Residuals and Diagnostics in Dirichlet Regression

Rafiq H. Hijazi
United Arab Emirates University - UAE

Abstract

Compositional data are rarely analyzed with the usual multivariate statistical methods. One approach to model such data is Dirichlet regression. We present various diagnostic methods for Dirichlet regression models. We discuss the use of quantile residuals to check the distributional assumptions. Measures of total variability and goodness of fit are proposed to assess the adequacy of the suggested models. An R-square measure based on Aitchison's distance is introduced. The likelihood distance is employed to identify the influential compositions. Finally, an example with real data is presented and discussed.

Keywords: Compositional data, Dirichlet regression, Aitchison's distance, explained variation, quantile residuals.

1 Introduction

Compositional data are non-negative proportions with unit-sum which occur in nearly all disciplines, but recognition and modelling of their basic structure have gotten particular attention in geology, chemistry, political science, business and economics. For example, a standard such data set notes the relative composition of sediments (sand, clay, silt) in an arctic lake (Aitchison, 1986). The usual multivariate covariances and correlations for such data can be misleading since the data are constrained to sum to one and hence the traditional multivariate statistical techniques can not be used.

Aitchison (1986) suggested an analysis based on the logratios of the compositional data so the traditional multivariate techniques can be applied on the transformed data. Campbell and Mosimann (1987) developed an alternative approach by extending the Dirichlet distribution to a class of Dirichlet Covariate Models (Dirichlet Regression).

The estimation in Dirichlet regression and the asymptotic properties of the estimates have been investigated by Campbell and Mosimann (1987) and Hijazi (2003). The examination of the residuals and the diagnostics is of great importance in regression analysis. In this paper, we propose the use of pseudo (quantile) residuals as a tool in checking the distributional assumption in Dirichlet regression and identifying the outlying compositions. Also, we propose different R^2 measures to assess the fit of the Dirichlet models to the compositional data. Two influence diagnostics based on Chi-square statistic and likelihood distance are presented to identify the influential compositions. Finally, an application to illustrate the use of the proposed techniques is introduced.

2 Dirichlet Regression

Let $\mathbf{x} = (x_1, ..., x_D)$ be a $1 \times D$ positive vector having Dirichlet distribution with positive parameters $(\lambda_1, ..., \lambda_D)$ with density function

$$f(\mathbf{x}) = \left(\Gamma(\lambda) / \prod_{i=1}^{D} \Gamma(\lambda_i) \right) \prod_{i=1}^{D} x_i^{\lambda_i - 1} \qquad (1)$$

where $\sum_{i=1}^{D} x_i = 1$ and $\lambda = \sum_{i=1}^{D} \lambda_i$.

A Dirichlet regression model is obtained by allowing the parameters of a Dirichlet distribution to change with a covariate. For a given covariate s, the parameters of a Dirichlet distribution $\mathcal{D}(\lambda_1, ..., \lambda_D)$ can be written as positive functions $h_j(s)$ of the covariate s. A different Dirichlet distribution is modelled for every value of the covariate, resulting in a conditional Dirichlet distribution with $\mathbf{x}|s$ is $\mathcal{D}(h_1(s), ..., h_D(s))$. The density function of this conditional distribution is

$$f(\mathbf{x}|s_i) = \left(\Gamma \left(\sum_{j=1}^{D} (h_j(s_i)) \right) / \prod_{j=1}^{D} \Gamma(h_j(s_i)) \right) \prod_{j=1}^{D} x_j^{h_j(s_i)-1}$$

and the conditional mean would be

$$E(\mathbf{X}|s) = \left(\frac{h_1(s)}{h(s)}, ..., \frac{h_D(s)}{h(s)} \right)$$

where $h(s) = \sum_{i=1}^{D} h_i(s)$.

The maximum likelihood method is used to estimate the parameters of the suggested model (Ronning, 1989; Campbell and Mosimann, 1987a; Hijazi, 2003). The asymptotic properties of the maximum likelihood estimates have been thoroughly investigated by Hijazi (2003).

Once the estimation has been accomplished, we focus on assessing the fit of the Dirichlet models to the compositional data. generally, the likelihood-based methods depend on the parametric assumption and a misspecification of the model may lead to inaccurate results. It is then of great importance to investigate the validity of the parametric assumption. Examination of the residuals, goodness-of-fit measures and influence diagnostics are widely used in model assessment in regression analysis. For Dirichlet regression, residuals and diagnostics are presented to investigate the goodness of fit of the estimated models.

3 Residuals

Based on the properties of Dirichlet distribution, the marginals of the Dirichlet distribution are single beta distributions. In others words, if $\mathbf{X} = (X_1, ..., X_D)$ is distributed as $\mathcal{D}(\lambda_1, ..., \lambda_D)$ and $\lambda = \sum_{i=1}^{D} \lambda_i$, then the random variable X_j has a beta distribution with parameters λ_j and $\lambda - \lambda_j$; i.e. $B(\lambda_j, \lambda - \lambda_j)$ for $1 \leq j \leq D$ with $F(x_j)$ being the cumulative distribution function. By probability integral transform, $p_j = F(x_j)$ follows a uniform distribution. The pseudo-residual, r_j, corresponding to the observation x_j is given by $r_j = \Phi^{-1}(p_j)$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution. If the Dirichlet distribution is the correct model with beta marginals, the pseudo residuals, r_j's, follow the standard normal distribution and can be treated as standardized residuals in linear regression (Zucchini and MacDonald, 1999).

4 R^2-Type Measures

In classical regression analysis, the coefficient of determination R^2 is used as a measure of explained variation. It has the interpretation as the proportion of explained variation in the dependent variable by the predictor variables of a given regression model. For Dirichlet regression, we need a numerical measure to evaluate model performance like the usual R^2 measure. In this section, we suggest three R^2 measures based on model likelihoods, total variability and sums of squares.

4.1 R^2-measure based on model likelihoods

Different R^2 measures have been proposed to evaluate regression models. The likelihood-ratio R^2 (R^2_L) has been widely used in the general linear models (Maddala, 1983), logistic regression (Magee, 1990) and Cox regression (Schemper, 1992). This measure is defined as

$$R^2_L = 1 - \left[\frac{L(\widehat{\theta}_0)}{L(\widehat{\theta}_1)}\right]^{\frac{2}{n}} \tag{2}$$

where $L(\widehat{\theta}_0)$ and $L(\widehat{\theta}_1)$ are the likelihoods of the constant and the covariate models respectively. The R^2_L measures the proportional improvement in the log-likelihood function due to the explanatory variable in the model, compared to the minimal "constant" model.

4.2 R^2-measure based on total variability

When introducing the logratio analysis, Aitchison (1986) suggested a measure of total variability based on the variation matrix of the transformed logratio data, $\mathbf{T}(\mathbf{x})$ defined as

$$\mathbf{T}(\mathbf{x}) = [\tau_{ij}] = [var\,\{\log(x_i/x_j)\}] \tag{3}$$

Obviously, $\mathbf{T}(\mathbf{x})$ is symmetric with zero diagonal elements. Aitchison's total variability measure $totvar(\mathbf{x})$ is defined as

$$totvar(\mathbf{x}) = \frac{1}{d}\sum_{i<j}[var\,\{\log(x_i/x_j)\}] = \frac{1}{2d}\sum \mathbf{T}(\mathbf{x}) \tag{4}$$

Aitchison compares the total variability of the observed data and the fitted data to obtain an R^2 measure (call it R^2_T) defined as

$$R^2_T = totvar(\widehat{\mathbf{x}})/totvar(\mathbf{x}) \tag{5}$$

where $\mathbf{x}$ is the observed data and $\widehat{\mathbf{x}}$ is the fitted data.

4.3 R^2-measure based on the sum of squares

The R^2 in ordinary least-squares regression is defined by R^2=1-SSE/SST, where SSE and SST denote the sum of the squared residuals and the sum of the squared distances from the mean, respectively. A general form of this R^2 is called the proportion of explained variation (PEV) and given by

$$PEV = \frac{\sum_i D(y_i) - \sum_i D(y_i|x_i)}{\sum_i D(y_i)} \tag{6}$$

where $D(y_i)$ and $D(y_i|x_i)$ represent a measure of the distance of y_i from a central location parameter, unconditional or conditional on a covariate x_i.

In the simplex geometry, Aitchison (1986) proposed Aitchison's distance (Δ_S) as a measure of the distance between two compositions. This distance is given by

$$\mathbf{\Delta_S}(\mathbf{X}, \mathbf{x}) = \left[\sum_{i=1}^{D}\left\{\log\frac{X_i}{g(\mathbf{X})} - \log\frac{x_i}{g(\mathbf{x})}\right\}^2\right]^{1/2} \tag{7}$$

where $g(y)$ is the geometric mean of the composition y. Aitchison, Barcelo-Vidal, Martin-Fernandez (2000) emphasize that (7) defines a metric on the simplex and has all the necessary properties required in compositional data analysis. Pawlowsky-Glahn and Egozcue (2002) investigated the invariance properties of $\mathbf{\Delta_S}$. Further, they showed that $\mathbf{\Delta_S}(\mathbf{X}, \xi)$ is minimized at the center ξ given by $\xi = agl(E[alr\mathbf{X}])$, where alr is the additive logratio transformation and agl is its inverse (Aitchison, 1986).

Now, let $\underline{\mathbf{X}}$ be a set of n compositions $\mathbf{x_1}, ..., \mathbf{x_n}$ and $\hat{\mathbf{x}}_\mathbf{1}, ..., \hat{\mathbf{x}}_\mathbf{n}$ be the fitted compositions. Then, the center of $\underline{\mathbf{X}}$ is given by

$$\mathbf{g} = \mathcal{C}\left(\left(\prod_{i=1}^{n} x_{i1}\right)^{\frac{1}{n}}, ..., \left(\prod_{i=1}^{n} x_{iD}\right)^{\frac{1}{n}}\right) \tag{8}$$

where $\mathcal{C}$ is the closure operation (Aitchison, 1986).

The compositional total sum of squares (CSST) and the compositional sum of squared residuals (CSSE) are then given by

$$CSST = \sum_{i=1}^{n} \Delta_S^2(\mathbf{x}_i, \mathbf{g}) \tag{9}$$

and

$$CSSE = \sum_{i=1}^{n} \Delta_S^2(\mathbf{x}_i, \hat{\mathbf{x}}_i) \tag{10}$$

Finally, our suggested R^2 measure; R_A^2 is given by

$$R_A^2 = 1 - \frac{CSSE}{CSST} \tag{11}$$

It is noteworthy that the R_A^2 measure can be used in logratio analysis as well.

5 Influence Diagnostics

Diagnostics in regression analysis include inspection of leverages and other influence measures like Cook's distance and likelihood displacement (Cook and Weisberg, 1982). Different methods have been suggested to detect the outliers in compositional data (Barcelô et al., 1996; Baxter, 1999) where visual assessment of compositions seems inappropriate. These methods are appropriate for unconditional compositions. When compositions depend on a covariate, residuals can be used to identify outlying compositions. Aitchison (1986) described the method of outliers detection in logratio analysis. We can use the joint distribution of the quantile residuals to identify the compositions with large Mahalanobis distance (Mardia et al., 1979) as outliers. Two measures are proposed to detect the influential compositions in Dirichlet regression. The first is based on Chi-squared statistic while the second is the likelihood distance.

5.1 Pearson Chi-Squared Statistic

Consider the composition $\mathbf{x} = (x_1, ..., x_D)$ where $\sum_{i=1}^{D} x_i = 1$, then we can treat this composition as multinomial probabilities. Thus, we can use the Pearson Chi-squared goodness of fit to compare the observed and the predicted compositions. For Dirichlet distribution with parameters $\Lambda = (\lambda_1, ..., \lambda_D)$, Boyles (1997) introduced the following modified chi-squared statistic

$$X^2 = (\lambda + 1) \sum_{i=1}^{D} \frac{(x_i - \mu_i)^2}{\mu_i} \tag{12}$$

where $\mu_i = \frac{\lambda_i}{\lambda}$ and $\lambda = \sum_{i=1}^{D} \lambda_i$.

Boyles showed that the later statistic is asymptotically distributed as chi-square with $D - 1$ degrees of freedom. The simulation studies have indicated that the use of the maximum likelihood estimates instead of the parameters will result in the same sampling distribution. The later statistic is then used to identify the outlying and influential compositions with large chi-squared values.

5.2 Likelihood distance

Let $\ell(\theta)$ be the log-likelihood function for the Dirichlet regression, the likelihood displacement (Cook el al., 1988) is defined by

$$LD_j = 2\left[\ell(\widehat{\theta}) - \ell(\widehat{\theta}_{(j)})\right] \tag{13}$$

where $\theta_{(j)}$ is the maximum likelihood estimate after deleting the j^{th} composition.

6 Example: Arctic lake sediments data

This example concerns the composition of 39 sediments taken from an Arctic lake. Each point represents the composition of sand, silt, and clay at different depths in a Arctic lake (Aitchison 1986). It is believed that the structure of this composition depends on the depth (s) where the sediment was taken. The ternary diagram in figure (1a) shows the distribution of the sediments. The effect of the depth as a covariate is expressed in figure (1b). Obviously, large values of the covariate are associated with a low proportion of the sand. Conversely, small values of the covariate correspond to large proportions of the sand and relatively higher proportions of silt and clay. The effect of the covariate is clearly curved as depicted in the ternary diagram. This relationship suggests that models based on constant parameters will not describe the variability well. Linear Dirichlet regression model with the water depth as a covariate is rejected in favor of the quadratic model based on likelihood ratio test (Casella and Berger 2002). The estimated parameters of the quadratic model are

$$\begin{aligned} \widehat{\lambda}_{sand} &= 5.240 - 0.072s + 0.001s^2 \\ \widehat{\lambda}_{silt} &= 3.426 - 0.203s + 0.011s^2 \\ \widehat{\lambda}_{clay} &= 3.635 - 0.391s + 0.013s^2 \end{aligned}$$

The observed and predicted compositions are shown is Figure (1a). The model shows a good fit for the compositions in the simplex. The fitted model follows the curvature of the compositions but fails to get closer to the 6-composition cluster in the top of the simplex. Now, we compare the performance of the model based on the marginal distributions. Figure (2a-c), shows the marginal distributions of the compositions and the fitted model against the water depth. The model appears to fit the three components well especially and follows the nonlinear form of the data quite well. Overall, the Dirichlet quadratic model shows a good performance in fitting the individual components as well as the compositions in the simplex.

After the fit of the model, the pseudo residuals are computed and used to produce the residuals plots and

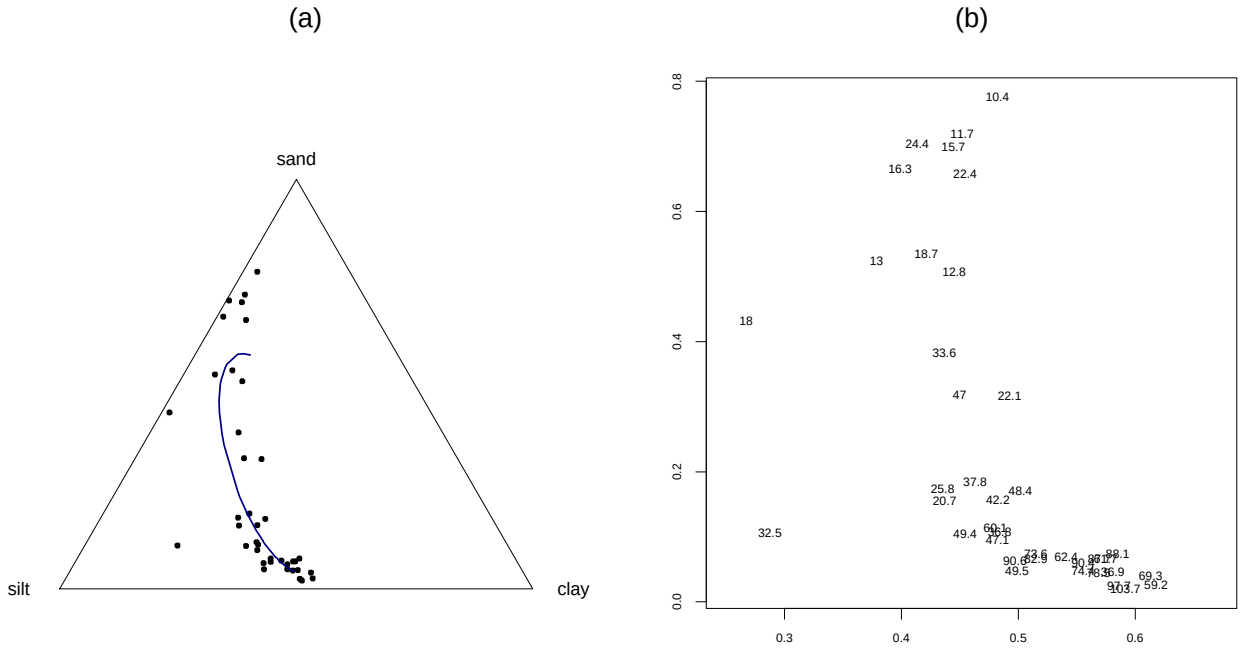

Figure 1: (a) Distribution of the sediments in the simplex (b) Distribution of sediments as a function of water depth

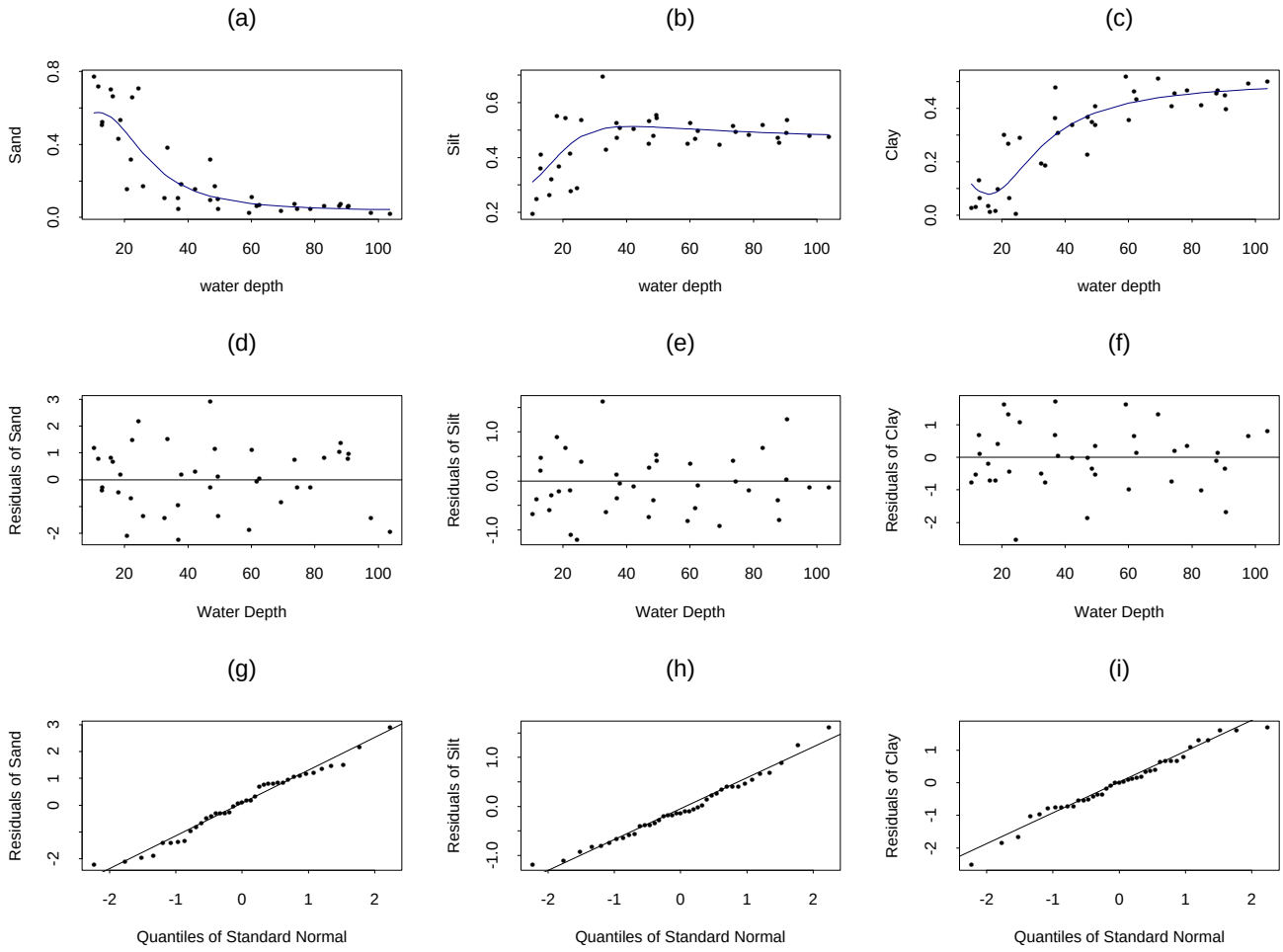

Figure 2: The marginal distributions, the residual plots and the normal probability plots for the sediments data

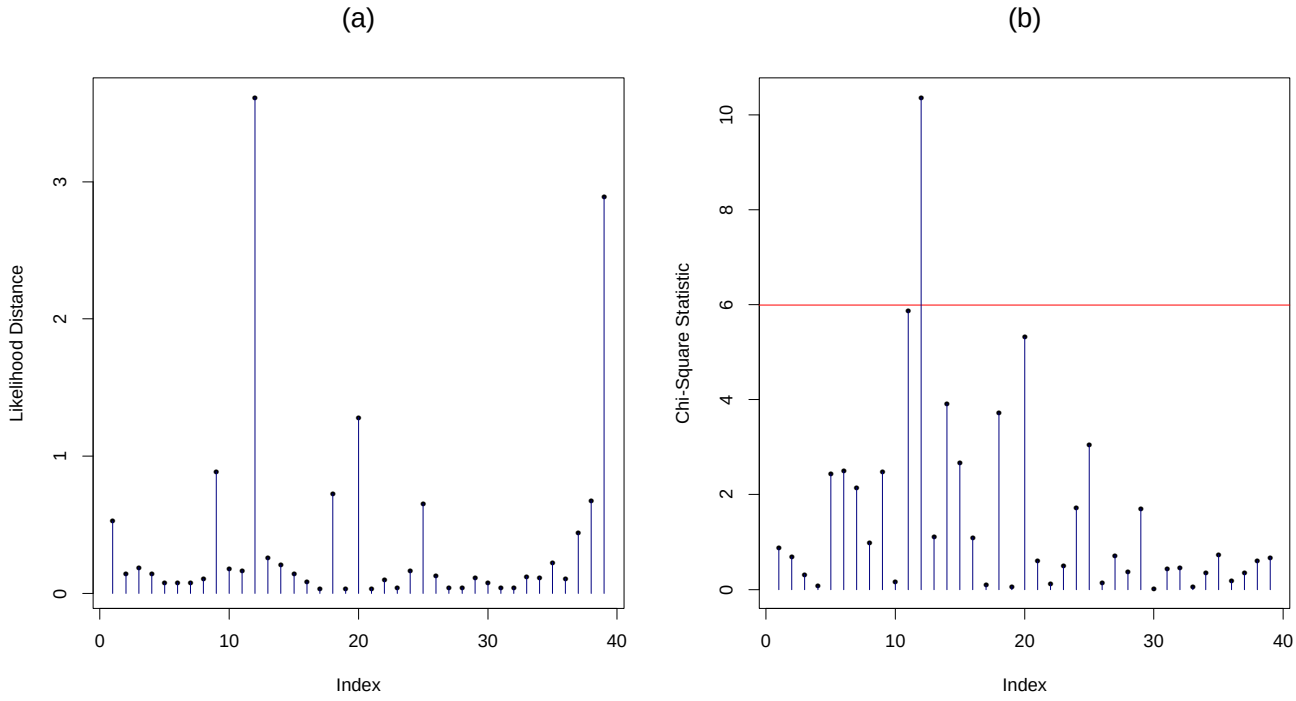

Figure 3: (a) The likelihood distance and (b) The chi-square statistic plots for the sediments data

the normal probability plots shown in the second and the third rows in Figure (2). An inspection of these plots reveals that there is no indication of violation of the parametric assumption or model misspecification.

The three proposed R^2 measures are R^2_L=97.56, R^2_T=53.78% and R^2_A= 60.65%. The likelihood-based measure is extremely high due to the strong dependence of the sediment compositions on the water depth. The other two measures are more conservative but show moderate percentage of explained variation by the Dirichlet model. The marginal plots of the compositions in Figure (2a-c) show moderate relationships between the compositions and the water depth. This is consistent with the values of R^2_T and R^2_A. Diagnostic plots are given in Figure (3). The two plots show that the 12th composition has the largest chi-square statistic and likelihood distance. This composition has the largest influence and it's the closest to the sides of the ternary diagram in Figure (1). The very small proportion of the clay in the composition (0.006) is inconsistent with the corresponding water depth (24.4) which explains the large influence of the composition. It is noteworthy that this composition has the largest Aitchison's and Mahalanobis distances and it has been identified as an outlier by Barcelô et al. (1996).

When we fitted the Dirichlet regression without the 12th composition, the estimated parameters of silt component show the highest change. The R^2 measures for the new model are 97.82%, 59.93% and 67.5% respectively. The most influential composition turns to be the one with the smallest proportion of sand.

7 Concluding Remarks

In this paper, we have investigated the residual analysis and diagnostics checking in modelling compositional data using Dirichlet regression. The quantile residuals are developed and used to check the parametric assumptions and severe misspecification of the model. Three R^2 measures have been proposed to assess the model and express the proportion of explained variation in the compositions by the covariate. Visual assessment of the outlying and influential compositions might be misleading. A modified Chi-squared statistic and the likelihood distance have been employed in identifying the influential compositions. Finally, an example with real compositional data was presented to illustrate the proposed techniques.

References

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, Chapman and Hall, New York.

Aitchison, J., Barcelo-Vidal, C., Martin-Fernandez, J. A. and Pawlowsky-Glahn, V. (2000), "Logratio analysis and compositional distance," *Mathematical Geology*, **32(3)**, 271–275.

Barcelô, C., Pawlowsky, V. and Grunsky, E. (1996), "Some aspects of transformations of compositional data and the identification of outliers," *Mathematical Geology*, **28(4)**, 501–518.

Baxter, M. (1999), "Detecting multivariate outliers in artefact compositional data," *Archaeometry*, **41(2)**, 321–338.

Boyles, R. (1997), "Using chi-square statistic to monitor compositional process data," *Journal of Applied Statistics*, **24(5)**, 589–602.

Campbell, G. , and Mosimann, J. E. (1987), "Multivariate methods for proportional shape," *ASA Proceedings of the Section on Statistical Graphics*, 10–17.

Casella, G. and Berger, R. (2002), *Statistical Inference*, Duxbury, California.

Cook, R. D., Peña, D., Weisberg, S. (1988), "The likelihood displacement: a unifying principle for influence measures," *Communications in Statistics: Theory and Methods*, **17**, 623–640.

Cook, R. and Weisberg, S. (1982), *Residuals and Influence in Linear Regression*, Champan and Hall, New York.

Hijazi, R. (2003), *Analysis of Compositional data using Dirichlet Covariate Models*, *PhD Dissertation*, The American University, Washington, DC.

Maddala, G. S. (1983), *Limited-Dependent and Quantitative Variables in Econometrics*, Cambridge, University Press.

Magee, L. (1990), "R^2 measures based on Wald and likelihood ratio joint significance tests," *American Statistician*, **44**, 250–253.

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

Pawlowsky-Glahn, V. and Egozcue, J. J. (2002), "BLU estimators and compositional data," *Mathematical Geology*, **34(3)**, 259–274.

Ronning, G. (1989), "Maximum likelihood estimation of Dirichlet distributions," *Journal of Statistical Computation and Simulation*, **32**, 215–221.

Schemper, M. (1992), "Further results on the explained variation in proportional hazards regression," *Biometrika*, **79**, 202–204.

Zucchini, W. and MacDonald, I. L. (1999), "Illustrations of the use of pseudo-residuals in assessing the fit of a model," *Proceedings of the 14th International Workshop on Statistical Modelling*, , Graz. Friedl, H. Berghold, A. and Kauermann, G., 409–416.