Understanding High-Risk Adolescent Obesity Through Variable Importance and Data Augmentation

GAME 2026

Amir Khorrami Chokami
Università di Cagliari

Bologna, 11 - 12 June 2026

Motivation

- Adolescent obesity is particularly concerning because of its long-term impact on adult health.
- In clinical practice, obesity is defined as body mass index (BMI) values exceeding 30 kg/m^2, corresponding to observations located in the upper tail of the BMI distribution.
- The Shapley values emerge as a tool from sensitivity analysis to tackle this challenge. However, the scarcity of data poses significant numerical challenges.
- We use a data augmentation procedure based on vine copulas.

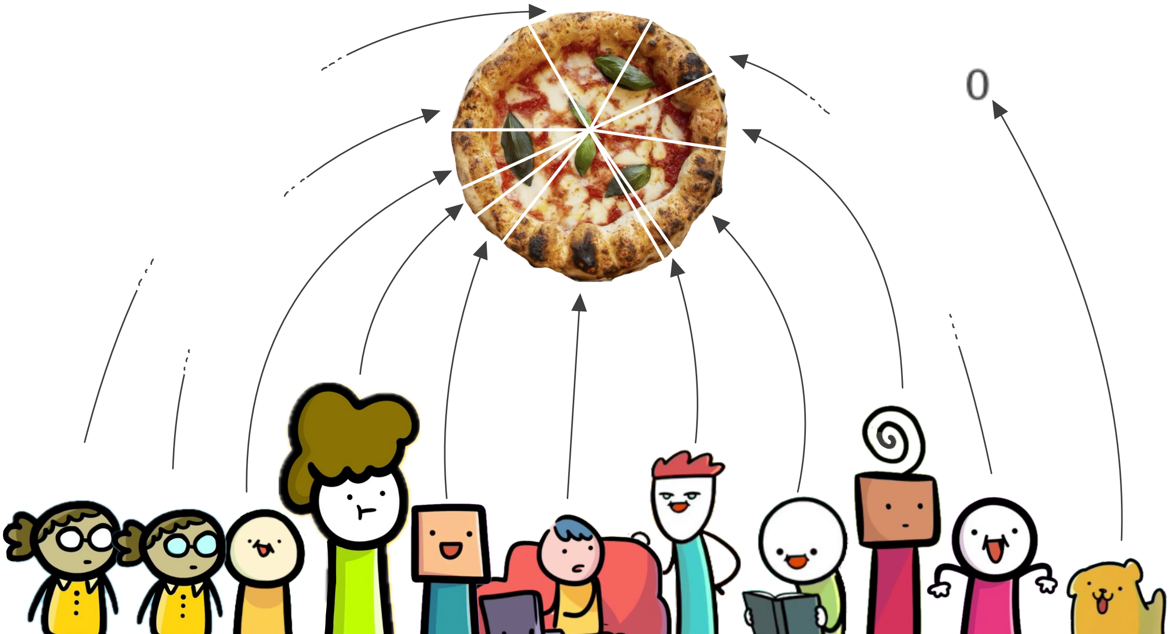
Table of Contents

The Shapley value

It is a method to allocate the total payoff of a game to each player in a coalition.
Consider a set of players $K = \{1, 2, \ldots, k\}$ and let $u \subseteq K$ be any possible coalition.
Call val : $2^K \to \mathbb{R}$ the value function of the game.

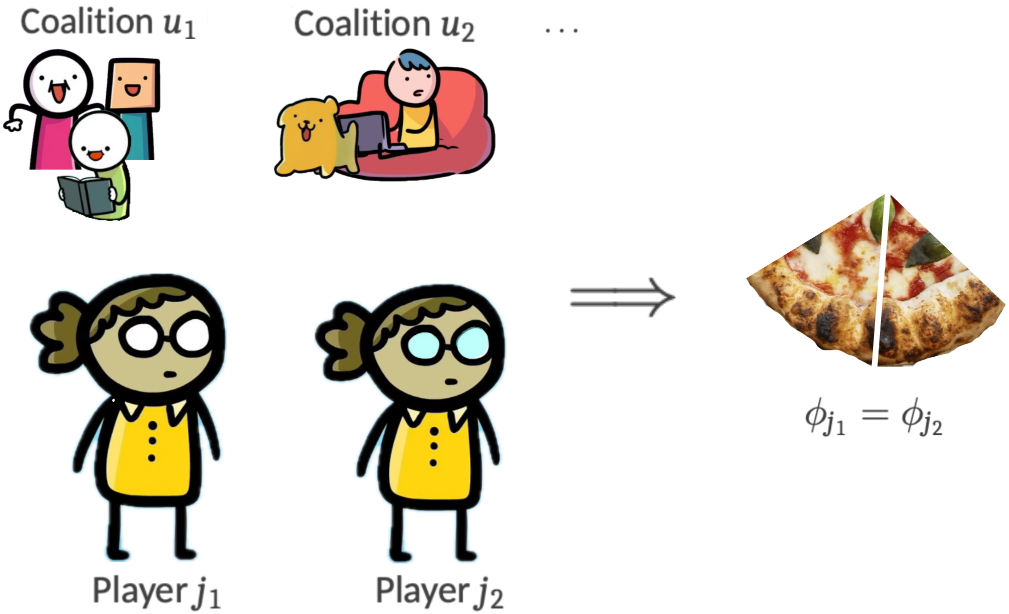
1. Efficiency

The sum of Shapley values of all players equates to the total value of the game $\sum_{j=1}^{k} \phi_j = \mathrm{val}(K)$.



2. Symmetry

If players j_1 and j_2 make the same marginal contribution to any coalition, *i.e.* $\text{val}(u \cup \{j_1\}) = \text{val}(u \cup \{j_2\})$ for all $u \subseteq K$ with $j_1, j_2 \notin K$, then $\phi_{j_1} = \phi_{j_2}$.



3. Dummy Property

If j is a dummy player, *i.e.* $\text{val}(u \cup \{j\}) = \text{val}(u)$ for all $u \subseteq K$, then $\phi_j = 0$.

Coalition u

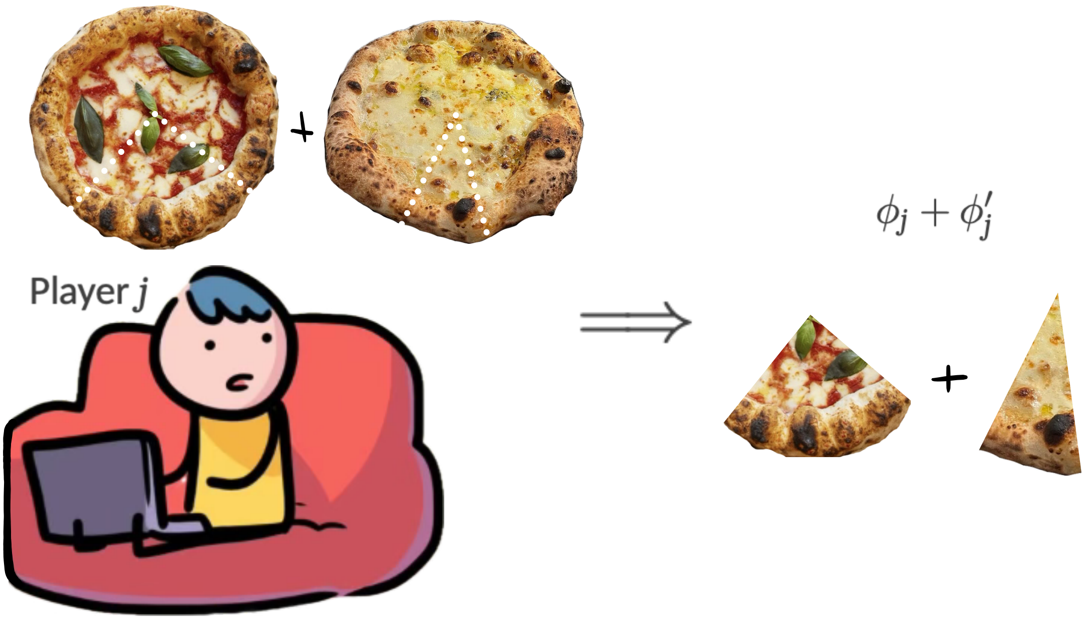
Coalition v

$\dots$

$\text{val}(u \cup \{\text{dog}\}) = \text{val}(u)$ $\quad$ $\text{val}(v \cup \{\text{dog}\}) = \text{val}(v)$ $\quad \Longrightarrow$

4. Additivity

If val and val$'$ have Shapley values ϕ and ϕ' respectively, then the game with value val(u) + val$'(u)$ has Shapley value $\phi_j + \phi'_j$ for all $j \in K$.



The Shapley value

The only value which satisfies the previous properties is (Shapley, 1953)

$$\phi_j = \sum_{u \subseteq K \setminus \{j\}} \frac{(k - |u| - 1)!|u|!}{k!} \left(\mathrm{val}(u \cup \{j\}) - \mathrm{val}(u)\right).$$

Table of Contents

The Shapley value for Variable Importance

- Key idea: the covariates are the *players* and the *total value* to be allocated is $\mathbb{V}ar\,(Y)$ use the Shapley values to fairly distribute the variance of the output among the covariates.
- The definition of the value function we consider for our purpose is

$$\mathrm{val}(u) = \mathrm{Var}(\mathbb{E}[Y \mid X_u]),$$

which quantifies the variance of Y explained by X_u.

- The covariates explaining a higher fraction of the variance are considered the most important.

Shapley Effects

- Appealing properties:
  1. $\phi_j \geq 0$ for all $j = 1, 2, \ldots, k$. This holds because val$(u \cup \{j\})$ − val(u) is positive as the variance explained increases when a new covariate is considered for all $u \subseteq K$.
  2. $\sum_{j=1}^{k} \phi_j = \mathbb{V}ar\,(Y)$, so the variance of the outputis fully allocated.
  3. Well-defined under any dependency structure among variables.

- We use the approach of Mase et al. (2019), who propose a procedure to estimate Shapley effects given data. Their approach is based on cohorts (observations with similar covariate values in the coalition u). Their sample mean estimates $\mathbb{E}[Y \mid X_u]$, with no predictive model.

- Problem: with scarce data, cohort estimates become inaccurate.

Shapley Effects

- Appealing properties:
  1. $\phi_j \geq 0$ for all $j = 1, 2, \ldots, k$. This holds because val$(u \cup \{j\})$ − val(u) is positive as the variance explained increases when a new covariate is considered for all $u \subseteq K$.
  2. $\sum_{j=1}^{k} \phi_j = \mathbb{V}ar\,(Y)$, so the variance of the outputis fully allocated.
  3. Well-defined under any dependency structure among variables.

- We use the approach of Mase et al. (2019), who propose a procedure to estimate Shapley effects given data. Their approach is based on cohorts (observations with similar covariate values in the coalition u). Their sample mean estimates $\mathbb{E}[Y \mid X_u]$, with no predictive model.

- Problem: with scarce data, cohort estimates become inaccurate.

Table of Contents

Copulas

Definition

A copula is a multivariate distribution function $C\colon [0,1]^d \to [0,1]$ with standard uniform margins, *i.e.* $C(1,\dots,1,u_i,1,\dots,1) = u_i$, $\forall i \in \{1,\dots,d\}$, $u_i \in [0,1]$.

Theorem (Sklar)

Any joint density factors as

$$f(x_1,\dots,x_d) = c(F_1(x_1),\dots,F_d(x_d))f_1(x_1)\cdots f_d(x_d)$$

for some d-dimensional copula C with copula density c.

Consequence: separate modeling of marginal distributions and the dependence structure using copulas, enabling flexibility in capturing complex dependencies.

Example - Czado and Nagler (2022)

Key tool: conditioning. Consider $d = 3$ and a trivariate density. We know

$$f\left(x_1, x_2, x_3\right) = f_{3|12}\left(x_3|x_1, x_2\right) f_{2|1}\left(x_2|x_1\right) f_1\left(x_1\right).$$

Here $F_{j|D}$ and $f_{j|D}$ are the conditional cdf and conditional pdf of $X_j|X_D$, respectively. Note

$$\begin{aligned} f_{3|12}\left(x_3|x_1, x_2\right) &= \frac{f_{13|2}\left(x_1, x_3|x_2\right)}{f_{1|2}\left(x_1|x_2\right)} \\ &= \frac{c_{13;2}\left(F_{1|2}\left(x_1|x_2\right), F_{3|2}\left(x_3|x_2\right); x_2\right)}{f_{1|2}\left(x_1|x_2\right)} f_{1|2}\left(x_1|x_2\right) f_{3|2}\left(x_3|x_2\right) \\ &= c_{13;2}\left(F_{1|2}\left(x_1|x_2\right), F_{3|2}\left(x_3|x_2\right); x_2\right) f_{3|2}\left(x_3|x_2\right) \end{aligned}$$

Example (Cont'd)

Since

$$f_{3|2}\left(x_3|x_2\right)=\frac{f\left(x_2,x_3\right)}{f_2\left(x_2\right)}=\frac{c_{23}\left(F_2\left(x_2\right),F_3\left(x_3\right)\right)}{f_2\left(x_2\right)}f_2\left(x_2\right)f_3\left(x_3\right)=c_{23}\left(F_2\left(x_2\right),F_3\left(x_3\right)\right)f_3\left(x_3\right)$$

then

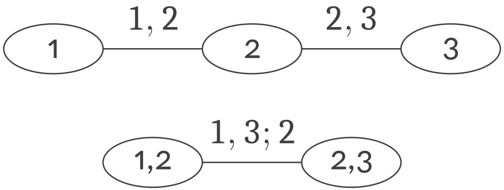
$$f_{3|12}\left(x_3|x_1,x_2\right)=c_{13;2}\left(F_{1|2}\left(x_1|x_2\right),F_{3|2}\left(x_3|x_2\right);x_2\right)c_{23}\left(F_2\left(x_2\right),F_3\left(x_3\right)\right)f_3\left(x_3\right).$$

Similarly $f_{2|1}\left(x_2|x_1\right)=c_{12}\left(F_1\left(x_1\right),F_2\left(x_2\right)\right)f_2\left(x_2\right)$, so that

$$\begin{aligned}f(x_1,x_2,x_3)&=f_{3|12}(x_3|x_1,x_2)f_{2|1}(x_2|x_1)f_1(x_1)\\&=c_{13;2}(F_{1|2}(x_1|x_2),F_{3|2}(x_3|x_2);x_2)\cdot c_{23}(F_2(x_2),F_3(x_3))\cdot c_{12}(F_1(x_1),F_2(x_2))\\&\qquad\cdot f_3(x_3)f_2(x_2)f_1(x_1)\end{aligned}$$

Vine Tree structure

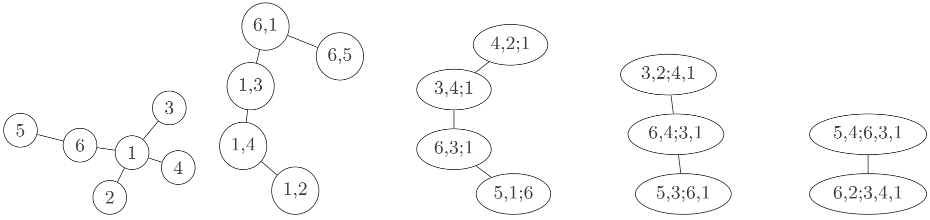
In three or more dimensions, it is helpful to represent the density factorization as a graph called vine tree structure (Bedford and Cooke, 2002).



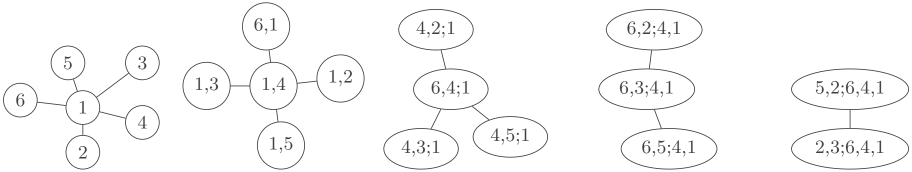
Graphical representation of $c_{23} - c_{12} - c_{13;2}$.

$$
\begin{aligned}
f(x_1, x_2, x_3) = c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) \\
\cdot c_{23}(F_2(x_2), F_3(x_3)) \\
\cdot c_{12}(F_1(x_1), F_2(x_2)) \\
\cdot f_3(x_3) f_2(x_2) f_1(x_1)
\end{aligned}
$$

Example ($d = 6$)



$$\text{R-vine: } f_{1,\dots,6} = c_{12}\, c_{13}\, c_{14}\, c_{16}\, c_{56}\, c_{15;6}\, c_{36;1}\, c_{34;1}\, c_{24;1}\, c_{23;14}\, c_{46;13}\, c_{35;16}\, c_{45;136}\, c_{26;134}\, c_{24;1346} \prod_{i=1}^{6} f_i$$



$$\text{C-vine: } f_{1,\dots,6} = c_{12}\, c_{13}\, c_{14}\, c_{15}\, c_{16}\, c_{24;1}\, c_{34;1}\, c_{45;1}\, c_{46;1}\, c_{26;14}\, c_{36;14}\, c_{56;14}\, c_{23;146}\, c_{25;146}\, c_{35;1246} \prod_{i=1}^{6} f_i$$

Data augmentation steps

To augment a small sample with vine copulas, we perform the following steps:

Step 1 Isolate the dependence structure by transforming the marginal distributions into the Uniform$(0, 1)$ distribution through the rank transformation $\left(\text{rank}(x) - \frac{1}{2}\right)/n$;

Step 2 Estimate the vine copula tree structure by maximum likelihood and select it based on the AIC with the `R` function `RVineStructureSelect` in the package `VineCopula`;

Step 3 Simulate new observation from the selected vine copula;

Step 4 Apply to the new data the empirical inverse cumulative distribution functions to recover the original scales of the marginals.

$\Rightarrow$ We have a larger sample from $f(y, x_1, x_2, \ldots, x_d)$.

More details in Khorrami Chokami and Rabitti (2025)

Table of Contents

Simulation Study

Model: $Y = 2X_1 + X_2 + 0.3X_3$, with $X_1 \sim \mathcal{N}(0,4)$, $X_2 \sim \mathcal{N}(0,1)$, $X_3 \sim \mathcal{N}(0,2.25)$, and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ 0 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}, \qquad \rho \in \{0, 0.5, 0.8\}.$$

Steps 1 - 4 were applied to simulate 2000 observations; procedure repeated 20 times.

Mean Squared Errors (No Data Augmentation vs Data Augmentation)

	$\rho = 0$		$\rho = 0.5$		$\rho = 0.8$	
n	No DA	DA	No DA	DA	No DA	DA
50	0.0128	**0.00318**	0.0104	**0.00301**	0.00767	**0.00363**
250	0.00370	**0.00163**	0.00350	**0.00210**	0.00304	**0.00208**
500	0.00275	**0.00158**	0.00259	**0.00168**	0.00211	**0.00184**

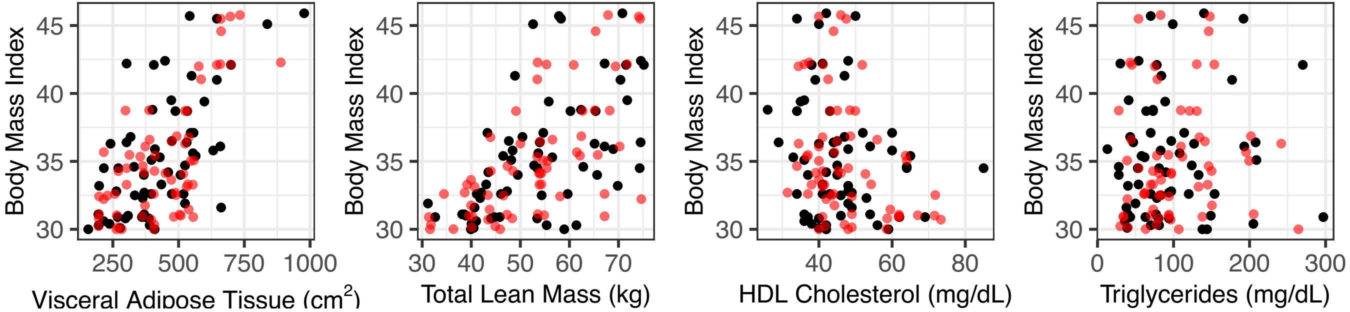
Table of Contents

Practical problem: Understanding Obesity

- Data are drawn from the National Health and Nutrition Examination Surveys , conducted by the Center for Disease Control and Prevention between 2013 and 2014.
- Civilian U.S. residents individuals aged between 3 and 18 years.
- The analysis focuses on 15 continuous variables, including:
  - Physical characteristics (*e.g.*, BMI, body measurements, lean and fat mass indicators)
  - Clinical biomarkers (*e.g.*, cholesterol, triglycerides, hematological parameters)
- The final dataset includes 62 observations.

Check adequacy

Figure: Scatterplots comparing the dependence between BMI and selected covariates in the observed and simulated data. Black dots represent the observed data, whereas red dots correspond to simulated observations generated from the selected vine copula.



Shapley effects

Figure: Shapley effects for each covariate, estimated from the augmented dataset.

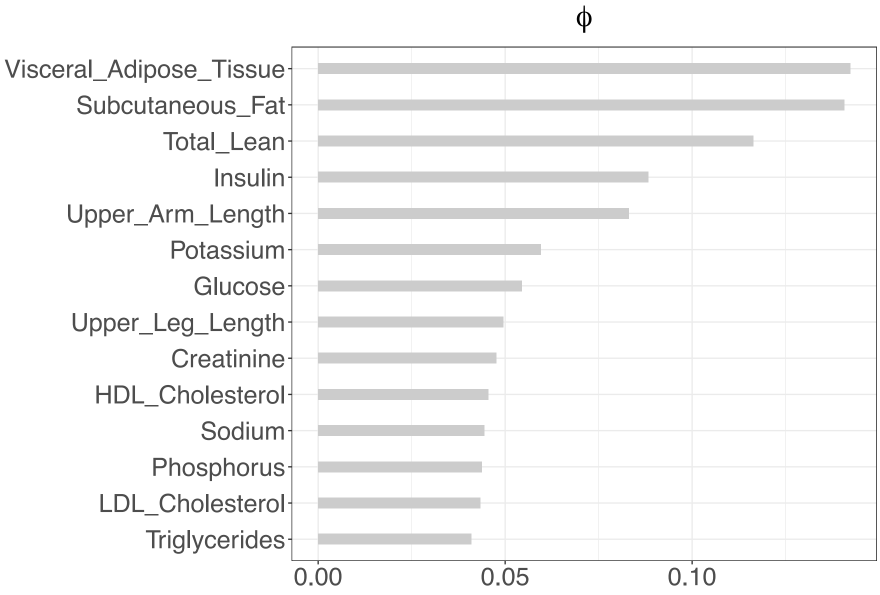


Table of Contents

Take-home message

- The objective of the work is to provide a procedure for describing the importance of covariates for tail values of the variable of interest.
- Shapley values are an index for this purpose, but the small sample size is a problem.
- We therefore rely on procedures for estimating the variable of interest and the dependence between covariates for tail events, to perform data augmentation and ensure an adequate sample size for correctly estimating the Shapley values.

Thank you!

amir.khorramichokami@unica.it

Bibliography

Bedford, T. and R. M. Cooke (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics 30*(4), 1031–1068.

Czado, C. and T. Nagler (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application 9*, 453–477.

Khorrami Chokami, A. and G. Rabitti (2025). A copula-based data augmentation strategy for the sensitivity analysis of extreme operational losses. *Quantitative Finance 25*(5), 841–849.

Mase, M., A. B. Owen, and B. B. Seiler (2019). Explaining black box decisions by shapley cohort refinement. *arXiv e-prints*.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn and A. W. Tucker (Eds.), *Contributions to the Theory of Games*. Princeton University Press.

The characters used were taken from the comics by Simone Albrigi – Scottecs Sio (@instasio).