Institute of Statistics

A beta partial least squares regression model: Diagnostics and application to mining data

Thesis

For the degree of Master in Statistics
Institute of Statistics
University of Valparaíso

Presented by:
Mauricio Hernán Huerta Aguiar

Advisors:
Dr. Víctor Leiva
Dr. Marco Riquelme

Valparaíso, Chile, 28-December-2016

Acknowledgements

I deeply dedicate this work to my mother, Elcira Aguiar, for the gift of life and early responsibility concerning to my education. I will never forget that she taught me to read and write, which granted my future learning process. To my father, Manuel Huerta, for giving me life lessons and supporting me at all time. Both of you have been fundamental pillars in my life and I thank to you whom I am today. I love both of you.

I also thank to my brothers, Carolina and Víctor, and to my brother-in-law, Manuel, for his pieces of advice every time I needed an orientation. Thanks to all of you I made the right decisions in my life.

I could not forget to mention my ex teacher from the college, Flavio Espinoza. Thank you to teach me the passion for mathematics and science since I was just a child.

I also thank my teacher and advisor from the University of Valparaíso, Víctor Leiva, and the guideline provided by Camilo Lillo and Marcelo Rodríguez. Thank you for showing me the path of an appropriate research, for your invaluable help in this work, as well as the offered knowledge and opportunities. I am eternally grateful to all of you.

To Dr. Marco Riquelme and every member of the Institute of Statistics of the University of Valparaíso for their academic support and training.

To my childhood friends, Gonzalo Vilches, Claudio Jiménez and Marcelo Fernández, for permanently supporting me. Thank you for crossing you in my life and staying with me always.

Abstract

Partial least squares (PLS) regression is a multivariate technique developed to solve the problem of multicollinearity and/or high dimensionality related to explanatory variables in multiple linear regression. PLS regression has been widely applied assuming normality, but this assumption is often violated in different practical problems. Particularly, if the response variable follows an asymmetric distribution or it is bounded into an interval, normality should be discarded. For example, if this response variable is restricted to values between zero and one, a beta distribution is more suitable for PLS modeling than the normal distribution. We consider a beta PLS regression and its diagnostics for modeling the proportion of kaolinite, a clay mineral present in rocks which is measured by infrared spectroscopy with wavelengths. We propose a residual used in the generalized additive models for location scale and shape and the Cook and Mahalanobis distances as diagnostic tools for this model. We illustrate the proposed methodology with real-world mining data. The analyses and results provided in this study based on the beta PLS regression model and its diagnostics may be of interest for the Chilean mining sector and for the world mining industry.

List of Symbols and Abbreviations

In this chapter, we define symbols and abbreviations used in the present work.

α	Significance level.
a	Shape parameter of the beta distribution.
$\underline{a}$	Vector of auxiliary regression coefficients.
AIC	Akaike information criteria.
ASD	Autosaved or autorecover.
$\underline{\beta}$	Vector of regression coefficients related to parameters to be estimated.
b	Shape parameter of beta distribution.
BIC	Bayesian information criteria proposed by Schwarz.
c	Optimal number of partial least squares components.
CDF	Cumulative distribution function.
CK	Coefficient of kurtosis.
$\mathrm{Corr}(\cdot,\cdot)$	Correlation coefficient.
$\mathrm{Cov}(\cdot,\cdot)$	Covariance operator.
CS	Coefficient of skewness.
CV	Coefficient of variation.
δ	Precision parameter of the reparameterized beta distribution.
d_i	Defined in (3.2) and (3.3).
η_i	Defined in (2.5).
$\underline{\varepsilon}$	Vector of model errors.
$\mathrm{E}(\cdot)$	Expectation operator.
ETL	Extract, transform and load.
$f(\cdot)$	Probability density function.
$F(\cdot)$	Cumulative distribution function.
$\Gamma(\cdot)$	Gamma function.
$g(\cdot)$	Link function.
GLM	Generalized linear models.
k	Number of unknown model parameters.
KDD	Knowledge discovering in databases
KS	Kolmogorov-Smirnov.
$\ell(\underline{\theta})$	Log-likelihood function.
μ	Mean parameter of the reparameterized beta distribution.
m_i	Defined in (3.4).
MD	Median.
ML	Maximum likelihood.
n	Size sample.

NIR	Near infrared.
OLS	Ordinary least squares.
$\Phi(\cdot)$	Standard normal CDF.
p	Number of explanatory variables (or covariates).
$\underline{p}$	OLS coefficients of regression defined in (2.14).
PCA	Principal component analysis.
PDF	Probability density function.
PLS	Partial least squares.
PP	Probability versus probability.
PRESS	Prediction error sum of squared.
$\underline{q}$	Loading vector.
QQ	Quantile versus quantile.
r_i^{RQ}	Randomized quantile residual.
RC	Relative change.
RQ	Randomized quantile.
Rbeta	Reparameterized beta.
$\mathrm{rk}(\boldsymbol{X})$	Rank of $\boldsymbol{X}$.
$\boldsymbol{\varsigma}$	Model error matrix.
$\boldsymbol{S}_T^{-1}$	Estimate of the variance-covariance matrix of $\boldsymbol{T}$.
$\boldsymbol{S}_X^{-1}$	Estimate of the variance-covariance matrix of $\boldsymbol{X}$.
S_{X_j}	Standard deviation of X_j.
S_Y	Standard deviation of $\underline{Y}$.
SD	Standard deviation.
SE	Standard error.
$\underline{\theta}$	Parameter vector to estimate.
$\boldsymbol{T}$	Matrix of PLS components.
$\underline{\bar{T}}$	Vector of means (centroid) of $\boldsymbol{T}$.
$\mathrm{Var}(\cdot)$	Variance operator.
$\boldsymbol{W}$	Weight matrix.
$\boldsymbol{X}$	Design matrix.
$\boldsymbol{X}^\star$	Defined in (2.9).
$\underline{X}$	Row vector of covariates.
$\underline{\bar{X}}$	Vector of means (centroid) of $\boldsymbol{X}$.
$\underline{x}$	Observations of $\underline{X}$.
Y	Beta distributed random variable.
$\underline{Y}$	Vector of independent response variables with Rbeta distribution.
$\hat{Y}_{i(i)}$	Estimate of Y_i without the case i.
$\underline{y}$	Observations of $\underline{Y}$.
$y_{(1)}$	Observed minimum value of $\underline{Y}$.
$y_{(n)}$	Observed maximum value of $\underline{Y}$.

Table of Contents

List of Figures

List of Tables

Objectives

The objectives of this study are the following.

Main objective:

To formulate a PLS beta regression and propose diagnostics methods.

Specific objectives:

1. To predict the proportion of kaolinite using NIR measurements by a beta PLS regression model.
2. To propose diagnostic tools to evaluate the performance of the beta PLS regression model.
3. To apply the proposed methodology based on beta PLS regression and diagnostics to real-world mining data.
4. To compare the beta PLS regression model to different existing models in the literature for predicting the proportion of kaolinite using NIR spectral data.

Chapter 1

Introduction

Partial least squares (PLS) regression is a multivariate technique that was developed for modeling with multicolinearity in explanatory (independent) variables (covariates) and/or when the number of parameters in the model is greater than the number of cases; see Hair et al. (2014). The PLS regression model combines multiple regression and principal component analysis (PCA). PLS is a method that reduces the amount of covariates to a smaller set of uncorrelated components, called PLS components (latent factors). The amount of PLS components allows us to maximize the covariance between the response (dependent) variables and covariates. In order to calculate the amount of components PLS, an optimization method is performed in which ordinary least squares (OLS) are calculated.

PLS regression was introduced by Wold (1975), which shows that if as many components PLS as covariates are selected, the estimated regression coefficients and their standard errors (SEs) are the same as OLS regression coefficients. PLS regression is widely used in bioinformatics, biology, chemometrics and medicine; see Wold et al. (2001), Fernandez et al. (2008), Land et al. (2011) and Bertrand et al. (2013). An extension of the PLS regression was proposed by Marx (1996) based on generalized linear models (GLM); see McCullagh and Nelder (1989) and Bastien et al. (2005). This model retains the logic of PLS regression, where the amount of PLS components is obtained by using the maximum likelihood (ML) estimation method. However, the method maintains the PLS regression methodology maximizing covariance between the covariates and responses.

The beta distribution has been widely use for modeling proportions or percentages. This distribution has good properties, is flexible and can be used in a wide range of applications; see, for example, Johnson et al. (1995). Bertrand et al. (2013) proposed an extension of the PLS regression based on the normal distribution for the beta distribution based on PLS-GLM. In order to optimally find the number of PLS components, either for the PLS regression model or for PLS-GLM, some methods focus on the prediction error sum of squared (PRESS), which is based on cross-validation; see Li et al. (2002). Other methods of selection in PLS-GLM focus on information criteria, which are based on the ML method; see Bastien et al. (2005).

Diagnostic analyses are a necessary stage in all statistical modeling once the estimation of parameters has been conducted. Such analyses are carried out to assess the suitability of the distributional assumptions and the sensitivity and stability of the parameter estimation. Diagnostics can be conducted by residuals and global influence methods, such as Cook and Mahalanobis distances. Residuals allow us to detect the distributional assumptions and to reveal extreme cases. Global influence methods remove cases and evaluate their effect on the fitted model globally; see Cook and Weisberg (1982) and Chatterjee and Hadi (1988). For the use of residuals and diagnostic methods in non-normal models, including beta regression, see, for example, Espinheira et al. (2008), Ferrari et al. (2011),

Paula et al. (2012), Leiva et al. (2014a,0), Garcia-Papani et al. (2016), Marchant et al. (2016), Santos-Neto et al. (2016), and Leao et al. (2017).

Kaolinite is a clay mineral, which is part of the group of industrial silicate minerals, collected in mining; see Orbovic and Huang (2012). Kaolinite is present in rock samples obtained by batch, which are measured by spectrometry near infrared (NIR) with wavelengths between 350 nm and 2500 nm, with 200 scans per sample in 20 seconds; see Burns and Ciurczak (2007). Data related to kaolinite are often measured between zero and one as a proportion. Then, these data can be adequately described by the beta distribution. However, as mentioned, proportion of kaolinite is explained by spectral variables, which usually correspond to a high number of covariates many of them correlated. Then, a beta PLS regression is suitable to model the proportion of kaolinite in function of spectral variables.

The remainder of this thesis presents the following structure: In Chapter 2, modeling aspects are is introduced. In Chapter 3, diagnostic methods to be proposed are derived, whereas in Chapter 4 the results on the modeling and diagnostics of the proportion of kaolinite are presented. Chapter 5 corresponds to the conclusions of this work.

Chapter 2

Background

In this chapter, we provide some preliminary concepts on beta regression and beta PLS regression model, which are useful to understand our methodology.

2.1 Beta regression model

The beta regression model is based on a reparameterization of the beta distribution; see Ferrari and Cribari-Neto (2004). The usual parameterization of the beta distribution (see Kotz and van Dorp, 2004) is denoted by $Y \sim \text{Beta}(a, b)$, with probability density function (PDF) given by

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \quad 0 < y < 1, \quad a > 0, \quad b > 0, \tag{2.1}$$

where Γ is the usual gamma function. Then mean and variance of Y are expressed as

$$\text{E}(Y) = \frac{a}{a+b}, \quad \text{Var}(Y) = \frac{ab}{(a+b+1)(a+b)^2}, \tag{2.2}$$

respectively. Ferrari and Cribari-Neto (2004) proposed the a reparameterization of the beta distribution based on its mean defined in (2.2), that we denote now by μ, and a precision parameter δ both given by

$$\mu = \frac{a}{a+b}, \quad \delta = (a+b). \tag{2.3}$$

From (2.3), note that $a = \mu\delta$ and $b = (1-\mu)\delta$. By using the reparameterization defined in (2.3), denoted here by $Y \sim \text{Rbeta}(\mu, \delta)$, we have that the mean and variance of Y are now expressed as

$$\text{E}(Y) = \mu, \quad \text{Var}(Y) = \frac{\mu(1-\mu)}{1+\delta}, \tag{2.4}$$

respectively. The PDF of Y given in (2.1) can be written, in this reparameterization, as

$$f(y; \mu, \delta) = \frac{\Gamma(\delta)}{\Gamma(\mu\delta)\Gamma((1-\mu)\delta)} y^{\mu\delta-1}(1-y)^{(1-\mu)\delta-1}, \quad 0 < y < 1, \quad 0 < \mu < 1, \quad \delta > 0.$$

Note that the variance of Y given in (2.4) is a function of μ. Therefore, such as the gamma and reparameterized Birnbaum-Saunders distributions (see Johnson et al., 1994; Leiva et al., 2014b), the reparameterized beta distribution allows us to model heteroscedasticity. Under this reparameterization, we can apply techniques based on GLM and the beta distribution.

Let $\underline{Y} = (Y_1, \ldots, Y_n)^\top$ be a sample (independent random variables) of a response $Y \sim \text{Rbeta}(\mu, \delta)$ distribution, and $\underline{y} = (y_1, \ldots, y_n)^\top$ a column vector with the observations

of $\underline{Y}$. In addition, let $\underline{X} = (X_1, \ldots, X_p)$ be a row vector of covariates whose observations are denoted by $\underline{x} = (x_1, \ldots, x_p)$. Then, a beta regression model based on GLM can be written as

$$\mathrm{E}(Y_i) = \eta_i = g(\mu_i) = \underline{x}_i^\top \underline{\beta}, \quad i = 1, \ldots, n, \tag{2.5}$$

where Y_i is the response, $\underline{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$ are the observations of covariates corresponding to the ith row vector of the design matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \tag{2.6}$$

$\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is a vector of unknown parameters to be estimated and g is a link function. Note that this function must be twice differentiable and surjective on the interval $(0, 1)$. There are several possible choices for the link function g. For instance, one can use the logit specification $g(\mu) = \log(\mu/(1-\mu))$, the probit function $g(\mu) = \Phi^{-1}(\mu)$, where Φ is the cumulative distribution function (CDF) of a standard normal random variable, the complementary $\log - \log$ link $g(\mu) = \log(-\log(1-\mu))$, the $\log - \log$ link $g(\mu) = -\log(-\log(\mu))$, among others; see McCullagh and Nelder (1989) and Atkinson (1985). The parameter estimation of the model defined in (2.5) is typically performed by the ML method. Then, the log-likelihood function for $\underline{\theta} = (\underline{\beta}^\top, \delta)^\top$ based on n observations $\underline{y}$ from the beta regression defined in (2.5) is given by

$$\begin{aligned} \ell(\underline{\theta}) \;=\; & \sum_{i=1}^{n} \log(\Gamma(\delta)) - \sum_{i=1}^{n} \log(\Gamma(\mu_i\delta)) - \sum_{i=1}^{n} \log(\Gamma((1-\mu_i)\delta)) \\ & + \sum_{i=1}^{n} (\mu_i\delta - 1)\log(y_i) + \sum_{i=1}^{n} ((1-\mu_i)\delta - 1)\log(1-y_i), \end{aligned} \tag{2.7}$$

where $\mu_i = g^{-1}(\eta_i) = g^{-1}(\underline{x}_i^\top \underline{\beta})$. The ML estimates of $\underline{\beta}$ and δ, $\widehat{\underline{\beta}}$ and $\widehat{\delta}$ say, respectively, are the solution of the system of equations given by

$$\frac{\mathrm{d}\ell(\underline{\theta})}{\mathrm{d}\underline{\beta}} = \underline{0}, \quad \frac{\mathrm{d}\ell(\underline{\theta})}{\mathrm{d}\delta} = 0.$$

For more details about inference in beta regression, see Cribari-Neto and Queiroz (2012). Since the corresponding ML estimators cannot be expressed in a closed form, these are typically computed by maximizing the log-likelihood function numerically with some quasi-Newton algorithm; see more details on nonlinear optimization in Nocedal and Wright (1999).

2.2 PLS regression models

PLS regression is a technique alternative to OLS regression, which is particularly useful when the covariates are correlated –inducting a multicollinearity problem, which is frequent in chemometrical applications of regression models– and/or when the number of covariates is greater than the number of cases ($p > n$). As mentioned, PLS regression mixes PCA and multiple regression, acting similarly as in PCA regression; Jolliffe (2002). However, unlike PCA regression, where the latent factors maximize the variance of the covariates, PLS regression extracts a set of latent factors that maximizes the covariance

between the response and the covariates. Then, as in multiple linear regression, the main goal of the PLS regression is to construct a linear model

$$\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{\varepsilon}, \tag{2.8}$$

where $\underline{Y}$ is the response, $\boldsymbol{X}$ is a design matrix with the values of covariates, $\underline{\beta}$ is a coefficient vector to be estimated and $\underline{\varepsilon}$ is the model error vector. Before performing PLS regression, we must center the observed values of the covariates by subtracting their mean and then we must scale them dividing by their standard deviations, that is, we work with

$$\boldsymbol{X}^\star = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{S_{X_1}} & \dots & \frac{x_{1p}-\bar{x}_p}{S_{X_p}} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{S_{X_1}} & \dots & \frac{x_{np}-\bar{x}_p}{S_{X_p}} \end{pmatrix} = \begin{pmatrix} x^\star_{11} & \dots & x^\star_{1p} \\ \vdots & \ddots & \vdots \\ x^\star_{n1} & \dots & x^\star_{np} \end{pmatrix}. \tag{2.9}$$

Thus, PLS regression produces a matrix of factor scores as linear combinations of the centered and scaled original covariates defined in (2.9). This matrix is given by

$$\boldsymbol{T} = \boldsymbol{X}^\star\boldsymbol{W}, \tag{2.10}$$

where

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & \dots & t_{1c} \\ \vdots & \ddots & \vdots \\ t_{n1} & \dots & t_{nc} \end{pmatrix}, \quad \boldsymbol{W} = \begin{pmatrix} w_{11} & \dots & w_{1c} \\ \vdots & \ddots & \vdots \\ w_{p1} & \dots & w_{pc} \end{pmatrix},$$

are factor (or PLS components) and weight matrices, respectively. Note that $c \leq \mathrm{rk}(\boldsymbol{X}) \leq p$, with c being the optimal number of PLS components to be selected suitably, reducing the dimensionality; see Section 2.4. All coefficient of correlation between the columns of the PLS component matrix $\boldsymbol{T}$ is zero, doing the columns to be orthogonal, solving the multicollinearity. Thus, the PLS regression model, which connects the linear model defined in (2.8) and the PLS components, is given by

$$\underline{Y} \;=\; \boldsymbol{X}\underline{\beta} + \underline{\varepsilon} = \boldsymbol{X}^\star\boldsymbol{W}\underline{q} + \underline{\varepsilon} = \boldsymbol{T}\underline{q} + \underline{\varepsilon}, \tag{2.11}$$

where $\underline{Y}, \boldsymbol{X}, \underline{\beta}$ and $\underline{\varepsilon}$ are given in (2.8), $\boldsymbol{T}, \boldsymbol{X}^\star$ and $\boldsymbol{W}$ are given in (2.10) and $\underline{q} = (q_1, \dots, q_c)^\top$ is a loading vector, corresponding to the PLS coefficients to be estimated. Once $\underline{q}$ is estimated, the model defined in (2.11) can be used as a predictive regression model. In order to compute the first PLS component, $\underline{t}_1 = \boldsymbol{X}^\star\underline{w}_1$ say, the weight

$$\mathrm{Cov}(Y, X_j) = \mathrm{Corr}(Y, X_j) S_Y S_{X_j}, \quad j = 1, \dots, p, \tag{2.12}$$

for the covariate X_j, must be determined, where $\mathrm{Cov}(Y, X_j), \mathrm{Corr}(Y, X_j), S_Y$ and S_{X_j} are the covariance, correlation coefficient and standard deviation of Y and X_j, respectively. Then, the covariate X_j is relevant in constructing $\underline{t}_1$, if (X_j, Y) are highly correlated and have sufficient variability. Note that the weight $\mathrm{Cov}(Y, X_j)$ defined in (2.12) is also the regression coefficient a_{1j} in an OLS regression of Y on the modified covariate $X_j/\mathrm{Var}(X_j)$, that is,

$$a_{1j} = \frac{\mathrm{Cov}(Y, X_j/\mathrm{Var}(X_j))}{\mathrm{Var}(X_j/\mathrm{Var}(X_j))} = \frac{\mathrm{Cov}(Y, X_j)/\mathrm{Var}(X_j)}{\mathrm{Var}(X_j)/(\mathrm{Var}(X_j))^2} = \mathrm{Cov}(Y, X_j), \tag{2.13}$$

Then, before obtaining the second PLS component, $\underline{t}_2$ say, a further multivariate regression model can be generated as

$$\boldsymbol{X}^\star \;=\; \underline{t}_1\underline{p}^\top + \boldsymbol{\varsigma}, \tag{2.14}$$

where $\boldsymbol{X}^\star$ is given in (2.9), the row vector $\underline{t}_1$ is the first PLS component, $\underline{p}^\top = (p_1, \ldots, p_n)$ is a column vector, related to the OLS coefficients to be estimated, and $\boldsymbol{\varsigma}$ is the corresponding model error matrix. Iteratively, the second PLS component is obtained by means of the vector $\underline{a}_2 = (a_{21}, \ldots, a_{2p})^\top$ corresponding to the regression of Y on $(T_1, X_j^\star)$, and so on for the other components; see Bastien et al. (2005). For practical use, the second component $\underline{t}_2$ must be expressed in terms of the covariate $X_j^\star$. This is possible due to that, when obtaining $\underline{t}_2$, in addition to the regression of Y on $(T_1, X_j^\star)$, the regression of $X_j^\star$ on $(T_1, X_{1j}^\star)$ is computed, permitting the residuals $x_{1j}^\star = x_j^\star - \widehat{p}_{1j} t_1$, obtained from (2.14), to be a function of the value $x_j^\star$ of $X_j^\star$.

Consequently, maximizing the vectors $\underline{a}_1, \ldots, \underline{a}_c$ defined as in (2.13), the covariance between Y and X_js is maximized and then the PLS regression model is obtained, differently as in PCA regression, where, as mentioned, the latent factors maximize the variance of the covariates; for more details on parameter estimation methods and algorithms in PLS regression, see Geladi and Kowalski (1986). The PLS approach is summarized in Algorithm 1; see details in Marx (1996).

Algorithm 1 PLS approach

1: Start $\boldsymbol{X}_0^\star = \boldsymbol{X}^\star$ and $\underline{y}_0 = \underline{y}$.
2: Carry out a for cycle from $h = 1$ to c following the steps:
   2.1 Obtain the direction of maximum covariance: $\underline{w}_h = \boldsymbol{X}_{h-1}^{\star\top} \underline{y}_{h-1}$.
   2.2 Project $\boldsymbol{X}^\star$ onto $\underline{w}_h$: $\boldsymbol{X}_{h-1}^\star \underline{w}_h$.
   2.2 Generate the hth normalized PLS component: $\underline{t}_h = (1/||\underline{w}_h||)\boldsymbol{X}_{h-1}^\star \underline{w}_h$.
   2.3 Establish the residual matrix: $\boldsymbol{X}_h^\star = \boldsymbol{X}_{h-1}^\star - \underline{t}_h \underline{p}_h^\top$.
3: End the "for cycle" obtaining $\boldsymbol{T}$ and $\boldsymbol{W}$.
4: Compute the PLS $\underline{\beta}$ coefficients using latent factors: $\underline{\beta} = \boldsymbol{W}(\boldsymbol{T}^\top \boldsymbol{X}^\star \boldsymbol{W})^{-1} \underline{y}$.

2.3 Beta PLS regression models

A beta PLS regression model can be obtained from equations (2.5) and (2.11) as

$$\mathrm{E}(Y_i) = \eta_i = g(\mu_i) = \underline{t}_i^\top \underline{q} = \underline{x}_i^{\star\top} \boldsymbol{W} \underline{q} = \underline{x}_i^\top \underline{\beta}, \quad i = 1, \ldots, n, \tag{2.15}$$

where $\underline{t}_i^\top = (t_{i1}, \ldots, t_{ic})$ is the ith row of $\boldsymbol{T}$, g, $\underline{x}_i^\top, \underline{\beta}$ are defined in (2.4) and $\underline{x}_i^\top$ is the ith row of $\boldsymbol{X}$ given in (2.6). Algorithm 2 allows us to determine the column factor vectors of $\boldsymbol{T}$, corresponding to the PLS components, $\underline{t}_h = (t_{1h}, \ldots, t_{nh})^\top$ say, for $h = 1, \ldots, c$. Then, the predictive beta PLS model is obtained; see Bertrand et al. (2013). In this algorithm, note that the coefficients a_{1j} defined in (2.13) can no longer be obtained in an analytical closed form but numerically.

2.4 Determining the number of PLS components

Since the beta PLS model is based on the ML estimation, model selection methods as the Akaike information (AIC) and Schwarz Bayesian (BIC) criteria may be employed. A smaller value of AIC (or BIC) indicates a better model. Li et al. (2002) mentioned that fixing the number of components in the first local minimum corresponds to the statistical model that best fits the data. The global minimum value should not be selected because

Algorithm 2 Getting a predictive beta PLS model

1: Calculate the first PLS component $\underline{t}_1$ following the steps:

1.1 Estimate the coefficient a_{1j} of $x_j^\star$ in the beta PLS regression of Y on $X_j^\star$, using (2.7) and (2.15), where $\mu_i = g^{-1}(a_{1j}x_{ji}^\star)$, with $i = 1, \ldots, n$, repeating this step for $j = 1, \ldots, p$, constructing so the vector $\underline{a}_1 = (a_{11}, \ldots, a_{1p})^\top$.

1.2 Determine the first column vector of $\boldsymbol{W}$ as $\underline{w}_1 = (1/||\underline{a}_1||)\underline{a}_1$.

1.3 Compute the first PLS component $\underline{t}_1 = (1/(\underline{w}_1^\top \underline{w}_1))\boldsymbol{X}^\star \underline{w}_1$.

2: Calculate the second PLS component $\underline{t}_2$ following the steps:

2.1 Estimate the coefficient a_{2j} of $x_j^\star$ in the beta PLS regression of Y on T_1 and $X_j^\star$, using (2.7) and (2.15), where $\mu_i = g^{-1}(a_{1j}t_{1i} + a_{2j}x_{ji}^\star)$, with $i = 1, \ldots, n$, repeating this step for $j = 1, \ldots, p$, constructing so the vector $\underline{a}_2 = (a_{21}, \ldots, a_{2p})^\top$.

2.2 Determine the second column vector of $\boldsymbol{W}$ as $\underline{w}_2 = (1/||\underline{a}_2||)\underline{a}_2$.

2.3 Establish the residual matrix $\boldsymbol{X}_1 = \boldsymbol{X}^\star - \underline{t}_1\underline{p}_1^\top$ of the OLS regression of $\underline{X}^\star$ on t_1.

2.4 Compute the second component $\underline{t}_2 = (1/(\underline{w}_2^\top \underline{w}_2))\boldsymbol{X}_1 \underline{w}_2$.

2.5 Express the component $\underline{t}_2$ in terms of $\boldsymbol{X}^\star$ as $\underline{t}_2 = \boldsymbol{X}^\star \underline{w}_2$.

3: State the optimal number $c < p$ of PLS components and calculate the other PLS components $\underline{t}_h$, for $h = 3, \ldots, c$, following the steps:

3.1 Estimate the coefficient a_{hj} of $x_j^\star$ in the beta PLS regression of Y on $T_1, \ldots, T_{h-1}$ and $X_j^\star$, using (2.7) and (2.15), where $\mu_i = g^{-1}(a_{1j}t_{1i} + \cdots + a_{(h-1)j}t_{(h-1)i} + a_{hj}x_{ji}^\star)$, with $i = 1, \ldots, n$, repeating this step for $j = 1, \ldots, c$, constructing so the vector $\underline{a}_h = (a_{h1}, \ldots, a_{hc})^\top$.

3.2 Determine the hth column vector of $\boldsymbol{W}$ as $\underline{w}_h = (1/||\underline{a}_h||)\underline{a}_h$.

3.3 Establish the residual matrix $\boldsymbol{X}_{h-1} = \boldsymbol{X}_{h-2} - \underline{t}_{h-1}\underline{p}_{h-1}^\top$ of the OLS regression of $\underline{X}^\star$ on $T_1, \ldots, T_{h-1}$.

3.4 Compute the hth component $\underline{t}_h = (1/((\underline{w}_h^\top \underline{w}_h))\boldsymbol{X}_{h-1}\underline{w}_h$.

3.5 Express the component $\underline{t}_h$ in terms of $\boldsymbol{X}^\star$ as $\underline{t}_h = \boldsymbol{X}^\star \underline{w}_h$.

4: Estimate the $\underline{q}$ loading vector, $\widehat{\underline{q}} = (\widehat{q}_1, \ldots, \widehat{q}_c)^\top$ say, based on: (i) the c PLS components, $\underline{t}_1, \ldots, \underline{t}_c$ say; (ii) the beta PLS model given in (2.15); and (iii) the log-likelihood function defined in (2.7).

5: Predict values for the response Y with the estimated beta PLS model given by $\widehat{y}_{\text{pred}} = \underline{x}_{\text{pred}}^\top \widehat{\underline{\beta}}$, where $\underline{x}_{\text{pred}}$ is a vector with the values of the covariates to be used for predicting y and $\widehat{\underline{\beta}}$ is the ML estimate of PLS $\underline{\beta}$ coefficients obtained from (2.15) as

$$\widehat{\underline{\beta}} = \boldsymbol{W}\widehat{\underline{q}}.$$

the information criteria will always be smaller as the number of components increases, concluding that a better model would have as components as variables; see Osten (1988). AIC and BIC are given, respectively, by

$$\text{AIC} \;=\; -2\ell(\widehat{\underline{\theta}}) + 2k, \quad \text{BIC} = -2\ell(\widehat{\underline{\theta}}) + k\log(n),$$

where $\ell(\widehat{\underline{\theta}})$ is the log-likelihood function of the model with parameter vector $\underline{\theta}$ evaluated $\underline{\theta} = \widehat{\underline{\theta}}$, n is to sample size and k is the number of unknown model parameters. AIC and BIC are functions of the log-likelihood, plus a component penalizing it.

When it is not possible to choose a local minimum value by using some information criteria, in order to select the number of PLS components, the PRESS method is recommended, which is a cross-validation method usually used in PLS regression to select the number of components. PRESS assesses the effect of removing a case in different modeling scenarios (different number of components). Thus, we proceed until that the n cases have been left out once, repeating the process until completing the c components and the a comparison must be carried out for choosing the number of components. In order to calculate the PRESS value for the component h, the expression

$$\text{PRESS}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{i(i)})^2, \quad h = 1, \ldots, c, \tag{2.16}$$

is used, where $\widehat{Y}_{i(i)}$ represents the estimate of Y_i with no the case i in the training model, such as a Jacknife estimate. To select the number of PLS components using PRESS, you can use a hypothesis test based on the F-statistic (see Osten, 1988) given by

$$F = \frac{(n - (h-1))(\text{PRESS}(h) - \text{PRESS}(h+1))}{\text{PRESS}(h+1)}, \tag{2.17}$$

where PRESS(h) is defined in (2.16), h is the number of PLS components and n is the sample size. The statistic defined in (2.17) is compared to a $(1-\alpha) \times 100$th percentile of the F distribution given by $F_{1-\alpha}(1; h, n-h+1)$; see Li et al. (2002).

Chapter 3

Diagnostics

In this chapter, we present diagnostic methods for the beta PLS regression. These methods correspond to residual analysis and Cook and Mahalanobis distance.

3.1 Residual analysis

In order to assess possible departures from the assumption of the model error, as well as to detect atypical observations, several types of residuals have been a very useful tool in regression models; see Leiva et al. (2016). However, to find residuals whose empirical distribution is close to normality in non-normal regression models is not an easy task.

Ferrari and Cribari-Neto (2004) used a standardized residual for the beta regression model. We propose a type of residual used in the generalized additive models for location scale and shape; see Dunn and Smyth (1996) and Stasinopoulos and Rigby (2007). This is the randomized quantile (RQ) residual defined by

$$r_i^{\mathrm{RQ}} = \Phi^{-1}\big(F(y_i; g^{-1}(\underline{t}_i\widehat{\underline{q}})\widehat{\delta}, (1 - g^{-1}(\underline{t}_i\widehat{\underline{q}}))\widehat{\delta})\big), \tag{3.1}$$

where $F(y;\mu,\delta) = \int_0^y f(u;\mu,\delta)\mathrm{d}u$ and Φ^{-1} is the inverse function of the standard normal cumulative distribution function (quantile function), with $f(y;\mu,\delta)$ being defined in (2.1), but based on the reparameterization given in (2.3). Note that the RQ residual defined in (3.1) must be compared with the normal distribution to evaluate its fitting to the data.

We propose to use a theoretical probability versus empirical probability (PP) plot to do this evaluation. In addition, note that the PP plot can be linked to the Kolmogorov-Smirnov (KS) test, by means of which acceptance bands may be constructed inside of this plot. Algorithm 3 summarizes this construction; see Leiva and Saunders (2015), Castro-Kuriss et al. (2016) and Lillo et al. (2016). Also, we use a theoretical quantile versus empirical quantile (QQ) plot as contrast based on the plot of simulated envelopes proposed by Atkinson (1985).

3.2 Cook distance

The Cook distance is useful to identify atypical cases measuring the effect on the estimated beta PLS coefficients, $\underline{q}$ say, by eliminating a case; see Cook (1977). In normal linear regression models, the influence of the case i can be measured by the classic Cook distance. A measure of influence equivalent to the classic Cook distance for PLS regression can be considered as

$$d_i = \frac{2}{c}(\ell(\widehat{\underline{\theta}}) - \ell(\widehat{\underline{\theta}}_{(i)})), \quad i = 1, \ldots, n, \tag{3.2}$$

Algorithm 3 Goodness of fit to any distribution.

1: Consider data $d_1, \ldots, d_n$ and order them as $d_{1:n}, \ldots, d_{n:n}$.
2: Estimate parameters $\underline{\theta}$ of the distribution by $\widehat{\underline{\theta}}$ with $d_1, \ldots, d_n$ and the ML method.
3: Compute $\widehat{v}_{j:n} = F(d_{j:n}; \widehat{\underline{\theta}})$, for $j = 1, \ldots, n$, with F being the corresponding cumulative distribution function.
4: Calculate $\widehat{y}_j = \Phi^{-1}(\widehat{v}_{j:n})$.
5: Obtain $\widehat{u}_{j:n} = \Phi(\widehat{z}_j)$, with $\widehat{z}_j = (\widehat{y}_j - \bar{y})/s_y$, $\bar{y} = \sum_{j=1}^{n} \widehat{y}_j/n$ and $s_y = (\sum_{j=1}^{n} (\widehat{y}_j - \bar{y})^2/(n-1))^{1/2}$.
6: Draw the PP plot with points $w_{j:n} = (j - 0.5)/n$ versus $\widehat{u}_{j:n}$, for $j = 1, \ldots, n$.
7: Specify a significance level α.
8: Construct acceptance bands according to $(\max\{w - d_{1-\alpha} + 0.5/n, 0\}, \min\{w + d_{1-\alpha} - 0.5/n, 1\})$, where $d_{1-\alpha}$ is the $(1-\alpha)\times 100$th percentile of the KS distribution (adapted) and w is a continuous version of $w_{j:n}$.
9: Determine the p-value of the KS statistic and reject the null hypothesis of the corresponding distribution for the specified significance level α based on this p-value.
10: Corroborate coherence between steps 8 and 9.

where c is the number of PLS components, $\ell(\underline{\theta})$ is the corresponding log-likelihood function and $\widehat{\underline{\theta}}_{(i)}$ is an estimation of the PLS coefficients without considering the case i. According to Williams (1987), the generalized Cook distance proposed in (3.2) can be approximated by

$$d_i = \frac{h_i (r_i^{\mathrm{RQ}})^2}{c(1-h_i)}, \quad i = 1, \ldots, n, \tag{3.3}$$

where r_i^{RQ} is the RQ residual defined in (3.1) and h_i is a diagonal element of the projection matrix. We consider the case i as influential if $d_i > 4/(n-c-1)$; for more information about this cut-off, see Fox (1991).

3.3 Mahalanobis distance

The Mahalanobis distance is the length between an observation and the centroid of a multivariate space (global average). This distance is used to identify outliers and given by $m_i = \sqrt{(\underline{X}_i - \bar{\underline{X}})^\top \boldsymbol{S}_X^{-1} (\underline{X}_i - \bar{\underline{X}})}$, for $i = 1, \ldots, n$, where $\underline{X}_i$ is a random vector, $\overline{\underline{X}}$ is a vector of mean (centroid) and $\boldsymbol{S}_X$ is an estimate of the variance-covariance matrix of $\boldsymbol{X}$. Since it is not possible to use the Mahalanobis distance as a criterion for detecting outliers when there are more variables that cases, the calculation of the Mahalanobis distance in PLS regression is based on the PLS component matrix, $\boldsymbol{T}$ say; see Varmuza and Filzmoser (2009). Thus, the Mahalanobis distance based on c components is defined by

$$m_i = \sqrt{(\underline{T}_i - \overline{\underline{T}})^\top \boldsymbol{S}_T^{-1} (\underline{T}_i - \overline{\underline{T}})}, \quad i = 1, \ldots, n, \tag{3.4}$$

where $\underline{T}_i$ is a random vector, $\overline{\underline{T}}$ is a vector of mean (centroid) and $\boldsymbol{S}_T$ is an estimate of the of variance-covariance matrix of the PLS component estimators. If the value of m_i is greater than the percentile $\chi^2_{1-\alpha}(c)$, then the case i is atypical.

Chapter 4

Application

This chapter provides an application of the methodology proposed in this work based on beta PLS modeling and diagnostics to kaolinite data used in mining.

4.1 Software and data

`R` is an open source and non-commercial software for statistical analysis and graphs. This software may be obtained freely at `http://www.r-project.org`. We implement a computational framework in `R` to analyze data with the methodology of modeling and diagnostics proposed in the thesis. The `R` codes employed in this application are available from the authors upon request. The NIR spectrometry data is loaded using the `get_spectra` command from the asdreader library; see Roudier (2016). The beta distribution can be symmetric or asymmetric depending on the values of its parameters. We use an adjusted boxplot for data following asymmetric distributions, which is implemented in `R` by the `adjbox` command of the `robustbase` package; see Rousseeuw et al. (2016). The beta PLS regression model is implemented in the `plsRbeta` package and may be performed by a command of the same name; see Bertrand et al. (2014a). This package uses the `betareg` package to estimate the regression parameters; see Cribari-Neto (2010). The `plsRbeta` package is based on the `plsRglm` package, which allows a distributional comparison of the beta PLS model with distributions of the exponential family (for example, the gamma distribution); see Bertrand et al. (2014b). For the diagnostic analysis, the Cook distance is implemented in the `base` package through the `cook.distance` command, whereas the Mahalanobis distance is implemented in the package `chemometrics` by the `Moutlier` command; see Filzmoser (2016).

We have 99 observations of the proportion of kaolinite (Y), which we call "kaolinite data" obtained by an anonymous mining company, for which were measured the reflectance through a process of NIR between 350 nm and 2500 nm. NIR spectrometry is one of the most widely used techniques in the identification of chemical compounds and in the determination of molecular structures. The spectra in the NIR are related to the changes that occur between the energy states of the molecules, so that the energy involved in this type of molecular transitions is represented in the region of the electromagnetic spectrum. For a molecule to absorb infrared radiation it must undergo a change in vibrational or rotational motion, which is measured and represented across a spectrum. One of the main advantages of NIR spectrometry is its low cost and short time to determine the presence of certain minerals in rock samples in the mining area compared to other techniques used to perform the same process, see Appendix A for more details. The aim is to predict the proportion of kaolinite based on reflectance measurements corresponding to spectral data ($\boldsymbol{X}$), which call "spectra data".

4.2 Business intelligence

The correct estimation of the proportion of kaolinite is very important in the business of mineral exploitation, because based on this estimate the decision is made to excavate or not. Analyzing rocks obtained in samples of minerals by using spectrometry is considerably less expensive and faster than the laboratory analyses. Business intelligence is summarized as the correct transformation of data into information, in order to make an appropriate decision.

"Business intelligence" begins (first stage) with the identification of the databases from which the structured data set will be built to be analyzed and stored in the "data warehouse". Second, it is the way of constituting the data set (stored in the data warehouse) which will be used to carry out the analytics (data processing) and to generate knowledge discovering in databases (KDD). Then, once the databases are identified, the data must be extracted, transformed and loaded in order to be analyzed. Thus, the next step of business intelligence (second stage) requires the use of a concept known as ETL (extract, transform and load). ETL corresponds to the process that will allow the organizations to move data from multiple sources, reformat, clean or purge and load them into the data warehouse to form the data set to be analyzed. Then, the third stage of business intelligence is the constitution of the data warehouse. In this way, once the data warehouse is constituted, the fourth stage is to perform the data analytic. This concept corresponds to a generic term used in the context of business intelligence and it refers to the techniques that allow us to process and analyze structured data, which permits them to be transformed into information. The fifth and final stage of business intelligence is to make decisions that support the business process based on the information obtained from the data and the experience of the decision makers. Figure 4.1 illustrates the steps of business intelligence. It is important to consider a correct use of the data by verifying that they are clean and with the necessary record for an adequate analysis. Failure to perform this process can lead to obtaining results totally out of the reality.

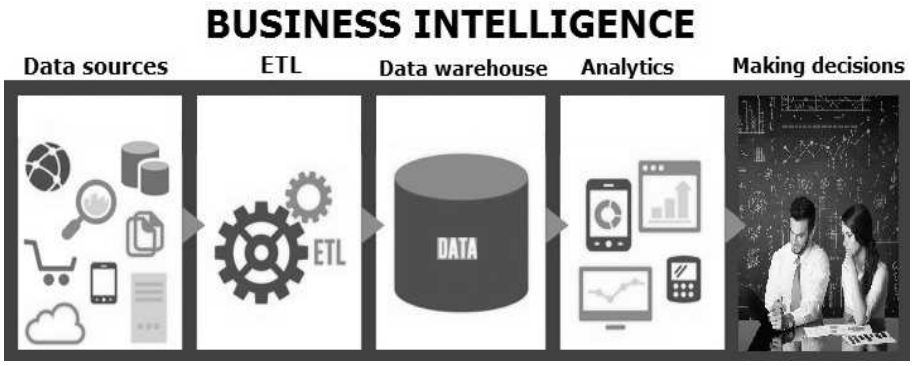

Figure 4.1: Scheme of business intelligence.

There are two different data sources for the 99 observations under study. The first one contains data on the real proportion of certain minerals in the samples of rocks, among which is kaolinite, alunite, biotite, gypsum, among others, in addition to a variable corresponding to an identifier of each one of the samples. The second data source contains the reflectance measurements of the 2151 different radiation waves obtained through NIR spectrometry, stored in 99 different files –one for each sample– in autosaved or autorecover (ASD) files, a special format delivered by the spectroscope. The name of each of these files corresponds to the identifier of each sample.

Once the data sources are identified, the ETL process carried out in this work consisted of extracting the data from each of the NIR spectroscopy samples by storing them in a database, and linking them to their corresponding mineral proportion measurements. Because the objective of this application is to predict the proportion of kaolinite, those variables corresponding to other types of minerals were not considered, thus obtaining our final data set to be analyzed in the remainder of this chapter. The Appendix B contains the executed routine to perform this ETL process.

4.3 Exploratory data analysis

We performance an exploratory analysis kaolinite data and their spectra to formulate a suitable model to describe them. Figure 4.2 (left) shows a histogram of the kaolinite proportion. In addition, Figure 4.2 (right) displays classical and adjusted box-plot for asymmetric data; see Rousseeuw et al. (2015). Furthermore, Table 4.1 presents a descriptive summary which includes the minimum value $y_{(1)}$, median (MD), mean $\bar{y}$, maximum value $y_{(n)}$, standard deviation (SD), and coefficients of variation (CV), skewness (CS) and kurtosis (CK) for kaolinite data (Y). From Figure 4.2 and Table 4.1, note that Y seems to follow a right asymmetric distribution. From the box-plots, we observe some outliers. Figure 4.3 represents 99 spectral data, with each line corresponding to 2151 covariates. Therefore, based on the nature of the phenomenon, the number of covariates and descriptive statistics, the beta PLS model seems to be a suitable model to predict the proportion of kaolinite.

Table 4.1: Descriptive statistics for kaolinite data.

n	$y_{(1)}$	MD	$\overline{y}$	$y_{(n)}$	SD	CV	CS	CK
99	0.0071	0.0820	0.1032	0.3803	0.0780	0.7560	1.3382	4.5088

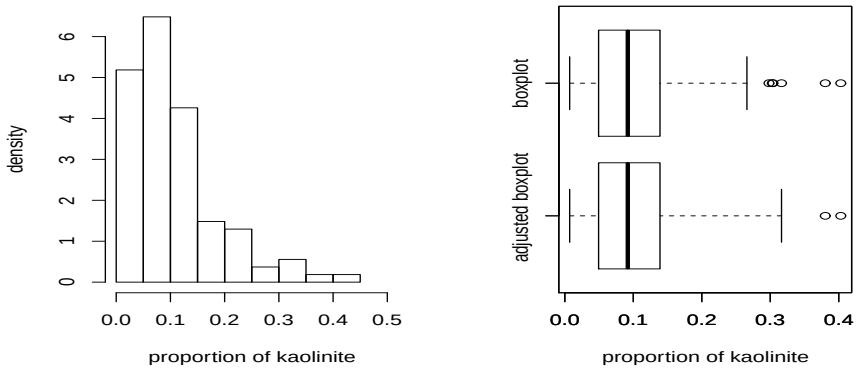

Figure 4.2: Histogram (left) and boxplots (right) for kaolinite data.

4.4 Modelling, estimation and inference

In order to establish whether the beta PLS model is appropriate to the kaolinite data, to determine the link function which must be used, and to choose the PLS component number, a comparison of different models and link functions is performed. The models to be compared are:

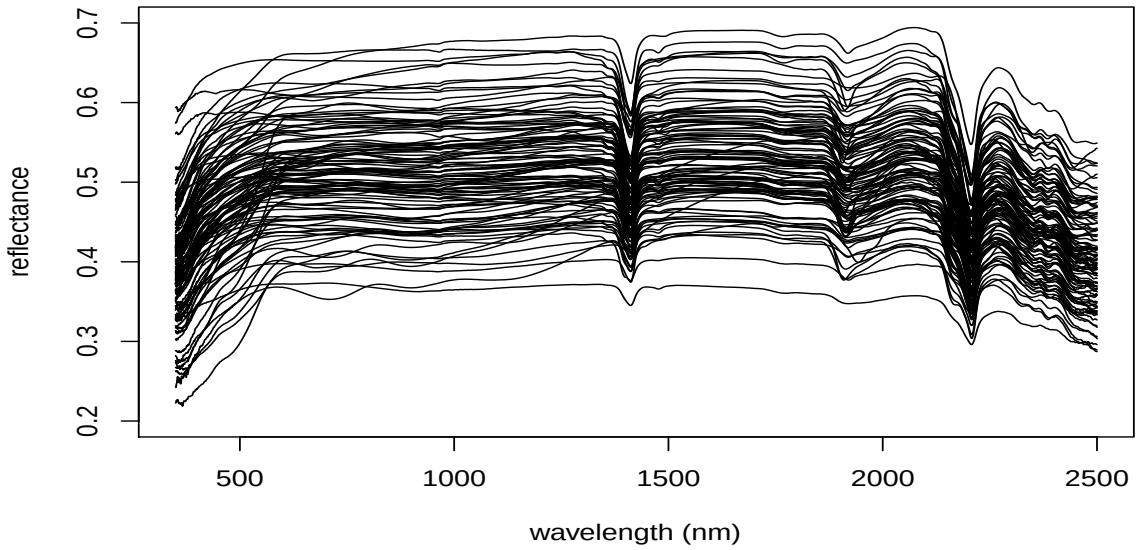

Figure 4.3: Plot of spectra for covariates related to kaolinite data.

- Beta PLS model with logit link function.
- Beta PLS model with logarithmic link function.
- Gamma PLS model with logarithmic link function.
- Gamma PLS model with identity link function.
- Normal (classical) PLS model with identity link function.
- Normal PLS model with logarithmic link function.

Figure 4.4 shows AIC (left) and BIC (right) values defined in (2.16) for different schemes of modeling (from 1 to 30 components) based on the mentioned PLS models. From Figure 4.4, note that the criterion is not conclusive, since no local minimum is detected by the AIC or BIC for the number of components. However, note that beta PLS model with logit link function seems to be the most appropriate model, regardless of the number of PLS components used for the model. As a local minimum is not detected by the ACI and BIC in Figure 4.4, the value of PRESS is used to obtain the optimum number of components to be used. Selection of the component numbers is formulated with the p-value criterion based on F-statistic. Figure 4.5 displays the PRESS values for beta PLS model, where a white circle shows the first significant p-value of the PRESS test defined in (2.17).

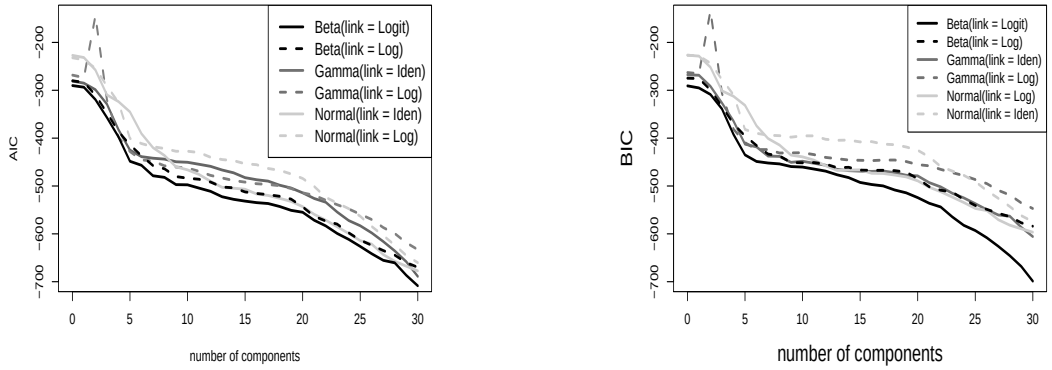

Figure 4.4: Number of components versus AIC (left) and BIC (right) values for the indicated PLS model and link function with kaolinite data.

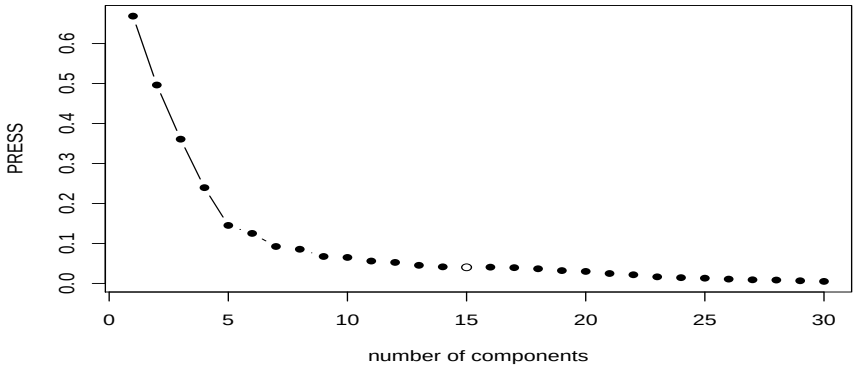

Figure 4.5: Number of components versus PRESS values for the beta PLS regression model with kaolinite data.

Based on the exploratory data analysis performed in Section 4.3, we assume the response $Y_i \sim \text{Rbeta}(\mu_i, \delta)$ for kaolinite data. Then, the systematic component of the regression model on the mean, by using the logit link function, is expressed as

$$\text{E}(Y_i) = \eta_i = g(\mu_i) = \underline{x}_i^\top \underline{\beta}, \quad i = 1, \ldots, 99, \tag{4.1}$$

where $\underline{x}_i = (1, x_{i1}, \ldots, x_{i2151})^\top$ are the observations of 2151 covariates, $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_{2151})^\top$ and $g(\mu) = \log(\mu/(1-\mu))$. The model defined in (4.1) can be written as a beta PLS model by

$$\begin{aligned} \text{E}(Y_i) &= g(\mu_i) = \log((\mu_i)/1 - \mu_i) \\ &= \underline{t}_i^\top \underline{q} = q_1 t_{i1} + \cdots + q_{15} t_{i15} \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{2151} x_{i2151}, \quad i = 1, 2, \ldots, 99, \end{aligned} \tag{4.2}$$

where $t_{i1}, \ldots, t_{i15}$ are obtained from the PLS components computed with Algorithm 2. The ML estimates of the PLS coefficient given in (4.2), and p-values of the t-tests, are presented in the Table 4.2. Note that all of the coefficient are significant to 5%. From the estimated coefficients provided in Table 4.2 for the model defined in (4.2), we can obtain the ML estimates $\widehat{\underline{\beta}}$ of $\underline{\beta}$. Due to restrictions of space, we only present the coefficients associated with $\widehat{\underline{\beta}}$ in Figure 4.6.

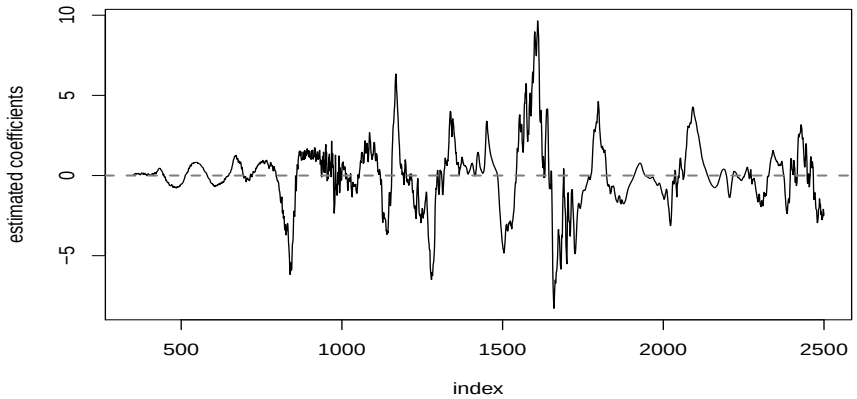

Figure 4.6: Index plot of the ML estimates of PLS $\underline{\beta}$ coefficients for kaolinite data.

4.5 Diagnostics and model checking

Model assumptions given in (4.1) and (4.2) are verified using RQ residuals defined in (3.1) for kaolinite data. Figure 4.7 shows that the autocorrelation assumption of the errors is verified, where neither outliers nor heterogeneity are detected. Figure 4.8 (left) sketches a PP plot with 95% acceptance bands to verify the distributional assumption of the model given in (4.1) and (4.2). Figure 4.8 (right) shows a QQ plot with simulated envelope to evaluate the normality fit to the RQ residuals. Note that KS p-value $= 0,1523$, which supports the normality assumption of the RQ residuals obtained from the beta PLS regression model. This figure does not show unusual features and assumption that the response follows an Rbeta distribution seems to be suitable. Diagnostics for the beta PLS regression model given in (4.1) and (4.2) are presented in Figure 4.9. From the left of this figure, corresponding to an index plot of the Cook distance defined in (3.3), the cases #33, #78 and #92 are detected as atypical and could be influential. Also, we analyze index plots of the Mahalanobis distance, defined in (3.4), in Figure 4.9 (right), from where the cases #20, #33, #61 and #78 are detected as atypical and then potentially influential. We now investigate the impact on the model inference when the cases detected, in both the Cook distance and Mahalanobis distance, are removed. Then, we again estimate the model parameters after removing the sets of cases $\{33\}$, $\{78\}$ and $\{33, 78\}$. Table 4.2 provides the relative changes (RCs) in the parameter estimates and in their corresponding estimated SEs, by using the kaolinite data. These RCs are calculated as $\mathrm{RC}_{\theta_{j(i)}} = |(\widehat{\theta}_j - \widehat{\theta}_{j(i)})/\widehat{\theta}_j| \times 100\%$ and $\mathrm{RC}_{\widehat{\mathrm{SE}}(\widehat{\theta}_j)_{(i)}} = |(\widehat{\mathrm{SE}}(\widehat{\theta}_j) - \widehat{\mathrm{SE}}(\widehat{\theta}_j)_{(i)})/\widehat{\mathrm{SE}}(\widehat{\theta}_j)| \times 100\%$, where $\theta_{j(i)}$ and $\widehat{\mathrm{SE}}(\widehat{\theta}_j)_{(i)}$ denote the ML estimates of θ_j and of the SE of the corresponding estimator, respectively, obtained after removing the case i, for $i = 1, \ldots, 99$ and $j = 1, \ldots, 16$. From Table 4.2, note that the cases #33 and #78 change more than 50% for some RCs. However, their significance is not modified for all coefficients, that is, there no inferential changes are found. The results presented in Table 4.2 provide evidence that the diagnostic measures derived in this article identify potentially influential points, but these do not affect the inference of the model. In summary, the diagnostic analysis based on Cook and Mahalanobis distances, and the RQ residual, confirm that the beta PLS regression model presented in (4.1) and (4.2) is not sensitive to the detected atypical cases and quite suitable for modeling the kaolinite data.

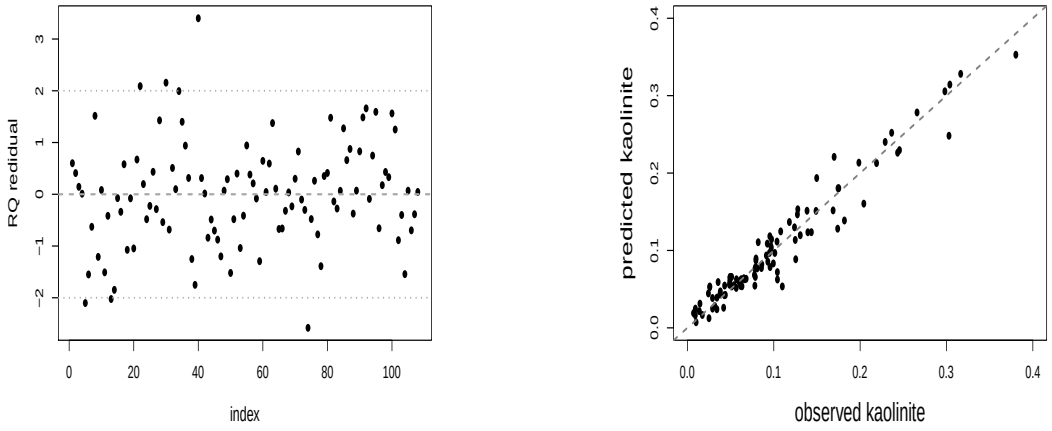

Figure 4.7: Index plot of the RQ residual (left) and predicted versus observed values (right) for kaolinite data.

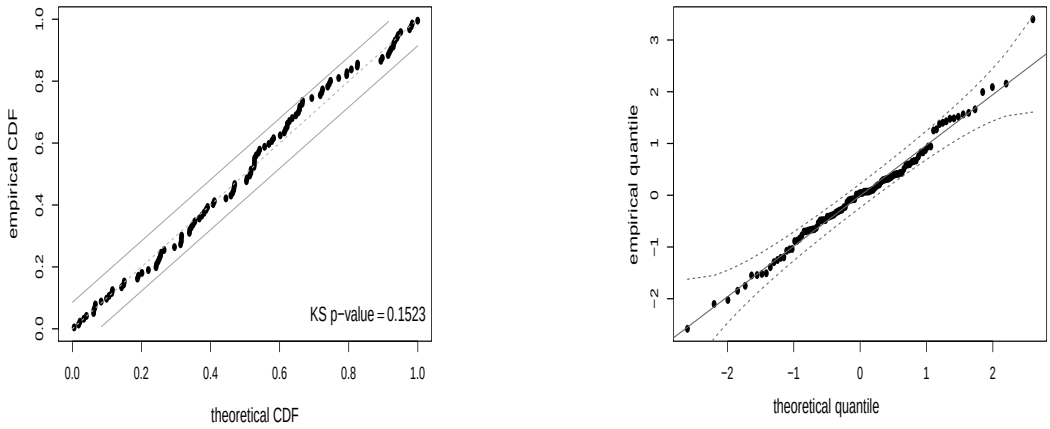

Figure 4.8: PP plot with 95% acceptance bands (left) and QQ plot and its simulated envelope (right) for RQ residuals with kaolinite data.

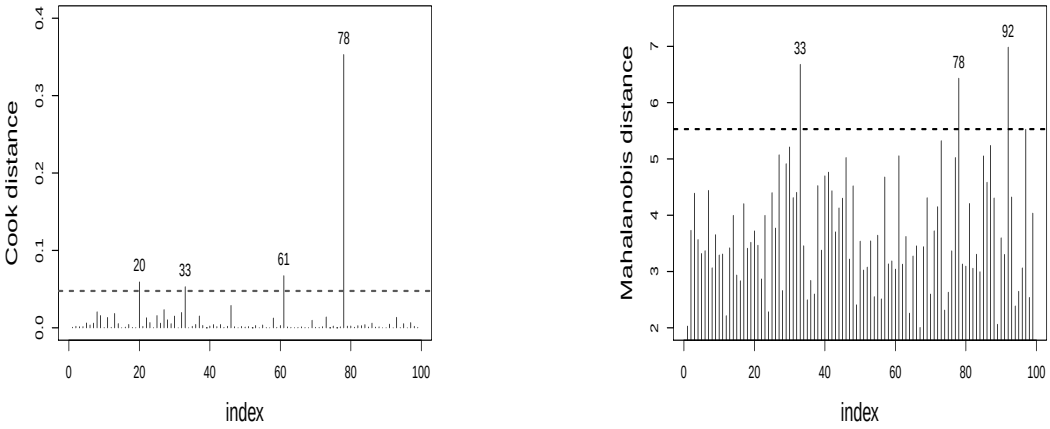

Figure 4.9: Index plots of the Cook (left) and Mahalanobis (right) distances with kaolinite data.

Table 4.2: RC on θ dropping the indicated case(s) for kaolinite data.

	None		{#33}				{#78}				{#33, #78}			
θ	$\widehat{\theta}$	p-value	$\widehat{\theta}$	p-value	$\mathrm{RC}_{\theta_{j(i)}}$	$\mathrm{RC}_{\mathrm{SE}(\widehat{\theta}_j)_{(i)}}$	$\widehat{\theta}$	p-value	$\mathrm{RC}_{\theta_{j(i)}}$	$\mathrm{RC}_{\mathrm{SE}(\widehat{\theta}_j)_{(i)}}$	$\widehat{\theta}$	p-value	$\mathrm{RC}_{\theta_{j(i)}}$	$\mathrm{RC}_{\mathrm{SE}(\widehat{\theta}_j)_{(i)}}$
q_1	0.007	< 0.01	0.007	< 0.01	1.140	0.758	0.008	< 0.01	8.458	4.987	0.008	< 0.01	8.705	3.220
q_2	0.047	< 0.01	0.047	< 0.01	1.601	0.647	0.050	< 0.01	7.154	4.518	0.051	< 0.01	8.951	4.710
q_3	0.175	< 0.01	0.185	< 0.01	5.594	6.146	0.209	< 0.01	19.297	3.408	0.226	< 0.01	28.938	12.680
q_4	0.185	< 0.01	0.181	< 0.01	2.203	0.335	0.156	< 0.01	15.347	15.684	0.153	< 0.01	17.290	15.331
q_5	0.202	< 0.01	0.206	< 0.01	2.184	2.518	0.162	< 0.01	19.949	0.773	0.160	< 0.01	20.646	0.076
q_6	0.197	< 0.01	0.204	< 0.01	3.404	0.725	0.136	< 0.01	31.271	13.353	0.143	< 0.01	27.732	13.148
q_7	0.248	< 0.01	0.246	< 0.01	0.720	0.803	0.161	0.001	35.199	11.321	0.164	0.001	33.680	10.572
q_8	0.397	< 0.01	0.417	< 0.01	5.077	3.930	0.487	< 0.01	22.627	22.406	0.568	< 0.01	42.836	25.178
q_9	0.154	0.015	0.208	0.004	35.376	13.625	0.122	0.043	20.911	0.420	0.125	0.040	19.004	3.835
q_{10}	0.226	0.015	0.227	0.014	0.409	0.528	0.373	0.003	65.215	36.829	0.412	0.001	82.419	33.981
q_{11}	0.767	< 0.01	0.672	< 0.01	12.312	14.573	0.373	0.018	51.331	28.499	0.548	< 0.01	28.513	37.979
q_{12}	0.759	< 0.01	0.734	< 0.01	3.272	16.075	0.310	0.023	59.123	36.450	0.282	0.013	62.894	47.522
q_{13}	0.743	0.001	0.980	< 0.01	31.916	18.893	0.450	0.005	39.442	28.441	0.526	0.015	29.182	2.674
q_{14}	0.909	0.003	0.706	0.013	22.363	7.178	0.895	0.002	1.530	7.923	0.906	0.005	0.372	4.731
q_{15}	1.073	0.002	0.804	0.007	25.133	14.049	1.283	< 0.01	19.490	12.821	0.964	< 0.01	10.187	21.486
δ	227.590	< 0.01	288.564	< 0.01	0.421	0.941	272.297	< 0.01	20.199	19.644	270.033	< 0.01	18.649	19.821

Chapter 5

Conclusions

Beta partial least squares regression model has proved to be quite appropriate for data modeling whose response corresponds to a proportion, which is coherent because the support of the response variable is in the range zero to one. Do not consider this distribution could lead to obtain predictions outside the natural range of the proportion, that is, negative estimates or greater than one. The use of the partial least squares regression models allows us to solve the usual problems of multicollinearity and/or high dimensionality in near infrared spectrometry data, thus obtaining reliable predictions about mineral content in different rock samples. We have proposed the use of randomized quantile residuals, which are commonly used in generalized additive models for location scale and shape models, closely related to the models proposed in the work. Furthermore, we have considered the Cook and Mahalanobis distances to detect outliers and influential cases, respectively, as the global diagnostic measures. We have performed modeling with data on the proportion of kaolinite, which is frequently considered in mining. The beta partial least squares regression model has been compared to other competitive partial least squares models showing a better performance and closer to reality for this kind of data. The analyses and results provided for the beta partial least squares regression model and its diagnostics may be of interest for the Chilean mining sector and for the world mining industry.

Appendix A

Reflectance spectrometry

Reflectance spectrometry is a technique that has been used since the beginning of the 20th century mainly by chemists and mining experts in order to identify certain compounds and minerals. However, since 1970, and due to advances in the field of electronics and optics, this technique of detection and analysis of certain compounds and mineral groups began to take a privileged place in topics of investigation and exploration of mineral resources.

Basically, spectrometry is a technique based on the behavior of electromagnetic field waves which are emitted, absorbed or reflected by a solid, liquid or gaseous body. All matter that is subject to radiation effects (such as a beam of light) undergoes a phenomenon of reflection and absorption of energy. Figure A.1 shows the behavior of a beam of light upon striking a given body in which one part of this beam of light is reflected and the other part propagates within the body being absorbed or transmitted. Both cases are manifested in the form of electromagnetic waves that can be measured and analyzed. Figure A.2 presents the electromagnetic spectrum divided by type and wavelength, such as gamma rays, X-rays, ultraviolet rays, visible area, infrared, among others.

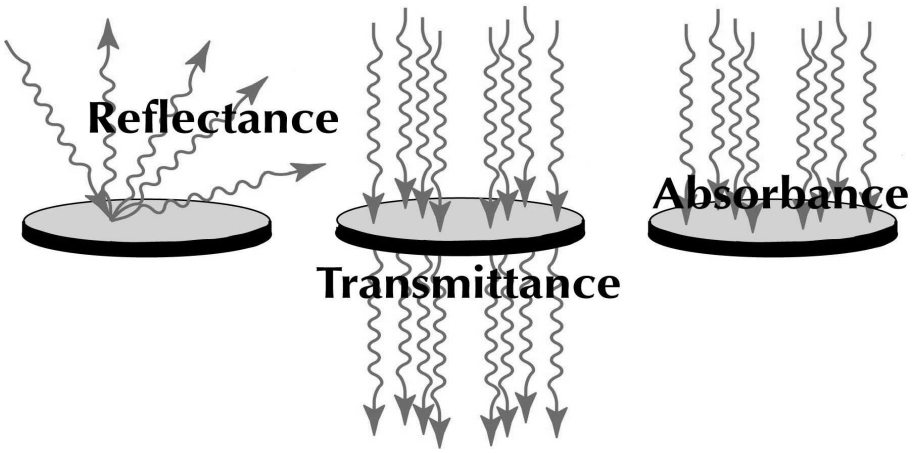

Figure A.1: Behavior of a beam of light on a body.

The absorption and reflection of energy of a molecule are due to the chemical and physical characteristics of this such as the distribution of its atoms, electrical composition, physical properties, etc. The spectrometry often used in the analysis of minerals in rocks employs the waves considered as the visible zone (between 350 nm and 780 nm), and the waves considered as near infrared (780 nm at 2500 nm). The absorption and reflection of the waves in these ranges are due to the vibration and rotation movements at the level of the atoms of each molecule subject to radiation. Thus, if the radiation frequency equals the natural vibration frequency of a given molecule, a change in the amplitude of the molecular

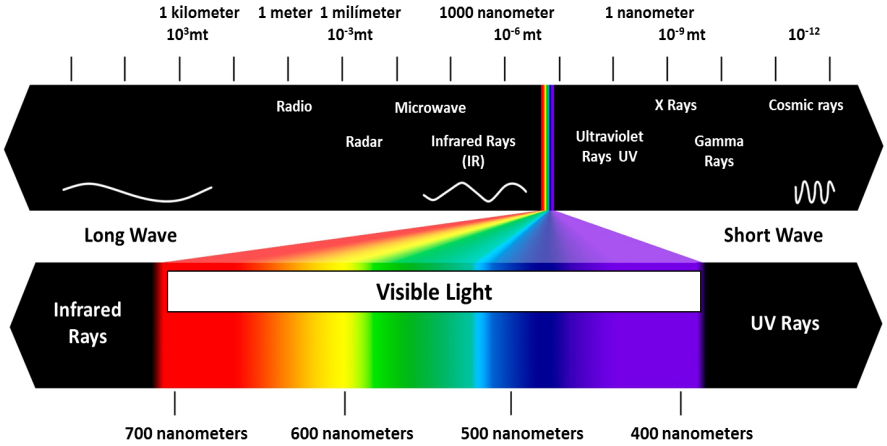

Figure A.2: Electromagnetic spectrum.

vibration is generated by absorbing the radiation. In this way, each different molecule will have a different spectrum. We can define a spectrum as a two-dimensional continuous graph whose horizontal axis represents the wavelength to which matter is subjected, and its vertical axis represents the percentage (or proportion) of reflectance. One of the main features that must be considered in any spectrum to determine and identify the presence or absence of its compounds is the absorption traits. These change the shape and depth across different wavelengths depending on the chemical composition of the analyzed sample, giving signs of presence of certain compounds as OH, H_2O, NH_4, CO_3, among others. These absorption ranges can be sharp, double and treble, simple and open, among others. Figure A.3 presents the main ranges of absorption that can be found in the analysis of spectra. For example, Figure A.4 shows some typical spectra of different minerals, illustrating their different shapes and positions with respect to the wavelength of their absorption traits.

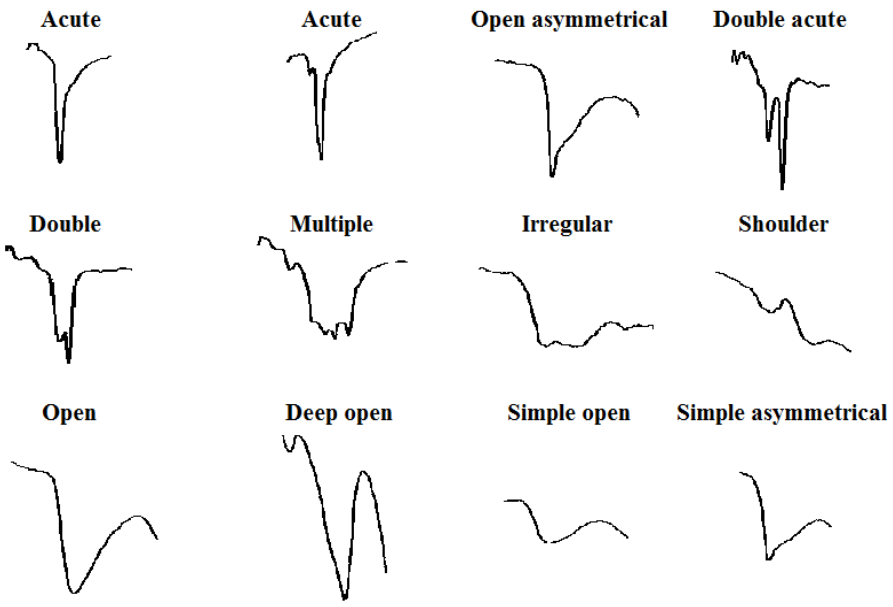

Figure A.3: Main absorption traits of an electromagnetic spectrum.

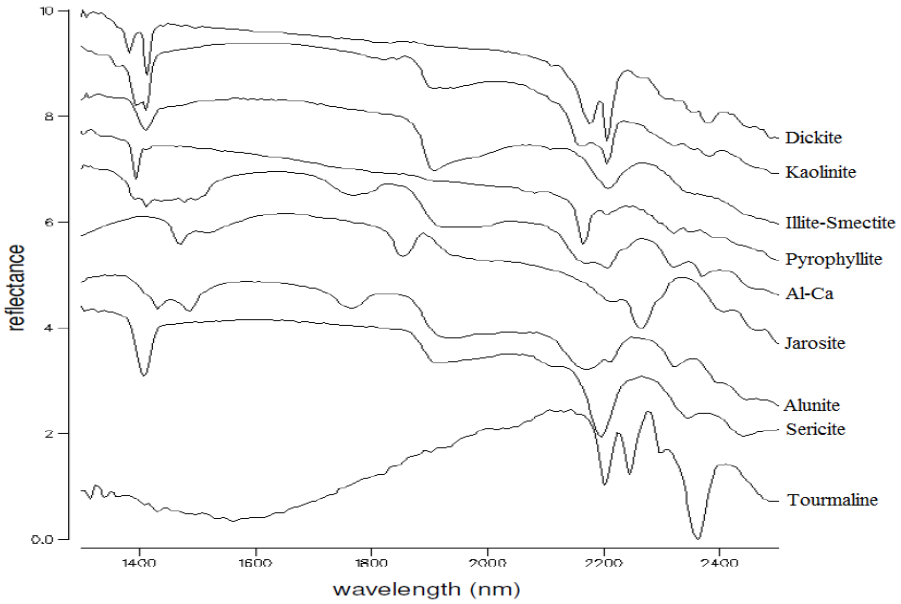

Figure A.4: Example of spectra of certain minerals.

It is worth to mention that, in practice, it is difficult to find pure samples of a particular compound, since the rocks usually present a mixture of several minerals. With the reflectance spectrometry method, it is possible to detect these combinations of minerals through the presence of different absorption ranges which are typical of certain minerals in the spectrum. Figure A.5 displays an example of the association of dickite and alunite by the representation of their spectra, observing that the features are well defined and combined in the sample.

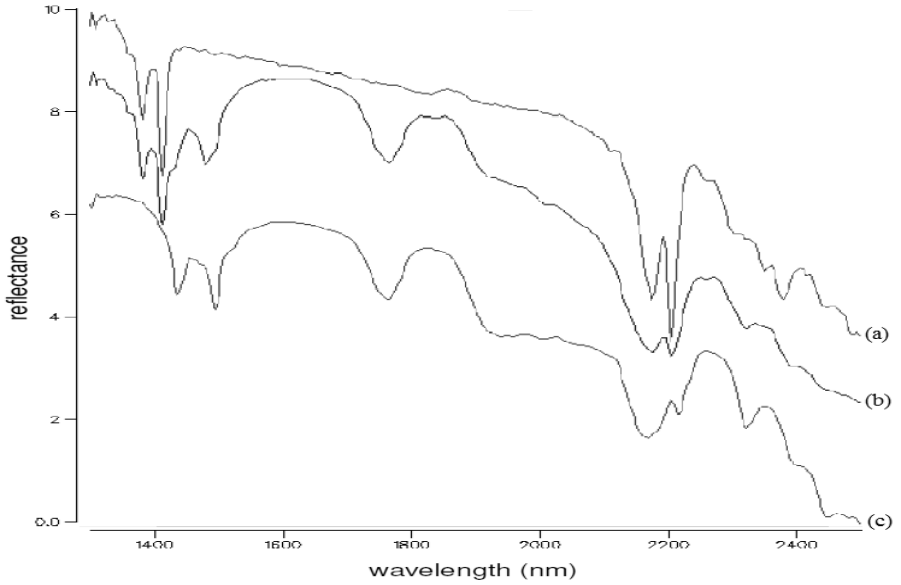

Figure A.5: Representation of (a) dickite (100%), (b) association of alunite-dickite and (c) alunite (100%) spectra.

It is important to mention that obtaining clean and accurate measurements of spectrometry lead to useful and reliable results. For this reason, the whole process of sampling and analysis of the samples must be correctly carried out. Therefore, for a correct interpretation of the spectra, the following considerations must be taken into account:

- Humidity: Water, like all chemical compounds, has well-defined spectral characteristics that can hide or dissolve the absorption traits of other minerals, generating inappropriate readings of the spectrum and consequently an imprecise interpretation of the sample subject to analysis. For this reason, it is important to consider the humidity of the rock in those minerals that do not contain water in its molecular structure.

- Irregular surface: It is important that the surface of the mineral to be analyzed is as regular as possible (flat) in order to avoid deformed spectra. The latter phenomenon is known as "noise". Some samples that may present this problem are very porous or fractured rocks.

- Color: Because the spectrometry method is based on the measurement of reflected waves by the minerals, the presence of certain dark minerals such as tourmaline can alter the levels of light absorption, hiding relevant data of other compounds of interest in the samples generating noise in the spectrum representation. The same happens when the minerals are translucent, such as gypsum, altering the levels of reflection in the measurement.

Some advantages of near infrared spectrometry are the following:

- It is a non-destructive or invasive technique.

- Solid, liquid and gaseous samples can be analyzed.

- The preparation of the sample is practically null.

- The analysis is fast.

- It has a very low cost.

- There is no need to use solvents, so it does not generate waste.

Appendix B

R CODES

```
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------
## A beta partial least squares regression model:
##   Diagnostics and application to mining data
##                    R codes
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------



#-------------------------------------------------------------------------
#-------------------------------------------------------------------------
## Used packages
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------

library(moments)
library(robustbase)
library(plsRbeta)
library(plsRglm)
library(betareg)
library(chemometrics)
library(asdreader)
```

```
#------------------------------------------------------------------------
#------------------------------------------------------------------------
## ETL process
#------------------------------------------------------------------------
#------------------------------------------------------------------------

setwd("C:/Users/Documents/betaPLS/ETL")
data      = read.csv2("minerals.csv", header = TRUE)
attach(data)
Caolinita = data[,"Caolinita"]/100

n     = length(Sample)
data2 = matrix(numeric(), nrow = n, ncol = 2152)

for(i in 1:n){
    name = Sample[i]
    fn        <- paste("C:/Users/Documents/betaPLS/ETL/",
              name,
   sep = "")
    s         <- get_spectra(fn)
    data2[i,] <- cbind(matrix(Caolinita[i]), s )
    print(i)
    }

colnames(data2) = c("Caolinita",
                    c(paste("v", c(350: 2500),
sep = "")))

write.csv2(data2, "NIR_data.csv")

#------------------------------------------------------------------------
#------------------------------------------------------------------------
## Function for implementing the classic PLS regression model
#------------------------------------------------------------------------
#------------------------------------------------------------------------

#------------------------------------------------------------------------
## Inputs:
#### Y: Response variable.
#### X: Matrix of covariates.
#### nc: Number of components.
## Outputs:
#### W: Weight matrix
#### T: PLS components matrix
#### Q: loading vector
#### coeff.scaled: Scaled regression coefficients.
#### coeff: Regression coefficients.
#### fitted: Fitted values
#------------------------------------------------------------------------
```

```
mypls<-function(Y,X,nc){
  pls<-NULL
  pls$y<-Y<-as.matrix(Y)
  pls$x<-X<-as.matrix(X)
  Resp<-scale(Y)
  Pred<-scale(X)
  A<-t(Pred)%*%Resp
  M<-t(Pred)%*%Pred
  C<-diag(ncol(Pred))
  pls$W<-W<-NULL
  pls$P<-P<-NULL
  pls$Q<-Q<-NULL

  for (i in 1:nc){
    q<-matrix(eigen(t(A)%*%A)$vectors[1,])
    w<-C%*%A%*%q
    w<-w/norm(w, type="f")
    p<-M%*%w
    c<-as.numeric(t(w)%*%M%*%w)
    p<-p/c
    q<-t(A)%*%w/c
    A<-A-c*p%*%t(q)
    M<-M-c*p%*%t(p)
    C<-C-w%*%t(p)
    pls$W<-W<-cbind(W,w)
    pls$P<-P<-cbind(P,p)
    pls$Q<-Q<-cbind(Q,q)
  }

  pls$T<-T<-X%*%W
  pls$coeff.scaled<-B<-W%*%t(Q)
  Ybar<-as.matrix(mean(Y), ncol=1)
  Xbar<-as.matrix(apply(X, 2, mean), ncol=1)
  Ysd<-as.matrix(sd(Y), ncol=1)
  Xsd<-diag(apply(X, 2, sd))
  B.originales<-solve(Xsd)%*%B%*%Ysd
  intercepto<-t(Ybar)-t(Xbar)%*%B.originales
  pls$coeff<-cbind(intercepto,t(B.originales))
  pls$fitted<-pred<-rep(intercepto, nrow(X))+X%*%B.originales
  pls
}
```

```
#------------------------------------------------------------------------
#------------------------------------------------------------------------
## Dataset
#------------------------------------------------------------------------
#------------------------------------------------------------------------

setwd("C:/Users/Documents/betaPLS")
data<-read.csv2("NIR_data.csv", dec=".")
data<-data[,-c(1,3)]
data$Caolinita<-data$Caolinita/100
attach(data)


#------------------------------------------------------------------------
#------------------------------------------------------------------------
## Exploratory data analysis
#------------------------------------------------------------------------
#------------------------------------------------------------------------

#------------------------------------------------------------------------
## Descriptive statistics
#------------------------------------------------------------------------

descriptive<-c(n      <- length(Caolinita),
               min    <- min(Caolinita),
               median <- median(Caolinita),
               mean   <- mean(Caolinita),
               max    <- max(Caolinita),
               sd     <- sd(Caolinita),
               skew   <- skewness(Caolinita),
               kurt   <- kurtosis(Caolinita),
               cv     <- sd/mean)

round(descriptive,4)
```

```
#---------------------------------------------------------------------------
## Histogram and boxplots
#---------------------------------------------------------------------------

setEPS()
postscript("fig01.eps", height=5)
par(mfrow=c(1,2))

hist(Caolinita,
     freq    = F,
     ylab    = "density",
     xlab    = "proportion of kaolinite",
     xlim    = c(0,0.5),
     cex.lab = 1.2)

boxplot(as.data.frame(cbind(rep(median,n),Caolinita)),
        names      = c("adjust boxplot","boxplot"),
        horizontal = T,
        xlab       = "proportion of kaolinite",
        cex.lab    = 1.2)

adjbox(as.data.frame(cbind(Caolinita,rep(median,n))),
       names=c("",""),
       add=T,
       horizontal=T)

dev.off()

#---------------------------------------------------------------------------
## Plot of spectra for covariates
#---------------------------------------------------------------------------

setEPS()
postscript("fig02.eps", height=5)
par(mfrow=c(1,1))

plot(350:2500,
     data[1,-1],
     type = "l",
     ylim = c(0.2,0.7),
     xlab = "wavelength (nm)",
     ylab = "reflectance")

for (i in 2:n){
     lines(350:2500, data[i,-1], type="l")
}

dev.off()
```

```
#------------------------------------------------------------------------
#------------------------------------------------------------------------
## Modelling, estimation and inference
#------------------------------------------------------------------------
#------------------------------------------------------------------------

#------------------------------------------------------------------------
## AIC and BIC
#------------------------------------------------------------------------

beta.logit   <- plsRbeta(Caolinita ~ .,
                         data   = data,
                         nt     = 30,
                         modele = "pls-beta",
                         link   = "logit")

beta.log     <- plsRbeta(Caolinita ~ .,
                         data   = data,
                         nt     = 30,
                         modele = "pls-beta",
                         link   = "log")

gamma.log    <- plsRglm(Caolinita ~ .,
                        data    = data,
                        nt      = 30,
                        modele  = "pls-glm-family",
                        family  = Gamma(link="log"))

gamma.ident  <- plsRglm(Caolinita ~ .,
                        data    = data,
                        nt      = 30,
                        modele  = "pls-glm-family",
                        family  = Gamma(link="identity"))

normal.ident <- plsRglm(Caolinita ~ .,
                        data    = data,
                        nt      = 30,
                        modele  = "pls-glm-family",
                        family  = gaussian(link="identity"))

normal.log   <- plsRglm(Caolinita ~ .,
                        data    = data,
                        nt      = 30,
                        modele  = "pls-glm-family",
                        family  = gaussian(link="log"))
```

```
#-------------------------------------------------------------------------
## Plot of AIC and BIC
#-------------------------------------------------------------------------

setEPS()
postscript("AIC.eps", height=5)
cex.lab=1.5
plot(beta.logit$AIC[1,],
     type = "l",
     lty  = 1,
     lwd  = 2,
     col  = "black",
     ylim = c(-700, -100),
     xlab = "number of components",
     ylab = "AIC")
lines(beta.log$AIC[1,], type="l", lty=2, lwd=2, col="black")
lines(gamma.ident$AIC[1,], type="l", lty=1, lwd=2, col="gray75")
lines(gamma.log$AIC[1,], type="l", lty=2, lwd=2, col="gray75")
lines(normal.ident$AIC[1,], type="l", lty=1, lwd=2, col="gray50")
lines(normal.log$AIC[1,], type="l", lty=2, lwd=2, col="gray50")
dev.off()

setEPS()
postscript("BIC.eps", height=5)
cex.lab=1.5
plot(beta.logit$BIC[1,],
     type = "l",
     lty  = 1,
     lwd  = 2,
     col  = "black",
     ylim = c(-700, -100),
     xlab = "number of components",
     ylab = "AIC")
lines(beta.log$BIC[1,], type="l", lty=2, lwd=2, col="black")
lines(gamma.ident$BIC[1,], type="l", lty=1, lwd=2, col="gray75")
lines(gamma.log$BIC[1,], type="l", lty=2, lwd=2, col="gray75")
lines(normal.ident$BIC[1,], type="l", lty=1, lwd=2, col="gray50")
lines(normal.log$BIC[1,], type="l", lty=2, lwd=2, col="gray50")
dev.off()
```

```
#-----------------------------------------------------------------------------
## Number of optimal PLS components: PRESS criteria
#-----------------------------------------------------------------------------

model1<-plsRbeta(Caolinita~.,
                 data   = data,
                 nt     = 30,
                 modele = "pls-beta")

mod1   <- model1$FinalModel
betas  <- cbind(mod1$coefficients$mean)
tt2    <- cbind(1,model1$tt)
pred   <- exp(tt2%*%betas)/(1 + exp(tt2%*%betas))
ynaive <- (mod1$y)
tt3    <- tt2[,-1]

tt_infl <- tt3
y_infl  <- ynaive

PRESS = numeric()

for(j in 1:30){
    predi = res = numeric()
    for(i in 1:n){
        mod1     = betareg(y_infl ~ tt_infl[,c(1:j)])
        tt_aux   = cbind(1,tt_infl[,c(1:j)])
        betas    = cbind(mod1$coefficients$mean)
        tt2      = tt_aux[i,]
        predi[i] = exp(tt2%*%betas)/(1 + exp(tt2%*%betas))
        res[i]   = y_infl[i] - predi[i]
        }
    PRESS[j] = sum(res^2)/n
    print(j)
    }

p_valuePRESS = numeric()

for(m in 1:30){
    statF          = (PRESS[m] - PRESS[m+1])/(PRESS[m+1]/(n - (m+1)))
    df_num          = 1
    df_den          = (n - (m+1))
    p_valuePRESS[m] = 1 - pf(statF, df1 = df_num, df2 = df_den)
    }

p_valuePRESS>0.05
```

```
#------------------------------------------------------------------------
## Plot of PRESS criteria
#------------------------------------------------------------------------

setEPS()
postscript("fig03.eps", height=5)
plot(PRESS*n,
     type = 'b',
     pch  = 16,
     ylab = "PRESS",
     xlab = "number of components", cex.lab = 1.5, cex = 1)
text(15, PRESS[15]*n + 1, "p-value\n >0.05", cex = 1)
points(15, PRESS[15]*n, col = 2, pch = 16)
dev.off()



#------------------------------------------------------------------------
#------------------------------------------------------------------------
## Modeling
#------------------------------------------------------------------------
#------------------------------------------------------------------------

model15<-plsRbeta(Caolinita~.,
                  data   = data,
                  nt     = 15,
                  modele = "pls-beta")



#------------------------------------------------------------------------
## Plot of predicted versus observed values
#------------------------------------------------------------------------

par(mfrow=c(1,1))
setEPS()
postscript("fig05.eps", height=5)
plot(Caolinita, predict(model15$FinalModel),
     xlab = "observed kaolinite",
     ylab = "predicted kaolinite",
     pch  = 16,
     xlim = c(0,0.4),
     ylim = c(0,0.4))
abline(a=0, b=1, lty=2, lwd=2, col="gray50")
dev.off()
```

```
#-------------------------------------------------------------------------
## Plot of estimated coefficients
#-------------------------------------------------------------------------

par(mfrow=c(1,1))
par(cex.lab=1.5)
setEPS()
postscript("fig08.eps", height=5)
plot(350:2500,model15$Coeffs[-1],
     type = "l",
     xlab = "index",
     ylab = "estimated coefficients")
dev.off()



#-------------------------------------------------------------------------
#-------------------------------------------------------------------------
## Diagnostics and model checking
#-------------------------------------------------------------------------
#-------------------------------------------------------------------------


#-------------------------------------------------------------------------
## Mahalanobis distance
#-------------------------------------------------------------------------

setEPS()
postscript("fig09.eps", height=5)

md<-Moutlier(model15$tt,
             quantile=0.975,
             plot=F)

plot(md$md,
     pch     = 16,
     xlab    = "index",
     ylab    = "Mahalanobis distance",
     cex.lab = 1.5,
     ylim    = c(2,7.5))
abline(h=md$cutoff, lty=2, lwd=2)

lab      <- md$md > md$cutoff
myLabels <- 1:n # choose appropriate labels
text(which(lab), md$md[lab], labels = myLabels[lab], pos = 3)
dev.off()
```

```
#-------------------------------------------------------------------------
## Cook distance
#-------------------------------------------------------------------------

cook_naive = as.numeric(cooks.distance(model15$FinalModel))
r2         = residuals(model15$FinalModel, "pearson")


mu = as.numeric(predict(model15$FinalModel))
ynaive = Caolinita
phi    = 205.2886
p      = phi*mu
q      = phi*(1-mu)
U      = pbeta(ynaive, p, q)
rQ     = qnorm(U)

cd     = (cook_naive/(r2*r2))*rQ*rQ

setEPS()
postscript("fig10.eps", height=5)
plot(cd,
     ylim = c(0,0.5),
     type = 'h',
     xlab = "index",
     ylab = "Cook's distance")
abline(h = 4/(n-15), col = "gray35", lty = 2, lwd = 2)

lab      <- cd > 4/(n-15)
myLabels <- 1:length(cd) # choose appropriate labels
text(which(lab), cd[lab], labels = myLabels[lab], pos = 3)
dev.off()

#-------------------------------------------------------------------------
## Estimated coefficients
#-------------------------------------------------------------------------

summary(model15$FinalModel)

estimated_coef<-round(as.data.frame(
                rbind(c(model15$FinalModel$coefficients$mean,
                model15$FinalModel$coefficients$precision),
                sqrt(diag(model15$FinalModel$vcov)))),4)

names(estimated_coef)<-c("$q_0$","$q_1$","$q_2$","$q_3$","$q_4$",
                         "$q_5$","$q_6$","$q_7$","$q_8$","$q_9$",
                         "$q_10$","$q_11$","$q_12$","$q_13$",
                         "$q_14$","$q_15$","$\\phi$")

write.csv2(estimated_coef,"estimated_coeffs.csv", row.names=F)
```

Bibliography

Atkinson, A. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press, Oxford, UK.

Bastien, P., Esposito, V., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics and Data Analysis*, 48:17–46.

Bertrand, F., Maumy-Bertrand, M., and Meyer, N. (2014a). *Partial Least Squares Regression for Beta Regression Models*. R package version 0.2.0.

Bertrand, F., Maumy-Bertrand, M., and Meyer, N. (2014b). *Partial Least Squares Regression for Generalized Linear Models*. R package version 1.1.1.

Bertrand, F., Meyer, N., Beau-Faller, M., El Bayed, K., Namer, I., and Maumy-Bertrand, M. (2013). Régression bêta PLS. *Journal de la Société Française de Statistique*, 154:143–159.

Burns, D. and Ciurczak, E. (2007). *Handbook of Near-Infrared Analysis*. CRC Press, London, UK.

Castro-Kuriss, C., Fierro, R., Leiva, V., and Saunders, S. C. (2016). On goodness of fit for cumulative damage distributions. *Under review in Applied Stochastic Models in Business and Industry*.

Chatterjee, S. and Hadi, A. (1988). *Sensitivity Analysis in Linear Regression*. Wiley, New York, US.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19:15–18.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, UK.

Cribari-Neto, F.and Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, 34:1–24.

Cribari-Neto, F. and Queiroz, M. (2012). On testing inference in beta regression. *Journal of Statistical Computation and Simulation*, 5:1–18.

Dunn, P. and Smyth, G. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:236–244.

Espinheira, P., Ferrari, S., and Cribari-Neto, F. (2008). Influence diagnostics in beta regression. *Computational Statistics and Data Analysis*, 52:4417–4431.

Fernandez, E., Valtuille, R., Willshaw, P., and Balzarini, M. (2008). Partial least squares regression: A valuable method for modeling molecular behavior in hemodialysis. *Annals of Biomedical Engineering*, 36:1305–1313.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31:799–815.

Ferrari, S., Espinheira, P., and Cribari-Neto, F. (2011). Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica*, 65:337–351.

Filzmoser, P.and Varmuza, K. (2016). *chemometrics: Multivariate Statistical Analysis in Chemometrics*. R package version 1.4.1.

Fox, J. (1991). *Regression Diagnostics: An Introduction*. Sage, Newbury Park, CA.

Garcia-Papani, F., Uribe-Opazo, M. A., Leiva, V., and Aykroyd, R. G. (2016). Birnbaum-Saunders spatial modelling and diagnostics applied to agricultural engineering data. *Stochastic Environmental Research and Risk Assessment* (in press).

Geladi, P. and Kowalski, B. (1986). Partial least squares regression: A tutorial. *Analytica Chimica Acta*, 1:1–17.

Hair, J., Hult, G., Ringle, C., and Sarstedt, M. (2014). *A Primer on Partial Least Squares Structural Equation Modeling*. Sage, London, UK.

Johnson, N., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, volume 1. Wiley, New York, US.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, volume 2. Wiley, New York, US.

Jolliffe, I. (2002). *Principal Component Analysis*. Wiley, New York, US.

Kotz, S. and van Dorp, J. (2004). *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*. World Scientific, Singapore.

Land, W., Ford, W., Park, J., Mathur, R., Nathan, H., Heine, J., Eschrich, S., Qiao, X., and Yeatman, T. (2011). Partial least squares (PLS) applied to medical bioinformatics. *Procedia Computer Science*, 6:273–278.

Leao, J., Leiva, V., Saulo, H., and Tomazella, V. (2017). Birnbaum-Saunders frailty regression models: Diagnostics and application to medical data. *Biometrical Journal* (in press).

Leiva, V., Ferreira, M., Gomes, M. I., and Lillo, C. (2016). Extreme value Birnbaum-Saunders regression models applied to environmental data. *Stochastic Environmental Research and Risk Assessment*, 30:1045–1058.

Leiva, V., Rojas, E., Galea, M., and Sanhueza, A. (2014a). Diagnostics in Birnbaum-Saunders accelerated life models with an application to fatigue data. *Applied Stochastic Models in Business and Industry*, 30:115–131.

Leiva, V., Santos-Neto, M., Cysneiros, F. J. A., and Barros, M. (2014b). Birnbaum-Saunders statistical modelling: A new approach. *Statistical Modelling*, 14:21–48.

Leiva, V. and Saunders, S. C. (2015). Cumulative damage models. *Wiley StatsRef: Statistics Reference Online*.

Li, B., Morris, J., and Martin, E. (2002). Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 64:79–89.

Lillo, C., Leiva, V., Nicolis, O., and Aykroyd, R. G. (2016). L-moments of the Birnbaum-Saunders distribution and its extreme value version: Estimation, goodness of fit and application to earthquake data. *Journal of Applied Statistics* (in press).

Marchant, C., Leiva, V., Cysneiros, F. J. A., and Vivanco, J. F. (2016). Diagnostics in multivariate generalized Birnbaum-Saunders regression models. *Journal of Applied Statistics*, 43:2829–2849.

Marx, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38:374–381.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, UK.

Nocedal, J. and Wright, S. (1999). *Numerical Optimization*. Springer, New York, US.

Orbovic, V. and Huang, Z. (2012). *Kaolinite: Occurrences, Characteristics, and Applications*. Nova, US.

Osten, D. (1988). Selection of optimar regression models via coss-validation. *Journal of Chemometrics*, 2:39–48.

Paula, G. A., Leiva, V., Barros, M., and Liu, S. (2012). Robust statistical modeling using the Birnbaum-Saunders-t distribution applied to insurance. *Applied Stochastic Models in Business and Industry*, 28:16–34.

Roudier, P. (2016). *asdreader: Reading ASD Binary Files in R*. R package version 0.1-2.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2015). *robustbase: Basic robust statistics*. R package version 0.92-5.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2016). *robustbase: Basic Robust Statistics*. R package version 0.92-6.

Santos-Neto, M., Cysneiros, F. J. A., Leiva, V., and Barros, M. (2016). Reparameterized Birnbaum-Saunders regression models with varying precision. *Electronic Journal of Statistics*, 10:2825–2855.

Stasinopoulos, D. and Rigby, R. (2007). Generalized additive models for location, scale and shape (GAMLSS). *Journal of Statistical Software*, 23:1–46.

Varmuza, K. and Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, Boca Raton, US.

Williams, D. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Journal of the Royal Statistical Society C*, 36:1181–191.

Wold, H. (1975). *Path models with latent variables: The NIPALS approach*. Academic Press, New York.

Wold, S., Sjöström, M., and Eirksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130.