Modelling structural zeros in compositional data

Michail Tsagris

Department of Economics, University of Crete, Rethymnon, Greece,

mtsagris@uoc.gr

Abstract

We present new model for analyzing compositional data with structural zeros. Inspired by Butler and Glasbey (2008) who suggested a model in the presence of zero values in the data we propose a model that treats the zero values in a different manner.

Keywords: compositional data, α-transformation, structural zeros

1 Introduction

Structural (and rounded) zeros are sometimes met in compositional data. The term structural refers to values which are truly zeros, for instance the percentage of money a family spends on smoking or alcohol. Rounded zeros on the other hand are very small values in some components which were rounded to zero. In geology for example the instrument which measures the composition of the elements has a detection limit. Values below that limit are not detected. This has two possible explanations; either the element is completely absent or had a value smaller than the detection limit of the instrument.

Ever since 1982 (Aitchison, 1982), the most widely used approach for compositional data analysis is the log-ratio approach. The nature of the logs though gives rise to a mathematical problem, the log of zero is undefined. This problem was dealt with simple imputation techniques such as imputation by a small value (Aitchison, 2003), or with substitution of the zero by a fraction of the detection limit (Palarea Albaladejo et al., 2005), or via the EM algorithm (Palarea-Albaladejo et al., 2007). If the zeros present are indeed rounded down only because the detection limit of the instrument was not that low, then these approaches can be used. However, even in this case, the true value could be lower than estimated. (Scealy and Welsh, 2011a) showed an example of the problem when these approaches are adopted. The smaller the imputed value is, the higher the

magnitude of the log-ratio transformed values are. If on the other hand the value is a true zero (not rounded), then any imputation technique is clearly not correct.

(Butler and Glasbey, 2008) proposed a latent Gaussian model for modelling zero values. They used a multivariate normal distribution in $\mathbb{R}^d$ to model the data. When a point was outside the simplex they projected it orthogonally onto the faces and vertices of the simplex. However this approach has the problem of sometimes assigning too much probability on the vertices and sometimes more than is necessary. Furthermore, the higher the dimensionality of the simplex, finding the correct regions to project the points lying outside the simplex becomes more difficult. Maximum likelihood estimation becomes more difficult also, but with the use of MCMC methods they managed to tackle the estimation problems. We propose a different model for handling zero values, which is inspired though from that model (Butler and Glasbey, 2008). Instead of using an orthogonal projection for the points lying outside the simplex we move them along the line connecting the points with the center of the simplex.

In this section we will discuss the issue of structural zeros, that is when the value observed is actually zero and is not due to a rounding error. We will suggest a new method for modelling structural zeros based on the multivariate normal distribution. It is a different projection than the one suggested by Butler and Glasbey (2008). Since the simplex has the form of a triangle (when $D = 3$), it seems that the projection of the points lying outside the simplex should be projected onto the boundaries of the simplex following a similar idea to the folded model (Tsagris and Stewart, 2018).

2 The α-transformation

2.1 The stay-in-the-simplex version of the α-transformation

The power transformation defined by Aitchison (2003) we saw earlier in Section **??** is

$$\mathbf{u} = \{u_i\}_{i=1,\ldots,D} = \left\{ \frac{x_i^{\alpha}}{\sum_{j=1}^{D} x_j^{\alpha}} \right\}_{i=1,\ldots,D}. \tag{1}$$

We shall call (1) the stay-in-the-simplex version of the α-transformation. Note that the map (1) is degenerate due to the constraints $\sum_{i=1}^{D} x_i = 1$ and $\sum_{i=1}^{D} u_i = 1$. In order to make (1) non-

degenerate we consider the version of (1) as follows

$$u_i\left\{(x_j)_{j=1}^d\right\} = \frac{x_i^\alpha}{\sum_{j=1}^d x_j^\alpha + \left(1 - \sum_{j=1}^d x_j\right)^\alpha} \quad i = 1, \ldots, d. \tag{2}$$

The (2) is presented to highlight that in fact we have $d = D - 1$ and not D variables. Thus, the Jacobian of (1) or (2) is not singular. The Jacobian of the stay-in-the-simplex version of the α-transformation (1) is equal to (Tsagris and Stewart, 2018)

$$|\mathbf{J}| = \alpha^d \prod_{i=1}^D \frac{x_i^{\alpha-1}}{\sum_{j=1}^D x_j^\alpha}. \tag{3}$$

2.2 The α-transformation

A centred and scaled version of (1) is defined as

$$B_\alpha(\mathbf{x}) = \mathbf{H}\frac{1}{\alpha}\left(D\mathbf{u} - \mathbf{1}_D\right), \tag{4}$$

where $\mathbf{u}$ is defined in (1), $\mathbf{1}_D$ is the D-dimensional vector of 1s and $\mathbf{H}$ is the Helmert sub-matrix (**??**). Note that (4) is simply a linear transformation of (1) and so any inference made on either of them should be the same. The Jacobian of the α-transformation (4) is equal to (Tsagris and Stewart, 2018)

$$|\mathbf{J}_\alpha| = D^{d+\frac{1}{2}} \prod_{i=1}^D \frac{x_i^{\alpha-1}}{\sum_{j=1}^D x_j^\alpha}.$$

Thus, we can say that the space of the α-transformation depends upon the value α. With (4) it is easy to see how the space increases to meet the whole of $\mathbb{R}^d$. As $\alpha \to 0$, $\mathbb{A}^d$ tends to the whole of $\mathbb{R}^d$, the space of the isometric log-ratio transformation. The new space defined in (**??**) is a function of α and is simply a linear expansion of the simplex with the redundant dimension having been discarded.

3 Zero-censored model

We will try to fit a multivariate normal distribution on the simplex which comprises of two components, one component for the data which lie inside the simplex and a second component for the

points lying on the faces. A key (possibly restrictive for some datasets) feature of the model is that we assign zero probability on the vertices. At first, we will use the α-transformation, with $\alpha = 1$, in order to escape the unit sum constraint. So in effect we center the simplex and multiply it by D, the number of components and then multiply from the left with the Helmert sub-matrix to remove the unit sum constraint. Then similarly to Butler and Glasbey (2008) we can write the log-likelihood as

$$\ell = \sum_{i=1}^{n_1} \log g\left(\mathbf{y}_i\right) + \sum_{i=1}^{n} \log |J| + \sum_{i=1}^{n_2} \log \int_{k_j}^{\infty} f_i\left(\mathbf{y}_i\right) dy_j, \tag{5}$$

where $\mathbf{y} = \mathbf{B}_1\left(\mathbf{x}\right)$ the α-transformation with $\alpha = 1$ (**??**), with $\mathbf{x} \in \mathbb{S}^d$, $g\left(.\right)$ is the density of the multivariate normal for the data lying inside the simplex, $f_i\left(.\right)$ is the density of the i-th point lying outside the simplex given that it is in a line going through the origin. The n_1 is the number of points lying inside the interior of the simplex and n_2 denotes the number of points on the faces of the simplex. The line integral refers to the i-th observation lying on the face the simplex for which the integral is calculated along the j-th component, with $j \in [1, ..., D]$, where D is the number of components. Finally, $|J|$ is the Jacobian determinant of the α-transformation with $\alpha = 1$.

The rationale is similar to the Butler and Glasbey (2008) model. We assume there is a latent multivariate normal distribution but we have observed the compositional data only. Zero values of compositional data imply that the values of the latent distribution were outside the simplex. An advantage of this model over the one suggested by Butler and Glasbey (2008) is that the likelihood is tractable for any number of dimensions.

The limitation of our suggested model is that can handle compositional vectors with zero values in only one of their components. We will need to calculate the line integral of this component in the multivariate density from that point to infinity. Therefore, the log-likelihood consists of the density inside the interior of the simplex and the density on the faces, thus assigning zero probability on the vertices (and to the edges when $D > 3$). We will express (5) in a more convenient way as

$$\begin{aligned} \ell &= -\frac{n_1}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n_1} \left(\mathbf{y}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_i - \boldsymbol{\mu}\right) \\ &\quad + \sum_{i=1}^{n_2} \log \int_{c_1}^{\infty} f_i\left(\mathbf{z}\right) dz_1 + \left(nd + \frac{n}{2}\right) \log D, \end{aligned} \tag{6}$$

where $n = n_1 + n_2$ is the full sample size. The vector inside the integral has changed from $\mathbf{y}_i$ to

$\mathbf{z}_i$ with $\mathbf{z} = \mathbf{B}\mathbf{y}$, where c_1 and $\mathbf{B}$ will be explained below in the Gram-Schmidt process. In order to calculate the line integral we will first perform a rotation towards one arbitrary direction. For convenience reasons we chose the first direction, the X-axis for instance in the two dimensions as seen in Figure 1. The rotation takes place by multiplying the vector with the zero component (after the α-transformation) by an orthonormal matrix. The matrix is calculated via the Gram-Schmidt orthonormalization process (Strang, 1988).

3.1 Gram-Schmidt orthonormalization process

The process in mathematical terms is described as follows. Suppose we have a vector $\mathbf{v}$ in R^d and we want to rotate it to the line defined by the unit vector $\mathbf{w} = (1, 0, \ldots, 0)^T$, with $\mathbf{w}$ in R^d. We have to find an orthonormal basis first using the Gram-Schmidt orthonormalization process. Let us denote the projection operation of a vector $\mathbf{v}$ onto $\mathbf{u}$ by

$$proj_{\mathbf{u}}(\mathbf{v}) = \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}.$$

Then the following operations will take place

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1 && && \text{and } \mathbf{e}_1 = \tfrac{\mathbf{u}_1}{\|\mathbf{u}_1\|}, \\
\mathbf{u}_2 &= \mathbf{v}_2 && -proj_{\mathbf{u}_1}(\mathbf{v}_2) && \text{and } \mathbf{e}_2 = \tfrac{\mathbf{u}_2}{\|\mathbf{u}_2\|}, \\
\mathbf{u}_3 &= \mathbf{v}_3 && -proj_{\mathbf{u}_1}(\mathbf{v}_3) && -proj_{\mathbf{u}_2}(\mathbf{v}_3) \quad \text{and } \mathbf{e}_3 = \tfrac{\mathbf{u}_2}{\|\mathbf{u}_2\|}, \\
&\vdots && && \vdots \\
\mathbf{u}_d &= \mathbf{v}_d && -\textstyle\sum_{i=1}^{d-1} proj_{\mathbf{u}_i}(\mathbf{v}_d) && \text{and } \mathbf{e}_d = \tfrac{\mathbf{u}_2}{\|\mathbf{u}_2\|},
\end{aligned}$$

where $\langle \cdot \rangle$ denotes the inner product of two vectors. Denote the matrix of the orthonormalized vectors $\mathbf{e}$ by

$$\mathbf{B} = [\mathbf{e}_1, \ldots, \mathbf{e}_d].$$

Then all we have to do to get $\mathbf{c}$ is $\mathbf{c} = \mathbf{B}\mathbf{w}$ and the first element of the vector $\mathbf{c}$ is the term c_1 we saw in (6).

Since the integral of (6) is with respect to the first variable of the multivariate normal we can

use the conditional normal to write (6) in a more attractive form as

$$\begin{aligned}\ell &= -\frac{n_1}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n_1}(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}) + \left(nd+\frac{n}{2}\right)\log D \\ &\quad + \sum_{i=1}^{n_2}\log\left[f_i\left(\mathbf{z}_{-1};\boldsymbol{\mu}_{iz^*},\boldsymbol{\Sigma}_{iz^*}\right)\int_{c_{1i}}^{\infty} f_i\left(z_{1i}|\mathbf{z}_{-1};\mu_{i,con},\sigma^2_{i,con}\right)dz_{1i}\right], \qquad (7)\end{aligned}$$

where $\mathbf{z}_{-1}$ means all elements except from the first one. $f_i\left(\mathbf{z}_{-1};\boldsymbol{\mu}_z^*,\boldsymbol{\Sigma}_z^*\right)$ is the density of the multivariate normal with parameters $(\boldsymbol{\mu}_z^*,\boldsymbol{\Sigma}_z^*)$ calculated at $\mathbf{z}_{-1}$ and $f_i\left(z_{1i}|\mathbf{z}_{-1};\mu_{i,con},\sigma^2_{i,con}\right)$ is the density of the conditional univariate normal with parameters $\left(\mu_{i,con},\sigma^2_{i,con}\right)$ calculated at z_{1i}. The conditional distribution of a multivariate normal is still a normal (Mardia et al., 1979) and the following relationships hold true

$$(\mathbf{X}_1,\mathbf{X}_2)\sim N_d\left((\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)^T,\boldsymbol{\Sigma}\right),\quad \text{then}\quad \mathbf{X}_1|\mathbf{X}_2\sim N_d\left(E\left(\mathbf{X}_1|\mathbf{X}_2\right),V\left(\mathbf{X}_1|\mathbf{X}_2\right)\right)$$

where

$$E\left(\mathbf{X}_1|\mathbf{X}_2\right)=\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{X}_2-\boldsymbol{\mu}_2\right)\quad\text{and}\quad V\left(\mathbf{X}_1|\mathbf{X}_2\right)=\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Hence, using these relationships we can calculate the parameters of the normal density appearing inside the integral of (7). Thus, we have the following relationships

$$\boldsymbol{\mu}_{iz}=\mathbf{B}_i\boldsymbol{\mu}\quad\text{and}\quad\boldsymbol{\Sigma}_{iz}=\mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i^T,$$

where the index i is used to indicate the i-th observation and $\mathbf{B}_i$ means the rotation matrix (calculated from the Gram-Schmidt orthonormalization process) for the i-th observation. The rotation matrix rotates the vector $\mathbf{y}_i$ to the line defined by the unit vector $\mathbf{v}=(1,0,...,0)^T$. Now,

$$\boldsymbol{\mu}_{iz^*}=\boldsymbol{\mu}_{iz,-1}\quad\text{and}\quad\boldsymbol{\Sigma}_{iz^*}=\boldsymbol{\Sigma}_{iz}[-1,-1]$$

and

$$\begin{aligned}\mu_{i,con} &= \boldsymbol{\mu}_{iz,1}-\boldsymbol{\Sigma}_{iz}[1,]\boldsymbol{\Sigma}_{iz}[-1,-1]^{-1}\boldsymbol{\mu}_{iz,-1}\\ \text{and}\quad\sigma^2_{i,con} &= \boldsymbol{\Sigma}_{iz}[1,1]-\boldsymbol{\Sigma}_{iz}[1,]\boldsymbol{\Sigma}_{iz}[-1,-1]^{-1}\boldsymbol{\Sigma}_{iz}[1,]^T,\end{aligned}$$

where $\boldsymbol{\Sigma}[-1,-1]$ means the matrix $\boldsymbol{\Sigma}$ without the first element and $\boldsymbol{\Sigma}[1,]$ indicates the first row of the matrix $\boldsymbol{\Sigma}$.

So, the idea is to multiply each vector by the rotation matrix $\mathbf{B}_i$ and rotate the data onto the first axis, thus the new vector is denoted by $\mathbf{c}_i = (c_{1i}, 0 \ldots, 0)$. Thus (7) can be written as

$$\begin{aligned} \ell &= \frac{n_1}{2} \log |2\pi\boldsymbol{\Sigma}| - 0.5 \sum_{i=1}^{n_1} (\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma} (\mathbf{y}_i - \boldsymbol{\mu}) + \\ & \sum_{i=1}^{n_2} \log f_i (\mathbf{0}; \boldsymbol{\mu}_{iz^*} \boldsymbol{\Sigma}_{iz^*}) + \sum_{i=1}^{n_2} \log \left[1 - \Phi \left(\frac{c_{1i} - \mu_{i,con}}{\sigma_{i,con}} \right) \right], \end{aligned} \tag{8}$$

where $\Phi(.)$ is the cumulative distribution of a standard normal random variable. The final form of the log-likelihood (8) is the form maximized numerically. The index i in each of the parameters (for the compositions which contained one zero value) indicates that each composition with one zero value had to be projected onto the face and thus its contribution to the parameters is different.

Figure 1 shows a graphical example of the rotation in R^2. The red line integral is calculated through a normal distribution whose parameters are rotated via the Gram-Schmidt orthonormalization process in the same way the black line was rotated to the red line. This is one example of a composition with a zero value in one of its components. In the sample, we have to sum all of these cases.

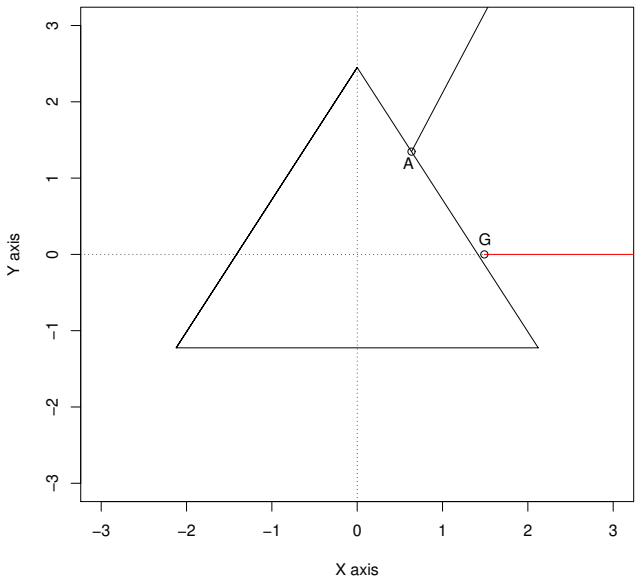

Figure 1: Ternary diagram showing the zero projection. We want to evaluate the line integral of the multivariate normal distribution from A to ∞ along the black line. For this reason we rotate the point onto the X-Axis and find the integral from G to ∞.

3.2 Example 1. Simulated data

Figure 2 shows a simulated example of the zero-censored model. Data of size 500 were generated from the following multivariate normal

$$N_2 \left(\begin{array}{cc} (0.625, & 0.821) \end{array} , \left(\begin{array}{cc} 0.149 & -0.200 \\ -0.200 & 1.523 \end{array} \right) \right).$$

When an observation fell outside the simplex it was "pulled" to the boundary, moving along the line connecting the point with the center of the simple, via the technique described in Tsagris and Stewart (2018). There were 197 such cases in the data. We applied the zero-censored model to the data by maximizing the log-likelihood (7). We estimated 194 zeros (194 compositional vectors having one element with a zero value). We generated 500 vectors from a multivariate normal and counted the number of vectors that fell outside the simplex. The estimated parameters of this normal distribution, used in this random vector generation, were

$$\hat{\boldsymbol{\mu}} = (0.656, 0.788) \quad \text{and} \quad \mathbf{S} = \left(\begin{array}{cc} 0.129 & -0.132 \\ -0.132 & 1.477 \end{array} \right).$$

Figure 2 shows the ternary plot of the data along with the contours of the zero-censored model calculated from the estimated parameters.

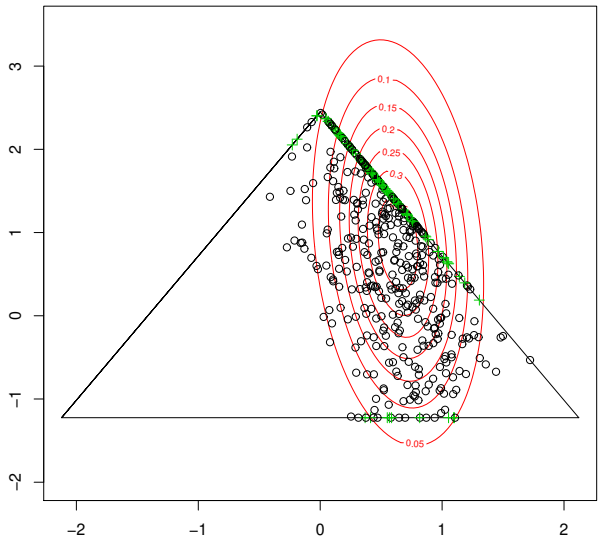

Figure 2: Ternary diagram where the green crosses indicate the points which lie on the boundaries of the simplex. The contour plots of the normal based on the estimated parameters are shown.

A key thing we have to mention about Figure 2 is that the contour lines look vertical but have in fact a negative slope. This is not seen because of the scaling of the ternary plot. The range of values of the x-axis is larger than the range of the simulated values in the first variable and thus the contour lines do not depict the negative slope they should.

3.3 Example 2. Time budget data

We will illustrate the performance of the zero-censored model using real data (Härdle and Hlávka, 2007). There are 28 individuals and for each person information about the time allocation in 10 activities is known. The individuals are identified according to gender, country where they live, professional activity, and matrimonial status. We are not interested in their categorization but in the amount of time each person spent on 10 categories of activities over 100 days (the total is $100 \times 24 = 2400$ hours fixed for every row) in 1976. The special feature of these data is that they contain some zero values. Some activities have zero allocation, for instance one woman did not spend even an hour on transportation linked to professional activity and four women did not spend

any hour on occupation linked to children. This means that we have five compositions which have one zero in one component only.

The estimated parameters are

$$\hat{\boldsymbol{\mu}} = (1.075, -0.030, 0.860, 0.3830.367, 0.222, -2.417, 0.568, -0.465) \quad \text{and}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.289 & 0.574 & 0.109 & 0.076 & 0.019 & 0.035 & 0.038 & 0.008 & 0.030 \\ 0.574 & 1.240 & 0.200 & 0.119 & 0.020 & 0.063 & 0.069 & -0.020 & 0.002 \\ 0.109 & 0.200 & 0.050 & 0.035 & 0.013 & 0.020 & 0.019 & 0.015 & 0.019 \\ 0.076 & 0.119 & 0.035 & 0.037 & 0.013 & 0.009 & 0.013 & 0.015 & 0.027 \\ 0.019 & 0.020 & 0.013 & 0.013 & 0.008 & 0.007 & 0.008 & 0.010 & 0.014 \\ 0.035 & 0.063 & 0.020 & 0.009 & 0.007 & 0.020 & 0.018 & 0.010 & 0.010 \\ 0.038 & 0.069 & 0.019 & 0.013 & 0.008 & 0.018 & 0.021 & 0.005 & 0.019 \\ 0.008 & -0.020 & 0.015 & 0.015 & 0.010 & 0.010 & 0.005 & 0.029 & 0.003 \\ 0.030 & 0.002 & 0.019 & 0.027 & 0.014 & 0.010 & 0.019 & 0.003 & 0.079 \end{bmatrix}.$$

3.4 Diagnostics for the zero-censored model

We have performed a similar goodness of fit diagnostic to the one Butler and Glasbey (2008) performed. We generated data from the fitted multivariate normal model and estimated the number of zeros in each component. For the first example with the simulated data we had 197 out of 500 vectors with one zero element, 171, 6 and 20 zeros in the first, second and third component respectively. The corresponding percentages are $(0.342, 0.012, 0.040)$. We estimated the percentages of the zero values in each component to be $(0.347, 0.008, 0.040)$ respectively based on $10,000,000$ simulated observations. We repeated the same procedure for the real data in the second example and the results are presented in Table 1.

Table 1: Observed and estimated number of zeros for every component

Components	prof	tran	hous	kids	shop	pers	eat	slee	tele	leis
Observed number of zeros	0	1	0	4	0	0	0	0	0	0
Estimated number of zeros	0.593	0.547	2.106	2.151	0.002	0.000	0.000	0.000	0.137	0.000

From Table 1 we see that there is evidence to support the hypothesis that the fit of the model is not to be rejected. We could also use the χ^2 test statistic as a discrepancy measure between the estimated and the observed frequencies and a p-value could be calculated via simulations or via the χ^2 distribution.

4 Conclusions

In this Chapter we developed two parametric models for compositional data modelling. Both of them are closely related to a multivariate normal model. The use of another multivariate model, such as the multivariate skew normal distribution (Azzalini and Valle, 1996) was also examined in the case of the inside-out model but was not exhibited here. The difficulty with this distribution is that more parameters need to be estimated, thus making the estimation procedure more difficult. This of course does not exclude the possibility of using this model.

The inside-out is a model more appropriate than the multivariate normal model for compositional data analysis because its support is the simplex. In Chapter 3 we saw the α-transformation and how it can be applied with the normal distribution. A drawback of that model is that it does not take account of the probability left outside the simplex. This is similar to the Box-Cox transformation, the support of the transformed data is not the whole of $\mathbb{R}^d$ as we have already mentioned. A possible solution would be to use truncation.

We took advantage of the α-transformation and proposed a more flexible model which is not based on truncation but on a folding technique. Thus it improves the α-normal distribution since it takes care of the probability left outside the simplex. As seen from the contour plots of the inside-out model and the application to real data it seems very good for modelling multimodal compositional data. Thus, we can say it is a multi-modal model, whose maximum number of modes depends on the number of the components and the estimated probability left outside the simplex. If on the other hand, there is little probability left outside the simplex, then the model becomes uni-modal.

A disadvantage is that location estimates for the compositional data do not exist in a closed form, thus empirical estimates based on simulation have to be calculated. Another disadvantage of this model is that it does not allow any zeros values to be present, since the Jacobian of the folding transformation is a function of the data. Similarly to Box-Cox transformation, the logarithm of the Jacobian of the α-transformation contains this term $(\alpha - 1)\sum_{j=1}^{n}\sum_{i=1}^{D}\log x_{ij}$. Thus, no zero

values are allowed.

The zero-censored model attacks the problem of zeros from a different perspective than the one Butler and Glasbey (2008) suggested. The data are projected on to the faces of the simplex using a non-orthogonal $1:1$ projection in contrast to the orthogonal Butler and Glasbey (2008) proposed. The advantage over Butler and Glasbey's approach is that it is not difficult to project the data onto the edges regardless of the dimension. Both models however share the same problem, that of estimating the parameters of the normal distribution which becomes harder as the dimension increases. Both the zero-censored model (8) and the Glasbey Butler and Glasbey (2008) model avoid the use of the log-ratio methodology or imputation of the zero values. A limitation of the zero-censored model is that it only allows for one zero per compositional vector. For instance if we have $D = 3$ or $D = 10$ components, only one zero should be present in each vector.

In any case, the question of modelling compositional data using any of these two examined models or the Butler and Glasbey (2008) model in the presence of covariates is still open. Scealy and Welsh (2011b) defined an alternative model based on the Kent distribution which offers the possibility for regression and handling zeros conveniently, at the cost of computational complexity.

Appendix

The Helmert sub-matrix

The Helmert matrix is a $D \times D$ orthogonal matrix. The Helmert sub-matrix has the first row omitted, hence is a $D-1 \times D$ matrix, the structure of which is presented below.

$$\mathbf{H}_{d,d+1} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & \cdots & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{\sqrt{i(i+1)}} & \cdots & \cdots & \frac{1}{\sqrt{i(i+1)}} & -\frac{i}{\sqrt{i(i+1)}} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 \\ \frac{1}{\sqrt{(D-1)D}} & \cdots & \cdots & \cdots & \frac{1}{\sqrt{(D-1)D}} & -\frac{D-1}{\sqrt{(D-1)D}} \end{pmatrix} \tag{A.1}$$

The space $\mathbb{A}^d$ of the α-transformation

The space $\mathbb{A}^d$ (**??**) of the α-transformation is an equilateral triangle in $\mathbb{R}^d$. The space of the α-transformation is defined by

$$\mathbb{A}^d = \left\{ \mathbf{H}\Gamma_\alpha(x_1), \ldots, \mathbf{H}\Gamma_\alpha(x_D)^T : -\frac{1}{\alpha} \leq \Gamma_\alpha(x_i) \leq \frac{D-1}{\alpha}, \sum_{i=1}^{D} \Gamma_\alpha(x_i) = 0 \right\},$$

where $(\Gamma_\alpha(x_1), \ldots \Gamma_\alpha(x_D)) = \Gamma_\alpha(\mathbf{x}) = \frac{1}{\alpha}(D\mathbf{u} - \mathbf{1}_D)$ and $\mathbf{H}$ is the Helmert sub-matrix (**??**).

The coordinates of the triangle in $\mathbb{R}^D$ are given by the following $D \times D$ matrix

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ldots & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

Therefore, the coordinates of the triangle after the α-transformation and before the Helmert sub-matrix ($\mathbf{H}$) multiplication are given by the $D \times D$ matrix

$$\mathbf{B} = \begin{pmatrix} \frac{D-1}{\alpha} & -\frac{1}{\alpha} & \ldots & \ldots & -\frac{1}{\alpha} \\ -\frac{1}{\alpha} & \frac{D-1}{\alpha} & -\frac{1}{\alpha} & \ldots & \vdots \\ \vdots & -\frac{1}{\alpha} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \frac{D-1}{\alpha} & -\frac{1}{\alpha} \\ -\frac{1}{\alpha} & \ldots & \ldots & -\frac{1}{\alpha} & \frac{D-1}{\alpha} \end{pmatrix}.$$

Finally, the coordinates of the new triangle after the Helmert sub-matrix (**??**) multiplication become

$$\mathbf{K} = \mathbf{HB} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \ldots & \ldots & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & \ldots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{\sqrt{i(i+1)}} & \ldots & \ldots & \frac{1}{\sqrt{i(i+1)}} & -\frac{i}{\sqrt{i(i+1)}} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 \\ \frac{1}{\sqrt{d(d+1)}} & \ldots & \ldots & \ldots & \frac{1}{\sqrt{d(d+1)}} & -\frac{d}{\sqrt{d(d+1)}} \end{pmatrix} \begin{pmatrix} \frac{D-1}{\alpha} & -\frac{1}{\alpha} & \ldots & \ldots & -\frac{1}{\alpha} \\ -\frac{1}{\alpha} & \frac{D-1}{\alpha} & -\frac{1}{\alpha} & \ldots & \vdots \\ \vdots & -\frac{1}{\alpha} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \frac{D-1}{\alpha} & -\frac{1}{\alpha} \\ -\frac{1}{\alpha} & \ldots & \ldots & -\frac{1}{\alpha} & \frac{D-1}{\alpha} \end{pmatrix}$$

$$\mathbf{K} = \begin{pmatrix} \frac{D}{\alpha\sqrt{2}} & -\frac{D}{\alpha\sqrt{2}} & 0 & 0 & \dots & \dots & 0 \\ \frac{D}{\alpha\sqrt{6}} & \frac{D}{\alpha\sqrt{6}} & -\frac{2D}{\alpha\sqrt{6}} & 0 & \dots & 0 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{D}{\alpha\sqrt{i(i+1)}} & \dots & \frac{D}{\alpha\sqrt{i(i+1)}} & -\frac{iD}{\alpha\sqrt{i(i+1)}} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{D}{\alpha\sqrt{dD}} & \dots & \dots & \dots & \dots & \frac{D}{\alpha\sqrt{dD}} & -\frac{dD}{\alpha\sqrt{dD}} \end{pmatrix} = \frac{D}{\alpha}\mathbf{H}.$$

We also have to note, that before the left multiplication by the Helmert sub-matrix, the space of the α-transformed data is singular and we shall denote it by $\mathbb{M}^d$, defined by

$$\mathbb{M}^d = \left\{ \Gamma_\alpha(\mathbf{x}) \,\middle|\, -\frac{1}{\alpha} \leq \Gamma_\alpha(x_i) \leq \frac{D-1}{\alpha}, \sum_{i=1}^{D} \Gamma_\alpha(x_i) = 0 \right\}. \tag{A.2}$$

References

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

Aitchison, J. (2003). *The statistical analysis of compositional data*. Reprinted by The Blackburn Press.

Azzalini, A. and Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715.

Butler, A. and Glasbey, C. (2008). A latent gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(5):505–520.

Härdle, W. and Hlávka, Z. (2007). *Multivariate statistics: exercises and solutions*. Springer Publishing Company, Incorporated.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. London: Academic Press.

Palarea Albaladejo, J., Fernández, M., Antoni, J., and Gómez García, J. (2005). alr approach for replacing values below the detection limit. In *Proceedings of the 2rd Compositional Data Analysis Workshop, Girona, Spain*.

Palarea-Albaladejo, J., Martín-Fernández, J., and Gómez-García, J. (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39(7):625–645.

Scealy, J. and Welsh, A. (2011a). Properties of a square root transformation regression model. In *Proceedings of the 4rth Compositional Data Analysis Workshop, Girona, Spain*.

Scealy, J. and Welsh, A. (2011b). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society. Series B*, 73(3):351–375.

Strang, G. (1988). *Linear algebra and its applications*. Thomson Learning Inc.

Tsagris, M. and Stewart, C. (2018). A folded model for compositional data analysis. *arXiv preprint arXiv:1802.07330*.