of the argument employed in the previous absolutely continuous case yields the same set of inequalities needed to guarantee criteria (i) and (ii).

If the conditional CDF $F(\cdot \,|\, \boldsymbol{x})$ has both continuous and discrete components, then one can apply the above constructions separately to the absolutely continuous part and to the jump part. This completes the proof of this proposition. □