LII Simpósio Brasileiro de Pesquisa Operacional
João Pessoa-PB, 3 a 5 de novembro de 2020

QUADRATIC LOGISTIC REGRESSION MODEL AND PRINCIPAL CURVES APPLIED TO DISCRIMINANT ANALYSIS: COMPARISON OF METHODS

Inácio Andruski-Guimarães
DAMAT – Departamento Acadêmico de Matemática
UTFPR – Universidade Tecnológica Federal do Paraná
Av. Sete de Setembro, 3165 Curitiba Paraná Brasil
andruski@utfpr.edu.br

Zaudir Dal Cortivo
UNINTER – Centro Universitário Internacional
Rua Luiz Xavier, 103 Curitiba Paraná Brasil
zdalcortivo@gmail.com

Jair Mendes Marques
UTP – Universidade Tuiuti do Paraná
Rua José Nicco, 179 Curitiba Paraná Brasil
jair.marques@utp.br

ABSTRACT

Studies on logistic regression usually consider the model with linear discriminant functions, but there are situations where quadratic discriminant functions work better. However, the quadratic logistic regression model involves the estimation of a great number of parameters. Another problem is that the model may not work when the data set has no overlapping. To deal with this issue we propose an extension of the hidden logistic model, whose advantage is the existence of estimators, even when there is no overlap. Principal curves, that pass through the middle of a *p*-dimensional data set, providing a summary of the data in $\mathbb{R}^p$, have been used for discriminant analysis. We propose an algorithm which extracts principal curves from the Fisher discriminant matrix, using the *k*-segments algorithm. This algorithm generates a polygonal line used to classify an observation in the class whose polygonal line has the shortest Euclidean distance by orthogonal projection.

KEYWORDS. Logistic Regression. Principal Curves. Discriminant Analysis.

TOPICS
EST&MP – Estatística e Modelos Probabilísticos
BDA – Big Data e Analytics
PM – Programação Matemática

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

1. Introduction

The logistic regression model is a method applied to model and evaluate the relationship between a categorical dependent variable and a set of explanatory variables, continuous or discrete. The performance of the Logistic Regression Model has been reported in many studies involving credit scoring, bankruptcy prediction, marketing applications and cancer classification, to cite just a few. Although almost all papers on logistic regression have only considered the classical logistic regression model with linear discriminant functions, there are situations where quadratic discriminant functions are useful, and work better, in terms of correct classification rates. However, the quadratic logistic regression model involves the estimation of a great number of unknown parameters, and this leads to computational difficulties, in terms of memory and computing time requirements, especially when there are a great number of explanatory variables. Another problem is that, while the logistic regression model works well for many cases, it may not work when the data set has no overlapping, a frequent problem in the logistic regression. To deal with this issue 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.

A principal curve is a smooth one-dimensional curve that passes through the middle of a p-dimensional data set, providing a non-linear summary of the data. Principal curves provide a smooth one-dimensional approximate curve for a set of data in R^p, and have been used for discrimination and classification. We use an algorithm in which the principal curves are extracted from the Fisher discriminant matrix of each class, using the k-segments algorithm. Segments are adjusted and connected in order to form a polygonal line; in the sequence, new segments are inserted at each iteration until an optimization criterion is reached. The proposed algorithm generates a polygonal line which is used to classify a new observation sample, labelling the observation in the class whose polygonal line has the shortest Euclidean distance by orthogonal projection.

This paper is organized as follows. First, we revisit the Classical Logistic Regression model (CLR). Next, we present an overview of the Quadratic Logistic Regression model (QLR) for polytomous response. Then we give a brief review of Fisher Linear Discriminant Function (FLDF). After that, we formulate the basic procedures in order to calculate Principal Curves (PC) and their application in discriminant analysis. Last, we apply the models on data sets taken from the literature and compare their performance in terms of correct classification rates. Finally, we give a brief conclusion about the results achieved.

2. Classical Logistic Regression Model

Let us consider a sample of n independent observations, available from the groups G_1 , … G_s , and a vector of explanatory variables, $\mathbf{x} = (x_0 , x_1 , \ldots x_p)$, where $x_0 \equiv 1$, 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:

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

The Classical Logistic Regression (CLR) Model assumes that the posterior probabilities are given by:

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

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

where $\mathrm{B}_i\mathbf{x} = \beta_{i0} + \sum_{j=1}^{p} \beta_{ij} x_{j_}$, $i = 1 , \ldots , s - 1$ and $\mathrm{B}_s = 0$. The *s-th* group is called reference group. The model (1) 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 $\mathbf{Y} = (Y_1 , \ldots , Y_n)^{\mathrm{T}}$ and $Y_i = (y_{1i} , \ldots , y_{si})$, with $y_{ki} = 1$, if $Y = k$, and $y_{ki} = 0$, otherwise. 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:

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

The Maximum Likelihood Estimator (MLE) $\widehat{\mathbf{B}}$ is obtained by setting the derivatives (4) to zero and solving for $\mathbf{B}$. 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. Albert & Anderson (1984) suggested a sample classification into three categories: Overlap, complete separation and quasi-complete separation, when perfect prediction occurs only for a few groups of observations. They also proved that the MLE do not exist for complete and quasi-complete separation. In this case, existing iterative methods fails to converge, or give a wrong answer.

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:

$$Q(G_k|\mathbf{x}) = \frac{exp(\Theta_k)}{\sum_{i=1}^{s} exp(\Theta_i)} \tag{5}$$

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

The model (5) involves $[(s - 1)(p + 1)]\left(1 + \frac{p}{2}\right)$ unknown parameters whose estimation 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 render an unworkable problem, in terms of memory and computing time requirements, so that a reduction dimension method can be useful to way out of this problem. As pointed out by Anderson (1975), the quadratic term also can be written as:

$$\Theta_k = \alpha_{k0} + \mathbf{x}^{\mathrm{T}} \mathbf{\Omega}_k \mathbf{x} + \alpha_k^{\mathrm{T}} \mathbf{x} \tag{6}$$

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 Anderson (1975), gives a quadratic term with a reduced number of parameters. This approximation is given by the spectral decomposition:

$$\mathbf{\Omega}_k = \sum_{j=1}^{p} \lambda_{jk} l_{jk} l_{jk}^{\mathrm{T}} \tag{7}$$

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:

$$\mathbf{\Omega}_k \cong \lambda_k l_k l_k^{\mathrm{T}} \ . \tag{8}$$

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

Because this approach is not convenient for computing, an alternative parameterization is suggested by Anderson (1975):

$$\Theta_k = \alpha_{k0} + \mu_k \left(d_k^{\mathrm{T}} \mathbf{x}\right)^2 + \alpha_k^{\mathrm{T}} \mathbf{x} \tag{9}$$

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

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. For example, if the explanatory variables are binary, the diagonal terms of $\mathbf{\Omega}$ are zero. We propose to use as covariates the principal components of the $(s - 1)(p + 1)$ matrix $\mathbf{I}(\Theta)$ whose elements are given by:

$$\frac{\partial^2 L(\Theta)}{\partial \beta_{jm} \partial \beta_{jm'}} = -\sum_{i=1}^{m} x_{m'i} x_{mi} \left[Q\left(G_j | \mathbf{x}\right)\right]\left[1 - Q\left(G_j | \mathbf{x}\right)\right] \tag{10}$$

and

$$\frac{\partial^2 L(\Theta)}{\partial \beta_{jm} \partial \beta_{j'm'}} = -\sum_{i=1}^{m} x_{m'i} x_{mi} \left[Q\left(G_j | \mathbf{x}\right)\right]\left[Q\left(G_{j'} | \mathbf{x}\right)\right] \tag{11}$$

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

The QLR model, as the CLR model, is not immune to complete, or quasi-complete, separation. In order to solve this issue Andruski-Guimarães (2016) suggested a generalization of the Hidden Logistic Regression (HLR) Model. The HLR Model assumes that the true response of a logistic regression model is unobservable, but there is an observable variable that is related to this response. We consider n unobservable independent variables $T_1 , \dots , T_n$, where each T_i has s values $\gamma_1 , \dots , \gamma_s$. Thus, we observe a variable $Y_i = j$, with $(Y_i = j | T_i = \gamma_k) = \delta_{jk}$, where $\delta_{jj} = max_{k=1,\dots,s}\{\delta_{jk}\}$ and $\sum_{j=1}^{s} \delta_{jk} = 1$. For comparison purposes, 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:

$$\tilde{y}_{ij} = \sum_{k=1}^{s} y_{ik}\delta_{kj} \ . \tag{12}$$

The purpose is to maximize the likelihood function:

$$L(\Theta|\tilde{\mathbf{Y}}, \mathbf{x}) = \prod_{i=1}^{n}\prod_{k=1}^{s}[Q(T_k|\mathbf{x}_i)]^{\tilde{y}_{ki}} \tag{13}$$

The log-likelihood function becomes:

$$\ell(\Theta|\tilde{\mathbf{Y}}, \mathbf{x}) = \sum_{i=1}^{n}\left[\sum_{j=1}^{s-1}\tilde{y}_{ji}\Theta_j - ln\left(1 + \sum_{j=1}^{s-1} exp(\Theta_j)\right)\right] \ . \tag{14}$$

Unlike the MLE for the CLR model, the MLE for the HLR model always exists, even when there is complete separation. We consider that the probability of observing the true status, given by $(Y_i = j|T_i = \gamma_j) = \delta_{jj}$, should be higher than 0.5, this is, $0.5 < \delta_{jj} < 1$. Furthermore, $\sum_{k=1,k\neq j}^{s}\delta_{jk} < \delta_{jj}$. Our default choice will be $\delta = 0.99$ and set $\delta_{jj} = \delta$ and $\delta_{jk} = \frac{1-\delta}{s-1}$.

Principal Component Analysis (PCA) is a method for dimensional reduction, and Principal Components are a set of linear combinations of the original covariates that retain most of the covariance structure of the data set. In order to reduce the dimension in the problem, with continuous covariates, as well as the computational costs, in terms of time for the parameter estimation, and without loss of accuracy, Andruski-Guimarães (2016) proposed to use as covariates a set of principal components of the exploratory variables. A generalization of the Principal Components Logistic Regression model (PCLR) for polytomous responses does not require a complex formulation. It begins by computing the covariance matrix **S**. Then the matrix **S** can be written as $\mathbf{S} = \mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V}$, where $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ and **V** is the matrix whose columns are the eigenvectors of **S** associated to the eigenvalues $\lambda_1, \ldots, \lambda_p$. Let $\mathbf{Z} = \mathbf{XV}$ so that $\mathbf{X} = \mathbf{ZV}^{\mathrm{T}}$. In this case:

$$x_{ik} = \sum_{j=1}^{p} z_{ij}v_{kj} \tag{15}$$

so that

$$P(G_t|\mathbf{Z}\mathbf{v}_i) = \frac{exp\left(\beta_{t0} + \sum_{k=1}^{p}\sum_{j=1}^{p} z_{ij}v_{kj}\beta_{tk}\right)}{\sum_{m=1}^{s} exp\left(\beta_{m0} + \sum_{k=1}^{p}\sum_{j=1}^{p} z_{ij}v_{kj}\beta_{mk}\right)} \tag{16}$$

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:

$$P(G_t|\mathbf{Z}\mathbf{v}_i) = \frac{exp\left(\beta_{t0} + \sum_{j=1}^{p} z_{ij}\gamma_{tj}\right)}{\sum_{i=1}^{s} exp\left(\beta_{i0} + \sum_{j=1}^{p} z_{ij}\gamma_{mj}\right)} \quad . \tag{17}$$

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

$$\Psi(G_k|\mathbf{Z}\mathbf{v}_i) = \frac{exp(\boldsymbol{\chi}_k)}{\sum_{i=1}^{s} exp(\boldsymbol{\chi}_i)} \tag{18}$$

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

The parameter estimation follows the same lines as that taken by the Quadratic Logistic Regression Model (QLR). One issue that has to be addressed is the choice of the principal components that will be used as covariates of the model obtained. We have used the first q principal components, with the largest variances, given by the eigenvalues. However, we should note that principal components with small eigenvalues can be as important as those with large eigenvalues. Furthermore, we should take 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 rates, which is an advantage. A disadvantage is that the estimates of the eigenvalues of **S** can be biased, and this bias is most pronounced when the eigenvalues of **S** tend toward equality, being less severe when they are highly disparate. Another disadvantage is that the classical PCA is not resistant to outlying observations. Approaches to deal with these issues can be found in Chipman & Gu (2005) and Croux *et al* (2013), to cite just a few.

4. Fisher Linear Discriminant Function

Let Y be a polytomous dependent variable, with s outcomes, a sample of n independent observations, available from the groups $G_1, \ldots G_s$, and a vector of explanatory variables given by $\mathbf{x}^{\mathrm{T}} = (x_1, \ldots x_p)$, in a matrix form given by:

$$\mathrm{X} = \begin{bmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{pn} \end{bmatrix}.$$

A matrix of discriminant spaces $Y = \mathbf{A}^{\mathrm{T}}\mathbf{x}$ contains the discriminant coefficients that maximize the ratio of the variances between the classes and within the classes, given by:

$$\frac{\boldsymbol{\omega}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{B}}\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{W}}\boldsymbol{\omega}} \tag{19}$$

where $\mathbf{S_B}$ is the covariance matrix between classes, given by:

$$\mathbf{S_B} = \sum_{i=1}^{s} (\bar{x}_i - \bar{\mathbf{x}})(\bar{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \tag{20}$$

where $\bar{x}_i$ is the mean vector for the *i*-th group and $\bar{\mathbf{x}}$ is the mean vector of the data set. The matrix $\mathbf{S_W}$ is the covariance matrix within each class, given by:

$$\mathbf{S_W} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{\mathbf{x}})^{\mathrm{T}} \tag{21}$$

where x_{ij} is the *j-th* observation of the *i-th* group. If $\boldsymbol{\omega}$ is a vector that maximizes the ratio of the variances, then $Y = \mathbf{x}^{\mathrm{T}}\mathbf{x}$ is the Fisher Linear Discriminant Function (FLDF). The function given by $Y = \boldsymbol{\omega}^{\mathrm{T}}\mathbf{x}$ allows an observation to be addressed to one of the *s* groups based on the discriminant score given by $\boldsymbol{\omega}^{\mathrm{T}}\mathbf{x}$. An observation $\mathbf{x}$ is addressed to the *k*-th group if, for all $i \neq k$,

$$\left|\boldsymbol{\omega}^{\mathrm{T}}\mathbf{x} - \boldsymbol{\omega}^{\mathrm{T}}\bar{\boldsymbol{x}}_{\boldsymbol{k}}\right| \leq \left|\boldsymbol{\omega}^{\mathrm{T}}\mathbf{x} - \boldsymbol{\omega}^{\mathrm{T}}\bar{\boldsymbol{x}}_{\boldsymbol{i}}\right| . \tag{22}$$

An advantage of the FLDF is the computational simplicity. Furthermore, the centroids and distances require easy calculations. A disadvantage is that, using centroids, groups with large variances are penalized.

5. Principal Curves

According to Hastie & Stuetzle (1989), principal curves are smooth one-dimensional curves that pass through the middle of a *p*-dimensional data set, providing a non-linear summary of the data and can be considered a generalization of linear principal components. They also defined principal curves to be those smooth curves that are self-consistent for a distribution or data set. Furthermore, in a principal curve, each point is the average of all data points that project to it. Given a *p*-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ with density $g_X(x)$, $f(\lambda)$ is a smooth curve (C^∞) in parametrized R^p on a closed interval. For each vector $\mathbf{x}$, $\lambda_{f(\mathbf{x})}$ is defined as the nearest point of the $\mathbf{x}$ curve. The $f(\lambda)$ curve is the principal curve for the distribution of the random vector $\mathbf{x}$ if:

$$E\left(\mathbf{x}|\boldsymbol{\lambda}_{\boldsymbol{f(\mathbf{x})}} = \boldsymbol{\lambda}\right) = f(\lambda) = \begin{bmatrix} f_1(\lambda) \\ \vdots \\ f_p(\lambda) \end{bmatrix} . \tag{23}$$

In (23) $\lambda_f : R^p \rightarrow R$ is the projection index and the function $f(\lambda)$ is the average of all orthogonally projected points on the curve, property known as self-consistency. Hastie and Stuetzle (1989) showed that a principal curve is a critical point of the square distance function, given by:

$$\sum_{j=1}^{p}\left(x_j - f_j(\lambda)\right)^2 . \tag{24}$$

In this sense, finding the projection of the data on the curve is equivalent to find the value of λ that minimizes the distance between the curve $f(\lambda)$ and $\mathbf{x}$, that is:

$$\lambda = argmin_{\lambda}\|f(\lambda) - \mathbf{x}\| . \tag{25}$$

According to Duchamp & Stuetzle (1996), the continuous multivariate distributions have infinite principal curves.

An algorithm for finding principal curves usually begins with the first linear principal component drawn from the data. The data are orthogonally projected on the principal component and then all points within an interval are used to calculate the average. The first approximation for the principal curve is obtained by using the averages of each sub-interval. The points are projected on the curve, the average of each sub-interval is calculated and the new approximation of the curve is obtained. This process is repeated until the desired convergence.

Definitions, and alternative methods, for estimating principal curves have been suggested in many studies. An approach presented by Kramer (1991) used auto-associative neural networks in order to obtain non-linear principal components. A probabilistic approach, given by Tibshirani (1992), proposed to minimize a penalized log-likelihood function. The Polygonal Line Algorithm, proposed by Kgl et al. (2000), starts with a single line and adds new vertices at each iteration. The position of the lines is updated for all vertices, so that the expected values of the squares of the distances of the points projected on the curve are minimized. The *k*-segments algorithm, proposed by Verbeek et al (2002), uses a probabilistic definition to find the principal components by maximizing a function in a similar way to the Tibshirani function. Line segments are fitted and connected to form polygonal lines, and then new segments are inserted until a performance criterion is met.

We use an algorithm proposed by Dal Cortivo & Marques (2015) in which the principal curve is extracted from the Fisher discriminant matrix of each class, using the *k*-segments algorithm. This algorithm generates the principal curve and a polygonal line, used to classify a new observation, labelled in the group whose polygonal line has the shortest Euclidean distance by orthogonal projection. The *k*-segments algorithm brings the center of the principal curve to the origin of the coordinate space, because it removes the average of the set of sampling data. In this way, all classes have the same center, that is, the origin of the orthogonal coordinate system. However, this fact may lead to a low performance classifier, because the algorithm uses the distance measurement of the vector $\mathbf{x}$ to the polygonal line of each group to perform the classification. In order to classify a vector $\mathbf{x}$ the polygonal line is adjusted for each group. If x_{ij} is the orthogonal projection on the i-th segment of j-th polygonal line, the vector $\mathbf{x}$ is addressed to the group with smallest distance to x_{ij}.

The following steps can describe the proposed classifier algorithm:

1. Start with the Fisher Linear Discriminant Function (FLDF) and get the matrix of discriminant scores, $\mathbf{A}$, that is, the matrix of eigenvectors that contains the discriminant coefficients that maximize the ratio of the variances between the groups and within them.
2. Perform $Y = \mathbf{A}\mathbf{x}^{\mathrm{T}}$ and separate Y per group G_i , $i = 1, \ldots, s$.
3. Apply the *k*-segments algorithm for each group, in order to obtain the matrices of nodes and their coordinates.
4. Determine the orthogonal point of projection, x_{ij} , $j = 1, \ldots, r$, where r is the number of segments per group and the projection of the vector $\mathbf{x}$ on each segment of G_i , $i = 1, \ldots, s$.
5. Calculate the Euclidean distances between x_{ij} and $\mathbf{x}$ for each segment of G_i and perform the value $d_i = min_{j = 1, \ldots, r} \| \mathbf{x} - x_{ij} \|$.
6. The sample observation $\mathbf{x}$ is addressed to the group that contains the shortest distance.

The line generated by the algorithm can determine a non-linear adjustment to elements of the group. This is an advantage because self-consistent points, which are also midpoints, form the principal curves.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

6. Applications

To compare the Quadratic Logistic Regression Model (QLR) with Principal Curves Model (PC) we have applied both models to two benchmark data sets, taken from the trade literature: Iris Data, taken from Fisher (1936), and Thyroid Data, taken from UCI repository of machine learning databases. The results achieved, in terms of correct classification rates, are given in the sequence.

Example 1: Iris Data. There are three groups of Iris flowers: 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 (x_1), Sepal Width (x_2), Petal Length (x_3) and Petal Width (x_4). In this paper, the reference group is Iris Virginica. This data set is an example of what we used to refer as partial separation. 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 variables x_3 and x_4 was found for the three groups.

The QLR model has 30 parameters. The correct classification rates for QLR and PC models are summarized in Table 1. The three models showed a good performance, which can be explained because the data set does not have a spherical distribution. 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.

Table 1 – Classification Matrix. Iris Data.

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
PC	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: Thyroid Data. There are 215 observations, three groups, Normal (G_1), Hypo (G_2) and Hyper (G_3) Functioning, and five variables, representing the T3-Resin Uptake Test (x_1), Total Serum Thyroxin (x_2), Total Serum Triiodothyronine (x_3), Basal Thyroid-Stimulating Hormone (TSH) (x_4), Maximal Absolute Difference of TSH value after injection of 200 micrograms of thyrotropin-releasing hormone as compared to the basal value (x_5). The reference group is hyper functioning (G_3). In this case, the QLR model involves 42 unknown parameters. Table 2 displays the classification matrix for the QLR, PC and PCQLR models. The QLR model had better performance than PCQLR model and PC model, however, 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 and classificatory performance.

Galoá { Este trabalho foi publicado utilizando o Galoá proceedings

Table 2 – Classification Matrix. Thyroid Data.

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	0.00	1.00
PC	G_1	0.98	0.01	0.01
	G_2	0.09	0.91	0.00
	G_3	0.00	0.00	1.00
PCQLR (3 p.c.)	G_1	1.00	0.00	0.00
	G_2	0.00	0.96	0.04
	G_3	0.00	0.01	0.99

7. Conclusion

The purpose of this paper was 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 Principal Curves Model, both applied to discriminant analysis. This paper does not intend to give a detailed explanation of theoretical aspects that involves neither Principal Curves nor the Quadratic Logistic Regression Model. In order to deal with the complete separation, we have used a generalization of the hidden logistic regression model and a model based on principal components, in order to reduce the number of parameters. For practical purposes, the main advantage of the hidden logistic regression model is the existence and uniqueness of estimators, and it involves neither arbitrary data manipulation nor complicated modifications to quadratic logistic model. The main advantage of the PC model is it that not only uses a central value, because self-consistent points, that are also midpoints, form the principal curves. The line generated by the algorithm can determine a non-linear adjustment to the elements of each class, which can improve the classification performance. A disadvantage of this approach is that the proposed algorithm requires other algorithms, which can increase the complexity in software development. With respect to the performance, we can see that the Quadratic Logistic Regression Model can provide classification rates slightly higher than the Principal Curves Model can. Furthermore, there are not any computational difficulties to implement the referred approaches. The results achieved suggest that the proposed approaches are a promising alternative to the classical logistic regression model and linear discriminant analysis.

References

Albert, A. & Anderson, J.A. (1984), On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*, pp. 1-10.

Anderson, J. A. (1975), Quadratic logistic discrimination. *Biometrika,* 62, pp. 149-154.

Andruski-Guimarães, I. (2016), Separation and multicollinearity in polytomous quadratic logistic regression model. *Revista Investigación Operacional*, *37*, pp. 13-21.

Chipman, H. A. & Gu, H. (2005), Interpretable dimension reduction. *Journal of Applied Statistics*, *32*, pp. 969-987.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

Croux, C., Filzmoser, P., Fritz, H. (2013), Robust sparse principal component analysis. *Technometrics*, 55, pp. 202-214.

Dal Cortivo, Z. & Marques, J.M. (2015), K-Segments classifier - A non-linear approach for the classification of sampling data. *WSEAS Transactions on Mathematics*, *14*, pp. 301-312.

Duchamp, T. & Stuetzle, W. (1996), Extremal properties of principal curves in the plane. *Annals of Statistics,* 24, pp. 1511-1520.

Fisher, R. A. (1936), The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 3*, pp. 179-188.

Hastie, T. & Stuetzle, W. (1989), Principal Curves. *Journal of the American Statistical Association,* 84, pp. 502-516.

Kgl, B., Krzyzak, A., Linder, T., Kenneth, Z. (2000), Learning and design of principal curves. *IEEE Transactions on Pattern Recognition Analysis and Machine Intelligence, 22* (3), pp. 281-297.

Kramer, M. A. (1991), Nonlinear components analysis using auto associative neural networks. *AIChE Journal, 37* (2), pp. 233-243.

Tibshirani, R. (1992), Principal curves revisited. *Statistics and Computing,* 2 (4), pp. 183-190.

UCI Machine Learning Repository [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science.

Verbeek, J. J., Vlassis, N., Krse, B. (2002), A *k*-segments algorithm for finding principal curves. *Pattern Recognition Letters, 23* (8), pp.1009–1017.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings