Principal Component Analysis Applied to Polytomous Quadratic Logistic Regression

Andruski-Guimarães, Inácio
UTFPR - Universidade Tecnológica Federal do Paraná, Departamento Acadêmico de Matemática
Rua sete de setembro, 3165
80930-201 Curitiba - Paraná - Brasil
E-mail: andruski@utfpr.edu.br

Let us consider a sample of n independent observations, available from the groups $G_1, G_2, ..., G_s$. Let $\underline{\mathbf{x}}$ the vector of explanatory variables, $\underline{\mathbf{x}}^T = (x_0, x_1, ..., x_p)$, 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:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & ... & x_{p1} \\ ... & ... & ... & ... \\ 1 & x_{1n} & ... & x_{pn} \end{bmatrix}$$

Classical Logistic Regression Model

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

$$P(G_k \mid \underline{\mathbf{x}}) = \frac{exp\left(\beta_{k0} + \sum\limits_{j=1}^{p}\beta_{kj}x_j\right)}{\sum\limits_{i=1}^{s} exp\left(\beta_{i0} + \sum\limits_{j=1}^{p}\beta_{ij}x_j\right)}$$

where $k = 1\,,\,2\,,\,...\,,\,s-1$ and $\underline{\mathbf{B}}_s = \mathbf{0}$. In this paper the group s is called reference group. The model involves $(s-1)(p+1)$ unknown parameters. The log-likelihood function is given by:

$$L(\underline{\mathbf{B}} \mid \mathbf{Y}, \underline{\mathbf{x}}) = \sum_{i=1}^{n}\sum_{k=1}^{s} Y_{ki} ln\left[P(G_k \mid \underline{\mathbf{x}}_i)\right]$$

where $\mathbf{Y} = (\mathbf{Y}_1\,,\,...\,,\mathbf{Y}_n)^T$ and $\mathbf{Y}_i = (Y_{1i}, ..., Y_{si})$, with $Y_{ki} = 1$ if $Y = k$, and $Y_{ki} = 0$ otherwise. Thus:

$$\frac{\partial}{\partial \beta_{kj}} L(\underline{\mathbf{B}} \mid \mathbf{Y}, \underline{\mathbf{x}}) = \sum_{i=1}^{n} x_{ij}\left(Y_{ki} - P(G_k \mid \underline{\mathbf{x}}_i)\right)$$

The Maximum Likelihood Estimator (MLE) $\hat{\underline{\mathbf{B}}}$ is obtained by setting the derivatives to zero and solving for $\underline{\mathbf{B}}$. The mostly used iterative method is the Newton-Raphson. In practice, the estimation of unknown parameters is affected by the data's properties. Albert and Anderson (1984) suggested a sample classification into three categories: complete separation, quasi-complete separation and overlap. They also proved that the MLE do not exist for complete and quasi-complete separation. Different approaches to deal with separation can be found in Heinze and Schemper (2002), Rousseeuv and Christmann (2003), for binary response, and Andruski-Guimarães and Chaves-Neto (2009), for polytomous response.

Our approach to solve the multiple group problem is to provide a simple and direct generalization of the Hidden Logistic Regression model, proposed by Rousseeuv and Christmann (2003). We consider

n unobservable independent variables $T_1, \ldots , T_n$, where each T_i has s possible values, $\gamma_i , \ldots , \gamma_s$. Thus, we observe $Y_i = j$ with a $P\left(Y_i = j \mid T_i = \gamma_k\right) = \delta_{jk}$ probability, where $\sum_{j=1}^{s}\delta_{jk} = 1$ and $\delta_{jj} = max_{k=1,\ldots,s}\left\{\delta_{jk}\right\}$.

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

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

Let us keep in mind that in the CLR model, $\delta_{jj} = 1$ and $\delta_{jk} = 0$, if $j \neq k$. The log-likelihood function is given by:

$$L\left(\underline{\theta} \mid \underline{\tilde{\mathbf{Y}}} , \underline{\mathbf{X}}\right) = \sum_{i=1}^{n}\left[\sum_{j=1}^{s-1}\tilde{y}_{ji}\mu_j - ln\left(1+\sum_{j=1}^{s-1}exp\left(\mu_j\right)\right)\right] ,$$

where $\mu_j = \theta_{j0} + \theta_{j1}x_1 + \theta_{j2}x_2 + \ldots + \theta_{jp}x_p$, $j = 1 , 2 , \ldots , s$ - 1.

The MEL estimators are the maximizers of the log-likelihood function, which is strictly concave. The main advantage of the HLR model is that the MEL estimator always exists and it is unique, even when the data set has no overlapping. According to Rousseeuv and Christmann (2003), Copas (1988) found that accurate the estimation of δ_0 and δ_1, in the binary case, is very difficult, unless n is extremely large. The symmetric approach is to chose a constant $\gamma > 0$ and to set $\delta_0 = \gamma$ and $\delta_1 = 1 - \gamma$, where γ is small so that terms in γ^2 can be ignored, and $\delta_0 < \hat{\pi} < \delta_1$, where $\hat{\pi}$, δ_0 and δ_1 are given by $\delta_1 = \frac{1+\hat{\pi}\delta}{1+\delta}$, $\delta_0 = \frac{\hat{\pi}\delta}{1+\delta}$, $\hat{\pi} = max\left\{\delta , min\left(1-\delta\, ;\, \bar{\pi}\right)\right\}$, $\bar{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

For a detailed explanation, and discussion, see Copas (1988) and Rousseeuv and Christmann (2003). In this paper, we consider that the probability of observing the true status, which is given by:

$$P\left(Y_i = j\,|\,T_i = \gamma_j\right) = \delta_{jj},$$

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 $\delta = 0.99$, and set $\delta_{jj} = \delta$ and $\delta_{jk} = \frac{1-\delta}{s-1}$.

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 can be given by:

$$Q\left(G_k \mid \underline{\mathbf{X}}\right) = \frac{exp\left(\underline{\chi}_k\right)}{\sum_{i=1}^{s} exp\left(\underline{\chi}_i\right)}$$

where

$$\underline{\chi}_k = \alpha_{k0} + \sum_{i=1}^{p}\alpha_{ki}x_i^2 + \sum_{i=p+1}^{{}_pC_2}\alpha_{ki}x_{j'}x_{j''} + \sum_{i={}_pC_2+1}^{{}_pC_2+p}\alpha_{ki}x_j$$

$k = 1 , 2 , \ldots , s-1$, $\underline{\chi}_s = \mathbf{0}$, and $j, j'' = 1 , 2 , \ldots , p$, $j' = 1 , 2 , \ldots , p-1$.

The model involves $\left[(s-1)\,(p+1)\right]\left(1+\frac{p}{2}\right)$ unknown parameters and the estimaton of these parameters follows the same lines as that taken by the CLR model. However, as pointed out by Anderson

(1975), 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, large number of parameters should be avoided, because of the risk of over-fitting. The quadratic term also can be written as:

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

where $\mathbf{\Omega}_k = \mathbf{V}_k^{-1} - \mathbf{V}_s^{-1}$, and $\mathbf{V}_k$ is the dispersion matrix in G_k, $k = 1, 2, ..., s-1$.
An approximation, proposed by Anderson (1975), gives a quadratic term with a reduced number of parameters. This approximation is given by the spectral decomposition:

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

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

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

In the sequence, each $l_j^T = (l_{j1}, ..., l_{jp})$ 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:

$$\underline{\chi}_k = \alpha_{k0} + \mu_k (d_k^T \underline{\mathbf{x}})^2 + \alpha_k^T \underline{\mathbf{x}}$$

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

Principal Components Analysis

Let us consider n observations of p continuous variables, given by the matrix $\mathbf{X}$. Since the observations $\underline{\mathbf{x}}$ can be standardized, the sample covariance matrix, $\mathbf{S}$, can be given by:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ & s_{22} & \cdots & s_{2p} \\ & & \ddots & \vdots \\ & & & s_{pp} \end{bmatrix} = \frac{1}{n-1} \mathbf{X}^{\mathrm{T}} \mathbf{X}.$$

The matrix $\mathbf{S}$ can be written as $\mathbf{S} = \mathbf{V}^{\mathbf{T}} \mathbf{\Lambda} \mathbf{V}$, where $\mathbf{\Lambda} = diag\,(\lambda_1, ..., \lambda_p)$ and $\mathbf{V}$ being orthogonal. Let $\mathbf{Z}$ the matrix whose columns are the principal components, given by $\mathbf{Z} = \mathbf{X}\mathbf{V}$, where $\underline{\mathbf{v}}_1, ..., \underline{\mathbf{v}}_p$ are the eigenvectors of the matrix $\mathbf{S}$, associated to the eigenvalues $\lambda_1 \geq ... \geq \lambda_p$, so that the matrix of observations can be written as $\mathbf{X} = \mathbf{Z}\mathbf{V}^{\mathbf{T}}$.

Furthermore, matrices $\mathbf{Z}$ and $\mathbf{V}$ can be written as:

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

and

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & ... & 0 & 0 & ... & 0 \\ 1 & v_{11} & ... & v_{1q} & v_{1(q+1)} & ... & v_{1p} \\ ... & ... & ... & ... & ... & ... & ... \\ 1 & v_{p1} & ... & v_{pq} & v_{p(q+1)} & ... & v_{pp} \end{bmatrix} = \left(\mathbf{V}_{(q)} | \mathbf{V}_{(r)}\right)$$

In order to improve the parameter estimation under multicollinearity, and to reduce the dimension of the problem, Aguilera, Escabias and Valderrama (2006) 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. Furthermore, cf. Barker and Brown (2001), estimates obtained via principal components can have smaller mean square error than estimates obtained through standard logistic regression.

The generalization of the PCLR model for polytomous responses does not require a complex formulation. We begin by computing the covariance matrix $\mathbf{S}$. Then the matrix $\mathbf{X}$ can be written as:

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

so that

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

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

Making $\gamma_{tj} = \sum_{k=1}^{p} v_{kj} \beta_{tk}$, the PCLR model extended to polytomous responses, with q principal components, is given by:

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

In order to estimate the principal components model's parameters, one can apply the Maximum Likelihood Method. With respecto to the QLR model, we propose to use as covariates the principal components of the $[(s-1)(p+1)]\left(1 + \frac{p}{2}\right)$ matrix $\mathbf{I}(\underline{\chi})$, whose elements are given by:

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

and

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

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

Simulations

In this section we consider a benchmark data set, Fatty Acid Composition Data, taken from Brodnjak – Vončina et al. (2005). The purpose is to compare the results provided by the two models, given by the Correct Classification Rate (CCR), defined as the percentage of observations that are correctly classified. There are 120 observations, five groups and seven variables, representing the percentage levels of seven fatty acids, namely palmitic, stearic, oleic, linoleic, eicosanoic and eicosenoic 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. The original data set have eight groups, and can be found in Brodnjak – Vončina et al. (2005). In this paper we use the group 5 (pumpkin oil) as the reference group. Table 1 displays the classification matrix for the CLR and PCLR models. Table 2 displays the classification matrix for the QLR and PCQLR models.

Table 1. 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	0.70	0.30
	G 5	0.15	0.00	0.05	0.05	0.75
PCLR (6 p.c.'s)	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 2. 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.'s)	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	0.06	0.91

Conclusions

The purpose with this job is to develop and implement a direct generalization for the Quadratic Logistic Regression model, for polytomous response, which allows the reduction of the dimensions in the problem, and to explore the performance of the model when compared to the Classical Logistic

Regression model with discriminant functions. In order to solve the problem that arises with the great number of unknown parameters, we have used the Principal Components Analysis, as well a generalization of the HLR model, to deal with the complete separation. We can see thet the PCA allows the reduction of the number of dimensions in a polytomous QLR model, with continuous variables and avoiding the multicollinearity of these variables. For practical purposes, the main advantage of the HLR model is the existence and uniqueness of estimators. Furthermore, there are not computational difficulties to implement both approaches. With respect to the performance, we can see that the QLR model can provide better classification rates than the CLR model.

In the future we intend to study the behavior of the models that were approached with respect to aspects such as their performance regarding data sets with a reduced number of observations and the bias of the estimators that were obtained.

REFERENCES (RÉFERENCES)

Aguilera, A. M., Escabias, M., Valderrama, M. J., Using principal components for estimating logistic regression with high-dimensional multicollinear data. Computational Statistics & Data Analysis 55, 1905-1924 (2006)

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

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

Andruski-Guimarães, I. and Chaves-Neto, A., Estimation in polytomous logistic model: comparison of methods, Journal of Industrial and Management Optimization, 5, pp. 239-252 (2009)

Barker, L. and Brown, C., Logistic regression when binary predictor variables are highly correlated, Satistics in Medicine, 20 (9-10), pp. 1431-1442, (2001)

Brodnjak – Vončina, D., Kodba, Z.C. and Novič, C., Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. Chemometrics and Intelligent Laboratory Systems 75, pp. 31-43, (2005)

Copas, J. B., Binary regression models for contaminated data. With discussion. Journal of Royal Statistical Society B, **50** (1988), 225–265.

Heinze, G. and Schemper, M., A solution to the problem of separation in logistic regression, Statistics in Medicine 21, 2409-2419 (2002)

Rousseeuw, P. J. and Christmann, A., Robustness against separation and outliers in logistic regression, Computational Statistics & Data Analysis 43, pp. 315-332 (2003)

RÉSUMÉ (ABSTRACT)

Is well known that the logistic regression model are a powerful method widely applied for modeling the relationship between a categorical dependent variable and a set of explanatory variables. Many papers on logistic regression have only considered the logistic regression model with linear discriminant functions, but there are situations where quadratic discriminant functions are useful, and works 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 independent variables. Furthermore, a great number of parameters should be avoided, because of the risk of overfitting. 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.