Explaining probabilistic predictions on the simplex with Shapley compositions

Anonymous Authors[1]

Abstract

Shapley values have been widely used for explaining a machine learning model's prediction by measuring the contribution of each feature to the prediction. Coming from game theory, the Shapley value has been designed for a one-dimensional function's codomain. However, for a multiclass probabilistic classifier, the output is a discrete probability distribution, over a set of more than two possible classes, and lives on a multidimensional simplex. In this case, the Shapley values are sometimes computed on each output dimension one-by-one, in an implicit one-vs-rest setting, ignoring the compositional nature of the output distribution. Indeed, elements of the simplex are known as compositional data and a discrete probability distribution can therefore be treated as such taking into account the relative information between probabilities. Using the Aitchison geometry of the simplex, this paper presents an initiative for a multidimensional extension on the simplex of the concept of Shapley value, named Shapley composition, for explaining probabilistic predictions in machine learning.

1. Introduction

Many machine learning approaches—like the one based on deep learning or random forests—are often regarded as black-boxes making them unreliable for real-life applications where the model's predictions have to be understood or explainable. These recent years, the interest in more explainable models and explainability methods has therefore increased in the machine learning literature (Angelov et al., 2021). One group of approaches, known as local explanation, aims to measure the contribution of each input feature to the computation of the model's output. Shapley values are widely used for this purpose (Štrumbelj & Kononenko, 2014; Datta et al., 2016), especially since the release of the SHAP toolkit (Lundberg & Lee, 2017)[1]. The Shapley values come from cooperative game theory where a group of players work together to maximize a payoff. A set of Shapley values uniquely distributes the payoff over all the players according to their individual contribution to the total. Shapley value is known for respecting a set of desired axiomatic properties (Shapley et al., 1953). For explaining a machine learning model's prediction, a player becomes a feature and the total payoff is the scalar output of the model.

The concept of Shapley value is designed for a one-dimensional function's codomain. In game theory, the function takes a coalition of players and gives a gain or payoff. In machine learning, the function takes a group of features from an instance and gives a one-dimensional output or prediction. For two-classes classification models, the output is one-dimensional and usually given by a sigmoid transformation. The Shapley value framework can therefore simply be applied to the logit transformation of the model output. The logit transformation maps the sigmoid codomain $]0, 1[$ to the real line $\mathbb{R}$. However, for classification tasks with more than two possible classes, the output of the model is multidimensional. This can be a discrete probability distribution or the output of a softmax layer commonly used in neural-networks. For a N-classes problem, these outputs live on a $(N-1)$-dimensional simplex. In this case, the Shapley value framework cannot be directly applied. Some apply the framework on each output probability one-by-one ignoring the structure of the simplex and the relative information between the probabilities.

Compositional data analysis (Aitchison, 1982; Pawlowsky-Glahn et al., 2015) aims at treating data—known as compositions—living on a simplex. Compositional data analysis tools have been applied for instance on geological, and chemical data but also to discrete probability distributions (Egozcue &

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] https://github.com/shap/shap

Pawlowsky-Glahn, 2011; Egozcue & Vera, 2018; Noé, 2023). This paper presents the use of the Aitchison geometry of the simplex, coming from the field of compositional data analysis, for extending the concept of Shapley value to the probability simplex, i.e. the space of discrete probability distributions. This can be used to properly explain multiclass probabilistic prediction.

The next section of the paper, Section 2, recalls the standard definition of the Shapley values, as applied to one-dimensional predictions in machine learning. Section 3 presents the necessary tools from compositional data analysis, in particular, the Aitchison geometry of the simplex. Section 4 is the main contribution of the paper which is mostly theoretical. It defines, using the Aitchison geometry, the concept of Shapley composition for extending the Shapley value framework to the multidimensional simplex. Section 5 shows, with simple examples and visualizations, how Shapley compositions can be used for explaining multidimensional probabilistic predictions in machine learning. Section 6 provides additional remarks and concludes the work.

2. The Shapley values in machine learning

This section recalls the theoretical formulation of the Shapley values for measuring the contribution of each feature on a machine learning prediction.

Let $f : \mathcal{X} \to \mathbb{R}$ be a learned model one wants to locally explain where $f(\boldsymbol{x})$ is the prediction on the instance $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$. Let Pr be the probability distribution over $\mathcal{X}$ of the data[2]. Let $S \subseteq \mathcal{I} = \{1, 2, \dots d\}$ be a subset of indices where d is the number of features. $\boldsymbol{x}_S$ refers to an instance $\boldsymbol{x}$ restricted to the features indicated by the indices in S.

When an instance $\boldsymbol{x}$ is observed, the expected value of the prediction is simply $\mathbb{E}[f(\boldsymbol{x}) \mid \boldsymbol{x}] = f(\boldsymbol{x})$. However, when only $\boldsymbol{x}_S$ is given with $\mathcal{S} \neq \mathcal{I}$, there is uncertainty about the non-observed features and we therefore compute the expected prediction given $\boldsymbol{x}_S$: $\mathbb{E}_{\mathrm{Pr}}[f(\boldsymbol{x}) \mid \boldsymbol{x}_S] = \int_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})\mathrm{Pr}(\boldsymbol{x} \mid \boldsymbol{x}_S)d\boldsymbol{x}$. The contribution of the feature indexed by $i \notin S$ in the prediction $f(\boldsymbol{x})$ given the known features indexed by S is given by:

$$c_{f,\boldsymbol{x},\mathrm{Pr}}(i, \boldsymbol{X}_S) = v_{f,\boldsymbol{x},\mathrm{Pr}}(\boldsymbol{X}_{S\cup\{i\}}) - v_{f,\boldsymbol{x},\mathrm{Pr}}(\boldsymbol{X}_S), \quad (1)$$

where v is known as the value function:

$$\begin{aligned} v_{f,\boldsymbol{x},\mathrm{Pr}} : 2^{\mathcal{I}} &\to \mathbb{R}, \\ S &\mapsto \mathbb{E}_{\mathrm{Pr}}[f(\boldsymbol{x}) \mid \boldsymbol{x}_S], \end{aligned} \quad (2)$$

where $2^{\mathcal{I}}$ is the set of all subsets of $\mathcal{I}$. This measures the contribution of the ith features with a particular coalition of features indexed by S. The whole contribution of the ith feature is computed by averaging this quantity over all possible coalitions as follows:

$$\phi_{f,\boldsymbol{x},\mathrm{Pr}}(i) = \frac{1}{d!}\sum_{\pi} c_{f,\boldsymbol{x},\mathrm{Pr}}(i, \pi_{\boldsymbol{X}}^{<i}), \quad (3)$$

where π is a permutation of the set $\mathcal{I}$ of indexes and $\pi_{\boldsymbol{X}}^{<i}$ is the features of $\boldsymbol{X}$ coming before the ith feature in the ordering given by π. For better clarity, the subscript $_{f,\boldsymbol{x},\mathrm{Pr}}$ will be dropped from the equations.

This quantity is known as the Shapley value for the ith feature. It comes from cooperative game theory and is known to be the only quantity respecting a set of desired axiomatic properties (Shapley et al., 1953). It is linear as a function of the model ($\alpha, \beta \in \mathbb{R}$): $\phi_{\alpha f+\beta g}(i) = \alpha\phi_f(i) + \beta\phi_g(i)$, and the "centered" learned model is additively separable with respect to the Shapley values:

$$f(\boldsymbol{x}) - \mathbb{E}_{\mathrm{Pr}}[f(\boldsymbol{X})] = \sum_{i=1}^{d} \phi_f(i), \quad (4)$$

which is known as the efficiency property.

The Shapley value is designed for a one-dimensional codomain of the function f. For explaining machine learning models which output multidimensional discrete probability distributions, like in multiclass classification, some people have been explaining each output dimension one-by-one, applying a logit transformation to the probabilities, resulting in a one-vs-rest comparison. However, this approach ignores the relative information between each probability and ignores the compositional nature of the discrete probability distributions. Indeed, the probabilistic output of a classifier lives on a multidimensional simplex. The latter is the sample space of compositional data briefly reviewed in the next section.

3. Compositional data

Compositional data carries relative information. Each element of a composition describes a part of some whole (Pawlowsky-Glahn et al., 2015) like vectors of proportions, concentrations, and discrete probability distributions. A N-part composition is a vector of N non-zero positive real numbers that sum to a constant k. Each element of the vector is a part of the whole k. The sample space of compositional data is the simplex: $\mathcal{S}^N = \left\{\boldsymbol{x} = [x_1, x_2, \dots x_N]^T \in \mathbb{R}_+^{*N} \,\middle|\, \sum_{i=1}^{N} x_i = k\right\}$. In a composition, only the relative information between

[2] Usually, this is unknown but expectations will be replaced by empirical averagings.

parts matters and John Aitchison introduced the use of log-ratios of parts to handle this (Aitchison, 1982). He defined several operations on the simplex which leads to what is called the Aitchison geometry of the simplex.

3.1. The Aitchison geometry of the simplex

John Aitchison defined an internal operation called perturbation, an external one called powering and an inner product (Aitchison, 2001):

- perturbation:

  $$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}\left([x_1 y_1, \dots x_N y_N]\right),$$

  seen as an addition between two compositions $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^N$,

- powering:

  $$\alpha \odot \boldsymbol{x} = \mathcal{C}\left([x_1^\alpha, \dots x_N^\alpha]\right),$$

  seen as a multiplication by a scalar $\alpha \in \mathbb{R}$,

- inner product:

  $$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j},$$

where $\mathcal{C}(\cdot)$ is the closure operator. Since only the relative information matter, scaling factors are irrelevant and a composition $\boldsymbol{x}$ is equivalent to $\lambda \boldsymbol{x} = [\lambda x_1, \lambda x_2, \dots \lambda x_N]$ for all $\lambda > 0$. This equivalence is materialized by the closure operator defined for $k > 0$ as: $\mathcal{C}\left(\boldsymbol{x}\right) = \left[\frac{k x_1}{\|\boldsymbol{x}\|_1}, \frac{k x_2}{\|\boldsymbol{x}\|_1}, \dots \frac{k x_N}{\|\boldsymbol{x}\|_1}\right]^T$, where $\boldsymbol{x} \in \mathbb{R}_+^{*N}$ and $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{N} |x_i|$.

This gives to the simplex a $(N-1)$-dimensional Euclidean vector space structure called Aitchison geometry of the simplex. In this paper, since we are interested in classifiers' outputs as discrete probability distributions, we restrict ourselves to the probability simplex where $k = 1$.

3.2. The isometric log-ratio transformation

An $(N-1)$-dimensional orthonormal basis of the simplex, referred to as an Aitchison orthonormal basis, can be built. The projection of a composition into this basis defines an isometric isomorphism between $\mathcal{S}^N$ and $\mathbb{R}^{N-1}$. This is known as an Isometric-Log-Ratio (ILR) transformation (Egozcue et al., 2003) and allows to express a composition into a Cartesian coordinates system preserving the metric of the Aitchison geometry. Within this real space, the permutation, the powering and the Aitchison inner product defined above are respectively the standard addition between two vectors, the multiplication of a vector by a scalar, and the standard inner product.

Given a composition $\boldsymbol{p} = [p_1, \dots p_N]^T \in \mathcal{S}^N$ we write its ILR transformation as $\tilde{\boldsymbol{p}} = \operatorname{ilr}(\boldsymbol{p}) = [\tilde{p}_1, \dots \tilde{p}_{N-1}]^T \in \mathbb{R}^{N-1}$. The ith element $\tilde{p}_i$ of $\tilde{\boldsymbol{p}}$ is obtained as: $\tilde{p}_i = \langle \boldsymbol{p}, \boldsymbol{e}^{(i)} \rangle_a$ where the set $\{\boldsymbol{e}^{(i)} \in \mathcal{S}^N, i = 1, \dots N-1\}$ forms an Aitchison orthonormal basis of the simplex. The choice of the basis will be discussed in Section 5.2.

4. Shapley composition on the simplex

In Section 2 we have briefly presented the standard definition of the Shapley values in the context of machine learning one-dimensional predictions explanation. In this Section, we will see how the Aitchison geometry can be used to extend the concept of Shapley value to the simplex for explaining multidimensional probabilistic predictions.

Let $\boldsymbol{f} : \mathcal{X} \to \mathcal{S}^N$ be a learned model, like a N-classes probabilistic classifier for instance, which outputs a probabilistic prediction on the $(N-1)$-dimensional probability simplex S^N. In order to properly consider the structure of the simplex and the relative information between the probabilities, the output of the model must be treated as compositional data using the operators and metrics defined by the Aitchison geometry of the simplex. We therefore rewrite the contribution and the value function of Equations 1 and 2 as follows:

$$\boldsymbol{c}_{\boldsymbol{f},\boldsymbol{x},\mathrm{Pr}}(i, \boldsymbol{X}_S) = \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\mathrm{Pr}}(\boldsymbol{X}_{S \cup \{i\}}) \ominus \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\mathrm{Pr}}(\boldsymbol{X}_S), \quad (5)$$

where $\boldsymbol{a} \ominus \boldsymbol{b}$ is the perturbation $\boldsymbol{a} \oplus ((-1) \odot \boldsymbol{b})$ which correspond to a substraction between compositions $\boldsymbol{a}$ and $\boldsymbol{b}$, and where:

$$\begin{aligned} \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\mathrm{Pr}} : 2^{\mathcal{I}} &\to \mathcal{S}^N, \\ S &\mapsto \mathbb{E}^{\mathcal{A}}_{\mathrm{Pr}}[\boldsymbol{f}(\boldsymbol{x}) \mid \boldsymbol{x}_S]. \end{aligned} \quad (6)$$

The $\mathcal{A}$ in superscript highlights the fact that the expectation is done with respect to the Aitchison measure, rather than the Lebesgue measure, which can simply be computed as: $\mathbb{E}^{\mathcal{A}}[\boldsymbol{Y}] = \operatorname{ilr}^{-1}\left(\mathbb{E}\left[\operatorname{ilr}\left(\boldsymbol{Y}\right)\right]\right)$, where $\mathbb{E}^{\mathcal{A}}$ refers to the expectation with respect to the Aitchison measure while $\mathbb{E}$ refers to the expectation with respect to the Lebesgue measure (Pawlowsky-Glahn et al., 2015).

The Shapley quantity expressing the contribution of the ith feature on a prediction can simply be expressed on the simplex as the composition $\boldsymbol{\phi}(i)$ given by:

$$\boldsymbol{\phi}_{\boldsymbol{f},\boldsymbol{x},\mathrm{Pr}}(i) = \frac{1}{d!} \odot \bigoplus_{\pi} \boldsymbol{c}_{\boldsymbol{f},\boldsymbol{x},\mathrm{Pr}}(i, \pi^{<i}_{\boldsymbol{X}}). \tag{7}$$

We call this quantity Shapley composition. It can be shown (in Appendix A) that the linearity and the efficiency properties naturally hold for the Shapley composition:

$$\begin{aligned} \boldsymbol{\phi}_{\alpha\odot\boldsymbol{f}(\boldsymbol{x})\oplus\beta\odot\boldsymbol{g}(\boldsymbol{x})}(i) &= \alpha \odot \boldsymbol{\phi}_{\boldsymbol{f}}(i) \oplus \beta \odot \boldsymbol{\phi}_{\boldsymbol{g}}(i), \\ \bigoplus_{i=1}^{d} \boldsymbol{\phi}_{\boldsymbol{f}}(i) &= \boldsymbol{f}(\boldsymbol{x}) \ominus \mathbb{E}^{\mathcal{A}}_{\mathrm{Pr}}[\boldsymbol{f}(\boldsymbol{X})]. \end{aligned} \tag{8}$$

This can be seen as a multidimensional extension of the Shapley value framework on the simplex. Here, the Shapley quantity is not a scalar anymore, this is a composition living on the probability simplex. In the next section, we will see in more detail how this can be used to explain the contribution of each feature to a multidimensional probabilistic prediction.

5. Explaining a prediction with Shapley compositions

Given a probabilistic prediction $\boldsymbol{f}(\boldsymbol{x}) \in \mathcal{S}^N$, the Shapley composition $\boldsymbol{\phi}_{\boldsymbol{f},\boldsymbol{x},\mathrm{Pr}}(i)$ describes the contribution of the ith feature on the prediction. The efficiency property shows how the probability distribution moves from the base distribution, i.e. the expected prediction regardless of the current input, to the prediction $\boldsymbol{f}(\boldsymbol{x})$. In the standard Shapley formulation recalled in Section 2, the prediction is one-dimensional such that the Shapley quantity is a scalar. In applications where there are more than two possible classes, the prediction is multidimensional such that the Shapley quantity is too. Both live in the same space: the probability simplex. In this section, we discuss how the set of Shapley compositions can be analysed to better understand the contribution and influence of each feature on the prediction.

5.1. Visualization

The Shapley compositions can be visualized in the Euclidean space isometric to the simplex thanks to the ILR transformation presented in Section 3.2. As we will see, this space has the advantage of being intuitive since it is a standard real $(N-1)$-dimensional vector space.

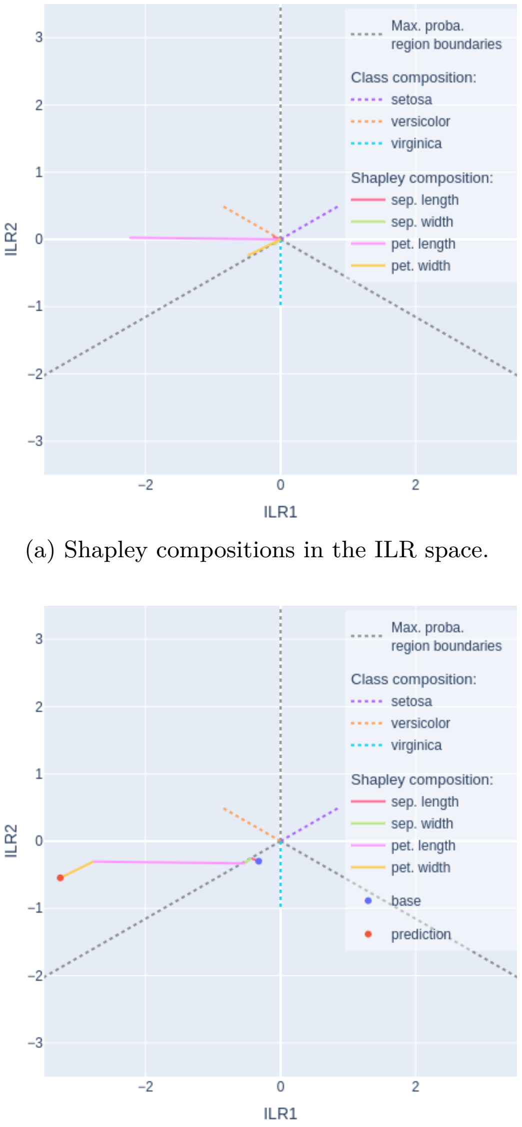
5.1.1. Three classes

In a three-classes case, the space of the predictions is 2-dimensional. We illustrate this example with the well-known Iris classification dataset consisting of a set of flowers described by 4 features: sepal length and width, and petal length and width. The aim of the classification task is to predict to which of the three species (setosa, versicolor and virginica) a flower belongs. In the present example, a Support Vector Machine (SVM) with a radial basis function (rbf) kernel and with pairwise coupling (Wu et al., 2003) is used as a probabilistic classifier. Figure 1 shows the explanation of the classifier prediction with one versicolor instance as input. Figure 1a shows the Shapley compositions in the ILR space and Figure 1b shows how they move the base distribution to the prediction. Having the highest norm, the petal length is the feature contributing the most to the prediction and moves the base distribution to the versicolor maximum probability decision region (maximum probability decision region boundaries are the dashed gray lines). Class-compositions are represented by color dashed vectors. A class-composition is defined as an unit norm composition going straight to the direction of one class and against all the others (see Appendix B for a formal definition). The Shapley composition for the petal length is orthogonal to the virginica class-composition, this suggests that this feature does not contribute to the predicted probability for this class. The Shapley composition for the petal width goes straight to the opposite direction of the setosa class-composition suggesting that this feature contributes to rejecting this class. The other Shapley compositions have a low norm suggesting that these features do not contribute to the prediction.

5.1.2. Four classes

In a four-classes example, the simplex is 3-dimensional. We illustrate this with a simple handwritten digit recognition task[3]. It consists of classifying an 8×8 image as representing one of the digits among 0, 1, 2 and 3. Since they are 64 pixels, considering each pixel as a feature would correspond to 64 Shapley compositions. This would be hard to visualize and analyse. We therefore reduce the number of features to 6 using a principal component analysis for better clarity and conciseness. A SVM with a rbf kernel and pairwise coupling is again used as a probabilistic classifier. The same explanation analysis as before can be applied here but within a 3-dimensional plot as illustrated in Figure 2. To better understand how this space is divided into four regions—each representing the maximum probability region for one class—one can think about the shape of a methane molecule. The hydrogens correspond to the vertices and the carbon to the center of a tetrahedron i.e. a 3-dimensional simplex.

[3] We use the scikit-learn's digits dataset (Pedregosa et al., 2011).



(a) Shapley compositions in the ILR space.

(b) The sum of the Shapley compositions in the ILR space from the base distribution to the prediction.

Figure 1. Shapley explaination in the ILR space for the classification of an Iris instance.

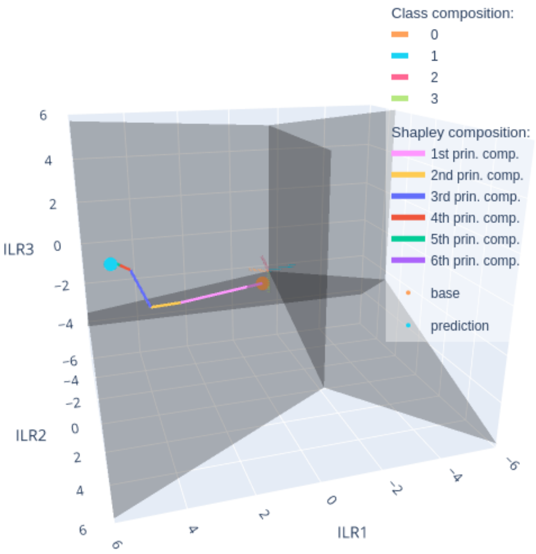


Figure 2. Shapley explaination in the 3-dimensional ILR space for a four classes digit recognition task. The Shapley compositions are summed in the ILR space from the base distribution to the prediction. The gray transparent walls markout the maximum probability decision regions.

The relative positions of the class-compositions in the ILR space are the same as the bonds between the carbon and hydrogen: the angles are ≈ 109.5°. In this example, the tested instance is a 0[4].

5.2. More classes: groups of parts and balances

When more than three classes are involved, all the dimensions of the ILR space cannot be visualized at once. However, 2 or 3-dimensional subspaces can still be visualized. In order to select the ILR components to investigate, one needs to understand what they refer to. In this section, we briefly discuss the interpretation of the ILR components.

A component of the ILR space can be interpreted as a balance, i.e. a log-ratio of two geometrical means of parts (Egozcue et al., 2003; Egozcue & Pawlowsky-Glahn, 2005; Pawlowsky-Glahn et al., 2015): one giving the central values of the probabilities in one group of classes and one for another group of classes. Therefore, a balance is here comparing the weight of two groups of classes. The set of balances is built such that they are geometrically orthogonal meaning they provide nonredundant information[5].

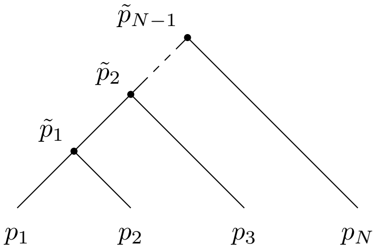
[4] More examples can be obtained from the notebooks released on the git repository: link not shared during the double-blind peer reviewing.

[5] Not to be confused with statistical uncorrelation (Pawlowsky-Glahn et al., 2015).

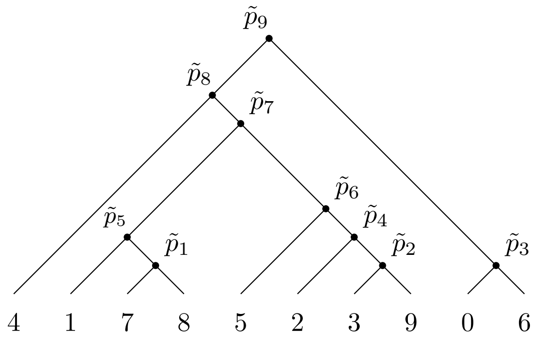
This can be illustrated by a sequential binary partition or bifurcation tree. Figure 3 gives two examples: 3a shows the bifurcation tree corresponding to the basis obtained with the Gram-Schmidt procedure as in (Egozcue et al., 2003) which is the one used in the examples of Figures 1 and 2 with respectively $N = 3$ and $N = 4$. Each node of the tree is a balance i.e. an ILR component. The first balance $\tilde{p}_1$ first compares the probability for class 1 with the probability for class 2. Each next balance then recursively compares the probability for the next class with the probabilities for the previous classes independently of all the others.

In some applications, one may be interested in particular comparisons of groups of classes not necessarily given by a basis in the form of Figure 3a. For instance, as in an example presented in (Egozcue & Pawlowsky-Glahn, 2005), if one wants to compare political parties or groups, it may be pertinent to have a balance comparing left and right-wing groups. However, there are sometimes no obvious relevant comparisons to study. For instance, in the handwritten digit recognition problem, one may want to compare odd with even numbers or prime with non-prime but being basically a shape recognition problem, and the shape of the numbers being independent of their arithmetic properties, these comparisons are not pertinent. Be that as it may, the choice of the basis must be left open, whether or not it is based on a relevant strategy.

Let's use the basis of Figure 3b for a 10-classes digit recognition task[6]. We comment, for conciseness, a single 2-dimensional subspace[7]. Let's have a look, in Figure 4, at the third and fifth ILR dimensions ($\tilde{p}_3$ and $\tilde{p}_5$). It is like saying we are only interested in comparing the probability assigned for class 0 with the probability assigned for class 6, and in comparing the probability assigned for class 1 with the group of probabilities assigned for classes 7 and 8. $\tilde{p}_3$ depends only on the probability for the digits 0 and 6 and $\tilde{p}_5$ depends only on the probabilities for the digits 1, 7 and 8. Therefore, the class-compositions for the other digits have a zero projection within this subspace and are therefore discarded in Figure 4. The class-compositions for 0 and 6 are orthogonal to the class-compositions for classes 1, 7 and 8. Indeed, the set of classes making the balance $\tilde{p}_3$ and the set of classes making $\tilde{p}_5$ have no intersection. In contrast, in the example of Figure 1, $\tilde{p}_1$ is comparing the probabilities for the class setosa with the probability for the class versicolor and $\tilde{p}_2$ is comparing the probabilities for the class virginica with the group of probabilities for setosa and versicolor. In this case, none of the class-compositions are orthogonal. In Figure 4, since $\tilde{p}_5$ is comparing 1 with the group of digits 7 and 8, the projection on this line of the class-compositions for 1 goes in an opposite direction than the one for the class-compositions for 7 and 8. The latter two are equal and half as long as the former. In this way, $\tilde{p}_5$ compares the probability for 1 with the group of probabilities for 7 and 8 with the same weight. In other words, in this subspace, the class-compositions for 7 and 8 are reweighted such that this group of two classes has the same weight as the group made of the single class 1.



(a) Bifurcation tree corresponding to the basis obtained with the Gram-Schmidt procedure as in (Egozcue et al., 2003) and used in the examples of Figures 1 and 2.



(b) Bifurcation tree used in the 10-classes digit recognition task discussed in Section 5.2 and in Figure 4.

Figure 3. Two examples of bifurcation tree.

Within this space, Shapley compositions can be explored like in the examples of Figures 1 and 2 keeping in mind that this is a subset of the full ILR space.

5.3. Angles, norms and projections

Some may find the analysis of the features contributions in cases with more than four classes tricky. In-

[6] In this example, the bifurcation tree is obtained with agglomerative clustering of classes by recursively merging a pair of classes based on the Mahalanobis distance in the classifier's output space, assuming that within a pair of classes, the class-conditional densities are logistic-normal (Aitchison & Shen, 1980) with the same covariance matrix.

[7] More visualisations can be obtaned from the shared python notebooks.

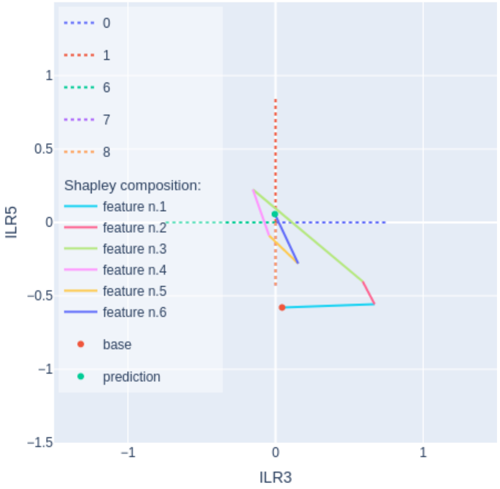


Figure 4. The sum of the Shapley compositions from the base to the prediction in the ILR subspace made of $\tilde{p}_3$ and $\tilde{p}_5$ for a test instance from class 2. $\tilde{p}_3$ compares the probability assigned for class 0 with the probability assigned for class 6 and $\tilde{p}_5$ compares the probability assigned for class 1 with the group of probabilities assigned for class 7 and 8. The color dashed vectors represent the class-compositions with non-zero projection.

deed, in this case, the ILR space cannot be visualized in a 2 or 3-dimensional plot and as discussed in Section 5.2, choosing which subspaces to visualize require a careful understanding of the balances and a careful building of the bifurcation tree. However, the Shapley explaination can be summarized by sets of angles, norms and projections. The norm of a Shapley composition gives the strength of the feature's contribution in the prediction. It gives the overall contribution of the feature on the prediction, regardless of its direction. The angle between two Shapley compositions can informs about their orthogonality. If two Shapley composition are orthogonal, this suggest the features are nonredundant. A negative angle would suggest that the features have an opposite influence on the prediction. The projection of a Shapley composition on the set of class-compositions informs in favor of, or against, which classes a feature is contributing. Appendix C, provides some examples of summarizing a Shapley explanation using the norm of Shapley compositions, angles between them and their projection on the class-compositions.

5.4. Histograms

If one found hard to visualize the proposed Shapley explanation in the ILR space, the Shapley composition can be visualized as histograms like discrete probability distributions. Figure 5 shows the Shapley compositions of the Iris classification example. The more uniform the histogram is, like for the sepal length and width, the less the contribution of the feature is. In opposite, the histogram for the petal length as a high value for the versicolor class, relatively to the others, confirming the contribution of the feature toward this class. For the petal width, the value for the class setosa is low relatively to the others which confirms the constribution of this feature against the class setosa.

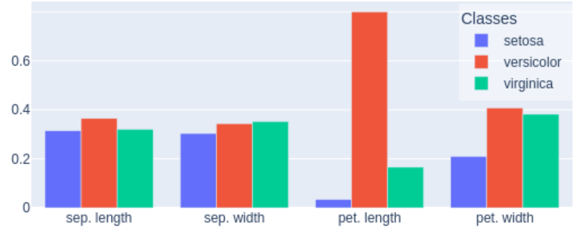


Figure 5. Shapley compositions visualized as histograms for the Iris classification example.

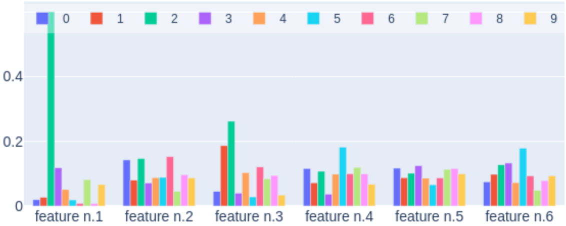


Figure 6. Shapley compositions visualized as histograms for the seven classes digit recognition example.

As another illustration, Figure 6 shows the Shapley compositions of the 10-classes digit recognition example. Contrary to the visualization of the compositions within the ILR space as discussed in Section 5.2, here, one can analyses all parts of each compositions within a single plot. In this example, the high of the class 2 for the first principal component confirms the contribution of this features toward this class.

5.5. About our implementation

In this work, the estimation algorithm we used to compute the Shapley compositions is an adaptation of Algorithm 2 in (Štrumbelj & Kononenko, 2014). Since the resulting Shapley compositions are approximations, the efficiency property does not necessarily hold. Each Shapley composition is therefore corrected following a similar method as in the sampling approximation in the SHAP toolkit (Lundberg & Lee, 2017)[8] to respect the efficiency propery. See Appendix D and

[8] https://github.com/shap/shap/blob/master/shap/explainers/_sampling.py

E for more details[9].

6. Discussion and conclusion

The use of standard Shapley value for explaining machine learning model with multidimensional probabilistic output has been rarely discussed in the literature. However, people computing a Shapley value on each output dimension one-by-one can be encounter. To be more precise, one may first found natural to compute a Shapley value on the logit of the probability for each classes resuilting in a N-dimensional vector of the Shapley values for a N-classes problem. Such vector will be refered here as Shapley vector. Even if the efficiency property holds, i.e. the sum of the element-wise logit of the base distribution with the Shapley vectors for each feature is equal to the element-wise logit of the prediction, the path from the base to the prediction may go out of the simplex which is conterintuitive. Indeed, this approach ignore the compositional nature of discrete probability distribution output by a classifier. Moreover, such strategy would require to run N explanation model contrary to our approach wich require a single explanation model.

As far as we know, this paper is the first work proposing an extension of the Shapley value framework to the multidimensional simplex for explaining probabilistic predictions in machine learning. We saw how the formalization of the standard Shapley value naturally extend to the simplex using the Aitchison geometry. The Shapley value therefore becomes a Shapley composition or distribution, i.e. a vector living on the probability simplex. Such Shapley composition informs on the contribution of a feature on the prediction. It tells how the feature moves the distribution from the base one to the predicted one on the simplex. The Aitchison geometry defines on the simplex an Euclidean vector space structure such that the Shapley compositions can be analysed through angles, norms and projections.

The machine learning literature about the use of Shapley value is prolific. Many estimation algorithms have been developped, many applications of the Shapley value have emerged, and large scales experiments have been conducted. In contrast, our paper presents limited experimental results as simple proofs of concept and illustrations. However, we believe this work lies down proper theoretical foundations to foster researches in explainable machine learning especially for multiclass classification.

[9] The estimation algorithm we use assumes that the features are independent. We leave the exploration of estimation algorithms with no features-independence assumption for future work.

References

Aitchison, J. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological), 44(2):139–177, 1982.

Aitchison, J. Simplicial inference. In Marlos A. G. Viana, D. S. P. R. e. A. S. S. o. A. M. i. S. (ed.), Algebraic Methods in Statistics and Probability, Contemporary Mathematics 287. American Mathematical Society, 2001.

Aitchison, J. and Shen, S. M. Logistic-normal distributions: Some properties and uses. Biometrika, 67 (2):261–272, 1980.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5):e1424, 2021.

Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE Symposium on Security and Privacy (SP), pp. 598–617, 2016. doi: 10.1109/SP.2016.42.

Egozcue, J. J. and Pawlowsky-Glahn, V. Groups of parts and their balances in compositional data analysis. Mathematical Geology, 37(7):795–828, 2005.

Egozcue, J. J. and Pawlowsky-Glahn, V. Evidence information in bayesian updating. Proc. International Workshop on Compositional Data Analysis, 05 2011.

Egozcue, J. J. and Vera, P.-G. Evidence functions: a compositional approach to information. SORT-Statistics and Operations Research Transactions, 1 (2):101–124, Dec. 2018.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. Isometric logratio transformations for compositional data analysis. Mathematical geology, 35(3):279–300, 2003.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc., 2017.

Noé, P.-G. Representing evidence for attribute privacy: bayesian updating, compositional evidence and calibration. PhD thesis, Université d'Avignon, 2023.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. Modeling and Analysis of Compositional Data. John Wiley & Sons, 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011.

Shapley, L. S. et al. A value for n-person games. pp. 307–317, 1953.

Štrumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, 41:647–665, 2014.

Wu, T.-F., Lin, C.-J., and Weng, R. Probability estimates for multi-class classification by pairwise coupling. Advances in Neural Information Processing Systems, 16, 2003.

A. Linearity and efficiency of the Shapley composition

In this section, we show the linearity of the Shapley composition with respect to the model prediction, and the efficiency property.

A.1. Linearity

The Shapley composition is linear, within the Aitchison geometry of the simplex, with respect to linear combination of models' predictions.

Proof. Let's consider the linear combination of predictions $\boldsymbol{h}(\boldsymbol{x}) = \alpha \odot \boldsymbol{f}(\boldsymbol{x}) \oplus \beta \odot \boldsymbol{g}(\boldsymbol{x})$. we want to check if:

$$\boldsymbol{\phi}_{\boldsymbol{h}}(i) = \alpha \odot \boldsymbol{\phi}_{\boldsymbol{f}}(i) \oplus \beta \odot \boldsymbol{\phi}_{\boldsymbol{g}}(i). \tag{9}$$

We have:

$$\begin{aligned}
\mathbb{E}^{\mathcal{A}}_{\Pr}[\boldsymbol{h}(\boldsymbol{x}) \mid \boldsymbol{x}_S] &= \mathrm{ilr}^{-1}\left(\mathbb{E}_{\Pr}[\mathrm{ilr}\left(\alpha \odot \boldsymbol{f}(\boldsymbol{x}) \oplus \beta \odot \boldsymbol{g}(\boldsymbol{x})\right) \mid \boldsymbol{x}_S]\right),\\
&= \mathrm{ilr}^{-1}\left(\mathbb{E}_{\Pr}[\alpha\, \mathrm{ilr}\left(\boldsymbol{f}(\boldsymbol{x})\right) + \beta\, \mathrm{ilr}\left(\boldsymbol{g}(\boldsymbol{x})\right) \mid \boldsymbol{x}_S]\right),\\
&= \mathrm{ilr}^{-1}\left(\alpha \mathbb{E}_{\Pr}[\mathrm{ilr}\left(\boldsymbol{f}(\boldsymbol{x})\right) \mid \boldsymbol{x}_S] + \beta \mathbb{E}_{\Pr}[\mathrm{ilr}\left(\boldsymbol{g}(\boldsymbol{x})\right) \mid \boldsymbol{x}_S]\right),\\
&= \alpha \odot \mathrm{ilr}^{-1}\left(\mathbb{E}_{\Pr}[\mathrm{ilr}\left(\boldsymbol{f}(\boldsymbol{x})\right) \mid \boldsymbol{x}_S]\right) \oplus \beta \odot \mathrm{ilr}^{-1}\left(\mathbb{E}_{\Pr}[\mathrm{ilr}\left(\boldsymbol{g}(\boldsymbol{x})\right) \mid \boldsymbol{x}_S]\right),\\
&= \alpha \odot \mathbb{E}^{\mathcal{A}}_{\Pr}[\boldsymbol{f}(\boldsymbol{x}) \mid \boldsymbol{x}_S] \oplus \beta \odot \mathbb{E}^{\mathcal{A}}_{\Pr}[\boldsymbol{g}(\boldsymbol{x}) \mid \boldsymbol{x}_S].
\end{aligned} \tag{10}$$

Therefore, $\boldsymbol{v}_{\boldsymbol{h},\boldsymbol{x},\Pr}(\boldsymbol{X}_S) = \alpha \odot \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(\boldsymbol{X}_S) \oplus \beta \odot \boldsymbol{v}_{\boldsymbol{g},\boldsymbol{x},\Pr}(\boldsymbol{X}_S)$, meaning that $\boldsymbol{v}$ is linear with respect to the learned function or model. The linearity of the contribution $\boldsymbol{c}$ naturally follows:

$$\begin{aligned}
\boldsymbol{c}_{\boldsymbol{h},\boldsymbol{x},\Pr}(i, \boldsymbol{X}_S) &= \boldsymbol{v}_{\boldsymbol{h},\boldsymbol{x},\Pr}(\boldsymbol{X}_{S\cup\{i\}}) \ominus \boldsymbol{v}_{\boldsymbol{h},\boldsymbol{x},\Pr}(\boldsymbol{X}_S),\\
&= \big(\alpha \odot \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(\boldsymbol{X}_{S\cup\{i\}}) \oplus \beta \odot \boldsymbol{v}_{\boldsymbol{g},\boldsymbol{x},\Pr}(\boldsymbol{X}_{S\cup\{i\}})\big) \ominus \left(\alpha \odot \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(\boldsymbol{X}_S) \oplus \beta \odot \boldsymbol{v}_{\boldsymbol{g},\boldsymbol{x},\Pr}(\boldsymbol{X}_S)\right),\\
&= \alpha \odot \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(\boldsymbol{X}_{S\cup\{i\}}) \oplus \beta \odot \boldsymbol{v}_{\boldsymbol{g},\boldsymbol{x},\Pr}(\boldsymbol{X}_{S\cup\{i\}}) \ominus \alpha \odot \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(\boldsymbol{X}_S) \ominus \beta \odot \boldsymbol{v}_{\boldsymbol{g},\boldsymbol{x},\Pr}(\boldsymbol{X}_S),\\
&= \alpha \odot \big(\boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(\boldsymbol{X}_{S\cup\{i\}}) \ominus \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(\boldsymbol{X}_S)\big) \oplus \beta \odot \big(\boldsymbol{v}_{\boldsymbol{g},\boldsymbol{x},\Pr}(\boldsymbol{X}_{S\cup\{i\}}) \ominus \boldsymbol{v}_{\boldsymbol{g},\boldsymbol{x},\Pr}(\boldsymbol{X}_S)\big),\\
&= \alpha \odot \boldsymbol{c}_{\boldsymbol{f},\boldsymbol{x},\Pr}(i, \boldsymbol{X}_S) \oplus \beta \odot \boldsymbol{c}_{\boldsymbol{g},\boldsymbol{x},\Pr}(i, \boldsymbol{X}_S).
\end{aligned} \tag{11}$$

And the linearity of the Shap composition:

$$\begin{aligned}
\boldsymbol{\phi}_{\boldsymbol{h}}(i) &= \frac{1}{d!}\bigoplus_{\pi} \boldsymbol{c}_{\boldsymbol{h},\boldsymbol{x},\Pr}(i, \pi^{<i}_{\boldsymbol{X}}),\\
&= \frac{1}{d!}\bigoplus_{\pi} \left(\alpha \odot \boldsymbol{c}_{\boldsymbol{f},\boldsymbol{x},\Pr}(i, \boldsymbol{X}_S) \oplus \beta \odot \boldsymbol{c}_{\boldsymbol{g},\boldsymbol{x},\Pr}(i, \boldsymbol{X}_S)\right),\\
&= \alpha \odot \left(\frac{1}{d!}\bigoplus_{\pi} \boldsymbol{c}_{\boldsymbol{f},\boldsymbol{x},\Pr}(i, \boldsymbol{X}_S)\right) \oplus \beta \odot \left(\frac{1}{d!}\bigoplus_{\pi} \boldsymbol{c}_{\boldsymbol{g},\boldsymbol{x},\Pr}(i, \boldsymbol{X}_S)\right),\\
&= \alpha \odot \boldsymbol{\phi}_{\boldsymbol{f}}(i) \oplus \beta \odot \boldsymbol{\phi}_{\boldsymbol{g}}(i).
\end{aligned} \tag{12}$$

□

A.2. Efficiency

The efficiency property naturally holds for Shapley compositions within the Aitchison geometry.

Proof.

$$
\begin{aligned}
\bigoplus_{i=1}^{d} \boldsymbol{\phi}_{\boldsymbol{f}}(i) &= \bigoplus_{i=1}^{d} \left(\frac{1}{d!} \odot \bigoplus_{\pi} \boldsymbol{c}(i, \pi_{\boldsymbol{X}}^{<i}) \right), \\
&= \frac{1}{d!} \odot \bigoplus_{i=1}^{d} \left(\bigoplus_{\pi} \left(\boldsymbol{v}(\pi_{\boldsymbol{X}}^{<i+1}) \ominus \boldsymbol{v}(\pi_{\boldsymbol{X}}^{<i}) \right) \right), \\
&= \frac{1}{d!} \odot \bigoplus_{i=1}^{d} \left(\underbrace{\left(\bigoplus_{\pi} \boldsymbol{v}(\pi_{\boldsymbol{X}}^{<i+1}) \right)}_{\boldsymbol{A}_{i+1}} \ominus \underbrace{\left(\bigoplus_{\pi} \boldsymbol{v}(\pi_{\boldsymbol{X}}^{<i}) \right)}_{\boldsymbol{A}_{i}} \right), \\
&= \frac{1}{d!} \odot \bigoplus_{i=1}^{d} \left(\boldsymbol{A}_{i+1} \ominus \boldsymbol{A}_{i} \right), \qquad (13) \\
&= \frac{1}{d!} \odot \left(\boldsymbol{A}_{d+1} \ominus \boldsymbol{A}_{1} \right), \text{ since we have a telescoping perturbation,} \\
&= \frac{1}{d!} \odot \left(\left(\bigoplus_{\pi} \boldsymbol{v}(\pi_{\boldsymbol{X}}^{<d+1}) \right) \ominus \left(\bigoplus_{\pi} \boldsymbol{v}(\pi_{\boldsymbol{X}}^{<1}) \right) \right), \\
&= \frac{1}{d!} \odot \left(\left(\bigoplus_{\pi} \boldsymbol{v}(\boldsymbol{X}) \right) \ominus \left(\bigoplus_{\pi} \boldsymbol{v}(\boldsymbol{X}_{\emptyset}) \right) \right), \\
&= \boldsymbol{v}(\boldsymbol{X}) \ominus \boldsymbol{v}(\boldsymbol{X}_{\emptyset}), \text{ since } d! \text{ is the number of different permutation,} \\
&= \boldsymbol{f}(\boldsymbol{x}) \ominus \mathbb{E}_{\mathrm{Pr}}^{\mathcal{A}}[\boldsymbol{f}(\boldsymbol{X})].
\end{aligned}
$$

□

B. Class-compositions

A k-class-compositions $\boldsymbol{c}^{(k)} \in \mathcal{S}^N$ is defined as an unit norm composition going straight to the direction of the kth class. This is a discrete probability distribution with maximum probability for the kth class and uniform values for the others. The ith part of the $\boldsymbol{c}^{(k)}$ is:

$$c_i^{(k)} = \begin{cases} 1-(N-1)p, & \text{if } i = k \\ p, & \text{otherwise,} \end{cases} \tag{14}$$

where $p < \frac{1}{N}$. We want the Aitchison norm of each class-composition to be one:

$$\begin{aligned}
\forall k \in \{1, \dots N\}, \qquad \|\boldsymbol{c}^{(k)}\|_a = 1 &\iff \sqrt{\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\log\frac{c_i^{(k)}}{c_j^{(k)}}\right)^2} = 1, \\
&\text{for clarity,} \\
\text{we drop the } (k) \text{ from the equations,} \\
&\iff \sqrt{\frac{1}{2N}\sum_{i=1}^{N}\left((N-1)\left(\log\frac{c_i}{p}\right)^2 + \left(\log\frac{c_i}{1-(N-1)p}\right)^2\right)} = 1, \\
&\iff \sqrt{\frac{1}{2N}2(N-1)\left(\log\frac{p}{1-(N-1)p}\right)^2} = 1, \\
\text{since } p < \frac{1}{N} \text{ and the norm should be positive:} \\
&\iff \sqrt{\frac{N-1}{N}}\log\frac{1-(N-1)p}{p} = 1, \\
&\iff p = \frac{\exp\left(-\sqrt{\frac{N}{N-1}}\right)}{1+(N-1)\exp\left(-\sqrt{\frac{N}{N-1}}\right)}.
\end{aligned} \tag{15}$$

To summarize, the ith part of a k-class-compositions $\boldsymbol{c}^{(k)} \in \mathcal{S}^N$ is given by:

$$c_i^{(k)} = \frac{1}{1+(N-1)\exp\left(-\sqrt{\frac{N}{N-1}}\right)}\left(\begin{cases} 1, & \text{if } i = k \\ \exp\left(-\sqrt{\frac{N}{N-1}}\right), & \text{otherwise,} \end{cases}\right). \tag{16}$$

In this way, $\boldsymbol{c}^{(k)}$ is going straight to the direction of class k and uniformly against all the others.

Table 1. Norm of the Shapley compositions, projection on the class-compositions and cosine similarity betweeen the Shapley compositions for the Iris classification example.

	norm	projection on the class-compositions			cosine similarity between Shapley compositions			
		setosa	versicolor	virginica	pet. length	pet. width	sep. length	sep. width
pet. length	2.20	-1.91	1.90	0.02	1	·	·	·
pet. width	0.51	-0.51	0.29	0.22	0.90	1	·	·
sep. length	0.13	-0.08	0.13	-0.05	0.93	0.69	1	·
sep. width	0.11	-0.11	0.04	0.07	0.77	0.97	0.49	1

Table 2. Norm of the Shapley compositions and Cosine similarity betweeen the Shapley composition for the 10-classes digit recognition example.

	norm	cosine similarity between Shapley compositions					
		feature n.1	feature n.2	feature n.3	feature n.4	feature n.5	feature n.6
feature n.1	4.00	1	·	·	·	·	·
feature n.2	1.15	-0.17	1	·	·	·	·
feature n.3	2.25	0.24	0.23	1	·	·	·
feature n.4	1.26	-0.23	0.23	0.09	1	·	·
feature n.5	0.60	0.18	-0.17	-0.01	-0.56	1	·
feature n.6	1.11	0.14	0.35	-0.19	-0.03	-0.48	1

C. Summarizing the explanation with norms, angles and projections of Shapley compositions

The Table 1 gives, for the Iris classification example of Figure 1, the norm of each Shapley compositions, the cosine similarity between them and their projection on the set of class-compositions. Having the highest norm, this confirms that the petal length is the feature that contributed the most on the prediction. Having a close to zero inner product with the virginica class-composition, shows that this feature did not contribute to the value of the probability of of this class. It contributes neither in favor, nor in the reject of this class. Having a positive inner product with the versicolor class-composition and a negative one with the setosa class-composition, this suggest that this feature is going in favor of the class versicolor and against setosa.

Table 2 shows, for the 10-classes digit recognition example, the norm of each Shapley composition and the cosine similarity between the, and Table 3 shows their projection of the set of class-compositions.
Add details...

Table 3. Projection of the Shapley compositions on the class-compositions for the 10-classes digit recognition example.

	projection on the class-compositions									
	0	1	2	3	4	5	6	7	8	9
feature n.1	1.65	-0.85	0.09	-0.57	-1.92	-0.34	1.74	1.96	0.22	-1.81
feature n.2	-0.50	0.35	0.11	0.08	-0.57	-0.29	-0.10	0.33	-0.14	0.74
feature n.3	0.36	0.45	0.23	1.20	-1.37	-0.56	0.91	-0.83	-0.31	-0.09
feature n.4	0.19	0.02	0.61	0.18	-0.53	0.46	-0.86	-0.17	-0.12	0.22
feature n.5	-0.03	-0.02	-0.22	-0.31	-0.02	0.13	0.42	-0.16	0.14	0.07
feature n.6	-0.17	0.34	-0.22	0.17	-0.12	-0.43	-0.50	0.75	0.34	-0.18

D. Estimation of the Shapley compositions

In this section, we present an adaptation of Algorithm 2 from (Štrumbelj & Kononenko, 2014) we used for the estimation the Shapley compositions.

Let d be the number of features. We want to optimally distribute the $m_{\max}$ drawn samples over the d features. Let $\hat{\boldsymbol{\phi}}_i$ be the estimation of the Shapley composition for the ith feature. We want to minimize the sum of squared errors: $\sum_{i=1}^{d} \|\hat{\boldsymbol{\phi}}_i \ominus \boldsymbol{\phi}_i\|_a^2$. Since $\hat{\boldsymbol{\phi}}_i$ is a (Aitchison) sample mean we have: $\tilde{\hat{\boldsymbol{\phi}}}_i \approx \mathcal{N}(\tilde{\boldsymbol{\phi}}_i, \frac{1}{m_i}\boldsymbol{\Sigma}^{(i)})$ and $\tilde{\hat{\boldsymbol{\phi}}}_i - \tilde{\boldsymbol{\phi}}_i \approx \mathcal{N}(\mathbf{0}, \frac{1}{m_i}\boldsymbol{\Sigma}^{(i)})$ where the tilde refers to the ILR transformation. Let $\boldsymbol{\Delta}_{\boldsymbol{i}} = \tilde{\hat{\boldsymbol{\phi}}}_i - \tilde{\boldsymbol{\phi}}_i$ and $Z_i = \|\hat{\boldsymbol{\phi}}_i \ominus \boldsymbol{\phi}_i\|_a = \|\tilde{\hat{\boldsymbol{\phi}}}_i - \tilde{\boldsymbol{\phi}}_i\|_2 = \|\boldsymbol{\Delta}_i\|_2$. The expectation of the sum of squared errors is:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{d} Z_i^2\right] &= \sum_{i=1}^{d} \mathbb{E}\left[Z_i^2\right], \\
&= \sum_{i=1}^{d} \mathbb{E}\left[\sum_{j=1}^{d-1} \Delta_{ij}^2\right], \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d-1} \mathbb{E}\left[\Delta_{ij}^2\right], \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d-1} \frac{1}{m_i} \Sigma_{jj}^{(i)}, \text{ since } \Delta_{ij} \approx \mathcal{Z}(0, \frac{1}{m_i}\Sigma_{jj}^{(i)}), \\
&= \sum_{i=1}^{d} \frac{1}{m_i} \operatorname{tr} \boldsymbol{\Sigma}^{(i)}.
\end{aligned}
\tag{17}
$$

When a sample is drawn, the feature for which the sample will be used for improving the Shapley composition estimation is chosen to maximize $\frac{\operatorname{tr}\boldsymbol{\Sigma}^{(i)}}{m_i} - \frac{\operatorname{tr}\boldsymbol{\Sigma}^{(i)}}{m_i+1}$. Like in (Štrumbelj & Kononenko, 2014), this is summarized in Algorithm 2.

Algorithm 1 Adaptation of the Algorithm 1 from (Štrumbelj & Kononenko, 2014) for approximating the Shapley composition of the ith feature, with model $\boldsymbol{f}$, instance $\boldsymbol{x} \in \mathcal{X}$ and m drawn samples.

```
Initialize $\boldsymbol{\phi_i} \leftarrow \mathrm{ilr}^{-1}(\mathbf{0})$
for 1 to $m$ do
  Randomly select a permutation $\pi$ of the set of indexes $\mathcal{I}$,
  Randomly select a sample $\boldsymbol{w} \in \mathcal{X}$,
  Construct two instances:
    • $\boldsymbol{b}_1$: which takes the values from $\boldsymbol{x}$ for the $i$th feature and the features indexed before $i$ in the order given by $\pi$, and takes the values from $\boldsymbol{w}$ otherwise,
    • $\boldsymbol{b}_2$: which takes the values from $\boldsymbol{x}$ the features indexed before $i$ in the order given by $\pi$, and takes the values from $\boldsymbol{w}$ otherwise.
  $\boldsymbol{\phi}_i \leftarrow \boldsymbol{\phi}_i \oplus \boldsymbol{f}(\boldsymbol{b}_1) \ominus \boldsymbol{f}(\boldsymbol{b}_2)$
end for
$\boldsymbol{\phi}_i \leftarrow \frac{\boldsymbol{\phi}_i}{m}$
```

Algorithm 2 Adaptation of the Algorithm 2 from (Štrumbelj & Kononenko, 2014) for approximating all the Shapley compositions by optimally distributing a maximum number of samples $m_{\max}$ over the d features, with model $\boldsymbol{f}$, instance $\boldsymbol{x} \in \mathcal{X}$ and $m_{\min}$ the minimum number of samples each feature estimation.

```
Initialization: $m_i \leftarrow 0$, $\boldsymbol{\phi}_i \leftarrow \mathbf{0}$, $\forall i \in \{1, \dots d\}$,
while $\sum_{i=1}^{d} m_i < m_{\max}$ do
  if $\forall i, m_i \leq m_{\min}$ then
    $j = \operatorname{argmax}_i \left( \frac{\operatorname{tr} \boldsymbol{\Sigma}^{(i)}}{m_i} - \frac{\operatorname{tr} \boldsymbol{\Sigma}^{(i)}}{m_i + 1} \right)$,
  else
    pick a $j$ such that $m_j < m_{\min}$,
  end if
  $\boldsymbol{\phi}_j \leftarrow \boldsymbol{\phi}_j$ + result of Algorithm 1 for the $j$th feature and $m = 1$,
  update $\boldsymbol{\Sigma}^{(j)}$ using an incremental algorithm,
  $m_j \leftarrow m_j + 1$
end while
$\boldsymbol{\phi}_i \leftarrow \frac{\boldsymbol{\phi}_i}{m_i}$, $\forall i \in \{1, \dots d\}$.
```

E. Adjustement of the estimated Shapley compositions for efficiency

In practice, the computation of the Shapley values has an exponential time complexity and we do not have necessarily access to the true distribution of the data. The Shapley values are therefore approximated using estimation algorithms like for instance the one presented in the previous appendix. However, since the obtained values are approximations, they do not necessarily respect the desired efficiency property. This point is often overlooked in the literature. In this section we write down an adjustment strategy of the estimated Shapley compositions for them to respect the efficiency property. This is a similar strategy as in the sampling approximation of the Shapley values in the SHAP toolkit (Lundberg & Lee, 2017)[10].

Let $\{\hat{\boldsymbol{\phi}}_i\}_{1 \leq i \leq d}$ be the estimated Shapley compositions (given by the Algorithm 2 in our experiments). Let $\boldsymbol{s}_{err} = \boldsymbol{f}(\boldsymbol{x}) \ominus \boldsymbol{f}_0 \ominus \bigoplus_{i=1}^{d} \hat{\boldsymbol{\phi}}_i$, where $\boldsymbol{f}_0$ is the base distribution, be the error composition on the pertubation of all Shapley compositions, i.e. the error making the efficiency property unfulfilled. In order to respect the efficiency property, we want this error to be the neutral element of the perturbation, i.e. the “zero” in the sense of the Aitchison geometry: the uniform distribution. We could simply perturb each estimated Shapley compositions by $\frac{1}{d} \odot \boldsymbol{s}_{err}$ however this would move each Shapley composition by the same amount while we want to allow the compositions with a higher estimation variance (i.e. with a precision likely to be lower) to move more than the ones with a smaller variance (i.e. with a precision likely to be higher).

We therefore weight the ith adjustment by a scalar $w_i = w\left(\operatorname{tr}\left(\boldsymbol{\Sigma}^{(i)}\right)\right)$, where w is an increasing function, and where $\sum_{i=1}^{d} w_i = 1$. Note that the vector of weight is actually a composition too. Similarly to the SHAP toolkit implementation, we choose w as:

$$w_i = w\left(\operatorname{tr}\left(\boldsymbol{\Sigma}^{(i)}\right)\right) = \frac{v_i}{1 + \sum_{j=1}^{d} v_j}, \text{ where } v_i = \frac{\operatorname{tr}\left(\boldsymbol{\Sigma}^{(i)}\right)}{\epsilon \max_j \operatorname{tr}\left(\boldsymbol{\Sigma}^{(j)}\right)}. \tag{18}$$

The ith estimated Shapley composition is then asjusted as follow:

$$\hat{\boldsymbol{\phi}}_i \leftarrow \hat{\boldsymbol{\phi}}_i \oplus \left(w_i \odot \boldsymbol{s}_{err}\right). \tag{19}$$

[10] https://github.com/shap/shap/blob/master/shap/explainers/_sampling.py

In this way, when ϵ goes to zero[11], the efficiency property is respected for the adjusted Shapley compositions and more weight is given to the adjusments of the Shapley compositions with a higher estimation variance.

[11] In our experiments, $\epsilon = 10^{-6}$.