Modeling asymmetric compositional data

Ana Beatriz Tozzo Martins¹, Vanderly Janeiro¹, Terezinha Aparecida Guedes¹, Robson Marcelo Rossi¹ and Antônio Carlos Andrade Gonçalves²

¹Departamento de Estatística, Universidade Estadual de Maringá, Av. Colombo, 5790, 87020-900, Maringá, Paraná, Brazil. ²Departamento de Agronomia, Universidade Estadual de Maringá, Maringá, Paraná, Brazil. *Author for correspondence. E-mail: abtmartins@uem.br

ABSTRACT. Compositional data belong to the simplex sample space, but they are transformed to the sample space of the real numbers using the additive log-ratio transformation to allow the application of standard statistical techniques. This study aims to model compositional skewed data of three soil components after additive log-ratio transformation. The current modeling was done for compositional data of sand, silt and clay (simplex), and bivariate data (real) using the standard skew theory with and without the inclusion of the covariate soil porosity. The analyses were run using the R statistical software and the package sn, and the goodness-of-fit was found after applying the covariate.

Keywords: compositional data, skew-normal distribution, parameter estimation.

Modelagem de dados composicionais assimétricos

RESUMO. Dados composicionais pertencem ao espaço amostral simplex, mas que são transformados para o espaço amostral dos números reais, através da transformação log-razão aditiva, permitindo a aplicação de técnicas estatísticas padrão. Esse trabalho tem por objetivo modelar dados composicionais de três componentes, que apresentam assimetria, após transformação log-razão aditiva. A modelagem foi feita para dados composicionais de areia, silte e argila (simplex), e para os dados bivariados (reais), utilizando-se a teoria normal assimétrica sem e com a inclusão da covariável porosidade do solo. As análises foram feitas por meio do ambiente estatístico R utilizando o pacote sn e o melhor ajuste ocorreu ao considerar a covariável na modelagem.

Palavras-chave: dados composicionais, distribuição normal assimétrica, estimação de parâmetros.

Introduction

Compositional data are vectors whose elements are proportions with summation equal to 1. The natural sample space is the unitary simplex with dimension equal to the number of elements which means a restrictive part of the real space. The modeling of these types of data can be done through the simplex or by the set of real numbers. Aitchison (1986), and Pawlowsky-Glahn and Olea (2004) described the additive log-ratio transformation (ALR) and the centered log-ratio (CLR) for data analysis, for example, as an analytical method in the real sampling space, but requiring a new data transformation to the original sample space. The current experiment uses the real space as the sampling space given the possibility of applying the theory developed by Martins et al. (2009), who applied a multivariate normal distribution to compositional data. Sometimes, however, the normal distribution does not provide the best fit of the data. Thus, the skew-normal distribution developed by Azzalini (1985) has been profitable for these cases and applied to several areas of research and development.

Azzalini (1985) formally introduced the skew-normal distribution after previous studies. Gupta et al. (2004) developed a survey on the different articles/authors as Branco and Dey (2001), Loperfido (2001), Azzalini and Capitanio (1999), Gupta and Chen (2001), Gupta and Brown (2001) and reported two experiments to characterize this distribution considering quadratic statistics. Considering the compositional data, the literature contains articles as reported by Aitchison et al. (2003) who characterized the shape of the logistic normal additive and its tests. Mâteu-Figueras et al. (2005) proposed the additive logistic skew-normal distribution to analyze compositional data, and highlighted the analytical difficulties to obtain the estimates. Mâteu-Figueras and Pawlowsky-Glahn (2007) used a measurement like Lebesgue compatible with the algebraic and geometric structure of the simplex to define the skew-normal distribution in such sample space.
The objective of this experiment was to fit a bivariate skew-normal model to compositional data after the ALR transformation. The application had a data set with 82 compositions of sand, silt and clay.

Material and methods

In the current analysis, we used a data set collected from an irrigated area under the central-pivot of the Areão Farm, campus of the Escola Superior de Agricultura – “Luiz de Queiroz” (ESALQ-USP). A quadrant was designed in the highest part of the hill (at the top of the slope) from which 82 soil samples were collected from 0 to 0.20 m using a regular squared net for sampling the soil at every 20 m. Every soil sample had the contents of sand, silt and clay evaluated.

Compositional data

The composition is a vector \( \mathbf{X} = [X_1, X_2, ..., X_B] \) with B components representing the proportions. The sample space is the simplex given by \( S = \{ \mathbf{X} \in \mathbb{R}^B / X_i > 0, i = 1, 2, ..., B; \sum X_i = 1 \} \).

However, it is possible for vectors with positive components measured in the same scale to become a composition by dividing every component by the summation of all of them.

Visual graphic of a sample with three components, for example, is done by a ternary diagram designed on equilateral triangle where each vertex represents an individual component. Lemos and Santos (1996 apud REICHARDT; TIMM, 2004), describes the soil classification using a diagram that divides the soil into classes so that a location in the diagram allows the classification, ranging from Sandy to Very Clay.

Avoiding the application of standard statistical techniques that produce inconsistent results because of the intrinsic correlation among the components, Aitchison (1986) proposed, amidst others, the log additive ratio transformation as the

\[
\text{alr}: \mathbb{R}^B \to \mathbb{R}^{B-1}
\]

\[
\mathbf{X} \to \text{alr}(\mathbf{X}) = \left( \ln \left( \frac{X_1}{X_B} \right), ..., \ln \left( \frac{X_{B-1}}{X_B} \right) \right)
\]

enabling the analysis in the real space and the inverse transformation that is the generalized additive logistic,

\[
\text{agl}(\text{alr}(\mathbf{X})) = \mathbf{X} = (\exp(\ln(X_1 / X_B)), ..., \exp(0))'
\]

and the results are transformed to the original scale. Similar to Martins et al. (2009) who applied the ALR transformation to obtain \( Y_1 = \ln(\text{Sand} / \text{Silt}) \) and \( Y_2 = \ln(\text{Clay} / \text{Silt}) \).

Skew-normal distribution

Azzalini (1985) defined a random variable Z as a standard skew-normal distribution if the density function was given by:

\[
f(z | \alpha) = 2\phi(z)\Phi(\alpha z), \quad \alpha \in \mathbb{R} \quad z \in \mathbb{R}
\]

where \( \phi(\cdot) \) and \( \Phi(\cdot) \) are the density function and the standard normal distribution with zero mean and variance 1, respectively, and \( \alpha \) is a parameter controlling the asymmetry of the distribution ranging from \( (-\infty, \infty) \). The expected value and the variance of \( Z \) are given by

\[
E(Z) = \frac{\sqrt{2}}{\sqrt{\pi}} \delta \quad \text{and} \quad \text{Var}(Z) = 1 - \frac{2}{\sqrt{\pi}} \delta^2
\]

with \( \delta \) related to \( \alpha \) by \( \delta(\alpha) = \frac{\alpha}{\sqrt{1 + \alpha^2}} \).

Based on Azzalini (2005), if \( Z \) has skew-normal distribution denoted by \( Z \sim \text{NA}(\alpha) \), and \( Y = \mu_{\nu} + \omega Z \), where \( \mu_{\nu} \in \mathbb{R} \) and \( \omega \in \mathbb{R}^+ \), thus \( Y \sim \text{NA}(\mu_{\nu}, \omega \alpha^2, \alpha) \).

The asymmetric coefficient of the skew-normal distribution based on Bayes and Branco (2007), is

\[
\lambda = \frac{\sqrt{\pi}}{4} \left( \frac{1}{\sqrt{1 + \alpha^2}} \right)^{3/2} \left( \frac{\alpha}{\sqrt{1 + \alpha^2}} \right)^{-3/2} \ldots
\]

\[
\lambda \in [-0.99527; 0.99527]
\]

which characterizes how and how much the distribution deviates from the symmetry.

After Azzalini and Dalla Valle (1996), Azzalini and Capitanio (1999) described the extension of the Equation (2) for the multivariate case, the skew-normal multivariate density function for a standard normal random vector \( Z \) with k dimension as the function

\[
f(z | \Omega) = 2\phi_k(z, \Omega)\Phi(\alpha z), \quad z \in \mathbb{R}^k
\]

where \( \phi_k(z, \Omega) \) is the k-dimensional normal density with vector of means equal to zero and correlation
matrix $\Omega_z$, $\Phi(\cdot)$ is the distribution function $N(0,1)$, univariate, and $\alpha$ a vector of $k$ dimension.

Specifically for two dimensions, Azzalini and Dalla Valle (1996) described the bivariate skew-normal density function of $Z' = (Z_1,Z_2)$ as

$$f(z_1,z_2) = 2f(z_1,z_2;\Omega_z)\Phi(\alpha z_1 + \alpha z_2),$$

where $\Omega_z = \begin{bmatrix} 1 & \omega^* \\ \omega & 1 \end{bmatrix}$

is the correlation matrix, and

$$\alpha_1 = \frac{\delta_1 - \delta_2 \omega^*}{[(1-(\omega^*)^2)(1-(\alpha^*)^2) - \delta_1^2 + 2\delta_1\delta_2 \omega^*]^{1/2}};$$

$$\alpha_2 = \frac{\delta_2 - \delta_1 \omega^*}{[(1-(\omega^*)^2)(1-(\alpha^*)^2) - \delta_2^2 + 2\delta_1\delta_2 \omega^*]^{1/2}}.$$ 

After introducing the location parameter $\mu \in \mathbb{R}$ and the scale $\sigma > 0$ in the Equation (2), we have the density function for a random variable $Y$ as

$$f(y;\mu,\sigma^2;\alpha) = \frac{1}{\sigma} \phi\left(\frac{y - \mu}{\sigma}\right) \Phi\left(\alpha \frac{y - \mu}{\sigma}\right),$$

$y, \mu, \alpha \in \mathbb{R}$ and $\sigma > 0$,

denoted as $Y \sim NA(\mu,\sigma^2,\alpha)$. 

In the multivariate case, from Azzalini (2005), the skew-normal distribution for a random vector $Y$ with $n$ observations $(y_1,...,y_n)$ can be

$$f(y;\mu,Y,\alpha) = 2(2\pi)^{-k/2} |\Omega_Y|^{-1/2}$$

$$\exp\left[-\frac{1}{2} (y - \mu)^\top \Omega_Y^{-1} (y - \mu)\right] \Phi\left(\alpha \Omega_Y^{-1} (y - \mu)\right),$$

in which $\Phi(\cdot)$ is defined as previously, the vector $k$-variate $\alpha$ controls the shape of the distribution and determines the direction of maximum asymmetry. Thus, we denote $Y \sim NA(\mu,Y,\alpha)$, so that

$$\Omega_Y = \begin{bmatrix} \delta_1 & \omega^* \\ \omega^* & \delta_2 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & \delta_1 & \omega^* \\ \delta_2 & 1 \omega^* & 1 \end{bmatrix}.$$ 

where $\omega^*$ is the square root of the $\text{diag}(\Omega_z)$, which denotes the diagonal matrix formed by $\Omega_z$. The vector can be written as $\delta = (1 + \alpha \Omega_z \alpha)^{-1/2} \Omega_z \alpha$, where $\Omega_z = (\omega^*)^{-1} \Omega_z (\omega^*)^{-1}$ is the correlation matrix associated with $\Omega_z$. Thus, the expected value and the variance of $Y$ are

$$E(Y) = \mu_z + \omega^* \mu_z$$

and

$$\text{Var}(Y) = \omega^* \mu_z \mu_z^\top \omega^*,$$

where $\mu_z$ is given by the Equation (3).

Based on Azzalini and Capitanio (1999), for $n$ independent observations $(y_1,...,y_n)$ sampled from $NA(\mu,Y,\alpha)$ with $i = 1,...,n$, the log-likelihood function for the random vector $Y$ is

$$l(y;\mu,Y,\alpha) = -\frac{1}{2} n \log |\Omega_Y| - \frac{1}{2} n \text{tr}(\Omega_Y^{-1} V) +$$

$$+ \sum_i \log \left(2 \Phi(\alpha (\omega^*)^{-1} (y_i - \mu))\right)$$

where

$$V = n^{-1} \sum (y_i - \mu)(y_i - \mu)^\top.$$

In the case of covariates

$$\mu_i = X_i \beta$$

where $\beta_{\text{pck}}$ is the parameter matrix and $X_{\text{cov}}$ is a matrix of covariates.

Testing the model significance for the distribution asymmetry, Azzaline and Capitanio (1999) suggested the log-likelihood ratio test to verify the normality. The null hypothesis is $\alpha = 0$ and the test statistics is:

$$2[l(\hat{\mu},\hat{\Omega},\hat{\alpha}) - l(\hat{\mu}^*,\hat{\Omega}^*,0)]$$

(5)

where $(\hat{\mu}^*,\hat{\Omega}^*)$ are the estimates of maximum likelihood of $(\mu,\Omega)$ under the assumption of normality. Under $H_0$, if the statistical test is higher than the $\chi^2$, the hypothesis of data normality is rejected.

In the current analysis, we fit the bivariate skew-normal model for the response variables $Y_1$ and $Y_2$ and
then, we fit the model using the soil porosity as the covariate. The analyses were carried out using the R Software (R DEVELOPMENT CORE TEAM, 2011) and adaptations of the functions of the package sn (AZZALINI, 2011).

We compared the model using the Akaike information criterion - AIC (AKAIKE, 1974) and corrected Akaike information criterion - AICc (BOZDOGAN, 1987) which are

\[
\text{AIC} = n_{par} - 2 \ln(l)
\]

and

\[
\text{AICc} = \text{AIC} + \frac{2n_{par}(n_{par} + 1)}{n - n_{par} - 1}
\]

respectively, where \( n_{par} \) is the number of parameter in the model and 1 is the maximum value of the likelihood function for the estimated model.

We verified the normality of the bivariate data using the quantil-quantil plots and the probability-probability plots (QQ and PP, respectively). The purpose of QQ plot is to calculate the expected value for the observed value of the variable based on the distribution, i.e. the normal. The PP plot, in turn, compares the empirical cumulative distribution of the variable with the theoretical cumulative distribution function as the best fit. In both plots, dots close to the line indicate that the data are following the distribution under study. We also showed the contours of the bivariate normal density for both models. Thereafter, we applied the Equation (1) to transform the bivariate data into compositional data. Finally, we presented a diagram of texture classification using the original compositions and the compositions estimated by the model.

Results and discussion

Table 1 lists the parameter estimates with and without the total soil porosity. Setting the limits of IC at 95% for the first model, the results of the average of the variable \( Y_2 \) as the skew components for the first model were significant, unlike those verified in the second model. However, the confidence intervals had limits close to zero maintaining the skew parameter of the model (Tables 1 and 2). This result is consistent with the responses in the Table 2 where the hypothesis of bivariate normality of data is also rejected for the model without the covariate.

In Figures 1 and 2, the QQ plots exhibited the dots closest to the line for the skew-normal model with the covariate. Similarly, the PP plots also confirmed the best goodness-of-fit to the skew-model when the total soil porosity is considered in the model.

Table 1. Estimates and confidence interval for parameters from the skew-normal model with and without the covariate.

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>L_95%</th>
<th>L_95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without covariate</td>
<td>( \beta_0 )</td>
<td>-0.1109</td>
<td>-0.2508</td>
<td>0.0290</td>
</tr>
<tr>
<td></td>
<td>( \beta_1 )</td>
<td>-0.6122</td>
<td>-0.7326</td>
<td>-0.4917</td>
</tr>
<tr>
<td></td>
<td>( \alpha_1 )</td>
<td>-6.4395</td>
<td>-12.5095</td>
<td>-0.3696</td>
</tr>
<tr>
<td></td>
<td>( \alpha_2 )</td>
<td>2.9699</td>
<td>-0.5065</td>
<td>6.4463</td>
</tr>
<tr>
<td>With covariate</td>
<td>( \beta_0 )</td>
<td>1.6251</td>
<td>1.1015</td>
<td>2.1488</td>
</tr>
<tr>
<td></td>
<td>( \beta_1 )</td>
<td>-0.0470</td>
<td>-0.0605</td>
<td>-0.0034</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 )</td>
<td>0.6607</td>
<td>0.2709</td>
<td>1.0504</td>
</tr>
<tr>
<td></td>
<td>( \beta_3 )</td>
<td>-0.0351</td>
<td>-0.0454</td>
<td>-0.0247</td>
</tr>
<tr>
<td></td>
<td>( \alpha_1 )</td>
<td>-7.2197</td>
<td>-15.0389</td>
<td>0.5996</td>
</tr>
<tr>
<td></td>
<td>( \alpha_2 )</td>
<td>3.0083</td>
<td>-0.7418</td>
<td>6.7585</td>
</tr>
</tbody>
</table>

Table 2. Estimates and confidence interval for the asymmetric coefficient for the skew-normal model with and without the covariate.

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>L_95% (( \lambda ))</th>
<th>L_95% (( \lambda ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without covariate</td>
<td>( \lambda_1 )</td>
<td>-0.9038</td>
<td>-0.9696</td>
<td>-0.0102</td>
</tr>
<tr>
<td></td>
<td>( \lambda_2 )</td>
<td>0.6623</td>
<td>-0.0248</td>
<td>0.9040</td>
</tr>
<tr>
<td>With covariate</td>
<td>( \lambda_1 )</td>
<td>-0.9214</td>
<td>-0.9774</td>
<td>0.0391</td>
</tr>
<tr>
<td></td>
<td>( \lambda_2 )</td>
<td>0.6883</td>
<td>-0.0077</td>
<td>0.9117</td>
</tr>
</tbody>
</table>

Obs: values calculated from the Equation (4) using the estimates of \( \alpha_i \) (i=1,2), Table 1. The models are compared by the Equation (5) that confirmed based on the previous results that the model with the covariate had a best goodness-of-fit due to the lower AIC and AICc values.

Table 3. Test statistics and p-value for the likelihood ratio test for bivariate normality.

<table>
<thead>
<tr>
<th>Model</th>
<th>Statistics</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>without covariate</td>
<td>14.6771</td>
<td>0.0007</td>
</tr>
<tr>
<td>with covariate</td>
<td>10.4633</td>
<td>0.0053</td>
</tr>
</tbody>
</table>

Table 4. Log-likelihood and comparison of the skew-normal models with and without the covariate.

<table>
<thead>
<tr>
<th>Models</th>
<th>Log-likelihood</th>
<th>AIC</th>
<th>AICc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without covariate</td>
<td>-4.6815</td>
<td>17.3629</td>
<td>17.8824</td>
</tr>
<tr>
<td>With covariate</td>
<td>14.7821</td>
<td>17.5642</td>
<td>16.4442</td>
</tr>
</tbody>
</table>
Figure 1. QQ-plots for the normal (a) and skew-normal distribution (b), PP-plots for the normal (c) and skew-normal (d) for the model without covariate.

Figure 2. QQ-plots for the normal (a) and skew-normal distribution (b), PP-plots for normal (c) and skew-normal distribution (d) with the covariate.
Figure 3. Contours of the bivariate model without (a) and with (b) the covariate.

Figure 4 illustrates the diagram of texture classification using the soil composition and the composition estimates by the model with the covariate. Thus, this soil is classified as Clay-loam to Very Clay, across the Clay. Otherwise, the model is satisfactory and the composition determines the soil classification as Clay-loam and Clay.

Conclusion

The results indicate that the bivariate skew-normal distribution is an alternative for modeling transformed (ALR) compositional soil data. The model is more appropriate using the soil porosity as covariate than the normal distribution. Experiments with particle size analysis can achieve the benefits of this methodology because of the inclusion of compositional data with three components.

Acknowledgements


References


Modeling asymmetric compositional data


License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.