Revista Colombiana de Estadística
January 2018, Volume 41, Issue 1, pp. 75 to 86
DOI: http://dx.doi.org/10.15446/rce.v41n1.63152

Finite Mixture of Compositional Regression With Gaussian Errors

Mixtura finita de una regresión composicional con errores gaussianos

Taciana Shimizu[a], Francisco Louzada[b], Adriano Suzuki[c]

Department of Applied Mathematics & Statistics, University of Sao Paulo, Sao Paulo, Brazil

Abstract

In this paper, we consider to evaluate the efficiency of volleyball players according to your performance of attack, block and serve, considering the compositional structure of the data related to the fundaments of this sport. In this way, we consider a finite mixture of regression model to compositional data. The maximum likelihood estimation of this model was obtained via an EM algorithm. A simulation study reveals that the parameters are correctly recovery. In addition, the estimators are asymptotically unbiased. By considering real dataset of Brazilian volleyball competition, we show that the model proposed presents best fit than the usual regression model.

***Key words*:** Compositional Data; Finite Mixture Regression; EM Algorithm.

Resumen

En este estudio evaluamos la eficiencia de los jugadores de voleibol de acuerdo con su desempeño de ataque, bloqueo y servicio, teniendo en cuenta la estructura composicional de los datos relacionados con los fundamentos de este deporte. Así, consideramos un modelo de regresión de mixtura finita para datos composicionales. La estimación de máxima verosimilitud fue obtenida via un Algoritmo EM. Un estudio de simulación revela que los parámetros son correctamente recuperados. Adicionalmente, los estimadores son asintóticamente insesgados. Considerando dados reales del campeonato de voleyball brasileño nosotros mostramos que el modelo propuesto presenta mejor ajuste que el modelo de regresión usual.

***Palabras clave*:** algoritmo EM; Datos Composicionales; mixtura finita.

[a]PhD. E-mail: taci_kisaki@yahoo.com.br
[b]PhD. E-mail: louzada@icmc.usp.br
[c]PhD. E-mail: suzuki@icmc.usp.br

1. Introduction

The performance of highlevel volleyball teams is considered fundamental for guarantee success at championships. Such performance may be related to efficiency of the players at the game. The knowledge about the main factors (for instance, the efficiency of the players) that affect the result of a game helps the decision-making of coaches, providing advantages for improving the skills of the teams. Hence, this is an important issue that must be analysed to contribute to the development of tactical and technical strategies.

Some studies about the efficiency of the volleyball players have been developed recently. For example, Bozhkova (2013) analyzed the efficiency of the best volleyball players based on the scoring winning points and the assisting actions, concluding that the attack is the most points-winning skill within the best volleyball players in the world. Pena, Guerra, Busca & Serra (2013) evaluated skills and factors that better predicted the outcomes of a regular seasons volleyball matches based on the logistic regression.

The points scored of the players like attack, block and serve have structure of compositional data which represent positive components, i.e., proportions of a whole. Compositional parts can be expressed in any scale without loss of information: accordingly, the sample space of representation of compositional data with a constant sum constraint is the simplex defined by $\mathbb{S}^D = \{(x_1, \ldots, x_D) : x_j > 0$ for $j = 1, \ldots, D$ and $\sum_{j=1}^{D} x_j = k\}$, where $k > 0$ and D is the number of variables (components).

Three essential principles of compositional data analysis are scale invariance, permutation invariance and subcompositional coherence (Aitchison 1986, Pawlowsky Glahn, Egozcue & Tolosana-Delgado 2015). Scale invariance means that a composition has information only about relative values. According to Aitchison & Egozcue (2005), such concept is easily formalized into a statement that all meaningful functions of a composition can be expressed in terms of a set of component ratios. The concept of permutation invariance is that if it provides same results when the components in the composition is changed. Finally, the subcompositional coherence can be summarized as: if we have two compositions, being one full compositions and another one a subcomposition of these full compositions, the inference about the relations within the common parts should be the same results (Aitchison & Egozcue 2005).

Aitchison & Shen (1980) and Aitchison (1986) introduced an appropriate theory for compositional data. The methodology involves transformations from restricted sample space simplex to well-defined real sample space $\mathbb{R}$. The general idea consists in the constraints that are removed, then standard statistical methods can be applied to the transformed observations. Such transformations were named by the additive logratio transformation (alr) and the centered logratio transformation (clr). Indeed, both alr and clr transformation represent coordinates with relation to the Aitchison geometry (Pawlowsky Glahn et al. 2015).

The use of multivariate normal distribution to compositional data can be found in Hron, Filzmoser & Thompson (2012), Egozcue, Daunis-I-Estadella, Pawlowsky-Glahn, Hron & Filzmoser (2011), among others.

Thus, our main motivation is to study the efficiency of the volleyball players through of the performance of attack, block and serve that result in point scoring in a game. The methodology of compositional data was applied in the points scored of the players during all the League: attack (x_1), block (x_2) and serve (x_3). Beyond that, it was considered a compositional regression model to study the relation between the fundaments and the associated covariates: z_1 is the percent of the team's efficiency in the reception and z_2 is the ratio of wins sets under losers sets, i.e., the higher the value of such ratio, the more likely the number of wins sets of the teams.

Preliminary, it was fitted a bivariate normal regression modelling for y_1 and y_2 independent random variables. Figure 1 shows the qq-plots of the fitting. Moreover, it was calculated the Shapiro-Wilks (SW) test, Kolmogorov-Smirnov (KS) test and Anderson-Darling (AD) test to verify the normality assumption of the data. The SW, KS and AD tests for y_1 presented p-value equal to 0.000 for all tests, rejected the null hypothesis that the sample came from an univariate normal distribution. On the other hand, the SW, KS and AD tests showed that y_2 follows an univariate normal distribution with p-values: 0.855, 0.902, 0.678, respectively.

According to the tests on the normality assumption for y_1 and y_2, a new approach for this data must have be considered, mainly for y_1. In this case, the mixture analysis is conducted to investigate the better fit for the the efficiency in points scoring by volleyball players.

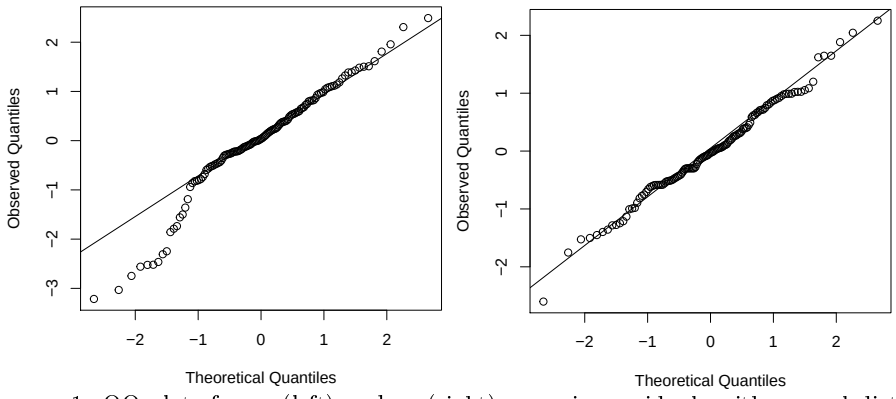

FIGURE 1: QQ-plots for y_1 (left) and y_2 (right) assuming residuals with normal distribution.

The aim of the paper is to introduce a Gaussian mixture regression model for compositional data with alr transformation, considering the multivariate structure of the data.

The methodology of finite mixture models has been much discussed in the literature. Quandt & Ramsey (1978) proposed such methodology in general form of switching regression. One of its advantages is to identify and relate populations

with two or more subpopulations. According to Miljkovic, Shaik & Miljkovic (2016), the variability of the variable may be explained better through by the investigation in a mixture of two or more distributions than a single distribution.

The paper is organized as follows. Section 2 introduces some preliminaries for compositional data and the methodology of Gaussian mixture regression model applied through the alr-coordinates, Sections 3 and 4 provide the results of the simulation study and application to a real data set related to the Brazilian Men's Volleyball Super League 2014/2015 and Section 5 ends the paper with some final remarks.

2. Methodology

First of all, the definition of compositional data is given below. Consider $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)^\top$ a compositional vector, x_i a positive value, for $i = 1, \ldots, D$ and $x_1 + x_2 + \cdots + x_D = 1$.

The operation *closure* assigns a constant sum representative to a composition. It divides each component of a vector by the sum of the components, rescaling of the initial vector to the constant sum 1. In mathematical terms, the definition is given by

Definition 1 (Closure)**.** For any vector of D strictly positive real components, $\boldsymbol{x} = [x_1, \ldots, x_D] \in \mathbb{R}_+^{\mathbb{D}}$, $x_i > 0$ for all $i = 1, \ldots, D$, the closure of $\boldsymbol{x}$ to $k > 0$ is defined as

$$C(\boldsymbol{x}) = \left[\frac{k.x_1}{\sum_{i=1}^{D} z_i}, \ldots, \frac{k.x_D}{\sum_{i=1}^{D} z_i} \right].$$

The family of the logratio coordinates is an alternative to lead with the constraints of compositional data, applying them before the statistical analysis. One of them was introduced by Aitchison (1986) called alr-coordinates. It is defined as

$$\begin{aligned} \text{alr} : \mathbb{S}^D &\to \mathbb{R}^{D-1} \\ \boldsymbol{y} = \text{alr}(\boldsymbol{x}) &= \left[\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right] = \zeta. \end{aligned}$$

The inverse alr-coordinates is given by

$$\boldsymbol{x} = \text{alr}^{-1}(\zeta) = C[\exp(\zeta_1), \ldots, \exp(\zeta_{D-1}), 1].$$

The alr-coordinates are not symmetric in the components, because the part x_D is in the denominator of the component logratios. Such coordinates $\zeta_i = \ln(x_i/x_D)$ are simple logratios and easily interpretable (Pawlowsky Glahn et al. 2015).

The regression model assuming alr-coordinates for the response variable is given by

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{z_i}\boldsymbol{\beta_1} + \boldsymbol{\epsilon_i}, \tag{1}$$

where $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{id})$ is a vector $(1 \times d)$ of response variables where $d = D - 1$ and D number of the components; $\boldsymbol{z_i}$ is a vector $(1 \times p)$ of covariates associated to the i-th sample; $\boldsymbol{\beta_0}$ is a vector $(1 \times d)$ intercepts; $\boldsymbol{\beta_1}$ is a matrix $(p \times d)$ of regression coefficients and $\boldsymbol{\epsilon_i}$ are random errors with distribution $N(0, \sigma^2)$, for $j = 1, \ldots, D-1$ and $i = 1, \ldots, n$.

In order to obtain the mixture structure for y_1 and univariate normal regression for y_2, the likelihood $L = L1 + L2$ for $\theta = (\pi_1, \ldots, \pi_K, \beta_{01}, \ldots, \beta_{0K}, \beta_{11}, \ldots, \beta_{1K}, \sigma_1^2, \ldots, \sigma_K^2)$ is

$$L_1(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi_k(y_{i1}|\mu_k, \sigma_k^2)$$
$$L_2(\theta) = \prod_{i=1}^{n} \phi(y_{i2}|\mu_2', \sigma_2^{2'}) \tag{2}$$

where $\phi(y|\mu_k, \sigma_k^2)$ is the normal distribution with mean $\mu_k = \beta_{0k} - \beta_{1k} z_i$ and variance σ^2, for $k = 1, \ldots, K$ and $i = 1, \ldots, n$.

The standard tool for estimate the parameters of mixture models is the EM algorithm, known for its applications in clustering and classifications models (McLachlan & Peel 2000). The simulation studies and statistical analysis of application were perfomed using `R` software (R Development Core Team 2013) through of the packages mixtools, maxLik and compositions.

2.1. EM Algorithm for Regression Model

The standard methods for finding maximum likelihood solution fail to solve the present setup. A powerful tool is to apply the EM algorithm proposed by Dempster, Laird & Rubin (1977).

The EM algorithm is an iterative method and the process of iterations is based on two steps, E (for expectation) and M (for maximization).

Following Faria & Soromenho (2010), the E-step calculates the Q-function which the expected value of the log likelihood function conditional on the parameter estimates and the observed data on the $(t+1)$th iteration,

$$Q(\theta, \theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(t)} \phi_k(y_{i1}|\mu_k, \sigma_k^2), \tag{3}$$

where for $i = 1, \ldots, n$ and $k = 1, \ldots, K$

$$w_{ik}^{(t)} = \frac{\pi_k^{(t)} \phi_k(y_{i1}|\mu_k, \sigma_k^2)}{\sum_{k=1}^{K} \pi_k^{(t)} \phi_k(y_{i1}|\mu_k, \sigma_k^2)}, \tag{4}$$

represents the posterior probability that the ith observation belongs to the kth component of the mixture.

In the M-step, the function (3) is maximized to obtain the updated estimates $\theta^{(t+1)}$. It follows that the M-step involves solving the following explicity equations expressed by,

$$\begin{aligned}
\hat{\pi}_k^{(t+1)} &= \frac{\sum_{i=1}^n w_{ik}^{(t)}}{n}, \\
\hat{\beta}_k^{(t+1)} &= (Z^\top W_k Z)^{-1} Z^\top W_k Y \text{ and} \\
\hat{\sigma}_k^{(t+1)} &= \frac{\sum_{i=1}^n w_{ij}^{(t)} (y_{1i} - z_i^\top \hat{\beta}_k^{(t+1)})^2}{\sum_{i=1}^n w_{ik}^{(t)}},
\end{aligned}$$

where Z is a $n \times (d+1)$ matrix of predictors, W_k is a $n \times n$ diagonal matrix with diagonal entries $w_{ik}^{(t)}$ and Y is a $n \times 1$ vector of response variable for $k = 1, \ldots, K$ (Faria & Soromenho 2010).

According to Migon, Gamerman & Louzada (2014), approximate $(1-\alpha)$ 100% confidence intervals for the parameters $\beta_{0j}, \beta_{1j}, \sigma_j, \beta_j$ are given by $\widehat{\beta}_{0j} \pm \xi_{\delta/2}\sqrt{Var(\widehat{\beta}_{0j})}$, $\widehat{\beta}_{1j} \pm \xi_{\delta/2}\sqrt{Var(\widehat{\beta}_{1j})}$, $\widehat{\sigma}_j \pm \xi_{\delta/2}\sqrt{Var(\widehat{\sigma}_j)}$ and $\widehat{\beta}_j \pm \xi_{\delta/2}\sqrt{Var(\widehat{\beta}_j)}$, where $\xi_{\delta/2}$ is the upper $\delta/2$ percentile of standard normal distribution and $j = 1, 2$.

We considered the discrimination criterion method based on log-likelihood function evaluated at the MLEs. Let m be the number of parameters to be fitted and $\hat{\boldsymbol{\theta}}$ the MLE's of $\boldsymbol{\theta}$, the discrimination criterion method is Akaike information criterion (AIC) computed through $AIC = -2l(\hat{\boldsymbol{\theta}}; \boldsymbol{x}) + 2m$.

3. Simulation Study

A simulation study via Monte Carlo methods was performed in order to study the asymptotic properties of the MLEs. It was simulated 1,000 samples of size $n = 100, 200, 300$ and 400 considering models with two components of mixture fixed in two types of cases $\boldsymbol{\pi_A} = (0.5, 0.5)$ and $\boldsymbol{\pi_B} = (0.2, 0.8)$. The true parameter values to perform this procedure were $\beta_{01} = -2, \beta_{02} = 5, \beta_{11} = 0.5, \beta_{12} = 0.5, \sigma_1 = 2$ and $\sigma_2 = 3$.

The data was generated randomly by the following scheme. A uniform random number $u \in (0, 1)$ was generated and the respective value was used to select a specific component k from mixture of regression models. Moreover, the associated covariate was generated through by $z_1 \sim$ Bernoulli (0.5) and a normal random ϵ_{ik} with mean 0 and variance σ_k^2, for $k = 1, 2$. Lastly, the value y_{1i} was calculated based on the values of z_1, ϵ_{ik}.

The criteria used to verify the performance of the algorithm were bias, standard deviation (SD), the mean square error (MSE) and coverage probability (CP). The coverage probability of confidence interval was computed through by bootstrapping, whereas the standard errors are not provided by the EM algorithm used in parameter estimation.

Tables 1 and 2 display the averages of the maximum likelihood estimates (Mean), standard deviation (SD), bias, mean square error (MSE) and coverage probability (CP) of the asymptotic 95% confidence intervals for the parameters considering two cases when $\boldsymbol{\pi_A} = (0.5, 0.5)$ and $\boldsymbol{\pi_B} = (0.2, 0.8)$. We can observe that the estimates are closer to the real value, besides the estimators are asymptotically unbiased for the parameters. According to the increase of the sample size, the MSE values decrease. Moreover, the coverage probabilities were stable.

TABLE 1: Simulated data. Mean, SD, bias, MSE and CP for estimates based on 1,000 generated samples of the two-component mixtures regression models.

Sample	Parameter	β_{01}	β_{02}	β_{11}	β_{12}	σ_1	σ_2	π_1	π_2
Size	Fixed value	-2	5	0.5	0.5	3	2	0.5	0.5
	Mean	-1.944	4.989	0.509	0.505	1.900	2.861	0.492	0.508
	SD	0.872	1.442	0.799	1.219	0.539	0.701	0.143	0.143
$n = 100$	Bias	0.055	-0.011	0.009	0.005	-0.100	-0.138	-0.008	0.008
	MSE	0.764	2.079	0.638	1.487	0.301	0.511	0.020	0.020
	CP	0.866	0.829	0.922	0.905	0.836	0.822	0.823	0.823
	Mean	-2.006	4.940	0.499	0.504	1.929	2.963	0.489	0.510
	SD	0.539	1.051	0.562	0.841	0.346	0.503	0.099	0.099
$n = 200$	Bias	-0.006	-0.060	-0.001	0.004	-0.070	-0.037	-0.010	-0.010
	MSE	0.290	1.107	0.315	0.707	0.125	0.254	0.010	0.010
	CP	0.920	0.883	0.928	0.924	0.903	0.857	0.885	0.885
	Mean	-2.009	4.915	0.521	0.500	1.947	2.998	0.490	0.510
	SD	0.424	0.799	0.434	0.655	0.269	0.428	0.082	0.082
$n = 300$	Bias	-0.009	-0.085	0.021	0.000	-0.053	-0.002	-0.009	0.009
	MSE	0.180	0.647	0.189	0.429	0.075	0.183	0.007	0.007
	CP	0.928	0.895	0.934	0.935	0.926	0.865	0.903	0.903
	Mean	-1.987	4.994	0.482	0.466	1.983	2.979	0.498	0.502
	SD	0.346	0.645	0.373	0.554	0.202	0.346	0.063	0.063
$n = 400$	Bias	0.013	-0.005	-0.018	-0.034	-0.017	-0.021	-0.002	0.002
	MSE	0.120	0.416	0.140	0.308	0.041	0.120	0.004	0.004
	CP	0.919	0.923	0.957	0.927	0.943	0.907	0.917	0.917

TABLE 2: Simulated data. Mean, SD, bias, MSE and CP for estimates based on 1,000 generated samples of the two-component mixtures regression models.

Sample	Parameter	β_{01}	β_{02}	β_{11}	β_{12}	σ_1	σ_2	π_1	π_2
Size	Fixed value	-2	5	0.5	0.5	3	2	0.2	0.8
	Mean	-0.847	5.594	0.567	0.468	2.196	2.502	0.352	0.648
	SD	2.318	1.433	1.605	1.366	1.087	0.819	0.247	0.247
$n = 100$	Bias	1.153	0.594	0.067	-0.032	0.196	-0.498	0.152	-0.152
	MSE	6.704	2.406	2.581	1.868	1.220	0.918	0.084	0.084
	CP	0.679	0.808	0.862	0.885	0.636	0.773	0.694	0.694
	Mean	-1.205	5.312	0.434	0.538	2.159	2.734	0.294	0.706
	SD	1.913	0.888	0.960	0.897	0.890	0.615	0.203	0.203
$n = 200$	Bias	0.794	0.312	-0.066	0.038	0.159	-0.265	0.094	-0.094
	MSE	4.291	0.885	0.927	0.806	0.817	0.448	0.050	0.050
	CP	0.765	0.871	0.921	0.924	0.750	0.842	0.794	0.794
	Mean	-1.352	5.212	0.452	0.489	2.160	2.798	0.277	0.723
	SD	1.747	0.750	0.785	0.648	0.800	0.539	0.188	0.188
$n = 300$	Bias	0.648	0.212	-0.048	-0.011	0.161	-0.202	0.077	-0.077
	MSE	3.473	0.607	0.619	0.420	0.665	0.332	0.041	0.041
	CP	0.800	0.891	0.932	0.938	0.788	0.862	0.826	0.826
	Mean	-1.480	5.151	0.480	0.506	2.139	2.856	0.260	0.740
	SD	1.590	0.638	0.701	0.500	0.728	0.472	0.171	0.171
$n = 400$	Bias	0.520	0.151	-0.020	0.006	0.139	-0.144	0.060	-0.060
	MSE	2.798	0.429	0.492	0.250	0.549	0.243	0.033	0.033
	CP	0.815	0.895	0.921	0.934	0.801	0.866	0.834	0.834

4. Application

We applied the proposed methodology a real data set where the sample corresponds to 127 players extracted from (*Brazilian Volleyball Confederation (CBV)* 2016). The data related to proportions of the volleyball players who participated of Brazilian Men's Volleyball Super League 2014/2015. The methodology of compositional data was applied in the points scored of the players during all the League which are considered components: attack (x_1), block (x_2) and serve (x_3). The associated covariates to the model are: z_1 is the percent of the team's efficiency in the reception and z_2 is the ratio of wins sets under losers sets, i.e., the higher the value of such ratio, the more likely the number of wins sets of the teams.

The main goal is to verify individually whether the fundaments (attack, block and serve) have relation to the associated covariates.

The ternary diagram (Figure 2) presents the three fundaments attack, block and serve. Such type of graphic represents a 3-part composition using a 2-dimensional plot (Van Den Boogaart & Tolosana-Delgado 2013). There is a concentration of points in direction to the attack component. Only some points are directed for block and serve components.

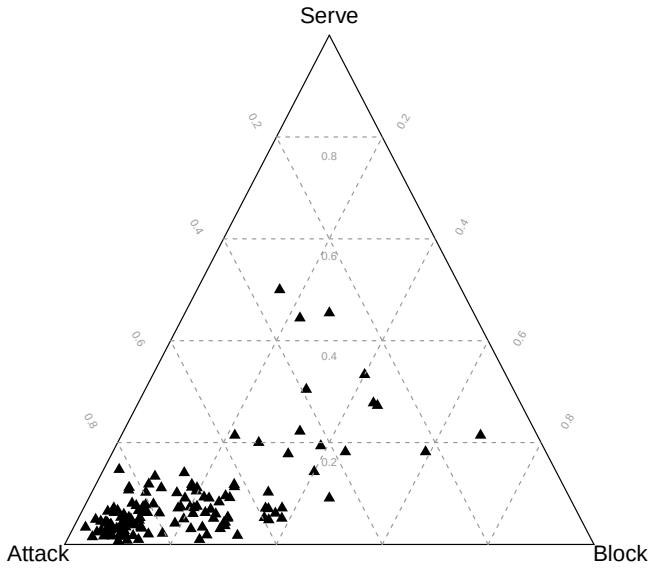

FIGURE 2: Ternary diagram for the components: attack, block and serve.

For sake of comparison, the discrimination criterion method was analysed based on log-likelihood function evaluated at the MLEs. Table 3 presents the maximum likelihood estimates and the result of AIC criteria for fitted models. Gaussian mixture model with 2 components has smallest value AIC indicating the best fit among the other models considered. We can observe that the mixing proportions by component of the 2-GM model are 0.118 and 0.882 reflecting how the data is distributed within each subpopulation. The model with 2-GM fitted better than others regressions for y_1 and based on the preliminary test of normality, the fit of the linear regression is adequate for y_2 (Figure 1). Such conclusions are corroborated by the behaviour of the fitting for the residuals of the 2-GM model in the Figure 3.

TABLE 3: Summary of the Maximum Likelihood Estimates for the parameters and comparison through the discrimination criterion of the bivariate normal (BN), 2-component Gaussian mixture (2-GM) and 3-component Gaussian mixture (3-GM) regressions for y_1 and y_2.

BN		2-GM		3-GM	
β_{01}	1.094	β_{01}	-2.670	β_{01}	-2.933
β'_{02}	-0.165	β_{02}	1.938	β_{02}	1.785
β_{11}	0.046	β_{11}	0.091	β_{03}	4.629
β'_{12}	0.037	β_{12}	0.028	β_{11}	0.101
β_{21}	-0.415	β_{21}	-0.534	β_{12}	0.029
β'_{22}	-0.344	β_{22}	-0.283	β_{13}	-0.024
σ_1	1.054	σ_1	0.479	β_{21}	-0.586
σ'_2	0.802	σ_2	0.736	β_{22}	-0.243
		π_1	0.118	β_{23}	0.771
		π_2	0.882	σ_1	0.517
				σ_2	0.616
				σ_3	0.156
				π_1	0.131
				π_2	0.816
				π_3	0.053
LogLik	-339.083		-327.405		-322.754
AIC	694.166		682.811		683.509

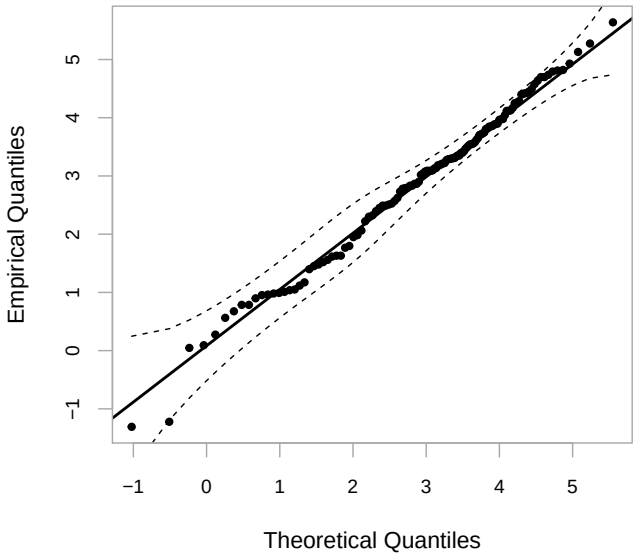

FIGURE 3: QQ-plot for the residuals of the 2-GM model.

5. Conclusions

This study provides a mixture compositional regression model to study the efficiency volleyball players. Based on the preliminary results, one of the variables, namely y_1, did not show good fit for a regression model with normal errors according to the tests of normality and the Figure 1. The Gaussian mixture compositional regression model was then developed and fitted to our dataset, corroborating with the preliminary results. Two approaches were considered, two and three components mixture regressions for the data of efficiency of volleyball players, according to the performance of the fundaments: attack, block and serve. Furthermore, the estimates of simulation study and the application for real dataset were obtained via an EM algorithm. The results pointed out that the fundaments of volleyball players are better described by using the compositional mixture model with two components, according to the discrimination criteria. Such approach considers the heterogeneous characteristics of the data.

Finally, the study's conclusions identified points in the attack as fundamental to highlight the effective teams through the estimates of proportions. As future work, following Egozcue & Pawlowsky-Glahn (2005) and Egozcue, Pawlowsky-Glahn, Mateu-Figueras & Barceló-Vidal (2003), the orthonormal coordinates (isometric logratio) can be incorporated in the finite mixture compositional model, instead of alr-coordinates, probable leading to some improvement.

Acknowledgements

The research is supported by the Brazilian organization FAPESP.

[Received: February 2017 — Accepted: September 2017]

References

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

Aitchison, J. & Egozcue, J. (2005), 'Compositional data analysis: Where are we and where should we be heading?', *Mathematical Geology* **37**(7), 829–850.

Aitchison, J. & Shen, S. M. (1980), 'Logistic-normal distributions: Some properties and uses', *Biometrika* **67**(2), 261–272.

Bozhkova, A. (2013), 'Playing efficiency of the best volleyball players in the world', *Research in Kinesiology* **41**(1), 92–95.

Brazilian Volleyball Confederation (CBV) (2016), http://www.cbv.com.br/v1/superliga1415/estatisticas-novo.asp?gen=m. Accessed: 2016-01-20.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1), 1–38.

Egozcue, J. J., Daunis-I-Estadella, J., Pawlowsky-Glahn, V., Hron, K. & Filzmoser, P. (2011), 'Simplicial regression. the normal model', *Journal of Applied Probability and Statistics* **6**(1), 87–108.

Egozcue, J. J. & Pawlowsky-Glahn (2005), 'Groups of parts and their balances in compositional data analysis', *Mathematical Geology* **37**(4), 795–828.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barceló-Vidal, C. (2003), 'Isometric logratio transformations for compositional data analysis', *Mathematical Geology* **35**, 279–300.

Faria, S. & Soromenho, G. (2010), 'Fitting mixtures of linear regressions', *Journal of Statistical Computation and Simulation* **80**(2), 201–225.

Hron, K., Filzmoser, P. & Thompson, K. (2012), 'Linear regression with compositional explanatory variables', *Journal of Applied Statistics* **39**(5), 1115–1128.

McLachlan, G. J. & Peel, D. (2000), *Finite Mixture Models*, Wiley series in probability and statistics, Wiley & Sons, New York.

Migon, H. S., Gamerman, D. & Louzada, F. (2014), *Statistical Inference: An Integrated Approach*, Chapman & Hall/CRC, London.

Miljkovic, T., Shaik, S. & Miljkovic, D. (2016), 'Redefining standards for body mass index of the us population based on brfss data using mixtures', *Journal of Applied Statistics* pp. 1–15.
*http://dx.doi.org/10.1080/02664763.2016.11683661

Pawlowsky Glahn, V., Egozcue, J. J. & Tolosana-Delgado, R. (2015), *Modeling and analysis of compositional data*, John Wiley & Sons.

Pena, J., Guerra, J. R., Busca, B. & Serra, N. (2013), 'Which skills and factors better predict winning and losing in high-level men's volleyball?', *Journal of Strength and Conditioning Research* **27**(9), 2487–2493.

Quandt, R. & Ramsey, J. (1978), 'Estimating mixtures of normal distributions and switching regression', *Journal of American Statistical Association* **73**, 730–738.

R Development Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*http://www.R-project.org

Van Den Boogaart, K. G. & Tolosana-Delgado, R. (2013), *Analyzing compositional data with R*, Springer.