Supplementary Materials :
The Shapley value of coalition of variables provides better estimates

February 5, 2021

This document is organized as follows:

A. **Proofs**

B. **Individual Shapley values for dummy variables**

C. **Plug-In estimator of Marginal expectation**

D. **Experimental details**

A Proofs

This section gathers all the proofs of the propositions and claims of the main paper.

2 Multiple ways of computing the Shapley value of categorical variables

2.1 The Shapley value of a categorical variable

Proposition 2.1. *For all $x \in \mathcal{X}$, and if $y_{1:K-1} = \mathcal{C}(y)$ then*

$$\left\{ \begin{array}{rcl} \phi_C(\tilde{f}; x, y_{1:K-1}) & = & \phi_Y(f; x, y) \\ \phi_X(\tilde{f}; x, y_{1:K-1}) & = & \phi_X(f; x, y) \end{array} \right. \tag{2.1}$$

Proof. As we consider only doable $(x, y_{1:K-1})$, then $\exists! y \in \{1, \ldots, K\}$ such that $\mathcal{C}(y) = y_{1:K-1}$. We have the coalition $C = \{1, \ldots, K-1\}$, and number of variables $p = K$, meaning

$$\phi_{\{1,\ldots,K-1\}}(\tilde{f}; x, y_{1:K-1}) = \frac{1}{2} \left\{ \frac{1}{\binom{1}{0}} \Delta(\tilde{f}; \emptyset, C) + \frac{1}{\binom{1}{1}} \Delta(\tilde{f}; \{X\}, C) \right\}$$

where

$$\begin{aligned} \Delta(\tilde{f}; \emptyset, C) &= E_{\tilde{P}} \left[\tilde{f}(X, Y_{1:K-1}) | Y_{1:K-1} = y_{1:K-1} \right] - E_{\tilde{P}} \left[\tilde{f}(X, Y_{1:K-1}) | \emptyset \right] \\ &= E_P \left[\tilde{f}\left(X, \mathcal{C}\left(Y\right)\right) | Y = y \right] - E_P \left[\tilde{f}\left(X, \mathcal{C}\left(Y\right)\right) \right] \\ &= E_P \left[f\left(X, Y\right) | Y = y \right] - E_P \left[f\left(X, Y\right) \right] \end{aligned}$$

Indeed

$$\begin{aligned} E_{\tilde{P}} \left[\tilde{f}(X, Y_{1:K-1}) | Y_{1:K-1} = y_{1:K-1} \right] &= \int \tilde{f}(x, y_{1:K-1}) dP(x | y_{1:K-1}) \\ &= \int \tilde{f}(x, y_{1:K-1}) \frac{dP(x, y_{1:K-1})}{P(y_{1:K-1})} \\ &= \int \tilde{f}(x, \mathcal{C}(y)) \frac{dP(x, \mathcal{C}(y))}{P(\mathcal{C}(y))} \\ &= \int f(x, y) \frac{dP(x, y)}{P(y)} \end{aligned}$$

In addition,

$$\begin{aligned} \Delta(\tilde{f}; \{X\}, C) &= E_{\tilde{P}} \left[\tilde{f}(X, Y_{1:K-1}) | X = x, Y_{1:K-1} = y_{1:K-1} \right] - E_{\tilde{P}} \left[\tilde{f}(X, Y_{1:K-1}) | X = x \right] \\ &= \tilde{f}(x, y_{1:K-1}) - E_P \left[\tilde{f}(X, \mathcal{C}\left(Y\right)) | X = x \right] \\ &= \tilde{f}(x, \mathcal{C}\left(y\right)) - E_P \left[\tilde{f}(X, \mathcal{C}\left(Y\right)) | X = x \right] \\ &= f(x, y) - E_P \left[f(X, y) | X = x \right] \end{aligned}$$

$$\begin{aligned} \phi_{\{1,\ldots,K-1\}}(\tilde{f}; x, y_{1:K-1}) = & \frac{1}{2} \left(E_P \left[f\left(X, Y\right) | Y = y \right] - E_P \left[f\left(X, Y\right) \right] \right) \\ & + \frac{1}{2} \left(f(x, y) - E_P \left[f(X, y) | X = x \right] \right) \end{aligned}$$

We can recognize that we have exactly $\phi_{\{1,\ldots,K-1\}}(\tilde{f}; x, y_{1:K-1}) = \phi_Y(f; x, y)$. From Equation 2.1, we derive

that $\phi_X(\tilde{f}; x, y_{1:K-1}) = \phi_X(f; x, y)$. □

4 Identifying Active and Null coalitions

4.1 Same Decision Probability

Proposition 4.1. *Let f a probabilistic predictor and its binary classifier C with threshold T, $Q_{S,x}$ the law of $X_{\bar{S}}|X_S = x_S$, then the $SDP_S(f; \boldsymbol{x})$ can be written explicitly with the reduced predictor:*

$$SDP_S(f; \boldsymbol{x}) = \frac{\mathbb{E}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}})] - \mathbb{E}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}})|f(x_S, X_{\bar{S}}) < T]}{\mathbb{E}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}})|f(x_S, X_{\bar{S}}) \geq T] - \mathbb{E}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}})|f(x_S, X_{\bar{S}}) < T]} \tag{4.1}$$

Proof. First note that $SDP_S(f; \boldsymbol{x}) = \mathbb{P}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}}) \geq T]$

$$\begin{aligned}\mathbb{E}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}})] = \mathbb{E}_{X_{\bar{S}}|X_S = x_S}[f(x_S, X_{\bar{S}})|\, f(x_S, X_{\bar{S}}) < T]\, \mathbb{P}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}}) < T] \\ + \mathbb{E}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}})|\, f(x_S, X_{\bar{S}}) \geq T]\, \mathbb{P}_{X_{\bar{S}}|X_S=x_S}[f(x_S, X_{\bar{S}}) \geq T]\end{aligned}$$

Rearranging the terms leads to equation 4.1. The proof of Proposition 4.3 is similar. □

Proposition 4.2. *Let a tree function define as $f(\boldsymbol{x}) = \sum_{m=1}^{M} f_m \mathbb{1}_{L_m}(\boldsymbol{x})$, $Q_{S,x}$ the law of $X_{\bar{S}}|X_S = x_S$. We can compute the conditional expectation of the euclidean distance with our Plug-In estimator:*

$$\begin{aligned}\mathbb{E}_{Q_{S,x}}[d(f(x_S, X_{\bar{S}}), f)] &= \mathbb{E}_{Q_{S,x}}\Big[(\sum_m f_m \mathbb{1}_{L_m}(x) - f)^2\Big] \\ &= \sum_m f_m^2 \mathbb{P}_{Q_{S,x}}\Big[L_m(x)\Big] + f^2 - 2f \sum_m f_m \mathbb{P}_{Q_{S,x}}\Big[L_m(x)\Big]\end{aligned}$$

Hence, we can approximate $\mathbb{P}_{Q_{S,x}}\Big[L_m(x)\Big] \simeq \frac{N(L_m)}{N(L_m^S)}$ as in Equation 3.1.

Proof. Since the leaves form a partition of the space, we can unfold the equation without holding the double terms :

$$\mathbb{E}_{Q_{S,x}}[d(f(x_S, X_{\bar{S}}), f)] = \mathbb{E}_{Q_{S,x}}\Big[(\sum_m y_m \mathbb{1}_{L_m}(x) - f)^2\Big] = \sum_m y_m^2 \mathbb{P}_{Q_{S,x}}\Big[L_m(x)\Big] + f^2 - 2f \sum_m y_m \mathbb{P}_{Q_{S,x}}\Big[L_m(x)\Big]$$

□

Proposition 4.3. *For any Random Forest regressor $F(\boldsymbol{x}) = \sum_{b=1}^{B} \frac{f^b(\boldsymbol{x})}{B}$ with $f^b(\boldsymbol{x}) = \sum_{m=1}^{M_b} f_{m,b} \mathbb{1}_{L_{m,b}}(\boldsymbol{x})$, the conditional expectation of the Euclidean distance is equal to*

$$\begin{aligned}\mathbb{E}_{X_{\bar{S}}|X_S=x_S}\Big[(\frac{\sum_b f_b(x)}{B} - f)^2\Big] = f^2 + \sum_{b=1}^{B}\sum_{m=1}^{M_b}(\frac{1}{M^2} f_{m,b}^2 - \frac{2f}{M} f_{m,b}) \mathbb{P}_{X_{\bar{S}}|X_S=x_S}(L_{m,b}) \\ + \frac{1}{M^2} \sum_{b,l=1 b\neq l}^{B} \sum_{i=1}^{M_b} \sum_{j=1}^{M_l} f_{m,b} f_{b,l} \mathbb{P}_{X_{\bar{S}}|X_S=x_S}(L_{i,b}(x) \cap L_{j,l}(x))\end{aligned}$$

Remark 4.1. We have an additional term that we need to estimate and approximate $\mathbb{P}_{X_{\bar{S}}|X_S=x_S}(L_{i,b}(x) \cap L_{j,l}(x))$, but we can still use the Plug-In estimator to compute it. Indeed,

$$\begin{aligned}\mathbb{P}_{X_{\bar{S}}|X_S=x_S}(L_{i,b}(x) \cap L_{j,l}(x)) &\approx \mathbb{P}(L_{i,b}(x) \cap L_{j,l}(x)|x_S \in L_{i,b}^S \cap L_{j,l}^S) \\ &\approx \frac{N(L_{i,b}(x), L_{j,l}(x))}{N(L_{i,b}^S, L_{j,l}^S)}\end{aligned} \tag{4.2}$$

Where

- $N(L_{i,b}(x), L_{j,l}(x))$ is the number of observations in the leaf $L_{i,b}(x)$ and $L_{j,l}(x)$
- $N(L^S_{i,b}, L^S_{j,l})$ is the number of observations satisfying the conditions $\boldsymbol{x}_S \in L^S_{i,b} \cap L^S_{j,l}$ across all the leaves of the tree.

Proposition 4.4. *If $X \sim \mathcal{N}(\mu, \Sigma)$, then $X_{\bar{S}}|X_S = x_S$ is also multivariate gaussian with mean $\mu_{\bar{S}|S}$ and covariance matrix $\Sigma_{\bar{S}|S}$ equal:*

$$\mu_{\bar{S}|S} = \mu_{\bar{S}} + \Sigma_{\bar{S},S}\Sigma^{-1}_{S,S}(x_S - \mu_S) \quad \textit{and } \Sigma_{\bar{S}|S} = \Sigma_{\bar{S}\bar{S}} - \Sigma_{\bar{S}S}\Sigma^{-1}_{SS}\Sigma_{S,\bar{S}}$$

B Individual Shapley values for dummy variables

We give some partial results for the Shapley Values of the modalities $Y = k$, based on the dummy encoding considered in section 2. Indeed equation 2.4 introduces $\phi_k(\tilde{f}, x, y_{1:K-1})$, and proposition 2.1 claims that their sum is different in all generality of the SV of Y. In this section, we give a deeper insight into these values and show that are related multiple comparisons between modalities.

We compute the Shapley Value at point $(x, y = i) = (x, 0, 0, \ldots, 1, \ldots, 0) = (x, \mathcal{C}(y))$: for ease of notation, we set $Y_0 = X$, and we compute also the Shapley values $\phi_k(\tilde{f}; x, y_{1:K-1})$ for $k = 1, \ldots, K-1$. We recall that we need to compute

$$\frac{1}{K}\sum_{k=0}^{K-1} \frac{1}{\binom{K-1}{k}} \sum_{\substack{Z \subseteq [\![1..K]\!] \setminus i \\ |Z| = k}} \Delta(\tilde{f}; Z, i).$$

where Δ denotes the difference between the value function evaluated at $Z \cup \{i\}$ and Z. If we examine the terms $\Delta(\tilde{f}; Z, i)$, the computation needs to take into account if $X = Y_0$ is part of the conditioning variable of not. Indeed, we have for each $k \geq 1$,

$$\sum_{\substack{Z \subseteq [\![0..K-1]\!] \setminus i \\ |Z| = k}} \Delta(\tilde{f}; Z, i) = \sum_{\substack{Z \subseteq [\![1..K-1]\!] \setminus i \\ |Z| = k}} \Delta(\tilde{f}; Z, i) + \sum_{\substack{Z' \subseteq [\![1..K-1]\!] \setminus i \\ |Z'| = k-1}} \Delta(\tilde{f}; Z' \cup \{0\}, i). \quad (1.3)$$

We start by computing the first term in the right hand side, and it involves only the dummies, and not the quantitative variable.

Proposition 1.1 (Computation of Contributions in Shapley without X)**.** *We compute the Shapley values of the variable Y_i , when we have the observations $(x, y_{1:K-1}) = (x, \mathcal{C}(i))$ for $i \in \{1, \ldots, K\}$. We consider any $Z' \subseteq [\![1..K-1]\!] \setminus i$, with $|Z'| = k \geq 1$ and $Z' = \{j_1, \ldots, j_k\}$. In that case,*

$$\Delta(\tilde{f}; Z, i) = E_P\left[f(X, Y)|Y = i\right] - E_P\left[f(X, Y)|Y \notin \{j_1, \ldots, j_k\}\right] \quad (1.4)$$

Proof. We have $Y_i = 1 \Leftrightarrow Y = i$, and for $Z' \subseteq [\![1..K-1]\!] \setminus \{0, i\}$, we consider $Z' = \{j_1, \ldots, j_k\}$, with $1 \leq j_1 < \cdots < j_k \leq K-1$,

$$\begin{aligned} E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_{j_1} = 0, \ldots, Y_{j_k} = 0, Y_i = 1\right] &= E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_i = 1\right] \\ &= E_{\tilde{P}}\left[\tilde{f}(Y_0, \mathcal{C}(Y))|Y_i = 1\right] \\ &= E_P\left[f(Y_0, Y)|Y = i\right] \end{aligned}$$

because for all $j_1, \ldots, j_{k-1} \neq i$, we have $\{Y_{j_1} = 0, \ldots, Y_{j_k} = 0, Y_i = 1\} = \{Y_i = 1\}$.
Moreover,

$$E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_{j_1} = 0, \ldots, Y_{j_k} = 0\right] = E_P\left[\tilde{f}(Y_0, \mathcal{C}(Y))|Y \neq j_1, \ldots, j_k\right]$$

Hence for $Z \subseteq [\![1..K-1]\!] \setminus i$, we have

$$\Delta(\tilde{f}; Z, i) = E_P\left[f(X,Y)|Y = i\right] - E_P\left[f(X,Y)|Y \notin \{j_1, \ldots, j_k\}\right].$$

□

The second term of the right hand side is given below.

Proposition 1.2 (Computation of Contributions in Shapley with X)**.** *We compute the Shapley values only for the variable Y_i , when we have the observations doable $(x, y_{1:K-1}) = (x, \mathcal{C}(i))$ for $i \in \{1, \ldots, K\}$. We consider any $Z' \subseteq [\![1..K-1]\!] \setminus i$, with $|Z'| = k-1 \geq 1$, and $Z' = \{j_1, \ldots, j_{k-1}\}$. In that case,*

$$\Delta(\tilde{f}; Z' \cup \{0\}, i) = E_P\left[f(X,Y)|X = x, Y = i\right] - E_P\left[f(X,Y)|X, Y \notin \{j_1, \ldots, j_{k-1}\}\right] \tag{1.5}$$

Proof. We assume that we have a subset $|Z'| = k-1$, such that $Z' \subseteq [\![1..K-1]\!] \setminus i$. This means that $Z' = \{j_1, \ldots, j_{k-1}\}$, with $1 \leq j_1, \ldots, j_{k-1} \leq K-1$. We

$$\begin{aligned} E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_0 = x, Y_{j_1} = 0, \ldots, Y_{j_{k-1}} = 0, Y_i = 1\right] &= E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_0 = x, Y_i = 1\right] \\ &= E_P\left[\tilde{f}(Y_0, \mathcal{C}(Y))|Y_0 = x, Y = i\right] \\ &= E_P\left[f(Y_0, \mathcal{C}(Y))|Y_0 = x, Y = i\right] \end{aligned}$$

and

$$\begin{aligned} E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_0 = x, Y_{j_1} = 0, \ldots, Y_{j_{k-1}} = 0\right] &= E_P\left[\tilde{f}(Y_0, \mathcal{C}(Y))|Y_0 = x, Y \notin \{j_1, \ldots, j_{k-1}\}\right] \\ &= E_P\left[f(Y_0, Y)|Y_0 = x, Y \notin \{j_1, \ldots, j_{k-1}\}\right] \end{aligned}$$

□

Finally, we can give several examples of the different increments involved in the Shapley values of each variable X or Y_k. If $k = 0$, then $Z' = \emptyset$ and

$$\Delta(\tilde{f}; Z', i) = \Delta(\tilde{f}; \emptyset, i) = E_P\left[f(X,Y)|Y = i\right] - E_P\left[f(X,Y)\right]$$

If $k = 1$, then $Z' = \{0\}$ or $Z' = \{j\} \neq \{i\}$,

$$\begin{aligned} \Delta(\tilde{f}; Z', i) = \Delta(\tilde{f}; 0, i) &= E_P\left[f(X,Y)|X = x, Y = i\right] - E_P\left[f(X,Y)|X = x\right] \\ \Delta(\tilde{f}; Z', i) = \Delta(\tilde{f}; \{j\}, i) &= E_P\left[f(X,Y)|Y = i\right] - E_P\left[f(X,Y)|Y \neq j\right] \end{aligned}$$

For $k = K-1$, $Z' = \{1, \ldots K-1\}$,

$$\Delta(\tilde{f}; \{1, \ldots K-1\}, i) = E_P\left[f(X,Y)|X = x, Y = i\right] - E_P\left[f(X,Y)|X = x, Y \neq i\right]$$

The propositions 1.1 and 1.2 show that the individual Shapley value for the variable (modality) Y_i is a weighted mean of the difference between classe i and group of classes:

$$\begin{cases} E_P\left[f(X,Y)|Y = i\right] - E_P\left[f(X,Y)|Y \notin \{j_1, \ldots, j_k\}\right] \\ E_P\left[f(X,Y)|X = x, Y = i\right] - E_P\left[f(X,Y)|X, Y \notin \{j_1, \ldots, j_{k-1}\}\right] \end{cases}$$

Finally, we can also compute the Shapley values of the other variables Y_j at point $(x, y = i)$, for $j \neq i$. In that case, the difference $\Delta(\tilde{f}; Z', j), j \neq i$ are of the type of

$$\begin{cases} E_P\left[f(X,Y)|Y \notin \{j, j_1, \ldots, j_k\}\right] - E_P\left[f(X,Y)|Y \notin \{j_1, \ldots, j_k\}\right] \\ E_P\left[f(X,Y)|Y = i\right] - E_P\left[f(X,Y)|Y = i\right] \\ E_P\left[f(X,Y)|X = x, Y \notin \{j, j_1, \ldots, j_k\}\right] - E_P\left[f(X,Y)|X, Y \notin \{j_1, \ldots, j_{k-1}\}\right] \\ E_P\left[f(X,Y)|X = x, Y = i\right] - E_P\left[f(X,Y)|X, Y = i\right] \end{cases}$$

The Shapley values computes a mean of the difference between different aggregation of modalities, that contains or not the variable of interest.

As a conclusion of this part, we see that the individual Shapley values $\phi_k(\tilde{f}; x, y_{1:K-1})$ perform a multiple comparison of the means obtained by aggregating the classes or modalities in various ways, looking at the presence or not of the modality k. These differences of means have weights $\frac{1}{\binom{K-1}{k}}$ where k is basically the number of classes of the variable Y that we aggregate.

Consequently the sum $\sum_{k=1}^{K} \phi_k(\tilde{f}; x, y_{1:K-1})$ is clearly different from the

$$\phi_Y(f; x, y) = \frac{1}{2}\left(E\left[f(X,Y)|Y = y\right] - E\left[f\left(X,Y\right)\right]\right) + \frac{1}{2}\left(f\left(x,y\right) - E\left[f\left(X,Y\right)|X = x\right]\right).$$

This latter has a much more global analysis that aims at measuring how the mean $E\left[f(X,Y)|Y = y\right]$ in the various classes changes w.r.t $E\left[f\left(X,Y\right)\right]$, while the individual Shapley focus on the difference between subgroups of classes.

C Plug-In estimator of Marginal expectation

As we have indicated in the paper, the Shapley Values can be computed with different probability $Q_{S,\boldsymbol{x}}$. In that section, we show that when we use the marginal distribution (as in the so-called interventional case), the previous estimators for tree-based models can be adapted straightforwardly.

We consider then decision tree

$$f(x) = \sum_{m=1}^{M} f_m \mathbb{1}_{L_m}(x)$$

and remark that the Marginal Shapley coefficients involve the computations of the marginal expectations $E_P\left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z)\right]$ for any subgroup of variables Z. On real data, we need to compute the conditional expectations, but we use the Tree approximations in order to replace

$$\begin{aligned} E_P\left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z)\right] &= \int\int \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) p(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{u}_Z) d\boldsymbol{u}_{\bar{Z}} d\boldsymbol{u}_Z \\ &= \int\int \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) p(\boldsymbol{u}_Z|\boldsymbol{u}_{\bar{Z}}) p(\boldsymbol{u}_{\bar{Z}}) d\boldsymbol{u}_Z d\boldsymbol{u}_{\bar{Z}} \\ &= \int \left\{\int p(\boldsymbol{u}_Z|\boldsymbol{u}_{\bar{Z}}) d\boldsymbol{u}_Z\right\} \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) p(\boldsymbol{u}_{\bar{Z}}) d\boldsymbol{u}_{\bar{Z}} \\ &= \int \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) p(\boldsymbol{u}_{\bar{Z}}) d\boldsymbol{u}_{\bar{Z}} \end{aligned}$$

This means that we just need the marginal distributions of the variables $\boldsymbol{X}_{\bar{Z}}$ in order to compute the expectations of the leaf. In the case of quantitative data, the leaf can be written $L_m = \prod_{i=1}^{p} [a_i^m, b_i^m]$, and we have by definition

$$\exists k \in Z,\, x_k \notin [a_k, b_k] \implies \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) = 0$$

We define the set of leafs compatible with condition $\boldsymbol{X}_Z = \boldsymbol{x}_Z$ as

$$C(Z, \boldsymbol{x}) = \left\{ m \in [1 \dots M] \, | L_m = \prod_{i=1}^{p} [a_i^m, b_i^m], \forall k \in Z, x_k \in [a_k^m, b_k^m] \right\}$$

We write for $m \in C(Z, \boldsymbol{x})$, $L_m = L_m^{\bar{Z}} \times L_m^Z$, with $L_m^{\bar{Z}} = \prod_{i \in \bar{Z}} [a_i^m, b_i^m]$ and $L_m^Z = \prod_{i \in Z} [a_i^m, b_i^m]$, this means that for all $m \in C(Z, \boldsymbol{x})$ we have

$$E_P \left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z) \right] = E_P \left[\mathbb{1}_{L_m^{\bar{Z}}}(\boldsymbol{X}_{\bar{Z}}) \right]$$

As an approximation, the conditional probability for $m \in C(Z, \boldsymbol{x})$ is computed as

$$\begin{aligned} E_P \left[\mathbb{1}_{L_m^{\bar{Z}}}(\boldsymbol{X}_{\bar{Z}}) \right] &= P \left(X_i \in [a_i^m, b_i^m], i \in \bar{Z} \right) \\ &\simeq \frac{N(L_m^{\bar{Z}})}{N} \end{aligned}$$

where $N(L_m^{\bar{Z}})$ is the number of observations in the (partial) leaf $L_m^{\bar{Z}}$. As a consequence we have

$$\begin{aligned} E_P \left[f(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z) \right] &= \sum_{m=1}^{M} \hat{y}_m E_P \left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z) \right] \\ &= \sum_{m \in C(Z, \boldsymbol{x})} \hat{y}_m E_P \left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z) \right] \\ &= \sum_{m \in C(Z, \boldsymbol{x})} \hat{y}_m E_P \left[\mathbb{1}_{L_m^{\bar{Z}}}(\boldsymbol{X}_{\bar{Z}}) \right] \\ &\simeq \sum_{m \in C(Z, \boldsymbol{x})} \hat{y}_m \frac{N(L_m^{\bar{Z}})}{N} \end{aligned}$$

D EXPERIMENTAL SETTINGS

All our experiments are reproducible and can be found on the github repository *Active Coalition of Variables*, https://github.com/acvicml/ACV.

1.1 Toy model of Section 2.2.1

Recall that the model is a linear predictor with categorical variables define as $f(X, Y) = B_Y X$ with $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma_y)$ and $\mathbb{P}(Y = y) = \pi_y$, $Y \in \{a, b, c\}$.
For the experiments in Figure 1 and 2, we set $\pi_y = \frac{1}{3}, \mu_y = 0 \forall \, y \in \{a, b, c\}$. We use a random matrices generated from a Wishart distribution. The covariance matrices are:
$\Sigma_a = \begin{bmatrix} 0.41871254 & -0.790061361 & 0.46956991 \\ -0.79006136 & 1.90865098 & -0.82571655 \\ 0.46956991 & -0.82571655 & 0.95835472 \end{bmatrix}$, $\Sigma_b = \begin{bmatrix} 0.55326081 & 0.11811951 & -0.70677924] \\ 0.11811951 & 2.73312979 & -2.94400196 \\ -0.70677924 & -2.94400196 & 4.22105088 \end{bmatrix}$, $\Sigma_c = \begin{bmatrix} 9.2859966 & 1.12872646 & 2.4224434 \\ 1.12872646 & 0.92891237 & -0.14373393 \\ 2.4224434 & -0.14373393 & 1.81601676 \end{bmatrix}$ for $y \in \{a, b, c\}$ respectively.
The coefficients are $B_a = [1, 3, 5], B_b = [-5, -10, -8], B_c = [6, 1, 0]$ and the selected observation in figure 1 values is $x = [-1.15151515, 3.50850731, -1.28257141, 1., 0., 0.]$

1.2 Model trained on Census Data of Section 2.2.2

The parameters of the Random Forest used in this section are: estimators=2, max depth=15, min samples leaf=10, random state=2021. It has an accuracy of 86% and was trained on 4 variables: Education, Marital

Status, Race, Workclass.

1.3 Linear model of Section 3.1.1

The data $\mathcal{D} = (x_i, z_i)_{1 \leq i \leq n}$ are generated from a linear regression $Z = B^t X$ with $X \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = 0$, $\Sigma = \begin{bmatrix} 3.29170917 & -1.05668255 & 3.13134267 \\ -1.05668255 & 7.1972724 & -3.26516336 \\ 3.13134267 & -3.26516336 & 3.72558998 \end{bmatrix}$, $B = [0.5, 10, -4]$, $n = 10000$.

We used a decision tree on $\mathcal{D}$ with the following parameters: min samples leaf=10, random state=2877. The Mean Squared Error (MSE) are MSE = 1.546 on Test Set and MSE = 0.934 on Training Set.

1.4 Toy model of Section 4.2

Model is same as 1.1 with the following parameters: $\pi_y = \frac{1}{3}, \mu_y = 0 \; \forall y \in \{a, b, c\}$. For covariance Σ_y, we used a random matrix generated from a Wishart distribution. The values used are:

$$\begin{bmatrix} 7.06021734 & 6.25437096 & 0.93988464 & -2.63006016 & -5.05313408 \\ 6.25437096 & 7.11169325 & 3.60447888 & -1.09159294 & -6.70063316 \\ 0.93988464 & 3.60447888 & 5.98234548 & 2.26837008 & -4.07866549 \\ -2.63006016 & -1.09159294 & 2.26837008 & 2.33853775 & 0.66676669 \\ -5.05313408 & -6.70063316 & -4.07866549 & 0.66676669 & 7.54928006 \end{bmatrix},$$

$$\begin{bmatrix} 1.10024448 & -2.05714423 & -0.70124936 & -0.54361427 & 0.35289562 \\ -2.05714423 & 6.55691981 & -0.67123557 & 2.23271169 & -0.73350145 \\ -0.70124936 & -0.67123557 & 2.47842554 & -0.23410955 & 1.11591119 \\ -0.54361427 & 2.23271169 & -0.23410955 & 1.26465508 & 0.3616061 \\ 0.35289562 & -0.73350145 & 1.11591119 & 0.3616061 & 3.12144679 \end{bmatrix},$$

$$\begin{bmatrix} 4.39282235 & 1.47065398 & 4.22447202 & -1.94787456 & -1.41195954 \\ 1.47065398 & 2.53622587 & 0.47361504 & -1.26155545 & 1.02024204 \\ 4.22447202 & 0.47361504 & 5.13977882 & -2.17268713 & -3.02963689 \\ -1.94787456 & -1.26155545 & -2.17268713 & 2.089194 & 1.39065915 \\ -1.41195954 & 1.02024204 & -3.02963689 & 1.39065915 & 3.29291599 \end{bmatrix} \text{ for } y \in \{a, b, c\} \text{ respectively. `}$$

The coefficients are: $B_a = [9, 5, -8, 0, 0], B_b = [0, 0, -6, -8, -9], B_c = [4, 9, 13, 0, 0]$ and the selected observation in figure 6 is $x = [2.0601348, 2.00732006, -0.90904155, -1.91465107, -3.7068906, 1., 0]$.

We use a decision tree with the following parameters: min samples leaf=5, min samples split=10, max depth=20, random state=2021. The MSE = 26.407, and on Test set, MSE= 8.643 on Training set.

1.5 Model used for Lucas Dataset in Section 5

In figure 2 and 3 we can find the graph and the probabilities defining the Bayesian network associated to the Lucas Dataset. We use a decision tree with the following parameters: min samples leaf=20, random state=212. The accuracy is 0.9230919439227646 on Training set and 0.9127666666666667 on Test set. The observation used in figure 9 is:

Smoking	Yellow_Fingers	Anxiety	Peer_Pressure	Genetics	Attention_Disorder	Born_an_Even_Day	Car_Accident	Fatigue	Allergy	Coughing
0	0	0	0	0	0	1	1	0	0	0

Figure 1: Observation used in figure 9

1.6 Example of calculation time of SV estimates with plugin estimator

We show below the elapsed time take to compute the SV for one observation, with an increasing number of variables d, for a single decision tree trained on dataset of size 20 000 generated by the toy model.

Dimension	3	4	5	6	7
Elapsed time	3.38 sec	34 sec	4.4 min	26.6 min	4.7 hours

Table 1: It was compute on a intel Core i7 10875H (8 Curs HT, 2.3Ghz, 5.1GHz Turbo, 16Mo cache) and 32 Go Ram.

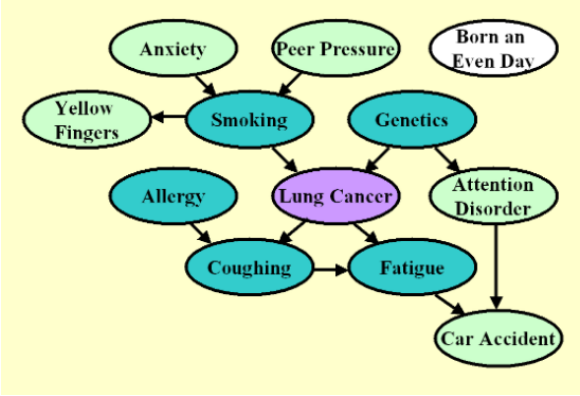

Figure 2: Bayesian network that represents the causal relationships between variables

P(Anxiety=T)=0.64277
P(Peer Pressure=T)=0.32997
P(Smoking=T|Peer Pressure=F, Anxiety=F)=0.43118
P(Smoking=T|Peer Pressure=T, Anxiety=F)=0.74591
P(Smoking=T|Peer Pressure=F, Anxiety=T)=0.8686
P(Smoking=T|Peer Pressure=T, Anxiety=T)=0.91576
P(Yellow Fingers=T|Smoking=F)=0.23119
P(Yellow Fingers=T|Smoking=T)=0.95372
P(Genetics=T)=0.15953
P(Lung cancer=T|Genetics=F, Smoking=F)=0.23146
P(Lung cancer=T|Genetics=T, Smoking=F)=0.86996
P(Lung cancer=T|Genetics=F, Smoking=T)=0.83934
P(Lung cancer=T|Genetics=T, Smoking=T)=0.99351
P(Attention Disorder=T|Genetics=F)=0.28956
P(Attention Disorder=T|Genetics=T)=0.68706
P(Born an Even Day=T)=0.5
P(Allergy=T)=0.32841
P(Coughing=T|Allergy=F, Lung cancer=F)=0.1347
P(Coughing=T|Allergy=T, Lung cancer=F)=0.64592
P(Coughing=T|Allergy=F, Lung cancer=T)=0.7664
P(Coughing=T|Allergy=T, Lung cancer=T)=0.99947
P(Fatigue=T|Lung cancer=F, Coughing=F)=0.35212
P(Fatigue=T|Lung cancer=T, Coughing=F)=0.56514
P(Fatigue=T|Lung cancer=F, Coughing=T)=0.80016
P(Fatigue=T|Lung cancer=T, Coughing=T)=0.89589
P(Car Accident=T|Attention Disorder=F, Fatigue=F)=0.2274
P(Car Accident=T|Attention Disorder=T, Fatigue=F)=0.779
P(Car Accident=T|Attention Disorder=F, Fatigue=T)=0.78861
P(Car Accident=T|Attention Disorder=T, Fatigue=T)=0.97169

Figure 3: Probabilities table used to generate Data