COTENABILITY AND CAUSALITY: EXPLAINING FEATURE IMPORTANCE USING SHAPLEY VALUES

Tony Liu

May 10, 2020

ABSTRACT

SHAP has become a popular post-hoc model interpretation method because of its flexibility and desirable theoretical axioms. However, in practice calculating Shapley values either assumes independence among model features which violates any correlation among features (which we define as *cotenability*), or assumes a conditional probability structure which does not capture causality and fails to satisfy key axioms. Here we present a graphical interpretation of Shapley values which clarifies assumptions made during Shapley value calculations. We then consider grouping features as a framework to make causal Shapley values cotenable, and explore their properties in both simulated and real data. The goal is to extend Shapley feature importances so that both cotenability and causality are captured, ultimately increasing interpretability of these explanations.

1 Motivation

As complex machine learning models are increasingly used in high-stakes settings such as criminal justice and healthcare, being able to explain model output is critical. Post-hoc additive explanation methods like LIME ([15]) are an attractive class of interpretability algorithms because of their flexibility in explaining individual predictions regardless of the underlying model and their seemingly simple and intuitive interpretation. With the SHAP framework unifying additive explanation methods with concepts from game theory ([12]), these additive algorithms can achieve theoretic axioms that are desirable properties for interpretable explanations, further evidence for their usefulness.

However, in practice the calculation of Shapley values is under-specified. Empirical Shapley values can fail some desirable theoretic axioms depending on different implementation methods [16]. The essence of the issue centers on whether to use a conditional or marginal distribution to sample features when generating an explanation for a feature. The recent discussion surrounding this choice([1, 2]) has highlighted two desiderata of model explanations that have to be traded off:

1. We want our explanation models to capture the causal effect of a feature on model output (satisfied in SHAP by using marginal distributions to sample features)
2. We want our explanation models to respect feature correlation structure in order to produce what we call *cotenable* explanations (satisfied in SHAP by using conditional distributions to sample features)

This trade-off has been implicitly explored by work studying the axiomatic properties of different Shapley value formulations. Our goal here is to clarify the trade-off using graphical interpretations of Shapley values, and explore how we can move towards satisfying both desiderata by making additional assumptions about the relationships between model features.

Related Work

The problem of Shapley values failing axiomatic properties depending on their formulation has been examined in detail in [16], including worked examples and proofs of correctness. We review the differences between their *Conditional Expectation* and *Random Baseline* Shapley value formulations, which respectively correspond to conditioning and not

conditioning on features, in Section 2.1. [9] provide a causal justification for the same arguments in distinguishing conditional and marginal distributions of features, framing them as the difference between *observational* vs. *interventional* distributions. [20] similarly take a graphical approach in justifying causal interpretations of partial dependence plots. Both [16] and [9] note that the original formulation of SHAP was ambiguous in its choice of feature distributions, without clear justification of using one over the other. The most recent work by the authors of SHAP have incorporated these critiques of their original formulation into their most recent implementation of the TreeSHAP algorithm [11].

In terms of desirable properties of model explanation, [13] propose a taxonomy of non-causal and causal interpretability where our desiderata loosely correspond to their categories of causal interpretability for "model-based interpretations" (desiderata 1), and causal interpretability for "verifying causal relationships discovered from data" (desiderata 2). Note however that in our desiderata, respecting feature correlation structure does not necessitate a causal interpretation – features may simply be correlated with one another, or grouped based on preferences of the practitioner (e.g. aggregating individual features into a composite metric such as a risk score). [3] modify the KernelSHAP algorithm to better respect feature correlations (desiderata 2), though as [9] note this work approximates Conditional Expectation Shapley, which is known to fail some Shapley axioms.

Aims

Our work presented here most closely follows the arguments presented in [9] and [20], building off their graphical interpretation of "interventional" Shapley values and extending them to accommodate contenable feature relationships. To do so, we make an assumption that features come in groups, either provided by underlying causal knowledge or correlations present within the data. We then report on how SHAP values can be interpreted in groups as well as explore their empirical properties through simulations and a case study.

The rest of this work is organized as follows:

> Section 2 gives the graphical foundations of both model-based causality (desiderata 1) and cotenability (desiderata 2) in Shapley values.
>
> Section 3 explores the behavior of Shapley values when features follow a group correlation structure through simulated data.
>
> Section 4 explores the interpretability benefits of grouping Shapley values through a case study using a real-world medical dataset.
>
> Section 5 concludes with discussions on limitations and future work.

2 Graphical Framework

2.1 Shapley Value Preliminaries

Here we review notation used for calculating Shapley values for model feature explanation. More details can be found in [12, 16, 11]. Given a collection of N features, a model $f : R^{|N|} \to R$, an example x, and a *set function* $v_{f;x} : 2^{|N|} \to R$ parameterized by f and x, we are interested in calculating the Shapley value $_i$ for feature i:

$$_i(f;x) = \sum_{S \subseteq N \setminus i} \frac{|S|!\ (|N| - |S| - 1)!}{|N|!} \left[v_{f;x}(S \cup \{i\}) - v_{f;x}(S) \right] \tag{1}$$

where the sum is taken over all possible subsets S of the feature set $N \setminus i$.

Alternatively, we can define the Shapley value for feature i as:

$$_i(f;x) = \sum_{R \in \Pi(N)} \frac{1}{|N|!} \left[v_{f;x}(R_i \cup \{i\}) - v_{f;x}(R_i) \right] \tag{2}$$

where $\Pi(N)$ are all possible orderings of the features N, and R_i is the set of all features that come before feature i in ordering R.

We note a few of the most useful Shapley value axioms:

efficiency:

$$f(x) = \quad ;(f) + \sum_{i} \quad i(f;x)$$

In words: the sum of the Shapley values for an example x equal the original model output $f(x)$.

consistency:

$$v_{f';x}(S) \quad v_{f';x}(S \setminus i) \quad v_{f;x}(S) \quad v_{f;x}(S \setminus i);\ \forall f; f^0; S \quad N \quad =) \quad \ i(f^0;x) \quad \ i(f;x)$$

In words: if a model changes such that some feature's contribution increases or stays the same, then the example's Shapley value for the feature does not decrease.

missingness:

$$v_{f;x}(S \ [\ i) = v_{f;x}(S);\ \forall S \quad N \quad =) \quad \ i(f;x) = 0$$

In words: features that do not impact the model's output have a Shapley value of 0.

In both formulations of the Shapley values above, we need to define the set function $v_{f;x}(S)$, as the axioms may or may not be satisfied depending on the choice of set function. We review two popular definitions of the set function below. Let S be the set of features we are interested in, X_S the collection of random variables associated with features in S, x_S a particular setting of the variables in X_S, and $B = N \setminus S$.

Conditional Expectation Shapley (CES) Set Function

The conditional expectation Shapley (CES) set function is defined as:

$$v_{f;x}(S) = E_{X_B | X_S}[f(X_N) | X_S = x_S] \tag{3}$$

where the expectation of the model output f is taken over a conditional distribution of X_B given a particular setting x_S of X_S from x.

As proved in [16], CES does not satisfy the Shapley axioms of *missingness* and *consistency*, unless additional assumptions about the feature distribution are made.

Interventional Shapley (IS) Set Function

The interventional Shapley (IS) set function (also called Random Baseline Shapley by [16]) is defined as:

$$v_{f;x}(S) = E_{X_B}[f(x_S; X_B)] \tag{4}$$

where the expectation of the model output f is taken over a marginal distribution of X_B, setting the variables in S to their corresponding values x_S.

As proved in [16], IS *does* satisfy the Shapley axioms of *missingness* and *consistency*.

Remarks

Note that implicit in both of these definitions of the set function is the knowledge of some probability distribution over the features. In practice these are often approximated by the empirical distribution of the training data.

2.2 Shapley Values as Graphs

We now give a graphical interpretation of the CES and IS set functions, first restating the arguments presented in Section 3 of [9] and then building upon them. For our worked example, we will consider a model f that takes three features as input $X_1; X_2; X_3$, representing the model output random variable as $Y = f(X_1; X_2; X_3)$. In addition, we also consider a latent variable Z that is a common parent of all the input features $X_1; X_2; X_3$ (Figure 1a). Later on, we will refer to Z as representing the "real world" when considering correlation structure among features. In the following examples, we are interested in the value of the set function for $f X_1 g$.

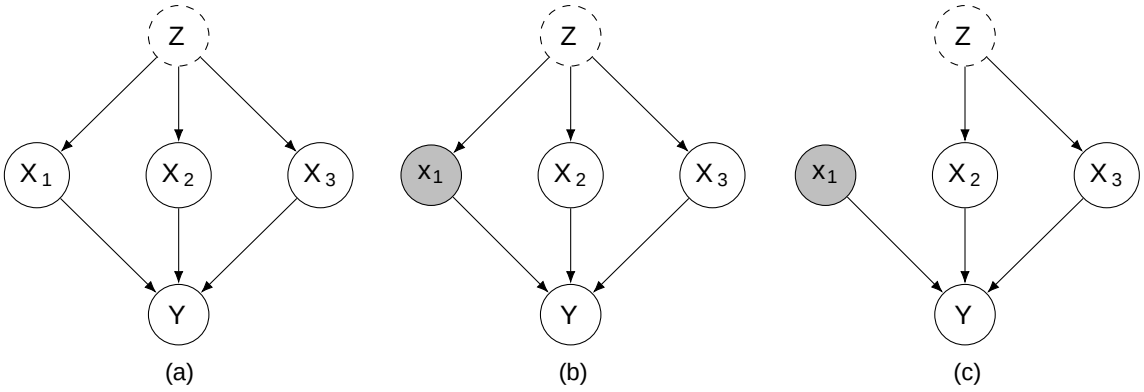


Figure 1: Graphical representation of model input/output and Shapley set functions. The dotted nodes Z represent latent variables that are not observable. Solid border nodes represent variables that we do observe, namely model input $(X_1; X_2; X_3)$ and model output Y. Shaded in nodes represent particular settings of X_1. Figure 1b represents the graphical state of the conditional expectation Shapley set function calculation for $\{X_1 = x_1\}$, while Figure 1c represents the graphical state of the interventional Shapley set function for the same set.

CES Graphically

Expanding Equation 3 for $v_{f;x}(f X_1 g)$, we have:

$$\begin{aligned} v_{f;x}(S) &= E_{X_B|X_S}[f(X_N)|X_S = x_S] \\ v_{f;x}(f X_1 g) &= E_{X_2;X_3|X_1}[f(X_1; X_2; X_3)|X_1 = x_1] \\ &= \int_{x_2}\int_{x_3} f(x_1; x_2; x_3)\, p(x_2; x_3|x_1)\, dx_2 dx_3 \end{aligned}$$

As noted in [9], the resulting state of our graphical representation is the observational distribution over the features (Figure 1b), as we are changing the distributions of X_2 X_3 by setting $X_1 = x_1$ because of the shared dependence Z. Using causal language, CES fails to capture the causal effect of setting $X_1 = x_1$ on Y because it does not control for the confounder Z.

IS Graphically

Expanding Equation 4 for $v_{f;x}(f X_1 g)$, we have:

$$\begin{aligned} v_{f;x}(S) &= E_{X_B}[f(x_S; X_B)] \\ v_{f;x}(f X_1 g) &= E_{X_2;X_3}[f(x_1; X_2; X_3)] \\ &= \int_{x_2}\int_{x_3} f(x_1; x_2; x_3)\, p(x_2; x_3)\, dx_2 dx_3 \end{aligned}$$

The resulting state of our graphical representation is the interventional distribution over the features (Figure 1c), as we are breaking the dependence between x_1 and $X_2; X_3$ by "forcing" $X_1 = x_1$. Again as noted by [9], we can use the backdoor criterion [14] to show that $v_{f;x}(f X_1 g)$ under IS is equivalent to the causal quantity $E[Y|do(X_1 = x_1)]$:

$$\begin{aligned} v_{f;x}(f X_1 g) &= \int^{x_2}\int^{x_3} f(x_1; x_2; x_3)\, p(x_2; x_3)\, dx_2 dx_3 \\ &= \int_{x_2}\int_{x_3} f(x_1; x_2; x_3)\, p(x_2; x_3|x_1)\, dx_2 dx_3; \text{ by independence} \\ &= E[f(X_1; X_2; X_3)|X_1 = x_1]; \text{ under the modified interventional graph (Figure 1c)} \\ &= E[Y|X_1 = x_1] = E[Y|do(X_1 = x_1)] \end{aligned}$$

IS is Model-Based Causality

Thus, the IS set function captures the causal effect of setting $X_1 = x_1$ on Y, which is a important and useful property for a feature explanation: we want to know how forcing a feature to be a certain value affects the model output, not simply observe the model output when the feature happens to be that value. We call any feature explanation technique that captures the causal effect of model inputs on the model output *model-based causality* explanation (desiderata 1 in Section 1).

2.3 Defining Cotenability

The other desirable feature explanation property we want to explore is the property of respecting any dependencies among features (desiderata 2 in Section 1). To give an example, suppose we are interpreting a linear regression for probability of mortality that takes as input an individual's height, weight, and body mass index (BMI). The regression coefficients for each feature are interpreted as "the effect on the model output when increasing the feature value by one unit, holding all other features constant." However, because of the complete dependence of BMI on height and weight, the interpretation of the coefficient for BMI loses its meaning because it is not possible to change BMI while holding both height and weight constant.

The idea of respecting dependencies has been explored in the philosophy of conditional logic: a conditional $a \Rightarrow b$ is true if b follows from a, together with a set of statements C such that it is not the case that $a \Rightarrow \neg c$, for any $c \in C$ [5]. The statements in C are called *cotenable premises*, and this framework can be mapped back onto our regression interpretation example: loosely speaking, a is increasing BMI by one unit, b is the change in model output, and C contains the premises that height is held constant (c_1), weight is held constant (c_2), and BMI is equal to weight over height squared (c_3). The conditional $a \Rightarrow b$ is not cotenable because if a is true, one of c_1, c_2, c_3 must be violated: $a \Rightarrow \neg c_i$ for at least one c_i.

We refer to this idea of an explanation respecting feature dependencies as *cotenability*. In terms of the graphical framework we have presented, a cotenable explanation respects the dependencies of the model inputs X (but not their parents Z) (Figure 2a). Note that both the choice of cotenable premises as well as the graphical structure Z must be supplied as input, and we can think of Z as the set of cotenable premises we wish to impose in our model explanation. The graphical structure of model inputs can reflect causal knowledge of "the real world," but do not necessarily have to – they can also reflect correlation among features or even dependencies among features that the practitioner may wish to impose.

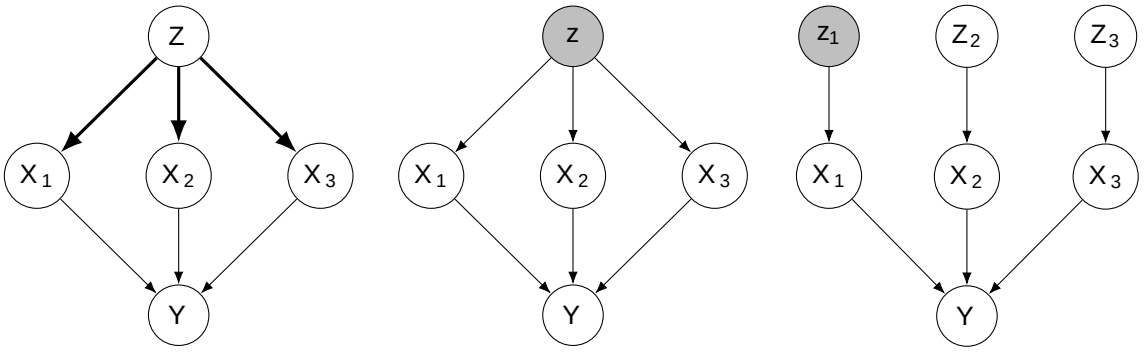


Figure 2: Graphical representation of cotenability concepts. The thick edges in Figure 2a indicate the dependencies that must be preserved in order for an explanation to be cotenable. Figure 2b represents a feature explanation that is both cotenable and causal. Figure 2c shows the assumption of feature independence needed for IS set functions to be contenable. Note that we assume that the dependencies between the features provided by Z are known (no longer dashed).

2.4 Cotenability and Causality

With the graphical definition of cotenability in hand, we see that CES set functions are cotenable while IS set functions are not, as IS removes any dependencies of the variables through the causal do-operator. On the other hand, IS by definition captures model-based causality while CES does not: see Section 3.3 and [Example 3.3 of 16]

for examples of how CES fails to capture the causal effect of features on the model output due to correlation among features.

An ideal explanation would have both feature cotenability and model-based causality, where an explanation could capture the causal effect of intervening Z_n on $E[Y|do(Z = z)]$ (Figure 2b). As noted above, some knowledge of the relationship between the parent features and the X model inputs must be provided in order to perform this cotenable intervention.

In practice, an assumption of feature independence is made in order to facilitate interpretation of Shapley values, as noted in Figure 2 and subsequent discussions of [9] (reproduced in our notation in Figure 2c): "it is necessary to emphasize that we are not talking about the causal relation between any features in the real world outside the computer, but only about causality of this technical input/output system". It is natural to next explore a relaxation of this assumption, namely that features have dependencies which would bring us closer to feature cotenability while still maintaining a causal interpretation.

Grouping Features

For the rest of this work, we will use the IS set function and make an assumption that features group (see Figure 3 in our simulation experiments for an example). These groups may have a causal interpretation, or they may simply reflect collections of features the practitioner wishes to consider together. Grouping features may be useful beyond respecting cotenability, as they can increase the interpretability of the feature explanations provided. In practice, features as input into machine learning models often already come in groups, such as the aggregation of individual image pixels into so-called "super pixels" or the aggregation of cardiovascular risk indicators into a single risk score.

We will first explore the robustness of Shapley values under group correlation structure via simulation, followed by a case study using the NHANES I mortality dataset that illustrates the subjective interpretability benefits of grouping features.

3 Grouped SHAP Simulations

We want to understand the behavior of Shapley values in the presence of grouped feature correlation structure. To this end, we utilize the experimental setup presented in [17], where they investigated the stability of feature importance scores across groups of features by simulating the structure of gene microarray data.

3.1 Experimental Setup

Data generation process

For a given sample size n, we generate three independent length n prototype vectors U, V, W corresponding to groups of features G_1, G_2, R respectively, from a mixture of two Gaussians with equal probabilities: $N(0, 0.1)$; f_1 $N(1, 0.1)$. Given a noise level p and a group size $|G_1|$, we then generate the features for group G_1 by randomly selecting p fraction of the components of U and adding Gaussian noise $N(0, 0.5)$, repeating for features $X_1, \ldots, X_{|G_1|}$. The same noising process is used to generate features corresponding to V and W. Thus, features within each group are correlated with each other (controlled by p) while they are uncorrelated with features from other groups.

The outcome variable y is a linear rule:

$$y = 1[5U + 4V \quad \overline{(5U + 4V)} + \quad > 0] \tag{5}$$

Where $N(0, 0.1)$ and $\overline{(5U + 4V)}$ is the mean of $(5U + 4V)$. Note that W does not carry any signal, and that U is more important for predicting the outcome than V (Figure 3). Again replicating [17] we let $n = 100$, $|R| = 50$, $|G_1| + |G_2| = 200$ and vary both $|G_1| = f1, 10, 100, 190, 199g$ and $p = f0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7g$. As $|G_1| + |G_2|$ is fixed, we are effectively varying the ratio of the number of features in groups G_1 and G_2. With this setup we will be able to measure how stable feature importances are across group sizes and noise levels.

Classifier and SHAP grouping method

We consider sums of Shapley values as a first-pass aggregation method for grouped features which we will define as (G) for a group of features G. Due to the efficiency axiom, we hypothesize that sums of Shapley values should be able to capture group-level feature importance, provided that the correct groupings of variables are chosen. In this simulation, we know exactly which features are grouped with each other, giving us "perfect" groupings. We use decision

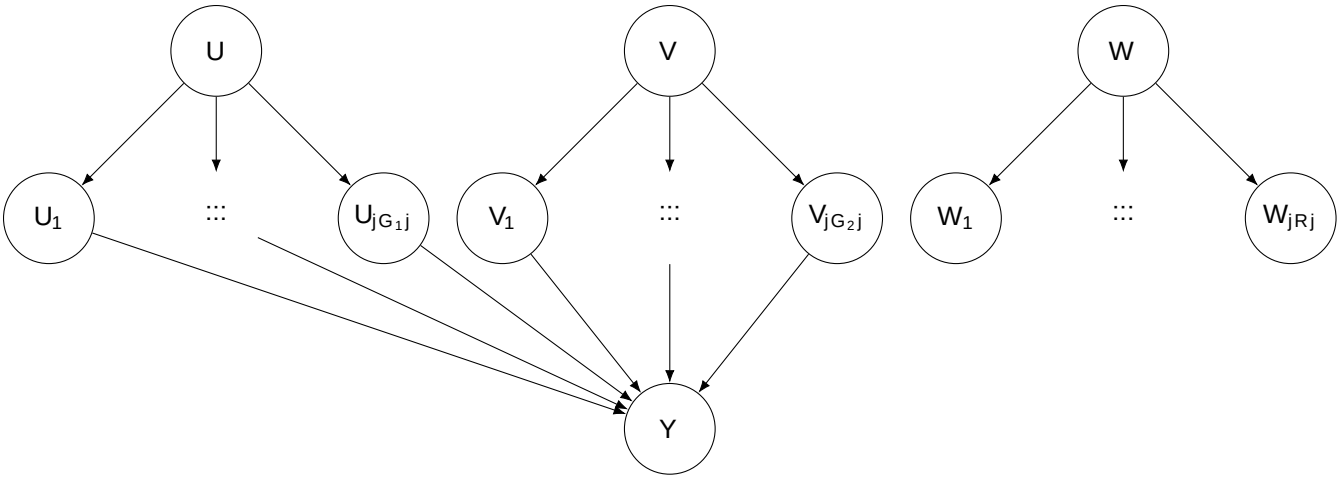


Figure 3: Graphical representation of our group correlation simulation. Features $U_i; V_j; W_k$ are generated from their corresponding prototype vectors $U; V; W$. W is not correlated with the outcome. Note that in this figure Y represents the true output function, as opposed to a learned machine learning model of the output $\hat{Y}$.

tree classifiers and the ISO set function TreeSHAP implementation for calculating Shapley values [11]. We use a 50/50 testing/training split and calculate Shapley values on the test data.

3.2 Results

Grouped Shapley values preserve relative feature importance

First, we see that in the absence of noise ($p = 0$), the Shapley value sums recover the Shapley values for the individual prototype vectors (Figure 4a and 4b, upper left), regardless of the size of group G_1. Furthermore, even after applying up to $p = 0.7$ noise, we see that $\phi(G_1) > \phi(G_2)$ across all $|G_1|$. Unlike the feature importance measures considered in [17], sums of Shapley values preserve the relative feature importance of the two groups G_1 and G_2 without having to cluster features before model training. This is key for interpretation as we want the more important feature groups to have higher Shapley value sums than less important feature groups, regardless of the feature group sizes.

Grouped Shapley values distort in the presence of noise

However, though relative ordering of the feature groups is preserved, the absolute sums of Shapley values change for certain group sizes as a function of the noise level. When $|G_1| = |G_2| = 100$ (grey lines in Figure 4a), the sums of the Shapley values correspond to the true Shapley values for each group, regardless of the noise level (Figure 4b). As $|G_1|$ decreases relative to $|G_2|$ however, we see the Shapley values for G_1 decrease as more noise is applied. Possible explanations for this result is increased variance of the group feature importance as the number of features within a group decreases and the decreasing gap between the importance of G1 and G2 as the noise level increases (Figure A.1). Additionally, we see that the total sums of the Shapley values across all three groups is the same (perhaps due to the efficiency properties of Shapley values), resulting in the Shapley value sum of G_2 increasing beyond the Shapley value for its prototype vector as the Shapley values for G_1 decrease. In these situations, it appears that sums of Shapley values "over-weigh" less important features in that are in larger groups as noise is applied, which seemingly violates the consistency axiom. Within the context of our simulation, Shapley value sums distort and behave somewhat surprisingly in the presence of noise, with further investigation needed into exactly why the Shapley value sums deviate from the prototype Shapley values.

4 Grouped SHAP Case Study: NHANES I

We also want to explore how grouping SHAP values can aid in the interpretation of feature importance. To do so, we examine the relationship between mortality and medical data from the National Health and Nutritional Examination Survey (NHANES I) dataset [7], building models to predict survival time and again using sums of Shapley values for model explanation.

(a)

(b)

Figure 4: Simulated Shapley sum results for varying feature group sizes and noise levels. a) depicts the mean absolute value of sums of Shapley values within each feature group as both a function of $|G_1|$ and the noise level σ. b) depicts the Shapley values for the prototype vectors G_1, G_2, R. Error bars displayed are bootstrapped 95% confidence intervals over 100 runs.

4.1 Experimental Setup

Data and feature correlation structure

There are 9,932 individuals in the NHANES I dataset with 27 predictors for mortality, including socioeconomic status, demographics, and nutritional lab readings. To understand the relationship between features, we first compute Pearson correlations between all of the predictors, and then perform a hierarchical clustering on the resulting correlations (Figure A.2). In practice, physicians often combine lab results or vital signs into a single composite measure for ease of interpretability and clinical threshold assessment, such as risk scores for diabetes [10] and cardiovascular disease [19]. We emulate this by providing qualitative groupings of the features that correspond to concepts that influence mortality

(Figure 5). Concepts such as "malnutrition" and "iron de ciency" are arguably more interpretable to a wider audience than the individual serum lab readings. Though features related to blood pressure ("hypertension") are nested within the "cardiovascular risk" block, we break them out as a separate group as they are a real-world illustration of feature cotenability: pulse pressure is the difference between systolic and diastolic blood pressure. We then use these groupings to inform our Shapley value sum aggregation.

Figure 5: Hierarchical clustering of NHANES I features with qualitative groupings. Cells in the clustermap are colored based on their Pearson correlation. Group labels are from a qualitative assessment of the feature group's effect on mortality by a medical student.

Model construction and feature generation

In addition to the interpretability bene ts of grouping features, we would like to establish some notion of feature importance stability when summing Shapley values. We would like to show that summed Shapley values produce the same relative (and ideally absolute) feature importances when compared to Shapley values of apriori grouped features. To explore grouped feature importance stability, we build two random forest models for predicting survival time: one which uses the individual features, and another which uses apriori aggregated features, standardized and unit weighted within our chosen groups. We do not include gender and age in our feature groupings because of their presumed large effect on mortality so as to not distort the importance of groups. Our nal feature groupings are "cardiovascular risk" sans blood pressure features, "blood pressure," "iron," "SES," and "nutrition," corresponding to the groups in Figure 5.

Missing features are mean imputed, with an additional indicator column for missingness included in the feature matrix. The default parameter settings in scikit-learn are used for training the random forest models, and a 80/20 train/test split is used. All model training details can be found in our code repository github.com/tliu526/group-shap .

Shapley values are calculated using TreeSHAP with the training data used for the interventional feature sampling. We report results on Shapley values for test data over ten sub-samples of size 500 to generate error bars.

(a) Individual Shapley Density Plot

(b) Individual Shapley Feature Importance

(c) Sums of Individual Shapley Density Plot

(d) Sums of Individual Shapley Feature Importance

Figure 6: Individual and summed Shapley values for NHANES features Figures 6a and 6c on the left are representative Shapley value density plots for one of the test set subsamples. Figures 6b and 6d on the right show the average impact on model output magnitude, with error bars indicating bootstrapped 95% con dence intervals over ten test set subsamples.

4.2 Results

Sums of Shapley values are qualitatively more interpretable

We see some clear interpretability bene ts to grouping features through summed Shapley values (Figure 6). First, when compared to the individual Shapley values, there are less features to consider which reduces cognitive load. Additionally, the group labels can transport more information to a wider audience. For example, "cardiovascular risk," which consists of BMI, serum cholesterol, and race, can be a meaningful label to individuals without medical expertise, and even medical professionals bene t from aggregations of multiple features into a single assessment or risk score [18]. Note that since we simply sum Shapley values, the individual Shapley feature importances are still available for model explanation if needed.

The "blood pressure" feature group is a notable instance of how Shapley value sums can aid interpretation. When treated as individual features, we see that systolic blood pressure, diastolic blood pressure, and pulse pressure all are less importance than the individual poverty index feature (Figure 6b). However, systolic and diastolic blood pressure are highly correlated as well as causally related to one another, and pulse pressure is completely dependent on the other two blood pressure readings as it is the difference between them. It is therefore more natural to consider them together

as a single feature. The aggregated "blood pressure" feature group then appears as the third most important feature in the model for predicting survival time behind age and gender (Figure 6d), which makes intuitive sense as hypertension is a contributing factor to some of the leading causes of death in the United States [6]. Thus, summing Shapley values according to pre-defined groups of variables that carry either semantic meaning, such as "CV risk," or respect dependencies present within the data and the real world, such as "blood pressure," increases overall interpretability of our model explanations.

Sums of Shapley values are comparable to Shapley values of apriori grouped features

We also find that sums of individual Shapley values ("sum-SHAP") and Shapley values of apriori grouped features ("apriori-SHAP") produce similar feature importances (Figure A.3). The two models trained have effectively identical test set performance (0.8083 vs 0.8084 risk prediction C-statistic for individual and grouped features) and produce identical rankings for the top four most important features: age, gender, blood pressure, and SES. Though the ordering of the subsequent features do not exactly align between sum-SHAP and apriori-SHAP, the feature importances are similar in magnitude with the exception of nutrition, and produce qualitatively similar density plot distributions on the same test subsample. For example, in both sum-SHAP and apriori-SHAP we see that lower levels of CV risk generally corresponds to lower model output (increased survival time), with two high CV risk individuals breaking the trend (Figures A.3a, A.3c). Despite not having "perfect" (independent) groupings like our simulation experiment, we see that sum-SHAP and apriori-SHAP give reasonably comparable results, providing some evidence for the feature importance stability of summed Shapley values.

5 Discussion

Here we have examined the trade-off between capturing causal effects of features on model output and respecting cotenability of the features for model explanation and feature importance. We formalized this trade-off in Shapley values through a graphical interpretation of the Conditional Expectation Shapley (CES) and Interventional Shapley (IS) set functions, showing that CES respects cotenability but not causality while IS respects causality but not cotenability. We then explored how we could move towards satisfying both cotenability and causality by using IS under the assumption that features come in pre-defined groups. We examined the behavior of grouped Shapley in the form of summed IS values through simulations and real medical data. Though further work is needed to determine the precise mathematical properties of summed Shapley values in the context of the original Shapley axioms, our experiments have demonstrated the benefits of grouping Shapley values: under reasonable conditions and an assumed grouped structure, sums of Shapley values preserve relative feature importance while also increasing interpretability.

5.1 Limitations

There are a number of limitations in both our graphical interpretation and experimental design we wish to highlight.

First, as we define cotenability in the context of our graphical models as the parents Z to the model input features X, there is flexibility in how the directed edges from Z to X may be interpreted due to graphical models being inherently under-specified. Though we note that this flexibility can be beneficial depending on the intent of the practitioner (Z can reflect causal knowledge, or just the correlation structure within the data), there is potential for the connections between Z and X to be misinterpreted as reflecting causal "ground truth" much like how correlational studies may be misleadingly couched in causal language. Any practitioner using this framework must be explicit in the assumptions underlying the definition of Z and what groupings among model features represent. Additionally, we acknowledge that assuming features come in groups, effectively a block correlation structure, is a simplifying assumption that may not hold in practice. In fact, we see in the NHANES I dataset that there is likely a hierarchical relationship among the features within our "cardiovascular risk" block (Figure 5).

By using the framework provided in [17] for our simulated data we are able to compare our results to theirs, demonstrating the benefits of Shapley value sums over the models and feature importance measures they consider. However, several aspects of the simulation could have been designed differently to better examine properties specific to Shapley values. The most prominent aspect that can be improved is the fact that the decision rule is linear in the prototype vectors. Due to the efficiency axiom, the Shapley value of a feature $X_i = x_i$ in linear models decomposes into a function of x_i and the univariate expectation $E[X_i]$ (Section B.1). This may make it difficult to determine precise properties of grouped Shapley values, as under a linear model the joint expectations $E[X_i; X_j]$ factor into $E[X_i]E[X_j]$ even when X_i and X_j are members of the same feature group. We aim to improve upon these simulations in future work.

5.2 Future Work

Group SHAP set function modification

Though the key assumption we make in order to satisfy contenability is that features come in pre-defined groups, we only take a simple first step exploring this in practice by summing Shapley values. A natural extension would be to modify the Shapley value calculation to only use coalitions that respect the pre-defined groups. Let C be a partition of the set of features N and $c \in C$ be a set of cotenable features. We can then define a modified Shapley value (borrowing notation from Equation 1):

$$\phi_c(f;x) = \sum_{S \subseteq C \setminus c} \frac{|S|!\ (|N| - |S| - 1)!}{|N|!} \left[v_{f;x}(S \cup c) - v_{f;x}(S) \right] \tag{6}$$

Where S is all possible subsets of the *sets* in $C \setminus c$. This formulation loses the key benefit of preserving individual Shapley values that is present when simply using sums as a grouping method, but could be worth exploring as a method for calculating grouped Shapley values. Another alternative formulation of the Shapley value calculation would be use the definition of Shapley values from Equation 1 but only consider subsets S that can be decomposed into a union of the contenable feature sets in C.

Consideration of alternative Shapley baselines

As discussed in [16], *Baseline Shapley* is another set function that satisfies all Shapley axioms, and only requires a definition of a baseline function as opposed to a full data distribution of the model features. The authors suggest that this formulation could be useful when particular features "immutable" as part of a model explanation. For example, when examining a model explanation for why someone was rejected for a loan, race should be part of the baseline function as it is not useful for a explanation to consider implausible interventions, as one cannot realistically change their race. This idea could be incorporated into our framework of contenability by forcing every group of variables to include any immutable features, such as age and race in our NHANES I case study. As noted above in Limitations, careful interpretation of the feature groups is needed especially since the sets of features are no longer disjoint, but this modification is also worth considering to further improve interpretability in certain situations.

Examining model stability in the context of feature explanation

Questions have been raised about the "stability" of model explanations provided by methods such as LIME and SHAP [4]. As an aside, we observed throughout our experiments that the stability of Shapley values depended on the "stability" of the underlying model – random forests surprisingly produced feature importances that varied widely as a function of the group sizes and initialization for our simulated classification task in Section 3. Another interesting line of research would be to formalize the notion of feature explanation stability in the context of model stability, including robust definitions of what it means for an explanation to be "stable."

5.3 Conclusion

Here we have explored what we believe to be a fundamental trade-off in model explanation and feature importance desiderata: causality vs. cotenability. As noted by the authors of SHAP, being "true to the model" (causality) vs. "true to the data" (cotenability) cannot be satisfied in general when there is arbitrary correlation among features [2]. Thus, additional domain expertise is often required for proper model explanation [20], much like traditional causal inference problems; see [8] for a recent work that highlights the benefits of domain expertise. Many interpretability methods to date only push towards one of our desiderata and often do not consider the role of domain expertise, so more progress is needed to bridge the gap. We hope that our work here on grouped Shapley values provides clarification to the causality vs. cotenability trade-off as well as a step towards a potential solution, so that future explainable AI methods can come closer to "cotenable causality."

Acknowledgements

Thanks to Lyle Ungar for helpful discussions around Shapley values and grouped feature interpretation, as well as for a great semester leading the XAI course. Thanks to everyone in the course for fun and stimulating discussions throughout the semester. Thanks to Christina Chen for providing interpretation of feature groups for the NHANES I dataset.

References

[1] Github issue: Possible problem with using conditional vs marginal expectation for dropped features in tree shap? https://github.com/slundberg/shap/issues/882.

[2] Github issue: Treeshap is not exact. https://github.com/christophM/interpretable-ml-book/issues/142.

[3] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

[4] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[5] Horacio Arlo-Costa. The logic of conditionals. https://plato.stanford.edu/entries/logic-conditionals/, 2007.

[6] CDC. Facts about hypertension. https://www.cdc.gov/bloodpressure/facts.htm.

[7] CDC. Nhanes i. national health and nutrition examination survey. https://wwwn.cdc.gov/nchs/nhanes/nhanes1/Default.aspx, 1974.

[8] Efstathios D Gennatas, Jerome H Friedman, Lyle H Ungar, Romain Pirracchio, Eric Eaton, Lara G Reichmann, Yannet Interian, José Marcio Luna, Charles B Simone, Andrew Auerbach, et al. Expert-augmented machine learning. *Proceedings of the National Academy of Sciences*, 117(9):4571–4577, 2020.

[9] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causality problem. *arXiv preprint arXiv:1910.13413*, 2019.

[10] Jaana Lindström and Jaakko Tuomilehto. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes care*, 26(3):725–731, 2003.

[11] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):2522–5839, 2020.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[13] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Ragliny, and Huan Liu. Causal interpretability for machine learning–problems, methods and evaluation. *arXiv preprint arXiv:2003.03934*, 2020.

[14] Judea Pearl. *Causality*. Cambridge university press, 2009.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[16] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. *arXiv preprint arXiv:1908.08474*, 2019.

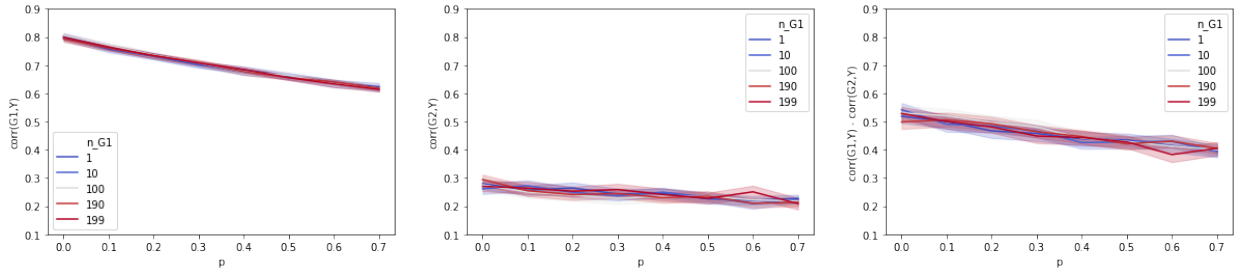
[17] Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

[18] Berk Ustun and Cynthia Rudin. Optimized risk scores. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1125–1134, 2017.

[19] Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

[20] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, pages 1–10, 2019.

A Additional Figures and Tables



(a) (b) (c)

Figure A.1: G_1 **and** G_2 **correlations with** Y **as noise varies.** A.1a and A.1b show the mean correlation between features in G_1 and G_2 respectively and the outcome Y as the noise level p increases. A.1c shows the difference in these correlations. We see that the difference in the correlations decreases as the noise level increases, showing that the relative importance of G_2 is increasing.

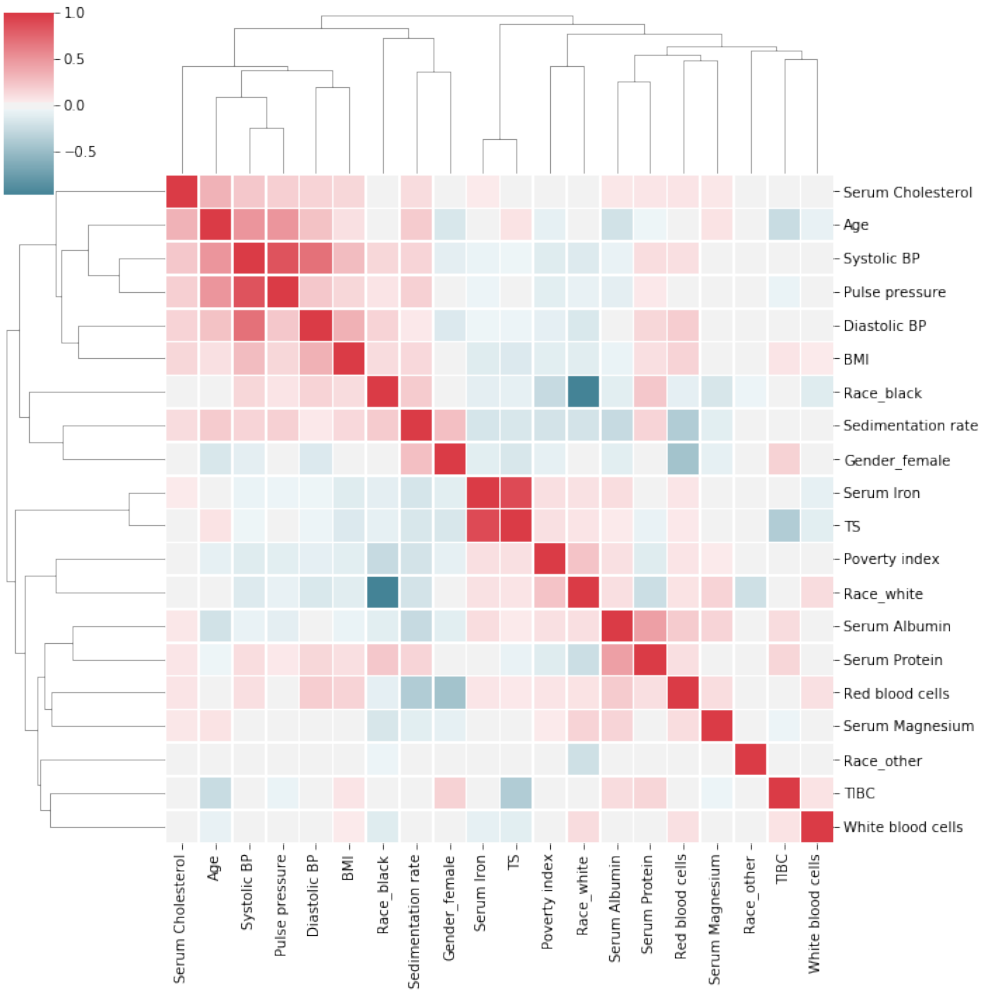


Figure A.2: **Hierarchical clustering of NHANES I feature correlation**. Cells in the clustermap are colored based on their Pearson correlation.