Submetido para TEMA

Análise de Componentes Principais Aplicada à Estimaçao de Parâmetros no Modelo de Regressão Logística Quadrático

Nome(s) do(s) Autor(es)[1], Departamento , Universidade , Endereço , Cidade

Resumo. A maioria dos estudos sobre o modelo de regressão logística politômica considera apenas o modelo com funções discriminantes lineares. Entretanto, funções discriminantes quadráticas podem ser de grande utilidade, além de apresentar melhores resultados. Porém, o modelo logístico quadrático envolve a estimação de um grande número de parâmetros desconhecidos, o que pode exigir um grande esforço computacional. Neste trabalho utiliza-se um conjunto de componentes principais das variáveis explanatórias a fim de reduzir as dimensões do modelo a ser estimado, com variáveis explanatórias contínuas, bem como os custos computacionais para a estimação de parâmetros na regressão logística quadrática politômica, sem perda de eficiência. Simulações com dois conjuntos de dados mostram que o modelo de regressão logística quadrático, com componentes principais, é computacionalmente viável, podendo produzir resultados melhores que aqueles obtidos pelo modelo de regressão logística clássico, em termos de taxas de classificações corretamente efetuadas.

Palavras-chave. Regressão Logística Politômica, Regressão Logística Quadrática, Análise de Componentes Principais.

1. Introdução

O Modelo de Regressão Logística é empregado para modelar a relação entre uma variável dependente categórica e um conjunto de variáveis explanatórias. Na literatura disponível a quase totalidade dos trabalhos sobre o modelo considera apenas funções discriminantes lineares, como [1], [2], [10] e [11], para citar alguns poucos. Entretanto, funções discriminantes quadráticas podem ser de grande utilidade, podendo também apresentar melhores resultados. Porém, o modelo logístico quadrático para variábel resposta politômica envolve a estimação de um grande número de parâmetros, o que pode exigir um elevado esforço computacional, especialmente quando há um número elevado de variáveis explanatórias no conjunto de dados. Neste trabalho propõe-se o uso de um conjunto de componentes principais das variáveis explanatórias a fim de reduzir as dimensões do modelo a ser estimado, com variáveis contínuas, bem como os custos computacionais para a estimação de parâmetros na regressão logística quadrática politômica, sem perda de eficiência.

[1]e-mail do autor

2. Modelo de Regressão Logística Clássico

Seja uma amostra de n observações independentes, distribuídas entre s grupos, $G_1, G_2, ..., G_s$. Seja $\underline{\mathbf{x}}$ o vetor de variáveis explanatórias, ou covariáveis, tal que, $\underline{\mathbf{x}}^T = (x_0, x_1, ..., x_p)$, onde $x_0 \equiv 1$, por conveniência. Seja Y a variável resposta politômica, com s respostas possíveis, cada resposta indicando o grupo, ou categoria, ao qual pertence a observação em questão. Na forma matricial, as n observações são dadas por:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & ... & x_{p1} \\ 1 & x_{12} & ... & x_{p2} \\ ... & ... & ... & ... \\ 1 & x_{1n} & ... & x_{pn} \end{bmatrix}$$

O Modelo de Regressão Logística Clássico (MRLC), para variável resposta politômica, assume que as probabilidades a *posteriori* são dadas por:

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

onde $k = 1\,,\,2\,,\,...\,,\,s-1$ e $\underline{\mathbf{B}}_s = \mathbf{0}$, considerando s como o grupo de referência. Há $(s-1)\,(p+1)$ parâmetros desconhecidos e a função de versossimilhança é:

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

onde $\mathbf{Y} = (\mathbf{Y}_1\,,\,...\,,\mathbf{Y}_n)^T$ e $\mathbf{Y}_i = (Y_{1i}, ..., Y_{si})$, com $Y_{ki} = 1$ se $Y = k$, e $Y_{ki} = 0$ em outro caso. A função log-verossimilhança é dada por:

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

Nestas condições, tem-se que:

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

O Estimador de Máxima Verossimilhança (EMV) $\hat{\underline{\mathbf{B}}}$ é obtido após igualar as derivadas a zero e resolver o sistema resultante em relação a $\underline{\mathbf{B}}$. O procedimento mais utilizado neste caso é o Método de Newton-Raphson.

Na prática, a estimação dos parâmetros desconhecidos do modelo logístico é sensível a certas características dos dados, especialmente no que se refere à sobreposição de grupos. Uma abordagem apresentada por [2] propõe a classificação do conjunto de dados em três categorias: separação completa, quando os grupos estão completamente separados, sobreposição parcial, quando apenas alguns grupos apresentam sobreposição e sobreposição completa, quando cada grupo compartilha informações iguais com todos os demais grupos. De acordo com [2], os estimadores de máxima verossimilhança podem ser calculados se, e somente se, houver sobreposição de grupos. Algumas abordagens para contornar o problema da separação entre grupos podem ser encontradas em [11] e [17], para variável resposta binária, e [4], para variável resposta politômica.

Para contornar o problema da separação completa, adotou-se neste trabalho uma generalização do Modelo de Regressão Logística Oculto (MRLO), método de estimação robusta proposto por [17] para variável resposta binária, e que tem como base abordagens apresentadas por [8] e [7]. Esta generalização considera que há n variáveis não observáveis, $T_1, \ldots, T_n$, onde T_i possui s valores possíveis, $\gamma_i, \ldots, \gamma_s$. Desta forma, tem-se $Y_i = j$ com probabilidade $P(Y_i = j \mid T_i = \gamma_k) = \delta_{jk}$, onde $\sum_{j=1}^{s} \delta_{jk} = 1$ e $\delta_{jj} = max_{k=1,\ldots,s}\{\delta_{jk}\}$.

O estimador de máxima verossimilhança para T_i, quando $Y_i = j$, é $\widehat{T}_{ML,i} = \gamma_j$. Para um modelo com n possíveis respostas y_{ij}, $i = 1, \ldots, n$ e $j = 1, \ldots, s$, onde $y_{ij} = 1$, se $Y_i = j$, e $y_{ij} = 0$, caso contrário, pode-se definir a variável dada por:

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

Para o MRLC, $\delta_{jj} = 1$ e $\delta_{jk} = 0$, se $j \neq k$. O objetivo é maximizar:

$$\ell\left(\underline{\mathbf{\Theta}} \mid \underline{\tilde{\mathbf{Y}}}, \underline{\mathbf{X}}\right) = \prod_{i=1}^{n}\prod_{j=1}^{s}\left[P\left(T_j \mid \underline{\mathbf{x}}_i\right)\right]^{\tilde{Y}_{ji}} .$$

A função log-verossimilhança fica:

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

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

Os estimadores de máxima verossimilhança são os valores que maximizam a função log-verossimilhança com relação a $\underline{\mathbf{\Theta}}$. Para maiores detalhes sobre a referida maximização, sugere-se consultar [4]. Na literatura disponível é possível encontrar diferentes abordagens visando a implementação de métodos de estimação robusta, apresentados por [10], [12] e [14], entre outros.

De acordo com [17], [7] observou que a estimação de δ_0 e δ_1, para variável resposta binária, pode ser bastante complexa e dispendiosa, sob o ponto de vista computacional, a menos que n seja muito grande. A abordagem simétrica consiste em escolher uma constante $\gamma > 0$ e tomar $\delta_0 = \gamma$ e $\delta_1 = 1-\gamma$, onde γ é tão pequeno que γ^2 possa ser ignorado, e $\delta_0 < \hat{\pi} < \delta_1$, onde $\hat{\pi}$, δ_0 e δ_1 são dados por $\delta_1 = \frac{1+\hat{\pi}\delta}{1+\delta}$, $\delta_0 = \frac{\hat{\pi}\delta}{1+\delta}$, $\hat{\pi} = max\left\{\delta\,,\, min\left(1-\delta\,;\,\bar{\pi}\right)\right\}$, $\bar{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Explicações mais detalhadas, bem como discussões, podem ser encontradas em [7], [17] e [12]. Neste trabalho considerou-se que a probabilidade de observar o verdadeiro estado, dada por $P\left(Y_i = j \,|\, T_i = \gamma_j\right) = \delta_{jj}$, deve ser superior a 0.5, isto é, $0.5 < \delta_{jj} < 1$, adicionalmente $\sum_{k=1,k\neq j}^{s}\delta_{jk} < \delta_{jj}$. Além disto, não é possível determinar o estimador dado por $\bar{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$, $j = 1\,,\,...\,,\,s$, uma vez que $\bar{\pi}_j$ pode ser menor que 0.5. Neste trabalho optou-se por escolher $\delta = 0.99$, e fazer $\delta_{jj} = \delta$ e $\delta_{jk} = \frac{1-\delta}{s-1}$.

3. Modelo de Regressão Logística Quadrático

A inclusão de termos quadráticos e multiplicativos na função linear do MRLC resulta no Modelo de Regressão Logística Quadrático (MRLQ), dado por:

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

onde $\underline{\chi}_k = \alpha_{k0} + \sum\limits_{i=1}^{p}\alpha_{ki}x_i^2 + \sum\limits_{i=p+1}^{{}_pC_2}\alpha_{ki}x_{j'}x_{j''} + \sum\limits_{i={}_pC_2+1}^{{}_pC_2+p}\alpha_{ki}x_j$, $k = 1\,,\,2\,,\,...\,,\,s-1$, $\underline{\chi}_s = \mathbf{0}$, e $j, j'' = 1\,,\,2\,,\,...\,,\,p$, $j' = 1\,,\,2\,,\,...\,,\,p-1$.
Este modelo envolve $\left[(s-1)\,(p+1)\right]\left(1+\frac{p}{2}\right)$ parâmetros desconhecidos, cuja estimação segue o mesmo raciocínio usado para obter os parâmetros do MRLC. Entretanto, caso haja um grande número de covariáveis, o número de parâmetos adicionais pode resultar em um problema de difícil resolução, do ponto de vista computacional, o que pode tornar de grande utilidade um método destinado a reduzir as dimensões do conjunto de dados. De acordo com [3], a expressão quadrática também pode ser apresentada na forma:

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

onde $\mathbf{\Omega}_k = \mathbf{V}_k^{-1} - \mathbf{V}_s^{-1}$, $\mathbf{V}_k$ é a matriz de covariâncias em G_k, $k = 1, 2, ..., s-1$, e $\mathbf{V}_s$ é a matriz de covariâncias em G_s.
Para reduzir o número de parâmetros, [3] sugere uma aproximação obtida através da decomposição espectral da matriz de informação, que resulta na expressão:

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

onde os λ_{jk} são os autovalores da matriz $\mathbf{\Omega}_k$, em ordem decrescente, $\lambda_{1k} \geq \lambda_{2k} \geq \ldots \geq \lambda_{pk}$, e l_{jk} são os respectivos autovetores. Neste caso, $\mathbf{\Omega}_k$ pode ser escrita como $\mathbf{\Omega}_k \cong \lambda_k l_k l_k^T$, onde λ_k é o maior autovalor. O passo seguinte é a normalização de cada $l_j^T = (\ell_{j1}, ..., \ell_{jp})$ sob as restrições:

$$\sum_{k=1}^{p} \ell_{jk}^2 = 1$$

Como esta abordagem pode ser pouco eficiente computacionalmente, sugere-se uma alternativa, que consiste em considerar a forma dada por:

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

onde $\mu_k = sgn(\lambda_k)$, $k = 1, ..., s-1$, $d_{kj} = \ell_{kj}/\sqrt{|\lambda_k|}$, $j = 1, ..., p$. A função log-verossimilhança é maximizada com relação a α_{kj} e d_{kj}, sem restrições, $2^{(s-1)}$ vezes para $\mu_k = \pm 1$. Na seqüência, toma-se como estimadores de máxima verossimilhança os maiores valores entre os $2^{(s-1)}$ valores da função log-verossimilhança, obtendo-se desta forma os estimadores para cada um dos $(s-1)p$ parâmetros. Contudo esta abordagem nem sempre é aplicável. Por exemplo, de acordo com [3] se o conjunto de dados contém variáveis binárias, os termos da diagonal da matriz de covariâncias são iguais a zero. Neste trabalho propõe-se utilizar como covariáveis as componentes principais da matriz de informação $\mathbf{I}(\underline{\mathbf{B}})$, de ordem $(s-1)(p+1)$, cujos elementos são dados por:

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

e

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

onde $j, j' = 1, 2, ..., (s-1)$ e $m, m' = 1, 2, ..., p$.

4. Modelo de Regressão Logística de Componentes Principais

A Análise de Componentes Principais (ACP) é um método utilizado para estudar a variância e a covariância através de combinações lineares das p variáveis envolvidas, e pode ser considerada uma ferramenta para reduzir a colinearidade entre as variáveis explanatórias e, também, a dimensão do conjunto de dados, pois permite expressar a maior parte da variabilidade através de q componentes principais, $q < p$.

Sejam n observações de p variáveis contínuas, dadas pela matriz $\mathbf{X}$, e seja a matriz de covariância amostral:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ & s_{22} & \dots & s_{2p} \\ & & \ddots & \vdots \\ & & & s_{pp} \end{bmatrix}.$$

As observações $\underline{\mathbf{x}}$ podem ser padronizadas, de modo que

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

A matriz $\mathbf{S}$ pode ser escrita como $\mathbf{S} = \mathbf{V}^{\mathbf{T}}\mathbf{\Lambda}\mathbf{V}$, onde $\mathbf{\Lambda} = diag\left(\lambda_1, ..., \lambda_p\right)$ e $\mathbf{V}$ ï¿½ uma matriz ortogonal. Seja $\mathbf{Z}$ a matriz cujas colunas são as componentes principais, dada por $\mathbf{Z} = \mathbf{X}\mathbf{V}$, onde $\underline{\mathbf{v}}_1, ..., \underline{\mathbf{v}}_p$ são os autovetores da matriz $\mathbf{S}$, associados aos autovalores $\lambda_1 \geq ... \geq \lambda_p \geq 0$, tal que a matriz de observações possa ser escrita como $\mathbf{X} = \mathbf{Z}\mathbf{V}^{\mathbf{T}}$, onde $x_{ij} = \sum_{k=1}^{p} z_{ik}v_{jk}$. Além disto, as matrizes $\mathbf{Z}$ e $\mathbf{V}$ podem ser escritas como:

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

e

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

Para melhorar a estimação de parâmetros do modelo logístico na presença de multicolinearidade, e reduzir a dimensão do conjunto de dados, [1] sugere usar como covariáveis do modelo logístico um conjunto reduzido de componentes principais das variáveis originais. Esta abordagem, chamada Modelo de Regressão Logística por Componentes Principais (MRLCP), fornece uma estimação bastante precisa dos parâmetros, no caso de existência de multicolinearidade, e toma por base uma abordagem proposta por [15]. Adicionalmente, de acordo com [5], estimadores obtidos através da ACP podem apresentar viés inferior ao apresentado por estimadores obtidos através de métodos mais comumente usados. Por outro lado, deve-se levar em consideração que, conforme [16], os estimadores dos autovalores da matriz $\mathbf{S}$ podem ter um grande viés quando tendem a ser iguais, ou muito próximos. Abordagens para reduzir este viés podem ser encontradas em [12].

A generalização do MRLCP para variável resposta politômica não exige uma formulação complexa. Inicialmente calcula-se a matriz de covariâncias $\mathbf{S}$. Desta forma, os elementos da matriz $\mathbf{X}$ podem ser representados por $x_{ik} = \sum_{j=1}^{p} z_{ij}v_{kj}$, tal que:

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

onde $i = 1, \ldots, s$, $j = 0, \ldots, p$, $t = 1, \ldots, s$ e $\beta_{sj} = 0$.

Fazendo $\gamma_{tj} = \sum_{k=1}^{p} v_{kj} \beta_{tk}$, o MRLCP para variável resposta politômica é dado por:

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

O Modelo de Regressão Logística Quadrático de Componentes Principais (MRLQCP) é:

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

onde $\underline{\chi}_k = \chi_{k0} + \sum_{i=1}^{p} z_{ij}\gamma_{kj}^2 + \sum_{i=p+1}^{{}_pC_2} z_{ij}\gamma_{kj'}\gamma_{kj''} + \sum_{i={}_pC_2+1}^{{}_pC_2+p} z_{ij}\gamma_{kj}$, para $k = 1, 2, \ldots, s-1$, $\underline{\chi}_s = \mathbf{0}$, e $j, j'' = 1, 2, \ldots, p$, $j' = 1, 2, \ldots, p-1$.

Para a formulação do MRLCP e do MRLQCP foram utilizadas as q primeiras componentes principais, com a percentagem acumulada da variância total não inferior a 95%. Sobre a seleção de componentes principais, deve-se ter em mente, conforme [13], que componentes principais com os menores autovalores podem ser tão úteis quanto aquelas com maiores autovalores, podendo, inclusive, gerar modelos discriminantes mais eficazes.

5. Aplicações

Para verificar a eficiência do modelo logístico quadrático para variável resposta politômica, bem como da utilização da Análise de Componentes Principais na estimação de parâmetros do modelo em questão, os modelos quadráticos obtidos foram testados em dois bancos de dados extraídos da literatura. O primeiro conjunto de dados foi obtido de [9], e contém medidas de 150 observações de flores de três espécies. O segundo conjunto, obtido de [6], contém 120 observações referentes aos teores de ácidos graxos de óleos vegetais de cinco variedades. Os resultados obtidos, comparados com relação à taxa aparente de erros, são apresentados na seqüência.

Exemplo 1: Iris. São três grupos: *Iris Setosa* (G_1), *Iris Versicolor* (G_2) e *Iris Virginica* (G_3), usado como grupo de referência. Em cada grupo há 50 observações e quatro variáveis explanatórias: Comprimento (x_1) e largura (x_2) da sépala e comprimento (x_3) e largura (x_4) da pétala, em milímetros. Sabe-se que o grupo G_1 é completamente separado dos demais. Além disto, há uma forte correlação entre as variáveis x_3 e x_4, com $r = 0,9629$. Neste exemplo, enquanto o MRLO requer a estimação de 10 parâmetros, ou cinco para cada função discriminante, o MRLQ envolve 30 parâmetros, 15 para cada função discriminante. Para os modelos com componentes principais foram selecionadas as duas componentes com maior percentagem da variância acumulada. Desta forma, o MRLCP envolve seis parâmetros, enquanto o MRLQCP envolve 12 parâmetros. A Tabela 1 mostra as variâncias (autovalores) e a percentagem acumulada da variância total para cada componente principal. As matrizes de classificações para o MRLO e para o MRLCP, com funções discriminantes lineares, são apresentadas na Tabela 2. As taxas de acertos para os modelos com discriminantes quadráticas, MRLCP e MRLQCP, são apresentadas na Tabela 3. Estes resultados mostram que o modelo quadrático com duas componentes principais apresentou as mesmas taxas de acerto obtidas pelo MRLQ. Entretanto, cabe ressaltar que o MRLQCP exigiu a estimação de um número significativamente menor de parâmetros.

Tabela 1: Iris. Variâncias (Autovalores) e Percentagem Acumulada da Variância Total.

Variância (λ)	2.92	0.91	0.15	0.021
Variância Total Acumulada (%)	72.96	95.81	99.48	100.00

Tabela 2: Matriz de Classificações. Iris. Funções Discriminantes Lineares.

Modelo	Grupo Observado	Grupo Predito		
		G 1	G 2	G 3
MRLO	G 1	1.00	0.00	0.00
	G 2	0.00	0.98	0.02
	G 3	0.00	0.02	0.98
MRLCP (2 c.p.'s)	G 1	1.00	0.00	0.00
	G 2	0.00	0.88	0.12
	G 3	0.00	0.10	0.90

Tabela 3: Matriz de Classificações. Iris. Funções Discriminantes Quadráticas.

Modelo	Grupo Observado	Grupo Predito		
		G 1	G 2	G 3
MRLQ	G 1	1.00	0.00	0.00
	G 2	0.00	0.98	0.02
	G 3	0.00	0.02	0.98
MRLQCP (2 c.p.'s)	G 1	1.00	0.00	0.00
	G 2	0.00	0.98	0.02
	G 3	0.00	0.02	0.98

Exemplo 2: Ácidos Graxos. São 120 observações, cinco grupos e sete variáveis explanatórias, representando os teores de sete ácidos graxos: palmítico, esteárico, oléico, linoléico, linolênico, eicosanoico e eicosenoico. Os cinco grupos considerados referem-se a óleos de: colza (G_1), girassol (G_2), amêndoa (G_3), milho (G_4) e abóbora (G_5), utilizado como grupo de referência. Cabe ressaltar que os teores de ácidos oléico e linoléico apresentam forte correlação, com $r = -0,9565$. Neste exemplo o modelo logístico linear contém quatro funções discriminantes, cada uma com oito parâmetros, isto é, 32 parâmetros desconhecidos. O modelo quadrático, por sua vez, requer a estimação de 144 parâmetros, ou 36 para cada função discriminante. A Tabela 4 mostra as variâncias (autovalores) e a percentagem acumulada da variância total para cada componente principal. Para a construção dos modelos MRLCP e MRLQCP foram utilizadas quatro componentes principais, totalizando 20 parâmetros para o MRLCP e 60 para o MRLQCP. A Tabela 5 apresenta as matrizes de classificações para o MRLO e para o MRLCP, com funções discriminantes lineares. As matrizes de classificações para o MRLQ e para MRLQCP são apresentadas na Tabela 6. Neste caso, embora o MRLQCP tenha apresentado taxas inferiores ao MRLQ, pode-se argumentar que estas mesmas taxas são superiores ao MRLO, com funções lineares.

Tabela 4: Ácidos Graxos. Variâncias (Autovalores) e Percentagem Acumulada da Variância Total.

Variância (λ)	3.91	1.08	0.93	0.79	0.21	0.08	0.00
Variância Total Acumulada (%)	55.85	71.84	84.66	95.90	98.83	99.99	100.00

Tabela 5: Matriz de Classificações. Ácidos Graxos. Funções Discriminantes Lineares.

Modelo	Grupo Observado	Grupo Predito G 1	G 2	G 3	G 4	G 5
MRLO	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	0.70
MRLCP (4 c.p.'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

Tabela 6: Matriz de Classificações. Ácidos Graxos. Funções Discriminantes Quadráticas.

Modelo	Grupo Observado	Grupo Predito G 1	G 2	G 3	G 4	G 5
MRLQ	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
MRLQCP (4 c.p.'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	0.05	0.92

6. Conclusão

O uso de componentes principais para substituir as variáveis explanatórias teve como resultado a redução das dimensões do conjunto de dados e, consequentemente, do número de parâmetros desconhecidos do modelo de regressão logística quadrático, com variável resposta politômica. Esta redução, juntamente com os resultados obtidos, em termos de taxas de classificação, permitem concluir que a ocorrência de multicolinearidade deixa de ser um problema e passa a ser elemento importante da solução, pois a sua ocorrência possibilita o uso de um conjunto reduzido de componentes principais, acarretando uma significativa redução do esforço computacional necessário à estimação dos parâmetros. Embora não tenha apresentado a mesma eficiência que o MRLQ, o modelo quadrático com componentes principais mostrou-se mais eficiente que o modelo clássico, com funções discriminantes lineares, o que demonstra a sua viabilidade como método de classificação. De acordo com [6], quando usadas na classificação de óleos vegetais, as componentes principais exigem uma inspeção mais elaborada. Os resultados obtidos mostram que as componentes principais, em conjunto com o modelo logístico, podem ser utilizadas sem maiores problemas na construição de modelos de apoio à tomada de decisões.

O problema decorrente da separação completa de grupos foi contornado através de uma generalização do modelo logístico oculto, possibilitando a estimação dos parâmetros mesmo nas referidas condições. A principal vantagem desta abordagem é a capacidade de encontrar valores finitos para os estimadores de máxima verossimilhança, tanto para o modelo clássico, com funções discriminantes lineares, como para o modelo logístico quadrático. Uma vantagem adicional está no fato de não haver maiores dificuldades para a implementação computacional. Com relação ao desempenho, é possível concluir, pelos exemplos apresentados, que o modelo logístico quadrático, com estimadores obtidos a partir das componentes principais das variáveis explanatórias, mostrou-se um método confiável para a análise e reconhecimento estatístico de padrões, podendo apresentar melhores taxas de classificação que o modelo clássico, com funções lineares.

Abstract. 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. 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.

Referências

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

[2] A. Albert, J. A. Anderson, On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, **71** (1984), 1–10.

[3] J.A. Anderson, Quadratic logistic discrimination, *Biometrika*, **62** (1975), 149–154.

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

[5] L. Barker, C. Brown, Logistic regression when binary predictor variables are highly correlated, *Satistics in Medicine*, **20** (9-10) (2001), 1431–1442.

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

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

[8] A. Ekholme, J. Palmgren, A model for binary response with misclassification. *GLIM 82 Proceedings of the International Conference on Generalized Linear Models*, (1982) 128–143.

[9] R.A. Fisher, The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **3** (1936), 179–188.

[10] D. Gervini, Robust adaptive estimators for binary regression models. *Journal of Statistical Planning and Inference*, **131**, (2005), 297–311 .

[11] G. Heinze, M. Schemper, A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21** (2002), 2409–2419.

[12] M. Hubert, K. van Driessen, Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, **45**, 2 (2004), 301–320.

[13] I.T. Jolliffe, A note on the use of principal components in regression. *Applied Statistics*, **31**, 3 (1982), 300–303.

[14] N. Kodzarkhia, G.D. Mishra, L. Reiersolmoen, Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference*, **98** (2004), 211-223.

[15] W. F. Massy, Principal component regression in exploratory statistical research. *Journal of American Statistical Association*, **60**, (1965), 234–246.

[16] G. J. McLachlan, "Discriminant Analysis and Statistical Pattern Recognition", p. 130. John Wiley & Sons, Inc., Hoboken, New Jersey, U.S.A. 2004.

[17] P.J. Rousseeuw, A. Christmann, Robustness against separation and outliers in logistic regression, *Computational Statistics & Data Analysis*, **43** (2003), 315–332.