Journal of Data Science 221-250 , DOI:10.6339/JDS.201804_16(2).0002

Modeling Compositional Regression With Uncorrelated and Correlated Errors: A Bayesian Approach

Taciana K. O. Shimizu[1], Francisco Louzada[2], Adriano K. Suzuki[3], Ricardo S. Ehlers[4]

[1]*Federal University of São Carlos and University of São Paulo*
[2]*University of São Paulo*

Abstract: Compositional data consist of known compositions vectors whose components are positive and defined in the interval (0,1) representing proportions or fractions of a "whole". The sum of these components must be equal to one. Compositional data is present in different knowledge areas, as in geology, economy, medicine among many others. In this paper, we propose a new statistical tool for volleyball data, i.e., we introduce a Bayesian anal- ysis for compositional regression applying additive log-ratio (ALR) trans- formation and assuming uncorrelated and correlated errors. The Bayesian inference procedure based on Markov Chain Monte Carlo Methods (MCMC). The methodology is applied on an artificial and a real data set of volleyball.

Key words:Compositional data, additive log-ratio transformation,inference Bayesian, correlated errors, MCMC.

1. Introduction

Compositional data are vectors of proportions specifying G fractions as a whole. Such data often result when raw data are normalized or when data is obtained as proportions of a certain heterogeneous quantity.

By definition, a vector x in the Simplex sample space is a composition, elements of this vector are components and the vectors set is compositional data (Aitchison (1986)). Therefore, for $x = (x_1, x_2, \cdots, x_G)'$ to be a compositional vector, x_i is non negative value, for $i = 1, \cdots, G$ and $x_1 + x_2 + \cdots + x_G = 1$. The first model addopted for the analysis of

such data was the Dirichlet distribution. However, it requires that the correlation structure is wholly negative, a fact that is not observed for compositional data, in which some correlations are positive (see for example, Aitchison (1982)).

Aitchison and Shen (1980) developed the logistic-Normal class of distributions transforming the G component vector x into a vector y in R^{G-1} and considering the Additive Log-Ratio (ALR) function. The use of Bayesian methods is a good alternative for the analysis of compositional data, for example Achcar and Obage (2005) considered Bayesian analysis using the ALR and Box-Cox transformations in regression models for compositional data assuming correlated errors with bivariate normal distributions ; Iyengar and Dey (1996) developed a complete Bayesian methodology to analyse such data with the implementation of Markov Chain Monte-Carlo methods, comparing with alternative methods as maximum likelihood estimates; Iyengar and Dey (1998) extended the last work (Iyengar and Dey,(1996)) applying Box-Cox transformations for compositional data; Tjelme- land and Lund (2003) defined a spatial model for compositional data in a Bayesian framework and discussed appropriate prior distributions; Tsagris (2014) performs supervised classification of compositional data using the k-NN algorithm.

Some researchers have focused on the performance indicators of volleyball, they have different objectives and statistical procedures. For example, Campos et al. (2014) studied the advantage of playing at home and the influence of performance indicators on the game score according to the number of sets, based on Brazilian and Italian elite women's volleyball leagues. Mesquita and Sampaio (2010) compared the volleyball game-related statistics by sex and analyzed all games of the male and female World Cup in 2007. Afonso et al. (2012) examined predictors of the setting zone in elite-level male volleyball. Rodriguez-Ruiz et al. (2011) analyzed the terminating actions (serve, attack, block and opponent errors) that led to point scoring. Although these studies were based on voleyball data, none of them considered methodology of compositional data.

Thus, the main purpose is to propose a new statistical tool for volleyball data, that is, a compositional regression model assuming correlated and uncorrelated normal errors based on Bayesian approach. Usually, the volleyball data (attack, block, serve and opponent error) have compositional restrictions, i.e., they have dependence structure, being that standard existing methods to analyze multivariate data under the usual assumption of multivariate

normal distribution (see for example, Johnson and Wichern (1998)) are not appropriate to analyze them.

We consider a real data set related to the first and second rounds matches of Brazilian Men's Volleyball Super League 2011/2012 obtained from the website (CBV, 2012). The data concern the teams that played and won the games in such rounds; more specifically, the points of the team that won each game were defined as composition and the proportions of each composition are the volleyball skills, as attack, block, serves and errors of the opposite team.

The points of the winning team in each game were obtained by four components. We denoted x1 the proportion of points in the attack, x2 the proportion of points in the block, x3 the proportion of points in the serve and x4 the proportion of points in the errors of the opposite team.

The paper is organized as follows: Section 2 introduces the formulation of regression model applied through the Additive Log-Ratio (ALR) transformation Bayesian analysis of the proposed model assuming correlated and uncorrelated Normal errors; Section 3 provides the results of the application to an artificial and a real data set related to the Brazilian Men's Volleyball Super League 2011/2012; finally, Section 4 concludes the paper with some final remarks.

2. Methodology

We can consider $y_{ij} = H\left(\frac{x_{ij}}{x_{iG}}\right), j = 1, \cdots, n$ and $j = 1, \cdots, g$, being $H(\bullet)$ the chosen transformation function that assures resulting vector has real components, where x_{ij} represents the $i-th$ observation for the $j-th$ component, such that$x_{i1} > 0, \cdots, x_{iG} > 0$ and $\sum_{j=1}^{G} x_{ij} = 1$, for $i = 1, \cdots, n$.

The ALR transformation for the analysis of compositional data is given by

$$y_{ij} = \left(\frac{x_{ij}}{x_{iG}}\right) = \log(\frac{x_{ij}}{x_{iG}}) \quad (1)$$

The regression model assuming ALR transformation for the response variables is given by

$$y_i = \beta_{zi} + \varepsilon_i \quad (2)$$

where $y_i = (y_{i1}, \cdots, y_{ig})^T$ is a vector $(g \times 1)$ of response variables where $g = G - 1$ and G is the number of components (compositional data); $\beta = (\beta_0, \beta_1, \cdots, \beta_g)$ is a matrix

$(g \times (p+1))$ of regression coefficients where p denotes the number of covariates; z_i is a vector $((p+1) \times 1)$ of associated covariates to the $i-th$ sample and ε_i are random errors, for $i = 1, \cdots, n$.The matrix formulation of the model

(2) is ,

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i2} \end{bmatrix} = \begin{bmatrix} \beta_{01} & \beta_{11} & \cdots & \beta_{p1} \\ \beta_{02} & \beta_{12} & \cdots & \beta_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{0g} & \beta_{1g} & \cdots & \beta_{pq} \end{bmatrix}_{g\times(p+1)} \begin{bmatrix} 1 \\ z_{i1} \\ \vdots \\ z_{ip} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{ig} \end{bmatrix}_{g\times 1}$$

In this paper, we are interested in a Bayesian analysis of model (2) with ALR transformation (1) applied to response variables and assuming a multivariate Normal distribution for the correlated and uncorrelated errors.

First of all, we assume uncorrelated errors for the model (2), i,e the error vector ε_{i1} follows a multivariate normal distribution $N_1(0, \sum 1)$, where 0 is a vector of zeros and $\sum 1$ is a variance-covariance matrix given by

$$\sum 1 = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_g^2 \end{pmatrix}$$

The likelihood function of parameters $v_1 = (\beta_{0j}, \beta_{lj}, \sigma_j^2)$ is given by

$$L(v_1) \propto \prod_{j=1}^{g} (\sigma_j^2)^{-n/2} \exp\left(-\frac{2}{2\sigma_j^2} \sum_{i=1}^{n} \epsilon_{ij}^2 \right) \quad (3)$$

where $\sum_{i=1}^{n} \epsilon_{ij}^2 = \sum_{i=1}^{n} (y_{ij} - \beta_{oj} - \beta_{ljzi})^2$, for $i = 1, \cdots, n, j = 1, \cdots, g$ and $l = 1, \cdots, p$.

An alternative approach for the analysis of compositional data is the use of Bayesian methods (see for example, Iyengar and Dey (1996); or Tjelmeland and Lund (2003)), especially considering Markov Chain Monte Carlo (MCMC) methods (see for example, Gelfand and Smith (1990)).

The Bayesian inference allows to associate previous knowledge of the parameters through a prior distribution. The Bayesian inference procedure for regression model (3) considers proper prior distributions guaranteeing proper posterior distributions. Furthermore, it was ensuring non-informative prior distributions according to the fixed hyperparameters. Thus, we assume the following prior distributions for the parameters v_1

$$\begin{aligned} \beta_{oj} &\sim N(a_{0j}, b_{oj}^2), \\ \beta_{lj} &\sim N(a_{lj}, b_{lj}^2), \\ \sigma_j^2 &\sim IG(c_j, d_j), \end{aligned} \tag{4}$$

where $\mathrm{IG(c, d)}$ denotes an Inverse-Gamma distribution with mean$\mathrm{d}/(\mathrm{c}-1)$ and variance $\mathrm{d}^2/[(\mathrm{c}-1)^2(\mathrm{c}-2)]$, for$\mathrm{c} > 2; \mathrm{a_{0j}}, \mathrm{b_{0j}}, \mathrm{a_{lj}}, \mathrm{b_{lj}}, \mathrm{c_j}$ and $\mathrm{d_j}$ are known hyperparameters,$\mathrm{j} = 1, \cdots, \mathrm{p}$.

All the parameters were assumed independent a priori.

Posterior summaries of interest for the model (3) assuming prior distributions (4) are given using simulated samples of the joint posterior distribution for $\mathrm{v_1}$ obtained using the Bayes formula, that is,

$$\pi(\beta_{0j}, \beta_{lj}, \sigma_j^2 | y) \alpha \prod_{j=1}^{g} \exp\left[-\frac{1}{2b_{0j}^2}(\beta_{0j} - a_{0j})^2\right] \times \prod_{j=1}^{g}\prod_{l=1}^{p} \exp\left[-\frac{1}{2b_{lj}^2}(\beta_{0j} - a_{0j})^2\right]$$

$$\times \prod_{j=1}^{g} (\sigma_j^2)^{-(c_j+1)} \exp(-d_j/\sigma_j^2) \times \prod_{j=1}^{g} (\sigma_j^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n} \epsilon_{ij}^2\right)$$

The full conditional densities using Gibbs sampling algorithm (Gelfand and Smith, (1990)) for each parameter are given by,

$$i)\ \pi(\beta_{0j} | \beta_{lj}, \sigma_j^2, y) \sim N\left[\frac{a_{oj} b_{oj} \sum_{i=1}^{n} \mu_i^{(j)}}{\sigma_j^2 + n b_{oj}^2}, \frac{b_{0j}^2 \sigma_j^2}{\sigma_j^2 + n b_{0j}^2}\right], \tag{5}$$

$$ii)\pi(\beta_{lj} | \beta_{0j}, \beta_{-l}, \sigma_j^2, y) \sim N\left[\frac{a_{lj}\sigma_j^2 + b_{lj}\sum_{i=1}^{n} z_{il}\theta_I^{(j)}}{\sigma_j^2 + b_{lj}^2 \sum_{i=1}^{n} z_{il}^2}, \frac{b_{Lj}^2 \sigma_j^2}{\sigma_j^2 + b_j^2 \sum_{i=1}^{n} z_{il}^2}\right] and, \tag{6}$$

$$iii)\pi(\sigma_j^2 | \beta_{0j}, \beta_{lj}, y) \sim IG\left[c_j + \frac{n}{2}, d_j + \frac{1}{2}\sum_{i=1}^{n} \epsilon_{ij}^2\right], \tag{7}$$

where $\mu_{\mathrm{i}}^{(\mathrm{j})} = \mathrm{y_{ij}} - \sum_{\mathrm{l}=1}^{\mathrm{p}} \beta_{\mathrm{lj}} \mathrm{z_{il}}, \theta_{\mathrm{i}}^{(\mathrm{j})} = \mathrm{y_{ij}} - \beta_{0\mathrm{j}}$ and $\epsilon_{\mathrm{ij}} = \mathrm{y_{ij}} - 0\mathrm{j} - \sum_{\mathrm{l}=1}^{\mathrm{p}} \beta_{\mathrm{lj}} \mathrm{z_{il}}$, for $\mathrm{i} = 1, \cdots, \mathrm{n}, \mathrm{j} = 1 \cdots, \mathrm{g}$ and $\mathrm{l} = 1, \cdots, \mathrm{p}$.

The estimation procedure considered joint estimation where all the model parameters are generated in the MCMC algorithm simultaneously. The above conditional densities (5), (6), (7) belong to known parametric density families. Posterior summaries of interest for each model are simulated through the Just Another Gibbs Sampler (JAGS) program (Plummer (2003)).

Another approach for the regression model is modeling the structure of covariance matrix. Here, we consider correlated errors for the model given in (3), i.e., ϵ_i represents the errors vector assumed to be dependent random variables with a multivariate normal distribution $N_g(0, \sum_2)$, where 0 is a vector of zeros and $\sum_2$ is a variance-covariance matrix given by

$$\sum_2 = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1g}\sigma_1\sigma_g \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2g}\sigma_2\sigma_g \\ \vdots & \vdots & \ddots & \vdots \\ \rho_g\sigma_1\sigma_g & \rho_{2g}\sigma_2\sigma_g & \cdots & \sigma_g^2 \end{pmatrix}, \tag{8}$$

where ρ_{12} is the correlation coefficient between ϵ_{i1} and ϵ_{i2} ; ρ_{1g} is the correlation coefficient between ϵ_{i1} and ϵ_{ig} and ρ_{2g} is the correlation coefficient between ϵ_{i2} and ϵ_{ig}.

Considering the assumptions above, the likelihood function of parameters $v_2 = (\beta_0, \beta_l, \sum_2)$ is given by

$$L(v_2) \propto \frac{1}{|\sum_2|^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\left(y_i - \beta_0 - \beta_{lz_i}\right)^T \middle| \sum_2^{-1}(y_i - \beta_0 - \beta_l z_i)\right]\right\}$$

for $i = 1, \cdots, n$ and $l = 1, \cdots, p$.

The following prior information is used for the Bayesian analysis,

$$\begin{aligned} \beta_o &\sim N(a_0, b_o^2 I), \\ \beta_l &\sim N(a_l, b_l^2 I), \\ \sum_2^{-1} &\sim W_g(m, M), \end{aligned} \tag{9}$$

where $W_g(m, M)$ denotes a Wishart prior distribution, m is the number of degrees of freedom and M is a prespecified precision matrix. Therefore, all the parameters were assumed independent a priori.

Posterior summaries of interest for the model defined by (3), but with correlated errors assuming priors distributions (9) are given using simulated samples of the joint posterior distribution for v_2 obtained using the Bayes formula, that is

$$\pi(\beta_0, \beta_l, \sum_2 | y) \alpha \exp\left[-\frac{1}{2}(\beta_0 - a_0)^T b_0^{-2}(\beta_0 - a_0)\right] \times \exp\left[-\frac{1}{2}(\beta_l - a_l)^T b_l^{-2}(\beta_l - a_l)\right] \times \left|\sum_2^{-1}\right|^{\frac{m-g-1}{2}} \exp\left\{-\frac{1}{2}tr(\sum_2^{-1}M^{-1})\right\} \times |\sum_2|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\left(y_i - \beta_0 - \beta_{lz_i}\right)^T \sum_2^{-1}(y_i - \beta_0 - \beta_l z_i)\right]\right\}.$$

The full conditional densities for each parameter are given by

$$i)\ \pi(\beta_0|\beta_l,\sum_2,y)\sim N\left[\left(\sum_2^{-1}\left[\sum_{i=1}^{n}y_i-\beta_l z_i\right]+\frac{I}{b_0}a_0\right)\left[n\sum_2^{-1}+\frac{I}{b_0^2}\right]^{-1},\left[n\sum_2^{-1}+\frac{I}{b_0^2}\right]^{-1}\right],$$

$$ii)\ \pi(\beta_0|\beta_{-l},\sum_2,y)\sim N\left[\left(\sum_2^{-1}\sum_{i=1}^{n}z_i\left[y_i-\beta_l\right]+\frac{a_l}{b_l}I\right)\left[\sum_{i=1}^{n}z_i^{\mathrm{T}}z_i\sum_2^{-1}+\frac{I}{b_l^2}\right]^{-1},\left[\sum_{i=1}^{n}z_i^{\mathrm{T}}z_i\sum_2^{-1}+\frac{I}{b_l^2}\right]^{-1}\right],$$

$$iii)\ \pi\left(\sum_2|\beta_0,\beta_l,y\right)\sim W_g\left[n+m,\left[\epsilon_i^{T}\epsilon_i+M^{-1}\right]^{-1}\right]$$

where $\epsilon_i=\left[y_{i1}-\beta_{01}-\sum_{l=1}^{p}\beta_{l1}z_{il},y_{i2}-\beta_{02}-\sum_{l=1}^{p}\beta_{l2}z_{il},\cdots,y_{ig}-\beta_{0g}-\sum_{l=1}^{p}\beta_{lg}z_{il},\right]$, for $i=1,\cdots,n$ and $l=1,\cdots,p$.

For the estimation procedure we consider joint estimation where all the model parameters are estimated simultaneously in the MCMC algorithm. Posterior summaries of interest for each model are simulated using standard MCMC methods through the Just Another Gibbs Sampler (JAGS) program (Plummer (2003)).

3. Application

This section reports a simulation study for the compositional data and illustrates an application of the proposed methodology through ALR transformation.

3.1 Simulation Study

A simulation study was conducted to illustrate the proposed methodology. The data was generated from multivarite normal distributions for both models (with uncorrelated and correlated errors). Assuming uncorrelated errors, the parameter values were fixed as $\beta_0=(0.5,-1,-1),\beta_1=(0.5,0.5,0.5),\beta_2=(0.5,0.5,0.5)$ and $\sigma^2=(1,1,1)$. For the case of correlated errors, we assume $\rho_{12}=\rho_{13}=\rho_{23}=0.5$. The covariates were generated by $z_1\sim$ Bernoulli (0.5) and $z_2\sim$ Normal (0.5, 0.1). We took the sample sizes $n=70,100,150$ and 300 where for each sample size we conducted 1,000 replicates. It was simulated 40,000 Gibbs samples using the rjags package (Plummer (2011)) interacting with R software (R (2011)), with a burn in of 25% of the size of the chain and we considered every 10th sample among the 30,000 Gibbs samples. Table shows the simulation results, i.e, mean, standard deviation (SD), bias, mean squared error (MSE) and credibility interval (CI). The CI was stable and close to the nominal coverage of 90%. The MSE of all the parameters decay towards zero as the sample size increases.

Table 1: Simulation Data. Summary of the posterior distributions for the models parameters assuming uncorrelated and correlated errors.

Sample Size	Uncorrelated Errors						Correlated Errors					
	Parameter	Mean	SD	Bias	MSE	CI	Parameter	Mean	SD	Bias	MSE	CI
n=70	β_{01}	0.5103	0.1793	0.0103	0.0322	0.900	β_{01}	0.5100	0.1794	0.0100	0.0322	0.898
	β_{02}	-1.0028	0.1725	-0.0028	0.0297	0.924	β_{02}	-0.9972	0.1733	0.0028	0.0300	0.908
	β_{03}	-0.9943	0.1793	0.0057	0.0321	0.899	β_{03}	-0.9907	0.1796	0.0093	0.0323	0.902
	β_{11}	0.4908	0.2287	-0.0092	0.0524	0.925	β_{11}	0.4911	0.2288	-0.0089	0.0524	0.925
	β_{12}	0.5063	0.2322	0.0063	0.0539	0.912	β_{12}	0.5008	0.2272	0.0008	0.0515	0.926
	β_{13}	0.4906	0.2352	-0.0094	0.0554	0.914	β_{13}	0.4894	0.2346	-0.0106	0.0551	0.906
	β_{21}	0.5025	0.1240	0.0025	0.0154	0.891	β_{21}	0.5026	0.1241	0.0026	0.0154	0.890
	β_{22}	0.5012	0.1300	0.0012	0.0169	0.884	β_{22}	0.5023	0.1275	0.0023	0.0163	0.890
	β_{23}	0.5053	0.1224	0.0053	0.0150	0.891	β_{23}	0.5058	0.1244	0.0058	0.0155	0.896
	σ_1	1.0232	0.1675	0.0232	0.0286	0.913	σ_1	1.0235	0.1674	0.0235	0.0286	0.916
	σ_2	1.0255	0.1775	0.0255	0.0321	0.899	σ_2	1.0256	0.1809	0.0256	0.0333	0.898
	σ_3	1.0167	0.1735	0.0167	0.0303	0.899	σ_3	1.0134	0.1725	0.0134	0.0299	0.911
							ρ_{12}	0.4947	0.0892	-0.0053	0.0080	0.912
							ρ_{13}	0.4941	0.0898	-0.0059	0.0081	0.906
							ρ_{23}	0.4931	0.0875	-0.0069	0.0077	0.912
n=100	β_{01}	0.4972	0.1476	-0.0028	0.0218	0.903	β_{01}	0.4967	0.1475	-0.0033	0.0218	0.905
	β_{02}	-1.0048	0.1527	-0.0048	0.0233	0.898	β_{02}	-1.0056	0.1545	-0.0056	0.0239	0.898
	β_{03}	-1.0120	0.1480	-0.0120	0.0220	0.909	β_{03}	-1.0123	0.1523	-0.0123	0.0233	0.895
	β_{11}	0.4981	0.2043	-0.0019	0.0417	0.898	β_{11}	0.4985	0.2042	-0.0015	0.0417	0.901
	β_{12}	0.5060	0.1996	0.0060	0.0399	0.900	β_{12}	0.5042	0.1964	0.0042	0.0385	0.915
	β_{13}	0.5048	0.1976	0.0048	0.0390	0.905	β_{13}	0.5045	0.1960	0.0045	0.0384	0.913
	β_{21}	0.5024	0.0989	0.0024	0.0098	0.912	β_{21}	0.5024	0.0989	0.0024	0.0098	0.910
	β_{22}	0.5020	0.1021	0.0020	0.0104	0.898	β_{22}	0.5029	0.1022	0.0029	0.0105	0.908
	β_{23}	0.5100	0.1034	0.0100	0.0108	0.883	β_{23}	0.5099	0.1042	0.0099	0.0109	0.898
	σ_1	1.0174	0.1500	0.0174	0.0228	0.887	σ_1	1.0178	0.1499	0.0178	0.0228	0.887
	σ_2	1.0265	0.1403	0.0265	0.0204	0.904	σ_2	1.0291	0.1407	0.0291	0.0206	0.910
	σ_3	1.0191	0.1426	0.0191	0.0207	0.913	σ_3	1.0246	0.1492	0.0246	0.0228	0.899
							ρ_{12}	0.4986	0.0744	-0.0014	0.0055	0.903
							ρ_{13}	0.4996	0.0746	-0.0004	0.0056	0.897
							ρ_{23}	0.4982	0.0782	-0.0018	0.0061	0.892

Sample Size	Uncorrelated Errors						Correlated Errors					
	Parameter	Mean	SD	Bias	MSE	CI	Parameter	Mean	SD	Bias	MSE	CI
n=150	β_{01}	0.4984	0.1183	-0.0016	0.0140	0.911	β_{01}	0.4980	0.1183	-0.0020	0.0140	0.910
	β_{02}	-1.0012	0.1183	-0.0012	0.0140	0.921	β_{02}	-1.0018	0.1191	-0.0018	0.0142	0.912
	β_{03}	-1.0075	0.1200	-0.0075	0.0144	0.903	β_{03}	-1.0070	0.1184	-0.0070	0.0141	0.918
	β_{11}	0.5016	0.1591	0.0016	0.0253	0.905	β_{11}	0.5019	0.1591	0.0019	0.0253	0.904
	β_{12}	0.5043	0.1550	0.0043	0.0240	0.908	β_{12}	0.5045	0.1539	0.0045	0.0237	0.918
	β_{13}	0.5127	0.1608	0.0127	0.0260	0.900	β_{13}	0.5122	0.1578	0.0122	0.0250	0.909
	β_{21}	0.4986	0.0819	-0.0014	0.0067	0.907	β_{21}	0.4987	0.0819	-0.0013	0.0067	0.903
	β_{22}	0.4983	0.0830	-0.0017	0.0069	0.891	β_{22}	0.4978	0.0819	-0.0022	0.0067	0.891
	β_{23}	0.5019	0.0803	0.0019	0.0064	0.914	β_{23}	0.5003	0.0813	0.0003	0.0066	0.910
	σ_1	1.0121	0.1133	0.0121	0.0130	0.906	σ_1	1.0125	0.1132	0.0125	0.0130	0.906
	σ_2	1.0063	0.1166	0.0063	0.0136	0.911	σ_2	1.0060	0.1143	0.0060	0.0131	0.910
	σ_3	1.0112	0.1139	0.0112	0.0131	0.905	σ_3	1.0115	0.1171	0.0115	0.0138	0.897
							ρ_{12}	0.4974	0.0622	-0.0026	0.0039	0.891
							ρ_{13}	0.4986	0.0592	-0.0014	0.0035	0.908
							ρ_{23}	0.4963	0.0599	-0.0037	0.0036	0.905
n=300	β_{01}	0.4974	0.0868	-0.0026	0.0075	0.892	β_{01}	0.4970	0.0869	-0.0030	0.0075	0.896
	β_{02}	-0.9975	0.0861	0.0025	0.0074	0.903	β_{02}	-0.9990	0.0860	0.0010	0.0074	0.905
	β_{03}	-1.0000	0.0848	0.0001	0.0072	0.903	β_{03}	-1.0003	0.0845	-0.0003	0.0071	0.910
	β_{11}	0.5011	0.1164	0.0011	0.0135	0.904	β_{11}	0.5014	0.1164	0.0014	0.0135	0.905
	β_{12}	0.4987	0.1094	-0.0013	0.0120	0.926	β_{12}	0.4992	0.1135	-0.0008	0.0129	0.897
	β_{13}	0.4975	0.1128	-0.0025	0.0127	0.907	β_{13}	0.4979	0.1153	-0.0021	0.0133	0.894
	β_{21}	0.4986	0.0583	-0.0014	0.0034	0.896	β_{21}	0.4986	0.0583	-0.0014	0.0034	0.896
	β_{22}	0.5010	0.0579	0.0010	0.0033	0.900	β_{22}	0.5001	0.0576	0.0001	0.0033	0.899
	β_{23}	0.5029	0.0577	0.0029	0.0033	0.894	β_{23}	0.5019	0.0580	0.0019	0.0034	0.905
	σ_1	1.0059	0.0822	0.0059	0.0068	0.901	σ_1	1.0063	0.0822	0.0063	0.0068	0.902
	σ_2	1.0018	0.0815	0.0018	0.0066	0.893	σ_2	1.0024	0.0847	0.0024	0.0072	0.883
	σ_3	1.0048	0.0842	0.0048	0.0071	0.898	σ_3	1.0048	0.0811	0.0048	0.0066	0.903
							ρ_{12}	0.4990	0.0436	-0.0010	0.0019	0.903
							ρ_{13}	0.4982	0.0427	-0.0018	0.0018	0.897
							ρ_{23}	0.4992	0.0447	-0.0008	0.0020	0.886

Table 2 shows the Bayesian criteria for the model assuming uncorrelated and correlated errors. The model assuming correlated errors is better when compared to the other model in all considered criteria.

Table 2: Simulation Data. Bayesian Criteria.

Sample Size	Model	Bayesian criteria		
		EAIC	EBIC	DIC
n=70	Uncorrelated errors	617.788	644.770	609.041
	Correlated errors	572.298	609.026	562.107
n=100	Uncorrelated errors	875.026	906.288	866.172
	Correlated errors	811.754	850.831	799.998
n=150	Uncorrelated errors	1299.335	1335.462	1290.421
	Correlated errors	1201.428	1246.588	1193.225
n=300	Uncorrelated errors	2575.710	2620.155	2566.762
	Correlated errors	2373.897	2429.453	2384.760

In order to verify the behavior of the MCMC implementation we provide some plots in the Appendix 1 (Additional Matter).

3.2 Real Data Application

In this section, we consider a Bayesian analysis of the real data set presented in the website Brazilian Volleyball Confederation (CBV) (2012) to illustrate an application of the proposed methodology, in particular, data related to proportions of the points volleyball teams. We apply the compositional data methodology to such set considering as components the proportions of the winning team points in 128 games of Brazilian Men's Volleyball Super League 2011/2012. This study was based on the four components: proportion of points in the attack(x_1), pro- portion of points in the block (x_2), proportion of points in the serve (x_3) and proportion of points in the errors of the opposite team (x_4).

On the other hand, it was considered five independent variables (covariates): player who scored more points in the game belongs to the winning team (z_1), the winning team has won League at least once in the last twelve years (z_2), percentage of excellent reception of the winning team in the game (z_3) and percentage of excellent defense of the loser team in the game (z_4).

We assume an additive log-ratio (ALR) transformation given by $y_{i1} = \log(x_{i1}/x_{i4}), y_{i2} = \log(x_{i2}/x_{i4})$ and $y_{i3} = \log(x_{i3}/x_{i4})$.

The likelihood function for the models with uncorrelated and correlated errors are given by $L(v_1)$ and $L(v_2)$, respectively

$$\mathrm{L}(v_1) \propto \prod_{j=1}^{3}\left(\sigma_j^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n}\varepsilon_{ij}^2\right), \tag{10}$$

and

$$L(v_2) \propto |\sum_2|^{-\frac{n}{2}} exp\left\{-\frac{1}{2R}\left[\frac{(1-\rho_{23}^2)}{\sigma_1^2}\sum_{i=1}^{n}\varepsilon_{i1}^2 + \frac{(1-\rho_{13}^2)}{\sigma_2^2} + \sum_{i=1}^{n}\varepsilon_{i2}^2 + \frac{(1-\rho_{12}^2)}{\sigma_3^2}\sum_{i=1}^{n}\epsilon_{i3}^2\right]\right\}$$

$$\times \exp\left\{-\frac{1}{2R}\left[2R_{12}\sum_{i=1}^{n}\epsilon_{i1}\epsilon_{i2} + 2R_{13}\sum_{i=1}^{n}\epsilon_{i1}\epsilon_{i3} + 2R_{23}\sum_{i=1}^{n}\epsilon_{i2}\epsilon_{i3}\right]\right\}, \tag{11}$$

where $R = 2\rho_{12}\rho_{13}\rho_{23} - (\rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2)$, $R_{12} = \frac{(\rho_{13}\rho_{23}-\rho_{12})}{\sigma_1\sigma_2}$, $R_{13} = \frac{(\rho_{12}\rho_{23}-\rho_{13})}{\sigma_1\sigma_3}$, $R_{23} = \frac{(\rho_{12}\rho_{13}-\rho_{23})}{\sigma_2\sigma_3}$ and $\sum_{i=1}^{n}\varepsilon_{ij}^2 = \sum_{i=1}^{n}\left(y_{ij} - \beta_{0j} - \beta_{1j}z_{i1} - \beta_{2j}z_{i2} - \beta_{3j}z_{i3} - \beta_{4j}z_{i4}\right)^2$, for $j = 1,2,3\ and\ i = 1,\cdots,128$.

The proposed model in (10) and the following independent non-informative prior distributions (4) were considered: $\beta_{0j} \sim \mathrm{N}(0, 10^3)$, $\beta_{lj} \sim \mathrm{N}(0, 10^3)$, $\sigma_j^2 \sim IG(0.1,100)$, where $l = 1,2,3,4\ and\ j = 1,2,3$. For proposed regression model with correlated errors (11), we considered $\sum_2^{-1} \sim W_3(3, R)$ (where R is) and the same independent proper prior distributions for c, β_{lj}, for $l = 1,2,3,4\ and\ j = 1,2,3$. It was simulated 100,000 Gibbs samples using the rjags package (Plummer (2011)) interacting with R software (R (2011)), in which the first 10,000 simulated samples were discarded to eliminate the effects of the initial values and we considered every 20th sample among the 90,000 Gibbs samples. The convergence was verified through Gelman-Rubin diagnostic. It shows values very close to 1 indicating convergence of the simulation algorithm.

According to Carlin and Louis (2009), the most basic tool for investigating model uncertainty is the sensitivity analysis, that is, making reasonable modifications to the assumption, recomputing the posterior quantities of interest and seeing if they have changed in a way that has practical impact on interpretations. Thus, we checked the sensitivity analysis for different choices of prior parameters (β_{0j}, β_{lj}, and σ_j^2, for $l = 1,2,3,4$) on the mean components by changing only on parameter at a time and keeping all other parameters constant to their default values. We observe that posterior summaries of the parameters do not present considerable difference and not affect the results.

Table 3 shows the posterior summaries for the parameters of the model (10) assuming uncorrelated and correlated errors. The convergence was verified through Gelman-Rubin diagnostic. It showed values very close to 1 indicating convergence of the simulation algorithm. Note that there is significant difference regarding to the proportions attack, block and serve points indicating by the estimated β_{11}, β_{31}, β_{42} and β_{43} for both models (uncorrelated and correlated errors), i.e., the player who scored in the game belongs to the winning team, percentage of excellent reception of the winning team and percentage of excellent defense of the loser team help it in these skills. Moreover, the estimated posterior means and standard deviations present similarity values for the both models (uncorrelated and correlated errors). We also observe that more parameters were significant in the correlated model than uncorrelated model, i.e., β_{12}, β_{13}, β_{21}, β_{22}, β_{32} and β_{41}.

Table 3: Summary of the posterior distributions for the models parameters assuming uncorrelated and correlated errors.

Uncorrelated Errors				Correlated Errors			
Parameter	Mean	Standard Deviation	Credibility Interval (90%)	Parameter	Mean	Standard Deviation	Credibility Interval (90%)
β_{01}	0.5471	0.2047	(0.2104; 0.8850)	β_{01}	0.5500	0.2036	(0.2118; 0.8873)
β_{02}	-1.9755	0.3496	(-2.5542; -1.4051)	β_{02}	-1.9696	0.3453	(-2.5384; -1.3936)
β_{03}	-0.9446	0.4724	(-1.7182; -0.1713)	β_{03}	-0.9415	0.4771	(-1.7364; -0.1609)
β_{11}	0.1741	0.0469	(0.0972; 0.2515)	β_{11}	0.1730	0.0474	(0.0957; 0.2508)
β_{12}	0.1444	0.0802	(0.0132; 0.2767)	β_{12}	0.1415	0.0805	(0.0102; 0.2746)
β_{13}	0.1905	0.1097	(0.0080; 0.3702)	β_{13}	0.1916	0.1098	(0.0096; 0.3703)
β_{21}	-0.0714	0.0450	(-0.1453; 0.0032)	β_{21}	-0.0720	0.0457	(-0.1464; 0.0039)
β_{22}	-0.1536	0.0764	(-0.2806; -0.0289)	β_{22}	-0.1532	0.0760	(-0.2786; -0.0297)
β_{23}	-0.0275	0.1047	(-0.1996; 0.1436)	β_{23}	-0.0273	0.1066	(-0.2029; 0.1473)
β_{31}	0.4285	0.1876	(0.1175; 0.7336)	β_{31}	0.4297	0.1872	(0.1233; 0.7381)
β_{32}	0.5257	0.3171	(0.0019; 1.0472)	β_{32}	0.5211	0.3185	(0.0001; 1.0465)
β_{33}	-0.2634	0.4346	(-0.9775; 0.4523)	β_{33}	-0.2656	0.4405	(-0.9912; 0.4606)
β_{41}	-0.5392	0.3034	(-1.0360; -0.0385)	β_{41}	-0.5440	0.3007	(-1.0380; -0.0436)
β_{42}	1.2414	0.5134	(0.3993; 2.0931)	β_{42}	1.2391	0.5072	(0.4001; 2.0750)
β_{43}	-1.9115	0.6980	(-3.0575; -0.7715)	β_{43}	-1.9177	0.6959	(-3.0666; -0.7668)
σ_1	0.0594	0.0078	(0.0478; 0.0731)	σ_1	0.0594	0.0078	(0.0478; 0.0731)
σ_2	0.1701	0.0222	(0.1372; 0.2089)	σ_2	0.1706	0.0219	(0.1378; 0.2093)
σ_3	0.3231	0.0425	(0.2606; 0.3986)	σ_3	0.3233	0.0420	(0.2611; 0.3992)
				ρ_{12}	0.3596	0.0789	(0.2262; 0.4845)
				ρ_{13}	0.2373	0.0851	(0.0939; 0.3746)
				ρ_{23}	0.1653	0.0878	(0.0169; 0.3068)

Table 4 presents the Bayesian model selection criteria expected Akaike information criterion (EAIC), expected Bayesian information criterion (EBIC) and deviance information

criterion (DIC). These results are suggesting that fitted regression model assuming correlated errors is the best choice (lower values EAIC, EBIC and DIC).

Table 4: Bayesian Criteria for the models parameters assuming uncorrelated and correlated errors.

Model	Bayesian Criteria		
	EAIC	EBIC	DIC
Uncorrelated errors	343.621	394.674	334.112
Correlated errors	342.964	394.017	333.469

Some plots to examine the behaviour of the chains of MCMC implementation are avaliable in the Appendix 2 (Additional Matter).

4. Concluding Remarks

In this paper, we present a Bayesian analysis for compositional regression model considering ALR transformation and assuming uncorrelated and correlated errors. The inferencial procedure for the parameters based on MCMC methods. The Bayesian approach has some advantages over other inference methods. We have that it allows to incorporate prior information about the parameter, it pro- vides results without reliance on asymptotic approximation, the great number of covariates and missing data are easily handed in the Bayesian framework.

Since studies about volleyball data do not consider the compositional data strucuture for the fundaments, here we applied the proposed methodology in order to verify, in the context of regression models, the relationship between fundaments of volleyball (without discarding the multivariate structure) and covariates observed on the volleyball games.

We analysed a real data set from percentages of winning volleyball team's points, in which it was considered multivariate data structure. A comparison study of models was carried out through model selection procedures based on a statistical criteria, i.e, the complete covariance matrix was estimated to evaluate the importance of correlations among the fundaments. Thus, the results indicate that the compositional regression model with correlated errors outperforms the model with uncorrelated errors, besides pointing out the advantage of considering the natural multivariate structure of the data.

References

[1] Achcar, J.A.; Obage, S.C. (2005). Uma abordagem Bayesiana para dados composicionais considerando erros correlacionados. Revista de MatemÃątica e EstatÃŋstica, 23, 95–107.

[2] Afonso, J.; Esteves, F.; Araujo, R.; Thomas, L.; Mesquita, I. (2012). Tactical determinants of setting zone in elite men's volleyball. Journal of Sports Science and Medicine, 11, 64–70.

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

[4] Aitchison, J. (1986). The statistical analysis of compositional data. Chapman & Hall.

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

[6] Brazilian Volleyball Confederation (CBV). (2012). Data set of Men's Volleyball Super League. Avaliable at:http://www.cbv.com.br/v1/superliga-1112/m-tabela.asp. Accessed August 23, 2012.

[7] Campos, F. A. D.; StanganÃľlli, L. C. R.; Campos, L. C. B.; Pasquarelli, B. N.; GÃşmez, M. A. (2014). Performance indicators analysis at Brazilian and Italian women's volleyball leagues according to game location, game outcome, and set number. Perceptual and motor skills, 118, 347–361.

[8] Carlin, B.P.; Louis, T.A. (2009). Bayesian Methods for Data Analysis. Chap- man & Hall/CRC Texts in Statistical Science, 3th. Edition.

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

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

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

[12] Johnson, R.; Wichern, D. (1998). Applied multivariate statistical analysis. New Jersey: Prentice Hall.

[13] Plummer, M. (2003). JAGS: A program for analysis Bayesian graphical mod- els using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC, 2003). Vol. 124. Technische Universit at Wien.

[14] Plummer, M. (2011). rjags: Bayesian graphical models using MCMC. r package version 3-3.

[15] R Development Core Team. (2011). R: A Language and Environment for Statis- tical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[16] Rodriguez-Ruiz, D.; Quiroga, M. E.; Miralles, J. A.; Sarmiento, S.; SaÃą, Y.; GarcÃŋa-Manso, J. M. (2011). Study of the technical and tactical variables determining set win or loss in top-level European men's volleyball. Journal of Quantitative Analysis in Sports, 7.

[17] Tjelmeland, H.; Lund, K.V. (2003). Bayesian modelling of spatial compositional data. Journal of Applied Statistics, 30, 87–100.

[18] Tsagris, M. (2014). The k-NN algorithm for compositional data: a revised approach with and without zero values present. Journal of Data Science, 12, 519–534

Received ; accepted .

[1] Taciana Kisaki Oliveira Shimizu Inter Program of Graduate Statistics
Federal University of S˜ao Carlos and University of S˜ao Paulo

[2] Francisco Louzada
Department of Applied Mathematics & Statistics, Sciences Institute of Mathematics and Com- puters, ICMC
University of S˜ao Paulo
Avenida Trabalhador S˜ao-carlense, 400, S˜ao Carlos, S˜ao Paulo, Brazil

[3] Adriano Kamimura Suzuki
Department of Applied Mathematics & Statistics, Sciences Institute of Mathematics and Com- puters, ICMC
University of S˜ao Paulo
Avenida Trabalhador S˜ao-carlense, 400, S˜ao Carlos, S˜ao Paulo, Brazil

[4] Ricardo Sandes Ehlers
Department of Applied Mathematics & Statistics, Sciences Institute of Mathematics and Com- puters, ICMC
University of S˜ao Paulo
Avenida Trabalhador S˜ao-carlense, 400, S˜ao Carlos, S˜ao Paulo, Brazil

Additional Matter of the paper: Modeling Compositional Regression with Uncorrelated and Correlated Errors

Shimizu, T.K.O., Louzada, F., Suzuki, A.K. and Ehlers, R.S.

Here we provide some plots about the MCMC implementation in both the simulation study (Section 3.1) and the application (Section 3.2) to volleyball data.

3.1 Appendix 1：Plots of MCMC implementation - Section

In this appendix we present a graphic visualization about the implementation in the simulation study for models assuming uncorrelated and correlated errors. The plots were about the last sample generated with sample size of $n = 70$.

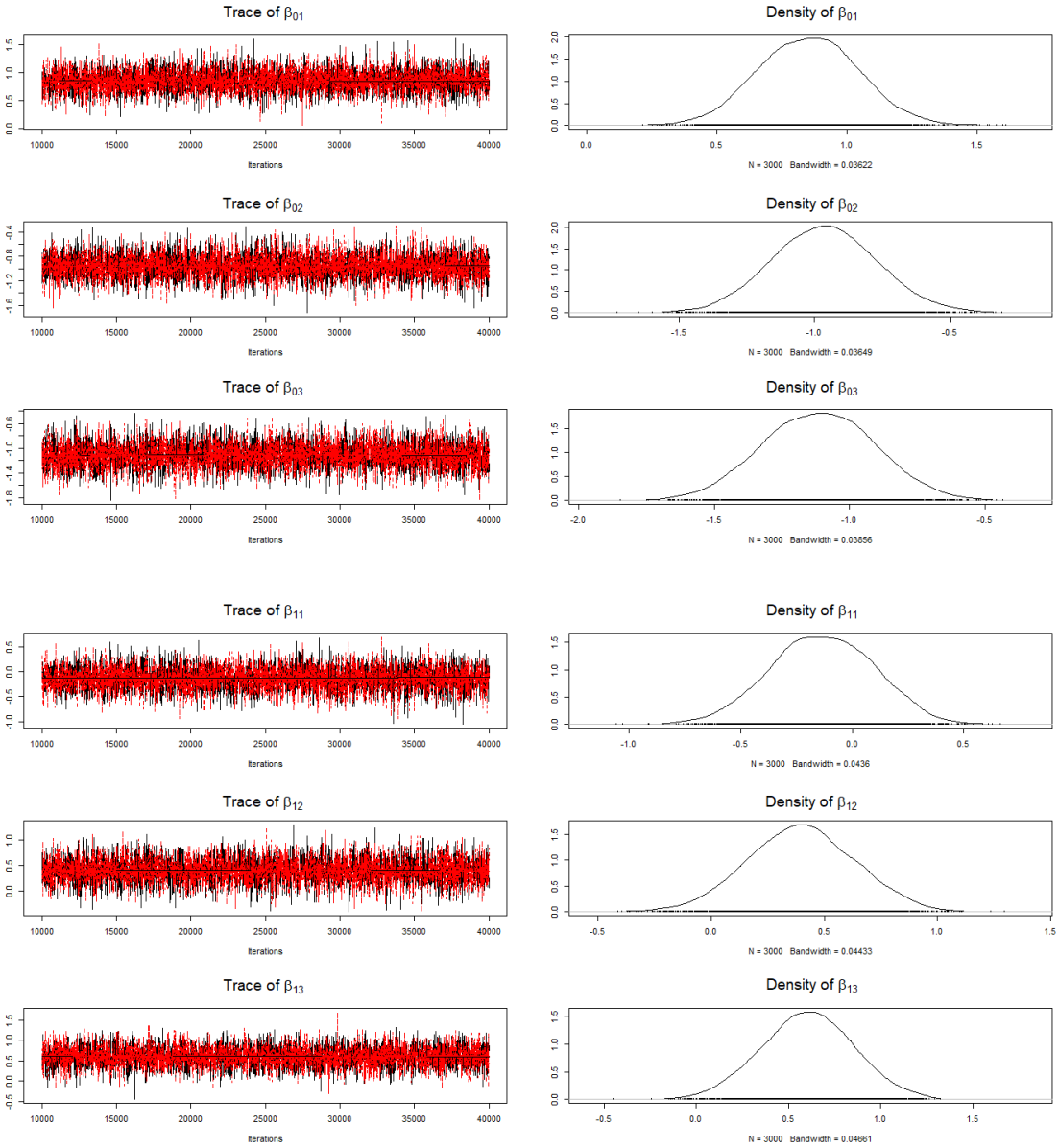

Figure 1 : Trace plots and density for posterior distribution of parameters (model with uncorrelated errors - Section 3.1).

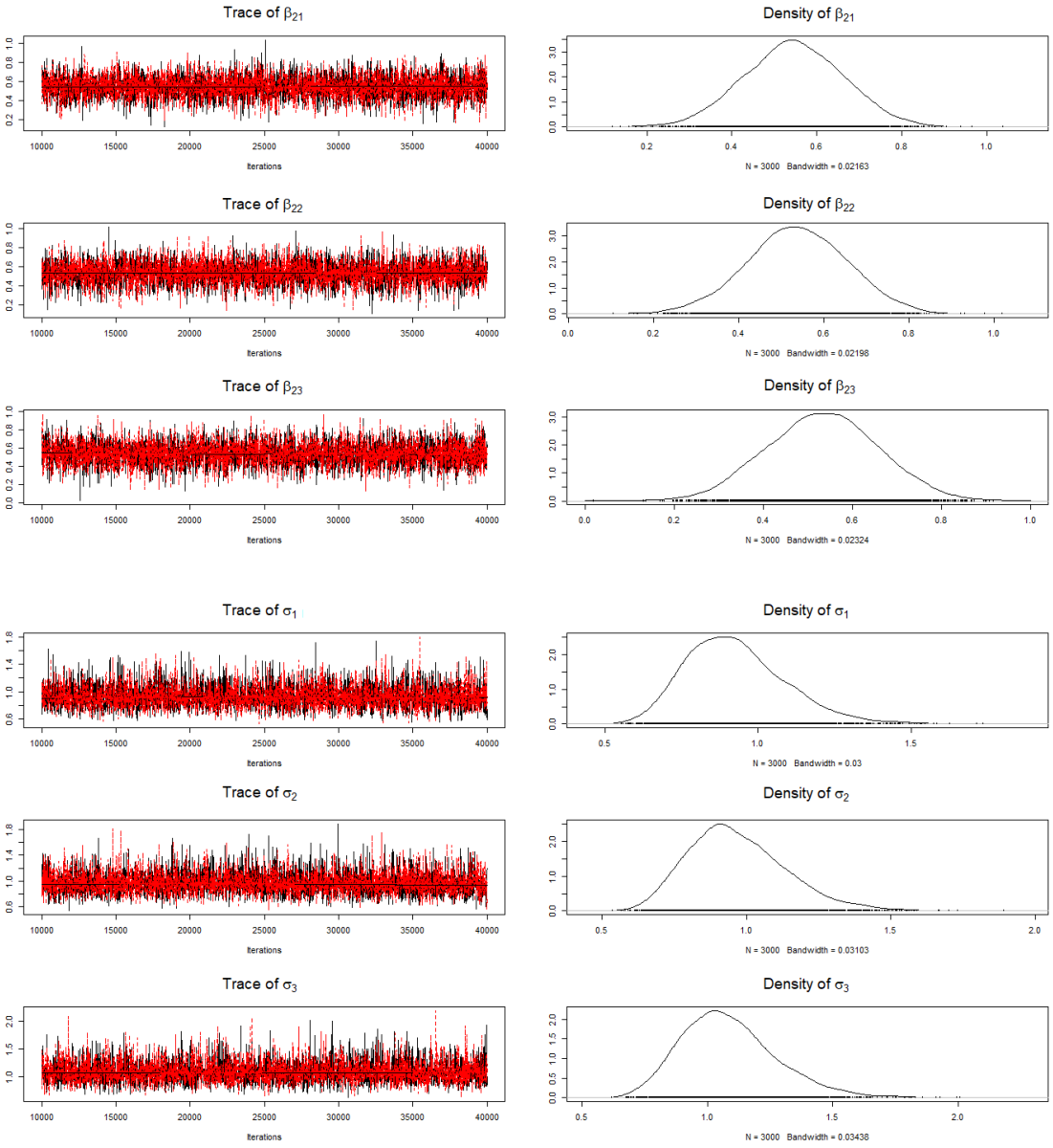

Figure 2 : Trace plots and density for posterior distribution of parameters (model with uncorrelated errors - Section 3.1).

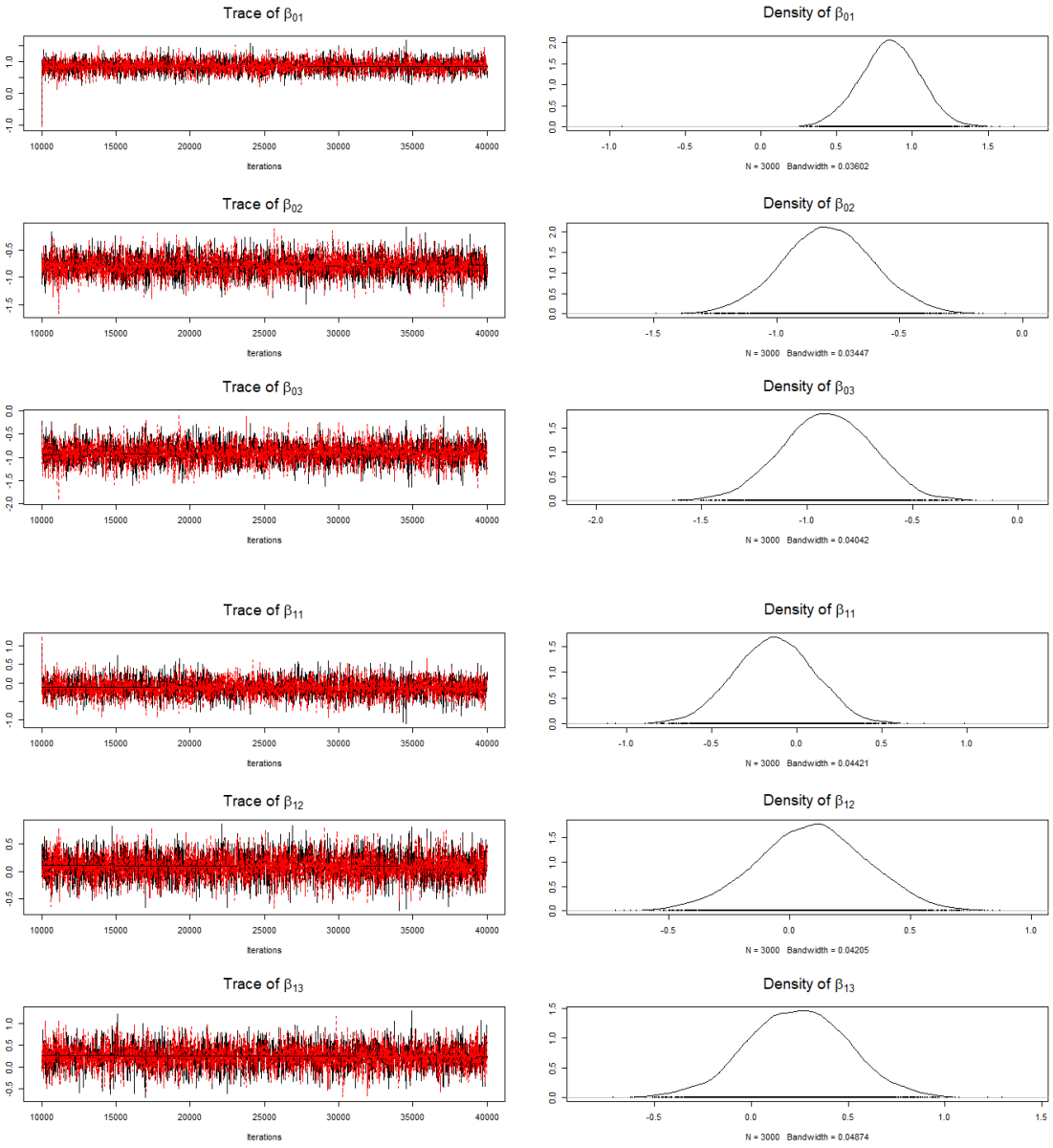

Figure 3：Trace plots and density for posterior distribution of parameters (model with correlated errors - Section 3.1).

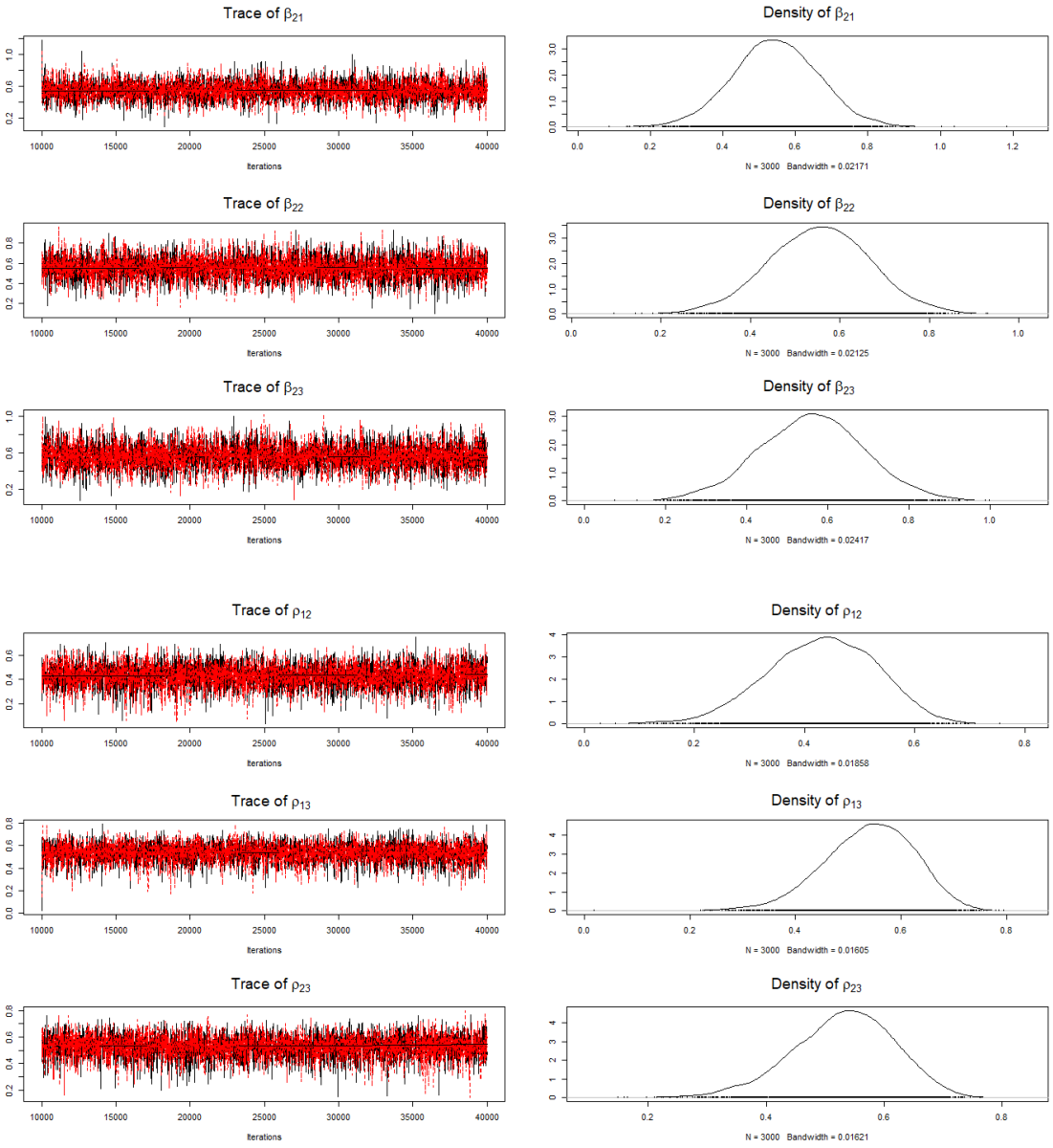

Figure 4：Trace plots and density for posterior distribution of parameters (model with correlated errors - Section 3.1).

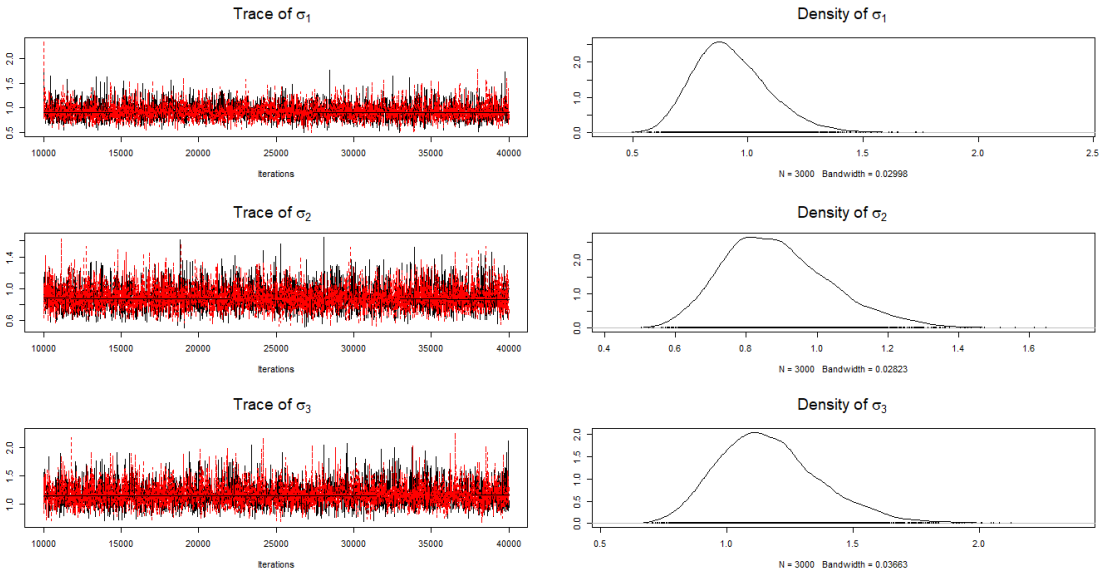

Figure 5 : Trace plots and density for posterior distribution of parameters (model with correlated errors - Section 3.1).

Figures 1, 2, 3, 4 and 5 show the behaviour of MCMC implementation for the parameters of model with uncorrelated and correlated errors for simulation studies. We observe that the chains converged for all the parameters (see trace plots). Also, the convergence was monitored through by Gelman-Rubin diagnostic, being that the values for all the parameters were around 1.

3.2 Appendix 2：Plots of MCMC implementation - Section

In this appendix we present a graphic visualization about the implementation in the application of volleyball data (Section 3.2) for models assuming uncorrelated and correlated errors.

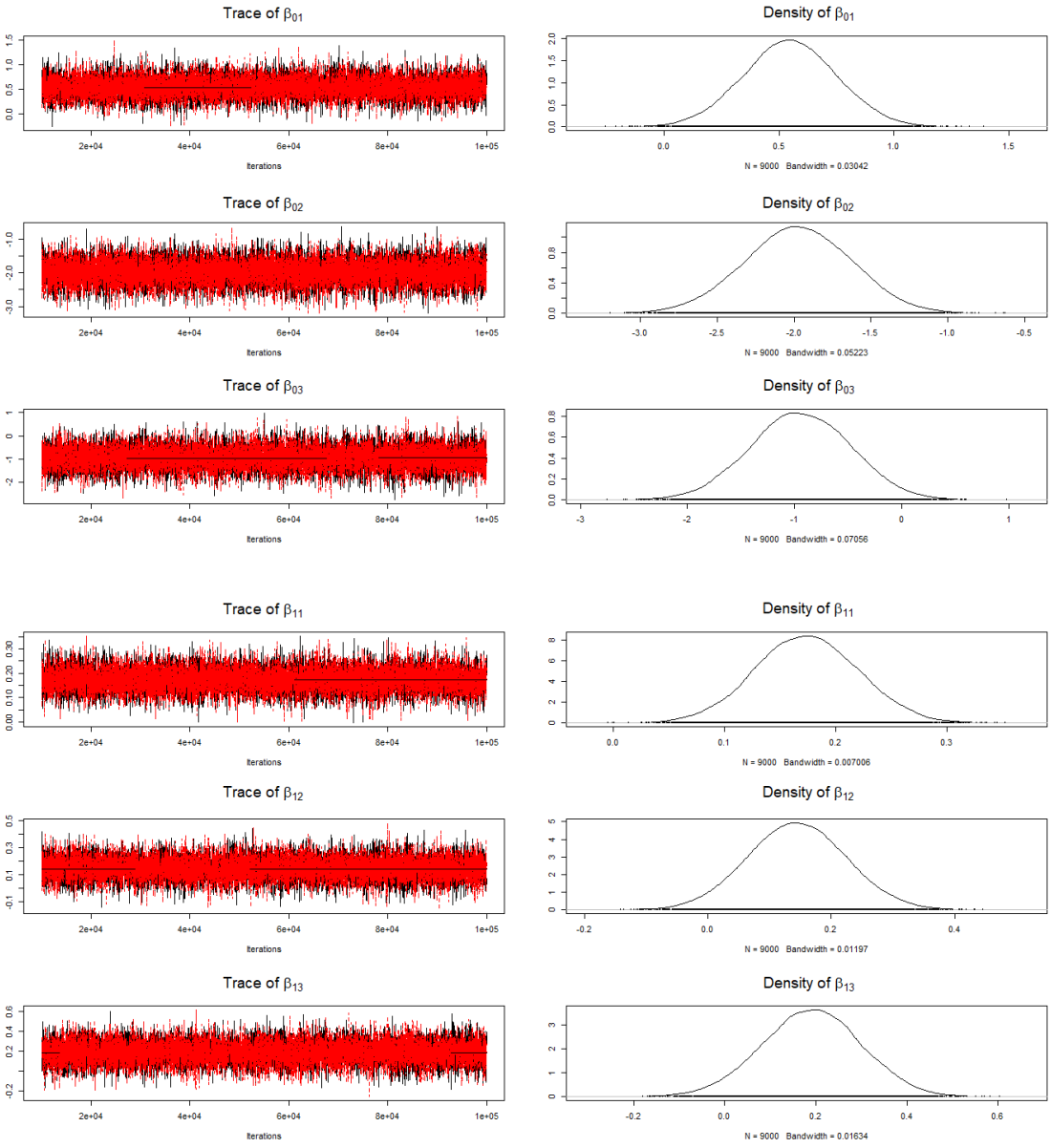

Figure 6：Trace plots and density for posterior distribution of parameters (model with uncorrelated errors - Section 3.2).

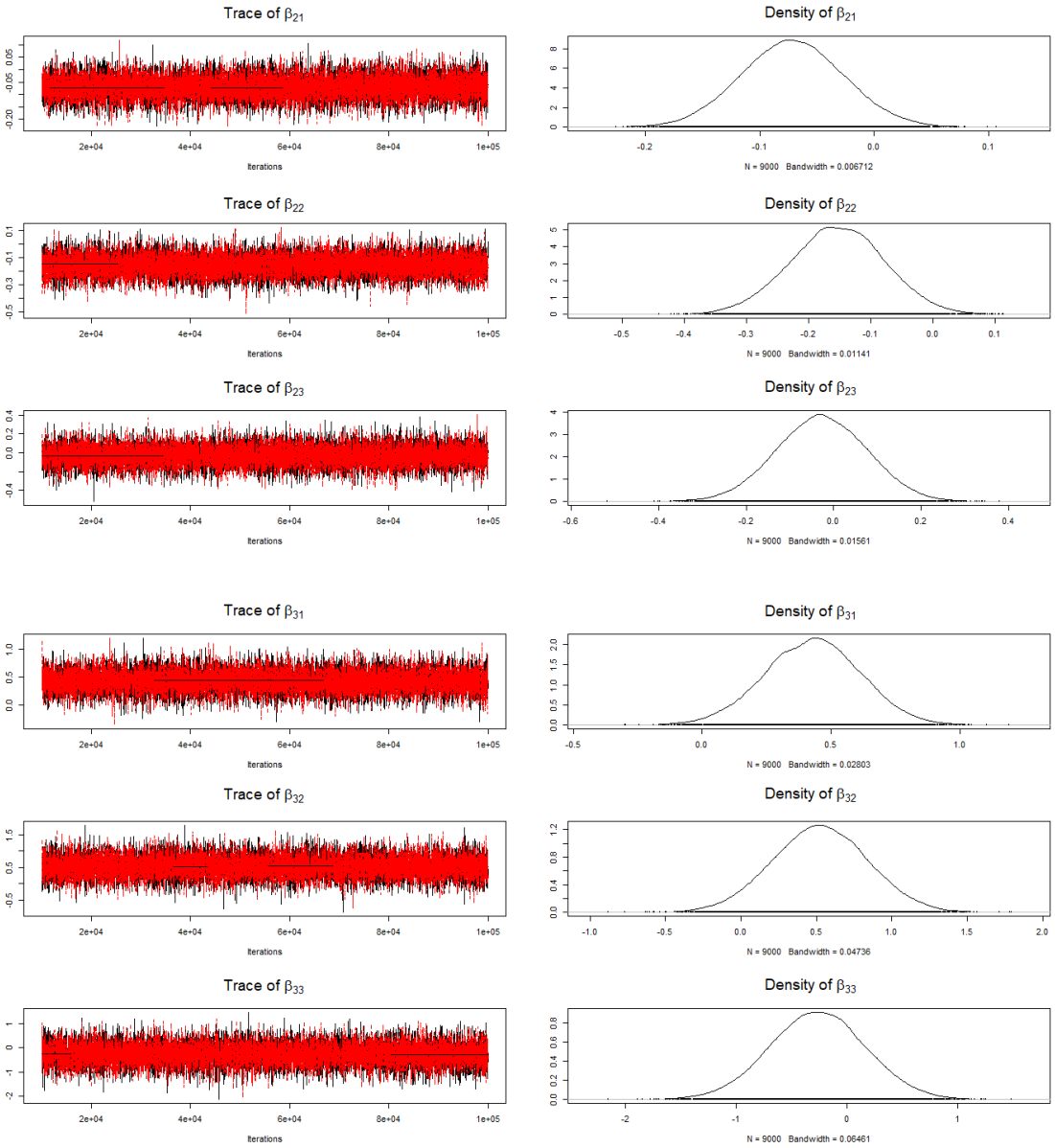

Figure 7：Trace plots and density for posterior distribution of parameters (model with uncorrelated errors - Section 3.2).

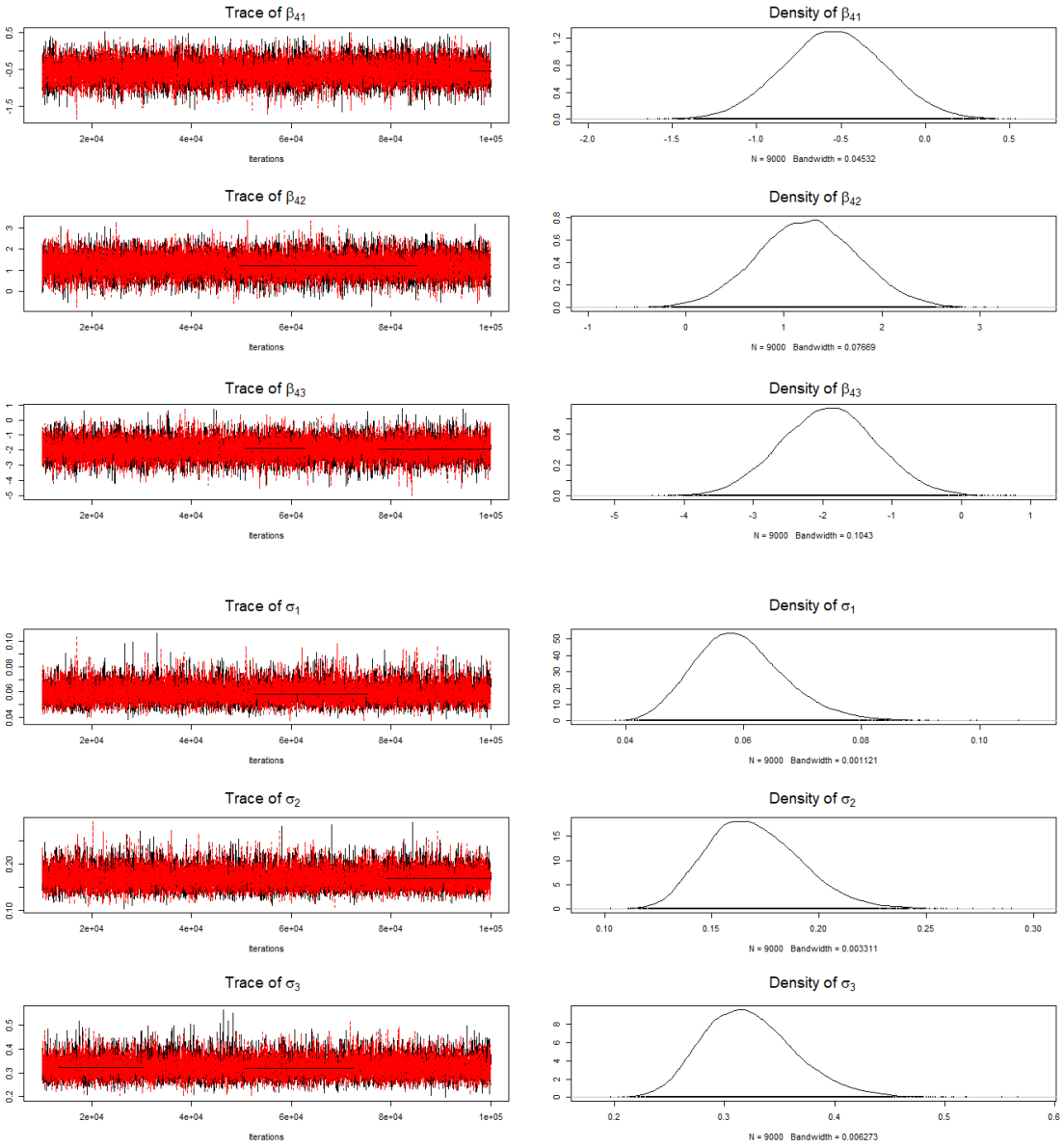

Figure 8：Trace plots and density for posterior distribution of parameters (model with uncorrelated errors - Section 3.2).

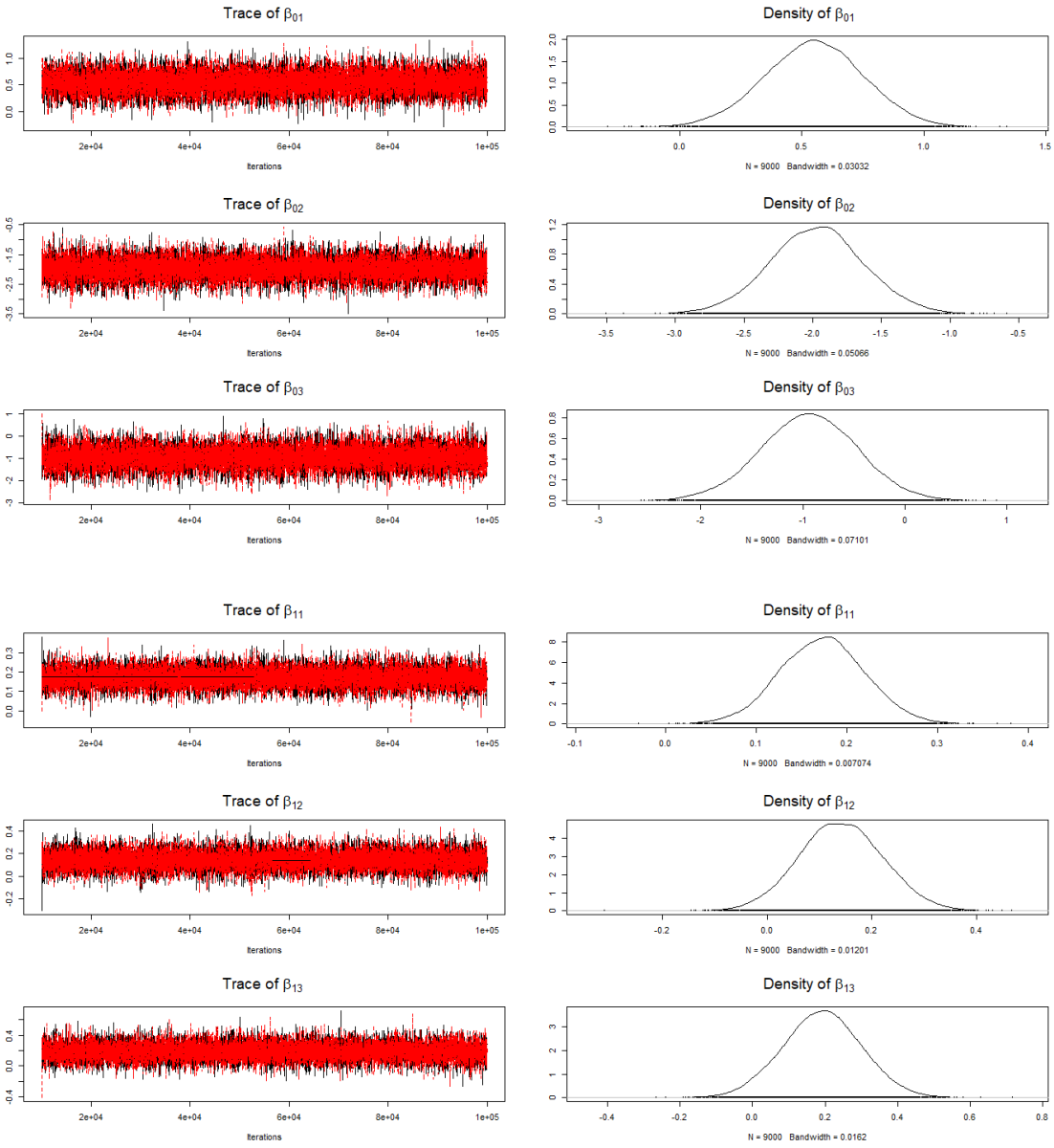

Figure 9：Trace plots and density for posterior distribution of parameters (model with correlated errors - Section 3.2).

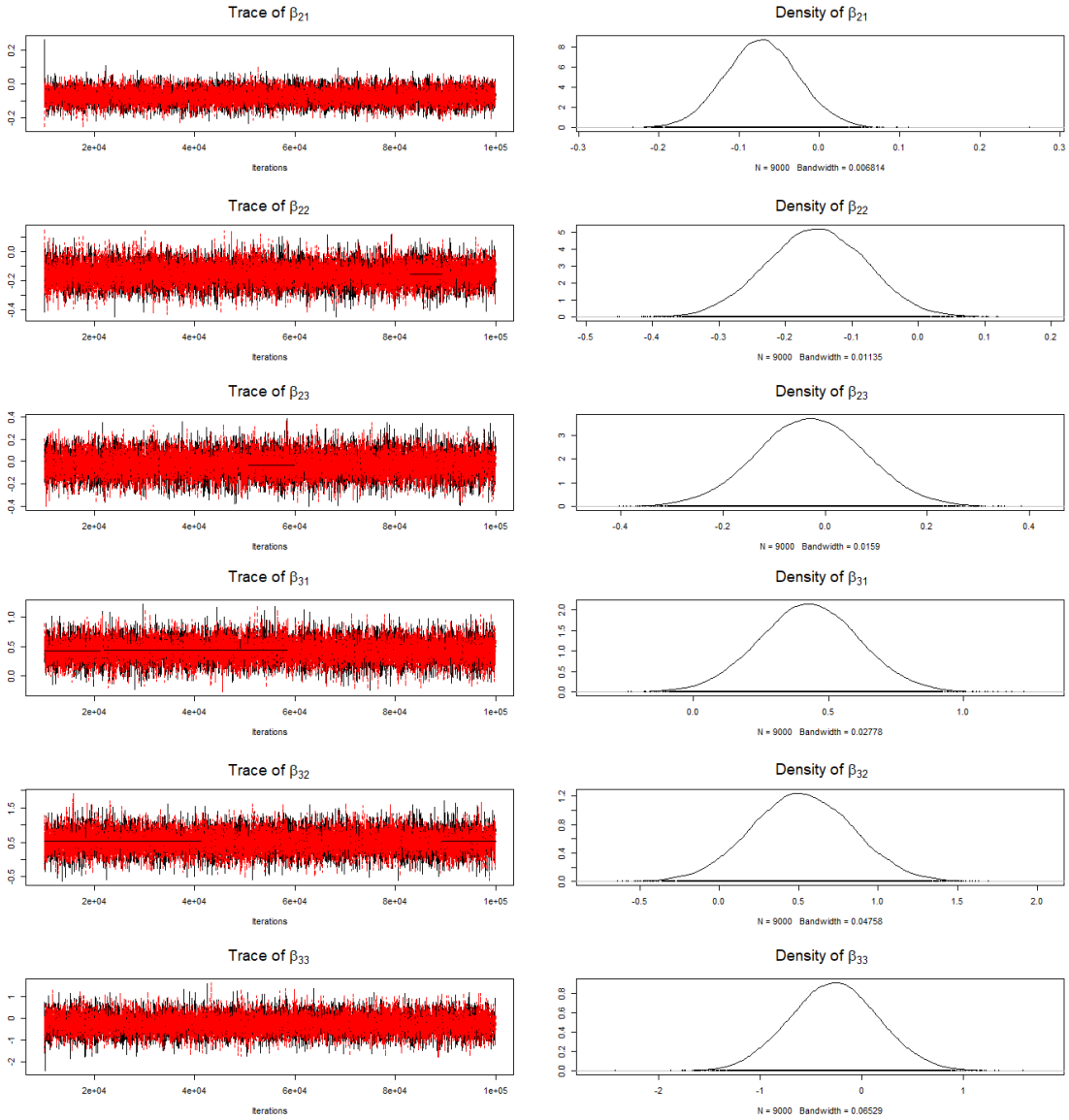

Figure 10：Trace plots and density for posterior distribution of parameters (model with correlated errors - Section 3.2).

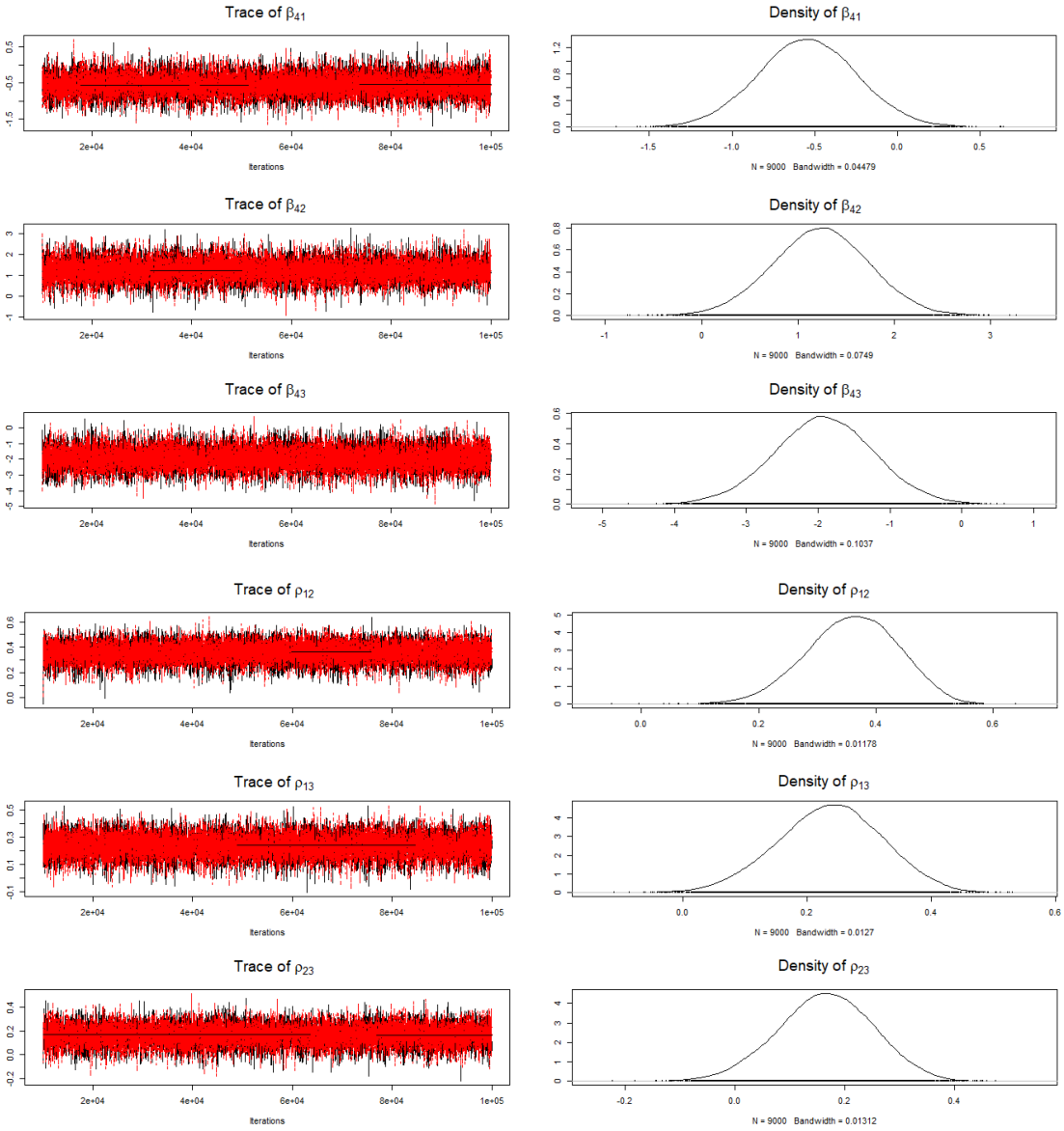

Figure 11：Trace plots and density for posterior distribution of parameters (model with correlated errors - Section 3.2)

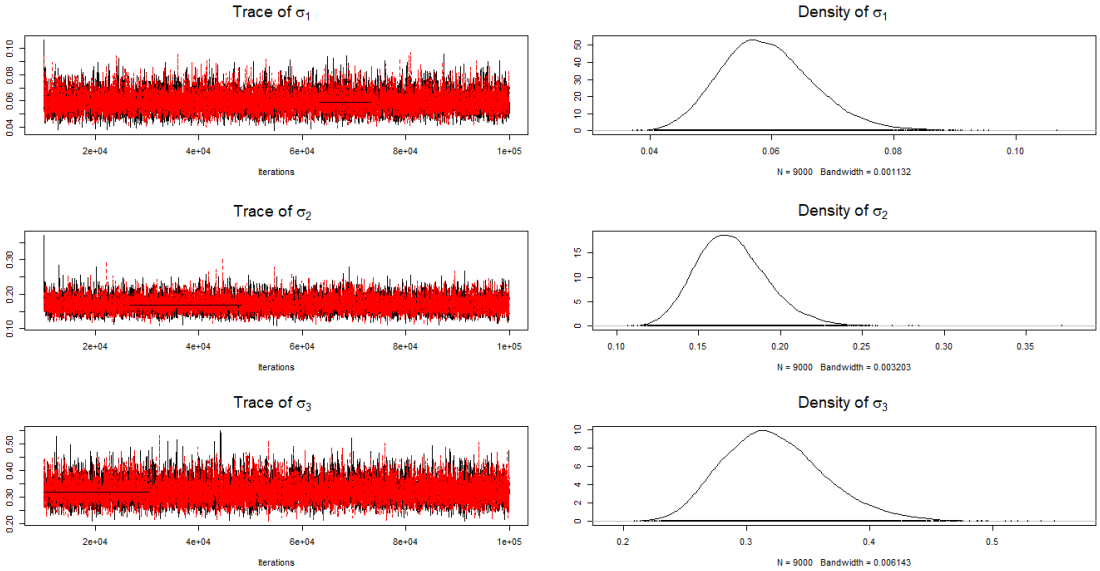

Figure 12：Trace plots and density for posterior distribution of parameters (model with correlated errors - Section 3.2).

Figures 6, 7, 8, 9, 10, 11 and 12 present the behaviour of MCMC implementation for the parameters of model with uncorrelated and correlated errors for application of volleyball data. We observe that the chains converged for all the parameters (see trace plots). Also, the convergence was monitored through by Gelman-Rubin diagnostic, being that the values for all the parameters were around 1.