# Compositional Data in the Presence of Covariate and Correlated Errors: A Bayesian Approach

Jorge Alberto Achcar [1]
Simone Cristina Obage
Teresa Cristina Martins Dias
Vera Lucia Damasceno Tomazella

Department of Statistics
Federal University of São Carlos
C.P. 676, 13565-905, São Carlos, SP., Brazil

**Abstract**

In this paper, we introduce a Bayesian analysis for compositional data considering additive log-ratio (ALR) and Box-Cox transformations assuming a multivariate normal distribution for correlated errors. These results generalize some existing Bayesian approaches assuming uncorrelated errors. We also consider the use of exponential power distributions for uncorrelated errors of the transformed compositional data. These models gives a better fit for compositional data. We illustrate the proposed methodology considering a real data set.

**Key Words:** Compositional data, Correlated errors, Bayesian analysisMarkov Chain Monte Carlo Methods.

# 1 Introduction

Compositional data are vectors of propositions specifying $G$ fractions as a whole. Thus, for $\mathbf{x} = (x_1, x_2, \cdots, x_G)'$ to be a compositional vector, we must have $x_i > 0$, for $i = 1, \cdots, G$ and $x_1 + x_2 + \cdots + x_G = 1$.

Compositional data often result when raw data are normalized or when data is obtained as proportions of a certain heterogeneous quantity. These conditions are usual in geology, economics and biology.

Standard existing methods to analyse multivariate data under the usual assumption of multivariate normal distribution (see for example, Johnson and Wichern, 1998) are not appropriate to analyse compositional data, since we have compositional restrictions.

Different modelling has been considered to analyse compositional data. A first model considered to analyse compositional data was given by the Dirichlet distribution, but this model requires that the correlation structure is wholly negative, a fact not observed for compositional data where some correlations are positive (see for example, Aitchison, 1982; or Aitchison, 1986).

[1]Corresponding author E-mail: jachcar@power.ufscar.br

Aitchison and Shen (1985) introduced the lognormal distribution to analyse compositional data, transforming the $G$ component vector $\mathbf{x}$ to a vector $\mathbf{y}$ in $R^{G-1}$ considering the additive log-ratio (ALR) function.

Rayens and Srinivasan (1991a, 1991b) extended the ALR transformation considering Box-Cox (1964) transformations as a generalization of the log-ratio function.

Usually we have great difficulties to get classical inference results for these models, especially in the presence of a vector of covariates. Alternatively, the use of Bayesian methods (see for example, Gelman *et al.*, 1995) is a good alternative to analyse compositional data (see for example, Iyengar and Dey, 1996, 1998; or Tjelmeland and Lund, 2003), especially considering Markov Chain Monte Carlo (MCMC) methods (see for example, Gelfand and Smith, 1990 or Roberts and Smith, 1993).

As an illustration of compositional data, we have in Table 1, a real data set comprising of sand, silt and clay compositions taken at various water depths in an Artic lake (see Coakley and Rust, 1968).

Table 1: Sand, silt, clay compositions taken at diferent water depths in an Artic lake

| Sample | Sand | Silt | Clay | Depth | Sample | Sand | Silt | Clay | Depth |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 77.5 | 19.5 | 3.0 | 10.4 | 21 | 9.5 | 53.5 | 37 | 47.1 |
| 2 | 71.9 | 24.9 | 3.2 | 11.7 | 22 | 17.1 | 48 | 34.9 | 48.4 |
| 3 | 50.7 | 36.1 | 13.2 | 12.8 | 23 | 10.5 | 55.4 | 34.1 | 49.4 |
| 4 | 52.2 | 40.9 | 6.6 | 13 | 24 | 4.8 | 54.7 | 41 | 49.5 |
| 5 | 70 | 26.5 | 3.5 | 15.7 | 25 | 2.6 | 45.2 | 52.2 | 59.2 |
| 6 | 66.5 | 32.2 | 1.3 | 16.3 | 26 | 11.4 | 52.7 | 35.9 | 60.1 |
| 7 | 43.1 | 55.3 | 1.6 | 18 | 27 | 6.7 | 46.9 | 46.4 | 61.7 |
| 8 | 53.4 | 36.8 | 9.8 | 18.7 | 28 | 6.9 | 49.7 | 43.4 | 62.4 |
| 9 | 15.5 | 54.4 | 30.1 | 20.7 | 29 | 4.0 | 44.9 | 51.1 | 69.3 |
| 10 | 31.7 | 41.5 | 26.8 | 22.1 | 30 | 7.4 | 51.6 | 40.9 | 73.6 |
| 11 | 65.7 | 27.8 | 6.5 | 22.4 | 31 | 4.8 | 49.5 | 45.7 | 74.4 |
| 12 | 70.4 | 29 | 0.6 | 24.4 | 32 | 4.5 | 48.5 | 47 | 78.5 |
| 13 | 17.4 | 53.6 | 29 | 25.8 | 33 | 6.6 | 52.1 | 41.3 | 82.9 |
| 14 | 10.6 | 69.8 | 19.6 | 32.5 | 34 | 6.7 | 47.3 | 45.9 | 87.7 |
| 15 | 38.2 | 43.1 | 18.7 | 33.6 | 35 | 7.4 | 45.6 | 46.9 | 88.1 |
| 16 | 10.8 | 52.7 | 36.5 | 36.8 | 36 | 6.0 | 48.9 | 45.1 | 90.4 |
| 17 | 18.4 | 50.7 | 30.9 | 37.8 | 37 | 6.3 | 53.8 | 39.9 | 90.6 |
| 18 | 4.6 | 47.4 | 48 | 36.9 | 38 | 2.5 | 48 | 49.5 | 97.7 |
| 19 | 15.6 | 50.4 | 34 | 42.2 | 39 | 2.0 | 47.8 | 50.2 | 103.7 |
| 20 | 31.9 | 45.1 | 23 | 47.0 | | | | | |

(The compositional data are given by proportions times 100)

In this paper, we introduce a Bayesian approach to analyse compositional data assuming different model structure for the error: the ALR transformation and their generalization

given by the Box-Cox transformation assuming correlated errors with multivariate normal distributions which generalizes the results of Iyengar and Dey (1996, 1998) for uncorrelated errors and also assuming the exponential power distribution (see for example Box and Tiao, 1973, p. 157) for the transformed ALR data assuming uncorrelated errors.

The paper is organized as follows: in section 2, we introduce the additive log-ratio (ALR) transformation for compositional data; in section 3, we introduce a Bayesian analysis for the compositional data considering the ALR transformation, the Box-Cox transformation and assuming an exponential power distribution for uncorrelated errors; in section 4, we present a Bayesian analysis for the Artic lake compositional data introduced in Table 1 and finally in section 5 we introduce some concluding remarks.

# 2 Compositional Data Assuming an Additive Log-Ratio (ALR) Transformation

To model compositional data, let us assume the regression model (see for example, Iyengar and Dey, 1996, 1998) given by

$$\mathbf{y}_i = \alpha + \Theta \mathbf{z}_i + \varepsilon_i \tag{1}$$

where $\mathbf{z}_i$ is a $(p \times 1)$ vector of covariates associated to the $i^{th}$ sample; $\alpha$ is a $(g \times 1)$ vector of intercept parameters, $\Theta$ is a $(g \times p)$ matrix of regression parameters; $\varepsilon_i$ is a vector of errors and $\mathbf{y}_i = (y_{i1}, \cdots, y_{ig})'$ is a $(g \times 1)$ vector of compositional data where $g = G - 1$ and $G$ is the number of compositional components. The compositional data is given by $y_{ik} = H(x_{ik}/x_{iG})$, for $i = 1, \cdots, n$ , $k = 1, \cdots, g$, where $H(\cdot)$ is a transformation function to have real components such that $x_{ik} > 0$ and $\sum_{k=1}^{G} x_{ik} = 1$.

The additive log-ratio (ALR) transformation is given by

$$y_{ik} = H\left(\frac{x_{ik}}{x_{iG}}\right) = \log\left(\frac{x_{ik}}{x_{iG}}\right) \tag{2}$$

where $i = 1, \cdots, n$ , $k = 1, \cdots, g$.

The ALR transformation (2) is a special case of the Box-Cox transformation,

$$y_{ik} = H\left(\frac{x_{ik}}{x_{iG}}\right) = \begin{cases} \frac{(x_{ik}/x_{iG})^{\lambda_k} - 1}{\lambda_k} & \text{if } \lambda_k \neq 0 \\ \log(x_{ik}/x_{iG}) & \text{if } \lambda_k = 0 \end{cases} \tag{3}$$

where $\lambda_k$ is an unknown parameter, $i = 1, \cdots, n$ , $k = 1, \cdots, g$.

As a special case, let us consider the compositional data of Table 1. Assuming the ALR transformation with $y_{i1} = \log(x_{i1}/x_{i3})$ and $y_{i2} = \log(x_{i2}/x_{i3})$, where $x_{i1}$ denotes the sand proportion, $x_{i2}$ denotes the silt proportion and $x_{i3}$ denotes the clay proportion for the $i^{th}$ water depth and the model (1), we have,

$$\begin{cases} y_{i1} = \alpha_1 + \theta_1 Z_i + \varepsilon_{i1} \\ y_{i2} = \alpha_2 + \theta_2 Z_i + \varepsilon_{i2} \end{cases} \tag{4}$$

where $i = 1, \cdots, n$, $Z_i$ is a covariate representing the water depth, $\mathbf{y}_i = (y_{i1}, y_{i2})'$ is the response vector and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})'$ has a bivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ with the variance-covariance matrix given by,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{5}$$

and $\rho$ is the correlation coefficient between $\varepsilon_{i1}$and $\varepsilon_{i2}$.

Assuming the model (4) , the likelihood function for $\mathbf{v}_1 = (\alpha_1, \alpha_2, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho)$ is given by,

$$L(\mathbf{v}_1) = \prod_{i=1}^{n} f(\mathbf{y}_i | \mathbf{v}_1) \tag{6}$$

where $\mathbf{y}_i = (y_{i1}, y_{i2})'$ has a bivariate normal distribution $N(\mu_i, \mathbf{\Sigma})$ where $\mu_i = (\mu_{i1}, \mu_{i2})'$,

with $\mu_{1i} = \alpha_1 + \theta_1 Z_i$, $\mu_{2i} = \alpha_2 + \theta_2 Z_i$ and $\mathbf{\Sigma}$ is the variance-covariance matrix given in (5) for $i = 1, \cdots, n$.

# 3 A Bayesian analysis for the compositional data

For a Bayesian analysis for the compositional data of Table 1, assuming the ALR transformation model (4) with a bivariate normal distribution for the errors $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})'$, let us assume the following prior distributions for the parameters:

$$\begin{array}{llll} \text{(i)} & \theta_k & \sim & N\left(a_k, b_k^2\right), \quad a_k, b_k^2 \text{ known;} \\ \text{(ii)} & \alpha_k & \sim & N\left(c_k, d_k^2\right), \quad c_k, d_k^2 \text{ known;} \\ \text{(iii)} & \sigma_k^2 & \sim & IG\left[e_k, f_k\right], \quad e_k, f_k \text{ known;} \\ \text{(iv)} & \rho & \sim & U\left[-1, 1\right], \end{array} \tag{7}$$

for $k = 1, 2$, where $N(\mu, \sigma^2)$ denotes a normal distribution with mean $\mu$ and variance $\sigma^2$; $IG(a, b)$ denotes an inverse gamma distribution with mean $b/(a-1)$ and variance $b^2/\left[(a-1)^2(a-2)\right]$, $a > 2$ and $U[c, d]$ denotes an uniform distribution in the interval $[c, d]$.

Other choices for the prior distributions could be considered as an inverse Wishart distribution to account for the covariance structure assuming $\sigma_k^2$ dependent, but the use of an inverse gamma distribution for $\sigma_k^2$ and an uniform prior for the correlation coefficient $\rho$ gives a great flexibility for practical use. In this way, we could incorporate prior opinion of experts

or to use standard empirical Bayesian methods to choose the values for the hyperparameters of the prior distributions.

Assuming prior independence among the parameters, the joint posterior distribution for $\mathbf{v}_1 = (\alpha_1, \alpha_2, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho)$ is given by

$$\begin{aligned}
\pi(\mathbf{v}_1 \mid D) \;\propto\; & \left\{ \prod_{k=1}^{2} \exp\left[ -\frac{1}{2b_k^2}(\theta_k - a_k)^2 \right] \right\} \\
& \times \left\{ \prod_{k=1}^{2} \exp\left[ -\frac{1}{2d_k^2}(\alpha_k - d_k)^2 \right] \right\} \\
& \times \left\{ \prod_{k=1}^{2} (\sigma_k^2)^{-(e_k+1)} e^{-f_k/\sigma_k^2} \right\} \left(\sigma_1^2\right)^{-n/2} \left(\sigma_2^2\right)^{-n/2} (1-\rho^2)^{-n/2} \\
& \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[ \frac{1}{\sigma_1^2} \sum_{i=1}^{n} \epsilon_{i1}^2 - \frac{2\rho}{\sigma_1 \sigma_2} \sum_{i=1}^{n} \epsilon_{i1}\epsilon_{i2} + \frac{1}{\sigma_2^2} \sum_{i=1}^{n} \epsilon_{i2}^2 \right] \right\}
\end{aligned} \tag{8}$$

where $D$ denote the data, $\varepsilon_{i1} = y_{i1} - \alpha_1 - \theta_1 Z_i$ and $\varepsilon_{i2} = y_{i2} - \alpha_2 - \theta_2 Z_i$ for $i = 1, \cdots, n$.

To get the posterior summaries of interest, we have difficulties to obtain samples from the posterior distribution for $\mathbf{v}_1$ given in (8) directly using analytical or numerical approximations. In this way, we use Markov chain simulation methods (the generalized Metropolis algorithm and the Gibbs Sampler).

The conditional posterior distributions needed for the Gibbs sampling algorithm (see for example, Gelfand and Smith, 1990) are given in an Appendix 1, section 6.1, at the end of this paper.

Observe that the assumption of correlated errors should be considered for a Bayesian analysis of composition data assuming the ALR transformation model (4) with a bivariate normal distribution for the errors, since in applied work we have this structure for compositional data. However, in some situations, we could assume in a second stage of the Bayesian analysis, uncorrelated errors, when we observe that zero is included in credible intervals for $\rho$. Assuming $\rho = 0$ (independent errors), the conditional posterior distributions for the Gibbs sampling algorithm have known distributions, given by,

$$\text{i)} \quad \pi(\theta_k \mid \nu_{(\theta_k)}, D) \sim N\left\{ \frac{a_k \sigma_k^2 + b_k^2 \sum_{i=1}^{n} Z_i u_i^{(k)}}{\sigma_k^2 + b_k^2 \sum_{i=1}^{n} Z_i^2}, \frac{b_k^2 \sigma_k^2}{\sigma_k^2 + b_k^2 \sum_{i=1}^{n} Z_i^2} \right\}$$

where $k = 1, 2$ and $u_i^{(k)} = y_{ik} - \alpha_k$ ; $i = 1, \cdots, n$;

$$\text{ii)} \quad \pi(\alpha_k \mid \nu_{(\alpha_k)}, D) \sim N\left\{ \frac{c_k \sigma_k^2 + d_k^2 \sum_{i=1}^{n} \xi_i^{(k)}}{\sigma_k^2 + n d_k^2}, \frac{d_k^2 \sigma_k^2}{\sigma_k^2 + n d_1^2} \right\} \tag{9}$$

where $k = 1, 2$ and $\xi_i^{(k)} = y_{ik} - \theta_k Z_i$ ; $i = 1, \cdots, n$.

$$\text{iii)} \quad \pi\left(\sigma_k^2 \mid \mathbf{v}_{(\sigma_k^2)}, D\right) \propto IG\left[e_k + \frac{n}{2}, f_k + \frac{1}{2}\sum_{i=1}^{n} \epsilon_{ik}^2\right]$$

where $k = 1, 2$ and $\epsilon_{ik} = y_{ik} - \alpha - \theta_k Z_i \ ; \ \ i = 1, \cdots, n$;

Since the conditional distributions for $\theta_k, \alpha_k$ and $\sigma_k^2$ are well known distributions, we use the Gibbs sampling algorithm to simulate samples for joint posterior distribution for $\theta_k, \alpha_k$ and $\sigma_k^2, k = 1, 2$.

In many applications of compositional data, the ALR transformation could not be well fitted by the data. In this case, we could consider the Box-Cox transformation (3) .

Assuming the Box-Cox transformation (3) we have $y_{i1}^{(\lambda_1)} = \left[(x_{i1}/x_{i3})^{\lambda_1} - 1\right]/\lambda_1$ if $\lambda_1 \neq 0$ and $y_{i1}^{(\lambda_1)} = \log(x_{i1}/x_{i3})$ if $\lambda_1 = 0$; also $y_{i2}^{(\lambda_2)} = \left[(x_{i2}/x_{i3})^{\lambda_2} - 1\right]/\lambda_2$ if $\lambda_2 \neq 0$ and $y_{i2}^{(\lambda_2)} = \log(x_{i2}/x_{i3})$ if $\lambda_2 = 0$.

With $y_{i1}^{(\lambda_1)}$ and $y_{i2}^{(\lambda_2)}$ in place of $y_{i1}$ and $y_{i2}$ in model (4), and also assuming a bivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ for the errors $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})$ where $\Sigma$ is given in (5), the likelihood function for $\mathbf{v}_2 = (\lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho)$ is given by,

$$\begin{aligned} L(\mathbf{v}_2) \ \propto \ & \left(\sigma_1^2\right)^{-n/2}\left(\sigma_2^2\right)^{-n/2}(1-\rho^2)^{-n/2}\left(\prod_{i=1}^{n} y_{i1}^{\lambda_1}\right)\left(\prod_{i=1}^{n} y_{i2}^{\lambda_2}\right) \\ & \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_1^2}\sum_{i=1}^{n}\epsilon_{i1}^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^{n}\epsilon_{i1}\epsilon_{i2} + \frac{1}{\sigma_2^2}\sum_{i=1}^{n}\epsilon_{i2}^2\right]\right\} \end{aligned} \tag{10}$$

where $\epsilon_{i1} = y_{i1}^{(\lambda_1)} - \alpha_1 - \theta_1 Z_i$, $\varepsilon_{i2} = y_{i2}^{(\lambda_2)} - \alpha_2 - \theta_2 Z_i$, $i = 1, \cdots, n$; $\left\{\prod_{i=1}^{n} y_{i1}^{\lambda_1}\right\}$ and $\left\{\prod_{i=1}^{n} y_{i2}^{\lambda_2}\right\}$ are the products of jacobians (see Box and Tiao, 1973).

For a Bayesian analysis, let us assume the prior distributions (7) for $\theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \alpha_1, \alpha_2$ and $\rho$, and

$$\lambda_k \sim N\left(g_k, h_k^2\right), \qquad g_k, \ h_k^2 \quad \text{known} \tag{11}$$

for $k = 1, 2$.

Assuming prior independence among the parameters, the conditional posterior distributions needed for the Gibbs sampling algorithm are given in Appendix 1, section 6.2 at the end of this paper.

Observe that, using Gibbs with Metropolis-Hastings algorithms, we find the posterior summaries of interest for $\lambda_1, \lambda_2$ and all other parameters of the model. If zero is included in the credible intervals for $\lambda_1$ and $\lambda_2$ we could reanalyze the data assuming the ALR transformation model. That is, we get better inference results for the compositional data.

We also could consider other distributions for the errors $\epsilon_{i1}$ and $\epsilon_{i2}$ in the ALR transformation model (4), to get better inference results for the compositional data.

In this case, let us assume independent errors $\varepsilon_{i1}$ and $\varepsilon_{i2}$ for the ALR transformation model (4) with exponential power distributions (see for example, Box and Tiao, 1973) given by,

$$\pi\left(\epsilon_{ik}|\sigma_1,\beta_k^*\right)=\frac{\omega\left(\beta_k^*\right)}{\sigma_k}\exp\left\{-c\left(\beta_k^*\right)\left|\frac{\epsilon_{ik}}{\sigma_k}\right|^{\frac{2}{1+\beta_k^*}}\right\} \tag{12}$$

where $-\infty<\epsilon_{ik}<\infty$ , $c\left(\beta_k^*\right)=\left\{\frac{\Gamma\left[\frac{3}{2}\left(1+\beta_k^*\right)\right]}{\Gamma\left[\frac{1}{2}\left(1+\beta_k^*\right)\right]}\right\}^{\frac{2}{1+\beta_k^*}}$, $\omega\left(\beta_k^*\right)=\frac{\left\{\Gamma\left[\frac{3}{2}\left(1+\beta_k^*\right)\right]\right\}^{\frac{1}{2}}}{\left(1+\beta_k^*\right)\left\{\Gamma\left[\frac{1}{2}\left(1+\beta_k^*\right)\right]\right\}^{\frac{3}{2}}}$ and $\sigma_j>0$, $-1<\beta_k^*\leq 1$ and $k=1,2$.

Some special cases of model (12) are given by,

$(i)$ If $\beta_k^*=0$, we have normal distributions for $\epsilon_{ik}$ with mean zero and variance $\sigma_k^2$, $k=1,2$;

(ii) If $\beta_k^*=1$, we have double exponential distributions for $\epsilon_{ik}$;

(iii) If $\beta_k^*\to -1$, we have uniform distributions for $\epsilon_{ik}$ in the interval $\left[-\sqrt{3}\sigma_k\ ,\ \sqrt{3}\sigma_k\right]$, $k=1,2$.

The likelihood function for $\mathbf{v}_3=(\alpha_1,\alpha_2,\theta_1,\theta_2,\beta_1^*,\beta_2^*,\sigma_1,\sigma_2)$ is given by

$$\begin{aligned} L(\mathbf{v}_3) \ \propto\ & \prod_{i=1}^{n}\frac{\omega\left(\beta_1^*\right)}{\sigma_1}\exp\left\{-c\left(\beta_1^*\right)\left|\frac{\epsilon_{i1}}{\sigma_1}\right|^{\frac{2}{1+\beta_1^*}}\right\} \\ & \times\frac{\omega\left(\beta_2^*\right)}{\sigma_2}\exp\left\{-c\left(\beta_2^*\right)\left|\frac{\epsilon_{i2}}{\sigma_2}\right|^{\frac{2}{1+\beta_2^*}}\right\}, \end{aligned} \tag{13}$$

where $\epsilon_{i1}=y_{i1}-\alpha_1-\theta_1 Z_i$, $\epsilon_{i2}=y_{i2}-\alpha_2-\theta_2 Z_i$, $\omega\left(\beta_1^*\right)$ and $c\left(\beta_1^*\right)$ are given in (12).

Let us assume the following prior distributions for the parameters:

$$\begin{array}{lrcll} \text{(i)} & \theta_k & \sim & N\left(a_k,b_k^2\right), & k=1,2,\ \ a_k,b_k^2 \ \text{ known;} \\ \text{(ii)} & \alpha_k & \sim & N\left(c_k,d_k^2\right), & k=1,2,\ \ c_k,d_k^2 \ \text{ known;} \\ \text{(iii)} & \sigma_k^2 & \sim & IG\left[e_k,f_k\right], & k=1,2 \quad e_k,f_k \ \text{ known;} \\ \text{(iv)} & \beta_k^* & \sim & U\left[-1,1\right], & k=1,2. \end{array} \tag{14}$$

Observe that if $\beta_k^*=0$ (normal distributions for the errors), the conditional distributions for $\theta_k$, $\alpha_k$ and $\sigma_k^2$ needed for the Gibbs sampling algorithm are given in (9).

A great simplification in the simulation of samples for the joint posterior distribution of interest is to use the software WinBugs (Spiegelhalter *et al.*, 1999) where we only need to specify the likelihood function and the prior distributions for the parameters.

# 4 Analysis of the Data of Table 1

To analyze the Artic lake data set of Table 1, let us assume the ALR transformation model with prior distributions (7) with $a_1 = a_2 = c_1 = c_2 = 0$, $b_1^2 = b_2^2 = d_1^2 = d_2^2 = 100000$; $e_1 = e_2 = f_1 = f_2 = 1000$.

Observe that we are considering very large values for the variances of the prior distributions, indicating prior ignorance about the parameters of the model. With this choice for the hyperparameters of the prior distributions, we get similar inference results considering standard frequentist methods. With good prior opinion, we could consider more informative prior distributions and more accurate Bayesian inference results.

Using the software WinBugs we simulated two parallel Gibbs sampling chains where we discarded the first 5000 iterations ("burn-in samples") to eliminate the effect of the initial values.

With a total of 5000 iterations, the chain started essentially in equilibrium, so the "burn-in" can be stopped. We also considered the samples $100^{th}$, $200^{th}$, $300^{th}, \cdots$ to eliminate the correlation among the samples.

In Figure 1, we have the plots for the Gibbs sampling algorithm. We observe convergence of the algorithm using these standard graphical methods. In Figure 2, we have the plots for some estimated autocorrelation functions. We observe approximately uncorrelated Gibbs samples.

The convergence of the Gibbs sampling algorithm also was monitored using the Gelman and Rubin (1992) method that uses the analysis of variance technique to determine if further iterations are needed.

In Table 2, we have the posterior summaries for the quantities of interest. We also have in Table 2, the potencial scale reduction index of Gelman and Rubin. We observe convergence of the Gibbs sampling algorithm since $\sqrt{\hat{R}} < 1.1$ in all cases.

Table 2: Posterior summaries (ALR transformation model)

| Parameters | Mean | 95% Credible interval | $\hat{R}$ |
|---|---|---|---|
| $\alpha_1$ | 2.690 | [1.8350 ; 3.4830] | 1.001 |
| $\alpha_2$ | 1.9600 | [1.4950 ; 2.4420] | 1.003 |
| $\theta_1$ | $-0.0624$ | $[-0.0772 ; -0.0474]$ | 1.001 |
| $\theta_2$ | $-0.0245$ | $[-0.0330 ; -0.0162]$ | 1.002 |
| $\sigma_1^2$ | 1.7730 | [1.1270 ; 2.7800] | 0.999 |
| $\sigma_2^2$ | 0.5849 | [0.3710 ; 0.9175] | 1.004 |
| $\rho$ | 0.8380 | [0.7220 ; 0.9160] | 1.010 |

Assuming the Box-Cox transformation model with adequate choices for the hyperparameters of the prior distributions (7) and (11) to have flat prior distribution for the parameters and following the same simulation steps as considered for the ALR transformation model, we have in Table 3, the posterior summaries for the quantities of interest. We also observe convergence for all parameters since $\hat{R} < 1.1$ for all cases.

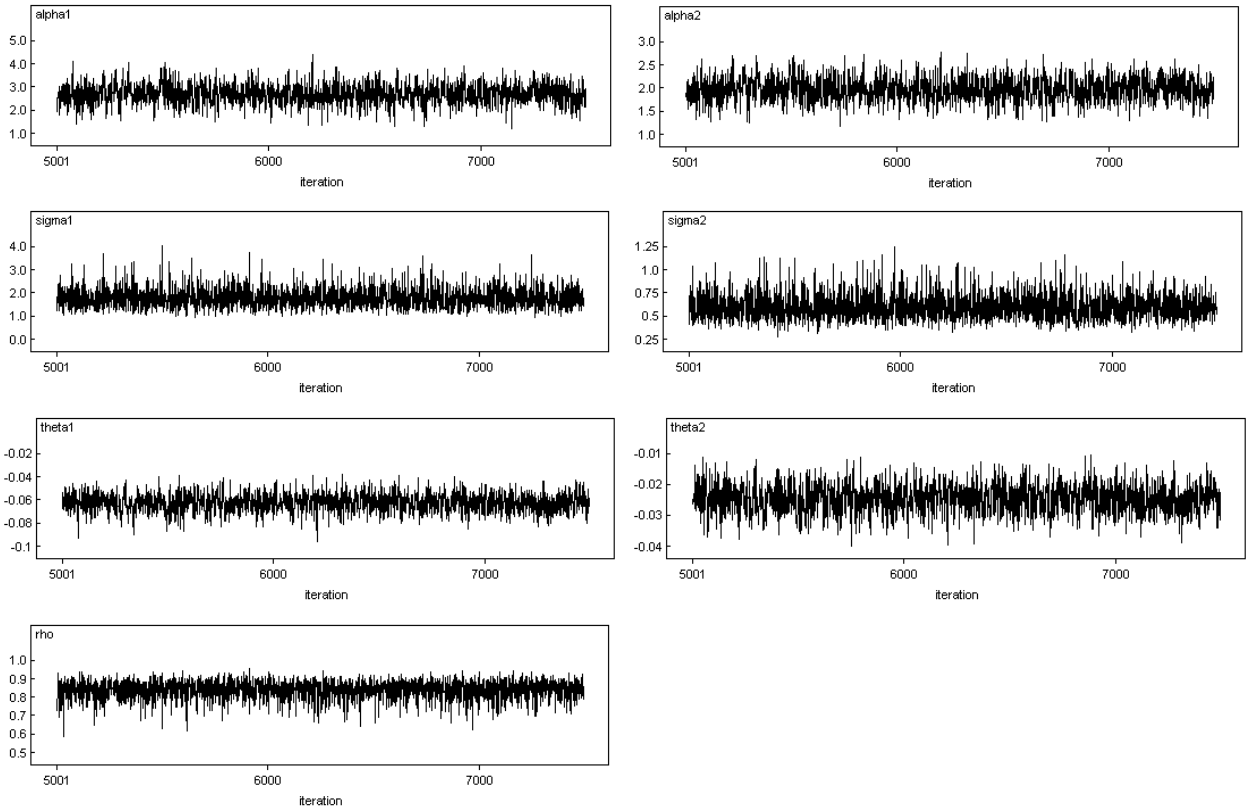


Figure 1: Sampled values for the parameters (ALR transformation model)

Table 3: Posterior summaries ( Box-Cox transformation model)

| Parameters | Mean | 95% Credible interval | $\hat{R}$ |
|---|---|---|---|
| $\alpha_1$ | 2.3150 | [1.544 ; 3.198] | 0.999 |
| $\alpha_2$ | 0.8948 | [0.699 ; 1.146] | 0.994 |
| $\theta_1$ | $-0.0638$ | $[-0.078 ; -0.049]$ | 1.001 |
| $\theta_2$ | $-0.0105$ | $[-0.014 ; -0.007]$ | 1.001 |
| $\sigma_1^2$ | 1.4360 | [0.902 ; 2.252] | 1.001 |
| $\sigma_2^2$ | 0.0466 | [0.022 ; 0.092] | 0.996 |
| $\lambda_1$ | $-0.1782$ | $[-0.295 ; -0.052]$ | 1.000 |
| $\lambda_2$ | $-0.9405$ | $[-1.251 ; -0.643]$ | 1.002 |
| $\rho$ | 0.8202 | [0.692 ; 0.908] | 1.010 |

From the results of Table 2 and Table 3, we observe that the water depths have significative effects on the proportions of the compositional data, since zero is not included in the 95% credible intervals for $\theta_1$and $\theta_2$.

Assuming independent errors for the ALR transformation model (4) with exponential power distributions (12), we also considered adequate choices for the hyperparameters of the

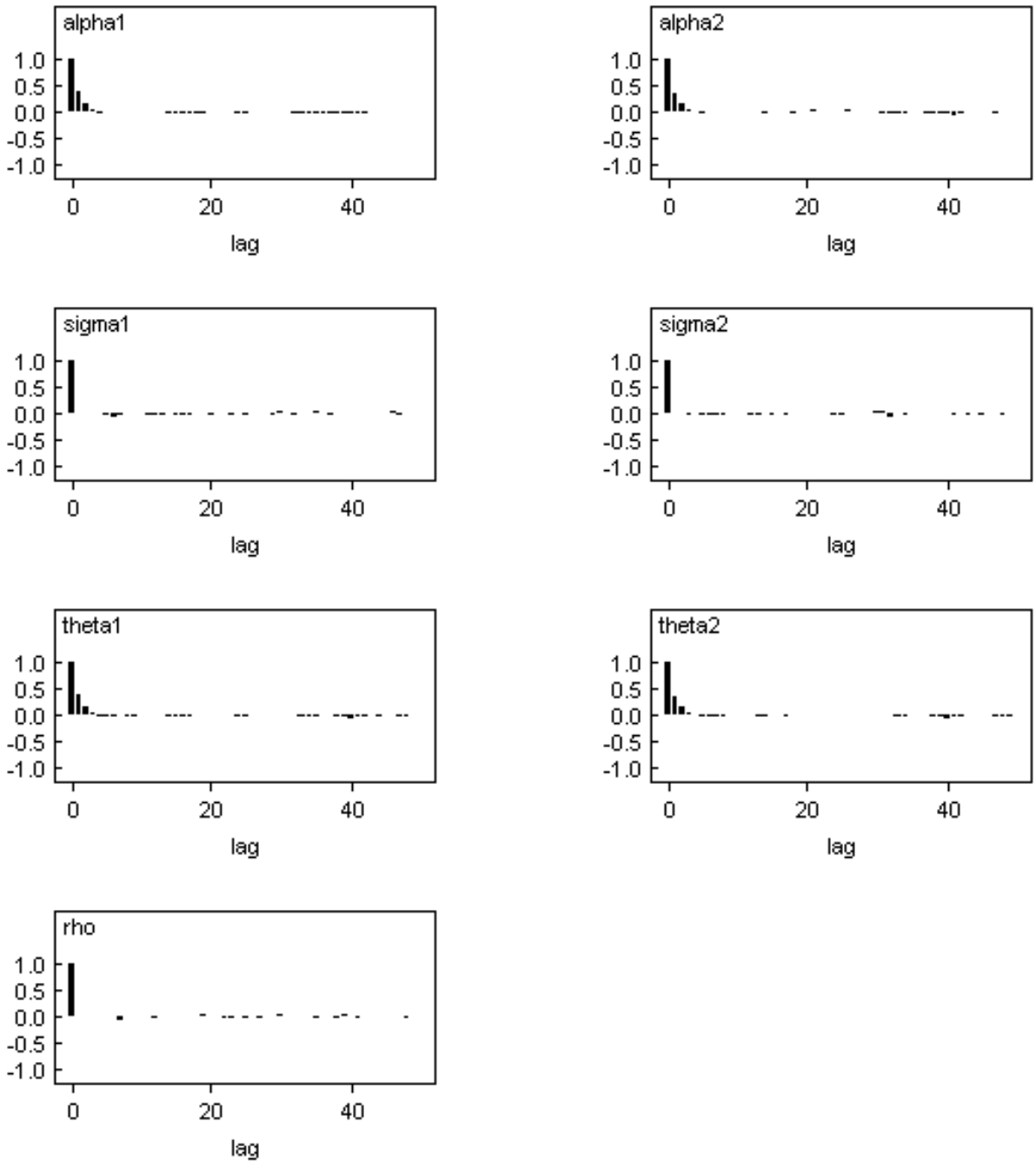


Figure 2: Estimated autocorrelation functions (ALR transformation model)

prior distributions (14) to have flat prior distributions for the parameters. In Table 4, we have the posterior summaries of interest. We also have in the Table 4, the potential scale reduction indexes of Gelman and Rubin.

Table 4: Posterior summaries ( exponential power distributions for the errors)

| Parameters | Mean | 95% Credible interval | $\hat{R}$ |
|---|---|---|---|
| $\alpha_1$ | 2.611 | [1.661 ; 3.612] | 1.001 |
| $\alpha_2$ | 1.511 | [0.990 ; 2.077] | 1.003 |
| $\theta_1$ | $-0.061$ | $[-0.078 ; -0.046]$ | 1.001 |
| $\theta_2$ | $-0.018$ | $[-0.027 ; -0.0105]$ | 1.002 |
| $\sigma_1$ | 1.809 | [1.706 ; 3.249] | 1.003 |
| $\sigma_2$ | 0.487 | [0.248 ; 0.952] | 1.000 |
| $\beta_1^*$ | 0.192 | [$-0.418$ ; 0.900] | 1.010 |
| $\beta_2^*$ | 0.769 | [0.293 ; 0.993] | 1.002 |

Assuming uncorrelated errors in the ALR transformation model (4) with normal distributions for the errors ($\beta_1^* = \beta_2^* = 0$), we have in Table 5, the posterior summaries of interest.

It is important to point out that the computing times for each simulation using the WinBugs software was very small (close to 20 minutes in each case).

Table 5: Posterior summaries ( normal distributions for the errors)

| Parameters | Mean | 95% Credible interval | $\hat{R}$ |
|---|---|---|---|
| $\alpha_1$ | 2.675 | [1.816 ; 3.513] | 1.001 |
| $\alpha_2$ | 1.962 | [1.475 ; 2.447] | 0.998 |
| $\theta_1$ | $-0.063$ | $[-0.078 ; -0.047]$ | 0.998 |
| $\theta_2$ | $-0.025$ | $[-0.034 ; -0.0158]$ | 0.998 |
| $\sigma_1$ | 1.828 | [1.147 ; 2.897] | 1.003 |
| $\sigma_2$ | 0.603 | [0.379 ; 0.9520] | 0.997 |

For model selection, we can use some existing adequacy measures as the Deviance Information Criterion (DIC) (Spiegelhalter *et al.*, 2000) or the Bayesian Information Criterion (BIC) that are aproximations for the Bayes factor.

Carlin and Louis (2000) introduced a modification of the BIC criterion given by,

$$BIC_i = -2E\left[\ln(\mathrm{L}\left(\theta_i|x, M_i\right)\right] + p_i \ln\left(n\right) \tag{15}$$

where $i$ indexes model; $n$ is the sample size and $p_i$ is the number of parameters under model $M_i$. Larger values of BIC indicate better models. Smaller values of DIC indicates better models.

In Table 6, we have the Monte Carlo estimates for the adequacy measures based on the simulated Gibbs samples for the following models: $M_1$ (ALR tansformation model with correlated errors); $M_2$ (ALR transformation model with uncorrelated errors); $M_3$ (Box-Cox transformation model with correlated errors) and $M_4$ (exponential power distribution for the ALR transformation model with uncorrelated errors).

Table 6: Adequacy measure for the models

| Models | DIC | BIC |
|---|---|---|
| $M_1$ | 178.483 | 26.994 |
| $M_2$ | 229.492 | 24.644 |
| $M_3$ | 127.244 | 30.166 |
| $M_4$ | 221.262 | 25.904 |

From the results of Table 6, we conclude that model $M_3$ (Box-Cox transformation model with correlated errors) is better fitted by the compositional data of Table 1 (smaller value for DIC and larger value for BIC).

# 5 Concluding Remarks

We consider in this paper, the modeling of compositional data in the transformed space and on the simplex. Usually, additive log-ratio (ARL) transformation model is considered to analyze correlated compositional data. Some generalizations of the ARL transformation model are considered assuming an exponential power distribution and the Box-Cox transformation with correlated errors. These results generalizes the results obtained by Iyengar and Dey (1996). We also consider different prior distributions for the errors.

We observed that the Bayesian methodology can be successfully employed to analyse compositional data using Markov Chain Monte Carlo methods, especially using some existing software. In this case, the software WinBugs (Spiegelhater *et al.*, 1999) gives a great simplification to simulate samples for the joint posterior distribution of interest for the compositional data. The codes of the WinBugs program are available under request for the first author.

It is important to point out that we could incorporate prior opinion of experts at the model building stage which yields more accurate inferences. Other possibility is to use empirical Bayesian methods.

The use of differents stages in the Bayesian analysis starting with a more general model like the Box-Cox transformed model or using the exponential power distribution for the errors, could be a good alternative to find better models to analyse compositional data. The Bayesian approach also has other strong advantages over other existing inference methods. Among these advantages, we have:

(i) The Bayesian approach does not rely on asymptotic results commonly used in the frequentist approach.

(ii) The presence of a great number of covariates is not a problem for the Bayesian approach based on Markov Chain Monte Carlo.

(iii) One important point in the statistical analysis for compositional data: the presence of missing values. This problem is easily handed in the Bayesian framework.

(iv) The Bayesian approach presents many Bayesian discrimination methods that can be used the search of a better model to analyse compositional data.

**Acknowledgements**: The author would like to thank the editor and a referee for their helpful suggestions and comments.

# 6 Appendix 1

## 6.1 Conditional posterior distribution for the Gibbs sampling algorithm (correlated errors in ARL model)

The conditional posterior distributions obtained from 8 are given by,

$$\text{i) } \pi\left(\theta_1 \mid v_{(\theta_1)}, D\right) \sim N\left\{u_{\theta_1}, \sigma^2_{\theta_1}\right\}$$

where $u_{\theta_1} = \frac{a_1\sigma_1^2\sigma_2(1-\rho^2)+\sigma_2 b_1^2\sum_{i=1}^n Z_i u_i^{(1)}-\sigma_1+b_1^2\rho\sum_{i=1}^n Z_i B_i^{(1)}}{\sigma_2\left[\sigma_1^2(1-rho^2)+b_1^2\sum_{i=1}^n Z_i^2\right]}$; $\sigma^2_{\theta_1} = \frac{b_1^2\sigma_1^2(1-\rho^2)}{\sigma_1^2(1-\rho^2)+b_1^2\sum_{i=1}^n Z_i^2}$; $u_i^{(1)} = y_{i1} - \alpha_1$ and $B_i^{(1)} = y_{i2} - \alpha_2 - \theta_2 Z_i$; $i = 1, \cdots, n$ ; $v_{(\theta_1)}$ is the vector of all of parameters except $\theta_1$ and $Z_i$ is the covariate;

$$\text{ii) } \pi\left(\theta_2 \mid v_{(\theta_2)}, D\right) \sim N\left\{u_{\theta_2}, \sigma^2_{\theta_2}\right\}$$

where $u_{\theta_2} = \frac{a_2\sigma_2^2\sigma_1(1-\rho^2)+\sigma_1 b_2^2\sum_{i=1}^n Z_i u_i^{(2)}-\sigma_2+b_2^2\rho\sum_{i=1}^n Z_i B_i^{(2)}}{\sigma_1\left[\sigma_2^2(1-\rho^2)+b_2^2\sum_{i=1}^n Z_i^2\right]}$; $\sigma^2_{\theta_2} = \frac{b_2^2\sigma_2^2(1-\rho^2)}{\sigma_2^2(1-\rho^2)+b_2^2\sum_{i=1}^n Z_i^2}$; $u_i^{(2)} = y_{i2} - \alpha_2$ and $B_i^{(2)} = y_{i1} - \alpha_1 - \theta_1 Z_i$; $i = 1, \cdots, n$;

$$\text{iii) } \pi\left(\alpha_1 \mid v_{(\alpha_1)}, D\right) \sim N\left\{u_{\alpha_1}, \sigma^2_{\alpha_1}\right\}$$

where $u_{\alpha_1} = \frac{c_1\sigma_1^2\sigma_2(1-\rho^2)+\sigma_2 d_1^2\sum_{i=1}^n \xi_i^{(1)}-\sigma_1 d_1^2\rho\sum_{i=1}^n C_i^{(1)}}{\sigma_2\left[\sigma_1^2(1-\rho^2)+nd_1^2\right]}$ ; $\sigma^2_{\alpha_1} = \frac{d_1^2\sigma_1^2(1-\rho^2)}{\sigma_1^2(1-\rho^2)+nd_1^2}$; $\xi_i^{(1)} = y_{i1} - \theta_1 Z_i$ and $C_i^{(1)} = y_{i2} - \alpha_2 - \theta_2 Z_i$; $i = 1, \cdots, n$;

$$\text{iv) } \pi\left(\alpha_2 \mid v_{(\alpha_2)}, D\right) \sim N\left\{u_{\alpha_2}, \sigma^2_{\alpha_2}\right\}$$

where $u_{\alpha_2} = \frac{c_2\sigma_2^2\sigma_1(1-\rho^2)+\sigma_1 d_2^2\sum_{i=1}^n \xi_i^{(2)}-\sigma_2 d_2^2\rho\sum_{i=1}^n C_i^{(2)}}{\sigma_2\left[\sigma_1^2(1-\rho^2)+nd_1^2\right]}$; $\sigma^2_{\alpha_2} = \frac{d_2^2\sigma_2^2(1-\rho^2)}{\sigma_2^2(1-\rho^2)+nd_2^2}$; $\xi_i^{(2)} = y_{i2} - \theta_2 Z_i$ and $C_i^{(2)} = y_{i1} - \alpha_1 - \theta_1 Z_i$; $i = 1, \cdots, n$;

$$\text{v) } \pi\left(\sigma_1^2 \mid v_{(\sigma_1^2)}, D\right) \sim (\sigma_1^2)^{-(e_1+1)} e^{-f_1/\sigma_1^2}\Psi_1(\mathbf{v}) \tag{16}$$

where $\Psi_1(\mathbf{v}) = \exp\left\{-\frac{n}{2}\ln(\sigma_1^2) - \frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_1^2}\sum_{i=1}^n \epsilon_{i1}^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^n \epsilon_{i1}\epsilon_{i2}\right]\right\}$, $\epsilon_{i1} = y_{i1} - \alpha_1 - \theta_1 Z_i$ and $\epsilon_{i2} = y_{i2} - \alpha_2 - \theta_2 Z_i$ ; $i = 1, \cdots, n$;

$$\text{vi) } \pi\left(\sigma_2^2 \mid \mathbf{v}_{(\sigma_2^2)}, D\right) \sim (\sigma_2^2)^{-(e_2+1)} e^{-f_2/\sigma_2^2}\Psi_2(\mathbf{v})$$

where $\Psi_2(\mathbf{v}) = \exp\left\{-\frac{n}{2}\ln(\sigma_2^2) - \frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_2^2}\sum_{i=1}^n \epsilon_{i2}^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^n \epsilon_{i1}\epsilon_{i2}\right]\right\}$, $\epsilon_{i1} = y_{i1} - \alpha_1 - \theta_1 Z_i$ and $\epsilon_{i2} = y_{i2} - \alpha_2 - \theta_2 Z_i$ ; $i = 1, \cdots, n$;

vii) $\pi(\rho \mid \mathbf{v}_{(\rho)}, D) \propto (1-\rho^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_1^2}\sum_{i=1}^{n}\epsilon_{i1}^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^{n}\epsilon_{i1}\epsilon_{i2} + \frac{1}{\sigma_2^2}\sum_{i=1}^{n}\epsilon_{i2}^2\right]\right\}.$

Since we do not have known forms for the conditional distributions for $\sigma_1^2$, $\sigma_2^2$ and $\rho$, we cannot use the Gibbs sampling to generate samples for $\sigma_1^2$, $\sigma_2^2$ and $\rho$. In this case, we use the Metropolis-Hastings algorithm (see for example, Roberts and Smith, 1993).

## 6.2 Conditional posterior distribution for the Gibbs sampling algorithm (Box-Cox transformation)

$$\text{i) } \pi\left(\lambda_1 | v_{(\lambda_1)}, D\right) \propto \exp\left\{-\frac{1}{2h_1^2}\left(\lambda_1 - g_1\right)^2\right\} \Psi_1\left(\mathbf{v}\right)$$

where

$$\begin{aligned}\Psi_1\left(\mathbf{v}\right) &= \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_1^2}\left(\sum_{i=1}^{n} y_{i1}^{(\lambda_1)2} - 2\alpha_1\sum_{i=1}^{n} y_{i1}^{(\lambda_1)} - 2\theta_1\sum_{i=1}^{n} Z_i y_{i1}^{(\lambda_1)}\right)\right.\right. \\ &\quad \left.\left. -\frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^{n} y_{i1}^{(\lambda_1)} D_i^{(2)} + \lambda_1\sum_{i=1}^{n}\ln y_{i1}\right]\right\}\end{aligned}$$

and $D_i^{(2)} = y_{i2}^{(2)} - \alpha_2 - \theta_2 Z_i, i = 1, \cdots, n.$

$$\text{ii) } \pi\left(\lambda_2 | v_{(\lambda_2)}, D\right) \propto \exp\left\{-\frac{1}{2h_2^2}\left(\lambda_2 - g_2\right)^2\right\} \Psi_2\left(\mathbf{v}\right) \tag{17}$$

where

$$\begin{aligned}\Psi_2\left(\mathbf{v}\right) &= \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_2^2}\left(\sum_{i=1}^{n} y_{i2}^{(\lambda_2)2} - 2\alpha_2\sum_{i=1}^{n} y_{i2}^{(\lambda_2)} - 2\theta_2\sum_{i=1}^{n} Z_i y_{i2}^{(\lambda_2)}\right)\right.\right. \\ &\quad \left.\left. -\frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^{n} y_{i2}^{(\lambda_2)} D_i^{(1)}\right] + \lambda_2\sum_{i=1}^{n}\ln y_{i2}\right\}\end{aligned}$$

and $D_i^{(1)} = y_{i1}^{(1)} - \alpha_1 - \theta_1 Z_i, i = 1, \cdots, n.$

The conditional posterior distributions for $\theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \alpha_1, \alpha_2$ and $\rho$ are given in (16) with $y_{i1}$ and $y_{i2}$ replaced by $y_{i1}^{(\lambda_1)}$ and $y_{i2}^{(\lambda_2)}$, respectively.

# References


[1] Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of Royal Statistical Society*, B, 139-177.

[2] Aitchison, J.; Shen, S. M. (1985). Logistic-normal distributions: Some properties and uses. *Biometrika*, 47:136-146.

[3] Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall.

[4] Box, G.E.P.; Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, B, 26: 211-52.

[5] Box,G. E. P.; Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading: Addison-Wesley.

[6] Carlin, B. P.; Louis, T. (2000). *Bayes and Empirical Bayes methods for data analysis*, $2^{nd}$ed, London: Chapman and Hall.

[7] Coakley, J. P.; Rust, B. R. (1968). Sedimentation in an Arctic lake. *Journal of Sedimentary Petrology*, 38:1290-1300.

[8] Gelfand, A. E.; Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

[9] Gelfand, A. E.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. (1995). *Bayesian Data Analysis*, Chapman & Hall., 85, 398-409.

[10] Gelman, A.; Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.

[11] Iyengar, M.; Dey, D. K. (1996). Bayesian analysis of compositional data. *Department of Statistics, University of Connecticut*, Storrs, CT 06269-3120.

[12] Iyengar, M.; Dey, D. K. (1998). Box-Cox transformations in Bayesian analysis of compositional data. *Environmetrics*, 9, 657-671.

[13] Johnson, R.; Wichern, D. (1982). *Applied multivariate statistical analysis*. New Jersey: Prentice Hall.

[14] Rayens, W. S.; Srinivasan, C. (1991a). Box-Cox transformations in the analysis of compositional data. *Journal of Chemometrics*, 5:227-239.

[15] Rayens, W. S.; Srinivasan, C. (1991b). Estimation in compositional data. *Journal of Chemometrics*, 5:361-374.

[16] Roberts, G. O.; Smith, A. F. M. (1993). Bayesian methods via the Gibbs sampler and related Markov Chain Monte Carlo methods, *Journal of the Royal Statistical Society*, B, Cambridge, v.55, n.1, 3-23.

[17] Spiegelhalter, D. J.; Thomas, A.; Best, N. G. (1999). *WinBugs: Bayesian inference using Gibbs sampling*. Cambridge: MRC Biostatistics Unit.

[18] Spiegelhalter, D. J.; Best, N. G.; Van der Linde, A. (2002). A Bayesian Measure of model complexity and fit (with discussion) *Journal of the Royal Statistical Society*, B, 64, 583-639.

[19] Tjelmeland, H.; Lund, K. V. (2003). Bayesian modelling of spatial compositional data, preprint n.1, *Journal of Applied Statistics*, 30, 87-100.