INÁCIO ANDRUSKI GUIMARÃES

MODELOS DE REGRESSÃO LOGÍSTICA OCULTO E DE COMPONENTES PRINCIPAIS PARA RECONHECIMENTO E CLASSIFICAÇÃO DE PADRÕES COM VARIÁVEL RESPOSTA POLITÔMICA

Tese apresentada como requisito parcial à obtenção do título de Doutor em Métodos Numéricos em Engenharia, área de concentração: Programação Matemática, sob a orientação do Prof. Dr. Anselmo Chaves Neto.

Curitiba - Paraná

2006

TERMO DE APROVAÇÃO

Inácio Andruski Guimarães

"Modelos de Regressão Logística Oculto e de Componentes Principais para Reconhecimento e Classificação de Padrões com Variável Resposta Politômica"

Tese aprovada como requisito parcial para obtenção do grau de Doutor em Ciências no Programa de Pós-Graduação em Métodos Numéricos em Engenharia – Área de Concentração em Programação Matemática, setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof. Anselmo Chaves Neto, D.Sc.
PPGMNE – UFPR

Profa. Maria Terezinha Arns Steiner, D. Eng.
PPGMNE – UFPR

Prof. Jair Mendes Marques, D.Sc.
PPGMNE – UFPR

Prof. Sebastião de Amorin, Ph. D.
Faculdade de Engenharia / UNICAMP

Prof. Júlio César Nievola, D. Eng.
PPGIA – PUC PR

Curitiba, 08 de dezembro de 2006.

À minha esposa Patrícia Accioly Calderari da Rosa, com amor.

AGRADECIMENTOS

Aos meus pais, pelos esforços jamais negados e pelos exemplos sempre oferecidos.

Ao meu orientador, Prof. Dr. Anselmo Chaves Neto, pela infinita paciência demonstrada diante das ocasionais dificuldades, provocadas ora pelas dificuldades inerentes ao trabalho, ora pelos equívocos do orientado.

Aos professores do Programa de Pós Graduação em Métodos Numéricos em Engenharia – PPGMNE, por todo o conhecimento transmitido.

Aos professores Maria Terezinha Arns Steiner, Sebastião Amorim, Jair Mendes Marques e Júlio César Nievola, pelas valiosíssimas sugestões para a melhoria do presente trabalho e também pelo exaustivo trabalho de revisão.

À secretária do CPGMNE, Maristela Bandil, pela eficiência sempre demonstrada e, sobretudo, pelo bom humor inesgotável.

Finalmente, aos funcionários da biblioteca do setor de ciências exatas e tecnológicas da UFPR, especialmente aos operadores do Programa de Comutação Bibliográfica – COMUT, pela eficiência e agilidade na obtenção dos trabalhos solicitados.

SUMÁRIO

LISTA DE QUADROS

LISTA DE FIGURAS

RESUMO

Este trabalho apresenta uma revisão dos métodos mais conhecidos e utilizados na estimação de parâmetros de Modelos de Regressão Logística aplicados a problemas de Reconhecimento de Padrões com variável resposta politômica. Também aborda o problema da separação de grupos, que é fundamental para o cálculo dos estimadores dos parâmetros dos modelos mencionados. O principal objetivo é comparar a eficiência de abordagens ao problema da obtenção de regras discriminantes a partir do Modelo de Regressão Logística Oculto, que é imune à separação de grupos em problemas com variável resposta binária, e também a partir do Modelo de Regressão Logística de Componentes Principais. As mencionadas abordagens consistem em estender os modelos citados a problemas com variável resposta politômica, de modo a apresentar uma alternativa original para a abordagem de problemas desta natureza. O desempenho dos modelos obtidos é avaliado mediante a aplicação dos mesmos a conjuntos de dados extraídos da literatura corrente, em comparação com o Modelo de Regressão Logística Clássico e Modelos de Regressão Logística Individualizados, além da Função Discriminante Linear de Fisher e de uma Rede Neural Artificial. O viés dos estimadores dos parâmetros do Modelo de Regressão Logística Oculto é estimado através do Método *Bootstrap*. O critério para comparação do desempenho dos modelos obtidos é a taxa de classificações incorretamente efetuadas pelos métodos mencionados, também chamada Taxa Aparente de Erros.

PALAVRAS CHAVES: Reconhecimento de Padrões, Variável Resposta Politômica, Separação de Grupos, Modelo de Regressão Logística Oculto, Modelo de Regressão Logística de Componentes Principais, Métodos de Reamostragem.

ABSTRACT

This job gives a review of the most widely used methods for estimating the parameters in the Logistic Regression Models, applied to Pattern Recognition with polytomous response variable, as well gives a brief review of some properties about data configuration, in order to compute the parameter estimates. The main goal is to compare the performance of these methods in building recognition rules based on the Hidden Logistic Regression Model, which is immune to any configuration of the data in binary case; as well the Principal Component Logistic Regression Model. We propose an extension of the models above to problems with polytomous response, in order to show an original approach to solve the parameter estimation problem when the groups are completely separated. The performance of the models is investigated through simulations and by applying it to some data sets taken from the trade literature, and compares with the performance obtained by the Classical Logistic Regression Model, Individualized Logistic Regression Model, Linear Discriminant Function and Artificial Neural Network. The bias of estimates in Hidden Logistic Regression Model is investigated through the Bootstrap Method. The criterion used to compare the resulting performance is the apparent error rate.

KEYWORDS: Pattern Recognition, Polytomous Response, Data Separation, Hidden Logistic Regression Model, Principal Component Logistic Regression Model, Resampling Methods.

1 INTRODUÇÃO

A principal motivação para este trabalho decorre do fato de que, a despeito da reconhecida eficiência do Modelo de Regressão Logística como método de Reconhecimento e Classificação de Padrões, é possível notar através da literatura disponível que estudos e aplicações envolvendo o método em questão empregam na sua grande maioria modelos com resposta binária. Mesmo os problemas pertinentes à estimação de parâmetros, particularmente para conjuntos com grupos cujas configurações comprometem os resultados, em geral são abordados a partir de modelos com resposta binária, sem que se dedique maior atenção aos problemas com variável resposta politômica e à respectiva metodologia envolvida. Também pode-se perceber, especialmente nos últimos cinco anos, que o emprego da Regressão Logística para modelos com variável resposta politômica é muito menos freqüente quando comparado a técnicas como, por exemplo, Redes Neurais Artificiais, em suas mais variadas configurações, Algoritmos Genéticos e, mais recentemente, Máquinas de Base Vetorial ("*Support Vector Machines*"). Sem desconsiderar a eficiência destas técnicas no Reconhecimento de Padrões, comprovada pelo grande número de trabalhos disponíveis na literatura atual, pode-se argumentar que a Regressão Logística tem um grande potencial como objeto de estudos, além de representar uma opção matematicamente consistente para a análise de dados categorizados. Também nota-se que em alguns campos de estudo, como na Medicina, por exemplo, o Modelo Logístico é o mais utilizado como método de discriminação, no que pode ser considerada uma abordagem padrão.

O propósito principal deste trabalho é apresentar um estudo comparativo de diferentes métodos de estimação de parâmetros em modelos de Regressão Logística, no que se refere à convergência dos mesmos para soluções finitas e também à eficiência dos modelos resultantes quando utilizados como regras discriminantes. Além disso, apresenta-se uma abordagem original baseada no Modelo de Regressão Logística Oculto e no Modelo de Regressão Logística de Componentes Principais, mediante a extensão dos mesmos para problemas com variável resposta politômica, e diferentes tipos de configurações de conjuntos de dados. A abordagem é desenvolvida de acordo com a metodologia empregada para encontrar as soluções obtidas por ambos os modelos para problemas com variável resposta dicotômica.

1.1 PROBLEMA

As técnicas estatísticas mais utilizadas no Reconhecimento Estatístico de Padrões são a Função Discriminante Linear de Fisher (FDL), a Regressão Logística e a Função Discriminante Quadrática (FDQ). Estas técnicas podem ser aplicadas a problemas que envolvem, por exemplo, análise de crédito, previsão de falências, detecção de fraudes em seguros e cartões de crédito, manutenção de equipamentos, diagnóstico médico, bem como em estudos biomédicos e epidemiológicos. A utilização da primeira presume que as matrizes de covariâncias dos grupos analisados sejam iguais, o que nem sempre ocorre na prática. A Regressão Logística é uma alternativa à Função Discriminante Linear, particularmente quando a suposição acerca das matrizes de covariâncias não é satisfeita, e pode ser aplicada a uma grande família de distribuições de probabilidade, envolvendo tanto variáveis discretas como contínuas. Na prática, contudo, a obtenção do Modelo de Regressão Logística pode ser prejudicada por determinadas características dos dados, que afetam o desempenho de procedimentos iterativos utilizados na estimação dos parâmetros desconhecidos. Estas características, que dizem respeito principalmente aos tipos de configuração dos conjuntos de dados, em especial à sobreposição de grupos, estão intimamente relacionadas à eficiência dos referidos processos iterativos, no que se refere à convergência dos mesmos no sentido de fornecer estimadores finitos. Alguns métodos tradicionalmente utilizados costumam falhar quando aplicados a problemas que envolvem conjuntos de dados com determinadas configurações, especialmente nos casos que envolvem variável resposta politômica, considerando que no caso de variável resposta dicotômica este problema pode ser contornado sem maiores dificuldades. A Função Discriminante Quadrática elimina a suposição de matrizes de covariâncias iguais, porém exige que os vetores aleatórios sejam oriundos de populações normais multivariadas, o que se configura um problema de igual magnitude.

1.2 OBJETIVOS

O principal objetivo deste trabalho é apresentar uma investigação do desempenho de diferentes métodos de estimação dos parâmetros para Modelos de Regressão Logística, em problemas com variável resposta politômica, ou multinomial. Mais especificamente, o que se pretende é avaliar o desempenho com relação à convergência de cada método para uma solução finita e apresentar uma síntese dos resultados obtidos, de modo a oferecer subsídios para a construção de modelos discriminantes baseados no Modelo de Regressão Logística. Os métodos abordados são: Modelo de Regressão Logística Clássico (MRLC); Modelos de Regressão Logística

Individualizados (MRLI), conforme a proposta de Begg e Grey (1984); Modelo de Regressão Logística Oculto (MRLO), apresentado por Rousseeuw e Christmann (2003), aqui generalizado para variável resposta politômica; e a utilização da Análise de Componentes Principais na obtenção dos estimadores de modelos com variável resposta politômica, tomando como ponto de partida a abordagem apresentada por Aguilera, Escabias e Valderrama (2006) para problemas com variável resposta binária. A mencionada generalização dos dois últimos métodos citados constitui uma abordagem inovadora, de fácil compreensão e rápida implementação computacional. Um objetivo secundário é a comparação da eficiência dos modelos abordados com a eficiência apresentada pela Função Discriminante Linear e por uma Rede Neural Artificial, com algoritmo de retro-propagação. Esta eficiência é avaliada por meio da taxa de classificações corretamente efetuadas pelos diferentes modelos, utilizando conjuntos de dados extraídos da literatura disponível, de modo a possibilitar a comparação dos resultados obtidos com aqueles eventualmente conhecidos por outros pesquisadores. A inclusão destas duas técnicas é motivada pela considerável quantidade de trabalhos publicados com o objetivo de avaliar a eficiência e o comportamento das mesmas quando aplicadas à Análise Discriminante. Com isto pretende-se fornecer subsídios para a utilização de qualquer das abordagens mencionadas como ferramenta de apoio à tomada de decisões.

1.3 ESTRUTURA DO TRABALHO

Este trabalho aborda inicialmente o conceito de separação de grupos e alguns aspectos referentes à medida de sobreposição dos mesmos para problemas com variável resposta binária. Também apresenta uma breve revisão de alguns conceitos e aspectos teóricos do Método do Custo Mínimo Esperado de Mistura (ECM). Em seguida apresenta uma breve explanação sobre o Modelo de Regressão Logística Clássico (MRLC) com variável resposta politômica e a estimação dos parâmetros, bem como o método de Newton – Raphson, destacando-se algumas propriedades importantes para a obtenção dos Estimadores de Máxima Verossimilhança (EMV). Também são revistos métodos alternativos ao Método da Máxima Verossimilhança, tais como os Modelos de Regressão Logística Individualizados (MRLI), e um método de estimação robusta, denominado Modelo de Regressão Logística Oculto (MRLO), aplicado a problemas com variável resposta dicotômica. Na seqüência apresenta-se uma síntese da aplicação do Método *Bootstrap* na estimação tanto dos parâmetros como do viés dos estimadores obtidos pelo Modelo de Regressão Logística Oculto. Uma rápida revisão de alguns aspectos teóricos sobre Função Discriminante Linear, bem como de aplicações da Programação Linear à Análise Discriminante Linear, e Redes Neurais Artificiais é apresentada na seqüência, aproveitando para expor algumas relações entre a última

técnica e o Modelo de Regressão Logística. O passo seguinte consiste em desenvolver uma extensão do Modelo de Regressão Logística Oculto para variável resposta politômica e, em seguida, a aplicação da Análise de Componentes Principais à estimação de parâmetros para o Modelo de Regressão Logística, que resulta no Modelo de Regressão Logística de Componentes Principais (MRLCP). Finalmente, a eficiência dos modelos obtidos é avaliada mediante a comparação com os resultados obtidos através das três técnicas. Para facilitar a comparação dos resultados obtidos com aqueles que porventura tenham sido obtidos por outros pesquisadores, optou-se por utilizar dados extraídos da literatura disponível.

2 REVISÃO DE LITERATURA

2.1 CONFIGURAÇÕES DOS CONJUNTOS DE DADOS

Uma questão de grande importância para o Reconhecimento Estatístico de Padrões envolve a configuração dos conjuntos de dados disponíveis para análise, especialmente quando a abordagem desenvolvida tem como base o Modelo de Regressão Logística. A mencionada configuração está diretamente relacionada à estimação dos parâmetros desconhecidos do modelo em questão, uma vez que a mesma tem influência sobre o desempenho de métodos numéricos tradicionalmente empregados para a obtenção dos referidos estimadores.

Sejam k grupos, populacionais ou amostrais, G_1 , G_2 , ... , G_k contendo n_1 , n_2 , ... , n_k observações, respectivamente, na forma $\underline{\mathbf{X}}^{\mathrm{T}} = (x_0 , x_1 , ... , x_p)$, onde $x_0 \equiv 1$, por conveniência, e as demais variáveis podem ser discretas ou contínuas. O problema aqui abordado é o do Reconhecimento e Classificação, isto é, dar uma descrição algébrica, ou gráfica, de características diferenciais das observações, com valores numéricos que permitam a máxima separação dos grupos estudados, além de encontrar uma regra que permita a alocação de uma nova observação em um dos grupos estudados.

Uma informação importante na configuração dos dados diz respeito à separação, ou sobreposição, dos k grupos estudados. Na Figura 2.1 são considerados como exemplo três grupos, isto é, $k = 3$, com $p = 2$ variáveis independentes, X_1 e X_2 . A Figura 2.1(a) mostra total sobreposição dos grupos. A Figura 2.1(b) ilustra separação, ou sobreposição, parcial. Na Figura 2.1(c) pode-se observar separação completa dos grupos. Estes conceitos foram formalizados por Albert e Anderson (1984) conforme o raciocínio mostrado a seguir.

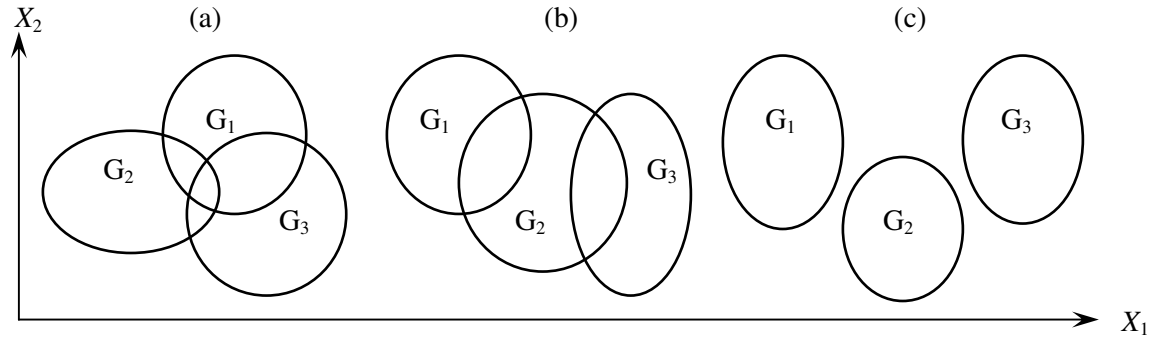

Figura 2.1 – Diferentes configurações de conjuntos de dados.

Seja a matriz **X**, de ordem $n \times (p + 1)$, de posto $(p + 1)$, por suposição, definida como:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{p1} \\ 1 & x_{12} & \ldots & x_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & \ldots & x_{pn} \end{bmatrix} . \qquad (2.1)$$

e que também costuma ser apresentada na forma:

$$\mathbf{X} = \left[\underline{\mathbf{1}} \quad \underline{\mathbf{x}}_1 \quad \ldots \quad \underline{\mathbf{x}}_p \right] . \qquad (2.1b)$$

Seja L_s o conjunto de linhas identificadoras das observações de G_s , $s = 1, 2, \ldots , k$. Diz-se que há separação completa entre os grupos se existe um vetor $\underline{\mathbf{B}} \in R^m$, onde $m = (k - 1)(p + 1)$, tal que, para todo $i \in L_j$, e para j , $t = 1, 2, \ldots , k$, $j \neq t$,

$$(\underline{\mathbf{B}}_j - \underline{\mathbf{B}}_t)^T \underline{\mathbf{X}}_i > 0 . \qquad (2.2)$$

Diz-se que há separação quase completa se

$$(\underline{\mathbf{B}}_j - \underline{\mathbf{B}}_t)^T \underline{\mathbf{X}}_i \geq 0 , \qquad (2.3)$$

com a igualdade valendo para, no mínimo, uma tripla (i , j , t). Sejam $j(i)$, o valor de j para o qual $i \in L_j$, e $Q(\underline{\mathbf{B}})$, o conjunto de indicadores das observações que satisfazem a igualdade em (2.3). Diz-se que estas observações são quase separadas em relação a $\underline{\mathbf{B}}$.

Finalmente, diz-se que há sobreposição dos grupos quando existe uma tripla (i , j , t) tal que

$$(\underline{\mathbf{B}}_j - \underline{\mathbf{B}}_t)^T \underline{\mathbf{X}}_i < 0 . \qquad (2.4)$$

Segundo Albert e Anderson (1984) a detecção da separação entre grupos pode ser abordada de duas maneiras distintas, algébrica ou empírica. A abordagem algébrica, segundo os autores

citados, foi relatada por J. Burridge (em um trabalho não publicado) e é baseada em conceitos da Programação Linear, seguindo o raciocínio mostrado na seqüência.

Sejam dois grupos distintos, G_1 e G_2 . Diz-se que há separação completa, ou quase completa, quando existe $\mathbf{B}_1 \neq \underline{0}$, tal que

$$A\mathbf{B}_1 \leq 0\,, \tag{2.5}$$

onde A é uma matriz $n \times (p + 1)$ com linhas $(-x_i)$, para $i \in L_1$, e x_i , para $i \in L_2$. Caso a inequação (2.5) não possa ser satisfeita com $\mathbf{B}_1 \neq \underline{0}$, diz-se que os grupos estão sobrepostos.

A desigualdade (2.5) também pode ser escrita na forma:

$$(A, I_n)(\mathbf{B}_1, \underline{T}) = 0\ . \tag{2.6}$$

onde I_n é a matriz identidade de ordem n, e $\underline{T}$ é um vetor com n variáveis de folga. Com relação à solução há três possíveis conclusões:

1. Se existe uma solução tal que $t_i > 0$ ($i = 1, 2, \ldots , n$), diz-se que os grupos estão completamente separados.
2. Se existe uma solução tal que $t_i \geq 0$ ($i = 1, 2, \ldots , n$), com a igualdade verificando-se para no mínimo um valor de i, diz-se que a separação é quase completa.
3. Se nenhuma das condições acima for verificada, diz-se que os grupos estão sobrepostos.

Em uma expansão do trabalho de Albert e Anderson, Santner e Duffy (1986) apresentam um modelo de Programação Linear que classifica os dados como (i) completamente separados, (ii) quase separados ou (iii) sobrepostos. Tal modelo é brevemente descrito a seguir.

Inicialmente considera-se o vetor $\mathbf{B}^{\mathrm{T}} = (\mathbf{B}^{\mathrm{T}}_1 , \ldots , \mathbf{B}^{\mathrm{T}}_{k-1})$, $\mathbf{B} \in R^m$, onde $m = (p + 1)(k - 1)$. Além disso, para cada i, $1 \leq i \leq n$, $j(i)$ representa o valor da variável resposta Y_i , isto é, $Y_i = j(i)$. Sejam os conjuntos A^C, de todos os vetores $\mathbf{B}$ que satisfazem (2.2), e A^Q, de todos os vetores $\mathbf{B}$ que satisfazem (2.3). Seja a matriz em blocos $\overline{\mathbf{X}}$, de dimensão $n(k - 2) \times m$, onde cada bloco tem dimensão $(k - 2) \times m$, e definida da seguinte forma:

1. Se $j(i) < k$, então uma linha é $\chi^{T}_{i}\, Q_{j(i)}$ e $(k-3)$ linhas são da forma $\chi^{T}_{i}\,\{Q_{j(i)} - Q_{t}\}$, considerando $t \in \{1, \ldots, k-2\} \setminus \{j(i)\}$; e
2. Se $j(i) = k$, então $(k-3)$ linhas são da forma $\chi^{T}_{i}\,\{-Q_{t}\}$, para $t \in \{1, \ldots, k-2\}$.

Aqui Q_t, para $1 \le t \le k-2$, é a matriz de dimensão $(p+1) \times m$, de elementos nulos ou unitários, tal que $Q_t\,\underline{\mathbf{B}} = \underline{\mathbf{B}}_t$ e χ_i é a i-ésima coluna de $\overline{\mathbf{X}}$. Seja $\mathbf{T}(\underline{\mathbf{B}})$ o produto cartesiano $\mathbf{X}\{\mathbf{T}_i(\underline{\mathbf{B}}): i \in Q^m\}$. O modelo de Programação Linear proposto pelos autores é da forma:

$$max \quad \sum_{i=1}^{n} z_i$$

$$sujeito \quad a \begin{cases} \left[\overline{\mathbf{X}} - \mathbf{I}_{\mathbf{n(k-2)}}\right]\left[\underline{\mathbf{B}} \;\; | \;\; \underline{\mathbf{S}}\right] = 0 \\ w_j \le \dfrac{s_j}{m} \quad (j = 1,\ldots, n(k-2)) \\ z_i\left(k - \dfrac{5}{2}\right) \le \displaystyle\sum_{j=(i-1)(k-2)+1}^{i(k-2)} w_j \quad (i = 1,\ldots, n) \end{cases}$$

onde $\underline{\mathbf{B}}$ é arbitrariamente escolhido, $s > 0$, w_j e $z_i \in \{0, 1\}$, $\mathbf{I}_{n(k-2)}$ é a matriz identidade, e $m > 0$ é escolhido de modo que $m \le s_j$ verifica-se para $s_j > 0$, sendo $\underline{\mathbf{S}} = \mathbf{T}_i^m \cup \{j(i)\}$.

De acordo com os autores, o modelo acima é sempre factível. Além disso, um componente de $\underline{\mathbf{S}}$ é positivo se, e somente se, a inequação $\overline{\mathbf{X}}\underline{\mathbf{B}} \ge 0$ é estritamente satisfeita. Também, $z_i = 1$ se, e somente se, Y_i é completamente separado. A função objetivo maximiza a cardinalidade do complemento de Q^m, onde $Q^m = \bigcap_{\underline{\mathbf{B}} \in A^Q} Q(\underline{\mathbf{B}})$.

2.2 MEDIDA DA SOBREPOSIÇÃO NO CASO BINÁRIO

Uma abordagem proposta por Christmann e Rousseeuw (2001) efetua a mensuração da sobreposição dos grupos para casos onde a variável resposta, Y, é binária, ou dicotômica. São definidos dois valores: $n_{overlap}$, que representa o menor número de observações cuja remoção é necessária para tornar impossível a existência dos EMV, e $n_{complete}$, que representa o menor número

de observações cuja remoção produz separação completa. Por definição, $n_{\text{overlap}} \leq n_{\text{complete}}$. A determinação de ambos os valores, e também dos conjuntos de índices correspondentes às mencionadas observações, é feita mediante a aplicação de um algoritmo proposto pelos autores, e cujo desenvolvimento prescinde de duas definições, apresentadas a seguir, relativas a um modelo de regressão linear aplicado a um conjunto de dados na forma que se segue:

$$Z_n = \{(x_{1,i}, x_{2,i}, ..., x_{p,i}, y_i); i = 1, 2, ..., n\} . \tag{2.7}$$

O objetivo é ajustar a cada y_i um hiperplano afim, pertencente ao espaço R^{p+1}, ou seja:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_p x_{p,i} \tag{2.8}$$

$$y_i = g\left(\underline{\mathbf{B}}^T (1, \underline{\mathbf{X}}_i)\right) \quad . \tag{2.9}$$

Definição 2.1: Diz-se que um vetor $\underline{\mathbf{B}}$ é não ajustado a Z_n se, e somente se, existe um hiperplano afim $V \in R^{p+1}$ que não contenha nenhum $\underline{\mathbf{X}}_i$ e tal que os resíduos

$$r_i(\underline{\mathbf{B}}) = y_i - g\left(\underline{\mathbf{B}}^T (1, \underline{\mathbf{X}}_i)\right) > 0 \tag{2.10}$$

para qualquer $\underline{X}_i$ em um dos subespaços e

$$r_i(\underline{\mathbf{B}}) = y_i - g\left(\underline{\mathbf{B}}^T (1, \underline{\mathbf{X}}_i)\right) < 0 \tag{2.11}$$

no outro subespaço.

Definição 2.2: A profundidade de regressão de um ajuste $\underline{\mathbf{B}}$ aos dados Z_n é o menor número de observações que necessitam ser removidas para tornar $\underline{\mathbf{B}}$ não ajustado, no sentido da Definição 2.1. O menor número de resíduos que necessitam de alteração de sinal é denotado por *rdepth*($\underline{\mathrm{B}}$, Z_n).

A profundidade de regressão é invariante com relação às transformações monótonas, no sentido de que é possível substituir y_i por uma função estritamente monótona $h(y_i)$, desde que g seja substituída por $(h \bullet g)$. Esta propriedade é válida, segundo Rousseeuw (1984), porque a

profundidade de regressão, conforme a Definição 2.1, depende apenas das variáveis independentes $\underline{\mathbf{X}}_i$ e dos sinais dos resíduos (2.10) e (2.11). Ainda de acordo com Rousseeuw (1984), esta propriedade não é verificada para a maioria dos métodos de estimação, entre os quais o Método dos Mínimos Quadrados (MMQ).

Para o caso de variável resposta binária, pode-se definir a profundidade de regressão através da substituição da função g pela função distribuição acumulada da distribuição logística. Neste caso, a medida é invariante com relação aos diferentes valores da variável resposta binária, e pode ser computada através de um algoritmo, desenvolvido por Rousseeuw e Hubert (1999), de ordem O($n$log(n)), para $p = 2$, ou de um algoritmo de ordem O(n^{p-1}log(n)), para $p \in \{3, 4\}$, desenvolvido por Rousseeuw e Struyf (1998). Ainda de acordo com Christmann e Rousseeuw (2001), a determinação do número mínimo exato de observações misturadas para p arbitrário, com base em um hiperplano afim, é um problema essencialmente NP-difícil.

O algoritmo proposto por Christmann e Rousseeuw (2001) consiste dos seguintes passos:

1. Ler o conjunto de dados na forma (2.7), considerando $y_i \in \{0, 1\}$, $i = 1, 2, \ldots, n$. Normalizar as variáveis $x_{1i}, x_{2i}, \ldots, x_{pi}$.
2. Determinar o número n_a de pontos distintos, na forma $(x^a_{j,1}, \ldots, x^a_{j,p-1}, y^a_j)$ em Z_n. Para cada $j \in \{1, \ldots, n_a\}$ contar o número t_j de pontos coincidentes, de modo que $n = \sum_{j=1}^{n_a} t_j$.

   Deste ponto em diante trabalha-se com o conjunto de dados agregados $Z^a_n = \{(x^a_{j,1}, \ldots, x^a_{j,p-1}, y^a_j; t_j); 1 \le j \le n_a\}$. Para computar n_{overlap} ou n_{complete}, os pontos coincidentes são contados como t_j pontos.
3. Se $p = 2$, aplicar o algoritmo exato para n_{overlap}, ou n_{complete}, ao conjunto de dados agregados. Ir para o passo 7.
4. Se $p > 2$, usar o algoritmo de aproximação baseado em projeções. Definir o número *NITER* de subgrupos a explorar. Iniciar o gerador de números aleatórios. Fixar os valores *NSIN* = 0, *ITER* = 1 e $n_{\text{overlap}} = n$, $n_{\text{complete}} = n$.
5. Explorar um subgrupo escolhido aleatoriamente, e de tamanho $(p - 1)$, de Z^a_n. Se o conjunto $\{(x^a_{j,1}, 1)', \ldots, (x^a_{j,p-1}, 1)'\}$ é linearmente dependente, fazer *NSIN* = *NSIN* + 1 e explorar o próximo subgrupo. Senão ir para o passo 6.

6. Projetar todos os x^a_j na direção $\underline{\mathbf{U}}$, ortogonal ao hiperplano dado pelo subgrupo. Agregar o conjunto bidimensional $\{(x^a_j \underline{\mathbf{U}}^T , y_j); j = 1,..., n_a\}$ e o correspondente valor de t_j , conforme definido no passo 2, e contar os pontos coincidentes. Computar n_{overlap} bidimensional. Se for menor que o atual valor de n_{overlap} , atualizá-lo, e proceder da mesma forma para n_{complete} . Fazer *ITER* = *ITER* + 1. Se *ITER* > *NITER*, ir para o passo 7. Senão, retornar ao passo 5.
7. Fornecer a aproximação resultante de n_{overlap} , ou n_{complete} , a direção correspondente de $\underline{U}$, e para $p > 2$ o número *NSIN* de subgrupos singulares encontrado.

De acordo com os autores, o algoritmo acima descrito tem a sua precisão e o seu tempo de computação fortemente afetados pelo número de subgrupos analisados. O número de variáveis independentes também contribui consideravelmente para o aumento do esforço computacional requerido para a mensuração da sobreposição. Além disso, o método em questão avalia a sobreposição entre as observações de dois grupos a cada vez, o que pode ser um sério obstáculo à sua utilização para problemas com três ou mais grupos de observações.

2.3 MÉTODO DO CUSTO MÍNIMO ESPERADO DE MISTURA

Sejam os grupos G_1 , G_2 , ... , G_k contendo n_1 , n_2 , ... n_k observações, respectivamente, na forma $\underline{\mathbf{X}}^T = (x_0 , x_1 , ... , x_p)$, onde $x_0 \equiv 1$ e as demais variáveis podem ser discretas ou contínuas. Sejam f_i ($\underline{\mathbf{X}}$) , a função densidade de probabilidade (f.d.p.) associada ao grupo G_i , $i = 1 , 2 , ... , k$; p_i , a probabilidade *a priori* do grupo G_i , e $C(j \mid i)$, o custo de classificar uma observação em G_j , quando a mesma pertence, de fato, ao grupo G_i , sendo $C(i \mid i) = 0$. Além disso, sejam R_j , o conjunto de observações $\underline{\mathbf{X}}$ classificadas como pertencentes ao grupo G_j , e a probabilidade $P(j \mid i)$, de classificar uma observação em G_j , quando a mesma pertence, de fato, ao grupo G_i , dada por:

$$P(j \mid i) = \int_{R_j} f_i(\underline{\mathbf{X}}) d\underline{\mathbf{X}} . \tag{2.12}$$

Também,

$$P(i \mid i) = 1 - \sum_{\substack{j=1 \\ j \neq i}}^{k} P(j \mid i) . \tag{2.13}$$

A classificação de uma observação do grupo G_i em um grupo G_j , $i \neq j$, tem um custo esperado de confusão dado por:

$$ECM(i) = \sum_{\substack{j=1 \\ j \neq i}}^{k} P(j \mid i) C(j \mid i). \tag{2.14}$$

Multiplicando-se cada *ECM* por sua respectiva probabilidade *a priori*, e efetuando-se a soma dos produtos, obtém-se:

$$ECM = \sum_{i=1}^{k} p_i \left[\sum_{\substack{j=1 \\ j \neq i}}^{k} P(j \mid i) C(j \mid i) \right] . \tag{2.15}$$

Então a obtenção de um procedimento ótimo de classificação consiste em escolher regiões R_1 , R_2 , ... , R_k , tais que (2.15) seja minimizada. Tais regiões são definidas pela classificação de uma observação $\underline{\mathbf{X}}$ no grupo G_j , $j = 1 , 2 , \ldots , k$, para o qual é mínima a soma definida por:

$$\sum_{\substack{i=1 \\ i \neq j}}^{k} p_i f_i(\underline{\mathbf{X}}) C(j \mid i) \quad . \tag{2.16}$$

Caso ocorra um empate, a observação pode ser classificada em qualquer dos grupos empatados.

De acordo com Johnson e Wichern (1988), quando os custos esperados forem os mesmos, sem perda de generalidade podem ser fixados em 1 (um), a soma (2.16) é pequena quando o termo omitido é grande. Desta forma, pode-se expressar a regra do custo esperado mínimo de confusão como:

Classificar $\underline{\mathbf{X}}$ como pertencente ao grupo G_j , $j = 1 , 2 , \ldots , k$, se

$$p_j f_j(\underline{\mathbf{X}}) > p_i f_i(\underline{\mathbf{X}}) \quad , \quad \forall \; i \neq j \quad . \tag{2.17}$$

A expressão (2.17) também pode ser apresentada como:

$$\ell np_j f_j(\underline{\mathbf{X}}) > \ell np_i f_i(\underline{\mathbf{X}}) \quad , \quad \forall \; i \neq j \quad . \tag{2.18}$$

Convém acrescentar que a regra anterior é idêntica à regra que maximiza a probabilidade *a posteriori*, $P(G_j \mid \underline{\mathbf{X}})$, dada por:

$$P(G_j \mid \underline{\mathbf{X}}) = \frac{p_j f_j(\underline{\mathbf{X}})}{\sum_{i=1}^{k} p_i f_i(\underline{\mathbf{X}})} \quad . \tag{2.19}$$

2.4 MODELO DE REGRESSÃO LOGÍSTICA

Sejam os grupos G_1 , G_2 , ... , G_k contendo n_1 , n_2 , ... , n_k , observações, respectivamente, na forma $\underline{\mathbf{X}}^{\mathrm{T}} = (x_0 , x_1 , ... , x_p)$, onde $x_0 \equiv 1$ e as demais variáveis, ou covariáveis, podem ser discretas ou contínuas. Seja Y_s , $s = 1, 2, ... , k$, a variável resposta, na forma $\mathbf{Y}^{\mathrm{T}} = (y_1 , y_2 , ... , y_n)$, que indica o grupo ao qual pertence cada observação. O Modelo de Regressão Logística assume que as probabilidades *a posteriori* têm a forma:

$$P(G_s \mid \underline{\mathbf{X}}) = \frac{\exp(\mu_s)}{\sum_{j=1}^{k} \exp(\mu_j)} \qquad (s = 1 , 2 , ..., k) \quad , \tag{2.20}$$

onde

$$\mu_s = \beta_{s0} + \beta_{s1} x_1 + \beta_{s2} x_2 + ... + \beta_{sp} x_p = \beta_{s0} + \sum_{i=1}^{p} \beta_{si} x_i = \underline{\mathbf{B}}_s^T \underline{\mathbf{X}} \qquad (s = 1, ..., k-1) \; , \tag{2.21}$$

e $\underline{\mathbf{B}}_k = \underline{0} \Rightarrow \mu_k = 0$.

Na forma anterior o *k*-ésimo grupo é adotado neste trabalho como grupo base, ou de referência. Convém ressaltar que alguns autores, como Hosmer e Lemeshow (1989), preferem tomar o grupo 1 como referência. Também há autores que consideram 0 (zero) como valor inicial para *s*, como Santner e Duffy (1986), por exemplo. Uma vez escolhido o grupo de referência, a estimação dos parâmetros segue um raciocínio análogo ao desenvolvido para modelos com resposta binária.

A forma (2.20) permite a modelagem da relação entre a variável resposta e o vetor $\underline{\mathbf{X}}$ de covariáveis, ou variáveis explanatórias. De acordo com McLachlan (1992), as primeiras aplicações do Modelo de Regressão Logística ocorreram no estudo prospectivo de doenças coronárias, por Cornfield (1962) e Truett, Cornfield e Kannel (1967). Nestes casos a estimação dos parâmetros seguia a suposição de normalidade. O problema da estimação em um contexto mais amplo foi considerado por Cox (1966, 1970), Day e Kerridge (1967) e por Walker e Duncan (1967). Ainda de acordo com McLachlan (1992), a discriminação logística é amplamente aplicável a uma grande variedade de famílias de distribuições, tais como:

1. Distribuições normais multivariadas com matrizes de covariâncias iguais.
2. Distribuições discretas multivariadas seguindo o modelo log-linear.
3. Distribuições conjuntas de variáveis aleatórias contínuas e discretas, não necessariamente independentes.

A regra mais simples para discriminação consiste em alocar uma observação $\underline{\mathbf{X}}$ no grupo G_s se, e somente se,

$$(\underline{\mathbf{B}}_s - \underline{\mathbf{B}}_t)^T \underline{\mathbf{X}} \geq 0 \quad (t = 1, 2, \ldots, k). \tag{2.22}$$

A função de verossimilhança condicional para k grupos pode ser expressa na forma:

$$\ell(\underline{\mathbf{B}}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[P(G_j \mid \underline{\mathbf{X}}_i) \right]^{Y_{ji}} \quad , \tag{2.23}$$

onde Y_{ji} é a variável resposta, indicadora do grupo ao qual pertence a i-ésima observação, isto é:

$$Y_{ji} = \begin{cases} 1 \, , \, se \; y_i = j \\ 0 \, , \, se \; y_i \neq j \end{cases} , \; i = 1 \, , \ldots , n, \; j = 1 \, , \ldots , k.$$

Do ponto de vista matemático é mais conveniente trabalhar com a função log-verossimilhança, dada por:

$$L(\underline{\mathbf{B}}) = \sum_{i=1}^{n}\left[\sum_{j=1}^{k-1} Y_{ji}\mu_j - \ell n\left(1+\sum_{j=1}^{k-1}\exp\mu_j\right)\right]. \qquad (2.24)$$

2.4.1 Estimadores de Máxima Verossimilhança

Os Estimadores de Máxima Verossimilhança (EMV) dos parâmetros são encontrados mediante a resolução do sistema de equações formado pelas derivadas parciais de (2.24) em relação a cada um dos $(k-1)(p+1)$ parâmetros desconhecidos, igualadas a zero, e cuja forma geral é:

$$\frac{\partial L(\underline{\mathbf{B}})}{\partial \beta_{jm}} = \sum_{i=1}^{n} x_{mi}\left[Y_{ji} - P(G_j \mid \underline{\mathbf{X}}_i)\right] \quad (j = 1, 2, \ldots, k-1, \text{ e } m = 0, 1, \ldots, p). \qquad (2.25)$$

O procedimento mais utilizado na obtenção dos EMV é o método de Newton-Raphson, que resulta na expressão dada por:

$$\underline{\mathbf{B}}^{(m+1)} = \underline{\mathbf{B}}^{(m)} + \left[I\left(\underline{\mathbf{B}}^{(m)}\right)\right]^{-1}\left[S\left(\underline{\mathbf{B}}^{(m)}\right)\right] \qquad (2.26)$$

onde $S(\underline{\mathbf{B}}^{(m)})$ é o vetor com $(k-1)(p+1)$ parâmetros, dados por (2.25) e $I(\underline{\mathbf{B}}^{(m)})$ é uma matriz quadrada, de ordem $(k-1)(p+1)$, cujos elementos são os negativos dos valores esperados para as derivadas parciais de segunda ordem, na forma que se segue:

$$\frac{\partial^2 L(\underline{\mathbf{B}})}{\partial\beta_{jm}\partial\beta_{jm'}} = -\sum_{i=1}^{n} x_{m'i}x_{mi}\left[P(G_j \mid \underline{\mathbf{X}}_i)\right]\left[1 - P(G_j \mid \underline{\mathbf{X}}_i)\right] \qquad (2.27)$$

$$\frac{\partial^2 L(\underline{\mathbf{B}})}{\partial\beta_{jm}\partial\beta_{j'm'}} = \sum_{i=1}^{n} x_{m'i}x_{mi}\left[P(G_j \mid \underline{\mathbf{X}}_i)\right]\left[P(G_{j'} \mid \underline{\mathbf{X}}_i)\right] \qquad (2.28)$$

onde $j, j' = 1, 2, \ldots, k-1$, e $m, m' = 0, 1, \ldots, p$.

A matriz de informação $I(\underline{\mathbf{B}}^{(m)})$ pode ser escrita na forma:

$$I(\underline{\mathbf{B}}) = \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} \qquad (2.29)$$

onde **V**, para variável resposta binária, é a matriz diagonal $n \times n$ de variâncias, isto é,

$$V_{ii} = P(G \mid \underline{\mathbf{X}}_i)[1 - P(G \mid \underline{\mathbf{X}}_i)] \quad . \tag{2.30}$$

Para variável resposta politômica a matriz de informação pode ser escrita na forma:

$$I(\underline{\mathbf{B}}) = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1(k-1)} \\ B_{21} & B_{22} & \cdots & B_{2(k-1)} \\ \cdots & \cdots & \cdots & \cdots \\ B_{(k-1)1} & B_{(k-1)2} & \cdots & B_{(k-1)(k-1)} \end{bmatrix} \quad , \tag{2.31}$$

onde

$$B_{ij} = \begin{cases} \underline{\mathbf{X}}^T \left[diag \left[P(G_s \mid \underline{\mathbf{x}}_i)(1 - P(G_s \mid \underline{\mathbf{x}}_i)) \right] \right] \underline{\mathbf{X}} & , \quad i = j \quad , \quad s = i \\ (-1)\underline{\mathbf{X}}^T \left[diag \left[P(G_i \mid \underline{\mathbf{x}}_i) P(G_j \mid \underline{\mathbf{x}}_i) \right] \right] & , \quad i \neq j \end{cases} \quad . \tag{2.32}$$

De acordo com Anderson (1972), Albert e Anderson (1984) e Albert e Lesaffre (1986), os estimadores para os parâmetros existem se, e somente se, houver sobreposição completa dos grupos. Aqui deve-se entender a existência no sentido de unicidade da solução, isto é, se os grupos não estão completamente sobrepostos, conforme a Figura 2.1(a), os estimadores obtidos pelo método da máxima verossimilhança não são únicos, ou tendem ao infinito, conforme os teoremas a seguir. Neste ponto é importante ressaltar a necessidade de escolher adequadamente o grupo base. Caso haja sobreposição parcial, conforme a Figura 2.1(b), os modelos poderão ser obtidos desde que G_2 seja escolhido como grupo base. Se a escolha recair sobre G_1 ou G_3 , os estimadores de máxima verossimilhança não serão encontrados. Para tais casos há algumas abordagens alternativas, que serão apresentadas adiante.

Teorema 2.1 – Se existe separação completa dos grupos de dados, então os estimadores de máxima verossimilhança para $\underline{\mathbf{B}}$ não existem, e

$$\max_{\beta} L(\underline{\mathbf{B}}) = 1 \quad .$$

Para provar o teorema anterior, Albert e Anderson (1984) definem inicialmente o conjunto A^C de todos os vetores $\underline{\mathbf{B}}$ que satisfazem a desigualdade (2.2), isto é:

$$(\underline{\mathbf{B}}_j - \underline{\mathbf{B}}_t)^{\mathrm{T}} \underline{\mathbf{X}}_i > 0 \quad .$$

Na seqüência consideram $\mathrm{B}(k) = k\underline{\mathbf{B}}$, onde $\underline{\mathbf{B}}$ pertence ao conjunto A^C e $k > 0$. Então a função verossimilhança pode ser escrita na forma:

$$L(\underline{\mathbf{X}}, k\underline{\mathbf{B}}) = \sum_{j=1}^{k} \sum_{i \in G_j} \ell n \left[\frac{1}{\sum_{t=1}^{k} \exp\left\{ - k(\underline{\mathbf{B}}_j - \underline{\mathbf{B}}_t)^T \underline{\mathbf{x}}_i \right\}} \right] . \qquad (2.33)$$

Os autores citados consideram o comportamento da função acima quando $k \to \infty$. Neste caso a desigualdade (2.2) ainda é satisfeita. Então os termos exponenciais tendem a zero, exceto aquele no qual a soma é sobre t quando $t = j$, onde k é igual a um. Desta forma $L(\underline{\mathbf{X}}, k\underline{\mathbf{B}}) \to 0$ no ponto de máximo absoluto, quando $k \to \infty$. Os autores concluem que o máximo absoluto da função é atingido no infinito, sobre a fronteira do espaço de parâmetros.

Teorema 2.2 – Se existe sobreposição dos grupos de dados, então os estimadores de máxima verossimilhança para $\underline{\mathbf{B}}$ existem e são únicos.

De acordo com os autores a prova para este teorema é dada por Silvapulle (1981), para modelos com resposta dicotômica. Para modelos com resposta politômica a demonstração tem como base o argumento de que $L(\underline{\mathbf{X}}, k\underline{\mathbf{B}}) \to -\infty$ e também que a função em questão é estritamente côncava.

As provas detalhadas para os teoremas citados podem ser encontradas em Albert e Anderson (1984). Uma importante referência adicional é o trabalho de Santner e Duffy (1986).

Outra abordagem ao problema da detecção de separação completa, ou quase completa, é defendida por Lesafre e Albert (1989). Segundo estes dois autores, as regras de identificação das referidas configurações poderiam tomar como base a resposta de um programa padrão de máxima verossimilhança. Como apoio ao seu argumento os referidos autores provam que a separação é

totalmente determinada pelo comportamento dos erros padrões dos estimadores no processo iterativo. Seguindo esta mesma linha de raciocínio, Heinze e Schemper (2002) afirmam que o problema da separação de grupos pode ser contornado monitorando-se a variância dos estimadores durante a execução do processo iterativo.

A separação completa dos grupos de dados não representa um grande problema quando a variável resposta é dicotômica, já que esta ocorrência em situações práticas pode ser contornada pela utilização de outros métodos de classificação, como a Função Discriminante Linear, por exemplo. Entretanto, para variável resposta politômica, a ausência de sobreposição pode tornar até mesmo impraticável a estimação dos parâmetros do modelo, especialmente em problemas envolvendo mais de três variáveis independentes ($p > 3$), já que a identificação de grupos totalmente separados torna-se mais complexa, particularmente quando há um número elevado de grupos envolvidos e de variáveis independentes. Nestas condições a implementação das soluções propostas por Santner e Duffy (1986) e por Christmann e Rousseeuw (2001) pode exigir um elevado esforço computacional, o que pode ser um obstáculo à sua utilização. Algumas abordagens alternativas, brevemente descritas a seguir, são sugeridas por Begg e Gray (1984) e por Rom e Cohen (1995). São técnicas que, além de permitir a estimação dos parâmetros, quando estes existem, possibilitam a identificação dos grupos totalmente separados do grupo de referência.

2.4.2 Modelos de Regressão Logística Individualizados

A primeira abordagem, apresentada por Begg e Gray (1984), propõe o uso de Modelos de Regressão Logística Individualizados (MRLI), na qual obtém-se uma série de modelos de Regressão Logística simples, em substituição ao modelo politômico. Seja a probabilidade:

$$P_{ji} = P(Y_{ji} = 1 \mid \underline{\mathbf{x}}_i) \quad , \quad j = 1,2,...,k\,. \tag{2.34}$$

No modelo politômico pode-se considerar que:

$$\ell n\left(\frac{P_{ji}}{P_{ki}}\right) = \underline{\tilde{\mathbf{B}}}^{T}\,\underline{\mathbf{x}}_{ii} \quad , \quad j = 1,2,...,k-1. \tag{2.35}$$

Para obter modelos individualizados, que comparam cada categoria com a categoria de referência, pode-se adotar um modelo na forma:

$$\ell n\left(\frac{Q_{di}}{Q_{ki}}\right) = \underline{\mathbf{A}}_d^{\mathrm{T}}\,\underline{\mathbf{x}}_i \quad , \quad d = 1, 2, ..., k-1 . \tag{2.36}$$

onde

$$Q_{di} = P(Y_{di} = 1 \mid \underline{\mathbf{x}}_i, Y_{di} + Y_{ki} = 1)$$

e

$$Q_{ki} = P(Y_{di} = 0 \mid \underline{\mathbf{x}}_i, Y_{di} + Y_{ki} = 1) .$$

Segundo os autores, é fácil verificar que os dois modelos são parametricamente equivalentes, isto é, $\underline{\mathbf{A}}_d = \underline{\mathbf{B}}_d$. De acordo com o teorema de Bayes:

$$Q_{di} = \frac{P_{di}}{P_{di} + P_{ki}} . \tag{2.37}$$

Também

$$\frac{P_{di}}{P_{ki}} = \frac{Q_{di}}{1 - Q_{di}} . \tag{2.38}$$

O método alternativo proposto pelos autores considera as probabilidades condicionais individualizadas dadas por:

$$IP_{di} = P\left[Y_{di} = 1 \mid \underline{\mathbf{x}}_i, Y_{di} + Y_{1i} = 1\right]$$

e

$$1 - IP_{di} = P\left[Y_{di} = 0 \mid \underline{\mathbf{x}}_i, Y_{di} + Y_{1i} = 1\right]$$

com $i = 1, 2, \ldots, n$ e $d = 2, \ldots, k$.

Então tem-se que:

$$\ell n\left[\frac{IP_{di}}{1 - IP_{di}}\right] = \ell n\left[\frac{P_{di}}{P_{1i}}\right] = \underline{\tilde{\mathbf{B}}}_d^{T}\,\underline{\mathbf{x}}_i \quad . \tag{2.39}$$

A abordagem apresentada permite, de acordo com os autores, ajustar $(k - 1)$ diferentes modelos de Regressão Logística para variável resposta binária, nos quais cada grupo d é comparado com o grupo de referência, aqui tomado como o Grupo 1. Se o procedimento for adotado para obter $(k - 1)$ modelos, os estimadores $\underline{\tilde{\mathbf{B}}}_1,...,\underline{\tilde{\mathbf{B}}}_{k-1}$, podem ser substituídos em (2.20). Além disso, se for empregado o Método da Máxima Verossimilhança, os estimadores citados serão assintoticamente não viesados. A eficácia do procedimento é avaliada a partir da eficiência assintótica relativa (EAR). Ainda, de acordo com os autores, a perda na EAR dos estimadores é muito pequena. Apesar disso, é possível que os estimadores não sejam todos encontrados. Uma situação típica é ilustrada pela Figura 2.1(b). Se o grupo G_1 for adotado como categoria de referência, o estimador $\underline{\tilde{\mathbf{B}}}_3$ não existirá, pois não há sobreposição entre G_1 e G_3 .

Embora o método seja analiticamente flexível, os autores levantam algumas questões a respeito do mesmo. A primeira diz respeito à eficiência assintótica dos estimadores com relação aos Estimadores de Máxima Verossimilhança do modelo politômico. Outra questão refere-se à eficiência das combinações lineares dos estimadores, especialmente das combinações de variáveis de diferentes modelos de regressão. Uma terceira questão é relativa à possibilidade de utilizar os modelos individualizados em testes de hipóteses efetuados com o objetivo de avaliar o impacto de uma variável explanatória sobre qualquer dos modelos. Finalmente, a última questão levantada envolve a possibilidade de desenvolver formulações práticas para obter intervalos de confiança para as probabilidades preditas, quando as variâncias dependem de todos os parâmetros.

De acordo com Hosmer e Lemeshow (1989), os estimadores obtidos através do método em questão são consistentes e não apresentam grande perda de eficiência, assumindo valores bastante próximos aos valores obtidos para os estimadores através do Método da Máxima Verossimilhança, conforme descrito em 2.1. Ainda de acordo com os autores citados, os modelos individualizados podem ser bastante úteis para tratar um problema que, embora não ocorra no caso de variável resposta binária, é comum em problemas com resposta politômica. Trata-se da situação na qual uma variável independente é significativa para apenas uma das funções discriminantes. Como o modelo (2.20) não é adequado ao tratamento deste problema, os Modelos Individualizados podem fornecer funções discriminantes que envolvem diferentes variáveis independentes. A significância em questão pode ser estimada através do Teste de Wald. Este teste fornece um estimador z_i , $i = 0$, 1 , ... , p, para cada estimador $\underline{\mathbf{B}}_i$, dado por:

$$z_i = \left[\frac{\underline{\mathbf{B}}_i}{SE(\underline{\mathbf{B}}_i)} \right]^2 . \qquad (2.40)$$

Alguns autores apontam problemas com o uso do Teste de Wald. De acordo com Menard (1995), quando o valor encontrado para o estimador é grande, o erro padrão é inflacionado, o que reduz o valor da estatística utilizada. Segundo Agresti (2002), o Teste da Razão das Verossimilhanças é mais confiável. A estatística deste teste é dada por:

$$-2\left[L(\underline{\mathbf{B}}) - L(\underline{\mathbf{B}}_i)\right] , \qquad (2.41)$$

onde $L(\underline{\mathbf{B}}_{\ i})$ é a log-verossimilhança para o modelo sem a i–ésima variável, e $L(\underline{\mathbf{B}})$ é a log-verossimilhança do modelo com todas as variáveis. Tanto (2.40) como (2.41) seguem distribuição Qui-Quadrado com $(n - p - 1)$ graus de liberdade.

Outra medida de adequação de ajustamento é a estatística χ^2_{arc} baseada na transformação arco-seno, dada por

$$\chi^2_{arc} = \sum_{i=1}^{n} 4\left[arcsen\sqrt{y_i} - arcsen\sqrt{P(Y \mid \underline{\mathbf{x}}_i)}\right] . \qquad (2.42)$$

Uma modificação para o método de Begg e Gray é sugerida por Rom e Cohen (1995). O método em questão tem por base a idéia de que comparações de pares de categorias podem aumentar as informações a respeito da razão (2.35), além de permitir a obtenção dos estimadores para os parâmetros desconhecidos. A idéia consiste em ajustar $k(k + 1)/2$ modelos binários e depois estimar os parâmetros através do Método de Mínimos Quadrados Ponderados, de modo a obter os Estimadores Individualizados Ponderados. Segundo os autores, embora o procedimento requeira maior esforço computacional, a EAR é maior que aquela observada para a abordagem de Begg e Gray. Além disso, quando comparado ao Método da Máxima Verossimilhança, é mais vantajoso para estudos que envolvem grandes conjuntos de dados, quando limitações computacionais podem tornar intratável o problema de estimação dos parâmetros.

2.4.3 Modelo de Regressão Logística Oculto

Com o objetivo de contornar o problema da não existência de estimadores de máxima verossimilhança em função da separação de grupos, Rousseeuw e Christmann (2003) apresentaram o Modelo de Regressão Logística Oculto (MRLO). Esse modelo foi assim denominado pela semelhança com a camada oculta de alguns modelos de Redes Neurais, como será visto adiante. Na sua concepção assume-se que, devido a um mecanismo estocástico, a verdadeira resposta de um Modelo de Regressão Logística é não observável, e que existe uma variável observável fortemente relacionada à verdadeira resposta. A estimação dos parâmetros do modelo resultante é realizada através de métodos de estimação robusta, tendo como referência conceitos formulados por Ekholm e Palmgren (1982) e Copas (1988).

A estimação robusta geralmente é abordada de duas formas, conforme Kodzarkhia, Mishra e Reiersølmoen (2001). Uma é baseada na minimização da Função Verossimilhança, enquanto a outra tem como base funções de influência. Na seqüência apresenta-se de forma sucinta a aplicação deste método ao Modelo de Regressão Logística com variável resposta dicotômica, seguindo o raciocínio apresentado por Kodzarkhia, Mishra e Reiersølmoen (2001).

Sejam Y_1 , Y_2 , ... , Y_i , ... , Y_n observações de uma distribuição de Bernoulli, $b(1\ ,\ p_i\ )$, por suposição geradas a partir de um modelo linear geral com vetor de variáveis explanatórias $\underline{\mathbf{x}}^{\mathrm{T}} = (1, x_1,..., x_p)$, vetor de parâmetros $\underline{\mathbf{B}}^{\mathrm{T}} = (\beta_0, \beta_1,..., \beta_p)$ e uma função L tal que:

$$p_i = P(Y_i = 1 \mid \underline{\mathbf{x}}_i) = L(\underline{\mathbf{B}}^{\mathrm{T}}\, \underline{\mathbf{x}})\,.$$

No Modelo de Regressão Logística, $L(z)$ é a função *sigmóide*, isto é, $L(z) = (1+e^{-z})^{-1}$. A função densidade de probabilidade condicional de Y dado $\underline{\mathbf{x}}$ é:

$$\ell(y \mid \underline{\mathbf{x}}) = L(\underline{\mathbf{B}}^{\mathrm{T}}\, \underline{\mathbf{x}})\delta(y-1) + [1 - L(\underline{\mathbf{B}}^{\mathrm{T}}\, \underline{\mathbf{x}})]\delta(y)\,, \qquad (2.43)$$

onde $\delta(\ y\ )$ é o valor Delta de Dirac.

Sejam $\tilde{Y}_1$, $\tilde{Y}_2$, ... , $\tilde{Y}_i$, ... , $\tilde{Y}_n$ observações de um modelo onde assume-se que Y_i é contaminada por uma variável $T_i \sim b(1\ ,\ q_i\ )$. Deste modo tem-se que:

$$\tilde{Y}_i = (1-\varepsilon_{i,n})Y_i + \varepsilon_{i,n}T_i \ ,$$

onde $\varepsilon_{i,n} \sim b\left(1, \nu/\sqrt{n}\right)$, $0 \le \nu$. A taxa de contaminação $\nu/\sqrt{n}$ justifica-se pela contigüidade de alternativas indexadas pelo parâmetro $0 < v$ com relação ao modelo ideal, isto é $v = 0$.

Para as amostras [Y_1 , ... , Y_n] e [T_1 , ... , T_n], a função densidade de probabilidade condicional de T dado $\underline{\mathbf{x}}$ é:

$$f(y \mid \underline{\mathbf{x}}) = F(\underline{\mathbf{x}})\,\delta(y-1) + [1-F(\underline{\mathbf{x}})]\,\delta(y) \ , \qquad (2.44)$$

onde $F(\underline{\mathbf{x}}) = P(T_i = 1 \mid \underline{\mathbf{x}}_i) = q_i$. Então $Y_i \sim Ber(\tilde{p}_i)$, sendo $\tilde{p}_i = p_i + \frac{\nu}{\sqrt{n}}(q_i - p_i)$.

Os estimadores robustos também podem ser definidos como solução para a equação:

$$\sum_{i=1}^{n} \underline{\mathbf{w}}_i \underline{\mathbf{x}}_i \left\{ Y_i - L\left(\underline{\mathbf{B}}^{\mathrm{T}} \underline{\mathbf{x}}_i\right) - c\left(\underline{\mathbf{B}}, \underline{\mathbf{x}}_i\right) \right\} = 0 \quad .$$

Se $\underline{\mathbf{w}}_i = \mathbf{1}$ e $c(\ .\ ) = 0$, a solução fornece os Estimadores de Máxima Verossimilhança. De acordo com Gervini (2005), os EMV têm variância assintótica mínima, embora sejam sensíveis a certas configurações dos dados. Observações com valores que destoam da região de maior concentração dos dados podem influenciar os estimadores; além disso, se tais observações estão associadas a respostas incorretas, os estimadores resultantes podem ser seriamente viesados. Para implementar a estimação robusta, o referido autor considera a função *deviance*, dada por:

$$d(y_i, \underline{\mathbf{x}}_i, \underline{\mathbf{B}}) = -2y_i \ell n \pi\left(\underline{\mathbf{B}}^{\mathrm{T}} \underline{\mathbf{x}}_i\right) - 2(1-y_i) \ell n\left[1 - \pi\left(\underline{\mathbf{B}}^{\mathrm{T}} \underline{\mathbf{x}}_i\right)\right] \quad . \qquad (2.45)$$

O estimador de máxima verossimilhança de $\underline{\mathbf{B}}$ é:

$$\hat{\underline{\mathbf{B}}} = \arg\min_{\underline{\mathbf{B}}} \sum_{i=1}^{n} d(y_i, \underline{\mathbf{x}}_i, \underline{\mathbf{B}}) \quad . \qquad (2.46)$$

O estimador acima deve satisfazer às condições de primeira ordem, dadas por:

$$\sum_{i=1}^{n} \frac{\left[y_i - \pi\left(\hat{\underline{\mathbf{B}}}^{\mathrm{T}} \underline{\mathbf{x}}_i \right) \right] \pi'\left(\hat{\underline{\mathbf{B}}}^{\mathrm{T}} \underline{\mathbf{x}}_i \right)}{\pi\left(\underline{\mathbf{B}}^{\mathrm{T}} \underline{\mathbf{x}}_i \right)\left[1 - \pi\left(\underline{\mathbf{B}}^{\mathrm{T}} \underline{\mathbf{x}}_i \right) \right]} \underline{\mathbf{x}}_i = 0 \quad . \tag{2.47}$$

Os estimadores robustos mais utilizados são dos tipos S (Schweppe) e M (Mallows). Os primeiros são definidos como soluções do sistema de equações

$$\begin{aligned} &\sum_{i=1}^{n} \psi_b \left[r\left(y_i, \underline{\mathbf{x}}_i, \hat{\underline{\mathbf{B}}}, \hat{V} \right) \left(\underline{\mathbf{x}}_i^{\mathrm{T}} \hat{V}^{-1} \underline{\mathbf{x}}_i \right)^{1/2} \right] \left(\underline{\mathbf{x}}_i^{\mathrm{T}} \hat{V}^{-1} \underline{\mathbf{x}}_i \right)^{-1/2} \underline{\mathbf{x}}_i = 0 \\ &\frac{1}{n} \sum_{i=1}^{n} \nu \left[\hat{\underline{\mathbf{B}}}^{\mathrm{T}} \underline{\mathbf{x}}_i \, , b\left(\underline{\mathbf{x}}_i^{\mathrm{T}} \hat{V}^{-1} \underline{\mathbf{x}}_i \right)^{-1/2} \right] \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^{\mathrm{T}} = \hat{V} \end{aligned} \quad , \tag{2.48}$$

onde $r\left(y_i, \underline{\mathbf{x}}_i, \hat{\underline{\mathbf{B}}}, \hat{V} \right) = y_i - \pi\left(\underline{\mathbf{B}}^{\mathrm{T}} \underline{\mathbf{x}}_i \right) - c\left(\hat{\underline{\mathbf{B}}}^{\mathrm{T}} \underline{\mathbf{x}}_i, b\left(\underline{\mathbf{x}}_i^{\mathrm{T}} \hat{V}^{-1} \underline{\mathbf{x}}_i \right)^{-1/2} \right)$ e $c(t\ ,\ b)$ é uma função de correção do viés. Além disso, $\nu(t,b) = \psi_b^2 \left[1 - \pi(t) - c(t,b) \right] \pi(t) + \psi_b^2 \left[-\pi(t) - c(t,b) \right] \left[1 - \pi(t) \right]$. De modo geral, escolhe-se ψ_b como a Função de Hubert, $\psi_b(t) = (-b) \vee (t \wedge b)$, para a qual:

$$c(t,b) = \begin{cases} b\pi(t) / \left[1 - \pi(t) \right] - \pi(t) \, , & t < 0 \, , b < 1 - \pi(t) \\ 1 - \pi(t) - b\left[1 - \pi(t) \right] / \pi(t) \, , & 0 < t \, , b < \pi(t) \\ 0 \, , & \textit{outro caso.} \end{cases}$$

Os estimadores tipo-M são definidos como soluções da equação dada por:

$$\sum_{i=1}^{n} \omega\left(\underline{\mathbf{x}}_i ; \hat{\eta} \right) \psi_b \left[y_i - \pi\left(\underline{\mathbf{B}}^{\mathrm{T}} \underline{\mathbf{x}}_i \right) - c\left(\hat{\underline{\mathbf{B}}}^{\mathrm{T}} \underline{\mathbf{x}}_i, b \right) \right] \underline{\mathbf{x}}_i = 0 \quad , \tag{2.49}$$

onde $\hat{\eta}$ é um vetor de parâmetros de perturbação, que podem ser medidas de posição ou de dispersão das covariáveis. Conforme Gervini (2005), a principal diferença entre os estimadores Tipo-S e Tipo-M é que neste último os pesos para as covariáveis e os resíduos são estimados independentemente, o que não se verifica para o Tipo-S.

Uma outra abordagem ao problema da estimação de parâmetros no caso de separação completa de grupos é proposta por Heinze e Schemper (2002). Os autores apresentam uma metodologia baseada no Escore Modificado de Firth, para estimação de máxima verossimilhança penalizada, cujo objetivo original é reduzir o viés dos estimadores. No Método da Máxima Verossimilhança, os estimadores para os parâmetros são as soluções da *função escore*:

$$\frac{\partial L(\underline{\mathbf{B}})}{\partial \underline{\mathbf{B}}} = U(\underline{\mathbf{B}}) = 0\,. \tag{2.50}$$

Na abordagem apresentada, a função acima é substituída por uma *Função Escore Modificada*, dada por:

$$U^{*}(\underline{\mathbf{B}}) = U(\underline{\mathbf{B}}) + \frac{1}{2} tr\left[\left(I(\underline{\mathbf{B}})\right)^{-1} \left\{ \frac{\partial I(\underline{\mathbf{B}})}{\partial \underline{\mathbf{B}}} \right\} \right] = 0 \quad . \tag{2.51}$$

Para modelos logísticos com resposta binária a Função Escore Modificada é dada por:

$$U^{*}(\underline{\mathbf{B}}) = \sum_{i=1}^{n} \{ y_i - \pi_i + h_i (1 - \pi_i) \} \underline{\mathbf{x}}_i = 0 \quad , \tag{2.52}$$

onde h_i é o elemento da i-ésima diagonal da matriz dada por:

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} \left(\mathbf{X}^{\mathrm{T}} \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{W}^{1/2} \tag{2.53}$$

e

$$\mathbf{W} = diag\left[\pi_i (1 - \pi_i) \right] \quad .$$

No mesmo trabalho, Heinze e Schemper (2002) apresentam um método alternativo para obter os estimadores. Os autores propõem que cada observação (y_i , $\underline{\mathbf{x}}_i$) seja transformada em duas observações, uma com resposta $\tilde{y}_i$ e outra com resposta $(1 - \tilde{y}_i)$, com pesos $(1 + h_i / 2)$ e $(h_i / 2)$, respectivamente. A contribuição das novas observações para a função escore é a mesma obtida com a abordagem baseada em (2.52). Os autores alertam que o método, embora apresente convergência, não é imune a problemas como multicolinearidade, por exemplo.

2.4.3.1 Modelo de Regressão Logística Oculto para Variável Resposta Dicotômica

De acordo com Rousseeuw e Christmann (2003), este modelo foi usado sob outra denominação por Copas (1988), utilizando uma abordagem distinta. Inicialmente os autores consideram uma situação na qual são possíveis apenas dois resultados para a variável resposta, "sucesso" (*s*) e "insucesso" (*f*). Assumem também que o verdadeiro estado *T* é não observável, ao contrário da variável *Y*, que é relacionada a *T*, conforme a Figura 2.2.

O estudo é baseado no seguinte raciocínio: Se o verdadeiro estado é $T = s$, observa-se $Y = 1$ com probabilidade $P(Y = 1 \mid T = s) = \delta_1$. Deste modo a probabilidade de má classificação é dada por $P(Y = 0 \mid T = s) = 1 - \delta_1$. Analogamente, se $T = f$, observa-se $Y = 0$ com probabilidade dada por $P(Y = 0 \mid T = f) = 1 - \delta_0$, com probabilidade de má classificação dada por $P(Y = 1 \mid T = f) = \delta_0$. Os autores assumem que a probabilidade de se observar o verdadeiro estado é superior a 50%, ou seja, $0 < \delta_0 < 0{,}5 < \delta_1 < 1$. Convém lembrar que no Modelo de Regressão Logística Clássico assume-se que $\delta_0 = 0$ e $\delta_1 = 1$.

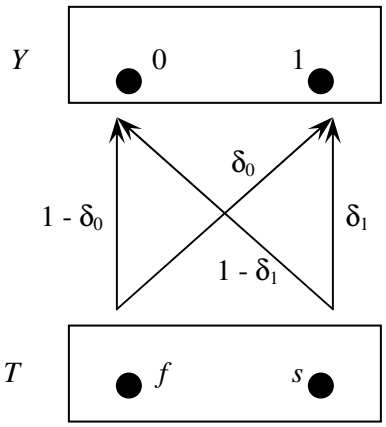

Figura 2.2 – Verdadeiro *T* não observável e resposta Y observável.

Na formulação do modelo considera-se que há *n* variáveis aleatórias independentes e não observáveis resultantes de um modelo de Regressão Logística, com $k = 2$ grupos. Então T_i tem distribuição de Bernoulli, com probabilidade de "sucesso" dada por $\Lambda(\underline{\mathbf{B}}^T \underline{\mathbf{x}}_i)$, $i = 1, 2, \ldots, n$, que é a função sigmóide e onde $\underline{\mathbf{B}}$ é um vetor de parâmetros finitos. A idéia é ilustrada na Figura 2.2.

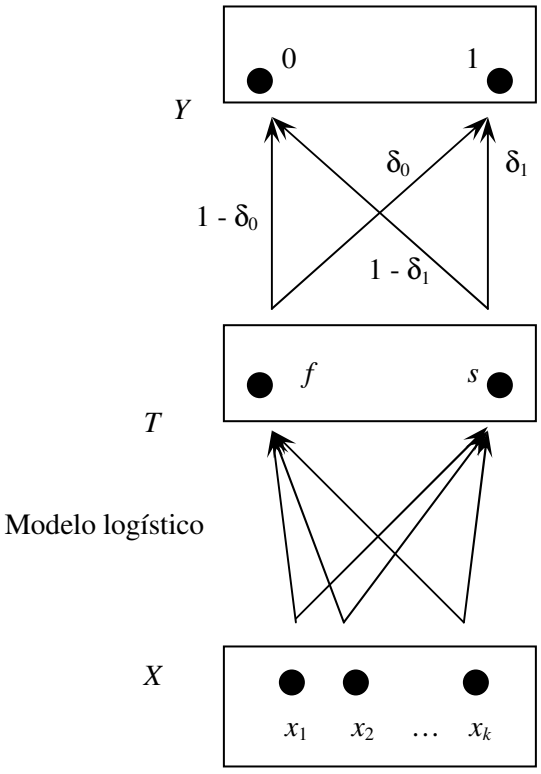

Figura 2.3 – Modelo de Regressão Logística Oculto.

O algoritmo ilustrado é chamado Modelo de Regressão Logística Oculto, já que o verdadeiro *status* T_i está oculto pela estrutura estocástica na parte superior da representação. Conforme os autores, este modelo pode ser interpretado como um tipo de rede neural com uma camada oculta correspondente à variável latente *T*.

Para obter os estimadores dos parâmetros do modelo de Regressão Logística os autores assumem que tanto *Y* como *T* têm distribuição de Bernoulli. Desta forma o estimador de máxima verossimilhança de *T*, dado $Y = y$, é:

$$\begin{aligned} \hat{T}_{ML}(Y=0) &= f \\ \hat{T}_{ML}(Y=1) &= s \end{aligned} \quad . \tag{2.54}$$

A probabilidade condicional de *Y* dado *T* é dada por:

$$\begin{aligned} P\left(Y=1 \mid \hat{T}_{ML}\right) &= \delta_0 \quad \text{se} \quad y=0 \\ P\left(Y=1 \mid \hat{T}_{ML}\right) &= \delta_1 \quad \text{se} \quad y=1 \end{aligned} \quad . \tag{2.55}$$

onde y é o valor observado de Y. Denotando (2.55) por $\tilde{Y}$ tem-se:

$$\tilde{Y} = (1-Y)\delta_0 + Y\,\delta_1$$

que, para n observações y_1 , y_2 , ... , y_n , fica:

$$\tilde{y}_i = (1-y_i)\delta_0 + y_i\delta_1 \quad . \tag{2.56}$$

Convém notar que no modelo clássico assume-se $\tilde{y}_i = y_i$, isto é $\delta_0 = 0$ e $\delta_1 = 1$.

Para ajustar um modelo de Regressão Logística às pseudo-observações $\tilde{y}_i$ utiliza-se o Método da Máxima Verossimilhança, a fim de maximizar a função de verossimilhança, que neste caso é dada por:

$$\ell(\underline{\theta} \mid \tilde{y}_1,\ldots,\tilde{y}_i) = \prod_{i=1}^{n} \left[\Lambda\left(\underline{\theta}^T \underline{\mathbf{x}}_i\right)\right]^{\tilde{y}_i} \left[1-\Lambda\left(\underline{\theta}^T \underline{\mathbf{x}}_i\right)\right]^{1-\tilde{y}_i} \quad . \tag{2.57}$$

A expressão (2.57) é chamada de Verossimilhança Estimada, pois não se conhece a verdadeira verossimilhança, que depende dos valores não observados t_1 , ... , t_n . Os valores que maximizam (2.57) são denominados Estimadores de Máxima Verossimilhança Estimada, e de acordo com os autores sempre existem e são finitos, ao contrário dos estimadores de máxima verossimilhança, conforme já foi exposto. Esta garantia de existência baseia-se inicialmente no fato de que os valores de $\tilde{y}_i$ pertencem ao intervalo (0 , 1).

O logaritmo da Função Verossimilhança (2.57) é:

$$L(\underline{\theta} \mid \tilde{y}_1,\ldots,\tilde{y}_n) = \sum_{i=1}^{n} \tilde{y}_i \ell n\left(\Lambda\left(\underline{\theta}^T \underline{\mathbf{x}}_i\right)\right) + (1-\tilde{y}_i)\ell n\left(1-\Lambda\left(\underline{\theta}^T \underline{\mathbf{x}}_i\right)\right) \quad . \tag{2.58}$$

A expressão (2.58) sempre existe quando $\underline{\theta}$ é finito. As derivadas parciais em relação a $\underline{\theta}$ resultam na função escore *p*-variada dada por:

$$S(\underline{\theta} \mid \tilde{y}_1,...,\tilde{y}_n) = \sum_{i=1}^{n} \left(\tilde{y}_i - \Lambda\left(\underline{\theta}^T \underline{\mathbf{x}}_i\right)\right)\underline{\mathbf{x}}_i \quad . \tag{2.59}$$

Após igualar a zero as expressões (2.59), obtém-se o sistema de equações cuja resolução fornece os estimadores procurados.

Uma propriedade do Estimador de Máxima Verossimilhança Estimada garante a sua existência, sempre que $0 < \delta_o < \delta_1 < 1$ e a matriz de dados tem posto $(p + 1)$. A prova para esta propriedade passa pelo fato de que o Hessiano de (2.58), dado por:

$$\frac{\partial}{\partial \underline{\theta}} S(\underline{\theta}) = -\sum_{i=1}^{n} \Lambda\left(\underline{\theta}^T \underline{\mathbf{x}}_i\right)\left(1 - \Lambda\left(\underline{\theta}^T \underline{\mathbf{x}}_i\right)\right)\underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T \ , \tag{2.60}$$

é uma matriz negativa definida, pois a matriz de dados tem posto $(p + 1)$. Além disso, a função (2.58) é estritamente côncava.

2.4.3.2 Escolha de δ_0 e δ_1

O problema continua com a escolha de valores adequados para δ_0 e δ_1. Citando Copas (1988), o autor relata que a estimação de δ_0 e δ_1 pode ser extremamente difícil, quando não impossível, a menos que *n* seja suficientemente grande. A abordagem simétrica usada por Copas (1988) consiste em escolher uma constante $\gamma > 0$ e fixar $\delta_0 = \gamma$ e $\delta_1 = 1 - \gamma$. A implementação computacional exige que γ seja suficientemente pequeno, de modo que os termos em γ^2 possam ser ignorados.

Finalmente, os autores chamam a atenção para o fato de que o método da máxima verossimilhança estimada não é suficientemente robusto para valores observados que destoam da região de maior concentração dos dados. A fim de aumentar a robustez do método sugere-se o uso de Estimadores de Máxima Verossimilhança Estimada Ponderados, definidos como a solução para a equação,

$$\sum_{i=1}^{n}\left(\tilde{y}_i - \Lambda\left(\underline{\theta}^T \underline{\mathbf{x}}_i\right)\right) w_i x_i = 0 \quad . \tag{2.61}$$

Os pesos w_i, que dependem apenas da distância de x_i até a região de maior concentração dos dados, são definidos por:

$$w_i = \frac{M}{\max\left\{RD^2\left(x_i^*\right), M\right\}} \quad , \tag{2.62}$$

onde, $RD(x_i^*)$ é a Distância Robusta de x_i e M é o 75º percentil de todos os valores RD^2 (x_i^*) , indicando que 25% dos pontos mais afastados recebem pesos inferiores a 1. Os valores em questão são obtidos através das expressões

$$RD\left(x_{ij}\right) = \sqrt{\left(x_{ij} - \bar{x}_j\right)^T S_j^{-1}\left(x_{ij} - \bar{x}_j\right)} \quad . \tag{2.63}$$

O peso para x_i é 1 se $RD\left(x_{ij}\right) \leq \sqrt{\chi^2_{p,0.975}}$ e 0, caso contrário.

Como estimadores da Distância Robusta sugere-se o Estimador Determinante de Covariância Mínima apresentado por Rousseeuw (1984), utilizando o algoritmo de Rousseeuw e Van Driessen (1996). Também é sugerido o algoritmo de Hubert, Rousseeuw e Verboven (2002) para Componentes Principais Robustos. Uma referência adicional é o trabalho de Hubert e Van Driessen (2004) sobre técnicas robustas aplicadas à Análise Discriminante, no qual as autoras estudam a obtenção de estimadores robustos para as Funções Discriminante Quadrática e Discriminante Linear de Fisher.

2.4.4 Análise de Componentes Principais Aplicada à Estimação de Parâmetros

Além da configuração dos dados, outros fatores podem afetar a obtenção de estimativas dos parâmetros. A existência de multicolinearidade entre as variáveis independentes, isto é, a existência de forte dependência entre as mesmas, pode ter efeitos sobre a precisão dos estimadores. Outro elemento a ser considerado é a existência de grande número de variáveis independentes, o que pode

exigir maior esforço computacional. Uma terceira questão envolve o tamanho da amostra e o seu efeito sobre o viés dos estimadores. Uma abordagem apresentada por Aguilera, Escabias e Valderrama (2006) consiste em utilizar a Análise de Componentes Principais (ACP) para reduzir o tamanho do conjunto de dados e a influência da multicolinearidade em problemas com variável resposta binária, tratando das duas primeiras questões. O passo seguinte é o ajuste de um modelo logístico às componentes principais, que substituem as variáveis originais. Neste sentido, o MRLCP pode ser interpretado como um método de substituição de variáveis, no qual as componentes principais substituem as variáveis originais.

Os autores citados apresentaram o Modelo de Regressão Logística de Componentes Principais (MRLCP), como uma extensão do Modelo de Regressão de Componentes Principais apresentado por Massy (1965) para o caso linear. Para atingir seu objetivo, os autores utilizam como covariáveis um conjunto de $s < p$ componentes principais das variáveis independentes, de modo a reduzir o tamanho do conjunto de dados originais. Também são propostos dois métodos para resolver o problema da escolha das componentes principais ótimas que devem ser incluídas no modelo. O primeiro inclui as componentes principais na ordem natural, dada pelas respectivas variâncias explicadas. O segundo método consiste em selecionar as componentes principais mediante um teste de razão de verossimilhanças. No mesmo trabalho são comparados os estimadores obtidos pelo método em questão com aqueles obtidos pelo Método de Mínimos Quadrados Parciais.

O primeiro passo na abordagem proposta é a escolha de um indicador de existência de multicolinearidade. Conforme Aguilera, Escabias e Valderrama (2006), se as variáveis independentes são todas contínuas, pode-se utilizar como indicador o coeficiente de correlação. O problema torna-se mais sério quando o modelo envolve variáveis não contínuas. Uma possível escolha para indicar a ocorrência de multicolinearidade pode ser alguma medida de concordância. De acordo com Hosmer e Lemeshow (1989), grandes valores para o erro padrão também podem servir como um alerta para o problema.

2.4.4.1 Formulação do Modelo de Regressão Logística de Componentes Principais

Sejam dois grupos de observações, G_1 e G_2 , associados à matriz de dados **X**, dada por:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{p1} \\ x_{12} & x_{22} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix},$$

e cuja matriz de covariâncias é:

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

onde

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \left(x_{ki} - \bar{x}_i \right) \left(x_{kj} - \bar{x}_j \right)$$

De acordo com Aguilera, Escabias e Valderrama (2006), pode-se considerar, sem perda de generalidade, que as observações estão centradas, isto é, apresentam médias iguais a zero. A matriz de covariância pode ser escrita como:

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

Seja **P** a matriz quadrada de ordem *p*, cujas colunas são os autovetores da matriz de covariâncias, associados aos autovalores $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, que são as variâncias das componentes principais correspondentes $Y_i = \underline{\mathbf{e}}_i^{\mathrm{T}} \underline{\mathbf{x}}$, i = 1,2, ..., p. Convém lembrar que a matriz de covariância **S**, simétrica e positiva definida, pode ser decomposta de acordo com o Teorema da Decomposição Espectral na forma $\mathbf{S} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$, onde $\boldsymbol{\Lambda} = diag(\lambda_1 , \dots , \lambda_p)$ contém os autovalores de **S** e a matriz ortogonal **P** contém os respectivos autovetores. Seja **Z** a matriz cujas colunas são as componentes principais previamente definidas, então **Z = XP**.

A definição do MRLCP começa pela formulação do modelo logístico em termos das componentes principais associadas à matriz de dados **X**, considerando que todas as variáveis são

contínuas e normalizadas. O Modelo de Regressão Logística para variável resposta dicotômica pode ser escrito na forma:

$$\pi_i = P(Y=1 \mid \underline{\mathbf{x}}_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right)}{1+\exp\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right)} \quad .$$

Considerando que $\mathbf{X} = \mathbf{ZP}^{\mathrm{T}}$, e substituindo $x_{ij} = \sum_{k=1}^{p} z_{ik} v_{jk}$ no modelo acima, o MRLCP para variável resposta binária pode ser escrito como:

$$\pi_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik} v_{jk}\beta_j\right)}{1+\exp\left(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik} v_{jk}\beta_j\right)} = \frac{\exp\left(\beta_0 + \sum_{k=1}^{p} z_{ik}\gamma_k\right)}{1+\exp\left(\beta_0 + \sum_{k=1}^{p} z_{ik}\gamma_k\right)} \quad , \qquad (2.64)$$

onde z_{ik} ($i = 1, \ldots, n$; $k = 1, \ldots, p$) são os elementos da matriz de componentes principais $\mathbf{Z} = \mathbf{XP}$, e $\gamma_k = \sum_{j=1}^{p} v_{jk}\beta_j$, $k = 1, \ldots, p$.

O modelo logístico também é considerado na sua forma matricial, em termos das transformações *logit* e das componentes principais, isto é:

$$\mathbf{L} = \mathbf{X}\underline{\mathbf{B}} = \mathbf{ZP}^{\mathrm{T}}\underline{\mathbf{B}} \quad , \qquad (2.65)$$

onde $\mathbf{Z}$ e $\mathbf{P}$ podem ser particionadas como:

$$\mathbf{Z} = \left(\begin{array}{cccc|cccc} 1 & z_{11} & \cdots & z_{1s} & z_{1s+1} & z_{1s+2} & \cdots & z_{1p} \\ 1 & z_{21} & \cdots & z_{2s} & z_{2s+1} & z_{2s+2} & \cdots & z_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & z_{n1} & \cdots & z_{ns} & z_{ns+1} & z_{ns+2} & \cdots & z_{np} \end{array}\right) = \left(\mathbf{Z}_{(s)} \middle| \mathbf{Z}_{(r)}\right)$$

e

$$\mathbf{P} = \left(\begin{array}{cccc|cccc} 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & v_{11} & \ldots & v_{1s} & v_{1s+1} & v_{1s+2} & \ldots & v_{1p} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & v_{p1} & \ldots & v_{ps} & v_{ps+1} & \ldots & \ldots & v_{pp} \end{array}\right) = \left(\mathbf{P}_{(s)} | \mathbf{P}_{(r)}\right)$$

Então $\mathbf{Z}_{(s)} = \mathbf{X}\mathbf{P}_{(s)}$ e $\mathbf{Z}_{(r)} = \mathbf{X}\mathbf{P}_{(r)}$, tal que:

$$\underline{\mathbf{B}} = \mathbf{P}_{(s)}\gamma_{(s)} + \mathbf{P}_{(r)}\gamma_{(r)} \quad .$$

Levando em consideração que o modelo (2.65) pode ser decomposto na forma:

$$\mathbf{L} = \mathbf{Z}_{(s)}\gamma_{(s)} + \mathbf{Z}_{(r)}\gamma_{(r)} \quad ,$$

o MRLCP, em termos das componentes principais, é obtido pela remoção na equação acima das últimas *r* componentes principais, ou seja:

$$\pi_{i(s)} = \frac{\exp\left(\gamma_0 + \sum_{j=1}^{s} z_{ij}\gamma_j\right)}{1 + \exp\left(\gamma_0 + \sum_{j=1}^{s} z_{ij}\gamma_j\right)} \quad , \quad i = 1 \, , \ldots , n. \qquad (2.66)$$

Os estimadores de máxima verossimilhança do MRLCP são:

$$\hat{\beta}_{(s)} = \mathbf{V}_{(s)}\hat{\gamma}_{(s)} \quad . \qquad (2.67)$$

A formulação apresentada utiliza as *s* primeiras componentes principais. Contudo, Aguilera, Escabias e Valderrama (2006), alertam que as componentes principais com as maiores variâncias não são necessariamente as mais eficientes para a predição, pois componentes com pequenas variâncias podem ser altamente correlacionadas com a variável resposta, razão pela qual devem ser consideradas como possíveis variáveis explicativas em um modelo otimizado.

2.5 VIÉS DOS ESTIMADORES

De acordo com autores como Anderson e Richardson (1979) e McLachlan (1992), por exemplo, o tamanho da amostra exerce grande influência sobre o viés dos Estimadores de Máxima Verossimilhança. Também de acordo com os autores citados, observa-se um aumento do viés quando os estimadores são obtidos a partir de amostras de tamanho reduzido. Uma explanação detalhada a respeito da redução do viés de Estimadores de Máxima Verossimilhança pode ser encontrada em Firth (1993). Uma comparação de métodos de estimação de parâmetros para o Modelo de Regressão Logística, em problemas com separação de grupos, pode ser encontrada em Heinze (2006). Neste trabalho o autor propõe a utilização de Estimadores de Máxima Verossimilhança Penalizada para reduzir o viés dos estimadores. O problema também é tratado por Bull, Mak e Greenwood (2002). Neste caso as autoras apresentam uma abordagem para redução do viés dos Estimadores de Máxima Verossimilhança em modelos da família exponencial para o Modelo de Regressão Logística Multinomial. A abordagem é testada em dois conjuntos de dados e, de acordo com as autoras, os estimadores de escores modificados para os modelos com resposta binária e politômica, neste caso com três grupos, têm viés médio próximo de zero e erro quadrático médio menor que aquele apresentado por outras abordagens. Em outro trabalho, Bull, Greenwood e Hauck (1997) utilizam o Método *Jackknife* para redução do viés em Modelos de Regressão Logística Politômica.

Conforme Anderson e Richardson (1979), o valor esperado para o Estimador de Máxima Verossimilhança $\hat{\underline{\mathbf{B}}}$ é dado por:

$$E\left(\hat{\underline{\mathbf{B}}}\right) = \underline{\mathbf{B}} + b\left(\underline{\mathbf{B}}\right) + \underline{\varepsilon} \qquad (2.68)$$

onde $b^{\mathrm{T}}(\underline{\mathbf{B}}) = [b_1(\underline{\mathbf{B}}) , \ldots , b_p(\underline{\mathbf{B}})]$ é o viés do estimador e $\underline{\varepsilon}$ é um vetor cujas componentes são todas $o(1 / n)$. Além disso,

$$b_t\left(\underline{\mathbf{B}}\right) = \frac{1}{2}\sum_{i,j,k=1}^{p} I^{it} I^{jk} \left\{ 2E\left(\frac{\partial\ \ell nL}{\partial \underline{\mathbf{B}}_j} \frac{\partial^2 \ell nL}{\partial \underline{\mathbf{B}}_i \partial \underline{\mathbf{B}}_k} \right) + E\left(\frac{\partial^3 \ell nL}{\partial \underline{\mathbf{B}}_i \partial \underline{\mathbf{B}}_j \partial \underline{\mathbf{B}}_k} \right) \right\} \quad , \quad (t = 1 , \ldots , p), \qquad (2.69)$$

onde

$$\left(I^{ij}\right)=\left[-\left\{E\left(\frac{\partial^2 \ell nL}{\partial \underline{\mathbf{B}}_i \partial \underline{\mathbf{B}}_j}\right)\right\}\right]^{-1} \quad . \tag{2.70}$$

Então os Estimadores de Máxima Verossimilhança Corrigida são dados por:

$$\tilde{\underline{\mathbf{B}}} = \hat{\underline{\mathbf{B}}} - b\left(\hat{\underline{\mathbf{B}}}\right) \quad . \tag{2.71}$$

Não é difícil perceber que a implementação de um algoritmo para estimar o viés dos estimadores, seguindo o raciocínio que conduz à expressão (2.71), deve levar em conta o esforço computacional exigido para sua execução, especialmente para conjuntos de dados de grandes dimensões.

2.5.1 Aplicações do Método *Bootstrap*

As aplicações do Método *Bootstrap* ao Reconhecimento Estatístico de Padrões incluem a estimação da taxa aparente de erros, que avalia o desempenho do modelo discriminante calculando a proporção de observações classificadas corretamente pelo modelo em questão, e também a estimação de parâmetros, tanto do modelo como do viés dos estimadores encontrados. A primeira aplicação mencionada é sugerida por Efron (1979), inclusive como alternativa à Análise de Lachenbruch, Lachenbruch e Mickey (1968). Uma explanação completa sobre a utilização do Método *Bootstrap* na correção do viés da Taxa Aparente de Erros pode ser encontrada em McLachlan (1992). Outros exemplos de aplicações do método em questão podem ser encontrados nos trabalhos de Jhun e Jeong (2000), para a construção de intervalos de confiança para proporções envolvendo populações multinomiais; e de Aerts e Claeskens (2001), para testar modelos envolvendo conjuntos de dados cujas variáveis não seguem um modelo paramétrico especificado.

O objetivo do Método *Bootstrap* é reamostrar o conjunto de dados para gerar réplicas que possam ser utilizadas na estimação de um parâmetro de interesse. A rigor obtém-se pseudo-réplicas, uma vez que são obtidas da amostra original seguindo um procedimento específico de reamostragem. Na estimação dos parâmetros de modelos de regressão, a aplicação do Método *Bootstrap* segue o raciocínio apresentado na seqüência. Conforme Efron (1979), o modelo geral de regressão geralmente é definido por:

$$Y_i = g_i(\underline{X}, \underline{\beta}) + \varepsilon_i \qquad (2.72)$$

onde $g(\ .\ )$ é uma função conhecida do vetor de parâmetros $\underline{\boldsymbol{\beta}}^{\mathbf{T}} = [\ \beta_1\ ,\ \ldots\ ,\ \beta_p\ ]$ e das variáveis explicativas $\underline{\mathbf{X}}^{\mathrm{T}} = [X_1, X_2, \ldots, X_p]$, enquanto $\varepsilon_i \sim_{\text{ind}} C$, $i = 1\ ,\ \ldots\ ,\ n$.

Normalmente, a informação que se tem a respeito de C é que está centrada em zero, isto é, $E_C(\varepsilon) = 0$ ou $Mediana_C(\varepsilon) = 0$. A partir de uma amostra observada para Y , dado $\underline{\mathbf{X}}$, utiliza-se algum método para estimar $\underline{\boldsymbol{\beta}}$, geralmente o Método dos Mínimos Quadrados, ou seja,

$$\hat{\beta} : \min_{\beta} \sum_{i=1}^{n} \left[y_i - g_i(\underline{x}, \underline{\beta}) \right]^2 \ , \qquad (2.73)$$

com o objetivo de estimar o vetor de parâmetros $\underline{\boldsymbol{\beta}}$ e obter alguma informação sobre a distribuição amostral de $\underline{\boldsymbol{\beta}}$.

A aplicação do Método *Bootstrap* pode ser efetuada pela definição de $\hat{C}$ como distribuição de probabilidade amostral dos resíduos $\hat{\varepsilon}_i$, isto é:

$$\hat{C} : mass\ \frac{1}{n} \quad para \quad \hat{\varepsilon}_i = x_i - g_i(\hat{\beta})\ .$$

Conforme Efron (1979), se alguma componente de $\underline{\boldsymbol{\beta}}$ é um parâmetro de posição para $g(\ .\ )$, então $\hat{C}$ tem média igual a zero. Caso contrário, e se a suposição de que $E_C(\varepsilon) = 0$ é bastante plausível, pode-se modificar $\hat{C}$ de modo a obter a média desejada. As amostras *bootstrap* são obtidas da seguinte forma: da amostra original dos resíduos, ε_1 , ε_2 , ... , ε_n , obtém-se com reposição B amostras de mesmo tamanho denominadas amostras *bootstrap*. Estas amostras de resíduos $\underline{\boldsymbol{\varepsilon}}^*$ são utilizadas para recompor os valores *bootstrap* da resposta,

$$Y_i^* = g_i(\hat{\beta}) + \varepsilon_i^* \quad , \quad i = 1,2, \ldots, n\ . \qquad (2.74)$$

Então, em cada amostra *bootstrap* das respostas, aplica-se o mesmo método de estimação, resultando para cada componente β do vetor de parâmetros o valor dado por:

$$\hat{\beta}^* : \min_{\beta} \sum_{i=1}^{n} \left[y_i^* - g_i(\underline{x}, \underline{\beta}) \right]^2 \quad . \tag{2.75}$$

E $\hat{\beta}^*$ é a estimativa *bootstrap* do parâmetro β. As B amostras *bootstrap* fornecem B estimativas *bootstrap* de β e, assim, pode-se estimar a distribuição amostral do estimador $\hat{\beta}$.

De acordo com Rousseeuw e Christmann (2003), citando Firth (1993), uma vez que o Método da Máxima Verossimilhança tende a superestimar a magnitude dos coeficientes não nulos para amostras reduzidas, uma correção para o viés dos estimadores requer algum tipo de "quebra" com relação a valores próximos de 0 (zero). Segundo os mesmos autores, os estimadores obtidos pelo Método da Máxima Verossimilhança Estimada apresentam esta propriedade, que se verifica na limitação ao intervalo (0 , 1) dos valores atribuídos às pseudo-observações $\tilde{y}_{ij}$.

Para estimar o viés dos estimadores fornecidos pelo MRLO, aplicou-se neste trabalho o Método *Bootstrap* para replicar o conjunto de dados e obter as estimativas *bootstrap*, utilizadas na estimação do viés. A aplicação do método é efetuada de acordo com o algoritmo dado a seguir.

Passo 1: A partir da amostra original, $\mathbf{X} = [\ \underline{\mathbf{x}}_1 , \underline{\mathbf{x}}_2 , \ldots , \underline{\mathbf{x}}_n\ ]$, selecionar B amostras independentes $\underline{\mathbf{X}}^*_1 , \underline{\mathbf{X}}^*_2 , \ldots , \underline{\mathbf{X}}^*_B$, todas de tamanho n, selecionadas com reposição.

Passo 2: Ajustar um modelo logístico a cada uma das amostras geradas no Passo 1, obtendo os vetores $\underline{\beta}^*_1 , \underline{\beta}^*_2 , \ldots , \underline{\beta}^*_B$ de parâmetros.

Passo 3: Os estimadores *bootstrap* são dados por:

$$\hat{\beta}^*_j = \frac{1}{B} \sum_{i=1}^{B} \beta^*_{ji} \quad . \tag{2.76}$$

Passo 4: O viés do estimador $\hat{\beta}_j$ é dado por:

$$b\left(\hat{\beta}_j\right) = \hat{\beta}_j - \hat{\beta}^*_j \quad . \tag{2.77}$$

2.6 FUNÇÃO DISCRIMINANTE LINEAR PARA MAIS DE DOIS GRUPOS

Sejam $k > 2$ grupos G_1, G_2, ... , G_k , com vetores médios $\underline{\mu}_1$, $\underline{\mu}_2$, ... , $\underline{\mu}_k$ e matrizes de covariâncias Σ_1 , Σ_2 , ... , Σ_k , respectivamente. O problema da discriminação linear entre $k > 2$ grupos não exige a suposição de que as populações são normais multivariadas e assume inicialmente a igualdade das matrizes de covariâncias, isto é, $\Sigma_1 = \Sigma_2 = ... = \Sigma_k = \Sigma$. Os estimadores para os vetores médios $\underline{\mu}_i$, $i = 1 , 2 , ... , k$, e para a matriz de covariâncias Σ são, respectivamente:

$$\underline{\bar{\mathbf{x}}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \tag{2.78}$$

e

$$\mathbf{S}_p = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \underline{\bar{\mathbf{x}}}_i\right)\left(x_{ij} - \underline{\bar{\mathbf{x}}}_i\right)^{\mathrm{T}} \tag{2.79}$$

O uso de $\mathbf{S}_p$ é apropriado, pois satisfaz a suposição de igualdade das matrizes de covariâncias. Sejam, também, a matriz *soma de produtos cruzados*,

$$\hat{\mathbf{B}} = \sum_{i=1}^{k} \left(\underline{\bar{\mathbf{x}}}_i - \underline{\bar{\mathbf{x}}}\right)\left(\underline{\bar{\mathbf{x}}}_i - \underline{\bar{\mathbf{x}}}\right)^{\mathrm{T}} \tag{2.80}$$

onde

$$\underline{\bar{\mathbf{x}}} = \frac{\sum_{i=1}^{k} n_i \underline{\bar{\mathbf{x}}}_i}{\sum_{i=1}^{k} n_i} \tag{2.81}$$

e a matriz positiva semidefinida: $\varphi = \mathbf{S}_p + \hat{\mathbf{B}}$. (2.82)

Seja a combinação linear:

$$Y = \underline{\mathbf{c}}^{\mathrm{T}} \underline{\mathbf{X}} \quad . \tag{2.83}$$

Neste caso tem-se, para a i-ésima população:

$$E(Y) = \underline{\mathbf{c}}^{\mathrm{T}} \underline{\mu}_i \tag{2.84}$$

e, para todas as populações:

$$V(Y) = \underline{\mathbf{c}}^{\mathrm{T}} \mathbf{S}_p \underline{\mathbf{c}} + \underline{\mathbf{c}}^{\mathrm{T}} \hat{\mathbf{B}} \underline{\mathbf{c}} \quad , \tag{2.85}$$

decomposta em duas parcelas, a primeira *dentro* e a segunda *entre* os grupos. Uma combinação linear para a qual $\underline{c}^{\mathrm{T}} \varphi \, \underline{c}$ é muito maior que $\underline{c}^{\mathrm{T}} \, \mathbf{S}_{\mathrm{p}} \, \underline{c}$ mostra que a variabilidade dentro dos grupos é inflacionada pelas diferenças na localização. O que se pretende é a maximização da razão:

$$\frac{\underline{\mathbf{c}}^{\mathrm{T}} \left(\mathbf{S}_p + \hat{\mathbf{B}} \right) \underline{\mathbf{c}}}{\underline{\mathbf{c}}^{\mathrm{T}} \mathbf{S}_p \underline{\mathbf{c}}} = 1 + \frac{\underline{\mathbf{c}}^{\mathrm{T}} \hat{\mathbf{B}} \underline{\mathbf{c}}}{\underline{\mathbf{c}}^{\mathrm{T}} \mathbf{S}_p \underline{\mathbf{c}}} \quad . \tag{2.86}$$

Seja a razão:

$$R(\underline{\mathbf{c}}) = \frac{\underline{\mathbf{c}}^{\mathrm{T}} \hat{\mathbf{B}} \underline{\mathbf{c}}}{\underline{\mathbf{c}}^{\mathrm{T}} \mathbf{S}_p \underline{\mathbf{c}}} \tag{2.87}$$

na qual assume-se que $\mathbf{S}_p$ é positiva definida.

De acordo com o teorema da decomposição espectral simultânea de duas matrizes, conforme Flury (1997), existem duas matrizes, uma não singular **H** e outra diagonal $\Lambda = \mathrm{diag}(\lambda_1 , \lambda_2 , \ldots , \lambda_p)$, $\lambda_i \geq 0$, tais que:

$$\mathbf{S}_p = \mathbf{H}\mathbf{H}^{\mathrm{T}} \qquad \text{e} \qquad \hat{\mathbf{B}} = \mathbf{H}\Lambda\mathbf{H}^{\mathrm{T}} \quad . \tag{2.88}$$

O uso do referido teorema é bastante conveniente do ponto de vista da implementação computacional, uma vez que a matriz, por ser simétrica, pode ter seus autovalores e autovetores determinados através do Algoritmo de Jacobi, conforme Kolman (1998). Sendo *m* o posto da matriz $\hat{\mathbf{B}}$, então *m* valores da diagonal de Λ são estritamente positivos, e é possível arranjar as colunas de **H** em ordem decrescente, isto é, $\lambda_1 \geq \ldots \geq \lambda_m > 0 = \lambda_{m+1} = \ldots = \lambda_p$. De acordo com (2.88) pode-se escrever:

$$R(\underline{\mathbf{c}}) = \frac{\underline{\mathbf{c}}^{\mathrm{T}}\mathbf{H}\Lambda\mathbf{H}^{\mathrm{T}}\underline{\mathbf{c}}}{\underline{\mathbf{c}}^{\mathrm{T}}\mathbf{H}\mathbf{H}^{\mathrm{T}}\underline{\mathbf{c}}} \quad . \tag{2.89}$$

A maximização da razão (2.89) pode ficar restrita aos vetores normalizados, isto é, assume-se que $\underline{\mathbf{d}}^{\mathrm{T}}\underline{\mathbf{d}} = 1$. Então o objetivo é maximizar:

$$\underline{\mathbf{d}}^{\mathrm{T}}\Lambda\underline{\mathbf{d}} = \sum_{i=1}^{m}\lambda_i d_i^2 \quad . \tag{2.90}$$

Como $\lambda_1 \geq \lambda_i$, $i \neq 1$, $\underline{\mathbf{d}}^{\mathrm{T}}\Lambda\underline{\mathbf{d}} = \sum_{i=1}^{m}\lambda_i d_i^2 = \sum_{i=1}^{p}\lambda_i d_i^2 \leq \lambda_1\sum_{i=1}^{p} d_i^2 = \lambda_1$. (2.91)

O máximo λ_1 é atingido para $\underline{\mathbf{d}} = \underline{\mathbf{e}}_1 = (1\ ,\ 0\ ,\ \ldots\ ,\ 0)^{\mathrm{T}}$, pois $\underline{\mathbf{e}}_1^{\mathrm{T}}\ \Lambda\ \underline{\mathbf{e}}_1 = \lambda_1$. Isto indica que a razão $R(\underline{\mathbf{c}})$ é maximizada pela escolha de:

$$\underline{\mathbf{c}} = \left(H^{\mathrm{T}}\right)^{-1}\underline{\mathbf{e}}_1 \quad . \tag{2.92}$$

Fazendo:

$$\underline{\mathbf{B}} = (\underline{\beta}_1,\ldots,\underline{\beta}_p) = \left(H^{\mathrm{T}}\right)^{-1} \tag{2.93}$$

tem-se uma combinação linear que maximiza $R(\underline{\mathbf{c}})$, dada por:

$$Y_1 = \underline{\beta}_1^{\mathrm{T}}\,\underline{\mathbf{X}} \quad . \tag{2.94}$$

A combinação linear (2.94), denominada *primeira função discriminante amostral*, fornece a melhor separação entre os k grupos, no sentido de maximizar a variabilidade *entre* os grupos, em relação à variabilidade *dentro* dos grupos.

Em uma segunda etapa o objetivo é maximizar novamente a razão (2.89), agora com uma restrição adicional, a nova combinação linear obtida não pode ser correlacionada com (2.94). Então o problema fica:

$$\begin{aligned} \max \quad & R(\underline{\mathbf{c}}) = \frac{\underline{\mathbf{c}}^{\mathrm{T}} \hat{\mathbf{B}} \underline{\mathbf{c}}}{\underline{\mathbf{c}}^{\mathrm{T}} \mathbf{S}_p \underline{\mathbf{c}}} \quad . \\ \text{s. a} \quad & \underline{\mathbf{c}}^{\mathrm{T}} S_p \underline{\beta}_1 = 0 \end{aligned} \tag{2.95}$$

Assumindo que $\underline{\mathbf{d}}^{\mathrm{T}}\underline{\mathbf{d}} = 1$, a restrição em (2.95) fica:

$$\underline{\mathbf{d}}^{\mathrm{T}} \underline{\mathbf{e}}_1 = 0 \quad . \tag{2.96}$$

Deste modo, o problema (2.96) torna-se:

$$\begin{aligned} max \quad & \underline{\mathbf{d}}^{\mathrm{T}} \Lambda \underline{\mathbf{d}} \\ s.a \quad & \underline{\mathbf{d}}^{\mathrm{T}} \underline{\mathbf{d}} = 1 \quad . \\ & \underline{\mathbf{d}}^{\mathrm{T}} \underline{\mathbf{e}}_1 = 0 \end{aligned} \tag{2.97}$$

Como $\lambda_2 \geq \ldots \geq \lambda_m > 0 = \lambda_{m+1} = \ldots = \lambda_p$, tem-se:

$$\underline{\mathbf{d}}^{\mathrm{T}} \Lambda \underline{\mathbf{d}} = \sum_{i=2}^{m} \lambda_i d_i^2 = \sum_{i=2}^{p} \lambda_i d_i^2 \leq \lambda_2 \sum_{i=2}^{p} d_i^2 = \lambda_2 \quad . \tag{2.98}$$

Agora, o máximo λ_2 é atingido para $\underline{\mathbf{d}} = \underline{\mathbf{e}}_2 = (0\ ,\ 1\ ,\ \ldots\ ,\ 0)^{\mathrm{T}}$, pois $\underline{\mathbf{e}}_2^{\mathrm{T}} \Lambda\ \underline{\mathbf{e}}_2 = \lambda_2$. Então a razão $R(\underline{\mathbf{c}})$ é maximizada pela escolha de:

$$\underline{\mathbf{c}} = \left(H^{\mathrm{T}}\right)^{-1} \underline{\mathbf{e}}_2 = \underline{\mathbf{B}}\underline{\mathbf{e}}_2 = \underline{\beta}_2 \quad . \tag{2.99}$$

A combinação linear:

$$Y_2 = \underline{\beta}_2^{\mathrm{T}} \underline{\mathbf{X}} \tag{2.100}$$

é chamada *segunda função discriminante amostral*. Uma generalização do processo fornece m funções discriminantes amostrais na forma de combinações lineares, isto é:

$$Y_j = \underline{\beta}_j^{\mathrm{T}} \underline{\mathbf{X}} \tag{2.101}$$

onde, $j = 1, 2, \ldots, m$, $m \leq min(p, k-1)$.

A regra de classificação baseada nas funções discriminantes amostrais consiste em classificar uma observação $\underline{X}$ no grupo, ou população, G_j se, para $i \neq k$,

$$\sum_{k=11}^{m} \left[\underline{\beta}_k (\underline{\mathbf{X}} - \bar{\underline{\mathbf{x}}}_k)\right]^2 \leq \sum_{k=1}^{m} \left[\underline{\beta}_k (\underline{\mathbf{X}} - \bar{\underline{\mathbf{x}}}_i)\right]^2 . \tag{2.102}$$

Neste trabalho, a Função Discriminante Linear é abordada com o único propósito de comparar o seu desempenho com o desempenho dos modelos de Regressão Logística abordados. Vale ressaltar que a FDL é objeto de estudo de muitos pesquisadores, servindo como ponto de partida para diferentes abordagens. Uma delas, utilizando a Programação Linear, é apresentada na seqüência a título de ilustração, e também para possibilitar uma comparação com o método apresentado por Santner e Duffy (1986).

2.6.1 Aplicações da Programação Linear à Análise Discriminante Linear

A utilização da Função Discriminante Linear supõe que as matrizes de covariâncias dos grupos analisados são iguais, o que nem sempre ocorre na prática. A violação desta suposição é ponto de partida para duas questões, relativas ao viés da função discriminante e à eficácia da mesma como método de discriminação e classificação. Duas abordagens baseadas na Programação Linear, apresentadas por Freed e Glover (1981) e por Lam e Moy (2003), são apresentadas na seqüência.

Na primeira abordagem, apresentada por Freed e Glover (1981) para dois grupos G_1 e G_2, de pontos A_i, deve-se determinar um vetor $\underline{\mathbf{X}}$ e um valor fronteiriço b tal que, tão próximo quanto possível,

$$\begin{aligned} A_i \underline{\mathbf{X}} &\leq b , \; A_i \in G_1 \\ A_i \underline{\mathbf{X}} &\geq b , \; A_i \in G_2 \end{aligned} \quad . \tag{2.103}$$

Seja a variável α_i , para medir o grau de acordo com o qual os membros do grupo A_i violam a fronteira entre os grupos. Então pode-se inserir uma solução na qual:

$$\begin{aligned} A_i \underline{\mathbf{X}} &\leq b + \alpha_i \ , A_i \in G_1 \\ A_i \underline{\mathbf{X}} &\geq b + \alpha_i \ , A_i \in G_2 \end{aligned} \quad . \tag{2.104}$$

Neste caso deve-se minimizar a soma das violações α_i da fronteira entre grupos, ou uma soma ponderada das violações $h_i \alpha_i$.

Adicionalmente, o hiperplano de separação, $A\underline{\mathbf{X}} = b$, será selecionado de modo que os pontos situados dentro da fronteira estejam tão distantes da mesma quanto possível. Seja d_i a distância do ponto A_i até a fronteira ajustada. Então é possível combinar dois objetivos: minimizar os desvios da fronteira e maximizar a soma ponderada destas distâncias, dada por $\Sigma k_i d_i$. Então o problema pode ser modelado como

$$Min \quad \sum_{i=1}^{n} h_i \alpha_i - \sum_{i=1}^{n} k_i d_i \tag{2.105}$$

$$s.a \quad \begin{cases} A_i \underline{\mathbf{X}} + d_i = b + \alpha_i \quad , \ A_i \in G_1 \\ -A_i \underline{\mathbf{X}} + d_i = -b + \alpha_i \ , A_i \in G_2 \end{cases}$$

No modelo acima as distâncias d_i são as variáveis de folga que transformam as inequações de (2.104) em equações.

O modelo (2.105) fornece uma solução na qual $d_i = 0$ sempre que o peso para minimizar a violação de fronteira exceda o peso para maximizar a distância de A_i até a fronteira ajustada e A_i violar a verdadeira fronteira, ou seja, se $\alpha_i > 0$, então $d_i = 0$, para todo $h_i > k_i$.

De acordo com Lam e Moy (2003), o desempenho de classificação da Programação Linear, especialmente quando a suposição de normalidade é violada, é superior ao da Função Discriminante Linear de Fisher (1936) para propósitos de classificação em muitos estudos experimentais. Os referidos autores apresentam um modelo que minimiza o desvio total dos escores de classificação

de todas as observações com relação às médias amostrais. O modelo de Programação Linear proposto para dois grupos, G_1 e G_2, é brevemente descrito a seguir.

Seja a_{ij} o valor do j–ésimo critério para a i–ésima observação na amostra; w_j o peso do j–ésimo critério; $\bar{a}_j^1$ e $\bar{a}_j^2$ os valores médios do j–ésimo atributo em G_1 e G_2, respectivamente; e sejam também d_i^+ , d_i^- e e_i^- as variáveis de desvio. O modelo proposto é formulado como

$$Min \quad \sum_{i \in G_1} \left(d_i^+ + d_i^- \right) + \sum_{i \in G_2} \left(e_i^+ - e_i^- \right)$$

$$s.a \quad \begin{cases} \sum_{j=1}^{q} \left(a_{ij} - \bar{a}_j^1 \right) w_j + d_i^- - d_i^+ = 0 \, , i \in G_1 \\ \sum_{j=1}^{q} \left(a_{ij} - \bar{a}_j^2 \right) w_j + e_i^- - e_i^+ = 0 \, , i \in G_2 \\ \sum_{j=1}^{q} \left(\bar{a}_j^1 - \bar{a}_j^2 \right) w_j \geq 1 \end{cases} \tag{2.106}$$

onde w_j não tem restrição de sinal, $d_i^+ \geq 0$, $d_i^- \geq 0$, e $e_i^- \geq 0$, para quaisquer i e j. A terceira restrição não apenas evita soluções inaceitáveis como também restringe a diferença entre os escores de classificação, no caso de dois grupos dados por (2.104), a valores iguais ou superiores a um.

$$\sum_{j=1}^{q} \bar{a}_j^1 w_j \quad \text{e} \quad \sum_{j=1}^{q} \bar{a}_j^2 w_j \, . \tag{2.107}$$

Enquanto o desvio entre os grupos é fixado, a função objetivo minimiza o desvio absoluto dentro de cada grupo em relação às suas médias. Em sua essência, segundo Lam e Moy (2003), o modelo acima é similar à Função Discriminante Linear, no sentido de maximizar a razão das variações entre grupos pelas variações dentro dos grupos.

Os valores para w_j obtidos após a resolução do modelo proposto podem ser usados para computar os escores de classificação para todas as observações, dados por

$$s_i = \sum_{i=1}^{q} a_{ij} w_j \tag{2.108}$$

O valor de corte c pode ser determinado também pela resolução do seguinte modelo de Programação Linear.

$$Min \quad \sum_{i \in \{G_1 \cup G_2\}} h_i \tag{2.109}$$

$$s.a \quad \begin{cases} s_i + h_i - c \geq 0 & i \in G_1 \\ s_i - h_i - c \leq 0 & i \in G_2 \end{cases}$$

onde $h_i \geq 0$ e c não tem restrição de sinal. Conhecido o valor de corte, as observações podem ser classificadas pelos seus escores de classificação.

2.7 REDES NEURAIS ARTIFICIAIS

As Redes Neurais Artificiais surgiram, de acordo com Fausett (1994), em meados da década de 50, tomando como ponto de partida as idéias apresentadas por McCulloch e Pitts (1943). A motivação era a necessidade de compreender o funcionamento do cérebro humano e reproduzir algumas de suas características, entre as quais o alto nível de interconexão e paralelismo maciço, isto é, muitos neurônios operando simultaneamente e tolerância a falhas, ou seja, o desempenho não é afetado de forma significativa por algum prejuízo porventura causado a alguns neurônios. A aplicação de Redes Neurais a problemas de reconhecimento de padrões pode ser observada em trabalhos como os de Guimarães e Chaves Neto (2006), que compara o desempenho de um modelo logístico, uma rede neural com algoritmo de retro-propagação e uma Função Discriminante Linear para classificação de padrões com resposta politômica; de Wilson e Sharda (1994), que utiliza Redes Neurais na previsão de falências; Schumacher, Roβner e Vach (1996), que apresenta uma comparação entre Redes Neurais e Regressão Logística, ilustrada pela aplicação das mencionadas técnicas ao diagnóstico de câncer de mama e ao estudo de problemas vasculares, e de Desai, Crook e Overstreet (1996) que compara Redes Neurais, análise discriminante linear e Regressão Logística na construção de modelos de escore de crédito. Uma fonte de referência bastante ilustrativa é o trabalho de Féraud e Clérot (2002), que apresenta uma metodologia para explicar a classificação de padrões através de Redes Neurais Artificiais. Outra fonte de referência para o estudo de

propriedades comuns às Redes Neurais Artificiais e ao Modelo de Regressão logística é o trabalho de Dreiseitl e Ohno-Machado (2002), que também aborda a aplicação das duas técnicas à pesquisa médica.

Uma Rede Neural Artificial consiste de um determinado número de elementos de processamento chamados *neurônios*, dispostos em *camadas*, aos quais são associados *pesos*. O tipo mais simples de rede neural, mostrado na Figura 2.4, é conhecido como *feedforward*, ou *perceptron* logístico, na terminologia de Schumacher *et al*. (1996).

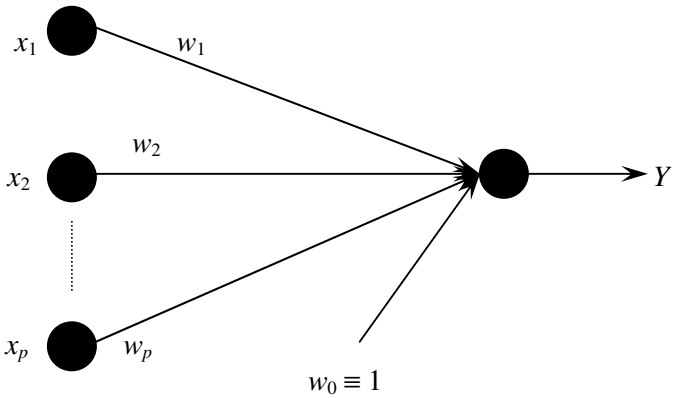

Figura 2.4 – *Perceptron* logístico.

O *perceptron* consiste de $(p + 1)$ unidades de entrada, onde x_0 tem entrada constante igual a 1, e uma unidade de saída. Os valores de entrada x_i são ponderados com pesos w_i , $i = 0, 1, \ldots , p$, e a soma das entradas ponderadas é transformada pela função logística, cujo gráfico é mostrado na Figura 2.5. Então o sinal de saída Y pode ser definido como função dos valores de entrada $\underline{\mathbf{X}}$ e dos pesos $\underline{\mathbf{W}}$, isto é,

$$Y = \frac{e^{\mu}}{1+e^{\mu}} \quad , \quad \mu = w_0 + \sum_{i=1}^{p} w_i x_i \tag{2.116}$$

O cálculo dos valores para os pesos é chamado *treinamento*, ou *aprendizagem*, da rede. O treinamento pode ser *supervisionado*, quando cada vetor do conjunto de entradas é associado a uma resposta e o objetivo é determinar a resposta correta para todos os vetores de entrada, ou *não supervisionado*, quando apenas o conjunto de entrada é fornecido e busca-se extrair propriedades de

acordo com determinadas representações internas. O referido *treinamento* é realizado no sentido de minimizar a *função erro*, definida como:

$$E(\underline{\mathbf{W}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.117}$$

onde

$$\hat{y}_i = f(\underline{\mathbf{X}}_i, \underline{\mathbf{W}}) = \frac{e^{\mu_i}}{1+e^{\mu_i}} \quad , \quad \mu_i = \underline{\mathbf{W}}' \, \underline{\mathbf{X}}_i . \tag{2.118}$$

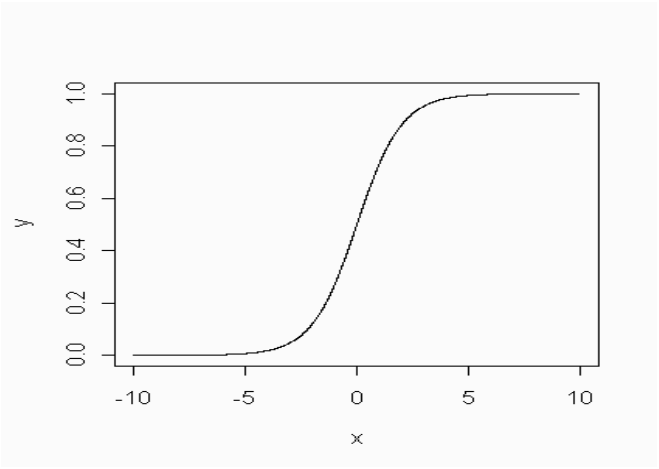

Figura 2.5 – Gráfico da Função Sigmóide, - 10 ≤ x ≤ 10 .

De acordo com Schumacher, Roβner e Vach (1996), a minimização nada mais é que uma aplicação do método dos mínimos quadrados. Neste processo utiliza-se o método conhecido como *retro-propagação* ("*back-propagation*"), que de acordo com os autores citados pode ser denominado como *retro-propagação de mínimos quadrados* (LS-BP), e é definido por:

$$\hat{w}^{(j+1)} = \hat{w}^{(j)} - \alpha \vec{\nabla} \left[E\left(\hat{w}^{(j)} \right) \right] \tag{2.119}$$

onde α é a taxa de aprendizagem. Esta formulação caracteriza (2.119) como um método de busca descendente. O modelo (2.116) envolve a função logística, o que torna o *perceptron* logístico semelhante ao modelo de Regressão Logística. De fato, Schumacher, Roβner e Vach (1996) sugerem que $\underline{\mathbf{W}}$ e $\underline{\mathbf{X}}$ sejam interpretados no mesmo sentido. Uma alternativa para a função de aprendizagem pode ser a Distância de Kullback – Leibler, na forma:

$$E^*(\underline{\mathbf{W}}) = \sum_{i=1}^{n} \left[y_i \ell n \frac{y_i}{\hat{y}_i} + (1 - y_i) \ell n \frac{1 - y_i}{1 - \hat{y}_i} \right] \tag{2.120}$$

onde $\hat{y}_i$ é definido por (2.116). Pela semelhança com o método da máxima verossimilhança, o método de retropropagação usando (2.116) também é conhecido como *retro-propagação de máxima verossimilhança* (ML-BP). Então a equação (2.119) pode ser escrita na forma

$$\hat{w}^{(j+1)} = \hat{w}^{(j)} - \alpha \vec{\nabla} \left[E^* \left(\hat{w}^{(j)} \right) \right] \quad . \tag{2.121}$$

O modelo do *perceptron* logístico pode ser estendido para s, $s > 1$, unidades de saída. A rede neural com tal configuração, cuja arquitetura é ilustrada na Figura 2.6, possui pesos w_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, s$, conectados às unidades de saída Y_j, na forma

$$Y_j = \frac{exp\left(w_{0j} + \sum_{i=1}^{n} w_{ij} x_i \right)}{\sum_{k=1}^{s} exp\left(w_{ok} + \sum_{i=1}^{n} w_{ik} x_i \right)} \quad . \tag{2.122}$$

2.7.1 Redes Neurais com Camadas Ocultas

Uma extensão do *perceptron* logístico consiste em adicionar uma camada com K neurônios, usualmente denominada camada oculta, conforme a Figura 2.7. Esta camada é estabelecida entre as unidades de entrada e de saída. A camada de entrada recebe os sinais, X_i, do conjunto de medidas. Estes sinais, ponderados com pesos v_{ik}, $i = 1, 2, \ldots, p$, $k = 1, 2, \ldots, K$, são enviados aos neurônios da camada oculta. Cada neurônio desta camada calcula sua ativação, w_{kj}, $j = 1, 2, \ldots, s$, e envia o sinal obtido para o neurônio da camada de saída, Y. O neurônio da camada de saída, por sua vez, calcula o seu sinal de ativação e o transforma em uma resposta, $\hat{Y}$, para o padrão fornecido, que é comparada com a resposta Y já conhecida, a fim de determinar o erro associado. Com base neste erro efetua-se uma atualização dos pesos. O processo é repetido para todos os padrões fornecidos, razão pela qual são também denominados *padrões de treinamento*. A atualização dos pesos é efetuada de acordo com a taxa de aprendizagem. A escolha do valor da referida taxa é decisiva para se alcançar um desempenho adequado da rede. Uma taxa constante,

porém de baixo valor, pode dificultar o trabalho de busca por um ponto de mínimo global, enquanto um valor alto pode desestabilizar o algoritmo nas proximidades do referido ponto.

A atualização dos pesos segue as equações:

$$i^{(p)} = \sum_j w_j x_j^{(p)} + w_0 \qquad \hat{Y}^{(p)} = \frac{exp\left(i^{(p)}\right)}{1 + exp\left(i^{(p)}\right)} \qquad (2.123)$$

$$\delta^{(p)} = Y^{(p)} - \hat{Y}^{(p)} \qquad \Delta w_j = \alpha x_j \delta^{(p)} \qquad w_j = w_j + \Delta w_j \qquad (2.124)$$

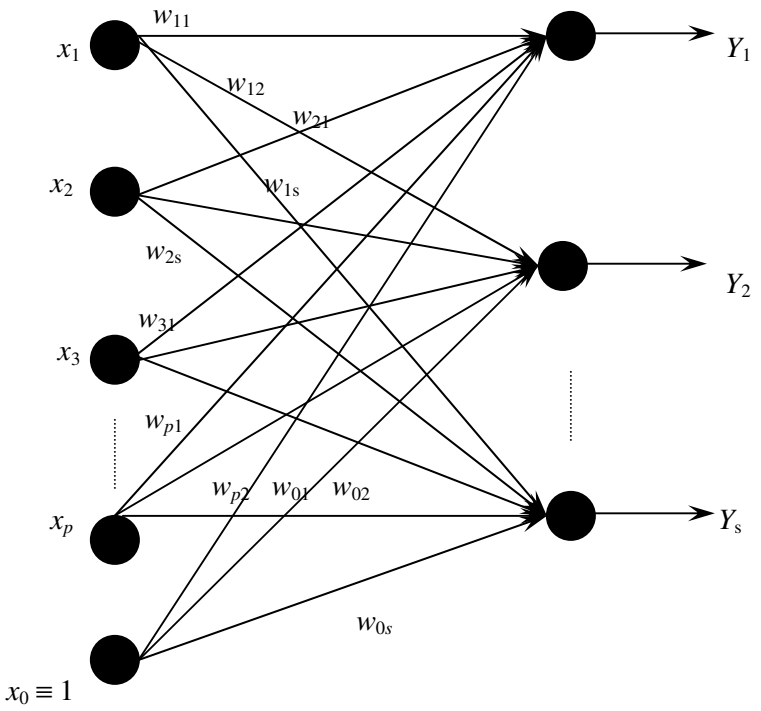

Figura 2.6 – *Perceptron* logístico para variável resposta politômica.

2.7.1.1 Algoritmo de Treinamento

Neste trabalho optou-se pela utilização de uma rede neural dotada de um *perceptron* logístico para variável resposta politômica com uma camada oculta, conforme a Figura 2.7. Esta opção foi motivada apenas pelo fato de ser esta arquitetura uma das mais utilizadas para o reconhecimento de padrões. Também é possível encontrar na literatura corrente um número considerável de trabalhos comparando o desempenho apresentado por redes com esta arquitetura com o desempenho de modelos baseados em Regressão Logística e funções discriminantes lineares, embora a maioria dos trabalhos restrinja-se a problemas com variável resposta dicotômica. O

algoritmo de treinamento segue o raciocínio apresentado por Fausett (1994), que chama a atenção para a importância da forma dos dados na escolha da função apropriada, entre outras questões. A derivada da função logística, na forma 2.116, pode ser escrita na forma:

$$f'(x) = f(x)(1 - f(x)) \tag{2.125}$$

Esta propriedade proporciona maior simplicidade na implementação computacional, já que não exige nenhum procedimento adicional para a avaliação da derivada.

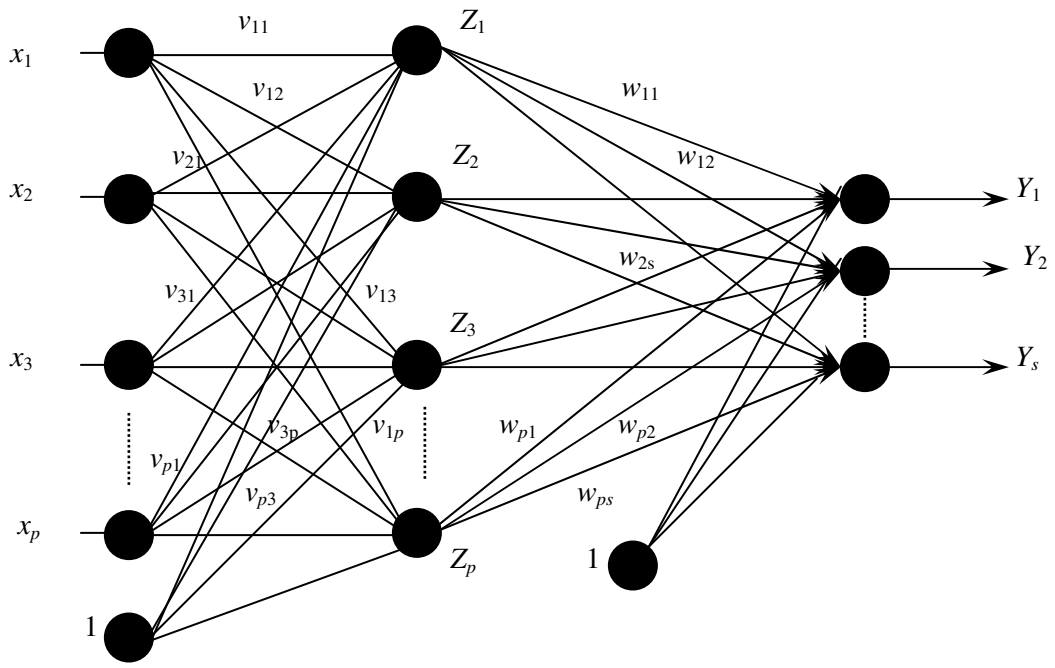

Figura 2.7 – *Perceptron* logístico para variável resposta politômica com uma camada oculta.

O algoritmo de treinamento segue os passos listados a seguir.

Passo 0. Iniciar os pesos.

Passo 1. Enquanto o critério de parada não for atendido, executar os passos 2 a 9.

Passo 2. Para cada par de treinamento (y_i , $\underline{\mathbf{X}}_i$), executar os passos 3 a 8.

Feedforward:

Passo 3. Cada unidade de entrada x_i, i = 1 , ... , p, recebe o sinal de entrada e o distribui para todas as unidades da camada oculta.

Passo 4. Cada unidade oculta z_i, i = 1 , ... , p, efetua a soma ponderada dos sinais de entrada,

$$z\_in_i = v_{0i} + \sum_{j=1}^{p} x_i v_{ij} \tag{2.126}$$

O sinal é computado pela função de ativação, gerando o sinal de saída

$$z_i = f\left(z\_in_i\right)$$

Este sinal é enviado a todas as unidades da camada de saída.

Passo 5. Cada unidade de saída Y_j, $j = 1$, ... , k, soma seus sinais ponderados de entrada,

$$y\_in_j = w_{oj} + \sum_{i=1}^{k} z_i w_{ij} \tag{2.127}$$

O sinal de saída é computado pela função de ativação,

$$y_j = f\left(y\_in_j\right)$$

Retro-propagação do erro:

Passo 6. Cada unidade de saída recebe o valor observado no conjunto de treinamento e computa o erro,

$$\delta_j = \left(yobs_k - yest_k\right) f'\left(y\_in_k\right) \tag{2.128}$$

Após computar o erro acima, cada unidade calcula o termo ponderado de correção

$$\Delta w_{ij} = \alpha \delta_j z_i \tag{2.129}$$

O termo de correção para o viés é dado por:

$$\Delta w_{i0} = \alpha \delta_j$$

Cada termo δ_j é enviado para as unidades da camada oculta.

Passo 7. Cada unidade Z_j soma suas entradas δ_j,

$$\delta\_in_j = \sum_{i=1}^{k} \delta_i w_{ij} \tag{2.130}$$

As entradas são multiplicadas pela derivada da função de ativação para calcular o erro de informação,

$$\delta_j = \delta\_in_j f'(z\_in_j) \tag{2.131}$$

Calcula-se o termo ponderado de correção:

$$\Delta v_{ij} = \alpha \delta_j x_i \tag{2.132}$$

e o termo de correção do viés:

$$\Delta v_{0j} = \alpha \delta_j$$

Atualização dos pesos e do viés

Passo 8. Cada unidade de saída atualiza o viés e seus pesos:

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \Delta w_{ij} \tag{2.133}$$

Cada unidade da camada oculta atualiza o viés e seus pesos:

$$v_{ij}^{(t+1)} = v_{ij}^{(t)} + \Delta v_{ij} \tag{2.134}$$

Passo 9. Testar o critério de parada.

2.7.1.2 Condições Iniciais

A importância da escolha adequada das condições iniciais, tanto para os pesos como para a taxa de aprendizagem, é apontada por autores como Schumacher, Roβner e Vach (1996) e Fausett (1994). De acordo com os primeiros autores, uma taxa de aprendizagem pequena pode diminuir a chance de encontrar um ponto de mínimo, além de dificultar o afastamento de pontos de mínimo locais. Por outro lado, uma taxa de aprendizagem grande tende a aumentar estas chances, mas o algoritmo pode tornar-se instável. Na escolha de valores iniciais para os pesos, no início do procedimento iterativo, Fausett (1994) recomenda que sejam evitados valores que possam zerar tanto a função de ativação como suas derivadas, uma vez que as mesmas são demasiado importantes para a atualização dos pesos durante a aplicação de procedimentos iterativos.

Neste trabalho, a Rede Neural utilizada teve seus pesos iniciados com valores pseudo-aleatórios do intervalo [- 0,5 ; 0,5]. Os valores foram simulados a partir de uma distribuição uniforme. Para efeito de comparação de desempenho, também foram adotados, em outras iterações, valores iniciais iguais a zero.

2.7.2 Vantagens e Desvantagens das Redes Neurais Apontadas na Literatura Disponível

Em que pese a eficiência relatada nas aplicações a problemas de reconhecimento e classificação de padrões, e demonstrada em numerosos artigos publicados na literatura corrente, o uso de Redes Neurais gera desconfiança entre alguns pesquisadores, principalmente pela carência de formalismo matemático. Segundo Faraggi e Simon (1995), o desenvolvimento das Redes Neurais tem sido conduzido em grande parte por pesquisadores não estatísticos, o que explica o pouco uso de técnicas estatísticas. Além deste fato, não há certeza sobre quando e sob quais condições o uso de Redes Neurais é preferível ao uso das técnicas estatísticas multivariadas aqui abordadas. De acordo com Schwarzer, Vach e Schumacher (2000) a aplicação incorreta de Redes Neurais pode levar a problemas como o ajuste de funções implausíveis e estimação viesada ou ineficiente. De acordo com Schumacher, Roβner e Vach (1996), apenas aplicações bem sucedidas são relatadas. Segundo os mesmos autores, citando comunicado do SAS – Institute (1994), muitos tipos de Redes Neurais são meras reinvenções de conhecidos métodos estatísticos, implementadas através de algoritmos ineficientes. Por outro lado, conforme White (1992), citado pelos mesmos autores, as Redes Neurais são atraentes, ao contrário dos métodos estatísticos. Ao discutir e comparar Redes Neurais com modelos de regressão, Warner e Misra (1996) apontam que as Redes

Neurais de múltiplas camadas não impõem nenhuma relação funcional entre variáveis dependentes e independentes. Pelo contrário, a relação funcional é determinada a partir dos dados no processo de determinação dos pesos. Segundo os autores, a vantagem deste processo é a conseqüente habilidade da rede neural para aproximar qualquer função contínua. A desvantagem do mesmo é a dificuldade para interpretar a rede neural obtida. Outra desvantagem apontada pelos autores é a lentidão da convergência para uma solução e sua dependência das condições iniciais. Aspectos dessa natureza são amplamente tratados por Intrator e Intrator (2001), que também apresentam uma metodologia para interpretar os resultados fornecidos por Redes Neurais aplicadas ao Reconhecimento de Padrões.

A literatura disponível está repleta de argumentos, tanto favoráveis como contrários, à utilização de Redes Neurais Artificiais. Esta discussão indica que o assunto é uma área ainda aberta a pesquisas, tanto no campo teórico como no campo das aplicações. O que não se pode ignorar é a utilidade desta técnica, bem como seu potencial, para a resolução de problemas de reconhecimento estatístico de padrões, razão pela qual é abordada neste trabalho. Assim como no caso da Função Discriminante Linear, o que se pretende é utilizar o desempenho da Rede Neural Artificial como referência para avaliar o desempenho dos modelos de Regressão Logística aqui abordados.

3 MODELOS DE REGRESSÃO LOGÍSTICA OCULTO E DE COMPONENTES PRINCIPAIS PARA RECONHECIMENTO E CLASSIFICAÇÃO DE PADRÕES COM VARIÁVEL RESPOSTA POLITÔMICA

Neste trabalho se faz a proposta da extensão do Modelo de Regressão Logística Oculto (MRLO) e do Modelo de Regressão Logística de Componentes Principais (MRLCP) para variável resposta politômica. A extensão do primeiro modelo tem como objetivo inicial verificar se o mesmo mantém, quando aplicado a problemas com variável resposta politômica, a sua principal propriedade, isto é, se o Método da Máxima Verossimilhança Estimada garante a existência dos Estimadores de Máxima Verossimilhança para quaisquer configurações dos conjuntos de dados, da mesma forma verificada para variável resposta dicotômica. O segundo objetivo é comparar o desempenho do modelo em questão com os desempenhos apresentados pelo Modelo de Regressão Logística Clássico (MRLC) e pelos Modelos de Regressão Logística Individualizados (MRLI). Um objetivo adicional é determinar o viés para cada estimador através do Método *Bootstrap*, utilizando o algoritmo apresentado em 2.5.1.

Com relação ao MRLCP, deseja-se investigar a sua eficiência não apenas na estimação de parâmetros, como também a sua contribuição para a eficiência do modelo obtido a partir das componentes principais, comparando o seu desempenho com os desempenhos dos modelos citados anteriormente. Cabe ressaltar que não foi dedicada atenção especial à escolha das componentes principais para compor os modelos obtidos.

3.1 MODELO DE REGRESSÃO LOGÍSTICA OCULTO PARA VARIÁVEL RESPOSTA POLITÔMICA

Para se aplicar o modelo em questão a problemas com variável resposta politômica, considera-se k variáveis não observáveis $T_1, \ldots, T_k$, assumindo os valores f_i ou s_i, $i = 1, \ldots, k$, conforme o raciocínio ilustrado na Figura 3.1, e que segue de perto a proposta de Rousseeuw e Christmann (2003) para variável resposta dicotômica.

Deste modo, se o verdadeiro estado é $T_j = s_j$, observa-se $Y = j$, $j = 1, \ldots, k$, com probabilidade dada por:

$$P\left(Y = j \mid T_j = s_j\right) = \delta s_j . \tag{3.1}$$

Analogamente, se $T_j = f_j$, observa-se $Y = j$,

$$P(Y = j \mid T_j = f_j) = \delta f_j . \tag{3.1}$$

Adicionalmente, tem-se que:

$$P(Y \neq j \mid T_j = s_j) = 1 - \delta s_j \tag{3.2}$$

e

$$P(Y \neq j \mid T_j = f_j) = 1 - \delta f_j \quad . \tag{3.3}$$

Assume-se, também, que $0 < \delta f_j < 0{,}5 < \delta s_j < 1, j = 1 , \ldots , k.$

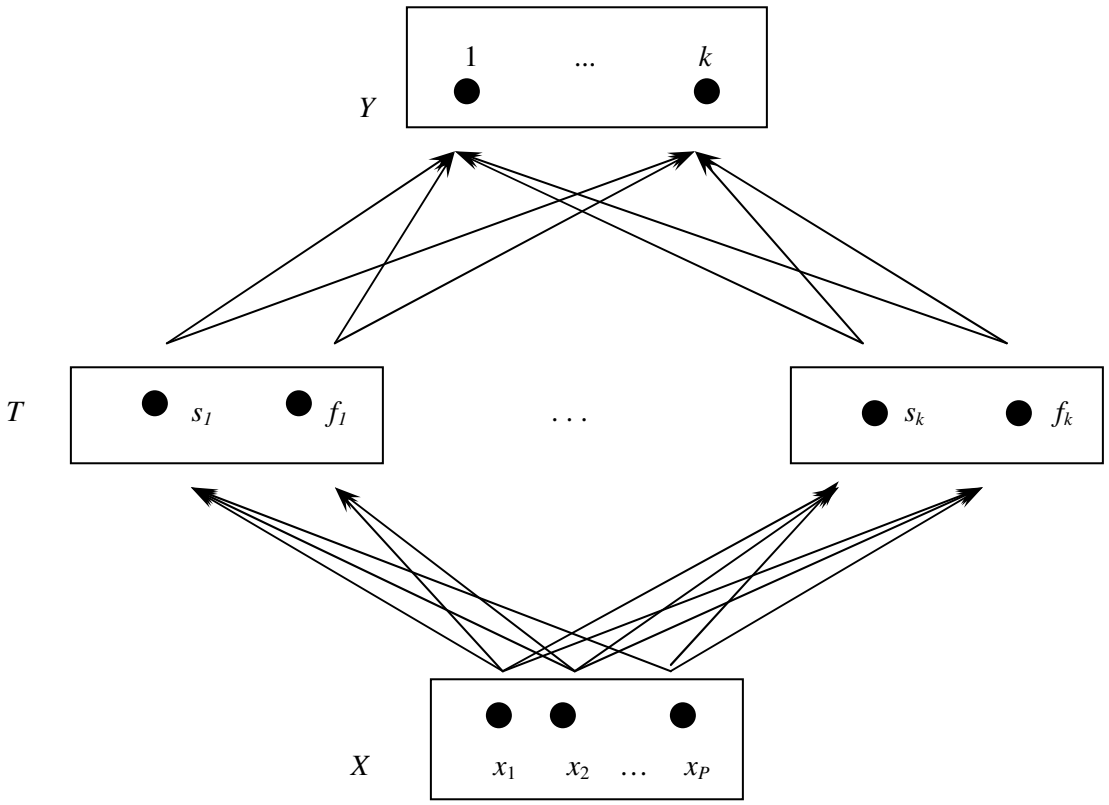

Figura 3.1 – Modelo de Regressão Logística Oculto Proposto para Variável Resposta Politômica.

Então, há $n \times k$ variáveis não observáveis T_{ij} , $i = 1$, ... , n, $j = 1$, ... , k, resultantes de k vetores $\underline{\boldsymbol{\theta}}_1$, ... , $\underline{\boldsymbol{\theta}}_k$, tais que:

$$\underline{\theta}_j = \left[\theta_{j0} \quad \theta_{j1} \quad \ldots \quad \theta_{jp}\right]^{\mathrm{T}}$$

onde $\underline{\boldsymbol{\theta}}_k = \underline{\mathbf{0}}$.

O estimador de máxima verossimilhança de T_j , dado $Y = y$, é dado por:

$$\begin{aligned} \hat{T}_{ML_j}(Y \neq j) &= f_j \\ \hat{T}_{ML_j}(Y = j) &= s_j \end{aligned} \quad . \tag{3.4}$$

A probabilidade condicional é dada então por:

$$\begin{aligned} P\left(Y = j \mid \hat{T}_{ML_j}\right) &= \delta s_j \quad \text{se} \quad y = j \\ P\left(Y = j \mid \hat{T}_{ML_j}\right) &= \delta f_j \quad \text{se} \quad y \neq j \end{aligned} \quad . \tag{3.5}$$

Desta forma pode-se definir a variável:

$$\tilde{y}_{ji} = (1 - y_{ji})\delta f_j + y_{ji}\,\delta s_j \ , \tag{3.6}$$

onde

$$y_{ji} = \begin{cases} 1 & \text{se} \quad Y_i = j \\ 0 & \text{se} \quad Y_i \neq j \end{cases} .$$

Agora o objetivo é ajustar às pseudo-observações $\tilde{y}_{ij}$ um Modelo de Regressão Logística na forma dada por:

$$P(G_s \mid \underline{\mathbf{X}}) = \frac{\exp(\theta_s)}{\sum_{j=1}^{k} \exp(\theta_j)} \qquad (s = 1, 2, \ldots, k) \ ,$$

onde $\theta_s = \theta_{s0} + \theta_{s1}x_1 + \theta_{s2}x_2 + \ldots + \theta_{sp}x_p = \underline{\theta}_s^T \underline{\mathbf{X}} \qquad (s = 1, 2, \ldots, k-1)$, e $\theta_k = \underline{\mathbf{0}}$.

Neste caso a função de verossimilhança estimada é dada por:

$$\ell\left(\underline{\theta} \mid \underline{\tilde{\mathbf{Y}}}\right) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[P(T_j \mid \underline{\mathbf{X}}_i)\right]^{\tilde{y}_{ji}} \quad . \tag{3.7}$$

Extraindo o logaritmo neperiano obtém-se a Função Log-Verossimilhança Estimada dada por:

$$L\left(\underline{\theta} \mid \underline{\tilde{\mathbf{Y}}}\right) = \sum_{i=1}^{n} \left[\sum_{j=1}^{k-1} \tilde{y}_{ji} \theta_j - \ell n \left(1 + \sum_{j=1}^{k-1} \exp \theta_j \right) \right] \tag{3.8}$$

Os estimadores de θ_j são os valores que maximizam (3.8). Neste trabalho utilizou-se o Método da Máxima Verossimilhança, e tem-se a equação:

$$\frac{\partial L(\underline{\theta})}{\partial \theta_{jm}} = \sum_{i=1}^{n} x_{mi} \left[\tilde{y}_{ji} - P\left(G_j \mid \underline{\mathbf{x}}_i\right)\right] \tag{3.9}$$

onde $j = 1 , \ldots , k - 1$ e $m = 0 , 1 , \ldots , p$.

Para verificar que (3.8) é estritamente côncava, basta considerar $\underline{\theta}_1$ e $\underline{\theta}_2$, vetores de um subespaço convexo de R^{p+1}. Diz-se, por definição, que uma função L, definida no mesmo subespaço, é *estritamente convexa*, para quaisquer $\underline{\theta}_1$, $\underline{\theta}_2$, $\underline{\theta}_1 \neq \underline{\theta}_2$, e para qualquer λ, $0 \leq \lambda \leq 1$, se:

$$L(\lambda \underline{\theta}_1 + (1-\lambda)\underline{\theta}_2) < \lambda\, L(\underline{\theta}_1) + (1-\lambda)\, L(\underline{\theta}_2) .$$

Também por definição, diz-se que uma função S, definida no mesmo subespaço referido acima, é *estritamente côncava* se $L = - S$ é estritamente convexa.

Sejam $\underline{\theta}_1$ e $\underline{\theta}_2$, vetores de um subespaço convexo de R^{p+1}, λ tal que $0 \leq \lambda \leq 1$, e a função $S\left(\underline{\theta} \mid \underline{\tilde{\mathbf{Y}}}\right) = -L\left(\underline{\theta} \mid \underline{\tilde{\mathbf{Y}}}\right)$. De (3.8) tem-se que:

$S\left((\lambda \underline{\theta}_1 + (1-\lambda)\underline{\theta}_2) \mid \underline{\tilde{\mathbf{Y}}}\right) =$

$$= -\sum_{i=1}^{n}\left[\sum_{j=1}^{k-1}\tilde{y}_{ji}\,\underline{\theta}_{2j}^{T}\underline{\mathbf{x}}_i + \lambda\,\tilde{y}_{ji}\,\underline{\theta}_{1j}^{T}\underline{\mathbf{x}}_i - \lambda\,\tilde{y}_{ji}\,\underline{\theta}_{2j}^{T}\underline{\mathbf{x}}_i - \ell n\left(1+\sum_{j=1}^{k-1}\exp\left(\lambda\,\underline{\theta}_{1j}^{T}+\underline{\theta}_{2j}^{T}-\lambda\underline{\theta}_{2j}^{T}\right)\underline{\mathbf{x}}_i\right)\right] \quad . \tag{3.10}$$

Da mesma forma:

$$\lambda S\left(\underline{\theta}_1 \mid \tilde{\underline{\mathbf{Y}}}\right)+(1-\lambda)S\left(\underline{\theta}_2 \mid \tilde{\underline{\mathbf{Y}}}\right)=$$

$$= -\sum_{i=1}^{n}\left[\sum_{j=1}^{k-1}\tilde{y}_{ji}\,(\underline{\theta}_{2j}^{T}+\lambda\underline{\theta}_{1j}^{T}-\lambda\underline{\theta}_{2j}^{T})\underline{\mathbf{x}}_i - \ell n\left(1+\sum_{j=1}^{k-1}\exp\left(\lambda\underline{\theta}_{1j}^{T}\underline{\mathbf{x}}_i\right)\right)-\ell n\left(1+\sum_{j=1}^{k-1}\exp\left(\underline{\theta}_{2j}^{T}-\lambda\underline{\theta}_{2j}^{T}\right)\underline{\mathbf{x}}_i\right)\right] \tag{3.11}$$

Para $\lambda = 0$ ou $\lambda = 1$, pode-se verificar que:

$$\lambda S\left(\underline{\theta}_1 \mid \tilde{\underline{\mathbf{Y}}}\right)+(1-\lambda)S\left(\underline{\theta}_2 \mid \tilde{\underline{\mathbf{Y}}}\right)= S\left((\lambda\underline{\theta}_1+(1-\lambda)\underline{\theta}_2)\mid \tilde{\underline{\mathbf{Y}}}\right)+\ell nk \quad ,$$

ou seja,

$$S\left((\lambda\underline{\theta}_1+(1-\lambda)\underline{\theta}_2)\mid \tilde{\underline{\mathbf{Y}}}\right)< \lambda S\left(\underline{\theta}_1 \mid \tilde{\underline{\mathbf{Y}}}\right)+(1-\lambda)S\left(\underline{\theta}_2 \mid \tilde{\underline{\mathbf{Y}}}\right) \quad .$$

Para $0 < \lambda < 1$ basta verificar que:

$$\lambda S\left(\underline{\theta}_1 \mid \tilde{\underline{\mathbf{Y}}}\right)+(1-\lambda)S\left(\underline{\theta}_2 \mid \tilde{\underline{\mathbf{Y}}}\right)=$$

$$= -\sum_{i=1}^{n}\left[\sum_{j=1}^{k-1}\tilde{y}_{ji}\left(\underline{\theta}_{2j}^{T}+\lambda\underline{\theta}_{1j}^{T}-\lambda\underline{\theta}_{2j}^{T}\right)\underline{\mathbf{x}}_i - \ell n\left(\left(1+\sum_{j=1}^{k-1}\exp\left(\lambda\underline{\theta}_{1j}^{T}\underline{\mathbf{x}}_i\right)\right)\left(1+\sum_{j=1}^{k-1}\exp\left(\underline{\theta}_{2j}^{T}-\lambda\underline{\theta}_{2j}^{T}\right)\underline{\mathbf{x}}_i\right)\right)\right]$$

$$> S\left((\lambda\underline{\theta}_1+(1-\lambda)\underline{\theta}_2)\mid \tilde{\underline{\mathbf{Y}}}\right) .$$

A matriz de informação, quadrada de ordem $(k-1)(p+1)$, pode ser escrita também na forma:

$$I(\underline{\theta})=\begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1(k-1)} \\ T_{21} & T_{22} & \cdots & T_{2(k-1)} \\ \cdots & \cdots & \cdots & \cdots \\ T_{(k-1)1} & T_{(k-1)2} & \cdots & T_{(k-1)(k-1)} \end{bmatrix} \quad , \tag{3.12}$$

Na forma (3.12) cada bloco T_{ij} é definido como:

$$T_{ij} = \begin{cases} \underline{\mathbf{X}}^T \left[diag \left[P(G_s \mid \underline{\mathbf{x}}_i)(1 - P(G_s \mid \underline{\mathbf{x}}_i)) \right] \right] \underline{\mathbf{X}} \quad , \quad i = j \quad , \quad s = i \\ (-1)\underline{\mathbf{X}}^T \left[diag \left[P(G_i \mid \underline{\mathbf{x}}_i) P(G_j \mid \underline{\mathbf{x}}_i) \right] \right] \quad , \quad i \neq j \end{cases} . \qquad (3.13)$$

Para a escolha de δs_j e δf_j optou-se por utilizar o mesmo procedimento adotado no modelo para variável resposta dicotômica. Neste trabalho assume-se que

$$\delta f_j = \gamma_j \quad \text{e} \quad \delta s_j = 1 - \delta f_j \quad ,$$

com $\gamma_j = 0{,}0001$. Com isto atende-se à necessidade da abordagem simétrica, isto é, escolher γ_j tal que o valor de $\| \gamma_j \|^2$ possa ser ignorado na fase de implementação computacional.

Finalmente, para aumentar a informação a respeito dos modelos obtidos, optou-se por estimar o viés dos estimadores obtidos através do Método *Bootstrap*, seguindo o algoritmo apresentado em 2.5.1. Cabe acrescentar que a aplicação do referido método visa apenas estimar o viés, e que a mesma não se constitui em uma abordagem destinada à redução do mesmo.

3.2 MODELO DE REGRESSÃO LOGÍSTICA DE COMPONENTES PRINCIPAIS PARA VARIÁVEL RESPOSTA POLITÔMICA

A extensão da Análise de Componentes Principais (ACP) a problemas com variável resposta politômica, também proposta neste trabalho, não requer uma formulação complexa. De fato, pode ser considerada como um método de substituição de variáveis, já que as p variáveis originais são substituídas por s componentes principais, $s \leq p$. Esta substituição tem como efeito mais evidente a redução do volume de dados, para $s < p$ colunas, o que acaba por exigir menor esforço computacional na fase de implementação do modelo. Outra questão de interesse refere-se ao comportamento do modelo com relação às diferentes configurações dos conjuntos de dados, mais especificamente quando há grupos completamente separados.

Sejam g grupos, G_1 , ... , G_g , de observações na forma da matriz de dados

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{p1} \\ x_{12} & x_{22} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}.$$

O MRLC para variável resposta politômica pode ser escrito na forma:

$$P(G_s \mid \underline{\mathbf{x}}_i) = \frac{\exp\left(\beta_{s0} + \sum_{k=1}^{p} \beta_{sk} x_{ki}\right)}{\sum_{j=1}^{g} \exp\left(\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ki}\right)}$$

O primeiro passo é o cálculo da matriz **S** de covariâncias e os correspondentes autovalores e autovetores. Este cálculo pode ser efetuado através do Algoritmo de Jacobi, já que a matriz é simétrica. Seguindo o mesmo raciocínio utilizado para o modelo com variável resposta dicotômica, pode-se fazer $x_{ki} = \sum_{j=1}^{p} z_{kj} v_{ij}$ e, em seguida, efetuar a substituição no modelo. Com isto tem-se que:

$$P(G_s \mid Z\underline{v}_i) = \frac{\exp\left(\beta_{s0} + \sum_{k=1}^{p}\sum_{j=1}^{p} z_{kj} v_{ij} \beta_{sk}\right)}{\sum_{j=1}^{g} \exp\left(\beta_{j0} + \sum_{k=1}^{p}\sum_{j=1}^{p} z_{kj} v_{ij} \beta_{jk}\right)} \qquad (3.14)$$

onde $i = 1, \dots, k, j = 0, \dots, p$ e $\beta_{kj} = 0$.

Fazendo $\gamma_{ij} = \sum_{q=1}^{p} v_{qj} \beta_{iq}$, e substituindo em (3.14), tem-se o Modelo de Regressão Logística de Componentes Principais (MRLCP) para variável resposta politômica, que pode ser escrito na forma:

$$P(G_s \mid \mathbf{Z}\underline{\mathbf{V}}) = \frac{\exp\left(\beta_{s0} + \sum_{j=1}^{p} z_j \gamma_{sj}\right)}{\sum_{i=1}^{k} \exp\left(\beta_{i0} + \sum_{j=1}^{p} z_j \gamma_{ij}\right)} \quad , \qquad (3.15)$$

onde $\gamma_{ij} = \sum_{q=1}^{p} v_{qj} \beta_{iq}$, $i = 1, \ldots, k$ e $j = 1, \ldots, p$.

Os parâmetros desconhecidos podem ser estimados através do Método da Máxima Verossimilhança, na mesma forma utilizada para o Modelo de Regressão Logística Clássico (MRLC). Neste trabalho optou-se pela construção do modelo a partir da matriz de dados com valores normalizados.

A escolha das componentes principais é abordada de forma mais detalhada por Aguilera, Escabias e Valderrama (2006). Os autores alertam que as componentes com maior variância não são necessariamente as melhores preditoras, já que componentes com pequena variância podem ser altamente correlacionadas com a variável resposta. Além disso, ainda segundo os autores, o Modelo de Regressão de Componentes Principais é alvo de críticas por parte de alguns autores, que usam como argumento o fato de que as componentes principais são obtidas sem levar em consideração a dependência entre a variável resposta e as variáveis explanatórias. Para resolver este problema, Aguilera, Escabias e Valderrama (2006) utilizam um procedimento que se inicia com um modelo sem componentes principais. Na seqüência adiciona-se a este modelo uma componente principal a cada passo, até que não haja nenhum ganho expressivo no desempenho do mesmo. Além do problema mencionado, alguns autores, como Hubert, Rousseeuw e Verboven (2002), por exemplo, apontam que os algoritmos comumente usados para a determinação das componentes principais demandam grande esforço computacional, sobretudo para grandes conjuntos de dados.

Com o objetivo de obter a melhor estimação possível para os parâmetros, os autores propõem diferentes critérios baseados em medidas distintas de eficiência dos parâmetros estimados. Inicialmente é definido o Erro Quadrático Médio do vetor de parâmetros, dado por:

$$MSEB_{(s)} = \frac{1}{p+1} \sum_{j=0}^{p} \left(\hat{\beta}_{j(s)} - \beta_j\right)^2 \quad . \qquad (3.16)$$

Na seqüência é definido o Máximo das Diferenças Absolutas dos parâmetros,

$$Max_{(s)} = Max_j \left\{ \left| \hat{\beta}_{j(s)} - \beta_j \right| \right\} \quad . \tag{3.17}$$

Finalmente, é definido o Erro Quadrático Médio das Probabilidades como,

$$MSEP_{(s)} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\pi}_{i(s)} - \pi_i \right)^2 \quad . \tag{3.18}$$

De acordo com os autores espera-se que pequenos valores para as três medidas indiquem melhor estimação dos parâmetros.

É importante ressaltar que o procedimento descrito foi utilizado em estudos que utilizaram conjuntos de dados gerados a partir de um modelo conhecido. Em seguida efetuou-se a comparação dos estimadores obtidos com os parâmetros do modelo, conhecidos *a priori*. Na prática, quando se trabalha com dados de observação, os parâmetros não são conhecidos, fato que impede o cálculo das medidas anteriores. Desta forma, pode-se utilizar como medida de ajuste a variância dos estimadores, dada por:

$$Var\left[\hat{\beta}_{(s)}\right] = \mathbf{V}_{(s)}^{\mathrm{T}} \left(\mathbf{Z}_{(s)}^{\mathrm{T}} \hat{\mathbf{W}}_{(s)} \mathbf{Z}_{(s)} \right)^{-1} \mathbf{V}_{(s)} \quad , \tag{3.19}$$

onde $\hat{\mathbf{W}}_{(s)} = diag\left(\hat{\pi}_{i(s)}\left(1 - \hat{\pi}_{i(s)}\right)\right)$.

Conforme Aguilera, Escabias e Valderrama (2006), geralmente as melhores simulações, isto é, com os menores valores para (3.16), apresentam grande acréscimo em suas variâncias estimadas.

Neste trabalho a Análise de Componentes Principais foi abordada com o objetivo de verificar a sua eficiência na estimação de parâmetros do Modelo de Regressão Logística, especialmente em casos nos quais os métodos conhecidos não apresentam convergência. Por este motivo não foi dedicada maior atenção à escolha das componentes principais que integrarão o modelo, sendo consideradas as *s* primeiras componentes principais tais que:

$$\left[\frac{\sum_{i=1}^{s}\lambda_i}{\sum_{j=1}^{p}\lambda_j}\right] \geq 0{,}95\,.$$

Como objetivo adicional, deseja-se também avaliar o desempenho do MRLCP incluindo diferentes componentes principais, a fim de verificar se o mesmo está sujeito a alguma influência quando o conjunto de dados originais é substituído pelas correspondentes componentes principais.

4 RESULTADOS E DISCUSSÕES

Para comparar a eficiência dos métodos abordados neste trabalho os diferentes modelos foram aplicados a conjuntos de dados extraídos da literatura disponível. A opção por tais conjuntos visa facilitar a comparação dos resultados obtidos com aqueles que eventualmente tenham sido alcançados por outros pesquisadores. Os mesmos conjuntos podem ser obtidos no endereço eletrônico www.fesppr.br/~inacio/BancosDeDados. A implementação computacional foi levada a efeito através de um programa escrito em linguagem *Visual Basic 6.0*®, executado em um computador *Hewlett-Packard*®, modelo *Pavillion b1040br*, processador *Intel*® *Pentium*® 4 – 2,93 Ghz.

O primeiro conjunto de dados, extraído de Hosmer e Lemeshow (1989), envolve variáveis estudadas em exames de mamografia, cujas características são apresentadas no Quadro 4.1. Os dados são resultantes da observação de 412 casos, sendo 104 pertencentes ao Grupo 1, 74 pertencentes ao Grupo 2 e 234 pertencentes ao Grupo 3, e podem ser obtidos na obra citada.

Quadro 4.1 – Variáveis observadas no conjunto MAMOGRAFIA.

Variável	Codificação	Abreviatura
Histórico de exame mamográfico (Variável Resposta)	1 – Nunca 2 – Há menos de um ano 3 – Há mais de um ano	ME
O exame é necessário apenas quando são apresentados os sintomas? (Pergunta feita à paciente)	1 – Concordo fortemente 2 – Concordo 3 – Discordo 4 – Discordo fortemente	SYMPT
Os benefícios do exame são perceptíveis? (Pergunta feita à paciente)	Escala variando de 5 a 20.	PB
Mãe ou irmã com histórico de câncer de mama.	0 – Não 1 – Sim	HIST
Sabe como efetuar o auto-exame?	0 – Não 1 – Sim	BSE
O exame de mamografia é confiável para diagnosticar novos casos de câncer de mama?	1 – Não é confiável 2 – É pouco confiável 3 – É muito confiável	DETC

O segundo conjunto de dados, extraído de Fisher (1936), contém 150 observações referentes às dimensões das sépalas e das pétalas de três espécies de íris, *iris setosa* (G_1), *iris versicolor* (G_2) e *iris virginica* (G_3). Para cada espécie foram efetuadas 50 observações, e as variáveis são descritas no Quadro 4.2. Este é, provavelmente, o mais conhecido banco de dados utilizado em trabalhos publicados na literatura disponível, tendo sido usado por Lesaffre e Albert (1989), entre outros, para demonstrar aspectos referentes à separação de grupos e suas conseqüências para a estimação de parâmetros. É bem sabido que o grupo 1, de exemplares da

espécie *iris setosa*, é completamente separado dos grupos 2 e 3, das espécies *iris versicolor e iris virginica*, respectivamente, conforme pode-se observar nas Figuras 4.1 e 4.2, que apresentam os espaços discriminantes para algumas combinações de variáveis independentes. Aqui também foi utilizado o Grupo 3 como grupo de referência.

Quadro 4.2 – Variáveis observadas no conjunto IRIS.

Variável	**Codificação (Domínio)**	**Abreviatura**
Espécies (Variável resposta).	1 – *Iris Setosa* 2 – *Iris Versicolor* 3 – *Iris Virginica*	Species
Comprimento da sépala	Medidas variando de 43 a 79 mm.	X_1
Largura da sépala	Medidas variando de 20 a 44 mm.	X_2
Comprimento da pétala	Medidas variando de 10 a 69 mm.	X_3
Largura da pétala	Medidas variando de 1 a 25 mm.	X_4

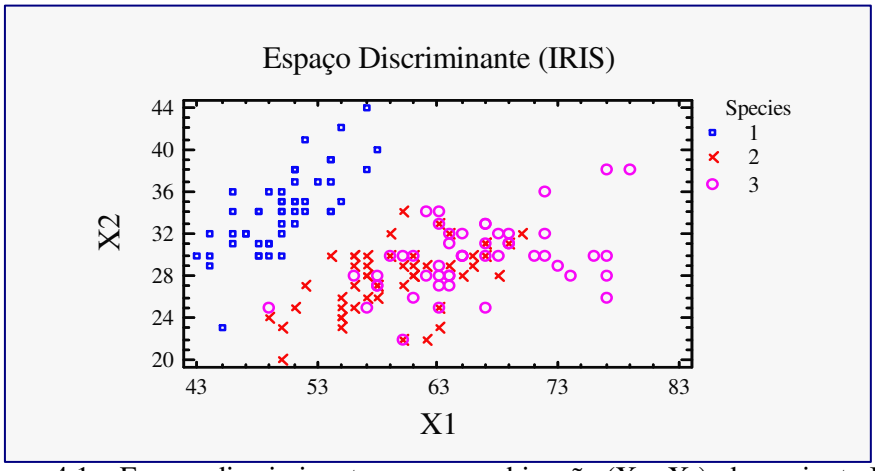

Figura 4.1 – Espaço discriminante para a combinação (X_1 , X_2), do conjunto IRIS.

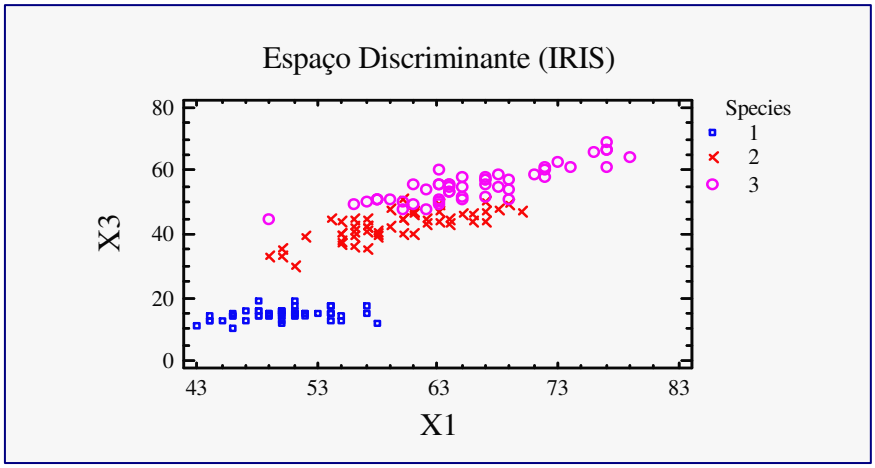

Figura 4.2 – Espaço discriminante para a combinação (X_1 , X_3), do conjunto IRIS.

O terceiro conjunto de dados contém 2567 observações correspondentes a valores anotados na inspeção de amostras de óleo isolante empregado em transformadores elétricos com tensão máxima de 69 kV, e foi apresentado por Paixão e Chaves Neto (2006). São definidos três grupos a partir da variável resposta ESTADO, sendo 1806 observações pertencentes ao Grupo 1, 114 observações pertencentes ao Grupo 2 e 647 observações pertencentes ao Grupo 3. Este conjunto foi utilizado em um estudo com o objetivo de construir uma regra discriminante que permita determinar o estado do óleo isolante. No estudo em questão os autores aplicaram a Função Discriminante Quadrática. As variáveis são mostradas no Quadro 4.3.

Quadro 4.3 – Variáveis observadas no conjunto ÓLEO ISOLANTE.

Variável	Domínio	Abreviatura
Estado do óleo (Variável resposta)	1 – Bom 2 – A recuperar 3 – A regenerar	ESTADO
Índice de neutralização	0,003 – 0,671 (mg KOH/g)	IN
Fator de potência	0,10 – 28,6 (%)	FP
Rigidez dielétrica	16 – 71 kV	RD
Tensão interfacial	11,5 – 51,2 (dina/cm^2)	TI
Teor de água	2 – 82 (ppm)	TA
Temperatura do óleo	5 – 85 (^{o} C)	TO

Foi também providenciada a construção de modelos de classificação baseados na Função Discriminante Linear e em Redes Neurais Artificiais. O desempenho dos modelos obtidos a partir das diferentes abordagens é avaliado mediante a comparação das taxas de classificações efetuadas corretamente, na forma de matrizes de classificações.

4.1 RESULTADOS PARA O CONJUNTO MAMOGRAFIA

Inicialmente foram obtidos o Modelo de Regressão Logística Clássico (MRLC), os Modelos de Regressão Logística Individualizados (MRLI) e o Modelo de Regressão Logística Oculto (MRLO). As variáveis SYMPT e DETC foram codificadas através das variáveis binárias SYMPT(1), SYMPT(2), SYMPT(3) e DETC(1), DETC(2), respectivamente. Desta forma para SYMPT = 4, por exemplo, tem-se SYMPT(1) = 1, SYMPT(2) = 1 e SYMPT(3) = 1. Se SYMPT = 2, tem-se SYMPT(1) = 0, SYMPT(2) = 1 e SYMPT(3) = 0. Utilizou-se como referência o grupo 3. Os estimadores obtidos são mostrados no Quadro 4.4. Os resultados obtidos pelos diferentes modelos são bastante próximos, tanto para os estimadores como para os respectivos erros padrões.

Quadro 4.4 – Estimadores para os Modelos de Regressão Logística Clássico (MRLC), Individualizados (MRLI) e Oculto (MRLO). Conjunto MAMOGRAFIA.

Função	Variável	MRLC	Erro Padrão	MRLI	Erro Padrão	MRLO	Erro Padrão
1	SYMPT(1)	2,1298	0,4818	2,1425	0,4901	2,1282	0,4655
	SYMPT(2)	0,3260	0,4801	0,3831	0,4858	0,3259	0,4636
	SYMPT(3)	0,2093	0,4884	0,1423	0,4956	0,2092	0,4721
	PB	- 0,2213	0,0754	- 0,2145	0,0766	- 0,2212	0,0748
	HIST	1,3670	0,4375	1,4153	0,4687	1,3663	0,4354
	BSE	1,2904	0,5300	1,3998	0,5384	1,2894	0,5206
	DETC(1)	- 0,9011	1,1265	- 1,0490	1,1268	- 0,8998	1,0910
	DETC(2)	- 0,0061	1,1613	- 0,1947	1,1667	- 0,0052	1,1254
	Intercepto	- 2,2816	1,4738	- 2,2568	1,4963	- 2,2804	1,4360
2	SYMPT(1)	1,1100	0,3623	1,1369	0,3626	1,1096	0,3581
	SYMPT(2)	0,0200	0,3559	0,0709	0,3572	0,0201	0,3515
	SYMPT(3)	0,2996	0,3663	0,3349	0,3702	0,2934	0,3616
	PB	- 0,1504	0,0762	- 0,1447	0,0755	- 0,1504	0,0756
	HIST	1,0660	0,4593	1,1573	0,4735	1,0655	0,4575
	BSE	1,0505	0,5150	1,0165	0,5158	1,0497	0,5063
	DETC(1)	0,6941	0,6870	0,5706	0,6869	0,6940	0,6829
	DETC(2)	0,9358	0,7132	0,7896	0,7174	0,9356	0,7084
	Intercepto	- 2,8915	1,1237	- 2,8282	1,1192	- 2,8900	1,1131

Os coeficientes das funções discriminantes lineares obtidas para o conjunto em questão são mostrados no Quadro 4.5. A Rede Neural Artificial utilizada neste trabalho segue o raciocínio mostrado na Figura 2.7. As camadas de entrada e oculta possuem *p* neurônios cada uma, onde *p* é o número de variáveis independentes. Os pesos foram iniciados com valores aleatórios seguindo uma distribuição uniforme $U(-0,5;0,5)$, o algoritmo utilizado para treinamento é do tipo retro-propagação, e segue os passos apresentados no Algoritmo 4.1.1.

Quadro 4.5 – Coeficientes das Funções Discriminantes Lineares. Conjunto MAMOGRAFIA

Variável	Função Discriminante	
	Primeira	Segunda
SYMPT(1)	1,4642	0,3706
SYMPT(2)	0,2740	- 0,6823
SYMPT(3)	0,2876	0,0721
PB	- 0,1822	- 0,0483
HIST	1,1809	0,1627
BSE	0,8690	1,4159
DETC(1)	0,0478	4,0925
DETC(2)	0,6207	2,9410
Autovalores	0,2845	0,0066

Quadro 4.6 – Matrizes de classificações observadas para o conjunto MAMOGRAFIA.

Modelo	Grupo Observado	Grupo Previsto		
		1	2	3
MRLO	1	0,1346	0,0000	0,8654
	2	0,0811	0,0405	0,8784
	3	0,0171	0,0000	0,9829
FDL	1	0,6154	0,2308	0,1538
	2	0,5000	0,2297	0,0405
	3	0,2051	0,1880	0,6068
RNA	1	0,8558	0,0000	0,1442
	2	0,2162	0,7432	0,0405
	3	0,0000	0,0128	0,9872

O Quadro 4.6 mostra que o desempenho do MRLO significativamente inferior, quando comparado à FDL e a uma RNA. Também é possível notar que os dois modelos logísticos, MRLC e MRLO, apresentaram desempenhos muito próximos entre si, embora demonstrem ser pouco úteis para o problema em questão.

4.2 RESULTADOS PARA O CONJUNTO IRIS

O Modelo de Regressão Logística Clássico (MRLC) não apresentou convergência, fato que já era esperado, em função da conhecida configuração do conjunto de dados. Os Modelos de Regressão Logística Individualizados (MRLI) apresentaram estimadores para a segunda função discriminante, que discrimina as observações do grupo 2 em relação ao grupo 3, mas não houve convergência para os estimadores da primeira função discriminante, que deveria discriminar as observações do grupo 1 em relação ao grupo 3, fato que também está de acordo com a mencionada configuração dos dados. Quanto ao Modelo de Regressão Logística Oculto (MRLO) não houve problemas de convergência, o que possibilitou a obtenção de todos os estimadores, confirmando a imunidade do método às diferentes configurações de dados. Para a implementação computacional utilizou-se $\gamma_j = 10^{-4}$. Os resultados são apresentados no Quadro 4.8. A Análise de Componentes Principais apresentou os autovalores e autovetores mostrados no Quadro 4.7. Estes valores foram obtidos a partir da matriz de dados normalizados. Os estimadores para o Modelo de Regressão Logística de Componentes Principais (MRLCP) não foram obtidos. A não existência dos estimadores para o MRLCP indica que a Análise de Componentes Principais, embora possibilite a redução do esforço computacional, não representa uma garantia contra o problema da não existência dos estimadores quando há separação completa de pelo menos um dos grupos.

Quadro 4.7 – Variâncias e autovetores do conjunto IRIS.

Variável	Autovetores			
	v_1	v_2	v_3	v_4
X_1	0,3614	0,6566	0,5820	0,3155
X_2	- 0,0845	0,7302	- 0,5979	- 0,3197
X_3	0,8567	- 0,1734	- 0,0762	- 0,4798
X_4	0,3583	- 0,0755	- 0,5458	0,7537
Variância (λ_i)	420,0053	24,1053	7,7688	2,3676
Percentagem	0,9246	0,0531	0,0171	0,0052

Quadro 4.8 – Estimadores para os Modelos de Regressão Logística Clássico (MRLC), Individualizados (MRLI) e Oculto (MRLO). Conjunto IRIS.

Função	**Variável**	**MRLC**	**Erro Padrão**	**MRLI**	**Erro Padrão**	**MRLO**	**Erro Padrão**
1	X_1	NE*	NE	NE	NE	0,4611	2,7904
	X_2	NE	NE	NE	NE	1,0454	2,0433
	X_3	NE	NE	NE	NE	- 1,4298	2,5094
	X_4	NE	NE	NE	NE	- 2,2577	4,7388
	Intercepto	NE	NE	NE	NE	36,0292	94,6970
2	X_1	NE	NE	0,2465	0,2394	0,2464	0,2377
	X_2	NE	NE	0,6681	0,4480	0,6541	0,4388
	X_3	NE	NE	- 0,9429	0,4737	- 0,9258	0,4599
	X_4	NE	NE	- 1,8286	0,9743	- 1,7886	0,9479
	Intercepto	NE	NE	42,6378	25,7077	41,5301	24,9025

* NE = Não Existe

Os coeficientes obtidos para as funções discriminantes lineares são apresentados no Quadro 4.9, juntamente com os respectivos autovalores. A Rede Neural utilizada segue o mesmo raciocínio exposto para o Conjunto MAMOGRAFIA. Os desempenhos dos modelos obtidos para o conjunto IRIS são apresentados no Quadro 4.10, e indicam ligeira superioridade do MRLO, mais precisamente para as observações do grupo 2, em relação à Função Discriminante Linear e à Rede Neural Artificial.

Quadro 4.9 – Coeficientes das Funções Discriminantes Lineares. Conjunto IRIS

Variável	**Função Discriminante**	
	Primeira	**Segunda**
X_1	- 0,0838	0,0024
X_2	- 0,1550	0,2187
X_3	0,2224	- 0,0941
X_4	0,2839	0,2868
Autovalores	32,1919	0,2854

Quadro 4.10 – Taxas de classificações efetuadas corretamente no conjunto IRIS.

Modelo	Grupo Observado	Grupo Previsto		
		1	2	3
MRLO	1	1,0000	0,0000	0,0000
	2	0,0000	0,9800	0,0200
	3	0,0000	0,0200	0,9800
FDL	1	1,0000	0,0000	0,0000
	2	0,0000	0,9600	0,0400
	3	0,0000	0,0200	0,9800
RNA	1	1,0000	0,0000	0,0000
	2	0,0000	0,9600	0,0400
	3	0,0000	0,0200	0,9800

4.3 RESULTADOS PARA O CONJUNTO ÓLEO ISOLANTE

O óleo mineral utilizado como isolante em transformadores elétricos é submetido a reações de oxidação devido à presença de oxigênio, água e metais. O acompanhamento e a manutenção da qualidade do óleo isolante têm por objetivo assegurar uma operação confiável dos transformadores. A verificação das medidas dos índices nem sempre é feita a tempo de se evitar panes ou mesmo a troca do equipamento. Atualmente, a técnica mais usada para a prevenção de falhas neste tipo de equipamento é a manutenção preditiva, caracterizada pela análise físico-química do óleo isolante utilizado. A avaliação é efetuada com base na interpretação de medidas realizadas através de ensaios físico-químicos e que são comparadas a limites admissíveis aplicados, conforme os Quadros 4.11, 4.12 e 4.13.

Quadro 4.11 – Índices de classificação para óleo isolante classificado como BOM.

Variável (Unidade)	**Tensão**		
	Até 69 kV	**69 a 240 kV**	**Acima de 240 kV**
Teor de água (ppm)	< 30	< 25	< 20
Rigidez dielétrica (kV)	> 30	> 30	> 35
Índice de neutralização (mg KOH/g)	< 0,15	< 0,15	< 0,10
Tensão interfacial (dina/cm^2)	> 20	> 20	> 22
Fator de potência (%)	< 15	< 15	< 15

Fonte: Paixão e Chaves Neto (2006).

Quadro 4.12 – Índices de classificação para óleo isolante classificado como A REGENERAR.

Variável (Unidade)	**Tensão**		
	Até 69 kV	**69 a 240 kV**	**Acima de 240 kV**
Índice de neutralização (mg KOH/g)	> 0,15	> 0,15	> 0,10
Tensão interfacial (dina/cm^2)	< 20	< 20	< 22
Fator de potência (%)	> 15	> 15	> 15

Fonte: Paixão e Chaves Neto (2006).

Quadro 4.13 – Índices de classificação para óleo isolante classificado como A RECUPERAR.

Variável (Unidade)	Tensão		
	Até 69 kV	69 a 240 kV	Acima de 240 kV
Teor de água (ppm)	> 30	> 25	> 20
Rigidez dielétrica (kV)	< 30	< 30	< 35
Índice de neutralização (mg KOH/g)	< 0,15	< 0,15	< 0,10
Tensão interfacial (dina/cm^2)	> 20	> 20	> 22
Fator de potência (%)	< 15	< 15	< 15

Fonte: Paixão e Chaves Neto (2006).

O conjunto em questão foi utilizado por Paixão e Chaves Neto (2006) para a construção de um modelo discriminante a partir da Função Discriminante Quadrática (FDQ). A função é dada por:

$$d_i^Q(\underline{x}) = -\frac{1}{2}\ell n|\Sigma_i| - \frac{1}{2}(\underline{x}-\mu_k)^T \Sigma_k^{-1}(\underline{x}-\mu_k) + \ell n p_i \qquad (4.1)$$

onde

μ_i = vetor médio da população Π_i .

Σ_i = matriz de covariâncias da população Π_i .

p_i = probabilidade *a priori* da observação pertencer à população Π_i .

$\underline{x}$ = vetor aleatório.

A regra de classificação consiste em alocar $\underline{x}$ em Π_k se $d_k^Q(\underline{\mathbf{x}}) = máx d_i^Q(\underline{\mathbf{x}})$, $i = 1, \ldots, k$.

As funções discriminantes têm a forma:

$$d_k = \sum_{i=1}^{p} \alpha_i x_i^2 + \sum_{i=1}^{p-1} \beta_i x_i x_{i+1} + \sum_{i=1}^{p} \gamma_i x_i + \kappa \qquad (4.2)$$

A Função Discriminante Quadrática obtida por Paixão e Chaves Neto (2006) apresentou a matriz de classificações mostrada no Quadro 4.14.

Quadro 4.14 – Matriz de Classificações da FDQ para o conjunto ÓLEO ISOLANTE.

		Grupo Previsto		
		1 – Bom	2 – A recuperar	3 – Regenerar
Grupo Observado	1	0,8793	0,0526	0,0681
	2	0,0000	0,9737	0,0263
	3	0,0680	0,0247	0,9073

Fonte: Paixão e Chaves Neto (2006).

O Modelo de Regressão Logística Clássico (MRLC) não apresentou convergência, indicando a completa separação de pelo menos um dos três grupos. Os Modelos de Regressão Logística Individualizados (MRLI) não apresentaram convergência para a segunda função discriminante, indicando que o Grupo 2 é completamente separado do Grupo 3, utilizado como grupo de referência, o que está de acordo com o desempenho do MRLC. O Modelo de Regressão Logística Oculto (MRLO) apresentou os estimadores mostrados no Quadro 4.15, juntamente com os respectivos erros padrões, confirmando a sua robustez em casos de separação total, ou parcial de grupos.

Quadro 4.15 – Estimadores para o MRLO. Conjunto ÓLEO ISOLANTE.

Variável	Função			
	Primeira		Segunda	
	Estimadores	Erro Padrão	Estimadores	Erro Padrão
Intercepto	- 991,2077	653,4131	-937,1940	657,4500
IN	4239,2423	1161,0326	4823,3180	1122,2514
FP	21,0500	10,5751	35,3202	10,1592
RD	- 6,0001	10,0635	- 11,0042	10,1592
TI	77,3480	36,3654	83,0050	36,3525
TA	- 17,2660	7,4260	- 24,1805	7,4655
TO	- 0,3208	4,6191	- 0,1756	4,6669
IN^2	- 1457,6189	463,4254	- 2826,8986	911,0983
IN × FP	- 10,4406	11,6354	- 44,4612	20,3344
IN × RD	3,6378	4,5702	2,5725	5,5612
IN × TI	- 218,6601	53,9111	- 231,0900	53,5668
IN × TA	5,6056	4,2733	4,9522	6,0819
IN × TO	3,6603	2,8594	5,1862	4,1339
FP^2	- 0,0707	0,0925	- 0,3890	0,1732
FP × RD	0,0482	0,0583	0,0469	0,0614
FP × TI	- 1,1316	0,3710	- 1,5083	0,3762
FP × TA	0,1050	0,0580	0,1035	0,0776
FP × TO	- 0,0232	0,0441	- 0,0386	0,0607
RD^2	- 0,0214	0,0171	- 0,0096	0,0179
RD × TI	0,3741	0,4939	0,3932	0,4943
RD × TA	- 0,0309	0,0259	0,0797	0,0263
RD × TO	0,0089	0,0147	0,0130	0,0163
TI^2	- 1,4385	0,3519	- 1,5262	0,3489
TI × TA	0,9727	0,3606	0,9800	0,3599
TI × TO	- 0,0207	0,2364	- 0,0213	0,2369
TA^2	- 0,0455	0,0155	0,0351	0,0156
TA × TO	- 0,0014	0,0165	- 0,0071	0,0177
TO^2	0,0014	0,0071	0,0001	0,0089

A matriz de classificações do MRLO é apresentada no Quadro 4.16. Todos os modelos contêm 27 variáveis independentes, resultantes de combinações efetuadas com as seis variáveis

independentes originais. O desempenho apresentado para as observações do grupo 2, confirma a completa separação do mesmo, já indicada pelo comportamento do MRLI, que não obteve estimadores para a segunda função discriminante. Para o conjunto em questão a QDF ajustada apresentou a matriz de classificações mostrada no Quadro 4.14, onde é possível notar que o MRLO apresentou desempenho superior à QDF em todos os grupos.

Quadro 4.16 – Matriz de Classificações do MRLO para o conjunto ÓLEO ISOLANTE.

Grupo Observado	Grupo Previsto		
	1 – Bom	2 – A recuperar	3 – Regenerar
1	0,9989	0,0006	0,0006
2	0,0000	1,0000	0,0000
3	0,0031	0,0000	0,9969

4.4 REPLICAÇÕES *BOOTSTRAP*

Com o objetivo de estimar o viés dos estimadores obtidos para o Modelo de Regressão Logística Oculto, algumas das simulações relatadas foram repetidas mediante a aplicação do Método *Bootstrap* para a obtenção de um Modelo de Regressão Logística *Bootstrap*, seguindo o algoritmo apresentado em 2.5.1. Os estimadores *bootstrap* encontrados para este modelo foram comparados aos estimadores obtidos pelo MRLO, com o viés dado pela diferença entre os respectivos valores.

Para o conjunto MAMOGRAFIA foram geradas 300 amostras *bootstrap*, que forneceram modelos cujos estimadores são mostrados no Quadro 4.17, juntamente com o viés encontrado para cada estimador. É possível perceber que nenhum dos estimadores encontrados apresenta viés elevado, em relação aos respectivos estimadores do Modelo de Regressão Logística Oculto.

Para o conjunto IRIS foram geradas 500 amostras *bootstrap*, e obtidos os estimadores apresentados no Quadro 4.18. Para estudar o comportamento do viés dos estimadores foi executada uma nova simulação com os dados do conjunto IRIS, desta vez incluindo no modelo apenas as variáveis X_1 e X_2 e gerando 500 amostras *bootstrap*. O desempenho do modelo obtido nesta simulação é apresentado no Quadro 4.19.

Quadro 4.17 – Estimadores para o Modelo de Regressão Logística Oculto (MRLO) e estimadores *Bootstrap*. Conjunto MAMOGRAFIA.

Função	**Variável**	**MRLO**	***Bootstrap***	**Viés**
1	SYMPT(1)	2,1282	2,1669	- 0,0387
	SYMPT(2)	0,3259	0,3240	0,0019
	SYMPT(3)	0,2092	0,2076	0,0016
	PB	- 0,2212	- 0,2216	0,0004
	HIST	1,3663	1,3726	0,0387
	BSE	1,2894	1,2496	0,0398
	DETC(1)	- 0,8998	- 1,2497	0,3499
	DETC(2)	- 0,0052	- 0,3578	0,3526
	Intercepto	- 2,2804	- 1,8872	- 0,3932
2	SYMPT(1)	1,1096	1,1409	- 0,0313
	SYMPT(2)	0,0201	0,0836	- 0,0635
	SYMPT(3)	0,2994	0,2497	0,0497
	PB	- 0,1504	- 0,1537	0,0033
	HIST	1,0655	1,0764	- 0,0109
	BSE	1,0497	1,0336	0,0161
	DETC(1)	0,6940	0,6435	0,0505
	DETC(2)	0,9356	0,8937	0,0419
	Intercepto	- 2,8900	- 2,8118	- 0,0782

Quadro 4.18 – Estimadores para o Modelo de Regressão Logística Oculto (MRLO) e estimadores *Bootstrap*. Conjunto IRIS.

Função	**Variável**	**MRLO**	***Bootstrap***	**Viés**
1	X_1	0,4611	0,2900	0,1711
	X_2	1,0454	0,6210	0,4244
	X_3	- 1,4342	- 0,8519	- 0,5823
	X_4	- 2,2577	- 0,7967	- 1,4610
	Intercepto	36,0292	31,2018	4,8274
2	X_1	0,2464	0,2436	0,0028
	X_2	0,6541	0,3509	0,3032
	X_3	- 0,9258	- 0,5666	- 0,3592
	X_4	- 1,7886	- 0,8716	- 0,9170
	Intercepto	41,5301	17,4460	24,0841

Quadro 4.19 – Taxas de classificações efetuadas pelo Modelo de Regressão Logística Oculto no conjunto IRIS, com as variáveis X_1 e X_2.

Grupo Observado	Grupo Previsto		
	1	2	3
1	1,0000	0,0000	0,0000
2	0,0000	0,7600	0,2400
3	0,0000	0,2600	0,7400

É possível perceber pelo Quadro 4.19 que, embora o modelo envolvendo apenas duas variáveis mantenha a capacidade de classificar corretamente a totalidade das observações do Grupo 1, apresentou uma significativa queda na eficiência ao classificar as observações pertencentes ao Grupo 2. Os estimadores obtidos para o MRLO e para o Modelo *Bootstrap* são apresentados no Quadro 4.20.

Quadro 4.20 – Estimadores para os Modelos de Regressão Logística Oculto (MRLO) e *Bootstrap*. Conjunto IRIS.

Função	**Variável**	**MRLO**	***Bootstrap***	**Viés**
1	X_1	- 3,0466	- 0,7239	- 2,3227
	X_2	2,5321	0,6369	1,8952
	Intercepto	85,6596	21,1293	64,5303
2	X_1	- 0,1902	- 0,1873	- 0,0029
	X_2	- 0,0403	- 0,0424	0,0021
	Intercepto	13,0381	12,9474	0,0907

Os resultados apresentados no Quadro 4.20 mostram que a redução da eficiência do modelo é acompanhada por uma redução no viés dos estimadores. Basta verificar que o viés dos estimadores da primeira função, com 100% de classificações corretas, é significativamente maior que o viés dos estimadores da segunda função, cuja taxa de eficiência é igual 76%, especialmente para o intercepto. De outra forma, há indícios de que o viés dos estimadores de um modelo é maior para o modelo com maior poder discriminante. Para verificar este comportamento foi providenciada a obtenção de regras discriminantes para outros dois conjuntos de dados. O primeiro foi extraído de Johnson e Wichern (1988) e contém 56 observações de amostras de petróleo extraídas de três diferentes tipos de solo e possui cinco variáveis explanatórias, apresentadas no Quadro 4.21.

Quadro 4.21 – Variáveis observadas no conjunto ÓLEO CRÚ.

Variável	**Codificação (Domínio)**	**Abreviatura**
Tipo de Solo (Variável resposta).	1 – *Argila Wilhelm* 2 – *Argila Sub-mulinia* 3 – *Argila Superior*	Grupo
Teor de vanádio (%)	Valores entre 1,2 e 11,0	X_1
Teor de Ferro (%)	Valores entre 5,6 e 52	X_2
Teor de Berílio (%)	Valores entre 0 e 1,5	X_3
Teor de Hidrocarbonetos Saturados (%)	Valores entre 3,06 e 9,25	X_4
Teor de Hidrocarbonetos Aromáticos (%)	Valores entre 2,22 e 13,01	X_5

Fonte: Johnson e Wichern (1988).

Em primeiro lugar foi providenciada a estimação de parâmetros para os quatro modelos de Regressão Logística abordados neste trabalho. Tanto o MRLC como o MRLI não apresentaram convergência. Os estimadores obtidos para o MRLO são mostrados no Quadro 4.22. Foram geradas

400 amostras *bootstrap*, que forneceram os estimadores mostrados também no mesmo quadro. Neste caso o MRLO classificou corretamente as 56 observações, sendo sete do Grupo 1, 11 do Grupo 2 e 38 do Grupo 3, tomado como grupo de referência. Na seqüência foi realizada uma simulação para calcular os estimadores de um MRLO que envolvesse apenas as variáveis X_1 , X_3 e X_5. Em seguida foram geradas 400 amostras *bootstrap*, obtendo-se os resultados mostrados no Quadro 4.23 e a matriz de classificações mostrada no Quadro 4.24.

Quadro 4.22 – Estimadores para o MRLO e estimadores *Bootstrap*. Conjunto ÓLEO CRÚ.

Função	**Variável**	**MRLO**	***Bootstrap***	**Viés**
1	X_1	- 18,5603	- 0,8455	- 17,7148
	X_2	4,8244	0,2795	4,5449
	X_3	- 118,2501	- 4,7558	- 113,4943
	X_4	39,9649	- 0,0020	39,9669
	X_5	- 3,6341	0,9459	- 4,5800
	Intercep.	- 254,5676	- 10,646	- 243,9216
2	X_1	- 17,1426	- 0,7232	- 16,4194
	X_2	1,5282	0,0531	1,4751
	X_3	- 195,0314	- 10,1938	- 184,8376
	X_4	59,1918	2,6895	56,5023
	X_5	- 8,7128	- 0,0993	- 8,6135
	Intercep.	- 199,6563	- 9,9869	- 189,6694

Quadro 4.23 – Estimadores para o Modelo de Regressão Logística Oculto (MRLO) e estimadores *Bootstrap*. Conjunto ÓLEO CRÚ.

Função	**Variável**	**MRLO**	***Bootstrap***	**Viés**
1	X_1	- 2,6341	- 1,6801	- 0,9540
	X_3	- 7,8109	- 6,9710	- 0,8399
	X_5	1,3550	0,9237	0,4313
	Intercep.	2,2940	2,2299	0,0641
2	X_1	- 1,3591	- 1,2151	- 0,1440
	X_3	- 8,4934	- 7,8298	- 0,6098
	X_5	0,3865	0,2025	0,1840
	Intercep.	7,1180	7,0028	0,1152

Quadro 4.24 – Taxas de classificações efetuadas pelo MRLO no conjunto ÓLEO CRÚ, com as variáveis X_1 e X_3 X_5 .

Grupo Observado	Grupo Previsto		
	1	2	3
1	1,0000	0,0000	0,0000
2	0,0909	0,5455	0,3636
3	0,0000	0,0000	1,0000

Nesta simulação também é possível notar que o viés dos estimadores é menor que aquele observado para o modelo que contém cinco variáveis, embora esta queda nos valores observados para o viés seja acompanhada por uma queda na eficiência dos modelos obtidos com apenas três variáveis independentes, especialmente do segundo modelo, que classificou corretamente pouco mais que a metade das observações do segundo grupo. Por outro lado, as duas simulações indicam que o viés dos estimadores não compromete, pelo menos aparentemente, a eficiência do modelo.

A Análise de Componentes Principais para o conjunto ÓLEO CRÚ apresentou as componentes principais e autovalores mostrados no Quadro 4.25. Os estimadores para o MRLCP com as três primeiras componentes principais são apresentados no Quadro 4.26, juntamente com os respectivos erros padrões. A matriz de classificações para o MRLCP é apresentada no Quadro 4.27.

Quadro 4.25 – Variâncias e autovetores do conjunto ÓLEO CRÚ.

Variável	Autovetores				
	v_1	v_2	v_3	v_4	v_5
X_1	0,5418	0,0320	0,3824	0,4995	0,5565
X_2	- 0,4971	- 0,0748	- 0,1885	0,8335	- 0,1304
X_3	0,1523	0,9376	- 0,0376	0,1218	- 0,2856
X_4	0,5823	- 0,3382	- 0,0235	0,2004	- 0,7112
X_5	- 0,3115	0,0011	0,9035	- 0,0272	- 0,2932
Variância	2,0837	1,0433	0,9460	0,6343	0,2928
Percentagem	41,67	20,87	18,92	12,69	5,86

Quadro 4.26 – Estimadores para o MRLCP. Conjunto ÓLEO CRÚ.

Função	**Componente**	**Estimador**	**Erro Padrão**
1	Intercepto	- 39,3597	41,0709
	1ª. Componente	- 30,5299	26,3815
	2ª. Componente	29,4016	38,5076
	3ª. Componente	- 7,8825	7,9746
2	Intercepto	- 1,0934	1,2281
	1ª. Componente	- 1,6196	0,5779
	2ª. Componente	- 6,5638	2,9371
	3ª. Componente	- 0,0403	0,2109

Quadro 4.27 – Taxas de classificações efetuadas pelo MRLCP no conjunto ÓLEO CRÚ, com as três primeiras componentes principais.

Grupo Observado	Grupo Previsto		
	1	2	3
1	1,0000	0,0000	0,0000
2	0,0000	0,7273	0,2727
3	0,0000	0,0000	1,0000

No Quadro 4.27 é possível notar que o MRLCP apresenta um desempenho inferior ao MRLO para as observações do grupo 2, embora tenha se mostrado uma alternativa válida para contornar o problema da separação completa de grupos. O conjunto em questão também foi utilizado por Johnson e Wichern (1988) para ilustrar a obtenção de uma Função Discriminante Linear.

O segundo conjunto de dados foi extraído de Brodnjak-Vončina, Kodba, Novič (2005) e contém 120 observações referentes a cinco classes de óleos vegetais. O objetivo das autoras é determinar a origem de amostras de óleos vegetais a partir dos teores de ácidos graxos presentes em cada um dos tipos de óleo vegetal. O conjunto original, disponível no trabalho citado, contém observações de oito tipos de óleo: abóbora, girassol, amêndoas, oliva, soja, colza, milho e de origem desconhecida ou misto. Do conjunto original foram excluídos três grupos, óleo de oliva, com três observações, óleo de soja, com sete observações, e óleo composto, ou misto, com apenas duas observações. Esta exclusão foi motivada apenas pelo pequeno número de observações em cada um dos grupos excluídos. As variáveis observadas e suas definições são apresentadas no Quadro 4.28, juntamente com algumas das características das mesmas.

Para a simulação com o conjunto ÁCIDOS GRAXOS foram consideradas as variáveis apresentadas no Quadro 4.28. No conjunto original as variáveis *Eicosanoic* e *Eicosenoic* assumem valores inferiores a 0,1. O conjunto em questão contém cinco grupos: G_1 (n_1 = 11 observações), G_2 (n_2 = 37), G_3 (n_3 = 26), G_4 (n_4 = 10) e o grupo de referência G_5 (n_5 = 36).

Quadro 4.28 – Variáveis observadas no conjunto ÁCIDOS GRAXOS.

Variável	**Definição (Domínio)**	**Abreviatura**
Classe (Variável resposta)	1 – Colza 2 – Girassol 3 – Amêndoas 4 – Milho 5 – Abóbora	CLASS
Teor de Ácido Palmítico (%)	Valores entre 3,8 e 13,1	Palmitic
Teor de Ácido Esteárico (%)	Valores entre 1,7 e 6,7	Stearic
Teor de Ácido Oléico (%)	Valores entre 22,3 e 80,6	Oleic
Teor de Ácido Linoléico (%)	Valores entre 11,3 e 66,1	Linoleic
Teor de Ácido Linolênico (%)	Valores entre 0,1 e 9,5	Linolenic
Teor de Ácido Eicosanóico (%)	Valores entre 0,0999 e 2,8	Eicosanoic
Teor de Ácido Eicosenóico (%)	Valores entre 0,0999 e 1,8	Eicosenoic

Fonte: Brodnjak-Vončina *et al*. (2005).

Os dados foram utilizados para a obtenção de uma Função Discriminante Linear, de uma Rede Neural Artificial e também para a estimação de parâmetros dos modelos de Regressão Logística aqui abordados. As matrizes de classificações são apresentadas no Quadro 4.29, no qual é possível observar que a FDL, a RNA e o MRLO apresentaram desempenho bastante inferior ao MRLCP, com quatro componentes principais.

Quadro 4.29 – Classificações efetuadas no conjunto ÁCIDOS GRAXOS.

Modelo	Grupo Observado	Grupo Previsto				
		1	2	3	4	5
FDL	1	0,0000	0,0000	0,0000	0,0000	1,0000
	2	0,0000	0,0000	0,0000	0,0000	1,0000
	3	0,0000	0,0000	0,0000	0,0000	1,0000
	4	0,0000	0,0000	0,0000	1,0000	0,0000
	5	0,0000	0,0000	0,0000	0,1944	0,8056
RNA	1	0,0000	0,1818	0,0000	0,0000	0,8182
	2	0,0000	0,5135	0,0541	0,0000	0,4324
	3	0,0000	0,0000	0,5385	0,0000	0,4615
	4	0,0000	0,1000	0,1000	0,5000	0,3000
	5	0,0000	0,1111	0,0833	0,1389	0,6667
MRLO	1	0,0000	0,0000	0,0000	0,6364	0,3636
	2	0,0000	0,0000	0,0000	0,9459	0,0541
	3	0,0000	0,0000	1,0000	0,0000	0,0000
	4	0,0000	0,0000	0,0000	0,7000	0,3000
	5	0,0000	0,0000	0,0278	0,2500	0,7222
MRLCP	1	0,4545	0,0000	0,0000	0,0000	0,5455
	2	0,0000	0,9189	0,0000	0,0000	0,0811
	3	0,0000	0,0000	1,0000	0,0000	0,0000
	4	0,0000	0,0000	0,0000	0,8000	0,2000
	5	0,1111	0,0556	0,0833	0,0556	0,6944

O MRLC não apresentou convergência, enquanto o MRLI não apresentou convergência apenas para a terceira função discriminante, o que está de acordo com a taxa de classificações apresentada pelo MRLO para as observações do Grupo 3. Os estimadores para os modelos são mostrados no Quadro 4.30. Para aplicar o Método *Bootstrap* foram geradas 300 amostras e os estimadores obtidos são apresentados no Quadro 4.30, juntamente com o viés dos estimadores em relação ao MRLO. O MRLCP foi construído com as cinco componentes principais mostradas no Quadro 4.31 e apresentou a maior eficiência, o que dá uma boa idéia do seu potencial na construção de regras discriminantes baseadas no Modelo de Regressão Logística através das componentes principais. Além do desempenho superior, o modelo obtido a partir das componentes principais mostra mais uma vez que pode eventualmente contornar o problema da separação completa de grupos.

Quadro 4.30 – Estimadores para o MRLI, MRLO e *Bootsratp*, com viés. Conjunto ÁCIDOS GRAXOS.

Função	Variáveis	MRLI	Erro Padrão	MRLO	Erro Padrão	*Bootstrap*	Viés
1	Intercepto	- 98,9055	107,1558	- 103,7708	107,4063	- 119,7021	15,9313
	Palmitic	1,5720	1,3175	1,6041	1,3513	1,8042	- 0,2001
	Stearic	- 0,0941	1,3616	- 0,1358	1,4076	0,0515	- 0,1873
	Oleic	0,9256	1,0708	0,9740	1,0822	1,1699	- 0,1959
	Linoleic	0,9491	1,1029	1,0097	1,0903	1,2093	- 0,1996
	Linolenic	1,2921	1,0750	1,3325	1,0468	1,2283	0,1042
	Eicosanoic	3,5625	4,2976	3,5678	4,0833	1,5486	2,0192
	Eicosenoic	- 1,4939	3,6686	- 1,3585	3,5988	- 0,1944	- 1,1641
2	Intercepto	1,2413	153,8729	13,3611	141,7017	- 51,1813	64,5424
	Palmitic	1,3511	1,474	1,2292	1,3275	1,6593	- 0,4301
	Stearic	1,1142	2,049	1,1061	2,0490	1,1899	- 0,0838
	Oleic	- 0,1526	1,5591	- 0,2768	1,4371	0,4318	- 0,7086
	Linoleic	- 0,2448	1,5960	- 0,3764	1,4627	0,3512	- 0,7276
	Linolenic	- 1,2777	1,8158	- 1,3835	1,6955	- 0,6082	- 0,7753
	Eicosanoic	1,5219	4,1871	1,8051	3,7158	- 0,7015	2,5066
	Eicosenoic	1,5388	4,7567	0,8945	4,1551	3,4550	- 2,5605
3	Intercepto	NE	NE	- 2579,9801	1695,6943	- 191,1927	- 2388,7874
	Palmitic	NE	NE	23,0618	14,9258	2,1674	20,8944
	Stearic	NE	NE	- 14,6793	14,4351	- 1,3957	- 13,2836
	Oleic	NE	NE	23,5866	15,8356	1,7723	21,8143
	Linoleic	NE	NE	30,3859	19,7337	2,2513	28,1346
	Linolenic	NE	NE	22,5048	16,2788	0,7502	21,7546
	Eicosanoic	NE	NE	- 1,4245	8,8502	- 2,2467	0,8222
	Eicosenoic	NE	NE	96,8737	71,8125	9,6328	87,2409
4	Intercepto	408,8112	418,0917	370,5995	375,6840	63,4565	307,1430
	Palmitic	- 6,1084	6,9035	- 5,5093	6,2583	- 0,8131	- 4,6962
	Stearic	- 0,5833	4,9011	- 0,6216	4,8528	- 0,7230	0,1014
	Oleic	- 4,2459	4,3936	- 3,8481	3,9510	- 0,6222	- 3,2259
	Linoleic	- 4,3607	4,4947	- 3,9576	4,0427	- 0,6844	- 3,2732
	Linolenic	- 1,7572	2,8222	- 1,5179	2,6002	- 0,2125	- 1,3054
	Eicosanoic	- 10,6680	10,2203	- 10,0232	9,7806	- 5,2921	- 4,7311
	Eicosenoic	- 2,7562	3,9845	- 2,7033	3,9584	- 0,2655	- 2,4378

É possível observar no Quadro 4.30 que a terceira e a quarta funções, justamente aquelas com maior eficiência, são que apresentam o maior viés para os estimadores, característica que não é observada para as duas primeiras funções, ambas com menor poder discriminante.

Uma nova simulação, desta vez com as variáveis *Palmitic*, *Oleic*, *Linolenic* e *Eicosenoic*, foi realizada e apresentou para o MRLO as taxas de classificação mostradas no Quadro 4.32. Para obter os estimadores *bootstrap* foram geradas 400 amostras. Os estimadores encontrados para o MRLO e para o Modelo *Bootstrap* são mostrados no Quadro 4.33, assim como o respectivo viés.

Quadro 4.31 – Variâncias e autovetores do conjunto ÁCIDOS GRAXOS.

Variável	Autovetores						
	v_1	v_2	v_3	v_4	v_5	v_6	v_7
Palmitic	0,3075	0,1919	0,7754	0,1112	0,0743	- 0,4814	0,1332
Stearic	0,4364	0,1992	0,2536	- 0,3569	0,2118	0,7290	0,0433
Oleic	-0,4253	- 0,2084	0,1760	- 0,5222	- 0,0947	- 0,0182	0,6802
Linoleic	0,4303	0,1190	- 0,4143	0,3513	0,1241	-0,0515	0,6983
Linolenic	- 0,3640	- 0,0018	0,3412	0,6608	- 0,2325	0,4806	0,1713
Eicosanoic	- 0,1670	0,8723	- 0,1226	- 0,1445	- 0,4154	- 0,0472	0,0245
Eicosenoic	- 0,4342	0,3242	- 0,0178	0,0812	0,8359	- 0,0262	0,0104
Variância	3,9092	1,0842	0,9325	0,7866	0,2053	0,0811	0,000098
Percentagem	55,85	15,49	13,32	11,24	2,93	1,16	0,0001

Quadro 4.32 – Matriz de classificações para o MRLO. Conjunto ÁCIDOS GRAXOS.

Grupo Observado	Grupo Previsto				
	1	2	3	4	5
1	0,0909	0,0000	0,0909	0,3636	0,4545
2	0,0000	0,3514	0,0000	0,5946	0,0541
3	0,0000	0,0000	0,9615	0,0000	0,0385
4	0,0000	0,0000	0,0000	0,8000	0,2000
5	0,0000	0,0833	0,0833	0,1667	0,6389

Quadro 4.33 – Estimadores para o MRLI, MRLO e *Bootstrap*, com viés. Conjunto ÁCIDOS GRAXOS, segunda simulação.

Função	**Variáveis**	**MRLO**	**Erro Padrão**	***Bootstrap***	**Viés**
1	Intercepto	- 8,8630	9,7859	- 6,3361	- 2,5269
	Palmitic	0,5749	0,6160	0,5121	0,0628
	Oleic	0,0037	0,1380	- 0,0020	0,0057
	Linolenic	0,4466	0,3165	0,2281	0,2185
	Eicosenoic	- 1,2817	2,0868	- 1,3606	0,0789
2	Intercepto	- 14,4956	5,1628	- 13,8513	- 0,6443
	Palmitic	1,7208	0,5608	1,5035	0,2173
	Oleic	0,0223	0,1136	0,0663	- 0,0440
	Linolenic	- 1,4332	0,5229	- 1,1720	- 0,2612
	Eicosenoic	3,8075	2,6844	3,4154	0,3921
3	Intercepto	17,9257	6,8214	11,8780	6,0477
	Palmitic	- 0,7117	0,6351	- 0,4639	- 0,2478
	Oleic	- 0,4258	0,1882	- 0,2689	- 0,1569
	Linolenic	- 0,5427	0,5280	- 0,5427	0,0000
	Eicosenoic	- 1,1970	2,5302	0,4790	- 1,6760
4	Intercepto	- 5,9831	23,2484	- 3,8587	- 2,1244
	Palmitic	- 0,9499	1,7930	- 0,4007	- 0,5492
	Oleic	0,0643	0,2623	0,0556	0,0087
	Linolenic	1,5165	0,6799	0,6868	0,8297
	Eicosenoic	- 3,9980	2,3275	- 2,4107	- 1,5873

Neste caso também é possível perceber uma significativa redução do viés em relação ao modelo *Bootstrap*, igualmente acompanhada por uma redução da taxa de observações corretamente classificadas pelo MRLO.

4.5 ABORDAGENS INDIVIDUALIZADAS

Com o objetivo de comparar o desempenho dos modelos obtidos através de diferentes abordagens, foi providenciada a divisão do conjunto IRIS em dois outros conjuntos. O primeiro, referido como IRIS – 13, contém as observações pertencentes aos grupos 1 e 3, enquanto o segundo, IRIS – 23, contém as observações dos grupos 2 e 3. Desta forma cada conjunto possui variável resposta binária. A cada um dos conjuntos aplicou-se os Modelos de Regressão Logística Clássico (MRLC) e Oculto (MRLO), ambos para variável resposta binária. Os estimadores obtidos são mostrados nos Quadros 4.34 e 4.35. Considerou-se $\gamma = 0{,}0001$.

Pode-se perceber que os estimadores obtidos individualmente, e também seus erros padrões, são bastante próximos aos obtidos pelos Modelos de Regressão Logística Clássico e Oculto para resposta politômica, mostrados no Quadro 4.4.

Quadro 4.34 – Estimadores para os Modelos de Regressão Logística Clássico (MRLC) e Oculto (MRLO). Conjunto IRIS – 13.

Variável	**MRLC**		**MRLO**	
	Estimador	**Erro Padrão**	**Estimador**	**Erro Padrão**
X_1	NE	NE	0,1390	3,3651
X_2	NE	NE	0,1437	3,2261
X_3	NE	NE	- 0,3689	2,7049
X_4	NE	NE	- 0,3390	4,8980
Intercepto	NE	NE	3,8527	125,6554

Quadro 4.35 – Estimadores para os Modelos de Regressão Logística Clássico (MRLC) e Oculto (MRLO). Conjunto IRIS – 23.

Variável	**MRLC**		**MRLO**	
	Estimador	**Erro Padrão**	**Estimador**	**Erro Padrão**
X_1	0,2465	0,2394	0,2457	0,2373
X_2	0,6681	0,4480	0,6527	0,4380
X_3	- 0,9429	0,4737	- 0,9239	0,4587
X_4	- 1,8286	0,9743	- 1,7853	0,9457
Intercepto	42,6378	25,7077	41,4616	24,8425

Os estimadores obtidos para os modelos correspondentes aos conjuntos Mamografia 13, contendo observações dos grupos 1 e 3, e Mamografia 23, contendo observações dos grupos 2 e 3, são apresentados nos Quadros 4.36 e 4.37, respectivamente.

Quadro 4.36 – Estimadores para MRLC e MRLO. Conjunto Mamografia – 13.

Variável	**MRLC**		**MRLO**	
	Estimador	Erro Padrão	Estimador	Erro Padrão
SYMPT(1)	2,1425	0,4901	2,1409	0,4898
SYMPT(2)	0,3831	0,4858	0,3829	0,4855
SYMPT(3)	0,1423	0,4956	0,1423	0,4953
PB	- 0,2145	0,0766	- 0,2144	0,0766
HIST	1,4153	0,4687	1,4148	0,4686
BSE	1,3998	0,5384	1,3990	0,5382
DETC(1)	- 1,0490	1,1268	- 1,0481	1,1262
DETC(2)	- 0,1974	1,1667	- 0,1970	1,1660
Intercepto	- 2,2568	1,4963	- 2,2553	1,4955

Quadro 4.37 – Estimadores para os Modelos de Regressão Logística Clássico (MRLC) e Oculto (MRLO). Conjunto Mamografia – 23.

Variável	**MRLC**		**MRLO**	
	Estimador	Erro Padrão	Estimador	Erro Padrão
SYMPT(1)	1,1369	0,3626	1,1362	0,3625
SYMPT(2)	0,0709	0,3572	0,0711	0,3571
SYMPT(3)	0,3349	0,3702	0,3346	0,3701
PB	- 0,1447	0,0755	- 0,1446	0,0755
HIST	1,1573	0,4735	1,1569	0,4735
BSE	1,0165	0,5158	1,0159	0,5156
DETC(1)	0,5706	0,6869	0,5705	0,6868
DETC(2)	0,7896	0,7174	0,7894	0,7173
Intercepto	- 2,8282	1,1192	- 2,8269	1,1190

Aqui também é possível notar que os estimadores obtidos são bastante próximos daqueles obtidos pelos modelos com resposta politômica, além da proximidade apresentada pelos modelos entre si. Os resultados encontrados nas duas simulações mostram que esta forma de abordagem pode ser utilizada como uma valiosa ferramenta para testar a consistência das soluções apresentadas pelas abordagens anteriores.

5 CONCLUSÕES

Nos últimos cinco anos tem-se percebido na literatura corrente o surgimento de um número considerável de abordagens alternativas para problemas de Reconhecimento e Classificação de padrões com variável resposta politômica baseadas em Algoritmos Genéticos, Redes Neurais Artificiais, Máquinas de Base Vetorial, e Análise Discriminante através da Programação Linear por Partes, entre outros exemplos. Em comparação com estas técnicas, é possível observar que o Modelo de Regressão Logística é abordado com menor freqüência. Convém destacar neste ponto os trabalhos de O'Brien e Dunson (2004) e Groenewald e Mokgatlhe (2005), ambos com enfoque na inferência Bayesiana. Algumas das técnicas mencionadas acima, com exceção talvez da Análise Discriminante Linear, são criticadas por alguns autores, que apontam desde a ausência de embasamento matemático até a falta de uma explanação melhor detalhada sobre as características e propriedades estatísticas das mesmas, embora não sejam poucos os trabalhos atestando a sua eficiência como modelos discriminantes. Além disso, em alguns campos de pesquisa, como a Medicina, por exemplo, a abordagem padrão é firmemente baseada em métodos estatísticos, como se pode observar pelo grande número de trabalhos publicados por pesquisadores da área. Nota-se também que muitas destas abordagens utilizam o Modelo de Regressão Logística como principal ferramenta de tomada de decisões e também que a maioria das aplicações envolve variável resposta dicotômica. Outro aspecto que merece ser destacado é a complexidade de alguns trabalhos a respeito do modelo em questão, muitas vezes além da compreensão de potenciais usuários que não possuem uma sólida formação em matemática.

A utilização de grandes conjuntos de dados exige certa cautela quanto ao uso de algumas abordagens aqui revisadas, como aquelas propostas por Santner e Duffy (1986) e Christmann e Rousseeuw (2001), por exemplo, principalmente no que diz respeito ao esforço computacional requerido pelas mesmas. Tais fatos devem ser considerados quando da utilização de uma metodologia baseada no Modelo de Regressão Logística e que exija uma verificação prévia da existência de sobreposição de grupos. Também é necessário ter em mente, especialmente em aplicações práticas, que o interesse de alguns pesquisadores concentra-se geralmente na eficiência do método utilizado como modelo discriminante e nos resultados fornecidos, enquanto as questões relativas à existência de estimadores são objetos de estudo de especialistas dedicados a esta área específica. Neste sentido é conveniente dispor de um método que forneça os estimadores procurados, sem sofrer qualquer influência ocasionada por eventuais características dos dados disponíveis e que seja efetivamente útil como ferramenta de apoio à tomada de decisões.

O Modelo de Regressão Logística Oculto (MRLO), aqui estendido para variável resposta politômica, segue rigorosamente a proposta apresentada por Rousseeuw e Christmann (2003), comprovadamente eficaz para problemas com variável resposta dicotômica. A abordagem utilizada consistiu basicamente em uma generalização que possibilita a aplicação do modelo em questão a problemas com variável resposta politômica. As simulações realizadas sobre conjuntos de dados disponíveis na literatura corrente mostram que a eficiência do MRLO não é afetada pelo número de grupos de observações, no sentido de fornecer estimadores para os parâmetros desconhecidos, seja qual for a configuração dos conjuntos de dados, além de não exigir algoritmos complexos ou dispendiosos para sua implementação computacional. Os resultados obtidos na comparação com duas outras técnicas, Função Discriminante Linear e Redes Neurais Artificiais, também podem ser um argumento em favor da viabilidade do MRLO no caso de variável resposta politômica.

As mesmas simulações mostram que o viés de cada estimador é maior para modelos ajustados a conjuntos de dados com pouca ou nenhuma sobreposição, justamente os casos nos quais o MRLO apresenta maiores índices de eficiência, em termos de classificações corretamente efetuadas. Convém lembrar que o viés é avaliado em relação aos Estimadores *Bootstrap*. Por outro lado, este fato mostra que o aumento do viés dos estimadores não afeta o desempenho do modelo obtido, como é possível observar nas matrizes de classificações apresentadas. Desta forma pode-se argumentar que o modelo apresentado pode ser considerado uma ferramenta confiável, além de matematicamente consistente, para a análise e reconhecimento estatístico de dados categorizados.

Os Modelos de Regressão Logística Individualizados (MRLI) mostraram que, embora não sejam imunes à separação completa de grupos, podem ser utilizados em estudos preliminares para fornecer uma descrição mais detalhada dos dados. Ao identificar um grupo totalmente separado, pode-se, por exemplo, removê-lo do conjunto de dados e concentrar a atenção aos grupos remanescentes, o que não é uma informação desprezível em termos de resultados práticos. Além disso, pode-se perceber que nas situações onde não há problemas para a convergência, os resultados obtidos concordam fortemente com aqueles obtidos pelo MRLC e pelo MRLO.

O Modelo de Regressão Logística de Componentes Principais (MRLCP) apresentou desempenho superior aos demais modelos em pelo menos uma das simulações apresentadas. Assim como o MRLO, este modelo foi apresentado como opção de abordagem a problemas com variável resposta dicotômica e, da mesma forma, mostrou-se uma opção igualmente viável para a abordagem

de problemas com variável resposta politômica, embora não seja imune à completa separação de grupos. A despeito deste fato, a possibilidade de escolher o número de componentes principais que integrarão o modelo logístico resultante faz do MRLCP uma ferramenta no mínimo adequada às finalidades aqui tratadas. Além disso, a escolha de diferentes componentes principais permite, como se verificou em alguns casos, contornar o problema causado pela completa separação de grupos, da mesma forma que é possível obter o mesmo efeito para o MRLC e para o MRLI quando se altera a combinação de variáveis explanatórias que devem integrar o modelo obtido.

A implementação computacional de todos os modelos abordados não se mostrou uma tarefa complexa. De fato, os algoritmos necessários podem ser escritos com relativa facilidade, sem contar que algumas sub-rotinas podem ser encontradas em numerosos livros ou trabalhos publicados, como o Método de Newton-Raphson, por exemplo, que é de longe o mais utilizado na estimação de parâmetros para o Modelo de Regressão Logística. A mesma facilidade de acesso à bibliografia é verificada para algoritmos empregados na Análise de Componentes Principais e também para os procedimentos de execução de operações matriciais, de larga utilização na estimação de parâmetros e na referida análise.

A respeito de um dos outros métodos aqui abordados, a Rede Neural Artificial (RNA), é possível notar que a literatura disponível está repleta de argumentos, muitos favoráveis e alguns contrários à sua utilização. Esta discussão indica que o assunto é uma área ainda aberta a pesquisas, tanto no campo teórico como no campo das aplicações práticas. O que não se pode ignorar é a utilidade desta técnica, bem como seu potencial, para a resolução de problemas de reconhecimento estatístico de padrões. Também é justo lembrar que a concepção do MRLO é inspirada na arquitetura, ou topologia, de uma RNA com uma camada oculta, conforme declaram os próprios pesquisadores que apresentaram o modelo em questão. O desempenho, em termos de classificações corretas, da RNA utilizada neste trabalho confirma a sua viabilidade como método discriminante, amplamente divulgada na literatura disponível. A Função Discriminante Linear (FDL), embora criticada por alguns autores por exigir suposições a respeito da matriz de covariâncias, suposições estas não confirmadas na prática, mostrou-se um método de considerável eficiência, além de servir como ponto de partida para o desenvolvimento de outras abordagens, utilizando, por exemplo, a Programação Linear. Além disso, é bem sabido que a FDL mostra-se mais eficiente quando aplicada a conjuntos de dados que contêm variáveis contínuas, mais especificamente variáveis com distribuição normal multivariada.

A utilização do Método *Bootstrap* deve levar em consideração o volume de dados tratados em cada problema, uma vez que o tempo de execução de um algoritmo baseado nesta técnica aumenta na medida em que são incluídas mais observações e, também, mais variáveis explanatórias. Também é conveniente lembrar que os problemas causados pelo viés dos estimadores têm efeito reduzido sobre os resultados, quando as amostras possuem grande número de observações. Neste trabalho optou-se por aplicar o método em questão apenas na estimação dos vieses dos estimadores calculados pelo MRLO por ser este o único modelo imune à completa separação de grupos.

Finalmente, pode-se concluir que os objetivos deste trabalho foram todos alcançados, tendo em vista inicialmente a ampla comparação entre diferentes métodos de estimação de parâmetros, seguida pelo desenvolvimento e implementação computacional dos mesmos, tendo como resultado uma abordagem original no tratamento do problema de reconhecimento e classificação de padrões com variável resposta politômica, através do Modelo de Regressão Logística Oculto e do Modelo de Regressão Logística de Componentes Principais.

REFERÊNCIAS

Aerts, M., Claeskens, G., **Bootstrap tests for misspecified models, with application to clustered binary data**. Computational Statistics and Data Analysis 36, pp. 383-401, 2001.

Aguilera, A. M., Escabias, M., Valderrama, M. J., **Using principal components for estimating logistic regression with high-dimensional multicollinear data**. Computational Statistics and Data Analysis 50, pp. 1905-1924, 2006.

Agresti, A., ***Categorical Data Analysis***. John Wiley & Sons, Inc. Hoboken, New Jersey, 2002.

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

Albert A., Lesaffre, E., **Multiple group logistic discrimination**. Comp. & Maths. with Applic. 12 , pp. 209-224, 1986.

Anderson, J. A., **Separate sample logistic discrimination**. Biometrika 59, pp. 19-35, 1972.

Anderson, J. A., Richardson, S. C., **Logistic discrimination and bias correction in maximum likelihood estimation**. Technometrics, vol. 21, pp. 71-78, 1979.

Begg, C. B., Gray, R., **Calculations of polychotomous logistic regression estimates using individualized regressions**. Biometrika 71, 1, pp. 11-18, 1984.

Brodnjak – Vončina, D., Kodba, Z. C., Novič, M., **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.

Bull, S. B., Greenwood, C. M. T., Hauck, W. W., **Jackknife bias reduction for polychotomous logistic regression**. Statistics in Medicine, 16, 5, pp. 545-560, 1997.

Bull, S. B., Mak, C., Greenwood, C. M. T., **A modified score function estimator for multinomial logistic regression in small samples**. Computational Statistics and Data Analysis 39, pp. 57-74, 2002.

Christmann, A., Rousseeuw, P. J., **Measuring overlap in binary regression**. Computational Statistics and Data Analysis 37, pp. 65-75, 2001.

Copas, J. B., **Binary regression models for contaminated data. With discussion**. Journal of Royal Statistic Society B 50, pp. 225-265, 1988.

Cornfield, J., **Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function approach**. Fed. Amer. Socs. Exper. Biol. Proc. Suppl., 11, pp. 58-61, 1962.

Cox, D. R., **Some procedures associated with the logistic qualitative response curve**. Research Papers on Statistics: Festschrift for J. Neyman, F. N. David (Ed.), New York: Wiley, pp. 55-71, 1966.

Cox, D. R., ***The Analysis of Binary Data***. First edition. London: Methuen, 1970.

Davison, A. C., Hinkley, D. V., ***Bootstraps methods and their application***. Cambridge University Press, 1997.

Day, N. E., Kerridge, D. F., **A general maximum likelihood discriminant**. Biometrics 23, pp. 313-324, 1967.

Desai, V. S., Crook, J. N., Overstreet Jr., G. A. **A comparison of neural networks and linear scoring models in the credit union environment**. European Journal of Operational Research 95,pp. 24-37, 1996.

Dreiseitl, S., Ohno-Machado, L., **Logistic regression and artificial neural network classification models: a methodology review**. Journal of Biomedical Informatics 35, 5-6, pp. 352-359, 2002.

Ekholm, A., Palmgren, J., **A model for binary response with misclassification**. In: Gil-Christ, R. (Ed.), GLIM-82, Proceedings of the International Conference on Generalized Linear Models. Springer, Heidelberg, pp. 128-143, 1982.

Efron, B., **Bootstrap methods: another look at jackknife**. Annals of Statistics 7, pp. 1-26, 1979.

Fausett, L., ***Fundamentals of Neural Networks – Architectures, algorithms, and applications***. New Jersey: Prentice Hall, Inc., 1994.

Faraggi, D., Simon, R., **The maximum likelihood neural network as a statistical classification model**. Journal of Statistical Planning and Inference 46, pp. 93-105, 1995.

Féraud, R., Clérot, F., **A methodology to explain neural network classification**. Neural Networks 15, pp. 237-246, 2002.

Firth, D., **Bias reduction of maximum likelihood estimates**. Biometrika 80, 1, pp. 27-38, 1993.

Fisher, R. A., **The use of multiple measurements in taxonomic problems**. *Annals of Eugenics* 7, pp. 179-188, 1936.

Flury, B., ***A first course in multivariate analysis***. Springer Verlag New York, Inc.,1997.

Freed, N., Glover, F., **Simple but powerful goal programming models for discriminant problems**. European Journal of Operational Research 7, pp. 44-60, 1981.

Gervini, D., **Robust adaptive estimators for binary regression models**. Journal of Statistical Planning and Inference 131, pp, 297-311, 2005.

Groenewald, P. C. N., Mokgatlhe, L., **Bayesian computation for logistic regression**. Computational Statistics and Data Analysis 48, pp. 857-868, 2005.

Guimarães, I.A., Chaves Neto, A. **Reconhecimento de padrões: Comparação de métodos multivariados e Redes Neurais**. Revista Negócios e Tecnologia da Informação 1, 1, pp.38-58, 2006.

Heinze, G., Schemper, M., **A solution to the problem of separation in logistic regression**. Statistic in Medicine 21, pp. 2409-2419, 2002.

Heinze, G., **A comparative investigation of methods for logistic regression with separated or nearly separated data**. Statistics in Medicine 25, 24, pp. 4216-4226, 2006.

Hosmer, D. W., Lemeshow, S., ***Applied logistic regression***. Wiley Interscience, New York, 1989.

Hubert, M., Rousseeuw, P.J., Verboven, S., **A fast method for robust principal components with applications to chemometrics**. Chemometrics Intelligent Laboratory Systems 60, pp. 101-111, 2002.

Hubert, M., Van Driessen, K., **Fast and robust discriminant analysis**. Computational Statistics and Data Analysis 45, pp. 301-320, 2004.

Intrator, O., Intrator, N., **Interpreting neural-network results: a simulation study**. Computational Statistics and Data Analysis 37, pp. 373-393, 2001.

Johnson, R. A., Wichern, D. W. ***Applied multivariate statistical analysis***. 2. ed. New Jersey: Prentice Hall International, Inc., 1988.

Jhun, M., Jeong, H. C., **Applications of bootstrap methods for categorical data analysis**. Computational Statistics and data Analysis 35, pp. 83-91, 2000.

Kodzarkhia, N., Mishra, G. D., Reiersølmoen, L., **Robust estimation in the logistic regression model**. Journal of Statistical Planning and Inference 98, pp. 211-223, 2001.

Kolman, B., ***Introdução à álgebra linear com aplicações***. 6ª ed. Rio de Janeiro. LTC – Livros Técnicos e Científicos Editora Ltda, 1998.

Lachenbruch, P., Mickey, R., **Estimation of error rates in discriminant analysis**. Technometrics 10, pp. 1-11, 1968.

Lam, K. F., Moy, J. W., **A piecewise linear programming approach to the two-group discriminant problem – an adaptation to Fisher's linear discriminant function model**. European Journal of Operational Research 145, pp. 471-481, 2003.

Lesaffre, E., Albert, A., **Partial separation in logistic discrimination**. Journal of Royal Statistical Society B 51, pp. 109-116, 1989.

Massy, W. F., **Principal component regression in explanatory statistic research**. J. Amer. Statist. Assoc. 60, pp. 234-246, 1965.

Menard, S., ***Applied logistic regression analysis***. Sage Publications. Series: Quantitative Applications 106, 1995.

McCulloch, W., Pitts, W. A., **A logical calculus of the ideas immanent in nervous activity**. Bulletin of Mathematical Biophysics, v. 5 , pp. 115-133, 1943.

McLachlan, G. J., ***Discriminant analysis and statistical pattern recognition***. John Wiley & Sons, New York, 1992.

O'Brien, S. M., Dunson, D. B., **Bayesian multivariate logistic regression**. Biometrics 60, pp. 739-746, 2004.

Paixão, L. A., Chaves Neto, A., **Avaliação de óleo isolante em transformadores com o emprego da análise discriminante quadrática**. Artigo aceito para apresentação no XVII Seminário Nacional de Distribuição de Energia Elétrica. Belo Horizonte, Minas Gerais. 21 – 25 de agosto de 2006.

Rom, M., Cohen, A., **Estimation in the polytomous logistic regression model**. Journal of Statistical Planning and Inference 43, pp. 341-353, 1995.

Rousseeuw, P. J., **Least median of squares regression**. Journal of American Statistical Association 79, pp. 871-880, 1984.

Rousseuw, P. J., Van Driessen, K., **A fast algorithm for the minimum covariance determinant estimator**. Technometrics 41, pp. 212-223, 1996.

Rousseeuw, P. J., Struyf, A., **Computing location depth and regression depth in higher dimensions**. Statist. Comput. 8, pp. 193-203, 1998.

Rousseeuw, P. J., Hubert, M., **Regression depth**. Journal of American Statistical Association 94, pp. 388-433, 1999.

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

Santner, T. J., Duffy, D. E., **A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models**. Biometrika 73, 3, pp. 755-758, 1986.

Schumacher, M., Roβner, R., Vach, W., **Neural networks and logistic regression: Part I**. Computational Statistics and Data Analysis 21, pp. 661-682, 1996.

Schwarzer, G., Vach, W., Schumacher, M., **On the misuses of artificial networks for prognostic and diagnostic classification in oncology**. Statistics in Medicine 19, pp. 541-561, 2000.

Silvapulle, M. J., **On the existence of maximum likelihood estimates for the binomial response models**. Journal of Royal Statistical Society B 43, pp. 310-313, 1981

Truett, J., Cornfield, J., Kannel, W.B., **A multivariate analysis of the risk of coronary heart disease in Framingham**. J. Chron. Dis. 20, pp. 511-524, 1967.

Walker, S. H., Duncan, D. B., **Estimation of the probability of an event as a function of several independent variables**. Biometrika 54, pp. 167-169, 1967.

Warner, B., Misra, M., **Understanding neural networks as statistical tools**. The American Statistician 50, 4, pp. 284-293, 1996.

White, H., ***Artificial neural networks: Approximation and learning theory***. Basil Blackwell, Oxford, 1992.

Wilson, R. L., Sharda, R., **Bankruptcy prediction using neural networks**. Decision Support Systems 11, pp. 545-577, 1994.

APÊNDICE I – ANÁLISE DE COMPONENTES PRINCIPAIS

A Análise de Componentes Principais (ACP) é utilizada no estudo da estrutura de variância-covariância através de combinações lineares das variáveis originais. Tem como objetivos a redução de dados e o auxílio à interpretação dos mesmos. De acordo com Johnson e Wichern (1988), a ACP não é uma finalidade em si, mas parte integrante de determinadas abordagens.

Embora o estudo da variabilidade exija p componentes, há situações nas quais boa parte desta variabilidade pode ser resumida por um número k, $k < p$, de componentes principais. Neste caso as p variáveis originais podem ser substituídas por k componentes principais, possibilitando a redução da matriz de dados de ordem $n \times p$ para uma matriz de ordem $n \times k$. A ACP também auxilia a expor relações entre as variáveis, relações estas que podem afetar fortemente os resultados esperados.

Sejam p variáveis aleatórias X_1 , X_2 , ... , X_p , e seja uma matriz de dados contendo n observações das referidas variáveis na forma

$$\begin{bmatrix} x_{11} & x_{21} & \dots & x_{p1} \\ x_{12} & x_{22} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix} \quad . \qquad \text{(AI.1)}$$

As componentes principais são combinações lineares das p variáveis aleatórias. Geometricamente estas combinações lineares podem ser interpretadas como uma mudança do sistema de coordenadas, através da rotação do sistema original, tomando X_1 , X_2 , ... , X_p como eixos coordenados, os quais representam as direções com máxima variabilidade.

Seja o vetor aleatório $\underline{\mathbf{X}} = \begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}^{\mathrm{T}}$, cuja matriz de covariâncias $\sum$ possui autovalores $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. Sejam também as combinações lineares

$$\begin{aligned} Y_1 &= l_1^{\mathrm{T}} \underline{\mathbf{X}} = l_{11} X_1 + \dots + l_{p1} X_p \\ &\dots \\ Y_p &= l_p^{\mathrm{T}} \underline{\mathbf{X}} = l_{1p} X_1 + \dots + l_{pp} X_p \end{aligned} \quad . \qquad \text{(AI.2)}$$

Então as variâncias e as covariâncias são dadas, respectivamente, por

$$Var(Y_i) = l_i^{\mathrm{T}} \Sigma l_i \tag{AI.3}$$

$$Cov(Y_i, Y_k) = l_i^{\mathrm{T}} \Sigma l_k \tag{AI.4}$$

As componentes principais são as combinações lineares não correlacionadas e cujas variâncias são tão grandes quanto possível. A primeira componente principal é a combinação linear que maximiza

$$Var(Y_1) = l_1^{\mathrm{T}} \Sigma l_1$$

Como a expressão pode ser alterada pela multiplicação por qualquer constante, pode-se eliminar tal indeterminação restringindo-se o problema a vetores unitários, isto é

$$\begin{aligned} &Max \quad l_1^{\mathrm{T}} \Sigma l_1 \\ &s.a \quad\; l_1^{\mathrm{T}} l_1 = 1 \end{aligned} \quad . \tag{AI.5}$$

A segunda componente principal, que não deve ser correlacionada com a primeira, é dada por

$$\begin{aligned} &Max \quad\; l_2^{\mathrm{T}} \Sigma l_2 \\ &s.a \quad\; l_2^{\mathrm{T}} l_2 = 1 \\ &\quad\; Cov(Y_1, Y_2) = 0 \end{aligned} \quad . \tag{AI.6}$$

A i – ésima componente principal é dada por

$$\begin{aligned} &Max \quad\; l_i^{\mathrm{T}} \Sigma l_i \\ &s.a \quad\; l_i^{\mathrm{T}} l_i = 1 \\ &\quad\; Cov(Y_i, Y_k) = 0 \,,\; k < i \end{aligned} \quad . \tag{AI.7}$$

Seja $\sum$ a matriz de covariâncias associada ao vetor $\underline{\mathbf{X}} = \left[X_1 \quad X_2 \quad ... \quad X_p\right]^{\mathrm{T}}$. Os pares de autovalores–autovetores de $\sum$ são $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$, onde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. A i – ésima componente principal é dada por

$$Y_i = e_i^{\mathrm{T}} \underline{\mathbf{X}} = e_{1i} X_1 + e_{2i} X_2 + \ldots + e_{pi} X_p \quad , \quad i = 1 , 2 , ... , p. \tag{AI.8}$$

Então

$$Var(Y_i) = e_i^{\mathrm{T}} \Sigma e_i = \lambda_i \quad , \quad i = 1 , 2 , ... , p. \tag{AI.9}$$

$$Cov(Y_i , Y_k) = e_i^{\mathrm{T}} \Sigma e_k = 0 \, , \, i \neq k \quad . \tag{AI.10}$$

Se $\lambda_i = \lambda_k$, $i \neq k$, há mais de uma opção para a escolha de e_i , e conseqüentemente Y_i não é única. Johnson e Wichern (1988) demonstram que as componentes principais são não correlacionadas e têm variâncias iguais aos autovalores de $\sum$.

Sejam Y_1 , Y_2 , ... , Y_p componentes principais associadas ao vetor $\underline{\mathbf{X}} = \left[X_1 \quad X_2 \quad ... \quad X_p\right]^{\mathrm{T}}$, com matriz de covariâncias $\sum$. Então

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} Var(Y_i). \tag{AI.11}$$

Isto significa que a variância total da população é

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \lambda_1 + \lambda_2 + \ldots + \lambda_p \, ,$$

de modo que a proporção da variância total explicada pela j – ésima componente principal é

$$\Pi_j = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \quad . \tag{AI.12}$$

Se uma grande proporção, algo entre 80% e 90%, pode ser atribuída a m, $m < p$, componentes principais, então estas m componentes podem substituir as p variáveis originais sem muita perda de informação.

Finalmente, o coeficiente de correlação entre uma componente principal Y_i e uma variável X_k é dado por:

$$\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad , \quad i, k = 1, \ldots, p. \tag{AI.13}$$

APÊNDICE II – MÉTODO *BOOTSTRAP*

INTRODUÇÃO

Esta técnica foi apresentada por Efron (1979) e é indicada para estimar a distribuição amostral de estatísticas a fim de medir a sua dispersão, entre outras aplicações. Deve ser usada quando não se dispõe de resultado analítico ou, quando existir, seja fortemente assintótico. Assim, é importante em especial quando se trabalha com amostras de tamanhos reduzidos. O método consiste em gerar um grande número de amostras com reposição, da amostra original, usando a função de distribuição empírica dos dados originais (ou resíduos de um modelo, ou outro procedimento). As amostras assim geradas podem posteriormente ser utilizadas na construção de uma estimativa da distribuição amostral da estatística de interesse. E, a partir dessa distribuição amostral *bootstrap* pode-se obter, com base em percentís, intervalos de confiança *bootstrap* com determinada probabilidade de cobertura para parâmetros, avaliação da variabilidade da estatística que estima um parâmetro, entre outras aplicações. A utilização dos dados originais para gerar mais dados lembra o truque utilizado pelo fictício Barão de Munchäusen, que conseguiu salvar-se de um naufrágio puxando a si mesmo pelos cadarços de suas botas, em inglês *bootstraps*. Cabe ressaltar que o principal objetivo do Método Bootstrap não é obter um aumento das informações trazidas pela amostra original, mas conseguir uma nova visão dos mesmos. A seguir apresenta-se uma introdução mais formal, adaptada de Davison e Hinkley (1997).

Seja uma amostra aleatória $\{x_1 , x_2 , ... , x_n\}$ das variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) $X_1 , X_2 , ... , X_n$, cujas funções densidades de probabilidades (p.d.f.'s) e funções distribuições acumuladas (c.d.f.'s) podem ser representadas, respectivamente, por $f(.)$ e $F(.)$. A referida amostra é utilizada para fazer inferências a respeito de um parâmetro θ, utilizando a estatística T, cujo valor amostral é t. As questões referentes à distribuição de probabilidade de T podem ser o valor do viés, o valor do erro padrão ou o intervalo de confiança para θ usando T.

Há duas situações distintas, a paramétrica e a não paramétrica. A primeira ocorre quando é possível usar um modelo matemático contendo constantes ajustáveis, ou parâmetros, para determinar $f(.)$. Neste caso o parâmetro θ é uma componente, ou função de Ψ. Quando não existe tal modelo, diz-se que a análise é não paramétrica e utiliza-se apenas o fato de que as variáveis

aleatórias X_j são i.i.d. Esta última pode ser útil na análise da robustez das conclusões obtidas através da análise paramétrica.

Na análise não paramétrica é muito importante o uso da distribuição empírica, que atribui a cada valor amostral x_j a mesma probabilidade n^{-1}. O estimador de F é a função distribuição empírica (EDF), denotada por $\hat{F}$, definida como a proporção amostral

$$\hat{F}(x) = \frac{\#\{x_j < x\}}{n}$$

ou, de modo mais formal, como

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} H(x - x_i) \qquad \text{(AII.1)}$$

onde $H(\,.\,)$ é a função passo unitária, ou seja

$$H(u) = \begin{cases} 0 \;, & u \le 0 \\ 1 \;, & 0 < u \end{cases} .$$

Os valores da EDF são fixados como $(0\ ,\ 1/n\ ,\ 2/n\ ,\ \ldots\ ,\ n/n)$. Deste modo a EDF é equivalente a esses pontos de acréscimo, os valores ordenados $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. Quando há valores repetidos na amostra, ocorrência comum em conjuntos de dados discretos, a EDF fixa probabilidades proporcionais à freqüência amostral dos mesmos.

SIMULAÇÃO PARAMÉTRICA

Seja um conjunto de dados $\{x_1\ ,\ x_2\ ,\ \ldots\ ,\ x_n\}$, para o qual há um modelo paramétrico. A CDF e a PDF podem ser denotadas por $F_0(x)$ e $f_0(x)$, respectivamente. Quando φ é estimado por $\hat{\varphi}$, quase sempre o estimador de máxima verossimilhança, tal substituição fornece o modelo ajustado, com CDF $\hat{F}(x) = F_{\hat{\varphi}}(x)$, que pode ser usado para calcular propriedades de T, algumas vezes de modo exato. Aqui se usa X^* para denotar a variável aleatória distribuída conforme o modelo ajustado.

Cálculos teóricos com o modelo ajustado podem ser muito complexos, além disso, algumas aproximações podem não ser disponíveis, ou confiáveis, até mesmo em função do tamanho reduzido da amostra. Neste caso uma possível alternativa é a estimação das propriedades a partir de conjuntos de dados simulados. Tais conjuntos podem ser denotados por X_1^* , X_2^* , ... , X_n^* , e são independentemente amostrados da distribuição ajustada $\hat{F}$. A estatística de interesse calculada a partir de um conjunto simulado é representada por T^*. Das R repetições obtém-se T_1^* , T_2^* , ... , T_R^*. Deste modo as propriedades de $(T - \theta)$ são estimadas a partir de T_1^* , T_2^* , ... , T_R^*. O estimador do viés de T, p.ex., dado por

$$b(F) = E(T \mid F) - \theta \quad ,$$

pode ser obtido como

$$B = b(\hat{F}) = E(T \mid \hat{F}) - t = E^*(T^*) - t \, .$$

Além disso

$$B_R = \frac{1}{R}\sum_{r=1}^{R} T_r^* - t = \overline{T}^* - t \, . \qquad \text{(AII.2)}$$

O parâmetro para o modelo é t, tal que $(T^* - t)$ é análogo a $(T - \theta)$. O estimador para a variância de T é dado por

$$V_R = \frac{1}{R-1}\sum_{r=1}^{R} \left(T_r^* - \overline{T}^*\right)^2 \, . \qquad \text{(AII.3)}$$

De acordo com Davison e Hinkley (1997) as aproximações acima são justificadas pela lei dos grandes números.

SIMULAÇÃO NÃO PARAMÉTRICA

Em algumas aplicações o modelo paramétrico pode ser desconhecido, embora seja sensato assumir que X_1 , X_2 , ... , X_n são i.i.d. de acordo com uma função distribuição F desconhecida. Neste

caso é possível usar a EDF $\hat{F}$ para estimar a CDF *F*, como se fosse um modelo paramétrico. Cálculos teóricos são possíveis em poucos casos, embora seja possível obter boas aproximações.

Uma vez que a EDF fixa probabilidades iguais para os dados $\{x_1 , x_2 , ... , x_n\}$, cada X^* é amostrado independentemente destes valores. Além disso, as amostras simuladas X_1^* , X_2^* , ... , X_n^* formam uma amostra aleatória tomada com reposição do mesmo conjunto. Esta modalidade de reamostragem é chamada *bootstrap* não paramétrica.

INTERVALOS DE CONFIANÇA

Uma das principais aplicações para um estimador *T* é o cálculo de limites de intervalos de confiança para o parâmetro θ. Em geral utiliza-se a aproximação normal para a distribuição de *T*, com média dada por (θ + β) e variância ν, onde β é o viés de *T*. Se β e ν são conhecidos, pode-se escrever

$$P(T \le t \mid F) \cong \Phi\left[\frac{t-(\theta+\beta)}{\nu^{1/2}}\right]$$

onde Φ(.) é a distribuição normal padrão. Se o quantil α da distribuição normal padrão é dado por $z_\alpha = \Phi^{-1}(\alpha)$, então um intervalo de confiança (1 – 2α) é

$$P\left(\beta + \nu^{1/2} z_\alpha \le T - \theta \le \beta + \nu^{1/2} z_{1-\alpha}\right) \cong 1 - 2\alpha . \qquad \text{(AII.4)}$$

Como na prática o viés e a variância não são conhecidos, ambos devem ser substituídos por estimadores. Tanto β como ν podem ser expressos como

$$\beta = b(F) = E(T \mid F) - t(F) \quad \text{e} \quad \nu = var(T \mid F) . \qquad \text{(AII.5)}$$

Supondo que *F* é estimada por $\hat{F}$, os referidos estimadores podem ser obtidos mediante a substituição de *F* por $\hat{F}$, isto é

$$\hat{\beta} = b(\hat{F}) = E(T \mid \hat{F}) - t(\hat{F}) \quad \text{e} \quad \hat{\nu} = var(T \mid \hat{F}) . \qquad \text{(AII.6)}$$

De acordo com Davison e Hinkley (1997), a aproximação normal para a obtenção dos estimadores em questão não apresenta problemas para grandes amostras. Caso o tamanho da amostra seja reduzido, a aproximação normal pode mostrar-se inadequada.

Se a distribuição de $(T - \theta)$ pode ser aproximada pela distribuição de $(T^* - t)$, então as probabilidades acumuladas podem ser estimadas pela EDF dos valores simulados $(t^* - t)$, ou seja, se $G(u) = P(T - \theta \leq u)$, então o estimador para $G(u)$ é dado por

$$\hat{G}_R(u) = \frac{1}{R}\sum_{r=1}^{R} I\left[t_r^* - t \leq u\right] . \qquad \text{(AII.7)}$$

onde I[E] é a função indicadora do evento E, igual a 1 quando E é verdadeiro e 0 quando E é falso. A aproximação (AII.7) contém duas fontes de erro, uma entre $\hat{G}$ e G, em função da variabilidade dos dados, e outra entre $\hat{G}_R$ e $\hat{G}$, devida a simulação finita.

Se forem utilizados estimadores *bootstrap* dos quantis para $(T - \theta)$, então um intervalo de confiança $(1 - 2\alpha)$ terá limites dados por

$$t - \left(t^*_{(R+1)(1-\alpha)} - t\right) \quad , \quad t - \left(t^*_{(R-1)\alpha} - t\right) . \qquad \text{(AII.8)}$$

Os limites acima são chamados limites de confiança *bootstrap* básicos, e sua acurácia depende do número R, de amostras *bootstrap*, e da concordância da distribuição de $(T^* - t)$ com a distribuição de $(T - \theta)$.

MODELOS DE REGRESSÃO

Conforme Efron (1979), um modelo de regressão geralmente é dado por

$$X_i = g_i(\underline{\mathbf{B}}) + \varepsilon_i \qquad \text{(AII.9)}$$

onde $g(\ .\ )$ é uma função conhecida do vetor de parâmetros $\underline{\mathbf{B}}^{T} = [\ \beta_1\ ,\ ...\ ,\ \beta_p\ ]$, enquanto $\varepsilon_i \sim_{ind} C$, $i = 1\ ,\ ...\ ,\ n$.

Normalmente, a informação que se tem a respeito de C é que está centrada em zero, talvez $E_C(\varepsilon) = 0$ ou $Mediana_C(\varepsilon) = 0$. A partir de uma amostra observada para X utiliza-se algum método para estimar $\underline{\mathbf{B}}$, geralmente o Método dos Mínimos Quadrados, ou seja,

$$\hat{\beta} : \min_{\beta} \sum_{i=1}^{n} \left[x_i - g_i(\beta)\right]^2 \quad , \tag{AII.10}$$

com o objetivo de obter alguma informação sobre a distribuição amostral de $\underline{\hat{\mathbf{B}}}$.

A aplicação do Método *Bootstrap* pode ser efetuada pela definição de $\hat{C}$ como distribuição de probabilidade amostral dos resíduos $\hat{\varepsilon}_i$, isto é

$$\hat{C} : mass\ \frac{1}{n} \quad para \quad \hat{\varepsilon}_i = x_i - g_i(\hat{\beta})\ .$$

De acordo com Efron (1979), se alguma componente de $\underline{\mathbf{B}}$ é um parâmetro de posição para $g(\ .\ )$, então $\hat{C}$ tem média igual a zero. Caso contrário, e se a suposição de que $E_C(\varepsilon) = 0$ é bastante plausível, pode-se modificar $\hat{C}$ de modo a obter a média desejada. As amostras *bootstrap* são dadas por

$$X_i^* = g_i(\hat{\beta}) + \varepsilon_i^* \quad . \tag{AII.11}$$

Para cada amostra aplica-se o mesmo método de estimação, então

$$\hat{\beta}^* : \min_{\beta} \sum_{i=1}^{n} \left[x_i^* - g_i(\beta)\right]^2$$

As amostras *bootstrap* obtidas podem então ser utilizadas para estimar a distribuição amostral de $\hat{\beta}^*$.