

# Compositional analysis of bivariate discrete probabilities

J. J. Egozcue<sup>1</sup>, J. L. Díaz-Barrero<sup>1</sup>, V. Pawlowsky-Glahn<sup>2</sup>

<sup>1</sup>Dep. Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Barcelona, Spain;

*juan.jose.egozcue@upc.edu*

<sup>2</sup>Dep. Informàtica i Matemàtica Aplicada, Universitat de Girona, Girona, Spain;

*vera.pawlowsky@udg.edu*

## Abstract

A joint distribution of two discrete random variables with finite support can be displayed as a two way table of probabilities adding to one. Assume that this table has  $n$  rows and  $m$  columns and all probabilities are non-null. This kind of table can be seen as an element in the simplex of  $n \cdot m$  parts. In this context, the marginals are identified as compositional amalgams, conditionals (rows or columns) as subcompositions. Also, simplicial perturbation appears as Bayes theorem. However, the Euclidean elements of the Aitchison geometry of the simplex can also be translated into the table of probabilities: subspaces, orthogonal projections, distances.

Two important questions are addressed: a) given a table of probabilities, which is the nearest independent table to the initial one? b) which is the largest orthogonal projection of a row onto a column? or, equivalently, which is the information in a row explained by a column, thus explaining the interaction? To answer these questions three orthogonal decompositions are presented: (1) by columns and a row-wise geometric marginal, (2) by rows and a column-wise geometric marginal, (3) by independent two-way tables and fully dependent tables representing row-column interaction. An important result is that the nearest independent table is the product of the two (row and column)-wise geometric marginal tables. A corollary is that, in an independent table, the geometric marginals conform with the traditional (arithmetic) marginals. These decompositions can be compared with standard log-linear models.

**Key words:** balance, compositional data, simplex, Aitchison geometry, composition, orthonormal basis, arithmetic and geometric marginals, amalgam, dependence measure, contingency table.

# 1 Introduction

From the very beginning of probability theory, joint probabilities of two discrete and finite support random variables are presented as two-way tables. Each cell in the table contains a probability (frequency) and, when normalised to probabilities, all cells add to one. The traditional elements of these tables are the marginals, obtained summing columns or rows, also adding to one; and the conditionals, which are easily obtained normalising to one each column or row. More advanced statistical tools of analysis of two-way contingency tables are the so-called log-linear models (Bishop et al., 1975; Haberman, 1978; Everitt, 1992) in which the table is decomposed into a product of tables: a constant table; tables associated with a column or with a row; tables related to interactions between rows and columns. This analysis is essentially statistical because the estimation of the different tables is normally carried out using maximum likelihood estimation. Also in the statistical framework, correspondence analysis (Benzecri, 1969; Greenacre, 1984) try to represent the interaction or linkage between rows and columns using principal components analysis and, importantly, giving a weight to each column and row depending on the data used to estimate the frequencies. Therefore, both log-linear and correspondence analysis are interpretable statistical models, but they do not rely upon structural characteristics of the joint distribution as the marginals or conditionals.

The present aim is to define structural elements of Discrete Bivariate Probability functions, DBP or DBP table for short, based on the Aitchison geometry of the simplex. In fact, a DBP has a number of positive components adding to one; the probabilities are assumed to be measured in a relative and symmetrical scale, and the natural operation between them can be identified with the perturbation in the simplex, i.e. the Bayes updating.

Section 2 briefly describes the Aitchison geometry of the simplex applied to the case of DBP and some subspaces associated with independent DBP, rows, and columns. Section 3 presents decompositions of DBP's based in the previous geometrical concepts and discusses some issues of interpretation. An example of orthogonal decomposition of a DBP is presented in Section 4.

## 2 Subspaces of discrete bivariate probability tables

### Basic operations

Consider  $(n, m)$ -arrays with positive entries. Generically, these kind of arrays are denoted as  $\mathbf{x}$ ,  $\mathbf{y}$ ,  $\mathbf{z}$ . For instance, the entries of the first one are denoted with a double index denoting row and column:  $x_{ij}$ . Proportional  $(n, m)$ -positive-arrays are considered equivalent. Each equivalence class is then represented by a DBP, closed to one. The extraction of the representative is called closure, and is denoted by  $\mathcal{C}\mathbf{x}$ . The set of DBP's can be identified with an  $(n \cdot m)$ -part simplex,  $\mathcal{S}^{nm}$ . Traditional elements in probability theory are easily identified with the corresponding concepts in the simplex. A conditional probability given the  $i$ -th row, (resp.  $j$ -th column) is simply the subcomposition in  $\mathcal{S}^m$  (resp.  $\mathcal{S}^n$ )  $\text{row}_i[\mathbf{x}] = \mathcal{C}[x_{i1}, x_{i2}, \dots, x_{im}]$  (resp.  $\text{col}_j[\mathbf{x}] = \mathcal{C}[x_{1j}, x_{2j}, \dots, x_{nj}]$ ). The brackets denote row vector despite of the previous character of row or column within the array  $\mathbf{x}$ . Marginals correspond to the concept of amalgamation (Aitchison, 1986). The marginal row of  $\mathbf{x}$  is defined as  $\text{mrgr}[\mathbf{x}] = \mathcal{C}[\sum_i x_{i1}, \sum_i x_{i2}, \dots, \sum_i x_{im}]$ , a composition in  $\mathcal{S}^m$ . Similarly, the marginal column is  $\text{mrgc}[\mathbf{x}] = \mathcal{C}[\sum_j x_{1j}, \sum_j x_{2j}, \dots, \sum_j x_{nj}]$  is a composition in  $\mathcal{S}^n$ . The operators that extract a row-vector from a DBP, i.e.  $\text{mrgc}$ ,  $\text{mrgr}$ ,  $\text{row}_i$ ,  $\text{col}_j$  etc., are followed by the argument in brackets. When the argument is in parenthesis it means that the result is a DBP in  $\mathcal{S}^{nm}$  (see Eq. 2).

The standard operations in the simplex, perturbation and powering, apply to DBP's. Components of perturbation,  $\mathbf{z} = \mathbf{x} \oplus \mathbf{y}$ , are  $z_{ij} = (x_{ij}y_{ij}) / \sum_{rs} x_{rs}y_{rs}$ , i.e. closed direct product of entries. Powering by a real constant  $\alpha$ ,  $\mathbf{z} = \alpha \odot \mathbf{x}$ , has entries  $z_{ij} = x_{ij}^\alpha / \sum_{rs} x_{rs}^\alpha$ . As is well-known (Aitchison et al., 2000; Pawlowsky-Glahn and Egozcue, 2001; Aitchison et al., 2002), the simplex is a vector space with these operations.

In this framework, perturbation has a direct interpretation: it represents the Bayes updating of probabilities. Consider the expression of a perturbation,  $\mathbf{z} = \mathbf{x} \oplus \mathbf{y}$ . Assume that  $\mathbf{x}$  is the prior joint probability of a bivariate family of events; and that  $\mathbf{y}$  is the probability of an experimental observation given the joint probability of the events, i.e. the likelihood of the observation. Then  $\mathbf{z}$  is equal to the posterior, i.e. conditional to the observation, joint probability.

Besides the vector space structure of the simplex, the Aitchison inner product provides a metric compatible with perturbation, thus structuring the simplex as an Euclidean space (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001). The Aitchison inner product for DBP's is

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i,j} \ln x_{ij} \cdot \ln y_{ij} - \frac{1}{nm} \left( \sum_{i,j} \ln x_{ij} \right) \cdot \left( \sum_{i,j} \ln y_{ij} \right). \quad (1)$$

The norm and distance are defined accordingly,

$$\|\mathbf{x}\|_a = (\langle \mathbf{x}, \mathbf{x} \rangle_a)^{1/2}, \quad d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a,$$

where  $\ominus \mathbf{y} = \oplus((-1) \odot \mathbf{y})$  is the inverse operation of perturbation or perturbation-subtraction, consisting in dividing element-wise and then closing the result.

## Row and column subspaces

The Euclidean structure of  $\mathcal{S}^{nm}$  permits to define subspaces and orthogonal projections of DBP's on them. The  $i$ -th row of a DBP,  $\mathbf{x} \in \mathcal{S}^{nm}$ , is the subcomposition  $\text{row}_i[\mathbf{x}] \in \mathcal{S}^m$ . However, this subcomposition can be identified with an orthogonal projection of  $\mathbf{x}$  onto a subspace of  $\mathcal{S}^{nm}$  of dimension  $m - 1$  (Egozcue and Pawlowsky-Glahn, 2005), denoted  $\mathcal{S}^{nm}(\text{row}_i)$ . In order to construct both the subspace and the projection, consider first an orthonormal basis in  $\mathcal{S}^m$  whose vectors are  $\mathbf{e}_k = \mathcal{C}(\exp[\xi_{k1}, \xi_{k2}, \dots, \xi_{km}])$ ,  $k = 1, 2, \dots, m - 1$ , where  $\text{clr}(\mathbf{e}_k) = [\xi_{k1}, \xi_{k2}, \dots, \xi_{km}]$  with  $\xi_{kj} = \ln(x_{kj}/g(\mathbf{e}_k))$ ;  $g(\cdot)$  denotes the geometric mean of the components of the argument; and the function  $\exp(\cdot)$  and  $\ln(\cdot)$  operate component-wise on the composition. Reference to the  $i$ -th row in  $\mathbf{e}_k$  has been removed because this basis is assumed to be the same for all the rows. An orthonormal basis in  $\mathcal{S}^{nm}(\text{row}_i)$  is

$$\mathbf{E}_{ik} = \mathcal{C} \exp \begin{pmatrix} 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \xi_{k1} & \xi_{k2} & \dots & \xi_{km} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad k = 1, 2, \dots, m - 1,$$

where the only non-null row of the matrix is the  $i$ -th row, and this row is now referenced in the subscript of  $\mathbf{E}_{ik}$ .

The orthogonal projection of  $\mathbf{x}$  onto  $\mathcal{S}^{nm}(\text{row}_i)$  is denoted by  $\text{row}_i(\mathbf{x})$ ; note that  $\text{row}_i[\mathbf{x}] \in \mathcal{S}^m$  denotes a row vector, whereas  $\text{row}_i(\mathbf{x}) \in \mathcal{S}^{nm}$  is a DBP table. This projection is

$$\text{row}_i(\mathbf{x}) = \bigoplus_{k=1}^{m-1} \langle \mathbf{x}, \mathbf{E}_{ik} \rangle_a \odot \mathbf{E}_{ik}. \quad (2)$$

A tedious computation (Egozcue and Pawlowsky-Glahn, 2005) shows that

$$\text{row}_i(\mathbf{x}) = \mathcal{C} \begin{pmatrix} g(\text{row}_i[\mathbf{x}]) & g(\text{row}_i[\mathbf{x}]) & \dots & g(\text{row}_i[\mathbf{x}]) \\ \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{im} \\ \dots & \dots & \dots & \dots \\ g(\text{row}_i[\mathbf{x}]) & g(\text{row}_i[\mathbf{x}]) & \dots & g(\text{row}_i[\mathbf{x}]) \end{pmatrix}, \quad (3)$$

where  $g(\text{row}_i[\mathbf{x}])$  denotes the geometric mean of the elements in the  $i$ -th row of  $\mathbf{x}$ . The DBP in Eq. (3) is also characterised by the following property: among the DBP's whose entries are equal except in the  $i$ -th row,  $\text{row}_i(\mathbf{x})$  is the nearest one to  $\mathbf{x}$  with respect to the Aitchison distance in  $\mathcal{S}^{nm}$ . Furthermore, using the isometric properties of the clr transformation, the orthogonality of  $\text{row}_i(\mathbf{x})$  and  $\text{row}_{i'}(\mathbf{x})$ ,  $i \neq i'$  is easily proven.

A similar development can be reproduced for columns leading to  $m$  mutually orthogonal subspaces of dimension  $n - 1$  associated with each column; accordingly, they are denoted  $\mathcal{S}^{nm}(\text{col}_j)$ . The subspaces associated with rows and columns are not orthogonal, as proven by the following counterexample.

Consider the DBP (2, 3)-array

$$\mathbf{x} = \begin{pmatrix} 0.05 & 0.30 & 0.15 \\ 0.10 & 0.20 & 0.20 \end{pmatrix}, \quad \text{mrg}_r[\mathbf{x}] = [0.15, 0.50, 0.35], \quad \text{mrg}_c[\mathbf{x}] = [0.50, 0.50],$$

with  $\|\mathbf{x}\|_a^2 = 2.008$ . The projection of  $\mathbf{x}$  onto the subspaces associated with the first row and first column are

$$\begin{aligned} \text{row}_1(\mathbf{x}) &= \mathcal{C} \begin{pmatrix} 0.05 & 0.30 & 0.15 \\ 2250^{1/3} \cdot 10^{-2} & 2250^{1/3} \cdot 10^{-2} & 2250^{1/3} \cdot 10^{-2} \end{pmatrix}, \quad \|\text{row}_1(\mathbf{x})\|_a^2 = 1.633, \\ \text{col}_1(\mathbf{x}) &= \mathcal{C} \begin{pmatrix} 0.05 & \sqrt{0.005} & \sqrt{0.005} \\ 0.10 & \sqrt{0.005} & \sqrt{0.005} \end{pmatrix}, \quad \|\text{col}_1(\mathbf{x})\|_a^2 = 1.764. \end{aligned}$$

The inner product of both projections is  $\langle \text{row}_1(\mathbf{x}), \text{col}_1(\mathbf{x}) \rangle_a = 1.364$ , which is non-null. Thus, they are not orthogonal. It corresponds to an angle of about 36 degrees.

## Geometric marginal row and column subspaces

A natural question is what is the orthogonal complement of the row (column) subspaces. The dimension is easily computed; as the total dimension of the simplex is  $nm - 1$  and the subspace of each row is  $m - 1$ , the remaining dimension is  $(nm - 1) - n(m - 1) = n - 1$  (resp.  $m - 1$  for the complement of the columns). These subspaces are denoted  $\mathcal{S}^{nm}(\text{row}^\perp)$  and  $\mathcal{S}^{nm}(\text{col}^\perp)$ . A basis of  $\mathcal{S}^{nm}(\text{row}^\perp)$  has the form

$$\mathbf{F}_k = \mathcal{C} \exp \begin{pmatrix} \eta_{1k} & \eta_{1k} & \dots & \eta_{1k} \\ \eta_{2k} & \eta_{2k} & \dots & \eta_{2k} \\ \dots & \dots & \dots & \dots \\ \eta_{nk} & \eta_{nk} & \dots & \eta_{nk} \end{pmatrix}, \quad k = 1, 2, \dots, n - 1,$$

where the vectors  $[\eta_{1k}, \eta_{2k}, \dots, \eta_{nk}]$ , for  $k = 1, 2, \dots, n - 1$ , constitute an orthonormal basis in  $\mathbb{R}^{n-1}$ . Orthogonality of these elements to the row projections can be checked computing the corresponding inner products, for which it holds  $\langle \mathbf{F}_k, \mathbf{E}_{ir} \rangle_a = 0$  for  $k = 1, 2, \dots, n - 1$ , and for  $i = 1, 2, \dots, n$ ,  $r = 1, 2, \dots, m - 1$ . The projection of  $\mathbf{x}$  onto  $\mathcal{S}^{nm}(\text{row}^\perp)$  is  $\text{row}^\perp(\mathbf{x}) = \bigoplus_{k=1}^{n-1} \langle \mathbf{x}, \mathbf{F}_k \rangle_a \odot \mathbf{F}_k$ , i.e.

$$\text{row}^\perp(\mathbf{x}) = \mathcal{C} \begin{pmatrix} g(\text{row}_1[\mathbf{x}]) & g(\text{row}_1[\mathbf{x}]) & \dots & g(\text{row}_1[\mathbf{x}]) \\ g(\text{row}_2[\mathbf{x}]) & g(\text{row}_2[\mathbf{x}]) & \dots & g(\text{row}_2[\mathbf{x}]) \\ \dots & \dots & \dots & \dots \\ g(\text{row}_n[\mathbf{x}]) & g(\text{row}_n[\mathbf{x}]) & \dots & g(\text{row}_n[\mathbf{x}]) \end{pmatrix}, \quad (4)$$

whose identical columns are, up to closure, the row geometric mean of the original DBP. The form of  $\text{row}^\perp(\mathbf{x})$  suggest the name *geometric marginal column* of  $\mathbf{x}$  for the vector

$$\text{gmrg}_c[\mathbf{x}] = \mathcal{C}[g(\text{row}_1[\mathbf{x}]), g(\text{row}_2[\mathbf{x}]), \dots, g(\text{row}_n[\mathbf{x}])].$$

The name of marginal is due to the fact that, when the geometric means in the DBP (4) are substituted by arithmetic means, the columns of (4) are equal to the traditional marginal column  $\text{mrg}_c[\mathbf{x}]$ . Similarly,

$$\text{gmrg}_r[\mathbf{x}] = \mathcal{C}[g(\text{col}_1[\mathbf{x}]), g(\text{col}_2[\mathbf{x}]), \dots, g(\text{col}_m[\mathbf{x}])],$$

denotes de *geometrical marginal row*. The projection  $\text{row}^\perp(\mathbf{x})$  ( $\text{col}^\perp(\mathbf{x})$ ) are called geometrical marginal column (row) DBP of  $\mathbf{x}$ .

The orthogonality of the row projections and the geometric marginal column DBP permits an orthogonal decomposition of  $\mathbf{x}$ ,

$$\begin{aligned}\mathbf{x} &= \text{row}^\perp(\mathbf{x}) \oplus \left( \bigoplus_{i=1}^n \text{row}_i(\mathbf{x}) \right) \\ &= \left( \bigoplus_{k=1}^{n-1} \langle \mathbf{x}, \mathbf{F}_k \rangle_a \odot \mathbf{F}_k \right) \oplus \left( \bigoplus_{i=1}^n \bigoplus_{k=1}^{m-1} \langle \mathbf{x}, \mathbf{E}_{ik} \rangle_a \odot \mathbf{E}_{ik} \right).\end{aligned}\tag{5}$$

A similar decomposition can be obtained for projections onto columns and the geometric marginal row. It can be expressed as

$$\mathbf{x} = \text{col}^\perp(\mathbf{x}) \oplus \left( \bigoplus_{j=1}^m \text{col}_j(\mathbf{x}) \right).\tag{6}$$

The subspaces of the geometric marginal row and column are orthogonal, i.e.  $\mathcal{S}^{nm}(\text{row}^\perp) \perp \mathcal{S}^{nm}(\text{col}^\perp)$ . To prove this fact, consider the clr representation of two DBP's:  $\mathbf{x}$  with equal rows, and  $\mathbf{y}$  with equal columns. Their clr representations have rows and columns adding to zero respectively, i.e.  $\sum \xi_i = 0$ ,  $\sum \eta_j = 0$ . The inner product has the form

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \text{clr}(\mathbf{x}), \text{clr}(\mathbf{y}) \rangle = \sum \begin{pmatrix} \xi_1 & \xi_2 & \dots & \xi_m \\ \xi_1 & \xi_2 & \dots & \xi_m \\ \dots & \dots & \dots & \dots \\ \xi_1 & \xi_2 & \dots & \xi_m \end{pmatrix} \times \begin{pmatrix} \eta_1 & \eta_1 & \dots & \eta_1 \\ \eta_2 & \eta_2 & \dots & \eta_2 \\ \dots & \dots & \dots & \dots \\ \eta_n & \eta_n & \dots & \eta_n \end{pmatrix} = 0,$$

where  $\sum(\cdot) \times (\cdot)$  denotes sum of the element-wise product of the matrices.

An important property of geometric marginals, not fulfilled by the arithmetic marginals, is the linearity under perturbation. Let  $\mathbf{x}$  and  $\mathbf{y}$  be DBP's in  $\mathcal{S}^{mn}$ . Then,

$$\text{gmrgr}[\mathbf{x} \oplus \mathbf{y}] = \text{gmrgr}[\mathbf{x}] \oplus \text{gmrgr}[\mathbf{y}] \quad , \quad \text{gmrgc}[\mathbf{x} \oplus \mathbf{y}] = \text{gmrgc}[\mathbf{x}] \oplus \text{gmrgc}[\mathbf{y}] ,$$

where the perturbations between marginals operate in  $\mathcal{S}^m$  and  $\mathcal{S}^n$  respectively.

## Independent subspace

An important goal when analysing a DBP is to identify relationships between rows and columns. Moreover, the part of the DBP that does not generate any relationship between rows and columns, called independent part, should be also identified. An independent DBP is generated by the diadic product of a row and a column. In the simplex context, this kind of DBP's is easily represented as a perturbation of two DBP whose rows, respectively columns, are equal. This set is called the independent subspace and it is denoted as  $\mathcal{S}_{\text{ind}}^{nm} = \mathcal{S}^{nm}(\text{row}^\perp) \oplus \mathcal{S}^{nm}(\text{col}^\perp)$ . Since both  $\mathcal{S}^{nm}(\text{row}^\perp)$  and  $\mathcal{S}^{nm}(\text{col}^\perp)$  are orthogonal subspaces of the simplex  $\mathcal{S}^{nm}$  with respective dimensions  $n-1$  and  $m-1$ , their perturbation has dimension  $(n-1) + (m-1)$ .

When analysing a DBP,  $\mathbf{x}$ , the orthogonal projection onto  $\mathcal{S}_{\text{ind}}^{nm}$  represents the part of  $\mathbf{x}$  in which rows and columns are unrelated. The remaining part is made of inter-relations between rows and columns. The independent part is readily computed as perturbation of geometric marginals

$$\mathbf{x}_{\text{ind}} = \text{row}^\perp(\mathbf{x}) \oplus \text{col}^\perp(\mathbf{x}) ,$$

which are easily computed from  $\mathbf{x}$ . This is not the independent DBP obtained as a perturbation of arithmetic marginals. In fact,  $\mathbf{x}_{\text{ind}} \neq \text{mrgc}(\mathbf{x}) \oplus \text{mrgr}(\mathbf{x})$  because

$$\text{row}^\perp(\mathbf{x}) \neq \text{mrgc}(\mathbf{x}) = (\text{mrgc}[\mathbf{x}]', \dots, \text{mrgc}[\mathbf{x}]') ,$$

$$\text{col}^\perp(\mathbf{x}) \neq \text{mrgr}(\mathbf{x}) = (\text{mrgr}[\mathbf{x}]', \dots, \text{mrgr}[\mathbf{x}]')'$$

The equality  $\mathbf{x}_{\text{ind}} = \text{mrgc}(\mathbf{x}) \oplus \text{mrgr}(\mathbf{x})$  only occurs whenever  $\mathbf{x}$  is an independent DBP itself, then,

$$\text{row}^\perp(\mathbf{x}) = \text{mrgc}(\mathbf{x}) , \quad \text{col}^\perp(\mathbf{x}) = \text{mrgr}(\mathbf{x}) .$$

In fact, in this case, the projection onto the row and column subspaces is null and, then, the projection in the independent subspace, being unique, corresponds to the product of marginals.

### 3 DBP analysis

Following the ideas of log-linear models, the goal is to decompose a DBP into perturbed parts in a meaningful way. A typical part should be an independent DBP table. Therefore, a first and important orthogonal decomposition of a DBP in  $\mathcal{S}^{nm}$  is

$$\mathbf{x} = \mathbf{x}_{\text{ind}} \oplus \mathbf{x}_{\text{int}} , \tag{7}$$

where

$$\mathbf{x}_{\text{ind}} = \text{gmrgc}(\mathbf{x}) \oplus \text{gmrgc}(\mathbf{x}) = \text{col}^\perp(\mathbf{x}) \oplus \text{row}^\perp(\mathbf{x}) .$$

An important issue of the analysis is to measure the distance of  $\mathbf{x}$  to  $\mathbf{x}_{\text{ind}}$ , its associate independent DBP. An appropriate measure of dependence is the squared-distance

$$\Delta^2(\mathbf{x}) = \|\mathbf{x}_{\text{int}}\|_a^2 = \|\mathbf{x}\|_a^2 - \|\mathbf{x}_{\text{ind}}\|_a^2 .$$

The dependence measure  $\Delta$  depends on the dimensions of the DBP and, therefore, the relative squared-norm may be more interpretable:

$$R_\Delta^2(\mathbf{x}) = \frac{\Delta^2(\mathbf{x})}{\|\mathbf{x}\|_a^2} , \quad 0 \leq R_\Delta^2 \leq 1 . \tag{8}$$

When  $R_\Delta^2(\mathbf{x}) = 1$ ,  $\mathbf{x}$  is a pure interaction DBP, whereas  $R_\Delta^2(\mathbf{x}) = 0$  means that  $\mathbf{x}$  is an independent DBP. Clearly,  $R_\Delta^2(\mathbf{x}_{\text{int}}) = 1$  and  $R_\Delta^2(\mathbf{x}_{\text{ind}}) = 0$ . It may be noted the differences and similarities of  $R_\Delta^2$  with the so-called *deviance* (Nedler and Wedderburn, 1972). Deviance can be defined as

$$D = 2 \sum_{i=1}^n \sum_{j=1}^m x_{ij} \ln \left( \frac{x_{ij}}{z_{ij}} \right) ,$$

where  $z_{ij}$  are the entries of the (arithmetic) independent DBP. The ratio  $x_{ij}/z_{ij}$  is essentially a (perturbation)-subtraction of the independent DBP. The weighted sum is a Kullback-Leibler divergence. From the present point of view, the independent DBP should be replaced by the geometric approach to the independent DBP, and the Kullback-Leibler divergence by the Aitchison-square-norm to get  $\|\mathbf{x}_{\text{int}}\|_a^2$  which is the proposed measure of dependence before normalisation.

Contributions of individual rows and columns may be also of interest. However, in the general case, the mentioned contributions are not orthogonal. In Section 2, two orthogonal decompositions were presented: row contributions and geometric marginal column or, alternatively, column contributions and geometric marginal row. The independent subspace is not orthogonal to any part in these decompositions. Being the projection onto the independent subspace of primary interest, a first step is the separation the independent contribution in (7). The remainder is made of interactions of rows and columns which cannot be included in an independent DBP. This remainder DBP is called interaction DBP. The DBP of pure interaction of rows and columns is then  $\mathbf{x}_{\text{int}} = \mathbf{x} \ominus \mathbf{x}_{\text{ind}}$ , whose geometric marginals are both neutral (all entries are equal). The arithmetic marginals are not neutral. Using decompositions (5) and (6) applied to  $\mathbf{x}_{\text{int}}$ , the following

orthogonal decompositions hold

$$\begin{aligned}\mathbf{x} &= \mathbf{x}_{\text{ind}} \oplus \left( \bigoplus_{i=1}^n \text{row}_i(\mathbf{x}_{\text{int}}) \right) \\ &= \mathbf{x}_{\text{ind}} \oplus \left( \bigoplus_{j=1}^m \text{col}_j(\mathbf{x}_{\text{int}}) \right).\end{aligned}$$

This means that the contribution of the  $i$ -th row to the square-norm of the interaction is  $\|\text{row}_i(\mathbf{x}_{\text{int}})\|_a^2$ ; and similarly for column contributions.

Additionally, special attention should be paid to interaction of rows and columns. But in this case there is not an orthogonal decomposition accounting for all row-column interactions. Row-column interactions in  $\mathbf{y} = \mathbf{x}_{\text{int}}$  may be defined in different but equivalent ways. A convenient definition is the *cross-contrast*: it is the balance of the  $(i, j)$ -cell against the other cells in the cross formed by the  $i$ -th row and the  $j$ -th column,

$$I_{\text{cross}}(i, j) = \sqrt{\frac{n+m-2}{n+m-1}} \ln \frac{y_{ij}}{\left( \prod_{i \neq r=1}^n y_{rj} \prod_{j \neq s=1}^m y_{is} \right)^{1/(n+m-2)}}.$$

These balances are not orthogonal, but their sum is zero, and the sum of all squares of them is proportional to the square-norm of  $\mathbf{x}_{\text{int}}$ ,

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^m I_{\text{cross}}(i, j) &= 0, \\ \sum_{i=1}^n \sum_{j=1}^m (I_{\text{cross}}(i, j))^2 &= \frac{(n+m)^2}{(n+m-1)(m+m-2)} \cdot \|\mathbf{x}_{\text{int}}\|_a^2.\end{aligned}\tag{9}$$

This is due to the very special properties of  $\mathbf{x}_{\text{int}}$  whose geometric marginals are neutral, i.e. the geometric mean of each row and each column is equal to the geometric mean of all entries of the interaction DBP. There is also a proportionality between the square cross-contrasts and the square *cell-interaction* of the  $i$ -th row and the  $j$ -th column. Cell-interaction is defined as the balance between the  $(i, j)$ -cell compared with all the remaining cells of DBP:

$$I_{\text{cell}}(i, j) = \sqrt{\frac{nm-1}{nm}} \ln \frac{y_{ij}}{\left( \prod_{(r,s) \neq (i,j)} y_{rs} \right)^{1/(nm-1)}}.$$

The sign of  $I_{\text{cell}}(i, j)$  indicates whether the interaction between row and column is constructive or destructive. The relationship of these balances to the clr coefficients was studied in Egozcue and Pawlowsky-Glahn (2006). From clr properties,  $\sum_{ij} I_{\text{cell}}(i, j) = 0$  and

$$\sum_{i=1}^n \sum_{j=1}^m (I_{\text{cell}}(i, j))^2 = \frac{nm}{nm-1} \cdot \|\mathbf{x}_{\text{int}}\|_a^2.\tag{10}$$

A way to figure out this (non-orthogonal) decomposition of the square-norm of  $\mathbf{y} = \mathbf{x}_{\text{int}}$  is to list (table-wise) the interaction in per units of square-norm, using either the cell-interactions or the cross-contrasts. Cell-interaction or cross-contrast simply logarithmically scale interaction of a row and a column as it may be visualised directly from the interaction DBP  $\mathbf{y} = \mathbf{x}_{\text{int}}$ . The fact that cross-contrasts and cell-interactions give equivalent information is an additional argument to consider them as a canonical way of describing interactions between rows and columns.

The decompositions of the square-norm  $\|\mathbf{x}_{\text{int}}\|_a^2$  given in (9) and in (10) correspond to non-orthogonal perturbation-decompositions of the interaction DBP. The original DBP can then be decomposed

$$\mathbf{x} = \mathbf{x}_{\text{ind}} \oplus \left( \bigoplus_{i=1}^n \bigoplus_{j=1}^m (a \cdot I_{\text{cross}}(i, j)) \odot \mathbf{c}_{ij} \right), \quad a = \frac{(n+m-2)(n+m-1)}{(n+m)^2}, \quad (11)$$

where the cross-interaction DBP, is

$$\mathbf{c}_{ij} = \mathcal{C} \exp \begin{pmatrix} 0 & \dots & -B & \dots & 0 \\ 0 & \dots & -B & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -B & \dots & A & \dots & -B \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -B & \dots & 0 \end{pmatrix}$$

where  $A = (n+m-2)^{1/2}(n+m-1)^{-1/2}$ ,  $B = ((n+m-2)(n+m-1))^{-1/2}$ , and the non-null row is the  $i$ -th one and the non-null column is the  $j$ -th one. Note that  $\|\mathbf{c}_{ij}\|_a = 1$  and they are a system of generators of the interaction matrices. The perturbation-decomposition (11) constitute a log-linear model for  $\mathbf{x}$  although, in general, the standard decompositions in such models do not coincide with the present one.

## 4 Example of orthogonal decomposition

The marks obtained by  $C = 227$  students in a medium level probability subject are considered. Most students study the subject for the first time (143). However, there is a number of them that try to pass the exam for the second (46) or third time (38). The results of the exam may be classified into five groups: those not contesting the exam (No Cont.); and the marks D(0-3.9), C(4-4.9), B(5-6.9), A(7-10), where the letters correspond approximately to anglo-saxon marks and the numeric intervals to Spanish marks over 10 points. The results can be organised in a two way table. Table 1 shows the original contingency table.

**Table 1:** Contingency table with the data from an examination of 227 students. By rows, number of exam trials; by columns, obtained mark.

counts	ORIGINAL			mrgc
	Num Contest			
Mark	1	2	3	
No Cont	68	8	17	93
D(0-3.9)	19	17	11	47
C(4-4.9)	29	7	5	41
B(5-6.9)	22	13	4	39
A(7-10)	5	1	1	7
mrgr	143	46	38	227

Cell probabilities can be estimated to obtain the corresponding DBP (Table 2). Each cell probability has been estimated as  $x_{ij} = (c_{ij} + 1/2)/(C + nm/2)$ , where  $c_{ij}$  are the counts in each cell,  $n = 5$  is the number of rows and  $m = 3$  is the number of columns. The estimation of such probabilities is not addressed here and the estimated ones are assumed to be known. The original DBP has Aitchison square norm 15.370 and the independent table obtained by multiplication of the arithmetic marginals 14.221. After dividing the DBP by this independent table and taking closure to one, an

**Table 2:** Exam results: original DBP table (left) with arithmetic and geometric marginals normalised to 1. Independent table obtained as the product of arithmetic marginals.

probability	ORIGINAL DBP					INDEPENDENT (ARITHMETIC)				
	Num Contest					Num Contest				
Mark	1	2	3	mrgrc	gmrgc	Mark	1	2	3	mrgrc
No Cont	0.292	0.036	0.075	0.403	0.352	No Cont	0.250	0.083	0.070	0.403
D(0-3.9)	0.083	0.075	0.049	0.207	0.256	D(0-3.9)	0.128	0.043	0.036	0.207
C(4-4.9)	0.126	0.032	0.023	0.181	0.173	C(4-4.9)	0.112	0.037	0.031	0.181
B(5-6.9)	0.096	0.058	0.019	0.173	0.180	B(5-6.9)	0.107	0.036	0.030	0.173
A(7-10)	0.023	0.006	0.006	0.036	0.038	A(7-10)	0.022	0.007	0.006	0.036
mrgr	0.620	0.207	0.173	1	0	mrgr	0.620	0.207	0.173	1
gmrgc	0.619	0.211	0.169	0	0.042					

interaction DBP, with square-norm 1.874, is obtained. The fact that  $15.370 < 14.221 + 1.874$  shows that this arithmetic decomposition is not orthogonal. Table 3 shows the column (left) and row

**Table 3:** Exam results: Geometric marginal DBP's.

probability	GMRGC					probability	GMRGR				
	Num Contest						Num Contest				
Mark	1	2	3	mrgrc	gmrgc	Mark	1	2	3	mrgrc	gmrgc
No Cont	0.117	0.117	0.117	0.352	0.352	No Cont	0.124	0.042	0.034	0.200	0.200
D(0-3.9)	0.085	0.085	0.085	0.256	0.256	D(0-3.9)	0.124	0.042	0.034	0.200	0.200
C(4-4.9)	0.058	0.058	0.058	0.173	0.173	C(4-4.9)	0.124	0.042	0.034	0.200	0.200
B(5-6.9)	0.060	0.060	0.060	0.180	0.180	B(5-6.9)	0.124	0.042	0.034	0.200	0.200
A(7-10)	0.013	0.013	0.013	0.038	0.038	A(7-10)	0.124	0.042	0.034	0.200	0.200
mrgr	0.333	0.333	0.333	1.000		mrgr	0.619	0.211	0.169	1.000	
gmrgc	0.333	0.333	0.333		0.053	gmrgc	0.619	0.211	0.169		0.056

(right) geometric marginal DBP. Marginal DBP's have all columns, resp. rows, equal. Therefore, one of the marginals of these marginal DBP's is constant and geometric and arithmetic marginals are equal. This property of equal arithmetic and geometric marginals propagates to their perturbation. The projection of the original DBP onto the independent subspace is the perturbation of the two marginal DBP's; it is shown in Table 4 (left) with its equal geometric and arithmetic marginals. This property does not hold for the independent DBP obtained from the arithmetic marginals (Fig. 2). Table 4 shows the independent and interaction (geometric) DBP's. Despite of the intuition that there is a large interaction between rows and columns in the original DBP, the orthogonal decomposition of the square-norm assigns 13.700 to the independent DBP and only 1.670 to the interaction, i.e. there is much more information in the independent DBP than in the interaction. Accordingly, the dependence measure (8) is  $R_{\Delta}^2 = 1.670/15.370 = 0.109$  in this case.

The interaction square norm can be decomposed into row-column interactions. As mentioned, this is not an orthogonal decomposition. Table 5 shows these interactions in per unit square-norm of the interaction DBP. Additionally, per units of square-norm are shown with the sign of the balance  $I_{\text{cross}}(i, j)$ . When the sign is positive the probability in the original DBP is greater than that one in the independent DBP, i.e. the interaction is constructive or positive. A negative sign indicates the a destructive interaction, i.e. the probability in the original DBP is less than the one in the independent DBP. The left table corresponds to the interaction extracted using the geometric criteria. The right table refers to the residual DBP when the arithmetic independent DBP is perturbation-subtracted from the original one. Note that this table is a little bit more informative than the interaction DBP (square-norm is  $1.874 > 1.670$ ). Comparison between the two interaction decompositions shows that there are cells where the percentages are multiplied or

**Table 4:** Exam results: Independent projection and interaction DBP's. The squared norms are 13.700, 1.670 respectively, adding up to the square norm of the original DBP.

probability		INDEPENDENT (GEOMETRIC)				probability		INTERACTION (GEOMETRIC)				
		Num Contest					Num Contest					
Mark		1	2	3	mrgc	gmrgr	Mark	1	2	3	mrgc	gmrgr
No Cont		0.218	0.074	0.060	0.352	0.352	No Cont	0.091	0.033	0.085	0.208	0.200
D(0-3.9)		0.159	0.054	0.043	0.256	0.256	D(0-3.9)	0.035	0.093	0.077	0.205	0.200
C(4-4.9)		0.107	0.037	0.029	0.173	0.173	C(4-4.9)	0.079	0.059	0.054	0.192	0.200
B(5-6.9)		0.112	0.038	0.031	0.180	0.180	B(5-6.9)	0.058	0.102	0.043	0.203	0.200
A(7-10)		0.023	0.008	0.006	0.038	0.038	A(7-10)	0.068	0.055	0.068	0.191	0.200
mrgr		0.619	0.211	0.169	1.000		mrgr	0.332	0.342	0.326	1.000	
gmrgr		0.619	0.211	0.169		0.045	gmrgr	0.333	0.333	0.333		0.063

**Table 5:** Exam results: Decomposition of the square-norm of interaction. Geometric, left; arithmetic, right. The square norm of interaction is shown with the sign of the cross-contrast. A negative (positive) sign means interaction reduces (increases) the probability with respect to the independent DBP;

Interaction (pu)		geometric			Interaction (pu)		arithmetic		
		Num Contest					Num Contest		
Mark		1	2	3	Mark	1	2	3	
No Cont		0.078	-0.254	0.051	No Cont	0.021	-0.331	0.007	
D(0-3.9)		-0.200	0.090	0.022	D(0-3.9)	-0.081	0.193	0.070	
C(4-4.9)		0.031	-0.003	-0.015	C(4-4.9)	0.013	-0.007	-0.032	
B(5-6.9)		-0.004	0.138	-0.094	B(5-6.9)	-0.002	0.145	-0.084	
A(7-10)		0.003	-0.013	0.003	A(7-10)	0.004	-0.007	0.002	

**Int norm2      1.670**                      **Int norm2      1.874**

divided by factors of 2, 3 or even 5. An interesting case is that of row D (bad result in the exam). The independent DBP (Tab. 4, left) explains a good deal of D-results, but interaction between the D-row and the 1-column (Tab. 5, left) points out an important destructive interaction ((-)0.200). This stands in contrast with the interaction between the D-row and the 2-column, which is relatively small ((+)0.090, a factor of 2) (Tab. 5, left). However, when examining these interactions in the arithmetic interaction DBP (Tab. 5, right), the result is reversed: interaction of D-row and 2-column is approximately double ((+)0.193) the interaction of D-row and 1-column ((-)0.081). Expressing these results in a colloquial way, the geometric approach states that *students contesting for the first time the examination get less bad results than others contesting for the second time*. However, the arithmetic analysis suggest the statement: *students contesting for the second time the examination increase their probability of a bad result*. Although both statements describe to a certain extent the same fact, there is a different stress in which is the column involved (first or second contest).

## 5 Conclusion

A discrete bivariate probability distribution (DBP) organised as a two way array is interpreted as a composition and thus represented in the simplex. The Aitchison geometry of the simplex is then applied to get orthogonal decompositions. The original DBP is expressed as the perturbation of a geometric marginal (column/row) and row/column-wise associated DBP's. An important result is that the perturbation of the geometric marginal row and column is the orthogonal projection of

the DBP onto the subspace of independent DBP's. This projection is not equal to the perturbation of the standard-arithmetic marginals. A proper measure of global dependence between rows and columns has been defined as the the square-norm not explained by the independent projection DBP over the square norm of the original DBP. Finally, the square-norm of the interaction DBP, i.e. the original DBP (perturbation)-minus the independent projection, is decomposed into row-column interactions represented by balances, thus allowing a complete analysis of dependence.

## Acknowledgements

This research has been supported by the Spanish Ministry of Education and Science under projects Ref.: 'Ingenio Mathematica (i-MATH)' No. CSD2006-00032 (Consolider – Ingenio 2010) and Ref.: MTM2006-03040.

## REFERENCES

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London UK. (Reprinted in 2003 with additional material by The Blackburn Press).
- Aitchison, J., C. Barceló-Vidal, J. J. Egozcue, and V. Pawlowsky-Glahn (2002). A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In U. Bayer, H. Burger, and W. Skala (Eds.), *Proceedings of IAMG'02 — The eighth annual conference of the International Association for Mathematical Geology*, Volume I and II, pp. 387–392. Selbstverlag der Alfred-Wegener-Stiftung, Berlin Germany.
- Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2000). Logratio analysis and compositional distance. *Mathematical Geology* 32(3), 271–275.
- Benzecri, J. P. (1969). *Statistical analysis as a tool to make patterns emerge from data*. In *Methodologies of pattern recognition* (S. Watanabe ed), Academic Press, New York USA.
- Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association* 96(456), 1205–1214.
- Bishop, Y., S. Fienberg, and P. Holland (1975). *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge, Mass. USA.
- Egozcue, J. J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology* 37(7), 795–828.
- Egozcue, J. J. and V. Pawlowsky-Glahn (2006). *Simplicial geometry for compositional data*. In: Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V., (eds) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London UK.
- Everitt, B. S. (1992). *The analysis of contingency tables*. 2nd ed. Chapman and Hall, London UK.
- Greenacre, M. J. (1984). *Correspondence analysis*. Academic Press, New York USA.
- Haberman, S. (1978). *Analysis of quantitative data*. Academic Press, New York USA.
- Nedler, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *J. Royal Statistical Society, A* 135, 370–384.
- Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)* 15(5), 384–398.