SEPARATION AND MULTICOLLINEARITY IN POLYTOMOUS QUADRATIC LOGISTIC REGRESSION MODEL

Inácio Andruski-Guimarães[1]
Departamento Acadêmico de Matemática, UTFPR – Universidade Tecnológica Federal do Paraná, Curitiba, Brasil.
Rua sete de setembro, 3065 Curitiba Paraná Brasil +55 041 3310-4650

ABSTRACT
The logistic regression model is used to model the relationship between a categorical dependent variable and a set of explanatory variables, continuous or discrete. Almost all papers on logistic regression have only considered the classical logistic regression model, with linear discriminant functions. But there are situations where quadratic discriminant functions are useful, and work better. However, the quadratic logistic regression model involves the estimation of a great number of unknown parameters, and this leads to computational difficulties when there are a great number of explanatory variables. Furthermore, if the groups of explanatory variables are completely separated, the maximum likelihood estimators of the unknown parameters do not exist. This paper proposes to use a set of principal components of the explanatory variables, in order to reduce the dimensions in the problem, with continuous independent variables, and the computational costs for the parameter estimation in polytomous quadratic logistic regression, without loss of accuracy. Examples on datasets taken from the literature show that the quadratic logistic regression model, with principal components, is feasible and, generally, works better than the classical logistic regression model with linear discriminant functions, in terms of correct classification rates.

KEYWORDS: Polytomous Logistic Regression, Quadratic Logistic Regression, Principal Components Analysis, Polytomous Response.

MSC: 62H30, 62H25, 62J02 , 68T10.

RESUMEN
El modelo de regresión logística es usado para modelar la relación entre una variable dependiente categórica y un conjunto de variables explicativas, contínuas o discretas. La mayoria de los artículos sobre regresión logística sólo han considerado el modelo de regresión logístico clásico, con funciones discriminantes lineales. Pero hay situaciones en las que las funciones discriminantes cuadráticas son útiles y funcionam mejor. Todavia, el modelo de regresión logístico cuadrático involvera la estimación de un número grande de parámetros desconocidos, y eso lleva a dificultades computacionales, sobretodo cuando hay un número grande de variables explicativas. Además, cuando hay separación entre los grupos de variables explicativas los estimadores de máxima verossimilitud de los parámetros desconocidos no existem. Este artículo propone usar un conjunto de componentes principales de las variables explicativas, con el fin de reducir las dimensiones del problema, con variables explicativas contínuas, y los costos computacionales para la estimación de parámetros en regresión logística cuadrática politómica, sin pérdida de precision. Ejemplos de conjuntos de datos tomados de la literatura muestram que el modelo de regresión logístico cuadrático, con componentes principales, es factible y generalmente funciona mejor que el modelo de regresión logístico clásico, con funciones discriminantes lineales, en términos de tasas de clasificación correctas.

1. INTRODUCTION

The logistic regression model is known as a powerful method widely applied to model the relationship between a categorical - or ordinal - dependent variable and a set of explanatory variables, or covariates, which may be either continuous or discrete. The accuracy of the Logistic Regression Model has been reported in many studies involving bankruptcy prediction, marketing applications and cancer classification, among other applications. Almost all papers on logistic regression have only considered the classical logistic regression model with linear discriminant functions, but there are situations where quadratic discriminant functions are useful, and work better. However, as pointed out by [3] and [5], the quadratic logistic regression model involves the estimation of

[1] andruski@utfpr.edu.br

a great number of unknown parameters, and this leads to computational difficulties, in terms of memory and computing time requirements, when there are a great number of explanatory variables. Furthermore, a great number of parameters should be avoided, because of the risk of over-fitting. Moreover, the logistic model becomes unstable when there is strong dependence, or multicollinearity, a phenomenon in which two or more explanatory variables are highly correlated. Another problem is that, while the logistic regression model works well for many situations, it may not work when the data set has no overlapping. These problems, multicollinearity and complete separation, are common in the logistic regression, and frequently occur together. In order to way out the problem that arises when the data set has no overlapping, [21] propose the Hidden Logistic Regression Model (HLR). Other approaches to deal with separation can be found in [13], for binary response, and [4] for polytomous response, to name just a few. In order to improve the parameter estimation under multicollinearity, and to reduce the dimension of the problem, [1] propose to use as covariates of the logistic regression model a reduced set of optimum principal components of the original covariates. This approach, called Principal Component Logistic Regression (PCLR) model, provide an accurate estimation of the parameters in the case of multicollinearity.

In this paper we propose to use a set of principal components of the explanatory variables, in order to reduce the dimensions in the problem, with continuous independent variables, and the computational costs for the parameter estimation in polytomous quadratic logistic regression, without loss of accuracy. To deal with separation we propose an extension of the hidden logistic model to polytomous response. The main advantage of this model is the existence and uniqueness of estimators, even when there is complete or quasi-complete separation. Furthermore, this paper gives an extension of the approach given, in a resumed way, by [5].

This paper is organized as follows. Section 2 consists of a brief review of the Classical Logistic Regression model (CLR). Section 3 presents an overview of the Quadratic Logistic Regression model (QLR) for polytomous response. In section 4 we extend an existing approach, called Principal Components Logistic Regression model (PCLR), developed to deal with multicollinearity in binary case, as well a generalization for the polytomous response case, and proposes the Principal Components Quadratic Logistic Regression model (PCQLR). In Section 5 we apply the QLR and PCQLR models on data sets taken from the literature and compare their performance with those that were obtained from the CLR and QLR models. Section 6 gives a brief conclusion and makes suggestions for future studies.

2. CLASSICAL LOGISTIC REGRESSION MODEL

Let us consider a sample of n observations, available from the groups G_1 , … , G_s, and a vector of explanatory variables $\mathbf{x}^{\mathrm{T}} = [x_0 , x_1 , \dots , x_p]$, where $x_0 \equiv 0$, for convenience. Let Y denote the polytomous dependent variable with s possible outcomes. We will summarize the n observations in a matrix form given by:

$$\mathrm{X} = \begin{bmatrix} 1 & \dots & x_{p1} \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_{pn} \end{bmatrix}.$$

The Classical Logistic Regression (CLR) Model assumes that the posterior probabilities have the form:

$$P(G_k|\mathbf{x}) = \frac{exp(\mathrm{B}_k\mathbf{x})}{\sum_{i=1}^{s} exp(\mathrm{B}_i\mathbf{x})} \tag{1}$$

where $\mathrm{B}_i\mathbf{x} = \beta_{i0} + \sum_{j=1}^{p} \beta_{ij} x_{j}$, $i = 1 , \dots , s-1$ and $\mathrm{B}_s = 0$. In this paper the group s is called reference group. The model involves $(s-1)(p+1)$ unknown parameters and the conditional likelihood function is:

$$L(\mathrm{B}|\mathbf{Y} , \mathbf{x}) = \prod_{i=1}^{n} \prod_{k=1}^{s} [P(G_k|\mathbf{x}_i)]^{y_{ki}} \tag{2}$$

where $\mathbf{Y} = (Y_1 , \dots , Y_n)^{\mathrm{T}}$ and $Y_i = (y_{1i} , \dots , y_{si})$, with $y_{ki} = 1$, if $Y = k$, and $y_{ki} = 0$, otherwise. Taking the logarithm, the log-likelihood function is given by:

$$\ell(\mathrm{B}|\mathbf{Y} , \mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{s} y_{ki} ln[P(G_k|\mathbf{x}_i)] \tag{3}$$

Thus:

$$\frac{\partial}{\partial \beta_{kj}} \ell(\mathrm{B}|\mathbf{Y} , \mathbf{x}) = \sum_{i=1}^{n} x_{ij}\big(y_{ki} - P(G_k|\mathbf{x}_i)\big) \tag{4}$$

The Maximum Likelihood Estimator (MLE) $\widehat{\mathrm{B}}$ is obtained by setting the derivatives to zero and solving for B. The solution is found using an iterative procedure, such as Newton-Raphson method.

In practice, the estimation of unknown parameters should take into account the possible configurations of the sample points. [2] suggested a sample classification into three categories: Overlap, complete separation and quasi-complete separation, when perfect prediction occurs only for a group of observations. They also proved that the MLE do not exist for complete and quasi-complete separation. In this case, existing iterative methods fail to converge, or give a wrong answer. We say that two groups are completely separated if there exists a vector $\mathrm{B} \in \mathbb{R}^m$, $m = (\mathrm{s}-1)(\mathrm{p}+1)$, such that for all $i \in G_j$, and j , $t = 1 , \dots , s$ $(j \neq t)$:

$$\left(\mathrm{B}_j - \mathrm{B}_t\right)^{\mathrm{T}} \mathbf{x}_i > 0 \tag{5}$$

We say that there is quasi-complete separation if, for all $i \in G_j$, and j , $t = 1 , \dots , s$ $(j \neq t)$:

$$\left(\mathrm{B}_j - \mathrm{B}_t\right)^{\mathrm{T}} \mathbf{x}_i \geq 0 \tag{6}$$

with equality for at least one (i , j , t) triplet. The points for which the equality holds are said to be quasi-separated.

In binary logistic regression, if there is complete separation, the MLE do not exist. However, in polytomous logistic regression, complete separation does not make the same sense, although the parameter estimation is not necessarily affected. According to [18], when there are more than two groups, separation can occur even when some groups overlap substantially. We say that two groups G_i and G_j are linearly separable if there exist a vector $\Omega = \left(\omega_1, \dots, \omega_p\right)$ and a real number δ such that $\Omega\mathbf{x}_k > \delta$, if $\mathbf{x}_k \in G_i$, and $\Omega\mathbf{x}_k < \delta$, if $\mathbf{x}_k \in G_j$, where i , $j = 1$, $\dots$, s $(i \neq j)$ and $k = 1, \dots , n$. When there are more than two groups, the difference between linear separability and separation becomes more important. In this case linear separability means the existence of a set of vectors $\{\Omega_1 , \dots , \Omega_s\}$ satisfying $s(s-1)$ inequalities given by:

$$\left(\Omega_j - \Omega_t\right)^{\mathrm{T}} \mathbf{x}_i \geq \delta \tag{7}$$

for all $i = 1 , \dots , n$ and j , $t = 1 , \dots , s$ $(j \neq t)$.

For the polytomous logistic regression model, [7] proposed an alternative method for the maximum likelihood estimation, called Individualized Logistic Regression (ILR) Model. For a response variable with s groups, they suggested fitting $(s-1)$ binary logistic regressions and, each time, comparing a group G_j , $j = 1 , \dots , s-1$, with the reference group, G_s. The coefficients for the polytomous logistic regression model are obtained from the $(s-1)$ separately fit logistic models. According the authors, the estimators that are obtained are consistent and will be approximately those from the CLR Model. Furthermore, according to [14], the ILR Model can be useful to select variables.

In order to way out the problem that arises when the data set has no overlapping, [21] propose the Hidden Logistic Regression Model (HLR). Other approaches to deal with separation can be found in [13], for binary response, and [4], for polytomous response, to name just a few.

3. QUADRATIC LOGISTIC REGRESSION MODEL

An extension of the linear logistic model is to include quadratic and multiplicative interaction terms. The Quadratic Logistic Regression (QLR) Model is given by:

$$Q\left(G_k|\underline{\mathbf{x}}\right) = \frac{exp(\Theta_k)}{\sum_{i=1}^{S} exp(\Theta_i)} \tag{8}$$

where $\Theta_k = \alpha_{k0} + \sum_{i=1}^{p} \alpha_{ki} x_i^2 + \sum_{i=p+1}^{(p-1)\frac{p}{2}} \alpha_{ki} x_{j'} x_{j''} + \sum_{i=(p-1)\frac{p}{2}}^{(p-1)\frac{p}{2}+p} \alpha_{ki} x_j$, $k = 1 , \dots , s-1$, $\Theta_s = 0$ and j , $j'' = 1$, $\dots$, p , $j' = 1 , 2 , \dots , p-1$.

The model involves $[(s-1)(p+1)]\left(1+\frac{p}{2}\right)$ unknown parameters and the estimation of these parameters follows the same lines as that taken by the Classical Logistic Regression Model (CLR). However, for a large number of independent variables, the number of extra parameters can be render an unworkable problem, so that a reduction dimension method can be useful to way out of this problem. Furthermore, a large number of parameters should be avoided, because of the risk of over-fitting. As pointed out by [3], the quadratic term also can be written as:

$$\Theta_k = \alpha_{k0} + \underline{\mathbf{x}}^T \mathbf{\Omega}_k \underline{\mathbf{x}} + \alpha_k^T \underline{\mathbf{x}} \tag{9}$$

where $\mathbf{\Omega}_k = \mathbf{V}_k^{-1} - \mathbf{V}_s^{-1}$, and $\mathbf{V}_k$ is the dispersion matrix in G_k , $k = 1 , \dots , s-1$. An approximation, proposed by [3], gives a quadratic term with a reduced number of parameters. This approximation is given by the spectral decomposition:

$$\mathbf{\Omega}_k = \sum_{j=1}^{p} \lambda_{jk} l_{jk} l_{jk}^T \tag{10}$$

where the λ_{jk} are the eigenvalues of $\mathbf{\Omega}_k$, in decreasing size, $\lambda_{1k} \geq \lambda_{2k} \geq \cdots \geq \lambda_{pk}$, and l_{jk} are the corresponding eigenvectors. In this case, $\mathbf{\Omega}_k$ can be given by:

$$\mathbf{\Omega}_k \cong \lambda_k l_k l_k^T \tag{11}$$

In the sequence, each $l_j^T = \left(l_{j1}, \dots, l_{jp}\right)$ is normed with the constraints $\sum_{k=1}^{p} l_{jk}^2 = 1$.

Since this approach is not convenient for computing, an alternative parameterization is suggested by [3]:

$$\Theta_k = \alpha_{k0} + \mu_k\left(d_k^T\underline{\mathbf{x}}\right)^2 + \alpha_k^T\underline{\mathbf{x}} \tag{12}$$

where $\mu_k = sgn(\lambda_k)$, $k = 1, \ldots, s-1$, $d_{kj} = {l_{kj}}\Big/{\sqrt{|\lambda_k|}}$, $j = 1, \ldots, p$.

The log-likelihood function is maximized with respect to the α_{kj} and d_{kj} unrestrictedly $2^{(s-1)}$ times for $\mu_k = \pm 1$ and to take as maximum likelihood estimates those values of the parameters which give the greatest of these $2^{(s-1)}$ values of the log-likelihood function. With this approximation, there are $(s-1)$ unknown parameters. However, this approach is not always applicable. If the independent variables are binary, the diagonal terms of $\boldsymbol{\Omega}$ are zero. In this paper we propose to use as covariates the principal components of the $(s-1)(p+1)$ matrix $I(\Theta)$ whose elements are given by:

$$\frac{\partial^2 L(\Theta)}{\partial\beta_{jm}\partial\beta_{jm'}} = -\sum_{i=1}^{m} x_{m'i}x_{mi}\left[Q\left(G_j|\underline{\mathbf{x}}\right)\right]\left[1 - Q\left(G_j|\underline{\mathbf{x}}\right)\right] \tag{13}$$

and

$$\frac{\partial^2 L(\Theta)}{\partial\beta_{jm}\partial\beta_{j'm'}} = -\sum_{i=1}^{m} x_{m'i}x_{mi}\left[Q\left(G_j|\underline{\mathbf{x}}\right)\right]\left[Q\left(G_{j'}|\underline{\mathbf{x}}\right)\right] \tag{14}$$

where j , $j' = 1, 2, \ldots, (s-1)$ and m , $m' = 1, \ldots, p$.

In this paper, it has been considered that the quadratic term is unlikely to be of major importance in determining the effectiveness of the discrimination compared to the linear term.

As the CLR model, the quadratic logistic model is not immune to complete separation. Our approach to solve the problem that arises when the data set has no overlapping, when there is complete, or quasi-complete separation, is to provide a simple and direct generalization of the Hidden logistic Regression (HLR) Model, a robust estimation method presented by [21]. This model was used by [9], under a different name. In related literature different approaches to implement robust estimation methods are given by [12] and [17], to name just a few. The HLR Model assumes that, due to an additional stochastic mechanism, the true response of a logistic regression model is unobservable, but there is an observable variable that is related to this response. Under this point of view, the true unobservable response is comparable to a hidden layer in a feed-forward neural network. In this paper we consider n unobservable independent variables $T_1, \ldots, T_n$, where each T_i has s values $\gamma_1, \ldots, \gamma_s$. Thus, we observe $Y_i = j$ with a $P(Y_i = j|T_i = \gamma_k) = \delta_{jk}$ probability, where $\delta_{jj} = max_{k=1,\ldots,s}\{\delta_{jk}\}$ and $\sum_{j=1}^{s}\delta_{jk} = 1$. Let us keep in mind that in the CLR Model $\delta_{jj} = 1$ and $\delta_{jk} = 0$, $j \neq k$.

The maximum likelihood estimator for T_i , if $Y_i = j$, is $\hat{T}_i = \gamma_j$. In a model with n responses y_{ij} , $i = 1, \ldots, s$ and $j = 1, \ldots, n$, with $y_{ki} = 1$, if $Y = k$, and $y_{ki} = 0$, otherwise, we can define the variable given by:

$$\tilde{y}_{ij} = \sum\nolimits_{k=1}^{s} y_{ik}\delta_{kj} \tag{15}$$

The purpose is to maximize:

$$L\left(\Theta|\tilde{\boldsymbol{Y}}, \mathbf{x}\right) = \prod_{i=1}^{n}\prod_{k=1}^{s}[Q(T_k|\mathbf{x}_i)]^{y_{ki}} \tag{16}$$

The log-likelihood function becomes:

$$\ell\left(\Theta|\tilde{\boldsymbol{Y}}, \mathbf{x}\right) = \sum_{i=1}^{n}\left[\sum_{j=1}^{s-1}\tilde{y}_{ji}\Theta_j - ln\left(1 + \sum_{j=1}^{s-1}exp\left(\Theta_j\right)\right)\right] \tag{17}$$

The MLE are the maximizers of the log-likelihood function, which is strictly concave. Unlike the MLE for the CLR model, the MLE for the HLR model always exists.

According to [21], [9] found that accurate estimation of δ, in the binary case, is very difficult, unless n is extremely large. For a detailed explanation, see [9], [15] and [21]. We consider that the probability of observing the true status, given by $\left(Y_i = j\middle|T_i = \gamma_j\right) = \delta_{jj}$, should be higher than 0.5, this is, $0.5 < \delta_{jj} < 1$. Furthermore, $\sum_{k=1,k\neq j}^{s}\delta_{jk} < \delta_{jj}$. Therefore, we cannot take the estimate given by $\bar{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$, $j = 1, \ldots, s$, once $\bar{\pi}_j$ can be smaller than 0.5. Our default choice will be δ = 0.99 and set $\delta_{jj} = \delta$ and $\delta_{jk} = \frac{1-\delta}{s-1}$.

4. PRINCIPAL COMPONENTS LOGISTIC REGRESSION MODEL

The Principal Components Analysis (PCA) is a method to explain the variance and covariance structure through linear combinations of the covariates and may be considered a tool for reducing the dimensionality of the data, as well as the multicollinearity among the independent variables.

Let us consider n observations of p continuous variables, given by the matrix $\boldsymbol{X}$, and the sample covariance matrix, given by:

$$\mathbf{S} = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \cdots & \ddots & \cdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}$$

The observations **x** can be standardized, so that

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\,.$$

The matrix **S** can be written as $\mathbf{S} = \mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V}$, where $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ being orthogonal. Consider the linear combinations given by $Y_j = \ell_{1j}x_1 + \cdots + \ell_{pj}x_j$, where $j = 1, \ldots, p$. The principal components are those uncorrelated linear combinations whose variances are as large as possible. The *j*-th principal component is the linear combination $\ell_j\mathbf{x}^{\mathrm{T}}$ that maximizes $Var(\ell_j\mathbf{x}^{\mathrm{T}})$ subject to $\ell_j\ell_j^{\mathrm{T}} = 1$.

Let **Z** be the matrix whose columns are the principal components, given by **Z** = **XV**, where $\mathbf{v}_1, \ldots, \mathbf{v}_p$ are the eigenvectors of the matrix **S**, associated to the eigenvalues $\lambda_1, \ldots, \lambda_p$, so that the matrix of observations can be written as $\mathbf{X} = \mathbf{Z}\mathbf{V}^{\mathrm{T}}$, where

$$x_{ij} = \sum_{k=1}^{p} z_{ik}v_{jk} \tag{18}$$

Furthermore, matrices **Z** and **V** also can be written as:

$$\mathbf{Z} = \begin{bmatrix} 1 & z_{11} & \ldots & z_{1(q+1)} & \cdots & z_{1p} \\ 1 & z_{21} & \ldots & z_{2(q+1)} & \cdots & z_{2p} \\ \ldots & \cdots & \ldots & \cdots & \ldots & \cdots \\ 1 & z_{n1} & \ldots & z_{n(q+1)} & \ldots & z_{np} \end{bmatrix} = \left(\mathbf{Z}_{(q)}|\mathbf{Z}_{(r)}\right)$$

and

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & \ldots & 1 & \ldots & 1 \\ 1 & v_{11} & \ldots & v_{1(q+1)} & \cdots & v_{1p} \\ \ldots & \cdots & \ldots & \cdots & \ldots & \cdots \\ 1 & v_{p1} & \ldots & v_{p(q+1)} & \ldots & v_{pp} \end{bmatrix} = \left(\mathbf{V}_{(q)}|\mathbf{V}_{(r)}\right)$$

In order to improve the parameter estimation under multicollinearity, and to reduce the dimension of the problem, [1] propose to use as covariates of the logistic regression model a reduced set of optimum principal components of the original covariates, as an extension of the model introduced by [19], in the linear case. This approach, called Principal Component Logistic Regression (PCLR) model, provide an accurate estimation of the parameters in the case of multicollinearity. Furthermore, according to [6], estimates obtained via principal components can have smaller mean square error than estimates obtained through standard logistic regression. But, according to [20], it is well known that the estimates of the eigenvalues of **S** are biased. This bias is most pronounced when the eigenvalues of **S** tend toward equality, being less severe when they are highly disparate.

A generalization of the PCLR model for polytomous responses can be found in [4] and does not require a complex formulation. It begins by computing the covariance matrix **S**. Then the matrix **S** can be written as:

$$x_{ik} = \sum_{j=1}^{p} z_{ij}v_{kj} \tag{19}$$

so that

$$P\left(G_t|\mathbf{Z}\underline{\mathbf{v}}_i\right) = \frac{exp\left(\beta_{t0} + \sum_{k=1}^{p}\sum_{j=1}^{p} z_{ij}v_{kj}\beta_{tk}\right)}{\sum_{m=1}^{s} exp\left(\beta_{m0} + \sum_{k=1}^{p}\sum_{j=1}^{p} z_{ij}v_{kj}\beta_{mk0}\right)} \tag{20}$$

where $i = 1, \ldots, s$; $j = 0, \ldots, p$; $t = 1, \ldots, s$ and $\beta_{sj} = 0$.

Setting $\gamma_{tj} = \sum_{k=1}^{p} v_{kj}\beta_{tk}$, the PCLR model, with linear discriminant functions, extended to polytomous responses is given by:

$$\left(G_t|\mathbf{Z}\underline{\mathbf{v}}_i\right) = \frac{exp\left(\beta_{t0} + \sum_{j=1}^{p} z_{ij}\gamma_{tj}\right)}{\sum_{i=1}^{s} exp\left(\beta_{i0} + \sum_{j=1}^{p} z_{ij}\gamma_{mj}\right)} \tag{21}$$

The Principal Components Quadratic Logistic Regression (PCQLR) is given by:

$$Q\left(G_k|\mathbf{Z}\underline{\mathbf{v}}_i\right) = \frac{exp\left(\underline{\chi}_k\right)}{\sum_{i=1}^{s} exp\left(\underline{\chi}_i\right)} \tag{22}$$

where $\underline{\chi}_k = \chi_{k0} + \sum_{i=1}^{p} z_{ij}\gamma_{kj}^2 + \sum_{i=p+1}^{(p-1)\frac{p}{2}} z_{ij}\gamma_{kj'}\gamma_{kj''} + \sum_{i=(p-1)\frac{p}{2}+1}^{(p-1)\frac{p}{2}+p} z_{ij}\,\gamma_{kj}$; $\underline{\chi}_s = \mathbf{0}$ and $k = 1, \ldots, s-1$; $j, j'' = 1, \ldots, p$; $j' = 1, \ldots, p-1$.

In order to estimate the parameters, one can apply the Maximum Likelihood Method. In the dichotomous case, [1] also propose two methods to solve the problem of choosing the optimum principal components that should be included in the model. In this paper we have used the first *q* principal components, with the largest variances, given by the eigenvalues. However, the interested reader should note that, according to [16], principal

components with small eigenvalues can be as important as those with large eigenvalues. A decision rule for discarding principal components, in linear regression, is given by [11]. Furthermore, we should taking into account that multicollinearity affects the accuracy of the parameters estimation, more precisely on the estimation of their standard errors, but it does not affect the performance, in terms of correct classifications.
In this paper the purpose is only to investigate the principal components model's classificatory performance in polytomous cases, using linear and quadratic forms, for practical purposes. In order to formulate the model, the first step was to obtain the principal components of the covariates. We have used the first q principal components, with the largest variances, including principal components in the natural order, given by the explained variability. In the sequence, we fitted the quadratic logistic model, using the selected principal components as covariates. With respect to the QLR model, we propose to use as covariates the first q principal components, with the largest variances, of the $\left[(s-1)(p+1)\left(\frac{p}{2}+1\right)\right]$ matrix I(Γ) , whose elements are given by:

$$\frac{\partial^2 L(\boldsymbol{\Gamma})}{\partial\gamma_{jm}\gamma_{jm'}} = -\sum_{i=1}^{n} x_{m'i}x_{mi}\left[Q\left(G_j|\mathbf{x}_i\right)\right]\left[1-Q\left(G_j|\mathbf{x}_i\right)\right] \tag{23}$$

and

$$\frac{\partial^2 L(\boldsymbol{\Gamma})}{\partial\gamma_{jm}\gamma_{j'm'}} = \sum_{i=1}^{n} x_{m'i}x_{mi}\left[Q\left(G_j|\mathbf{x}_i\right)\right]\left[Q\left(G_{j'}|\mathbf{x}_i\right)\right] \tag{24}$$

where j , j' = 1 , 2 , ... , (s – 1) and m , m' = 1 , 2 , ... , p. The parameter estimation follows the same lines as that taken by the CLR model with linear discriminant functions.

5. APPLICATIONS

In this section we consider two benchmark data sets, taken from the trade literature. Iris Data, taken from [10], and Fatty Acid Composition Data, taken from [8]. We have applied the CLR model, PCLR model, QLR model and PCQLR model to both data sets. A computer program that implements the approaches previously described was written in *Visual Basic 6.0* and was run on *HP Pavilion dv6* computer. The purpose is to compare the results provided by the four models, given by the Correct Classification Rate (CCR). The results achieved, in terms of correct classification rates, are given in the sequence.
Example 1: Iris Data. There are three groups: Iris Setosa (G_1), Iris Versicolor (G_2) and Iris Virginica (G_3). For each group there are 50 observations and four independent variables: Sepal Length, Sepal Width, Petal Length and Petal Width, all measured in *mm*. The reference group is Iris Virginica. It is well known that two groups, Iris Versicolor and Iris Virginica, overlap and form a cluster completely separated from Iris Setosa. Furthermore, a high correlation (r = 0.9629) between the Petal Lenght and Petal Width was found for the three groups. The cross-correlation matrix is given by:

$$R_1 = \begin{bmatrix} 1 & -0.1176 & 0.8718 & 0.8179 \\ -0.1176 & 1 & -0.4284 & 0.3661 \\ 0.8718 & -0.4284 & 1 & 0.9629 \\ 0.8179 & 0.3661 & 0.9629 & 1 \end{bmatrix}$$

There is no MLE for the CLR model. In this example, the HLR model has two discriminant functions and 10 parameters. The QLR model has 30 parameters. Table 1 displays the principal components and their cumulative percentage of the total variance. In order to build the PCLR and PCQLR models, two principal components were selected. In this case, the PCLR model requires six parameters and the PCQLR model requires 12 parameters. The correct classification rates for HLR and PCLR models are summarized in Table 2. The correct classification rates for QLR and PCQLR models are summarized in Table 3. It should be noted that the classificatory performance of PCQLR model, with 12 parameters, was equal to the classificatory performance of the QLR model, with 30 parameters. Furthermore, PCQLR model had better results than PCLR model. When we compare one model with the other, we should to take in mind that the purpose in this case is to reduce the number of unknown parameters, without loss of accuracy.

Table 1. Classification Matrix. Iris data. Linear Discriminant Functions.

Model	Observed Group	Allocated Group		
		G_1	G_2	G_3
HLR	G_1	1.00	0.00	0.00
	G_2	0.00	0.98	0.02
	G_3	0.00	0.02	0.98
PCLR (1 p.c.)	G_1	1.00	0.00	0.00
	G_2	0.00	0.88	0.12
	G_3	0.00	0.10	0.90

Table 2. Iris data. Variances (eigenvalues)

Variance (λ)	2.9185	0.9140	0.1468	0.0207
Cumulative Percentage of Total Variance	72.96	95.81	99.48	100

Table 3. Classification Matrix. Iris data. Quadratic Discriminant Functions.

Model	Observed Group	Allocated Group		
		G_1	G_2	G_3
QLR	G_1	1.00	0.00	0.00
	G_2	0.00	0.98	0.02
	G_3	0.00	0.02	0.98
PCQLR (3 p.c.)	G_1	1.00	0.00	0.00
	G_2	0.00	0.98	0.02
	G_3	0.00	0.02	0.98

Example 2: Fatty Acid Data. There are 120 observations, five groups and seven variables, representing the percentage levels of seven fatty acids, namely palmitic (x_1), stearic (x_2), oleic (x_3), linoleic (x_4), linolenic (x_5), eicosanoic (x_6) and eicosenoic (x_7) acids. In this paper we consider five groups: rapeseed (G_1), sunflower (G_2), peanut (G_3), corn (G_4) and pumpkin (G_5) oils. In this paper the reference group is (G_1) (pumpkin oil). The original data set have eight groups, and the complete table of the original data can be found in [8]. There is a high correlation between oleic and linoleic acids ($r = -0.9565$). Table 4 displays the classification matrix for the QLR and PCQLR models. Table 5 displays the principal components and their cumulative percentage of the total variance. Table 6 displays the classification matrix for the QLR and PCQLR models. In this case, there is no MLE for the CLR model; the HLR model has four discriminant functions and 32 unknown parameters. The QLR model involves 144 unknown parameters. Keeping four principal components, the PCLR model involves 20 parameters, and the PCQLR model involves 60 parameters. Table 4 displays the principal components and their cumulative percentage of the total variance.

Table 5 displays the classification matrix for the HLR and PCLR models and Table 6 displays the classification matrix for the QLR and PCQLR models. In this case, the PCQLR model, with 60 parameters, had better performance than HLR model, and similar performance to the QLR model, with 144 parameters.

Table 5. Classification Matrix. Fatty acid data. Linear Discriminant.

Model	Observed Group	Allocated Group				
		G_1	G_2	G_3	G_4	G_5
HLR	G_1	0.64	0.00	0.00	0.00	0.36
	G_2	0.00	0.95	0.00	0.00	0.05
	G_3	0.00	0.00	1.00	0.00	0.00
	G_4	0.00	0.00	0.00	1.00	0.00
	G_5	0.15	0.00	0.05	0.05	0.75
PCLR (6 p.c.)	G_1	0.64	0.00	0.00	0.00	0.36
	G_2	0.00	0.95	0.00	0.00	0.05
	G_3	0.00	0.00	0.96	0.00	0.04
	G_4	0.00	0.00	0.00	0.80	0.20
	G_5	0.17	0.06	0.03	0.06	0.68

Table 5A. Fatty acid data. Variances (eigenvalues)

Variance (λ)	3.9092	1.0842	0.9325	0.7866	0.2053	0.0811	0.0001
% of Total Variance	55.85	71.84	84.66	95.90	98.83	99.99	100

According to [8], the Principal Component Analysis (PCA) was successful to distinguish clusters of different oil samples, but it is not suitable for an automatic prediction of vegetable oil classes, because PCA requires a visual inspection and the final decision has to be made by an expert. Based on the results achieved by the PCLR and PCQLR models, we believe that the PCA can be a useful tool to develop and implement an automated method for classification of vegetable oils, because it gives an accurate estimation of the parameters of a polytomous quadratic logistic model, with a high classificatory performance.

Table 6. Classification Matrix. Fatty acid data. Quadratic Discriminant.

Model	Observed Group	Allocated Group				
		G_1	G_2	G_3	G_4	G_5
QLR	G_1	0.82	0.00	0.00	0.00	0.18
	G_2	0.00	1.00	0.00	0.00	0.00
	G_3	0.00	0.00	1.00	0.00	0.00
	G_4	0.00	0.00	0.00	1.00	0.00
	G_5	0.00	0.00	0.00	0.00	1.00
PCQLR (6 p.c.)	G_1	0.73	0.00	0.00	0.00	0.27
	G_2	0.00	1.00	0.00	0.00	0.00
	G_3	0.00	0.00	1.00	0.00	0.00
	G_4	0.00	0.00	0.00	0.90	0.10
	G_5	0.00	0.03	0.00	0.05	0.92

6. CONCLUSION

The purpose of this paper was simply to develop and implement a simple and direct generalization for the Quadratic Logistic Regression Model, for polytomous response, which allows the reduction of the number of unknown parameters in the problem, and to explore the performance of the model when compared to the Classical Logistic Regression model with linear discriminant functions. This paper does not intend to give a detailed explanation of theoretical aspects which involves neither Principal Components Analysis (PCA) nor the quadratic logistic model. In order to solve the problem that arises when there are multicollinearity and a large number of unknown parameters, we have used the PCA, as well a generalization of the Hidden Logistic Regression Model (HLR), to estimate the unknown parameters in case of complete separation. We have concentrated on a comparison of classical logistic model to quadratic logistic model, with estimates obtained via PCA. We can see that the PCA allows the reduction of the number of dimensions, and of unknown parameters, in a polytomous quadratic logistic model, with continuous covariates and avoiding the multicollinearity of these variables, without loss of accuracy. Furthermore, this approach provides a simple and easy-to-implement solution to the problem that arises when there is multicollinearity among the independent variables. For practical purposes, the main advantage of the HLR model is the existence and uniqueness of estimators, and, in addition, it involves neither arbitrary data manipulation nor complicated modifications to both classical and quadratic logistic models. Furthermore, there are not any computational difficulties to implement the referred approaches. With respect to the performance, we can see that the Quadratic Logistic Regression Model and the Principal Components Quadratic Logistic Model can provide better classification rates than the Classical Logistic Regression Model. The results achieved suggest that the proposed approach is a promising alternative to the classical logistic regression model, when a large number of dimensions have to be considered.

ACKNOWLEDGEMENTS
The author would like to thank very much the two anonymous referees for their helpful and valuable comments and suggestions.

RECEIVED DECEMBER, 2014
REVISED JUNE, 2015

REFERENCES

[1] AGUILERA, A. M., ESCABIAS, M., VALDERRAMA, M. J., (2006): Using principal components for estimating logistic regression with high-dimensional multicollinear data. **Comput. Statist. & Data Anal. 55**, 1905-1924.
[2] ALBERT, A. AND ANDERSON, J.A., (1984): On the existence of maximum likelihood estimates in logistic regression models. **Biometrika,** 71, . 1-10.
[3] ANDERSON, J. A., (1975): Quadratic logistic discrimination. **Biometrika,** 62, . 149-154.
[4] ANDRUSKI-GUIMARÃES, I. and CHAVES-NETO, A., (2009): Estimation in polytomous logistic model: comparison of methods. **Journal of Industrial and Management Optimization,** 5,. 239-252.
[5] ANDRUSKI-GUIMARÃES, I., (2011): Principal Component Analysis Alied to Polytomous Quadratic Logistic Regression. **Proc. 58th World Statistical Congress**, Dublin.
[6] BARKER, L. and BROWN, C., (2001): Logistic regression when binary predictor variables are highly correlated, **Statist. Med.,** 20, . 1431-1442.
[7] BEGG, C. B. and GRAY, R., (1984): Calculation of polychotomous logistic regression parameters using individualized regressions. **Biometrika,** 71, . 11-18.

[8] BRODNJAK-VONČINA, D., KODBA, Z.C. and NOVIČ, C., (2005): Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. **Chemometr. Intell. Lab. Syst., 75**, . 31-43.
[9] COPAS, J. B., (1988): Binary regression models for contaminated data. With discussion. **J. R. Stat. Soc. Ser. B Stat. Methodol.,** 50, . 225-265.
[10] FISHER, R. A., (1936): The use of multiple measurements in taxonomic problems. **Annals of Eugenics,** 3, . 179-188.
[11] FOUCART, T., (2000): A decision rule for discarding principal components in regression. **J. Statist. Plann. Inference,** 89**,** . 187-195.
[12] GERVINI, D., (2005): Robust adaptive estimators for binary regression models. **J. Statist. Plann. Inference,** 131, . 297-311.
[13] HEINZE, G. and SCHEMPER, M., (2002): A solution to the problem of separation in logistic regression. **Statist. Med.,** 21, . 2409-2419.
[14] HOSMER, D W. and LEMESHOW, S., (2000): **Applied Logistic Regression**. Wiley , New
[15] HUBERT, M. and VAN DRIESEN, K., (2004): Fast and robust discriminant analysis. **Comput. Statist. & Data Anal.,** 45, . 301-320.
[16] JOLLIFFE, I. T., (1982): A note on the use of principal components in regression. **Applied Statistics, 31, 3**, . 300-303.
[17] KODZARKHIA, N., MISHRA, G. D. and REIERSOLMOEN, L., (2004): Robust estimation in the logistic regression model. **J. Statist. Plann. Inference**, 98,. 211-223.
[18] LESAFFRE, E. and ALBERT, A., (1989): Partial separation in logistic discrimination**. J. R. Stat. Soc. Ser. B Stat. Methodol., 51**, 1, . 109-116.
[19] MASSY, W. F., (1965): Principal component regression in exploratory statistical research. **J. Amer. Statist. Assoc., 60**, . 234-246.
[20] MCLACHLAN, G. J., (2004): **Discriminant Analysis and Statistical Pattern Recognition**. John Wiley & Sons, Inc., Hoboken, New Jersey, U.S.A.

[21] ROUSSEEUW, P. J. and CHRISTMANN, A., (2003): Robustness against separation and outliers in logistic regression, **Comput. Statist. & Data Anal.** 43,. 315-332.