COMMON FUNCTIONAL DECOMPOSITIONS CAN MISATTRIBUTE DIFFERENCES IN OUTCOMES BETWEEN POPULATIONS

Manuel Quintero
MIT IDSS
mquint@mit.edu

William T. Stephenson
MIT Lincoln Laboratory*
william.stephenson@ll.mit.edu

Advik Shreekumar
MIT Economics
adviks@mit.edu

Tamara Broderick
MIT EECS
tbroderick@mit.edu

ABSTRACT

In science and social science, we often wish to explain why an outcome is different in two populations. For instance, if a jobs program benefits members of one city more than another, is that due to differences in program participants (particular covariates) or the local labor markets (outcomes given covariates)? The Kitagawa-Oaxaca-Blinder (KOB) decomposition is a standard tool in econometrics that explains the difference in the mean outcome across two populations. However, the KOB decomposition assumes a linear relationship between covariates and outcomes, while the true relationship may be meaningfully nonlinear. Modern machine learning boasts a variety of nonlinear functional decompositions for the relationship between outcomes and covariates in one population. It seems natural to extend the KOB decomposition using these functional decompositions. We observe that a successful extension should not attribute the differences to covariates — or, respectively, to outcomes given covariates — if those are the same in the two populations. Unfortunately, we demonstrate that, even in simple examples, two common decompositions — functional ANOVA and Accumulated Local Effects — can attribute differences to outcomes given covariates, even when they are identical in two populations. We provide a characterization of when functional ANOVA misattributes, as well as a general property that any discrete decomposition must satisfy to avoid misattribution. We show that if the decomposition is independent of its input distribution, it does not misattribute. We further conjecture that misattribution arises in any reasonable additive decomposition that depends on the distribution of the covariates.

1 INTRODUCTION

Motivating Examples.

1. A worker at a government health department is reviewing patient mortality rates (Y) at two hospitals, H and K. He notices that the mortality rate is lower in hospital K ($\mathbb{E}_K[Y] < \mathbb{E}_H[Y]$). Is mortality lower because K receives patients who are easier to treat? Or is K more effective at providing care? If he can determine which explanation is more accurate, he may be able to better allocate training or resources across the two hospitals.
2. The mayor of City K compares the results of a job training program to a similar one in City H. She notices that program graduates in her city have lower post-program income (Y) than those in City H ($\mathbb{E}_K[Y] < \mathbb{E}_H[Y]$). Is income lower because program graduates in City K are meaningfully

*DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

different from those in City H? Or do jobs in City K tend to pay workers less than jobs in City H? If she can figure out why this difference occurs, perhaps she can modify the job training program or its recruitment strategy to make it more effective.

Many scientific questions reduce to similar comparisons between two populations. After observing differences, analysts often want to ask why these differences occur. One reason might be that the populations differ on meaningful traits. In our first example, perhaps the distribution of X is unequal: say, both hospitals provide the same standard of care, but hospital K's patient population is healthier and naturally has lower mortality rates. In this case, covariates X drive the difference. Or perhaps the patients in both hospitals are similar, but the medical staff in hospital K are particularly skilled at treating serious conditions, such as pneumonia or heart attacks. In this case, outcomes given covariates $Y \mid X$ drive the difference. A useful explanation for mean differences between populations would distinguish between these possibilities, as well as describe which aspects of the covariates or outcomes given covariates explain the difference.

The Kitagawa-Oaxaca-Blinder (KOB) decomposition (Kitagawa, 1955; Oaxaca, 1973; Blinder, 1973) is widely used in the econometrics literature to solve exactly this problem. The KOB decomposition separates a difference of means into components that depend on the distribution of covariates, X, and those that depend on the conditional expectation of outcomes given covariates, $\mathbb{E}[Y|X]$. However, it relies on parametric linear models of $\mathbb{E}[Y|X]$, which are likely inadequate to describe the complex and heterogeneous relationships that may arise in practice (Fortin et al., 2011; Bach et al., 2024). A natural generalization of the KOB decomposition would allow for non-linear or nonparametric models for $\mathbb{E}[Y|X]$. Such an approach could account for shifts in the distribution of X through a generic step-wise transformation that moves the distribution of X from population H to population K. Importance can be assigned to individual features in the change in conditional expectation $\mathbb{E}[Y|X]$ through the use of additive functional decomposition methods.

Such a generalization requires a choice of the functional decomposition. Fortunately, modern machine learning offers multiple options. For example, the functional ANOVA (FANOVA) (Stone, 1994; Huang, 1998; Hooker, 2004; 2007; Agrawal & Broderick, 2023) and Accumulated Local Effects (ALE) (Apley & Zhu, 2020) decompositions have been widely used in sensitivity analysis (Chastaing et al., 2012; Antoniadis et al., 2021), machine learning interpretability (Lengerich et al., 2020; Limmer et al., 2024), finance (Liang & Cai, 2022; Belhadi et al., 2021), and environmental and climate sciences (Huang et al., 2023; Peichl et al., 2021; Hill et al., 2023).

The success of such decompositions makes them seem like natural choices for use in explaining a difference in means. However, we demonstrate that common functional decompositions—the FANOVA and ALE—are ill-suited for this task in that they can misattribute differences stemming from a changing X to differences from changing $Y \mid X$. Throughout, we will use "misattribution" as a shorthand for differences in X attributed to $Y \mid X$. We characterize when FANOVA makes this misattribution and provide a characterization of when a general decomposition will misattribute in discrete settings (i.e., when the covariate space is countable). In particular, we show that misattribution occurs whenever the functional decomposition depends on the input covariate distribution. We conjecture and partially prove that this result holds in continuous settings as well.

The remainder of this paper is structured as follows. In Section 2, we review the Kitagawa-Oaxaca-Blinder (KOB) decomposition and discuss functional decomposition methods commonly used in machine learning, such as FANOVA and ALE. In Section 3, we define our generalized decomposition framework for difference in means, extending the KOB decomposition to non-linear models. In Section 4, we show through simple examples that FANOVA and ALE misattribute effects, and we characterize when FANOVA fails in practical cases of interest. In Section 5, we provide a general characterization of when a functional decomposition misattributes effects. Finally, in Section 6, we discuss the implications of our findings.

2 RELATED WORK AND NOTATION

2.1 NOTATION

Throughout this work, we let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the feature space of the column random vector $X = (X_1, \ldots, X_d)^T$. In general, we assume $\mathcal{X}$ is a subset of $\mathbb{R}^d$ (continuous case), except for when we explicitly state that $\mathcal{X}$ is countable (discrete case). We write $x = (x_1, \ldots, x_d)^T \in \mathcal{X}$ for a realization of X. Uppercase letters (X, X_i) thus denote random variables, while lowercase letters (x, x_i) denote specific realizations of these. We denote by $Y \in \mathbb{R}$ the outcome. When referring to distributions over the joint (X, Y), we use capital letters such as H or K. Subscripts indicate marginals or conditionals: H_X is the marginal distribution of X under H; H_i is the marginal distribution of the i-th coordinate X_i; and $H_{1:i}$ is the joint distribution of $(X_1, \ldots, X_i)$. Probability densities or mass functions are denoted by lowercase letters, such as $h(x)$ or $k(x)$. All probability measures are defined on the Borel σ-algebra of $\mathcal{X}$ in the continuous case or on the power set of $\mathcal{X}$ in the discrete case.

2.2 KITAGAWA-OAXACA-BLINDER DECOMPOSITION

The Kitagawa-Oaxaca-Blinder (KOB) decomposition provides a framework for explaining differences in means between two populations by decomposing them into components attributable to differences in covariates and conditional expectations. In its original formulation, KOB assumes the covariate space is $\mathcal{X} = \mathbb{R}^d$ and that the covariates X have a linear relationship with the outcome $Y \in \mathbb{R}$:

$$\mathbb{E}_{H_{Y|X}}[Y \mid X] = X^T \beta_H \quad \text{and} \quad \mathbb{E}_{K_{Y|X}}[Y \mid X] = X^T \beta_K,$$

where $H_{Y|X}$ is the conditional distribution of $Y \mid X$ in population H, and similarly for population K. The vectors $\beta_H, \beta_K \in \mathbb{R}^d$ are the regression coefficients defining the linear relationship $\mathbb{E}[Y \mid X]$ for each population. The KOB decomposes the difference $\mathbb{E}_K[Y] - \mathbb{E}_H[Y]$ as

$$\underbrace{\mathbb{E}_{K_X}[\mathbb{E}_{K_{Y|X}}[Y \mid X] - \mathbb{E}_{H_{Y|X}}[Y \mid X]]}_{Y \mid X \text{ effect}} + \underbrace{\mathbb{E}_{K_X}[\mathbb{E}_{H_{Y|X}}[Y \mid X]] - \mathbb{E}_{H_X}[\mathbb{E}_{H_{Y|X}}[Y \mid X]]}_{\text{Covariate effect}} \tag{1}$$

$$= \sum_{j=1}^{d} \underbrace{\mathbb{E}_{K_X}[X_j](\beta_j^K - \beta_j^H)}_{Y \mid X \text{ effect for the } j\text{th covariate}} + \sum_{j=1}^{d} \underbrace{(\mathbb{E}_{K_X}[X_j] - \mathbb{E}_{H_X}[X_j])\beta_j^H}_{\text{Covariate effect for the } j\text{th covariate}}. \tag{2}$$

In the next section, we introduce a natural extension of the KOB decomposition that retains a similar interpretation, but allows for more general forms of $Y \mid X$ by using functional decompositions. The existing literature provides several such functional decompositions; however, we focus only on additive decompositions that decompose functions into additive components, as they provide a natural analogue of the additive form of $Y \mid X$ in the KOB decomposition. Other functional decomposition methods, such as Partial Dependence Plots (Friedman, 2001), do not offer additive decompositions and thus cannot be immediately incorporated into KOB-like decompositions. We next review two particularly common additive functional decompositions: FANOVA and ALE.

2.3 FANOVA

FANOVA measures the importance of features in determining the output of a function and in identifying underlying additive interactions between subsets of variables (Hooker, 2004). It provides a natural representation of a functional in terms of low-order components (Hooker, 2007) by stating that a square-integrable function $f(x)$ with $x \in \mathcal{X} = \mathbb{R}^d$ can be written uniquely as $f(x) = \sum_{S \in 2^{[d]}} \mathcal{L}(f, K_X, S)(x)$, where $2^{[d]}$ denotes the

power set of $[d] = \{1, 2, \ldots, d\}$ and K_X is a general measure of the covariates. Then, the components are jointly defined as

$$\{\mathcal{L}(f, K_X, S)(x) \mid S \in 2^{[d]}\} = \underset{\{\mathcal{L}(f,K_X,S)\in L^2(\mathbb{R}^d)\}_{S\in 2^{[d]}}}{\arg\min} \int \left(\sum_{s\in 2^{[d]}} \mathcal{L}(f, K_X, S)(x) - f(x) \right)^2 k(x)dx, \tag{3}$$

subject to *hierarchical orthogonality conditions* among the components:

$$\forall S \in 2^{[d]}, \forall V \subsetneq S : \int \mathcal{L}(f, K_X, S)(x)\mathcal{L}(f, K_X, V)(x)k(x)dx = 0, \tag{4}$$

where $\subsetneq$ denotes a proper subset.

Note that the FANOVA component corresponding to a subset S depends only on the covariates in S, as it is constructed to capture their contribution separately from the rest. However, for the sake of generality in defining a functional decomposition, we express it as a function of the full covariate vector. The same applies to the ALE decomposition below.

2.4 Accumulated Local Effects (ALE)

ALE is another additive functional decomposition method that is particularly suitable for visualizing the effects of predictors (Apley & Zhu, 2020). Although ALE is defined more generally—allowing for non-differentiable f and extending to dimensions $d > 2$—the case for $d = 2$ with a differentiable model, $f(x_1, x_2) = \mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2]$, suffices for our illustrative purposes. The ALE main effect component for the first covariate X_1 is then defined as:

$$\mathcal{L}(f, K_X, \{1\})(x) = \int_{x_{\min,1}}^{x_1} \mathbb{E}_{K_2}\left[\frac{\partial f(z_1, X_2)}{\partial z_1}\right] dz_1 - \text{constant}, \tag{5}$$

where $\frac{\partial f(x_1,x_2)}{\partial x_1}$ denotes the partial derivative of f with respect to the first component, $x_{\min,1}$ is a lower bound of the support of K_1, and *constant* is a centering constant such that $\mathbb{E}_{K_X}[\mathcal{L}(f, K_X, \{1\})(X)] = 0$. The term $\mathcal{L}(f, K_X, \{2\})$ is defined similarly; for the definition of $\mathcal{L}(f, K_X, \{1, 2\})$ and for the $d > 2$ case, see (Apley & Zhu, 2020).

3 Additive decompositions of population differences

As discussed in Section 1, a desirable extension of KOB would allow for arbitrary flexible regression models by extending it to non-linear functional forms. Recall the KOB decomposition in Equation 2 separates a difference in means into a $Y \mid X$ effect and a covariate effect. To extend the KOB decomposition to more flexible models, we assume a general regression model $f : \mathcal{X} \to \mathbb{R}$ is fitted such that $f^K(x) = \mathbb{E}_{K_{Y|X}}[Y \mid X = x]$, and similarly for population H.[1] Our goal is to decompose Equation 1 into smaller, interpretable components just as in the KOB decomposition. To achieve this goal and in the spirit of FANOVA and ALE discussed in Section 2, we assume a generic additive functional decomposition, denoted by $\mathcal{L}$, which operates on arbitrary functions f of the covariates, probability measures of the covariates H_X, and subsets of features S. We assume this decomposition yields an additive representation that holds for all $x \in \mathcal{X} \subseteq \mathbb{R}^d$:

$$f(x) = \sum_{S\in 2^{[d]}} \mathcal{L}(f, H_X, S)(x). \tag{6}$$

[1]We write an equality $f^K(x) = \mathbb{E}_{K_{Y|X}}[Y \mid X = x]$ for the purpose of exposition. In practice, the fitted $f^K(x)$ will contain approximation error, and our results apply to decompositions of $f^K(x)$ rather than the exact expectation $\mathbb{E}_{K_{Y|X}}[Y \mid X = x]$.

Given such an additive functional decomposition, it is straightforward to extend the KOB decomposition. We define two types of swaps, analogous to the terms in the KOB decomposition. First, we can swap out the distribution of each one-dimensional covariate of X at a time; we call such terms *the difference due to change in* X. Second, we can swap out the functional decomposition terms of f^H for those of f^K; we call such terms *the difference due to change in* $Y \mid X$. We define this KOB extension below:

Definition 1. *Let $\mathcal{S}$ define an ordering of all subsets $S \in 2^{[d]}$; we refer to the ith subset in this ordering as $\mathcal{S}_i$. We define the importance decomposition to be:*

$$\mathbb{E}_K[Y] - \mathbb{E}_H[Y] = \sum_{i=1}^{|\mathcal{S}|} \delta_{S_i}^{Y|X} + \sum_{j=1}^{d} \delta_j^X, \tag{7}$$

$$\text{where: } \delta_{S_i}^{Y|X} := \mathbb{E}_{H_X}\left[\mathcal{L}(f^K, K_X, S_i)(X)\right] - \mathbb{E}_{H_X}\left[\mathcal{L}(f^H, H_X, S_i)(X)\right],$$
$$\delta_j^X := \mathbb{E}_{K_{1:j|j+1:d}H_{j+1:d}}\left[f^K(X)\right] - \mathbb{E}_{K_{1:j-1|j:d}H_{j:d}}\left[f^K(X)\right].$$

Note that $\delta_{S_i}^{Y|X}$ holds the covariate distribution fixed at H_X, and changes whether the S_i term of $\mathbb{E}[Y \mid X]$ comes from H or K. We therefore call $\delta_{S_i}^{Y|X}$ the difference due to the dependence of $Y \mid X$ on feature subset S_i. Likewise, the distributions over covariates in δ_j^X differ in whether X_j follows a distribution determined by H or K. We therefore call δ_j^X the difference due to the change in distribution of covariate j.

Definition 1 is an extension of the KOB decomposition from Section 2, which also defines differences from swapping out distributions of covariates, as well as differences in swapping out (a model for) $Y \mid X$. The main difference is that Definition 1 uses a generic additive decomposition of $Y \mid X$, whereas the KOB decomposition assumes a linear model.

This decomposition—like the KOB decomposition—makes a series of specific choices: first swapping S_1, then S_2, ... then finally swapping $S_{|\mathcal{S}|}$, and then swapping covariate one, then covariate two, etc. Why not swap S_2 first? Why not swap covariate three immediately after S_1? In general, there is no reason to prefer any one ordering, and different orderings will produce different results. With no preferred ordering of swaps, one may prefer to average over all possible orderings and report the resulting averages as the definitions of $\delta_{S_i}^{Y|X}$ and δ_j^X.[2] Our results here apply to any fixed order; we leave the extension to averaging over all orderings as future work.

4 Failure of existing functional decompositions

Once a user has specified the functional forms of $f^H(x)$ and $f^K(x)$, the only decision to be made before using Definition 1 is the choice of functional decomposition $\mathcal{L}$. At first glance, options such as ALE or FANOVA from Section 2.3 and Section 2.4 seem like excellent choices: they provide additive decompositions of generic functions with properties that make them well-suited for understanding functions in other applications. However, we show that a broad class of functional decompositions, including FANOVA and ALE, are inappropriate for explaining population differences in the sense of Definition 1, despite their great success in other applications. In particular, we characterize when such decompositions incorrectly state that differences stem from changes in $Y \mid X$.

Recall that Definition 1 defines $\delta_{S_i}^{Y|X}$ to be the *difference due to the dependence of $Y \mid X$ on feature subset S_i*. Suppose that the distributions of $Y \mid X$ are in fact identical across the two populations: $f^K = f^H = f$. In

[2] Shorrocks (2013) describes such averages as applying logic of Shapley values to functional decompositions.

this situation, any reasonable decomposition should lead us to believe there is no difference due to differences in $Y \mid X$; that is, $\delta_{S_i}^{Y|X} = 0$. Unfortunately, we present examples where FANOVA and ALE can misattribute differences in X to differences in $Y \mid X$.

4.1 Examples where FANOVA and ALE misattribute

To begin with, we formalize what we mean by misattribution. Note that when $f^K = f^H = f$, $\delta_S^{Y|X}$ reduces to

$$\Delta(\mathcal{L}, f, H_X, K_X, S) := \delta_S^{Y|X} = \mathbb{E}_{H_X}\left[\mathcal{L}(f, K_X, S)(X) - \mathcal{L}(f, H_X, S)(X)\right]. \tag{8}$$

With eq. (8) in mind, we define misattribution as follows:

Definition 2. *We say that a functional decomposition $\mathcal{L}$ misattributes effects of $Y \mid X$ if $\Delta(\mathcal{L}, f, H_X, K_X, S) \neq 0$ for any f, H_X, K_X, S.*

In the following examples, we show that FANOVA and ALE misattribute effects of $Y \mid X$ in simple scenarios; we present here a summary of the examples, and the full details are provided in Appendix A.

Example 1 (FANOVA)**.** *Consider the case where the fitted model is additive and has two covariates: $f^K = f^H = f(x) = x_1 + x_2$. Suppose that population H has covariates X_1, X_2, with $\mathbb{E}_{H_X}[X_1] = \mathbb{E}_{H_X}[X_2] = 0$ while in population K, $\mathbb{E}_{K_X}[X_1] = \mu$ and $\mathbb{E}_{K_X}[X_2] = 0$ for $\mu \neq 0$. In both populations, X_1 and X_2 are independent and have finite variance.*

Then, following the FANOVA decomposition in Equation 3, the components for each subset satisfy the following for each population:

$$\mathcal{L}(f, H_X, \emptyset)(x) = 0, \quad \mathcal{L}(f, H_X, \{1\})(x) = x_1, \quad \mathcal{L}(f, H_X, \{2\})(x) = x_2, \quad \mathcal{L}(f, H_X, \{1,2\})(x) = 0,$$

$$\mathcal{L}(f, K_X, \emptyset)(x) = \mu, \quad \mathcal{L}(f, K_X, \{1\})(x) = x_1 - \mu, \quad \mathcal{L}(f, K_X, \{2\})(x) = x_2, \quad \mathcal{L}(f, K_X, \{1,2\})(x) = 0.$$

Hence, the difference in means due to differences in $Y \mid X$ for the component of $S = \{1\}$ is given by:

$$\Delta(\mathcal{L}, f, H_X, K_X, \{1\}) = \mathbb{E}_{H_X}\left[\mathcal{L}(f, K_X, \{1\})(X) - \mathcal{L}(f, H_X, \{1\})(X)\right] = \mathbb{E}_{H_X}[X_1 - \mu - X_1] = -\mu \neq 0.$$

Since this term is not equal to zero, FANOVA misattributes effects to $Y \mid X$ in this example.

Similarly, the following example demonstrates that ALE misattributes in a simple scenario.

Example 2 (ALE)**.** *Let $f^K = f^H = f(x) = x_1 x_2$, and assume for population H, $X_1 \sim N(1,1)$, $X_2 \sim N(0,1)$, and for population K, $X_1 \sim N(0,1)$, $X_2 \sim N(\mu, 1)$, where $\mu \neq 0$. Assume we observed data $\{(x_{i,1}^j, x_{i,2}^j)\}_{i=1}^n$, with n sufficiently large, for $j = H, K$. Following the ALE decomposition in Equation 5, we can compute the centered components for each population:*

$$\mathcal{L}(f, H_X, \{1\})(x) = 0, \quad \mathcal{L}(f, H_X, \{2\})(x) = (x_2 - x_{\min,2}^H) - (-x_{\min,2}^H) = x_2,$$

$$\mathcal{L}(f, K_X, \{1\})(x) = \mu(x_1 - x_{\min,1}^K) - (-\mu x_{\min,1}^K) = \mu x_1, \quad \mathcal{L}(f, K_X, \{2\})(x) = 0.$$

Thus, the difference in means due to differences in $Y \mid X$ for $S = \{1\}$, is given by:

$$\begin{aligned}\Delta(\mathcal{L}, f, H_X, K_X, \{1\}) &= \mathbb{E}_{H_X}[\mathcal{L}(f, K_X, \{1\})(X) - \mathcal{L}(f, H_X, \{1\})(X)] = \mathbb{E}_{H_X}[\mu X_1 - 0] \\ &= \mu \mathbb{E}_{H_X}[X_1] = \mu \cdot 1 = \mu \neq 0.\end{aligned}$$

Since this term is not equal to zero, ALE misattributes effects to $Y \mid X$ in this example.

Examples 1 and 2 show that both ALE and FANOVA can misattribute differences in the covariates to differences in $Y \mid X$. With the knowledge that common functional decompositions like FANOVA or ALE *can* misattribute effects, it behooves us to understand how widespread this behavior is. Does misattribution happen frequently for common functional decompositions? And what properties of functional decompositions will prevent misattribution? Practitioners need answers to these questions to confidently use methods similar to the one described in Definition 1. In the next sections, we provide partial answers to these questions. In Section 4.2, we argue that FANOVA misattributes in the sense of Definition 1 in almost all cases of practical interest, rendering it unsuitable in practice. And in Section 5, we conjecture—and partially prove—that any reasonable functional decomposition that depends on the input covariate distribution will misattribute.

4.2 When does FANOVA misattribute effects?

In the last section, we presented an example in which FANOVA misattributes differences in X to differences in $Y \mid X$. However, without a precise characterization of when this misattribution occurs, one might think it is specific to the example rather than a general phenomenon. We now provide theoretical characterizations of when FANOVA misattributes effects in cases of practical interest. Specifically, we first show that FANOVA does not misattribute when the lower-dimensional components computed for population K have a mean of zero under population H. We then demonstrate that this condition is highly unrealistic, even in simple cases such as affine functions, and becomes even more restrictive when we allow for more flexibility.

Theorem 1 (FANOVA attribution)**.** *Let $\mathcal{L}$ denote the FANOVA decomposition. Then, for any Lebesgue measurable function f, any pair of probability measures H_X and K_X, and any subset of the covariates $S \in 2^{[d]} \setminus \{\emptyset\}$, we have*

$$\Delta(\mathcal{L}, f, H_X, K_X, S) = 0$$

if and only if

$$\mathbb{E}_{H_X}\Big[\mathcal{L}(f, K_X, S)(X)\Big] = 0. \tag{9}$$

Proof. By definition, we have $\Delta(\mathcal{L}, f, H_X, K_X, S) = \mathbb{E}_{H_X}[\mathcal{L}(f, K_X, S)(X) - \mathcal{L}(f, H_X, S)(X)]$. The mean-zero property of the FANOVA components (see Appendix Lemma 1) implies that $\mathbb{E}_{H_X}[\mathcal{L}(f, H_X, S)(X)] = 0$. Thus, $\Delta(\mathcal{L}, f, H_X, K_X, S) = \mathbb{E}_{H_X}[\mathcal{L}(f, K_X, S)(X)]$. It follows immediately that $\Delta(\mathcal{L}, f, H_X, K_X, S) = 0$ if and only if $\mathbb{E}_{H_X}[\mathcal{L}(f, K_X, S)(X)] = 0$. ■

Theorem 1 states that the expectation under population H of the component computed under population K must have a mean of zero. This suggests a close relationship between the two distributions or that the moments of the marginal distribution must satisfy specific conditions for Equation 9 to hold. If these conditions hold in practical scenarios, then FANOVA could indeed be a viable option. Our next set of results indicates that Equation 9 unfortunately cannot be reasonably expected to hold in practice.

We start by studying a particularly simple case—when f is a given affine function. We show that even in this case, the conditions under which FANOVA does not misattribute are very restrictive.

Theorem 2 (FANOVA affine class)**.** *Let $X_1, \ldots, X_M$ be independent random variables, and let H_X and K_X be two probability measures. Consider a function $f : \mathbb{R}^M \to \mathbb{R}$ given by $f(x) = \sum_{m=1}^{M} a_m\, b_m(x_m)$, where each coefficient $a_m \in \mathbb{R} \setminus \{0\}$ and each basis function $b_m : \mathbb{R} \to \mathbb{R}$ is measurable for $m = 1, \ldots, M$. Then, we have*

$$\textit{for all } S \in 2^{[M]} \setminus \{\emptyset\}, \quad \mathbb{E}_{H_X}\Big[\mathcal{L}(f, K_X, S)(X)\Big] = 0, \tag{10}$$

if and only if

$$\textit{for all } m \in \{1, \ldots, M\}, \quad \mathbb{E}_{H_X}\Big[b_m(X_m)\Big] = \mathbb{E}_{K_X}\Big[b_m(X_m)\Big]. \tag{11}$$

See Appendix B for the proof. Equation 11 is a fairly restrictive condition, as we expect covariate means to vary between populations; e.g., the proportion of the workforce with high school diplomas will likely not be *identical* between two cities H and K.

We can now see why FANOVA misattributes effects in Example 1. There, the function $f(x) = x_1 + x_2$ corresponds to the affine class in Theorem 2, with coefficients $a_1 = a_2 = 1$ and basis functions $b_1(x_1) = x_1$, $b_2(x_2) = x_2$. Since the expectation condition in Equation 11 does not hold ($\mathbb{E}_{H_X}[X_1] \neq \mathbb{E}_{K_X}[X_1]$), FANOVA assigns a nonzero difference to $S = \{1\}$, leading to misattribution.

A more revealing version of Theorem 2 would extend to richer basis representations by using multiple basis functions per dimension rather than a single one and allowing for correlated covariates. However, we conjecture that this would further restrict the relationship between the moments of the distributions H_X and K_X, imposing increasingly stringent requirements on them. In other words, adding flexibility to our model of $\mathbb{E}[Y \mid X]$ comes at the cost of restricting the set of populations where the decomposition remains accurate. This culminates in our next result, which shows that placing minimal restrictions on $\mathbb{E}[Y \mid X]$ imposes maximal restrictions on the distribution of X.

Theorem 3 (FANOVA unrestricted)**.** *Let $\mathcal{M}(\mathcal{X})$ denote the set of measurable functions on the covariate space. Suppose that H_X and K_X are probability measures such that H_X is absolutely continuous with respect to K_X ($H_X \ll K_X$). Then, we have*

$$\text{for all } f \in \mathcal{M}(\mathcal{X}) \text{ and for all } S \in 2^{[d]} \setminus \{\emptyset\}, \quad \mathbb{E}_{H_X}\Big[\mathcal{L}(f, K_X, S)(X)\Big] = 0,$$

if and only if

$$H_X = K_X, \quad K_X\text{-almost surely.}$$

See Appendix B.3 for a proof.

Theorem 3 says that if we want FANOVA to never misattribute for a given pair of distributions—that is, not misattribute for every measurable function f and every nonempty subset S of the covariates—then it is necessary and sufficient that the input covariate distributions are identical (i.e., $H_X = K_X$, up to a K_X-null set). Equivalently, if $H_X \neq K_X$, then there exists at least one problematic measurable function f and nonempty subset S for which FANOVA misattributes to $Y \mid X$. In practice, we generally compare distinct populations (i.e., $H_X \neq K_X$), implying that FANOVA will misattribute in settings where f is one of the problematic cases. Theorem 3 does not characterize the problematic f, suggesting that knowledge or assumptions about f could rule out misattribution in some applications. A more practical result would characterize the set of problematic f for a particular set of input densities; we leave this as a direction for future work. A more concerning result would instead give conditions under which misattribution can occur for any given f; we also leave this as a direction for future work.

5 When do functional decompositions misattribute effects?

Given the pessimistic results in Section 4, one may be reasonably concerned that *any* functional decomposition $\mathcal{L}$ may misattribute, making the decomposition of Definition 1 of little value, as practitioners would never know when to trust its outputs. To resolve this problem, we attempt to characterize what properties of the functional decomposition $\mathcal{L}$ cause misattribution. We show that under regularity conditions, a functional decomposition $\mathcal{L}$ does not misattribute the effects of $Y \mid X$ if and only if its derivative with respect to the probability measure is orthogonal to K in the L^2 sense. Furthermore, we prove that Definition 1 does not lead to misattribution if $\mathcal{L}$ is independent of its input distribution. For a reverse direction statement, we conjecture that under reasonable assumptions on $\mathcal{L}(f, K_X, S)$, the function f, and the distributions H_X and K_X, a functional decomposition $\mathcal{L}$ does not misattribute the effects of $Y \mid X$ if it does not depend on its input distribution.

We now aim to characterize conditions on the functional $\mathcal{L}$ under which $\Delta(\mathcal{L}, f, H_X, K_X, S)$ does or does not equal zero for all f, H_X, K_X. We start by studying the discrete case and leave the continuous generalization as a conjecture.

5.1 THE DISCRETE SETTING

First, suppose that $\mathcal{X}$ is a countable space so that the covariates of the random vector $X = (X_1, \ldots, X_d)^T$ are discrete. Let $k(x)$ and $h(x)$ be the probability mass functions corresponding to the probability measures K_X and H_X, respectively, with shared support on $\mathcal{X}$. For example, in healthcare, $\mathcal{X}$ might represent patient categories based on age or pre-existing conditions, while $f(x)$ could denote the expected recovery time, and $k(x)$, $h(x)$ represent the proportions of patients in different hospitals.

Before stating our main theorem of this section, we impose the following regularity conditions on the class of functional decompositions we consider.

Assumption 1. *The map $K_X \mapsto \mathcal{L}(f, K_X, S)$ is twice continuously differentiable with respect to K_X, and its second derivative is uniformly bounded. For a mathematical formulation see Appendix Assumption 1.*

For any fixed measurable function f and subset $S \in 2^{[d]}$, we denote by $\mathcal{J}_{K_X}(f, H_X, S)$ the Jacobian matrix of the mapping $K_X \mapsto \mathcal{L}(f, K_X, S)$, with respect to K_X, evaluated at $K_X = H_X$.

We now state our main result for the discrete case, which characterizes when a functional decomposition $\mathcal{L}$ will never misattribute the effects of $Y \mid X$ to changes in X.

Theorem 4 (Discrete characterization)**.** *Under Assumption 1 and for all $H_X, K_X \in \Sigma^{\circ}$, we have*

$$\Delta(\mathcal{L}, f, H_X, K_X, S) := \delta_S^{Y|X} := \mathbb{E}_{H_X}\Big[\mathcal{L}(f, K_X, S)(X) - \mathcal{L}(f, H_X, S)(X)\Big] = 0,$$

if and only if

$$\textit{for all } K_X \in \Sigma^{\circ}, \quad \mathcal{J}_{K_X}\mathcal{L}(f, K_X, S) = \mathbf{c}\,\mathbf{1}^T, \quad \textit{for some } \mathbf{c} \in \mathbb{R}^d.$$

Remark: *The condition $\mathcal{J}_{K_X}\mathcal{L}(f, K_X, S) = \mathbf{c}\,\mathbf{1}^T$ implies that all columns of the Jacobian are identical, so that its rank is* 1.

See Appendix C.1 for the proof.

Theorem 4 shows that if we require a functional decomposition $\mathcal{L}$ to never to misattribute the effect of $Y \mid X$ for *any* distribution K_X, its dependence on K_X becomes severely restricted. Concretely, if the *average* change of $\mathcal{L}$ is zero for every pair of distributions $H_X, K_X \in \Sigma^{\circ}$, then all the columns of the Jacobian of $\mathcal{L}$ with respect to K_X must be the same. This structure means that $\mathcal{L}$ cannot distinguish between different probability distributions, which implies that Δ must be zero. As the following corollary shows, this characterization implies $\mathcal{L}$ must be constant in its second argument across values in Σ°.

Corollary 1. *Under the same assumptions as Theorem 4, $\mathcal{L}$ will not misattribute the effects of $Y \mid X$ if and only if $\mathcal{L}(f, K_X, S) = \mathcal{L}(f, H_X, S)$ for all $K_X, H_X \in \Sigma^{\circ}$, i.e., the functional $\mathcal{L}$ is constant with respect to the distribution over covariates.*

Proof. From Theorem 4, there will be no misattribution if and only if $\mathcal{J}_{K_X}(f, H_X, S) = \mathbf{c}\mathbf{1}^T$ for some $\mathbf{c} \in \mathbb{R}^d$. From the Mean Value Theorem, $\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) = \mathcal{J}_{K_X}\mathcal{L}(f, \tilde{H}_X, S)(K_X - H_X)$, and since for all $K_X \in \Sigma^{\circ}$, $\mathcal{J}_{K_X}(f, K_X, S) = \mathbf{c}\mathbf{1}^T$, we have $\mathbf{c}\mathbf{1}^T(K_X - H_X) = 0$, implying $\mathcal{L}(f, K_X, S) = \mathcal{L}(f, H_X, S)$. ■

That is, for a decomposition not to misattribute, $\mathcal{L}$ must be constant in Σ°, meaning it is completely unresponsive to changes in K_X. We note that this may be unnecessarily restrictive in practice. In particular,

in most applications, we are not concerned with *every* possible redistribution of probability mass but rather with specific structured changes that carry meaningful information. Still, as the following example shows, there is at least one existing decomposition that satisfies the conditions of Corollary 1.

Example 3 (Non-generalized FANOVA (Hooker, 2004))**.** *The non-generalized FANOVA decomposes $f(x)$ using a uniform distribution U over the unit hypercube, independent of H_X or K_X. When $f^K = f^H = f$, the decomposition remains constant over Σ^{o}, i.e., $\mathcal{L}(f, K_X, S) = \mathcal{L}(f, U, S) = \mathcal{L}(f, H_X, S)$. Consequently, its Jacobian is zero for all $K_X \in \Sigma^{\text{o}}$, trivially satisfying $\mathcal{J}_{K_X}\mathcal{L}(f, K_X, S) = \mathbf{c}\,\mathbf{1}^T$.*

In contrast, we conjecture that *generalized* FANOVA does satisfy the conditions of Corollary 1.

Conjecture 1. *Suppose f is non-constant, and let $\mathcal{L}(f, K_X, S)$ be the FANOVA decomposition. Then, $\mathcal{L}$ satisfies Assumption 1, and there exist $H_X, K_X \in \Sigma^{\text{o}}$ such that it misattributes effects of $Y \mid X$.*

Put together, Example 3 and Corollary 1 tell a story that is exactly the opposite of typical observations in the functional decomposition literature. In particular, that generalized FANOVA and ALE depend on their input distributions is tyipcally viewed as beneficial; indeed, this is a major motivator for the work of Apley & Zhu (2020) and Hooker (2007). One reason for this benefit is that typical functions f of interest are often machine learning models fit to training data drawn from covariate distribution K_X. Many flexible machine learning models have arbitrary behavior outside the support of the training data; thus methods like the non-generalized FANOVA that integrate with respect to the uniform distribution may integrate over nonsensical values of f. ALE and FANOVA resolve this issue by integrating over the covariate distribution K_X.

These results highlight a tension in the design of functional decomposition methods: non-use of the covariate distribution K_X may result in strange behavior by integrating over nonsensical values of f, while use of the covariate distribution may result in nonsensical decompositions of differences between two populations. We leave as future work an attempt to resolve this seeming contradiction.

In practice, many applications involve continuous distributions, where densities vary smoothly rather than being restricted to discrete points. For example, in economic models, income distributions are continuous, and in healthcare, biomarkers like blood pressure or cholesterol levels are measured on a continuous scale. To extend our characterization to these cases, we analyze the continuous setting in Appendix C.2.

6 Conclusion

In this work, we present a natural extension of the Kitagawa-Oaxaca-Blinder decomposition for explaining differences in means to non-linear models of the conditional expectation. We note that functional decompositions like FANOVA and ALE seem at first glance like excellent candidates to incorporate into our KOB extension. However, we provide simple counterexamples showing that both FANOVA and ALE can incorrectly assign differences in the distribution of covariates X to differences in the outcome-given-covariates, $Y \mid X$. We further provide characterizations of when FANOVA misattributes for practical cases of interest, as well as a general property that any discrete decomposition should satisfy to never misattribution: the decomposition must be constant across all distributions. For the general continuous case, we show that if the decomposition is independent of its input distribution, it does not misattribute. For a reverse statement, we conjecture that the same will hold as in the discrete case: any reasonable functional decomposition method that depends on its input distribution in a meaningful way will misattribute.

Our findings highlight a fundamental limitation: methods effective for single-population analysis may be unreliable for comparing differences between populations. Our work also suggests that the requirements for single-population decomposition and two-population difference decomposition may diverge, highlighting the importance of developing new methods to decompose the difference in means. Future work should explore how to develop decompositions that avoid misattribution while preserving interpretability in real-world applications.

ACKNOWLEDGMENTS

The authors are grateful to Raj Agrawal for his contributions and insights during the initial stages of this project. This research was supported in part by an ONR Early Career Grant. Advik Shreekumar was supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Combatant Commands under Air Force Contract No. FA8702-15-D-0001 or FA8702-25-D-B001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Combatant Commands.

REFERENCES

Raj Agrawal and Tamara Broderick. The skim-fa kernel: high-dimensional variable selection and nonlinear interaction discovery in linear time. *Journal of Machine Learning Research*, 24(27):1–60, 2023.

Anestis Antoniadis, Sophie Lambert-Lacroix, and Jean-Michel Poggi. Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, 206:107312, 2021. doi: 10.1016/j.ress.2020.107312. URL https://www.sciencedirect.com/science/article/pii/S0951832020308073.

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4): 1059–1086, September 2020. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12377.

V. I. Averbukh and O. G. Smolyanov. The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys*, 22:201–258, 1967. doi: 10.1070/RM1967v022n02ABEH001223.

Philipp Bach, Victor Chernozhukov, and Martin Spindler. Heterogeneity in the us gender wage gap. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(1):209–230, 2024.

Amine Belhadi, Swapnil S. Kamble, V. Mani, et al. An ensemble machine learning approach for forecasting credit risk of agricultural smes' investments in agriculture 4.0 through supply chain finance. *Annals of Operations Research*, 2021. doi: 10.1007/s10479-021-04366-9.

Alan S. Blinder. Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources*, 8(4):436, 1973. ISSN 0022166X. doi: 10.2307/144855.

Gaelle Chastaing, Fabrice Gamboa, and Clémentine Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis. *Electronic Journal of Statistics*, 6(none), January 2012. ISSN 1935-7524. doi: 10.1214/12-EJS749.

Ward Cheney. *Analysis for Applied Mathematics*, volume 208 of *Graduate Texts in Mathematics*. Springer, 2001. doi: 10.1007/978-1-4613-0101-2. Chapter 3: Calculus in Banach Spaces.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, New York, 2nd edition, 1999. Theorem 3.8, "The Lebesgue-Radon-Nikodym Theorem".

Nicole Fortin, Thomas Lemieux, and Sergio Firpo. Decomposition Methods in Economics. In *Handbook of Labor Economics*, volume 4, pp. 1–102. Elsevier, 2011. ISBN 978-0-444-53450-7. doi: 10.1016/S0169-7218(11)00407-2.

Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29 (5):1189–1232, 2001.

Kathryn E. Hill, Stuart C. Brown, Alice Jones, Damien Fordham, and Robert S. Hill. Modelling climate using leaves of Nothofagus cunninghamii—overcoming confounding factors. *Sustainability*, 15(9):7603, 2023. doi: 10.3390/su15097603. URL https://doi.org/10.3390/su15097603.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *arXiv preprint arXiv:1112.1788*, 2011. URL https://arxiv.org/pdf/1112.1788.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 575–580. ACM, 2004.

Giles Hooker. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, September 2007. ISSN 1061-8600, 1537-2715. doi: 10.1198/106186007X237892.

Feini Huang, Wei Shangguan, Qingliang Li, Lu Li, and Ye Zhang. Beyond prediction: An integrated post-hoc approach to interpret complex model in hydrometeorology. *Environmental Modelling & Software*, 167:105762, 2023. ISSN 1364-8152. doi: 10.1016/j.envsoft.2023.105762. URL https://www.sciencedirect.com/science/article/pii/S1364815223001482.

Jianhua Z. Huang. Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, 26(1), February 1998. ISSN 0090-5364. doi: 10.1214/aos/1030563984.

Evelyn M Kitagawa. Components of a Difference Between Two Rates. *Journal of the American Statistical Association*, pp. 1168–1194, 1955.

F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Wozniakowski. On decompositions of multivariate functions. *Mathematics of Computation*, 79:953–966, 2010. doi: 10.1090/S0025-5718-09-02281-4.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying Interaction Effects with the Functional ANOVA: An Efficient Algorithm for Recovering Identifiable Additive Models. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 119:5979–5989, 2020. URL https://par.nsf.gov/servlets/purl/10298432.

Longyue Liang and Xuanye Cai. Time-sequencing european options and pricing with deep learning – analyzing based on interpretable ale method. *Expert Systems with Applications*, 187:115951, 2022. doi: 10.1016/j.eswa.2021.115951.

Steffen Limmer, Steffen Udluft, and Clemens Otte. Neural-anova: Model decomposition for interpretable machine learning, Aug 2024. URL https://www.researchgate.net/publication/383308412_Neural-ANOVA_Model_Decomposition_for_Interpretable_Machine_Learning.

Ronald Oaxaca. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14(3):693, October 1973. ISSN 00206598. doi: 10.2307/2525981.

M. Peichl, S. Thober, L. Samaniego, B. Hansjürgens, and A. Marx. Machine-learning methods to assess the effects of a non-linear damage spectrum taking into account soil moisture on winter wheat yields in germany. *Hydrology and Earth System Sciences*, 25:6523–6545, 2021. doi: 10.5194/hess-25-6523-2021. URL https://doi.org/10.5194/hess-25-6523-2021.

Sharif Rahman. A generalized anova dimensional decomposition for dependent probability measures. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):670–697, 2014a. doi: 10.1137/120904378. URL https://doi.org/10.1137/120904378.

Sharif Rahman. Approximation errors in truncated dimensional decompositions. *Mathematics of Computation*, 83(290):2799–2819, 2014b. doi: 10.1090/S0025-5718-2014-02883-4. URL https://www.ams.org/journals/mcom/2014-83-290/S0025-5718-2014-02883-4/S0025-5718-2014-02883-4.pdf.

Anthony F. Shorrocks. Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *The Journal of Economic Inequality*, 11(1):99–126, March 2013. ISSN 1569-1721, 1573-8701. doi: 10.1007/s10888-011-9214-z.

I. M. Sobol. Theorems and examples on high dimensional model representations. *Reliability Engineering & System Safety*, 79:187–193, 2003. doi: 10.1016/S0951-8320(02)00229-6.

Charles J. Stone. The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation. *The Annals of Statistics*, 22(1), March 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325361.

Eberhard Zeidler. *Nonlinear Functional Analysis and Its Applications: I: Fixed-Point Theorems*. Springer-Verlag, New York, 1986. ISBN 978-0387964992.

A FANOVA AND ALE MISATTRIBUTION EXAMPLE

A.1 EXAMPLE 1: FANOVA

Consider the setting where the covariates X_1 and X_2 are independent, and let $f(x) = x_1 + x_2$, be the model used for both populations. Assume that for population H we have $\mathbb{E}_{H_X}[X_1] = 0$ and $\mathbb{E}_{H_X}[X_2] = 0$ and for population K we have $\mathbb{E}_{K_X}[X_1] = \mu \neq 0$ and $\mathbb{E}_{K_X}[X_2] = 0$. By Lemma 2, under independence the FANOVA decomposition is equivalent to the recursive formula (see Equation 13) for the Hoeffding–Sobol decomposition (Sobol, 2003; Kuo et al., 2010) for a general probability measure. Then, the FANOVA components are computed as follows:

$$\mathcal{L}(f, H_X, \emptyset)(x) = \mathbb{E}_{H_X}[X_1 + X_2] = 0.$$

$$\mathcal{L}(f, H_X, \{1\})(x) = \mathbb{E}_{H_X}[X_1 + X_2 \mid X_1 = x_1] - \mathcal{L}(f, H_X, \emptyset)(x) = x_1 + \mathbb{E}_{H_X}[X_2] - 0 = x_1.$$

$$\mathcal{L}(f, H_X, \{2\})(x) = \mathbb{E}_{H_X}[X_1 + X_2 \mid X_2 = x_2] - \mathcal{L}(f, H_X, \emptyset)(x) = \mathbb{E}_{H_X}[X_1] + x_2 - 0 = x_2.$$

$$\begin{aligned}\mathcal{L}(f, H_X, \{1,2\})(x) &= \mathbb{E}_{H_X}[X_1 + X_2 \mid X_1 = x_1, X_2 = x_2] - \mathcal{L}(f, H_X, \{1\})(x) - \mathcal{L}(f, H_X, \{2\})(x) - \mathcal{L}(f, H_X, \emptyset)(x) \\ &= (x_1 + x_2) - x_1 - x_2 - 0 = 0.\end{aligned}$$

Thus, the FANOVA components for population H are:

$$\mathcal{L}(f, H_X, \emptyset)(x) = 0, \quad \mathcal{L}(f, H_X, \{1\})(x) = x_1, \quad \mathcal{L}(f, H_X, \{2\})(x) = x_2, \quad \mathcal{L}(f, H_X, \{1,2\})(x) = 0.$$

Similarly, we compute the FANOVA components for population K:

$$\mathcal{L}(f, K_X, \emptyset)(x) = \mathbb{E}_{K_X}[X_1 + X_2] = \mu.$$

$$\mathcal{L}(f, K_X, \{1\})(x) = \mathbb{E}_{K_X}[X_1 + X_2 \mid X_1 = x_1] - \mathcal{L}(f, K_X, \emptyset)(x) = x_1 + \mathbb{E}_{K_X}[X_2] - \mu = x_1 + 0 - \mu = x_1 - \mu.$$

$$\mathcal{L}(f, K_X, \{2\})(x) = \mathbb{E}_{K_X}[X_1 + X_2 \mid X_2 = x_2] - \mathcal{L}(f, K_X, \emptyset)(x) = \mathbb{E}_{K_X}[X_1] + x_2 - \mu = \mu + x_2 - \mu = x_2.$$

$$\begin{aligned}\mathcal{L}(f, K_X, \{1,2\})(x) &= \mathbb{E}_{K_X}[X_1 + X_2 \mid X_1 = x_1, X_2 = x_2] - \mathcal{L}(f, K_X, \{1\})(x) - \mathcal{L}(f, K_X, \{2\})(x) - \mathcal{L}(f, K_X, \emptyset)(x) \\ &= (x_1 + x_2) - (x_1 - \mu) - x_2 - \mu = 0.\end{aligned}$$

Thus, the FANOVA components for population K are:

$$\mathcal{L}(f, K_X, \emptyset)(x) = \mu, \quad \mathcal{L}(f, K_X, \{1\})(x) = x_1 - \mu, \quad \mathcal{L}(f, K_X, \{2\})(x) = x_2, \quad \mathcal{L}(f, K_X, \{1,2\})(x) = 0.$$

Finally, the difference in the FANOVA effects attributed to $Y \mid X$ for the subset $S = \{1\}$ is given by

$$\Delta(\mathcal{L}, f, H_X, K_X, \{1\}) = \mathbb{E}_{H_X}\Big[\mathcal{L}(f, K_X, \{1\})(X) - \mathcal{L}(f, H_X, \{1\})(X)\Big] = \mathbb{E}_{H_X}[X_1 - \mu - X_1] = -\mu \neq 0.$$

Since this term is nonzero, FANOVA misattributes effects to $Y \mid X$ in this example.

A.2 Example 2: ALE

Recall from Section 2.4 that for a differentiable model $f(x_1, x_2)$ the ALE main effect for X_1 is defined by

$$\mathcal{L}(f, K_X, \{1\})(x) \;=\; \int_{x_{\min,1}}^{x_1} \mathbb{E}_{K_X}\Big[\frac{\partial f(z_1, X_2)}{\partial z_1}\Big] dz_1 \;-\; \text{constant},$$

with an analogous definition for X_2. Here, the constant is chosen so that the ALE effect has mean zero over the observed data, we will denote these constants by c_i^j for $i = \{1, 2\}$ and $j = \{H, K\}$. Consider the function $f(x_1, x_2) = x_1 x_2$, which is used for both populations. The relevant partial derivatives are $\frac{\partial f}{\partial x_1} = x_2$ and $\frac{\partial f}{\partial x_2} = x_1$.

For population H assume $X_1 \sim N(1, 1)$ and $X_2 \sim N(0, 1)$. Then, because X_1 and X_2 are independent, $\mathbb{E}_{H_X}[X_2 \mid X_1 = z] = \mathbb{E}_{H_X}[X_2] = 0$. Thus, the ALE component for X_1 is

$$\mathcal{L}(f, H_X, \{1\})(x) \;=\; \int_{x^H_{\min,1}}^{x_1} 0\, dz \;-\; c_1^H \;=\; 0 - c_1^H.$$

In practice, we compute the constants by setting them equal to the empirical mean of the uncentered ALE component $\mathcal{L}(f, H_X, \{2\}) - c_2^H$, assuming a large sample size n so that the Central Limit Theorem provides a good approximation. Theoretically, we take c_1^H such that $\mathbb{E}_{H_X}[\mathcal{L}(f, H_X, \{1\})(X)] = 0$, which gives $c_1^H = 0$.

Similarly, for X_2, since $\mathbb{E}_{H_X}[X_1 \mid X_2 = z] = \mathbb{E}_{H_X}[X_1] = 1$, we obtain

$$\mathcal{L}(f, H_X, \{2\})(x) \;=\; \int_{x^H_{\min,2}}^{x_2} 1\, dz \;-\; c_2^H \;=\; (x_2 - x^H_{\min,2}) \;-\; c_2^H.$$

Choosing c_2^H so that $\mathbb{E}_{H_X}[\mathcal{L}(f, H_X, \{2\})(X)] = 0$ forces $c_2^H = -x^H_{\min,2}$, and thus

$$\mathcal{L}(f, H_X, \{2\})(x) = x_2.$$

Hence,

$$\mathcal{L}(f, H_X, \{1\})(x) = 0 \;\text{ and }\; \mathcal{L}(f, H_X, \{2\})(x) = x_2.$$

For population K assume $X_1 \sim N(0, 1)$ and $X_2 \sim N(\mu, 1)$ with $\mu \neq 0$. Then, by independence, $\mathbb{E}_{K_X}[X_2 \mid X_1 = z] = \mathbb{E}_{K_X}[X_2] = \mu$ and

$$\mathcal{L}(f, K_X, \{1\})(x) \;=\; \int_{x^K_{\min,1}}^{x_1} \mu\, dz \;-\; \text{constant} \;=\; \mu(x_1 - x^K_{\min,1}) \;-\; c_1^K,$$

where the constant c_1^K solves $\mathbb{E}_{K_X}[\mathcal{L}(f, K_X, \{1\})(X)] = 0$, thus $c_1^K = -\mu x^K_{\min,1}$. Similarly, for X_2, since $\mathbb{E}_{K_X}[X_1 \mid X_2 = z] = \mathbb{E}_{K_X}[X_1] = 0$, we obtain

$$\mathcal{L}(f, K_X, \{2\})(x) \;=\; \int_{x^K_{\min,2}}^{x_2} 0\, dz \;-\; \text{constant} \;=\; 0 \;-\; c_2^K,$$

where $c_2^K = 0$. Thus,

$$\mathcal{L}(f, K_X, \{1\})(x) \;=\; \mu(x_1 - x^K_{\min,1}) - \left[-\mu\, x^K_{\min,1}\right] \;=\; \mu\, x_1 \;\text{ and }\; \mathcal{L}(f, K_X, \{2\})(x) \;=\; 0 - 0 \;=\; 0.$$

The difference in the ALE effects attributed to $Y \mid X$ for the change in the covariate $S = \{1\}$ is given by

$$\Delta(\mathcal{L}, f, H, K, \{1\}) \;=\; \mathbb{E}_{H_X}\Big[\mathcal{L}(f, K_X, \{1\})(X) - \mathcal{L}(f, H_X, \{1\})(X)\Big] \;=\; \mathbb{E}_{H_X}\left[\mu\, X_1 - 0\right] \;=\; \mu\, \mathbb{E}_{H_X}[X_1] \;=\; \mu \cdot 1 \;=\; \mu \neq 0.$$

Since this term is not equal to zero, ALE misattributes effects to $Y \mid X$ in this example.

B FANOVA

In this section, we formalize the statements from Section 4.2 regarding the characterization of when FANOVA misattributes.

B.1 Notation and assumptions

We assume a general probability measure P_X, which will denote either H_X or K_X, defined on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ denotes the Borel sigma algebra. We assume the measure P_X is absolutely continuous with respect to a σ-finite reference measure λ, with density $p_X = \frac{dP_X}{d\lambda}$, usually the counting or Lebesgue measure.

The associated Hilbert space is:

$$L^2(P_X) = \left\{ g : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \int_{\mathbb{R}^d} g^2(x) p_X(x)\, d\lambda(x) < \infty \right\},$$

with inner product and norm defined as:

$$\langle g_1, g_2 \rangle_{L^2(P_X)} = \int_{\mathbb{R}^d} g_1(x) g_2(x) p_X(x)\, d\lambda(x), \quad \|g\|_{L^2(P_X)} = \sqrt{\langle g, g \rangle_{L^2(P_X)}}.$$

For the specific cases where $P_X = H_X$ or $P_X = K_X$, we denote their densities as $h(x) = \frac{dH_X}{d\lambda}$ and $k(x) = \frac{dK_X}{d\lambda}$, respectively.

We denote by $\mathcal{M}(\mathcal{X})$ the space of λ-measurable functions on the covariates, representing flexible regression models for the conditional expectation of $Y \mid X$. That is,

$$\mathcal{M}(\mathcal{X}) = \{ f : \mathcal{X} \to \mathbb{R} \mid f \text{ is } \lambda\text{-measurable} \}.$$

Note that, since functions in $\mathcal{M}(\mathcal{X})$ are λ-measurable, they are also measurable with respect to probability measures that are absolutely continuous with respect to λ. We use $\subseteq$ to denote a subset of or equal to a set, and $\subsetneq$ to indicate a proper subset of a set.

When working with the Lebesgue measure, we will write dx in place of $d\lambda(x)$ for readability, particularly in cases where integrals and densities are defined with respect to λ, while keeping dx for standard notation in functionals and expectations.

B.2 Proof of Theorem 2

The following lemma offers an alternative formulation of the orthogonality conditions in Equation 4, ensuring that all components have zero mean with respect to their corresponding input distributions.

Lemma 1 (Lemma 1 (Hooker, 2007))**.** *The orthogonality conditions in Equation 4 hold over $L^2(\mathbb{R}^d)$ if and only if the integral conditions*

$$\forall S \subseteq 2^{[d]},\ \forall i \in S, \quad \int \mathcal{L}(f, K_X, S)(x) k(x)\, dx_i\, dx_{-S} = 0 \tag{12}$$

are satisfied.

Equations 12 are sometimes referred to as the *Weak Annihilating Conditions* (Rahman, 2014a), and we will refer to these later.

Corollary 2. *All FANOVA components have mean zero under their input distribution:* $\mathbb{E}_{K_X}[\mathcal{L}(f, K_X, S)(X)] = 0$.

Proof. The proof follows directly from Lemma 1 by integrating out the marginal distribution of the covariates not in S. ■

The following lemma shows that under the assumption of independence, the FANOVA decomposition is equivalent to the standard result of the Hoeffding-Sobol decomposition (Sobol, 2003; Kuo et al., 2010) for a general probability measure.

Lemma 2. *Let $X_1, \ldots, X_m$ be independent random variables with joint probability distribution K_X. Then, the solution to the FANOVA decomposition problem, as defined in Problem 3, is equivalent to the following recursive formula:*

$$\mathcal{L}(f, K_X, S)(x) = \begin{cases} \mathbb{E}_{K_X}[f(X)], & \text{if } S = \emptyset, \\ \mathbb{E}_{K_X}[f(X) \mid X_S] - \sum_{V \subsetneq S} \mathcal{L}(f, K_X, V)(X), & \text{if } S \neq \emptyset. \end{cases} \tag{13}$$

Proof. Under the assumption of independence, the Weak Annihilating Conditions in Equation 12 are equivalent to the stronger condition:

$$\int \mathcal{L}(f, K_X, S)(x)\, k_i(x_i)\, dx_i = 0, \quad \forall\, i \in S, \tag{14}$$

which follows by the independence assumption by writing $k(x) = \prod_{j=1}^{d} k_j(x_j)$ and integrating out the variables not in S.

Using this result and integrating the additive representation $f(x) = \sum_{S \in 2^{[d]}} \mathcal{L}(f, K_X, S)(x)$ against the set $-S$, we get the recursive formula. For any non-empty subset $S \in 2^{[d]}$, we integrate:

$$\begin{aligned} \mathbb{E}_{K_X}[f(X) \mid X_S] \underset{(indep.)}{=} \int f(x) \prod_{i \in -S} k_i(x_i)\, dx_{-S} &= \int \left(\sum_{V \in 2^{[d]}} \mathcal{L}(f, K_X, V)(x) \right) \prod_{i \in -S} k_i(x_i)\, dx_{-S} \\ &= \sum_{V \in 2^{[d]}} \int \mathcal{L}(f, K_X, V)(x) \prod_{i \in -S} k_i(x_i)\, dx_{-S}. \end{aligned} \tag{15}$$

If $V \cap (-S) = \emptyset$, i.e., $V \subseteq S$, then $\mathcal{L}(f, K_X, V)(x)$ depends only on x_S. Consequently,

$$\int \mathcal{L}(f, K_X, V)(x) \prod_{i \in -S} k_i(x_i)\, dx_{-S} = \mathcal{L}(f, K_X, V)(x) \int \prod_{i \in -S} k_i(x_i)\, dx_{-S} = \mathcal{L}(f, K_X, V)(x).$$

If $V \cap (-S) \neq \emptyset$ but $V \not\subseteq S$, then $\mathcal{L}(f, K_X, V)(x)$ depends on at least one coordinate in $-S$. Due to Equation 14 and the independence of $k(x)$, we have:

$$\int \mathcal{L}(f, K_X, V)(x) \prod_{i \in -S} k_i(x_i)\, dx_{-S} = 0.$$

Thus, Equation 15 becomes

$$\mathbb{E}_{K_X}[f(X) \mid X_S] = \sum_{V \subseteq S} \mathcal{L}(f, K_X, V)(x) = \sum_{V \subsetneq S} \mathcal{L}(f, K_X, V)(x) + \mathcal{L}(f, K_X, S)(x)$$

$$\iff \mathcal{L}(f, K_X, S)(x) = \mathbb{E}_{K_X}[f(X) \mid X_S] - \sum_{V \subsetneq S} \mathcal{L}(f, K_X, V)(x).$$

Similarly, for $S = \emptyset$, taking the expectation over all variables X yields the corresponding formula: $\mathcal{L}(f, K_X, \emptyset) = \mathbb{E}_{K_X}[f(X)]$.

■

Lemma 2 was previously shown by Rahman (2014b, Corollary 4.6) in the context of the generalized ANOVA dimensional decomposition, which reduces to the standard FANOVA dimensional decomposition. We restate it here using the specific terminology of FANOVA and provide a concrete proof to derive the reduced formula solution.

We can now formally prove Theorem 2 which characterizes the conditions under which FANOVA does not misattribute for a given affine function of the covariates f.

Proof of Theorem 2. ($\Rightarrow$) Suppose that

$$\mathbb{E}_{H_X}\left[\mathcal{L}(f, K_X, S)(X)\right] = 0, \quad \text{for all } S \in 2^{[d]} \setminus \{\emptyset\}.$$

By Lemma 2 we have that under the independence assumption, FANOVA reduces to the recursive form solution in Equation 13. Then, for $S \neq \emptyset$, the component would be:

$$\mathcal{L}(f, K_X, \emptyset)(x) = \mathbb{E}_{K_X}[f(X)] = \sum_{m=1}^{M} a_m \mathbb{E}_{K_X}[b_m(X_m)].$$

For the single effects, $S = \{m\}$, the additive components would take the form:

$$\mathcal{L}(f, K_X, \{m\})(x) = \mathbb{E}_{K_X}[f(X)|X_m] - \mathcal{L}(f, K_X, \emptyset)(x).$$

Expanding this, and using the independence assumption, we get:

$$\begin{aligned}\mathcal{L}(f, K_X, \{m\})(x) &= \left(a_m b_m(x_m) + \sum_{j \neq m} a_j \mathbb{E}_{K_X}[b_j(X_j)]\right) - \left(\sum_{j=1}^{M} a_j \mathbb{E}_{K_X}[b_j(X_j)]\right) \\ &= a_m \left(b_m(x_m) - \mathbb{E}_{K_X}[b_m(X_m)]\right), \quad \forall\, m = 1, \ldots, M.\end{aligned}$$

Lastly, for S with $|S| \geq 2$, we have that the recursive formula for the Functional ANOVA components is given by

$$\mathcal{L}(f, K_X, S)(x) = \mathbb{E}_{K_X}[f(X) \mid X_S] - \sum_{V \subsetneq S} \mathcal{L}(f, K_X, V)(x).$$

Where, $\mathbb{E}_{K_X}[f(X) \mid X_S] = \sum_{m\in S} a_m b_m(x_m) + \sum_{m\notin S} a_m \mathbb{E}_{K_X}[b_m(X_m)]$ and the sum of lower-order terms takes the form:

$$\begin{aligned}\sum_{V\subsetneq S} \mathcal{L}(f, K_X, V)(x) &= \mathcal{L}(f, K_X, \emptyset) + \sum_{m\in S} \mathcal{L}(f, K_X, \{m\})(x) \\ &= \sum_{m=1}^{M} a_m \mathbb{E}_{K_X}[b_m(X_m)] + \sum_{m\in S} a_m \big(b_m(x_m) - \mathbb{E}_{K_X}[b_m(X_m)]\big) \\ &= \sum_{m\in S} a_m b_m(x_m) + \sum_{m\notin S} a_m \mathbb{E}_{K_X}[b_m(X_m)].\end{aligned}$$

Substituting into the recursive formula:

$$\mathcal{L}(f, K_X, S)(x) = \Big(\sum_{m\in S} a_m b_m(x_m) + \sum_{m\notin S} a_m \mathbb{E}_{K_X}[b_m(X_m)]\Big) - \Big(\sum_{m\in S} a_m b_m(x_m) + \sum_{m\notin S} a_m \mathbb{E}_{K_X}[b_m(X_m)]\Big) = 0.$$

That is,

$$\mathcal{L}(f, K_X, S)(x) = 0, \quad \forall\, S \text{ with } |S| \geq 2.$$

Now, by hypothesis, for every $m = 1, \ldots, M$, we have:

$$0 = \mathbb{E}_{H_X}\left[\mathcal{L}(f, K_X, \{m\})(X)\right] = \mathbb{E}_{H_X}\left[a_m \left(b_m(X_m) - \mathbb{E}_{K_X}[b_m(X_m)]\right)\right] \iff 0 = a_m \left(\mathbb{E}_{H_X}[b_m(X_m)] - \mathbb{E}_{K_X}[b_m(X_m)]\right).$$

We also assumed that $a_m \neq 0$ for all $m = 1, \ldots, M$. Therefore, we conclude that

$$\mathbb{E}_{H_X}[b_m(X_m)] = \mathbb{E}_{K_X}[b_m(X_m)], \quad \text{for all } m = 1, \ldots, M.$$

($\Leftarrow$) Suppose that $\mathbb{E}_{H_X}[b_m(X_m)] = \mathbb{E}_{K_X}[b_m(X_m)]$ for each $m = 1, \ldots, M$. Then,

$$\mathbb{E}_{H_X}\left[\mathcal{L}(f, K_X, \{m\})(X_m)\right] = \mathbb{E}_{H_X}\left[a_m \left(b_m(X_m) - \mathbb{E}_{K_X}[b_m(X_m)]\right)\right] = a_m \left(\mathbb{E}_{H_X}[b_m(X_m)] - \mathbb{E}_{K_X}[b_m(X_m)]\right) = 0.$$

■

B.3 Proof of Theorem 3

Note that in Theorem 3 we have assumed that any measurable function of the covariates satisfies an additive decomposition. However, this assumption can be relaxed by imposing a couple of conditions, such as the probability measure being dominated by a product measure (Hoffman et al., 2011, Equation C.1), and a boundedness assumption on the densities (Hoffman et al., 2011, Equations C.2 or C.3). Under these assumptions, it follows that for any square-integrable measurable function of the covariates, there exist functions such that the original function admits an additive representation; see Hoffman et al. (2011, Theorem 1).

To show Theorem 3, we first state the definition of mean-zero functions and then prove Lemma 3 and Lemma 4, which will serve as intermediate steps for the main proof.

Definition 3 (Mean-zero square integrable functions)**.** *We denote by $W(K_X)$ the space of mean-zero functions in $L^2(K_X)$ with respect to the probability measure K_X:*

$$W(K_X) = \left\{\phi \in L^2(K_X) : \int_{\mathcal{X}} \phi(x) k(x)\, d\lambda(x) = 0\right\}.$$

The following lemma states that the orthogonal complement of the space of mean-zero square-integrable functions with respect to a probability measure is the space of almost surely constant functions.

Lemma 3. *Let K_X be a probability measure on $\mathbb{R}^d$ with density $k(x)$ with respect to a reference measure λ. The orthogonal complement of $W(K_X)$ in $L^2(K_X)$ is the space of constant functions, that is:*

$$W(K_X)^\perp = \left\{ f \in L^2(K_X) : f(x) = c,\ K_X\text{-a.s.} \right\},$$

where $c \in \mathbb{R}$ is a constant.

Proof. Let

$$V = \left\{ f \in L^2(K_X) : f(x) = c,\ K_X\text{-a.s.} \right\}.$$

We will prove that $W(K_X)^\perp = V$ by first showing that $V \subseteq W(K_X)^\perp$. Let $f \in V$; then $f(x) = c$ for K_X-almost every $x \in \mathcal{X}$. Thus, for any $\phi \in W(K_X)$, we have

$$\mathbb{E}_{K_X}[f(X)\phi(X)] = c \cdot \mathbb{E}_{K_X}[\phi(X)] = 0.$$

It remains to show that $W(K_X)^\perp \subseteq V$. Let $f \in W(K_X)^\perp$; then $\mathbb{E}_{K_X}[f(X)\phi(X)] = 0$ for every $\phi \in W(K_X)$. In particular, take an arbitrary K_X-measurable set $A \subseteq \mathcal{X}$ and define $\psi_A(x) = \chi_A(x) - K_X(A)$. Where $\chi_A(x)$ is the indicator function over A and $K_X(A)$ is the probability of the event A under the distribution K_X. Clearly, ψ_A belongs to $W(K_X)$. Moreover,

$$\begin{aligned} 0 = \mathbb{E}_{H_X}[f(X)\psi(X)] &= \mathbb{E}_{H_X}[f(X)\chi_A(X)] - \mathbb{E}_{H_X}[f(X)K_X(A)] \\ &= \int_A f(x)\, dK_X - K_X(A) \int_{\mathcal{X}} f(x)\, dK_X. \end{aligned}$$

$$\iff \int_A f(X)\, dK_X = K_X(A) \int_{\mathcal{X}} f(x)\, dK_X. \tag{16}$$

Define $\mu(A) = \int_A f(x)\, dK_X(x)$, which is a signed measure with respect to K_X. By the Radon–Nikodym theorem for signed measures, f is the unique Radon–Nikodym derivative $d\mu/dK_X$. Since we also have, by Equation 16 that

$$\mu(A) = K_X(A) \int_{\mathcal{X}} f(x)\, dK_X = K_X(A) \cdot c = c \int_A dK_X = \int_A c\, dK_X,$$

it follows by the uniqueness of the Radon–Nikodym derivative that $f(x) = c$ K_X-almost surely. Therefore, $V = W(K_X)^\perp$. ■

The following lemma shows that the space of mean-zero functions is equivalent to the span of hierarchical orthogonal functional ANOVA components, obtained by varying the covariate functions over the entire space of measurable functions.

Lemma 4. *Suppose $f \in \mathcal{M}(\mathcal{X})$ and let $\mathcal{L}(f, K_X, S)$ denote the functional ANOVA component corresponding to the subset of covariates S. Define $\mathcal{F} = \{\mathcal{L}(f, K_X, S)(x) : S \in 2^{[d]}, S \neq \emptyset,\ f \in L^2(K_X)\}$. Then, span$\{\mathcal{F}\} = W(K_X)$.*

Proof. ($\subseteq$) Let $\phi(x) = \sum_{i=1}^n c_i\, \mathcal{L}(f_i, K_X, S_i)(x_{S_i})$, where $c_i \in \mathbb{R}$, $f_i \in L^2(K_X)$, and $S_i \in 2^{[d]} \setminus \{\emptyset\}$.

Since $\mathbb{E}_{K_X}\left[\mathcal{L}(f_i, K_X, S_i)(X_{S_i})\right] = 0$, we have $\mathbb{E}_{K_X}[\phi(X)] = 0$. Moreover, ϕ is square integrable since each summand $\mathcal{L}(f_i, K_X, S_i)(x_{S_i})$ is square integrable. Thus, $\phi \in W(K_X)$ and span$\{\mathcal{F}\} \subseteq W(K_X)$.

($\supseteq$) Take an arbitrary $\phi \in W(K_X)$. Because $W(K_X) \subseteq L^2(K_X)$, we have that ϕ is square integrable and so has an additive decomposition:

$$\phi(x) = \sum_{S \in 2^{[d]}} \mathcal{L}(\phi, K_X, S)(x) = \sum_{S \in 2^{[d]} \setminus \{\emptyset\}} \mathcal{L}(\phi, K_X, S)(x) + \mathbb{E}_{K_X}[\phi(X)] = \sum_{S \in 2^{[d]} \setminus \{\emptyset\}} \mathcal{L}(\phi, K_X, S)(x),$$

where the last equality holds because $E_{K_X}[\phi(X)] = 0$. Each term on the right-hand side of the last expression belongs to $\mathcal{F}$. Thus, $\phi \in \text{span}\{\mathcal{F}\}$. ■

With these two results, we now proceed to prove Theorem 3.

Proof of Theorem 3. ($\Rightarrow$) Suppose $\mathbb{E}_{H_X}[\mathcal{L}(f, K_X, S)(X)] = 0$. By the mean zero property of FANOVA, $\mathbb{E}_{K_X}[\mathcal{L}(f, K_X, S)(X)] = 0$. Then,

$$\begin{aligned}
\mathbb{E}_{H_X}[\mathcal{L}(f, K_X, S)(X)] &= \mathbb{E}_{K_X}[\mathcal{L}(f, K_X, S)(X)] \\
&\iff 0 = \mathbb{E}_{K_X}[\mathcal{L}(f, K_X, S)(X)] - \mathbb{E}_{H_X}[\mathcal{L}(f, K_X, S)(X)] \\
&\iff 0 = \mathbb{E}_{K_X}[\mathcal{L}(f, K_X, S)(X)] - \mathbb{E}_{K_X}\left[\mathcal{L}(f, K_X, S)(X)\frac{h(x)}{k(x)}\right] \\
&\iff 0 = \int \mathcal{L}(f, K_X, S)(X)\left(1 - \frac{h(x)}{k(x)}\right) dK_X, \quad \text{for all } f \in \mathcal{M}(\mathcal{X}) \text{ and for all } S \neq \emptyset.
\end{aligned} \tag{17}$$

By Lemma 4, as we vary f and S over the space $\mathcal{M}(\mathcal{X}) \times \left(2^{[d]} \setminus \{\emptyset\}\right)$, the components $\{\mathcal{L}(f, K_X, S)\}_{(f,S)}$ span the mean-zero space; that is, $\text{span}\{\mathcal{F}\} = W(K_X)$. Therefore, Equation 17 is equivalent to

$$0 = \int \mathcal{L}(f, K_X, S)(x)\left(1 - \frac{h(x)}{k(x)}\right) dK_X = 0, \quad \text{for all } \mathcal{L}(f, K_X, S)(x) \in W(K_X).$$

Thus, $\left(1 - \frac{h(x)}{k(x)}\right) \in W(K_X)^{\perp}$. i.e., $(1 - \frac{h(x)}{k(x)})$ is orthogonal to all zero-mean functions in $L^2(K_X)$. Furthermore, by Lemma 3 we know that the orthogonal space to $W(K_X)$ is the space of K_X-almost surely constant functions. Thus, there must exist a constant $c \in \mathbb{R}$ such that $1 - \frac{h(x)}{k(x)} = c$, K_X-a.s. Finally, noting that $\int dK_X - \int \frac{h(x)}{k(x)} dK_X = \int c\, dK_X$, we have $c = 0$, K_X-a.s. Therefore,

$$\frac{h(x)}{k(x)} = 1 \quad \Rightarrow \quad h(x) = k(x), \quad K_X\text{-a.s.} \quad \Rightarrow \quad H_X = K_X, \quad K_X\text{-a.s.}$$

The only if part ($\Leftarrow$) follows by the mean-zero property of $\mathcal{L}(f, K_X, S)(x)$ under K_X. ■

C Mathematical framework and Section 5 results

In this section, we describe in detail the mathematical background necessary for Section 5, along with the additional notation and assumptions required to prove our main results and state our main conjecture.

C.1 The discrete case

We now formalize the assumptions discussed in Section 5.1 and formally prove Theorem 4. First, let

$$\Sigma = \left\{z \in \mathbb{R}^d : z_i \geq 0 \ \forall i,\ \mathbf{1}^T z = 1\right\},$$

denote the standard $d-1$ simplex and let Σ^{o} denote its interior. Following Definition 3 for the space of mean-zero square integrable functions of the covariates, we denote by W the space of mean-zero vectors:

$$W = \left\{\phi \in \mathbb{R}^d : \mathbf{1}^T\phi = 0\right\}.$$

We make the following mild regularity conditions on the functional decomposition $\mathcal{L}$.

Assumption 1. *The following hold:*

1. ***Twice differentiable**. For any f and S, the map $K_X \to \mathcal{L}(f, K_X, S)$ is twice continuously differentiable.*
2. ***Uniformly bounded condition**. There exists a constant $M > 0$ such that $\sup_{(f,K_X,S)} \|\mathcal{H}_{K_X}\mathcal{L}(f, K_X, S)\|_{\mathrm{op}} \leq M$, where $\mathcal{H}_{K_X}\mathcal{L}(f, K_X, S)$ is the Hessian of $\mathcal{L}(f, K_X, S)$ with respect to K_X.*

Before proving Theorem 4, we show two auxiliary Lemmas: Lemma 5 states that for any vector in the interior of the simplex and any mean-zero vector, there always exists a small perturbation along the mean-zero vector that keeps the perturbed vector within the simplex. Lemma 7 serves as an intermediate step in proving Proposition 1, which characterizes matrices satisfying a specific condition that the Jacobian of a general functional decomposition must satisfy.

Lemma 5. *For every $x \in \Sigma^{\mathrm{o}}$ and $\phi \in W$, there exists an open interval $I \subsetneq \mathbb{R}$ containing 0 such that for all $\varepsilon \in I$, $x + \varepsilon\phi \in \Sigma^{\mathrm{o}}$.*

Proof. Let $y = x + \varepsilon\phi$. We need to show that $y \in \Sigma^{\mathrm{o}}$ for a suitable choice of ε within some interval I. Note that $\mathbf{1}^\top y = \mathbf{1}^\top x + \varepsilon\mathbf{1}^\top\phi = 1$, which implies that y satisfies the constraint $\mathbf{1}^\top y = 1$. Thus, it remains to verify that $y_i > 0$ for all i.

If $\phi_i > 0$, then $x_i + \varepsilon\phi_i > 0$ holds for all ε such that $\varepsilon > -\frac{x_i}{\phi_i}$, which is also satisfied by taking

$$\varepsilon > \max_{i:\phi_i>0}\left\{-\frac{x_i}{\phi_i}\right\}.$$

If $\phi_i < 0$, then $x_i + \varepsilon\phi_i > 0$ holds for all ε such that $\varepsilon < -\frac{x_i}{\phi_i}$, which is also satisfied by taking

$$\varepsilon < \min_{i:\phi_i<0}\left\{-\frac{x_i}{\phi_i}\right\}, \quad \text{note } -\frac{x_i}{\phi_i} > 0.$$

If $\phi_i = 0$, then any $\varepsilon \in \mathbb{R}$ satisfies $y_i > 0$. Therefore, by choosing

$$\varepsilon \in \left(\max_{i:\phi_i>0}\left\{-\frac{x_i}{\phi_i}\right\}, \min_{i:\phi_i<0}\left\{-\frac{x_i}{\phi_i}\right\}\right) =: I,$$

we ensure that $y_i > 0$ for all i. Thus, $y \in \Sigma^{\mathrm{o}}$. ■

Remark 1. *Lemma 5 is equivalent to stating that for a given $x \in \Sigma^{\mathrm{o}}$ there exists some point $y \in \Sigma^{\mathrm{o}}$ and $\varepsilon > 0$ such that $\phi = \frac{1}{\varepsilon}(y - x)$. That is, we can recover any vector $\phi \in W$ given an initial x and a suitable pair $(\varepsilon, y) \in \mathbb{R}_{++} \times \Sigma^{\mathrm{o}}$.*

Lemma 6. *$W^\perp = span\{\mathbf{1}\}$.*

Proof. ($\supseteq$) Let $u = \alpha\mathbf{1}$ for some scalar $\alpha \in \mathbb{R}$. For any $w \in W$, we have $\mathbf{1}^\top w = 0$. Then:

$$u^\top w = (\alpha\mathbf{1})^\top w = \alpha(\mathbf{1}^\top w) = \alpha \cdot 0 = 0.$$

Thus, $\text{span}\{\mathbf{1}\} \subseteq W^\perp$.

($\subseteq$) Let $u \in W^\perp$, that is $u^\top w = 0$ for all $w \in W$. Consider the vectors $e_j - e_k$ (where e_j is the j-th standard basis vector) with $j \neq k$. Note that: $\mathbf{1}^\top(e_j - e_k) = 1 - 1 = 0$. Hence, $e_j - e_k \in W$. Since $u \in W^\perp$, it follows that:

$$u^\top(e_j - e_k) = 0 \implies u_j - u_k = 0 \implies u_j = u_k.$$

As j and k were arbitrary, all coordinates of u are equal. Thus, there exists a scalar α such that $u = \alpha\mathbf{1}$. This implies $u \in \text{span}\{\mathbf{1}\}$. Therefore, $W^\perp = \text{span}\{\mathbf{1}\}$.

■

Lemma 7. *Let $w \in \mathbb{R}^d$. If $x^\top w = 0$, for all $x \in \Sigma^{\text{o}}$, then $w = 0$.*

Proof. We proceed by contradiction. Suppose $w \neq 0$; we will show that there exists some vector $x \in \Sigma^{\text{o}}$ such that $x^\top w \neq 0$.

Assume there is a component i such that $w_i > 0$. Take $x_i = 1 - \alpha$ and $x_{j\neq i} = \frac{\alpha}{d-1}$ for some $\alpha \in (0, 1)$. Clearly, $x \in \Sigma^{\text{o}}$. We will show that, for a valid α, $x^\top w > 0$. We start by noting that:

$$x^\top w = (1-\alpha)w_i + \sum_{j\neq i} \frac{\alpha}{d-1} w_j > 0 \iff w_i - \alpha w_i + \frac{\alpha}{d-1}\left(\sum_{j\neq i} w_j\right) > 0$$

$$\iff w_i + \alpha\left(\frac{\sum_{j\neq i} w_j}{d-1} - w_i\right) > 0.$$

Let $M = \frac{\sum_{j\neq i} w_j}{d-1} - w_i$. Then, we have the following subcases:

Case 1: If $M > 0$, then $w_i + \alpha M > w_i > 0$ for all $\alpha > 0$.

Case 2: If $M < 0$, then $\alpha < \frac{-w_i}{M}$, where $\frac{-w_i}{M} > 0$. Thus, any $\alpha \in \left(0, \frac{-w_i}{M}\right)$ would imply that $x^\top w > 0$.

The case where $w_i < 0$ for some i follows by a completely analogous argument. Therefore, if $w \neq 0$, we can always find an $x \in \Sigma^{\text{o}}$ such that $x^\top w \neq 0$.

Hence, it must be that $w = 0$. ■

Proposition 1. *Let $x, y \in \Sigma^{\text{o}}$ and $A \in \mathbb{R}^{d\times d}$. Then,*

$$x^\top A(y - x) = 0, \text{ for all } x, y \in \Sigma^{\text{o}},$$

if and only if

$$A = \mathbf{c}\mathbf{1}^\top, \text{ for some } \mathbf{c} \in \mathbb{R}^d. \tag{18}$$

Proof. Consider a fixed $x \in \Sigma^{\text{o}}$. By Lemma 5, for any $\phi \in W$, there exists $I \subsetneq \mathbb{R}$ such that $\forall \varepsilon \in I$, $y = x + \varepsilon\phi \in \Sigma^{\text{o}}$. Thus, $\frac{y-x}{\varepsilon} = \phi \in W$, and as we vary $(\varepsilon, y) \in I \times \Sigma^{\text{o}}$, we recover any $\phi \in W$ (see the proof of Lemma 5).

Since $x \neq y$, we have $\varepsilon \neq 0$. Therefore,

$$0 = x^\top A(y - x) = x^\top A\varepsilon\phi \implies x^\top A\phi = 0, \forall \phi \in W.$$

It follows that $x^\top A \in W^\perp$ for each $x \in \Sigma^{\text{o}}$. By Lemma 6, we know that $W^\perp = \text{span}\{\mathbf{1}\}$. We claim that if

$$x^\top A \in \text{Span}\{\mathbf{1}\} \text{ for all } x \in \Sigma^{\text{o}}, \text{ then } A = \mathbf{c}\mathbf{1}^\top, \text{ for some } \mathbf{c} \in \mathbb{R}^d.$$

Suppose $x^\top A \in \text{Span}\{\mathbf{1}\}$. This implies there exists $c(x) \in \mathbb{R}$ such that $x^\top A = c(x)\mathbf{1}^\top$. Let a_j denote the j-th column of A. Then, $x^\top a_j = c(x)$ for all $j \in \{1, \dots, d\}$. In particular, for $i \neq j$, we have

$$x^\top a_i = x^\top a_j \iff x^\top (a_i - a_j) = 0, \text{ for all } x \in \Sigma^{\text{o}}. \tag{19}$$

Applying Lemma 7 to the vector $w = a_i - a_j \in \mathbb{R}^d$, we know that Equation 19 implies $w = a_i - a_j = 0$. Thus, $a_1 = a_2 = \cdots = a_d$, which means $A = \mathbf{c}\mathbf{1}^\top$ for some $\mathbf{c} \in \mathbb{R}^d$. ■

Now we proceed to prove our main discrete characterization result.

Proof of Theorem 4. ($\Rightarrow$) Since $\mathcal{L}(f, K_X, S)$ is continuously differentiable, by the Mean Value Theorem there exists $\tilde{H_X}$ on the line segment between K_X and H_X, such that

$$\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) = \mathcal{J}_{K_X}\mathcal{L}(f, \tilde{H_X}, S)(K_X - H_X), \tag{20}$$

where $\mathcal{J}_{K_X}\mathcal{L}(f, \tilde{H_X}, S)$ is the Jacobian of $\mathcal{L}(f, K_X, S)$ with respect to K_X.

Consider the path $\gamma(t) = H_X + t\,(\tilde{H_X} - H_X)$, $t \in [0, 1]$. Then $\gamma(0) = H_X$ and $\gamma(1) = \tilde{H_X}$. By the fundamental theorem of calculus (in vector form), we have

$$\mathcal{J}_{K_X}\mathcal{L}(f, \tilde{H_X}, S) - \mathcal{J}_{K_X}\mathcal{L}(f, H_X, S) = \int_0^1 \mathcal{H}_{K_X}\mathcal{L}\big(f, \gamma(t), S\big)\left[\tilde{H_X} - H_X\right] dt.$$

Taking the operator norm on both sides, we obtain

$$\begin{aligned}
\left\|\mathcal{J}_{K_X}\mathcal{L}\big(f, \tilde{H_X}, S\big) - \mathcal{J}_{K_X}\mathcal{L}\big(f, H_X, S\big)\right\| &= \left\|\int_0^1 \mathcal{H}_{K_X}\mathcal{L}\big(f, \gamma(t), S\big)\,(\tilde{H_X} - H_X)\,dt\right\| \\
&\leq \int_0^1 \left\|\mathcal{H}_{K_X}\mathcal{L}\big(f, \gamma(t), S\big)\right\| \|\tilde{H_X} - H_X\|\,dt \\
&\leq \int_0^1 M\,\|\tilde{H_X} - H_X\|\,dt, \quad \text{(by the uniform bound } \leq M) \\
&= M\,\|\tilde{H_X} - H_X\| \int_0^1 dt \;=\; M\,\|\tilde{H_X} - H_X\|.
\end{aligned}$$

Thus we have

$$\left\|\mathcal{J}_{K_X}\mathcal{L}(f, \tilde{H_X}, S) - \mathcal{J}_{K_X}\mathcal{L}(f, H_X, S)\right\| \;\leq\; M\,\|\tilde{H_X} - H_X\|.$$

That is,

$$\mathcal{J}_{K_X}\mathcal{L}(f, \tilde{H_X}, S) = \mathcal{J}_{K_X}\mathcal{L}(f, H_X, S) + O\big(\|\tilde{H_X} - H_X\|\big).$$

Substituting back into Equation 20, we get

$$\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) = \mathcal{J}_{K_X}\mathcal{L}(f, H_X, S)(K_X - H_X) + O\big(\|K_X - H_X\|^2\big). \tag{21}$$

By Lemma 5, for a fixed h and for any $\phi \in W$, there exists $I \subsetneq \mathbb{R}$ such that $\forall \varepsilon \in I$, $k = h + \varepsilon\phi \in \Sigma^{\text{o}}$. Since $h \neq k$ we have that $\varepsilon \neq 0$. Thus, $\frac{y-x}{\varepsilon} = \phi \in W$, and as we vary $(\varepsilon, y) \in I \times \Sigma^{\text{o}}$, we recover any $\phi \in W$, see proof of Lemma 5. Substituting in Equation 21 we have that

$$\mathcal{L}(f, K_X, S) - \mathcal{L}(f, K_X + \varepsilon\phi, S) = \mathcal{J}_{K_X}\mathcal{L}(f, K_X + \varepsilon\phi, S)(\varepsilon\phi) + O\big(\varepsilon^2\|\phi\|^2\big)$$

$$\iff \lim_{\varepsilon \to 0} \frac{\mathcal{L}(f, K_X, S) - \mathcal{L}(f, K_X + \varepsilon\phi, S)}{\varepsilon} = \lim_{\varepsilon \to 0} \left\{ \mathcal{J}_{K_X} \mathcal{L}(f, K_X + \varepsilon\phi, S)\phi + O\big(\varepsilon \|\phi\|^2\big) \right\}$$

$$\underset{(1)}{\iff} \lim_{\varepsilon \to 0} \frac{\mathcal{L}(f, K_X, S) - \mathcal{L}(f, K_X + \varepsilon\phi, S)}{\varepsilon} = \lim_{\varepsilon \to 0} \left\{ \mathcal{J}_{K_X} \mathcal{L}(f, K_X + \varepsilon\phi, S)\phi + O\big(\varepsilon \|\phi\|^2\big) \right\}$$

$$\underset{(2)}{\iff} \lim_{\varepsilon \to 0} \frac{\mathcal{L}(f, K_X, S) - \mathcal{L}(f, K_X + \varepsilon\phi, S)}{\varepsilon} = \mathcal{J}_{K_X} \mathcal{L}(f, K_X, S)\phi$$

$$\underset{(3)}{\iff} \mathbb{E}_{H_X} \left[\lim_{\varepsilon \to 0} \frac{\mathcal{L}(f, K_X, S)(X) - \mathcal{L}(f, K_X + \varepsilon\phi, S)(X)}{\varepsilon} \right] = h^\top \mathcal{J}_{K_X} \mathcal{L}(f, K_X, S)\phi$$

$$\underset{(4)}{\iff} \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{E}_{H_X} \left[\mathcal{L}(f, K_X, S)(X) - \mathcal{L}(f, K_X + \varepsilon\phi, S)(X) \right] = h^\top \mathcal{J}_{K_X} \mathcal{L}(f, K_X, S)\phi$$

$$\underset{(5)}{\iff} 0 = h^\top \mathcal{J}_{K_X} \mathcal{L}(f, K_X, S)\phi, \quad \forall \phi \in W.$$

Where (1) follows by dividing by ε and taking the limit as it goes to zero; (2) follows from the continuity of the first derivative and the definition of $O(\cdot)$; (3) follows from taking expectations under the probability density H_X; (4) follows from interchanging limits and expectations; and (5) follows from the hypothesis.

Finally, the result follows from Proposition 1, which implies that $\mathcal{J}_{K_X} \mathcal{L}(f, K_X, S)) = \mathbf{c}\mathbf{1}^T$ for some $\mathbf{c} \in \mathbb{R}^d$.

($\Leftarrow$) If $\mathcal{J}_{K_X} \mathcal{L}(f, K_X, S) = \mathbf{c}\mathbf{1}^T$ for any K_X, then by the mean value theorem, substituting into Equation 20, we obtain $\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) = \mathbf{c}\mathbf{1}^T(K_X - H_X) = \mathbf{c}(1 - 1) = 0$. Thus, $\mathcal{L}(f, K_X, S) = \mathcal{L}(f, H_X, S)$ for all H_X, K_X, and consequently, $\mathbb{E}_{K_X}[\mathcal{L}(f, K_X, S)(X) - \mathcal{L}(f, H_X, S)(Xs)] = 0$ for all $H_X, K_X \in \Sigma^{\text{o}}$.

■

C.2 THE CONTINUOUS SETTING

In this section, we introduce the continuous setting and motivate its relevance in a more expository manner; a more formal treatment is provided in the following appendix section.

In section 5.1, Theorem 4 and Corollary 1 show that, in the discrete case, a functional decomposition $\mathcal{L}$ that never misattributes effects must be constant with respect to the distribution over covariates. We now analyze the continuous setting, introducing pertinent regularity assumptions to study how $\mathcal{L}$ responds to perturbations in the input distribution.

Namely, we assume $\mathcal{L}(f, K_X, S)$ is a continuous functional in its first argument f, Lebesgue measurable in its second argument, K_X, and square integrable, in the L^2 sense, for all triplets (f, K_X, S). For example, our first condition is satisfied in cases such as in FANOVA, when $\mathcal{L}$ is the integral operator with respect to any probability measure absolutely continuous with respect to the Lebesgue measure. The third assumption is identical to those in FANOVA and ALE, which both require $\mathcal{L}$ to belong to the space of square integrable functions, L^2. Lastly, we assume that the densities $k(x)$ belong to the space of compactly supported functions, which we denote by $\mathcal{P}(\mathcal{X})$. Throughout, we use the notation $K_X \in \mathcal{P}(\mathcal{X})$ or $k(x) \in \mathcal{P}(\mathcal{X})$ interchangeably to refer to the probability measure and its corresponding density—this abuse of notation will be clear from context. The definition of ALE already assumes compactly support densities. In practice, most distributions can be restricted to a compact region (e.g., age, income, and years of education are all bounded).

We parametrize perturbations around a density $k(x)$ as $k(x) + \phi(x)$, for admissible[3] functions ϕ. We denote by $\mathcal{D}_{K_X}$ the set of admissible perturbation functions of K_X. Throughout, we may write $k(x) + \phi(x)$ or $K_X + \phi$ interchangeably to denote such perturbations—again this is a mild abuse of notation and will be clear from context. Under an additional condition, assuming that $\mathcal{L}$ is continuously differentiable as a function of ϕ, we ensure that we can approximate $\mathcal{L}(\cdot, \phi)$ with a linear approximation around zero:

$$\mathcal{L}(\cdot, \phi) \approx \mathcal{L}(\cdot, 0) + D_\phi \mathcal{L}(\cdot, 0)[\phi].$$

Where $\mathcal{L}(\cdot, \phi)$ is short notation for $\mathcal{L}(f, K_X + \phi, S)$ and $D_\phi \mathcal{L}(\cdot, 0)$ is the Fréchet derivative of $\mathcal{L}$ with respect to the function ϕ evaluated at $\phi = 0$. The Fréchet derivative is an operator, and $D_\phi \mathcal{L}(\cdot, 0)[\phi]$ denotes it taking ϕ as input. See Definition 4 and Appendix D.2 for a more rigorous discussion of the perturbation functions. Although we have not yet verified that FANOVA and ALE satisfy continuous differentiability with respect to perturbations, our conditions are mild, so we conjecture that this is the case. We now attempt to characterize the behavior of the functional $\mathcal{L}$ under small perturbations of the distribution K_X.

Theorem 5. *Assume the above regularity conditions on $\mathcal{L}$ (see Assumptions 2 and 3 in the Appendix). Let $K_X \in \mathcal{P}(\mathcal{X})$, and let $\mathcal{D}_{K_X}$ denote the set of admissible perturbation functions of K_X. If for all $\phi \in \mathcal{D}_{K_X}$, we have*

$$\mathbb{E}_{K_X + \phi} \left[\mathcal{L}(f, K_X, S)(X) - \mathcal{L}(f, K_X + \phi, S)(X) \right] = 0,$$

then

$$\mathbb{E}_{K_X} \left[D_\phi \mathcal{L}(\cdot, 0)[\phi](X) \right] = 0, \quad \textit{for all } \phi \in \mathcal{D}_{K_X}. \tag{22}$$

See Appendix D.3 for the proof.

Theorem 6. *Under the assumptions of Theorem 5, if a functional decomposition $\mathcal{L}(f, K_X, S)$ does not depend on its input distribution (i.e., $\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) = 0$ for all f, K_X, H_X, S), then it does not misattribute effects of $Y \mid X$.*

Proof. The proof is straightforward: by definition, if $\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) = 0$ for all f, K_X, H_X, S, then $\Delta(\mathcal{L}, f, K_X, H_X, S) = 0$. ■

We verify that when the KOB decomposition's assumptions are met, it satisfies this theorem. Assuming that $Y \mid X$ remains unchanged and examining Equation 2, $\beta_H = \beta_K$ and the difference in means simplifies to the sum of the covariate effects. Since Δ depends solely on $Y \mid X$, its value is zero.

We conjecture a reverse direction of Theorem 6, suggesting that under reasonable assumptions, a functional decomposition $\mathcal{L}$ will not misattribute effects if it does not depend on its input distribution. Specifically, one might hope that when allowing K_X to range over the probability space, Equation 22 would imply that $D_\phi \mathcal{L}(\cdot, 0)[\phi](x)$ is constant—in a similar way to the discrete case characterized in Theorem 4. This, in turn, would imply that $\mathcal{L}(\cdot, \phi)$ is invariant under perturbations of concentration; in other words, it is locally constant everywhere and, therefore, $\mathcal{L}$ does not depend on its input distribution in a meaningful way—analogous to Corollary 1.

Conjecture 2. *Under the same regularity conditions as in Theorem 5. If for all $K_X \in \mathcal{P}(\mathcal{X})$ and all $\phi \in \mathcal{D}_{K_X}$, we have*

$$\mathbb{E}_{K_X + \phi} \left[\mathcal{L}(f, K_X, S)(x) - \mathcal{L}(f, K_X + \phi, S)(x) \right] = 0.$$

Then,

$$\mathcal{L}(f, K_X, S)(x) = \mathcal{L}(f, H_X, S)(x), \quad \textit{for all } K_X, H_X \in \mathcal{P}(\mathcal{X}).$$

[3]We require that ϕ be square integrable and that $k(x) + \phi$ be a valid probability density; see Appendix D.2 for details.

While we have not yet fully proved Conjecture 2, we feel it is intuitively sensible: if a decomposition $\mathcal{L}$ does not misattributes effects of transport for *any* distribution, then it must be constant with respect to its input distribution.

Our Examples 1 and 2, together with Section 4.2 and Theorem 6, underscore that popular decomposition methods, such as FANOVA and ALE, are not suitable for explaining differences between two populations under Definition 1, highlighting the need to develop novel decomposition techniques to tackle this problem.

D MATHEMATICAL FRAMEWORK: THE CONTINUOUS SETTING

We now develop the mathematical definitions and assumptions introduced in Appendix C.2 needed to prove Theorem 5 and to work toward Conjecture 2. We also provide a precise definition of an *admissible perturbation* and show that such perturbations exist for any compactly supported density.

D.1 ADDITIONAL NOTATION

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a compact set of possible covariate values, equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{X})$. Let $\mathcal{C}_0(\mathcal{X})$ denote the set of continuous functions on $\mathcal{X}$. In what follows, we focus on probability measures whose densities are continuous, strictly positive, and supported on $\mathcal{X}$. We denote by $\mathcal{P}(\mathcal{X})$ the space of such probability measures; formally,

$$\mathcal{P}(\mathcal{X}) = \Big\{ P : \forall A \in \mathcal{B}(\mathcal{X}),\ P(A) = \int_A p(x)\,\mathrm{d}x,\ p(x) \in \mathcal{C}_0(\mathcal{X}),\ p(x) > 0\ \forall x \in \mathcal{X},\ \int_{\mathcal{X}} p(x)\,\mathrm{d}x = 1 \Big\}.$$

As in Appendix B, we can think of these densities as the Radon-Nikodym derivatives of probability measures that are absolutely continuous with respect to an underlying measure. Since we now focus only on the Lebesgue measure—though our work applies to any underlying measure—we use dx instead of $d\lambda(x)$ for clarity.

We make the following regularity and basic assumptions on the functional decomposition $\mathcal{L}(f, K_X, S)$.

Assumption 2. *The following hold:*

1. ***Continuity**: For any (K_X, S), the map $f \to \mathcal{L}(f, K_X, S)$ is continuous for almost all $f \in \mathcal{M}(\mathcal{X})$.*
2. ***Measurability**: For any (f, S), the map $K_X \to \mathcal{L}(f, K_X, S)$ is Lebesgue measurable for all $K_X \in \mathcal{P}(\mathcal{X})$.*
3. ***Integrability**: The map $(f, K_X, S) \mapsto \mathcal{L}(f, K_X, S)$ belongs to $L^2(\mathcal{X}, \lambda)$, for all $(f, K_X, S) \in \mathcal{M}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \times 2^{[d]}$.*

Where we have used the usual notation $L^2(\mathcal{X}, \lambda)$ to denote the space of square-integrable functions over X with respect to a measure λ, we now omit λ from the notation and write $L^2(\mathcal{X})$ to refer specifically to integration with respect to the Lebesgue measure, making the measure explicit otherwise.

D.2 ADMISSIBLE PERTURBATION FUNCTIONS

To define the admissible perturbation functions mentioned in Appendix C.2, we first need to define Fréchet differentiability.

Definition 4 (Fréchet differentiability; Cheney (2001))**.** *Let $f : D \to Y$ be a mapping from an open set D in a normed linear space X into a normed linear space Y. Let $x \in D$. If there exists a bounded linear map $A : X \to Y$ such that*

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - Ah\|}{\|h\|} = 0,$$

then f is said to be Fréchet differentiable at x, or simply differentiable at x. Furthermore, A is called the Fréchet derivative of f at x.

Definition 5 (Admissible perturbation function)**.** *We say a continuous function $\phi \in L^2(\mathcal{X})$ is an admissible perturbation of the probability measure K_X, if $k(x) + \phi(x)$ is a density of a distribution in $\mathcal{P}(\mathcal{X})$ and has full support everywhere X.*

We denote by $\mathcal{D}_{K_X}$ the set of admissible perturbation functions of K_X: $\mathcal{D}_{K_X} = \{\phi \in L^2(\mathcal{X}) : k(x)+\phi(x) > 0 \text{ and } \int_{\mathcal{X}}(k(x) + \phi(x))\, dx = 1\}$. We show that $\mathcal{D}_{K_X} \neq \{0\}$ for all $K_X \in \mathcal{P}(\mathcal{X})$.

Lemma 8. *For any distribution $K_X \in \mathcal{P}(\mathcal{X})$, there exist an admissible perturbation function different than zero.*

Proof. Let any smooth compactly supported function $\psi \in L^2(\mathcal{X})$. Then, we can take the define the function

$$\tilde{\phi}(x) = \psi(x) - \frac{1}{\lambda(\mathcal{X})} \int_{\mathcal{X}} \psi(y) dy$$

such that $\tilde{\phi}(x) \in \mathcal{W}(\mathcal{X})$, that is, $\int_{\mathcal{X}} \tilde{\phi}(x)\, dx = 0$. To ensure the positivity requirement, we can take a function $\phi(x) = \varepsilon\tilde{\phi}(x)$, for $\varepsilon > 0$, which still is in $L^2(\mathcal{X})$ and integrates to zero. Such $\varepsilon > 0$ must satisfy that for a given density $K_X(x)$,

$$K_X(x) + \phi(x) = K_X(x) + \varepsilon\tilde{\phi}(x) > 0 \iff \varepsilon\tilde{\phi}(x) > -K_X(x),\ \forall\, x \in \mathcal{X}. \tag{23}$$

Whenever $\tilde{\phi}(x) > 0$, Equation 23 is always satisfied. Thus, the only relevant case is when $\tilde{\phi}(x) < 0$, for which Equation 23 is satisfied if and only if

$$\varepsilon < \frac{-K_X(x)}{\tilde{\phi}(x)},\ \forall\, x \in \mathcal{X} \text{ such that } \tilde{\phi}(x) < 0.$$

Or equivalently,

$$\varepsilon \leq \frac{\inf_{x\in\mathcal{X}} K_X(x)}{\sup_{x\in\mathcal{X}} |\tilde{\phi}(x)|},$$

where by assumption the right hand side is strictly positive. Thus $\phi(x)$ is an admissible perturbation function of $K_X(x)$. ■

Note that for any fixed density K_X, we can parameterize the functional decomposition in terms of $\phi(x)$ as follows: $\mathcal{L}(f, \phi, S) = \mathcal{L}(f, K_X + \phi, S) : \mathcal{M}(\mathcal{X}) \times \mathcal{D}_{K_X}(X) \times 2^{[d]} \to L^2(\mathbb{R}^S)$. For this parameterization, in addition to Assumption 2, we need to assume the continuous differentiability of $\mathcal{L}$ as a function of ϕ (see Assumption 3) to ensure that $\mathcal{L}$ is Fréchet differentiable as a map from the Banach space $L^2(\mathcal{X})$ into the Banach space $L^2(\mathcal{X}_S)$ (Zeidler, 1986; Averbukh & Smolyanov, 1967).

Assumption 3. *The map $\phi \to \mathcal{L}(\cdot, \phi(x))$ is continuously differentiable as a map from $L^2(\mathcal{X})$ into $L^2(\mathcal{X}_S)$.*

Under this new assumption, we can linearly approximate $\mathcal{L}(\cdot, \phi)$ around $\phi = 0$ with a linear and bounded functional.

$$\mathcal{L}(\cdot, \phi) = \mathcal{L}(\cdot, 0) + D_\phi\mathcal{L}(\cdot, 0)[\phi] + o(\|\phi\|_{L^2}).$$

Where $D_\phi\mathcal{L}(\cdot, 0)[\phi]$ is the Fréchet derivative of $\mathcal{L}$ with respect to the function ϕ evaluated at the zero function and $o(\|\phi\|_{L^2})$ represents a higher-order functional that vanishes faster than $\|\phi\|_{L^2}$ as $\phi \to 0$. More formally, for any $\delta > 0$, there exists a $\tau > 0$ such that if $\|\phi\|_{L^2} < \tau$, then $|o(\|\phi\|_{L^2})| \leq \delta\|\phi\|_{L^2}$.

Remark 2. *The Fréchet derivative is a linear and bounded functional which operates on functions $\phi \in L^2(\mathcal{X})$. That is, there exist a constant $C > 0$ such that,*

$$\|D_\phi \mathcal{L}(\cdot, 0)[\phi]\|_{L^2} \leq C\|\phi\|_{L^2}.$$

D.3 Proof of Theorem 5

We first show some lemmas that will be useful through the proof of Theorem 5.

Lemma 9. *Given our assumptions, for any $K_X \in \mathcal{P}(\mathcal{X})$ and $\phi \in \mathcal{W}(\mathcal{X})$, the following integrals are finite.*

$$\left| \int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\phi](x))\ \phi(x)\, dx \right| < \infty, \tag{24}$$

$$\left| \int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\phi](x))\, \phi(x)\, k(x)\, dx \right| < \infty. \tag{25}$$

Furthermore,

$$\left| \int_{\mathcal{X}} o(\|\phi\|_{L_2})(x)(k(x) + \phi(x))\, dx \right| = o(\|\phi\|_{L^2}) \tag{26}$$

Proof. $k(x)$ is continuous and compactly supported on X, then by a direct consequence of the extreme value theorem, it is bounded: there exists a $B > 0$ such that $\sup_{x\in\mathcal{X}} |k(x)| \leq B_k < \infty$; by a similar argument, $\sup_{x\in\mathcal{X}} |\phi(x)| \leq B_\phi < \infty$. We first show Equation 24:

$$\begin{aligned}
\left| \int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot; 0)[\phi](x))\, k(x)\, dx \right| &\leq \int_{\mathcal{X}} |D_\phi \mathcal{L}(\cdot; 0)[\phi](x)|\, k(x)\, dx \\
&\leq \left(\int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot; 0)[\phi](x))^2\ dx \right)^{1/2} \left(\int_{\mathcal{X}} k(x)^2\, dx \right)^{1/2} \\
&\leq C \cdot \|\phi\|_{L^2} \cdot B_K \cdot \sqrt{\lambda(\mathcal{X})} \\
&\leq C \cdot B_\phi \cdot B_K \cdot \lambda(\mathcal{X}) < \infty.
\end{aligned}$$

To show Equation 25:

$$\begin{aligned}
\left| \int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot; 0)[\phi](x))\, \phi(x)\, dx \right| &\leq \int_{\mathcal{X}} |D_\phi \mathcal{L}(\cdot; 0)[\phi](x)|\, |\phi(x)|\, dx \\
&\leq \left(\int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot; 0)[\phi](x))^2\ dx \right)^{1/2} \left(\int_{\mathcal{X}} \phi(x)^2\, dx \right)^{1/2} \\
&\leq C \cdot \|\phi\|_{L^2} \cdot B_\phi \sqrt{\lambda(\mathcal{X})} \\
&\leq B_\phi^2 \cdot C \cdot \lambda(\mathcal{X}).
\end{aligned}$$

To show Equation 26: For any $\delta > 0$, there exist a $\tau > 0$ such that if $\|\phi\|_{L_2} < \tau$, then $o(\|\phi\|_{L^2}) \leq \delta\|\phi\|_{L_2}$, thus:

$$\begin{aligned}
\left| \int_{\mathcal{X}} o(\|\phi\|_{L_2})(x)(k(x) + \phi(x))\, dx \right| &\leq \int_{\mathcal{X}} |o(\|\phi\|_{L_2})(x)|(k(x) + \phi(x))\, dx \\
&\leq (B_K + B_\phi) \int_{\mathcal{X}} |o(\|\phi\|_{L_2})|\, dx \\
&\leq (B_K + B_\phi) \cdot \delta\|\phi\|_{L_2} \lambda(\mathcal{X}) \\
&= o(\|\phi\|_{L_2}).
\end{aligned}$$

■

Lemma 10. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a measurable set with finite Lebesgue measure $\lambda(\mathcal{X}) < \infty$. Then, the orthogonal complement of $\mathcal{W}(\mathcal{X})$ in $L^2(\mathcal{X})$ is the space of constant functions on $\mathcal{X}$; that is,*

$$\mathcal{W}(\mathcal{X})^\perp = \left\{ f \in L^2(\mathcal{X}) : f(x) = c, \text{ a.e. on } \mathcal{X} \right\}.$$

Proof. Let

$$V = \left\{ f \in L^2(\mathcal{X}) : f(x) = c, \text{ a.e. on } \mathcal{X} \right\}.$$

We will prove that $\mathcal{W}(\mathcal{X})^\perp = V$ by first showing that $V \subseteq \mathcal{W}(\mathcal{X})^\perp$. Let $f \in V$; then, for any $\psi \in \mathcal{W}(\mathcal{X})$, we have

$$\int_{\mathcal{X}} f(x)\psi(x)\,dx = c \int_{\mathcal{X}} \psi(x)\,dx = 0.$$

It remains to show that $\mathcal{W}(\mathcal{X})^\perp \subseteq V$. Let $f \in \mathcal{W}(\mathcal{X})^\perp$, then $\int_{\mathcal{X}} f(x)\psi(x)\,dx = 0$ for any $\psi(x) \in \mathcal{W}(\mathcal{X})$. In particular, we can take an arbitrary measurable set $A \subseteq \mathcal{X}$ and define

$$\psi_A(x) = \chi_A(x) - \frac{\lambda(A)}{\lambda(\mathcal{X})}$$

where $\chi_A(x)$ is the indicator function over A and λ is the Lebesgue measure. Thus,

$$\begin{aligned} 0 = \int_{\mathcal{X}} f(x)\psi_A(x)\,dx &= \int_{\mathcal{X}} f(x)\chi_A(x)\,dx - \int_{\mathcal{X}} f(x)\frac{\lambda(A)}{\lambda(\mathcal{X})}\,dx \\ \Leftrightarrow \int_A f(x)\,dx = \int_{\mathcal{X}} f(x)\frac{\lambda(A)}{\lambda(\mathcal{X})}\,dx &= \lambda(A)\left(\frac{\int_{\mathcal{X}} f(x)\,dx}{\lambda(\mathcal{X})}\right) \end{aligned} \tag{27}$$

Define $\mu(A) = \int_A f(x)\,dx$, which is a signed measure absolutely continuous with respect to the Lebesgue measure. On one hand, by the Radon-Nikodym Theorem for signed measures (Folland (1999); Theorem 3.8), $f(x)$ is the Lebesgue integrable Radon-Nikodym derivative. On the other, by Equation 27:

$$\mu(A) = \lambda(A) \cdot c, \text{ for any measurable set A} \subseteq \mathcal{X}, \tag{28}$$

where $c = \left(\frac{\int_{\mathcal{X}} f(x)\,dx}{\lambda(\mathcal{X})}\right)$. By the Lebesgue almost everywhere uniqueness of the Radon-Nikodym derivative, we have form Equation 28 and definition of μ that

$$f(x) = c, \text{ a.e. } x \in \mathcal{X}.$$

Therefore, $f \in V$ and $\mathcal{W}(\mathcal{X})^\perp \subseteq V$. ■

We can now proceed to prove Theorem 5, which we hope to use in proving our main Conjecture 2 in future work.

Proof of Theorem 5. By assumption $\mathbb{E}_{K_X+\phi}\left[\mathcal{L}(f, K_X, S)(X) - \mathcal{L}(f, K_X + \phi, S)(X)\right] = 0$, for all $\phi \in \mathcal{D}_{K_X}$. i.e.,

$$\begin{aligned} 0 &= \int_{\mathcal{X}} \left(\mathcal{L}(f, K_X, S)(x) - \mathcal{L}(f, K_X + \phi, S)(x)\right)\left(k(x) + \phi(x)\right)dx, \\ &= \int_{\mathcal{X}} \left(\mathcal{L}(\cdot, 0)(x) - \mathcal{L}(\cdot, \phi)(x)\right)\left(k(x) + \phi(x)\right)dx, \\ &= -\int_{\mathcal{X}} \left[D_\phi \mathcal{L}(\cdot, 0)[\phi](x) + o(\|\phi\|_{L^2}(x))\right]\left(k(x) + \phi(x)\right)dx, \end{aligned}$$

$$\iff 0 = \int_{\mathcal{X}} \left[D_\phi \mathcal{L}(\cdot, 0)[\phi](x) + o(\|\phi\|_{L^2})(x)\right] (k(x) + \phi(x))\, dx. \tag{29}$$

Then, by Lemma 9, we can split the integrals, and rewrite Equation 29 as:

$$\int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\phi](x))\, k(x)\, dx + \int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\phi](x))\phi(x)\, dx = -\int_{\mathcal{X}} o(\|\phi\|_{L^2})(x)(k(x) + \phi(x))\, dx.$$

Since this equation must hold for all $\phi \in \mathcal{D}_{K_X}$, we can proceed as in the proof of Lemma 8. Specifically, let $\phi(x) = \varepsilon\psi(x)$ for sufficiently small $\varepsilon > 0$ and $\psi(x) \in \mathcal{W}(\mathcal{X})$. Furthermore, by Lemma 9, we know the following: $\int_{\mathcal{X}} o(\|\phi\|_{L^2})(x)(k(x) + \phi(x))\, dx = o(\|\phi\|_{L^2})$. Note also that $o(\|\varepsilon\psi\|_{L^2}) = o(\varepsilon\|\psi\|_{L^2}) = o(\varepsilon)$ since $\|\psi\|_{L^2} < \infty$, then the above equation simplifies to:

$$\int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\varepsilon\psi](x))\, k(x)\, dx + \int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\varepsilon\psi](x))\varepsilon\psi(x)\, dx = o(\varepsilon).$$

Where by $o(\varepsilon)$, we mean a constant that goes to zero faster than ε. By the linearity of the Fréchet derivative, we can take ε out of the operator, divide by it, and since $\frac{o(\varepsilon)}{\varepsilon} = o(1)$, we obtain:

$$\int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\psi](x))\, k(x)\, dx + \varepsilon \int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\psi](x))\psi(x)\, dx = o(1).$$

Taking $\varepsilon \to 0$, we get that the first integral is equal to zero:

$$\int_{\mathcal{X}} (D_\phi \mathcal{L}(\cdot, 0)[\psi](x))\, k(x)\, dx = \mathbb{E}_{K_X}[D_\phi \mathcal{L}(\cdot, 0)[\psi](x)] = 0, \quad \text{for all } \psi \in \mathcal{W}(\mathcal{X}). \tag{30}$$

■