RT-MAE 2005-24

Diagnóstico em análise discriminante.

Reigada, S. M. B.
and
Elian, S. N.

Palavras-Chave: diagnóstico, análise discriminante, pontos influentes, distância de Mahalanobis, probabilidade de classificação incorreta.
Classificação AMS: 62H30.

- Agosto de 2005 -

Diagnóstico em análise discriminante

Sueli Maria Beltrame Reigada[1]

Silvia Nagib Elian[2]

[1] **E-mail: smbreigada@uol.com.Br**

[2] **Departamento de Estatística**
Instituto de Matemática e Estatística,
Universidade de São Paulo
C.P. 66281 – São Paulo, SP Brasil
E-mail: selian@ime.usp.br

Resumo

Este trabalho apresenta medidas de diagnóstico em análise discriminante em duas situações principais: para duas populações com matriz de covariância constante e duas populações com matrizes de covariância desiguais (análise discriminante quadrática). É estudada ainda a abordagem Bayesiana para a detecção de observações influentes. No final do trabalho examina-se também as medidas de diagnóstico para a análise discriminante múltipla, ou seja, quando se tem mais que duas populações. O estudo é complementado com a aplicação das principais medidas estudadas a um conjunto de dados reais.

Palavras-chave: diagnóstico, análise discriminante, pontos influentes, distância de Mahalanobis, probabilidade de classificação incorreta.

1. Introdução

A literatura sobre análise discriminante bem como sobre diagnóstico é muito vasta, mas a pesquisa na determinação de pontos influentes em análise discriminante ainda é pouco explorada.

O estatístico frequentemente se depara com o problema de como detectar e tratar observações aparentemente atípicas. Esse problema fica mais complicado quando se analisa dados multivariados, uma vez que um *outlier* multivariado não é fácil de ser definido, dificultando a sua identificação. Pode surgir de um erro grosseiro em um de seus componentes ou de pequenos erros em todos eles. Além disso, o *outlier* multivariado pode distorcer medidas de locação e escala como por exemplo, vetor de médias e matriz de covariâncias.

Essa complexidade faz com que os procedimentos de detecção sejam elaborados com o cuidado de se proteger contra certos tipos de problemas como por exemplo, distorções da correlação, em que a presença do *outlier* pode indicar que as variáveis são relacionadas quando na realidade não o são. Além do mais, deve-se ter em mente que uma observação julgada discrepante para um propósito, pode ser considerada normal para outro.

O objetivo deste trabalho é apresentar as medidas de diagnóstico em análise discriminante definidas pelos principais pesquisadores do assunto nos últimos 27 anos.

Essas medidas indicam pontos que podem estar influenciando na regra de alocação ou na probabilidade de classificação incorreta, ou em ambos, cabendo ao pesquisador a decisão de mantê-los ou não em seu estudo.

O presente artigo está esquematizado da seguinte forma: na Seção 2 serão apresentadas as medidas de diagnóstico em análise discriminante para duas populações e matriz de covariância constante. Inicialmente são abordadas as medidas baseadas na função de influência de Hampel (1974). Em seguida, são apresentadas duas estatísticas fundamentais das quais várias medidas dependem. Medidas relacionadas com a probabilidade de classificação incorreta também são mencionadas.

A Seção 3 trata da detecção de observações influentes no contexto Bayesiano. Sob essa abordagem, são definidas medidas de influência globais, locais e individuais.

Na Seção 4, são apresentadas as medidas de diagnóstico utilizadas na análise discriminante quadrática, que é utilizada geralmente quando as matrizes de covariância são diferentes, apesar de ser uma análise menos robusta.

A Seção 5 se destina a abordar as medidas que detectam observações influentes em análise discriminante múltipla, ou seja, quando existem mais que duas populações.

Uma aplicação das medidas de diagnóstico a um conjunto de dados reais encontra-se na Seção 6 e a Seção 7 finaliza o trabalho apresentando algumas conclusões.

2. Diagnóstico em análise discriminante com duas populações e matriz de covariância constante

2.1 Função de influência em análise discriminante

Na análise de diagnóstico, quando o procedimento estatístico adotado envolve estimação, uma estratégia comum na determinação de pontos influentes é proceder ao ajuste com e sem o ponto suspeito e comparar as estimativas obtidas. Nessa linha de estudo, Campbell (1978) sugeriu o uso da função de influência, proposta por Hampel (1974).

O objetivo principal da função de influência teórica é calcular a influência de um particular ponto x no parâmetro de interesse, de modo que valores altos para essa função indicariam que x tem grande influência no parâmetro.

Em análise de grupos multivariados, caso em que o parâmetro estudado envolve mais que uma população, a função de influência é determinada excluindo uma observação de apenas um grupo.

Campbell (1978) considera o caso em que o vetor aleatório X tem distribuição normal p-variada com média μ_j e matriz de covariância Σ, para X pertencendo à população π_j, j = 1, 2, e utiliza f unção de influência definida a seguir

" Suponha g populações com funções distribuição $F_1, F_2, ..., F_k, ..., F_g$, seja $\theta = T(F_1, F_2, ..., F_k, ..., F_g)$, um parâmetro geral, obtido a partir das funções de distribuição F_k, k = 1, ..., g. Para um valor x do vetor aleatório X, a função de influência $I(x;\theta)$ é definida como:

$$I(x;\theta) = \lim_{\varepsilon \to 0}\left(\frac{\tilde{\theta}-\theta}{\varepsilon}\right) \quad \text{em que} \quad 0<\varepsilon<1, \quad \tilde{\theta} = T(\tilde{F}) \quad , \quad \tilde{F} = (F_1, ..., \tilde{F}_k, ..., F_g)$$

e $\tilde{F}_k = (1-\varepsilon)F_k + \varepsilon\,\delta_x$ é a função distribuição F_k após sofrer perturbação no ponto x, sendo δ_x uma função de distribuição que assume probabilidade 1 no ponto x."

Considera-se inicialmente perturbações em Σ^{-1} e μ_j, sendo que $\Sigma = w_1\Sigma_{F_1} + w_2\Sigma_{F_2}$. Σ_{F_1} é a matriz de covariâncias para a população π_1, Σ_{F_2} é a matriz de covariâncias para a população π_2, $w_1 + w_2 = 1$ e $w_j > 0$. Os pesos w_1 e w_2 são adotados uma vez que as amostras podem ter tamanhos diferentes, então $w_1 = \frac{n_1}{n_1+n_2}$ e $w_2 = \frac{n_2}{n_1+n_2}$.

Tomando a função linear discriminante $y = l'x = \delta'\Sigma^{-1}x$ em que $\delta = \mu_1 - \mu_2$, e a distância de Mahalanobis $\Delta^2 = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$, calcula-se os parâmetros após a exclusão de um elemento da população π_1, (ver Reigada (2005, pg. 8)), e usando a notação * sobrescrito para os parâmetros após perturbação tem-se:

$$\mu_1^* = \mu_1 + \varepsilon(x-\mu_1) = \mu_1 + \varepsilon z \qquad \text{com} \quad z = x - \mu_1$$

$$\delta^* = \delta + \varepsilon z \quad , \quad \Sigma^* = (1-\varepsilon w_1)\Sigma + \varepsilon w_1 zz' \quad \text{e} \quad \Sigma^{-1^*} \equiv (1+\varepsilon w_1)\Sigma^{-1} - \varepsilon w_1 \Sigma^{-1} zz'\Sigma^{-1}.$$

Para obter a função de influência para a distância de Mahalanobis $\Delta^2 = \delta'\Sigma^{-1}\delta$, obtém-se primeiramente $\Delta^{2*} \equiv (1+\varepsilon w_1)\Delta^2 + 2\varepsilon\phi - \varepsilon w_1\phi^2$ em que $\phi = \delta'\Sigma^{-1}\mathbf{z}$, e a partir daí, a função de influência $I(\mathbf{x};\Delta^2) = w_1\Delta^2 + 2\phi - w_1\phi^2$.

Pode-se calcular a função de influência para as médias discriminantes $\mathbf{l}'\boldsymbol{\mu}_j$: $I(\mathbf{x};\mathbf{l}'\boldsymbol{\mu}_1) = w_1\mathbf{l}'\boldsymbol{\mu}_1 + \phi - w_1\phi\,\eta_1 + \eta_1$ em que $\eta_1 = \mathbf{z}'\Sigma^{-1}\boldsymbol{\mu}_1$, e para os coeficientes da função discriminante $\mathbf{l} = \Sigma^{-1}\delta$: $I(\mathbf{x};\mathbf{l}) = w_1\mathbf{l} + (1 - w_1\phi)\Sigma^{-1}\mathbf{z}$, mas como essa quantidade é um vetor, a fim de se obter um valor numérico calcula-se a função de influência para $\mathbf{l}'\mathbf{l}$: $I(\mathbf{x};\mathbf{l}'\mathbf{l}) = 2w_1\mathbf{l}\mathbf{l}' + 2(1 - w_1\phi)\mathbf{l}'\Sigma^{-1}\mathbf{z}$ (Reigada (2005), pg. 14 e 15).

Sendo $\mathbf{l} = \Sigma^{-1}\delta$ o vetor de coeficientes discriminantes, ϕ pode ser escrito como $\mathbf{l}'\mathbf{z}$ que é equivalente à diferença entre o escore discriminante $\delta'\Sigma^{-1}\mathbf{x}$ e sua média discriminante para a primeira população, $\mathbf{l}'\boldsymbol{\mu}_1$:

$$\phi = \mathbf{l}'\mathbf{z} = \delta'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) = \delta'\Sigma^{-1}\mathbf{x} - \delta'\Sigma^{-1}\boldsymbol{\mu}_1 = \delta'\Sigma^{-1}\mathbf{x} - \mathbf{l}'\boldsymbol{\mu}_1.$$

Calculando a esperança e variância de ϕ, para $\mathbf{x}$ pertencendo à população π_1 que foi perturbada, tem-se: $E(\phi) = E(\delta'\Sigma^{-1}\mathbf{z}) = 0$ e $Var(\phi) = \Delta^2$, portanto $\phi \sim N(0, \Delta^2)$, e $\phi_p = \frac{\phi}{\Delta}$ tem distribuição normal padrão. Tomando $\phi = \phi_p\Delta$, a função de influência para Δ^2 pode ser escrita em termos de ϕ_p como: $I_p(\mathbf{x};\Delta^2) = w_1\Delta^2 + 2\Delta\phi_p - w_1\Delta^2\phi_p^2$.

A função de influência pode ser considerada como uma variável aleatória, uma vez que é uma transformação do vetor aleatório $\mathbf{X}$.

Se $\mathbf{X}$ tem distribuição normal multivariada, está demonstrado na Seção A-2 do Apêndice que a distribuição de probabilidades da função de influência para a distância de Mahalanobis (Δ^2) tem assimetria negativa.

Tomando a função de influência para a distância de Mahalanobis com ϕ não padronizado, $I(\mathbf{x};\Delta^2) = w_1\Delta^2 + 2\phi - w_1\phi^2$ e derivando em relação a ϕ, obtém-se o ponto de máximo para $\phi = \frac{1}{w_1}$, assim, $I_{max}(\mathbf{x};\Delta^2) = w_1\Delta^2 + \frac{1}{w_1}$.

Campbell (1978) sugere o uso de $I_m(\mathbf{x};\Delta^2) = I_{max}(\mathbf{x};\Delta^2) - I(\mathbf{x};\Delta^2)$, dada por $I_m(\mathbf{x};\Delta^2) = w_1^{-1}(1 - 2w_1\Delta\phi_p + w_1^2\Delta^2\phi_p^2)$.

Esta variável aleatória é sempre positiva e, como $I(\mathbf{x};\Delta^2)$ tem assimetria negativa, segue que $I_m(\mathbf{x};\Delta^2)$ tem distribuição positivamente assimétrica.

Multiplicando e dividindo $I_m(\mathbf{x};\Delta^2)$ por ϕ_p^2 obtém-se $I_m(\mathbf{x};\Delta^2) = w_1\Delta^2(\phi_p - w_1^{-1}\Delta^{-1})^2$.

Como $\phi_p \sim N(0,1)$ então $\left(\phi_p - w_1^{-1}\Delta^{-1}\right) \sim N_\phi\left(-w_1^{-1}\Delta^{-1}, 1\right)$ e usando o Resultado A-3 do Apêndice, segue que $\left(\phi_p - w_1^{-1}\Delta^{-1}\right)^2 = \left(\phi_p - w_1^{-1}\Delta^{-1}\right)1^{-1}\left(\phi_p - w_1^{-1}\Delta^{-1}\right)$ tem distribuição quiquadrado não central com 1 grau de liberdade e parâmetro de não centralidade $\left(w_1^{-1}\Delta^{-1}\right)^2$.

Portanto $\frac{I_m\left(x;\Delta^2\right)}{w_1\Delta^2}$ tem uma distribuição quiquadrado não central com 1 grau de liberdade e parâmetro de não centralidade $\left(w_1^2\Delta^2\right)^{-1}$.

Observa-se então que a distribuição nula de $I_m\left(x;\Delta^2\right)$ é quiquadrado com um grau de liberdade, sendo assim a distribuição esperada na inexistência de pontos discrepantes.

2.2 Principais estatísticas no diagnóstico em análise discriminante

Fung (1992) apresentou duas estatísticas fundamentais no diagnóstico em análise discriminante. Uma delas é a medida ϕ apresentada na seção anterior. Tomando a regra discriminante de Fisher em que uma observação x é alocada na população π_1 se $(\bar{x}_1 - \bar{x}_2)'S^{-1}x \geq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'S^{-1}\bar{x}_1 + \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'S^{-1}\bar{x}_2$, chega-se a $\hat{\phi} \geq -\frac{1}{2}D^2$, em que $\hat{\phi} = (\bar{x}_1 - \bar{x}_2)'S^{-1}(x - \bar{x}_1)$ e D é a estimativa para a distância de Mahalanobis.

Portanto, a medida ϕ é importante tanto na discriminação quanto na análise de influência.

A segunda medida é $d_{ji}^2 = (x_{ji} - \bar{x}_j)'S^{-1}(x_{ji} - \bar{x}_j)$ $i = 1, \cdots, n_j$, $j = 1, 2$. Healy (1968) utiliza essa estatística em testes de normalidade e também com o objetivo de detectar *outliers* sob normalidade.

Assim como o resíduo e o ponto de alavanca em regressão, Fung (1992) aponta d_{ji}^2 e $\hat{\phi}_{ji} = \hat{l}'(x_{ji} - \bar{x}_j)$ como estatísticas básicas para detectar *outliers* e observações influentes em análise discriminante, uma vez que várias medidas de diagnóstico são escritas em função dessas duas estatísticas.

Para facilitar a notação serão usados os termos d_i^2 e ϕ_i nas próximas seções. Fung (1995-a) mostra que, assintoticamente, d_i^2 é distribuída como χ_p^2 (p = dimensão do vetor X) e $\frac{\hat{\phi}_i}{D}$ tem distribuição $N(0,1)$. Esses resultados assintóticos são muito importantes uma vez que calculando-se estas estatísticas para os elementos amostrais, valores maiores que os respectivos valores críticos da distribuição quiquadrado e da distribuição normal são indicadores de pontos influentes, ver Reigada (2005, pg. 25).

2.3 Medidas relacionadas com a probabilidade de classificação incorreta

Em análise discriminante com duas populações, a probabilidade de classificação incorreta (PCI) é definida como $PCI = P(1/2)(1-q) + P(2/1)\,q$ em que

$P(i/j)$ = probabilidade de classificar uma observação em π_i quando ela é de π_j

q = probabilidade *a priori* de π_1

$1-q$ = probabilidade *a priori* de π_2.

Quando as probabilidades *a priori* são iguais, $PCI = \frac{1}{2}P(2/1)+\frac{1}{2}P(1/2)$. Sabe-se que $PCI = \Phi\left(\frac{-\Delta}{2}\right)$, em que $\Phi(.)$ é a função de distribuição acumulada da distribuição normal padrão.

Fung(1992) propõe uma medida de diagnóstico com base na diferença das estimativas das probabilidades de classificação incorreta definida como

$DPCI_i = \left[\Phi\left(-\frac{D_{(i)}}{2}\right)\right]-\left[\Phi\left(-\frac{D}{2}\right)\right]$ em que $D_{(i)}$ é a estimativa para a distância de Mahalanobis sem a observação i do grupo 1.

Após vários cálculos obtém-se $DPCI_i \cong \frac{\Phi\left(-\frac{1}{2}D\right)}{4D(n_1-1)^2}\left[\left(1-w_1\hat{\phi}_i\right)^2\left(d_i^2-\hat{\phi}_i^2/D^2\right)+\frac{1}{4}\hat{\phi}_i^2\right]$.

que depende dos valores de $\hat{\phi}_i$ e d_i^2.

Critchley e Vitiello (1991), independentemente de Fung(1992) determinaram duas estatísticas principais que compõe as medidas de influência. São elas a "atipicalidade", que exerce o papel da medida d_i^2 e a diferença entre o escore discriminante linear da observação e o escore discriminante linear de sua média amostral, que corresponde à medida ϕ_i. Os autores determinaram uma medida de influência com o objetivo principal de considerar o efeito em Δ^2 produzido pela eliminação de uma única observação i da amostra proveniente de π_1. Desse modo, chegam à medida

$\hat{\Delta}^2_{1(i)} = \left(\frac{n_1+n_2-1}{n_1+n_2}\right)\left[\hat{\Delta}^2-\frac{1}{w_1(n_1-1)}+\frac{\left\{e(x_{1i})-w_1^{-1}\right\}^2}{w_1^{-1}(n_1-1)-\hat{\alpha}_1^2(x_{1i})}\right]$, que está relacionada com a probabilidade de classificação incorreta visto que $PCI = \Phi\left(\frac{-\Delta}{2}\right)$.

2.4 Medidas baseadas na média quadrática

Na análise discriminante com duas populações, uma boa regra de alocação é aquela que minimiza o custo médio de classificação incorreta ou seja, consiste em alocar x_0 (de origem desconhecida) em π_1 se $o(x_0) = \frac{\Pr(x_0 \in \pi_1 / d)}{\Pr(x_0 \in \pi_2 / d)} > \frac{c(1/2)}{c(2/1)}\cdot\frac{1-q}{q}$, e em π_2 caso contrário,

sendo que $c(i/j)$ = custo de classificar uma observação em π_i quando de fato ela pertence a π_j, i, $j = 1, 2$, $i \neq j$. Aplicando a função logaritmo obtém-se

$$(\mu_1 - \mu_2)'\Sigma^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left[\frac{c(1/2)}{c(2/1)} \cdot \frac{1-q}{q}\right].$$

Tomando essa regra discriminante de Fisher, assumindo os custos de classificação incorreta iguais e usando estimativas das quantidades desconhecidas, aloca-se x_0 em π_1 se :

$\hat{l}'\left[x_0 - \frac{(\bar{x}_1 + \bar{x}_2)}{2}\right] > \ln\left(\frac{1-q}{q}\right)$. A quantidade $\hat{l}'\left[x_0 - \frac{(\bar{x}_1 + \bar{x}_2)}{2}\right] - \ln\left(\frac{1-q}{q}\right)$ é denominada logaritmo da razão das probabilidades estimadas (*log-odds*). Na maioria dos casos, admite-se $1 - q = q$, então essa quantidade reduz-se à função discriminante para a observação x_0.

Fung (1995-a) propôs a medida $E(DLO_i)^2$ que é a média do quadrado da diferença entre os *log-odds* da amostra toda e da amostra sem a observação i :

$$E\left(\hat{\gamma}'y - \hat{\gamma}'_{(i)}y\right)^2 \quad \text{em que} \quad \hat{\gamma}'y - \hat{\gamma}'_{(i)}y = \left(-\hat{l}'\frac{(\bar{x}_1 + \bar{x}_2)}{2} + \hat{l}'x\right) - \left(-\hat{l}'_{(i)}\frac{(\bar{x}_{1(i)} + \bar{x}_2)}{2} + \hat{l}'_{(i)}x\right).$$

Para esse cálculo foi utilizado $E(DLO_i)^2 = Var(DLO_i) + E^2(DLO_i)$, chegando-se a

$$E(DLO_i)^2 = (w_1 B_1 + w_2 B_2)^2 + V$$

em que $B_1 = \frac{A_1}{2} - \frac{A_2}{2}$ e $B_2 = \frac{-A_1}{2} - \frac{A_2}{2}$

$$A_1 = \frac{D^2}{n-2} + cf_i\left(\hat{\phi}_i\right)^2 + ch\hat{\phi}_i - chf_i d_i^2\hat{\phi}_i$$

$$A_2 = c.\ h.\ \left(\hat{\phi}_i - h d_i^2\right)g_i$$

$$V = \frac{D^2}{(n-2)^2} + c\left(f_i\hat{\phi}_i + hg_i\right)\left[\frac{2\hat{\phi}_i}{n-2} + \left(f_i\hat{\phi}_i + hg_i\right)cd_i^2\right]$$

$$c = \frac{n-3}{n-2}, \quad f_i = \frac{b}{1+bd_i^2}, \quad b = \frac{-n_1}{(n_1-1)(n-2)}, \quad h = \frac{1}{n_1 - 1}, \quad g_i = \frac{1}{1+bd_i^2}.$$

Essa medida também é função de $\hat{\phi}_i$ e d_i^2, e detecta observações com grande influência no logaritmo da razão das probabilidades estimadas (escore discriminante quando $1 - q = q$).

Adicionalmente, Fung (1996-b) obteve medidas de diagnóstico condicionais ao fato que a observação x está próxima do plano discriminante, uma vez que essas observações são difíceis de classificar . Fung (1998) mostra que a medida condicional tomada para probabilidades *a priori* iguais é assintoticamente proporcional à medida *DPCI*, proposta em 1992 pelo mesmo autor.

3. Detecção de observações influentes no contexto Bayesiano

3.1 Densidades preditivas e divergência de Kullback-Leibler

Utilizando densidades preditivas, que são a base da análise discriminante Bayesiana, estuda-se, nesta seção, a influência que as observações têm sobre aspectos inferenciais tais como alocação de futuras observações, separação entre populações e determinação da probabilidade de uma observação pertencer a certa população. Para esse estudo são analisadas estatísticas de influências definidas por Johnson (1987). Em adição à notação já utilizada, sejam

o vetor de parâmetros (μ_1, μ_2, Σ), que será denominado por θ,

X_{ji} com distribuição normal p-variada condicional aos parâmetros,

d, definido como o conjunto de todos os dados observados X_{ji}, $j = 1, 2$, $i = 1, 2 \ldots n_j$,

$p(\theta/\pi_j)$, $j = 1,2$, a densidade *a priori* para θ sob π_j,

$p(\theta/d, \pi_j)$, densidade *a posteriori* para θ sob π_j e

Y vetor aleatório de observações futuras, independente do conjunto de dados,

com distribuição condicional aos parâmetros $N_p(\mu_j, \Sigma)$ sob π_j.

No enfoque Bayesiano, o problema de previsão de observações futuras é resolvido utilizando-se $f(y/d)$, que é a densidade *a posteriori* da quantidade a ser prevista, Y, dada a amostra observada d.

Essa densidade, denominada densidade preditiva *a posteriori* de Y dado d, é definida como sendo a esperança *a posteriori* da densidade amostral para Y:

$$f_j(y) = f(y/d, \pi_j) = \int N(y/\theta, \pi_j) p(\theta/d, \pi_j) d\theta, \quad j = 1, 2,$$

esperança essa tomada com relação à distribuição *a posteriori* de θ dada a amostra d.

Se o vetor de parâmetros θ é conhecido, a densidade preditiva para Y (sob π_j) será a densidade amostral $N(./\theta, \pi_j)$.

Quando θ é desconhecido, Johnson (1987) obtém densidades preditivas aproximadas, substituindo os parâmetros da densidade normal pelas correspondentes estimativas. Outra possibilidade seria o uso da *priori* de referência

$$p(\theta/\pi_j) \propto |\Sigma|^{-\frac{1}{2}(p+1)} \quad j = 1, 2, \tag{3.1}$$

sendo que, com base nessa distribuição, Geisser (1964, 1966) obteve a densidade preditiva:

$$Y/x, \pi_j \sim St\left(\nu, \bar{x}_j, \left(1 + n_j^{-1}\right)\tilde{S}\right)$$

que é a distribuição *t-Student* p-variada com ν graus de liberdade, vetor de locação $\bar{x}_j$ e matriz de dispersão $\left(1 + n_j^{-1}\right)\tilde{S}$, em que

$$\nu = n - p - 1, \quad \bar{x}_j = \sum_{i=1}^{n_j} \frac{x_{ji}}{n_j} \quad \text{e} \quad \tilde{S} = \nu^{-1} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)(x_{ji} - \bar{x}_j)' = \nu^{-1}(n-2)S.$$

Com o objetivo de avaliar a discrepância entre duas densidades preditivas, Johnson (1987) usou a medida de divergência de Kullback-Leibler (1951):

$$I(f_i, f_j) = E_{f_i} \ln\left(f_i f_j^{-1}\right) = \int_{\Re^p} f_i(y) \ln\left(\frac{f_i(y)}{f_j(y)}\right) dy .$$

Se f_i é a distribuição $N(\mu_i, \Sigma_i)$, verifica-se que:

$$2I(f_i, f_j) = (\mu_i - \mu_j)' \Sigma_j^{-1} (\mu_i - \mu_j) + tr\Sigma_i \Sigma_j^{-1} - \ln\left|\Sigma_i \Sigma_j^{-1}\right| - p .$$

Quando f_i é a densidade *t-Student* com η_i graus de liberdade, vetor de locação α_i e matriz de dispersão λ_i, não é possível calcular $I(f_i, f_j)$ explicitamente. Face à complexidade na solução das integrais envolvidas no cálculo, aproxima-se tal densidade por uma densidade normal multivariada apropriada.

Johnson (1987) propõe aproximar a densidade *t-student* por uma densidade normal com parâmetros μ_j e Σ_j de modo a minimizar $I(St(\eta_j, \alpha_j, \lambda_j), N(\mu_j, \Sigma_j))$. Verifica-se que o ínfimo dessa expressão é atingido para $\mu_j = \alpha_j$ e $\Sigma_j = \frac{\eta_j}{\eta_j - 2}\lambda_j$.

Dessa forma, a estimativa normal ótima para a densidade *t-Student* $St(\eta_j, \alpha_j, \lambda_j)$ é

$$\tilde{f}_j = N\left(\alpha_j, \frac{\eta_j}{(\eta_j - 2)}\lambda_j\right).$$

Do ponto de vista de inferência, a correta alocação do vetor p-variado y, bem como a determinação de $p(y)$, depende do grau de separação entre as populações, assim, esses objetivos estão relacionados entre si.

Quando o objetivo é alocar y em π_1 ou π_2, a regra ótima consiste em alocar y em π_1 se:

$$o(y) = \frac{\Pr(y \in \pi_1 / d)}{\Pr(y \in \pi_2 / d)} > \frac{c(1/2)}{c(2/1)} \cdot \frac{1-q}{q} \text{ . Verifica-se que, } o(y) = \frac{q}{(1-q)} \cdot \frac{f_1(y)}{f_2(y)}.$$

Quando os parâmetros são conhecidos, utiliza-se

$$\ln(o(y)) = \ln\frac{q}{(1-q)} + (\mu_1 - \mu_2)' \Sigma^{-1} y - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2),$$

que é a função discriminante de Fisher quando as probabilidades q e $1 - q$ forem iguais.

Se θ é desconhecido, os parâmetros são substituídos por suas estimativas. Utilizada a *priori* de referência (3.1) calcula-se o logaritmo da razão das probabilidades estimadas (*log odds*) *a posteriori* como:

$$\ln(o(y)) = \ln\frac{q}{1-q} + \ln\left(\frac{St(y / v, \bar{x}_1, \kappa_1 \tilde{S})}{St(y / v, \bar{x}_2, \kappa_2 \tilde{S})}\right) \text{ com } \kappa_1 = 1 + n_1^{-1} \quad \text{e} \quad \kappa_2 = 1 + n_2^{-1}.$$

3.2 Medida de influência global, local e individual

Os conceitos de medida de influência global, local e individual foram definidos por Johnson (1987). Uma medida de influência é dita global quando avalia o quanto uma observação afeta dois ou mais aspectos da análise. Uma vez que a densidade preditiva

$f_j(\mathbf{y}/\mathbf{d},\pi_j)=\int N(\mathbf{y}/\theta,\pi_j)p(\theta/\mathbf{d},\pi_j)d\theta$ envolve vários elementos – funções densidade, probabilidades, dados, parâmetros – analisando o efeito que as observações têm sobre a densidade preditiva, obtém-se medidas de influência globais.

Conforme descrito na Seção 3.1, para medir o efeito de eliminar a observação x_i, $i = 1, 2, ..., n_j$, na j-ésima densidade preditiva, Johnson (1987) emprega a medida de divergência simétrica de Kullback-Leibler. Na impossibilidade de calcular $J(f_j, f_{j(i)})$ explicitamente, é necessário usar aproximações.

Se utilizada a *priori* de referência (3.1), a densidade preditiva obtida é $St(\eta_j,\alpha_j,\lambda_j)$, em que $\eta_j = n-p-1$, $\lambda_j = \left(1+n_j^{-1}\right)\tilde{S}$ e $\alpha_j = \bar{x}_j$. A aproximação normal ótima para esta densidade é $N\left(\alpha_j, \frac{\eta_j}{(\eta_j-2)}\lambda_j\right)$, sendo que $\alpha_j = \bar{x}_j$, e a matriz de covariância é aproximada por $\frac{n-2}{n-p-3}S$.

A medida de influência é local quando mede o efeito de uma observação em um objetivo inferencial específico, tal como por exemplo, alguma estatística de separação. Para medir o efeito de uma observação no logaritmo da razão das probabilidades estimadas, Johnson (1987) considerou a esperança do quociente dessas medidas na presença e ausência da observação. Tal medida, denominada esperança do logaritmo da razão das probabilidades estimadas é definida como:

$ELO_i = E\ \ln\left\{\frac{o(\mathbf{y})}{o_{(i)}(\mathbf{y})}\right\}$ sendo que a esperança é calculada em relação à densidade preditiva com todos os dados, $f(\mathbf{y}) = qf_1(\mathbf{y}) + (1-q)f_2(\mathbf{y})$ e $o(\mathbf{y}) = \frac{q}{(1-q)}\frac{f_1(y)}{f_2(y)}$.

Se a estatística ELO_i for positiva, indica que a propensão para alocar uma futura observação y em π_1 é menor depois de eliminada a observação i.

Para medir o efeito da observação i no logaritmo da razão das probabilidades estimadas amostral, o autor considera o logaritmo da razão das probabilidades estimadas relativo amostral, definido como:

$$LORAM_i = n^{-1}\sum_{j=1}^{2}\sum_{k=1}^{n_j}\ln\left\{\frac{o(\mathbf{x}_{jk})}{o_{(i)}(\mathbf{x}_{jk})}\right\},\quad i = 1, ..., n_1.$$

4. Observações influentes em análise discriminante quadrática

Nesta seção, são propostas medidas de diagnóstico em análise discriminante quadrática, caso em que as matrizes de covariância são diferentes.

Partindo de duas populações normais p-variadas com vetor de médias μ_j e matriz de covariância Σ_j, e considerando amostras aleatórias dessas duas populações, sejam $\bar{x}_j$, $\hat{\Sigma}_j$, $j = 1, 2$ os estimadores não viciados dos vetores de médias e matrizes de covariância.

Assumindo probabilidades *a priori* e custos de classificação incorreta iguais para as duas populações, a regra discriminante quadrática aloca uma observação x_0 (de origem desconhecida) para o grupo com maior estimativa da densidade, ou seja aloca x_0 em π_1 se

$$(x_0 - \bar{x}_1)' \hat{\Sigma}_1^{-1} (x_0 - \bar{x}_1) + \log|\hat{\Sigma}_1| < (x_0 - \bar{x}_2)' \hat{\Sigma}_2^{-1} (x_0 - \bar{x}_2) + \log|\hat{\Sigma}_2| \qquad (4.1)$$

e caso contrário, aloca x_0 em π_2.

Uma vez que a regra discriminante quadrática (4.1) depende da função densidade, analisa-se o efeito das observações nas densidades estimadas.

Considerando-se a eliminação da observação l do grupo 1, têm-se as estimativas $\bar{x}_{1(l)}$ e $\hat{\Sigma}_{1(l)}$, sem essa observação, produzindo a densidade $\hat{f}_{1(l)}$, normal multivariada com média $\bar{x}_{1(l)}$ e matriz de covariância $\hat{\Sigma}_{1(l)}$. Nesta análise, as estimativas de parâmetros da população 2 não são afetadas.

Fung (1996-a) propõe o uso de medidas que levem em conta a estrutura da análise discriminante quadrática, ou seja, a probabilidade estimada que uma observação x pertença ao grupo j determinada por $\hat{p}_j(x) = \dfrac{\hat{f}_j(x)}{\hat{f}_1(x) + \hat{f}_2(x)}$, para $j = 1, 2$.

Estuda-se o efeito da observação l nessa probabilidade através do cálculo de $E_x\left[\hat{p}_j(x) - \hat{p}_{j(l)}(x)\right]^2$ $j = 1, 2$, em que $\hat{p}_{j(l)}(x)$ é a probabilidade estimada que uma observação x pertença ao grupo j depois de eliminada a observação l. Calculando-se essa esperança com base na distribuição empírica, define-se a medida de diagnóstico de probabilidade quadrática relativa :

$$PQR_l = \frac{1}{n} \sum_{j=1}^{2} \sum_{k=1}^{n_j} \left[\hat{p}_1(x_{jk}) - \hat{p}_{1(l)}(x_{jk})\right]^2, \quad l = 1, 2, \cdots, n_1.$$

Para medir a influência da l-ésima observação no logaritmo da razão das probabilidades estimadas, a medida proposta por Fung (1996-a) é:

$$LOQR_l = \frac{1}{n} \sum_{j=1}^{2} \sum_{k=1}^{n_j} \left[\log\left\{\frac{\hat{p}_1(x_{jk})}{\hat{p}_2(x_{jk})}\right\} - \log\left\{\frac{\hat{p}_{1(l)}(x_{jk})}{\hat{p}_{2(l)}(x_{jk})}\right\} \right]^2 \text{ que pode ser escrita como}$$

$$LOQR_l = \frac{1}{n} \sum_{j=1}^{2} \sum_{k=1}^{n_j} R_j^2(x_{jk}) \text{ em que}$$

$$2R_j(x_{jk}) = -\log|\hat{\Sigma}_1| - (x_{jk} - \bar{x}_1)' \hat{\Sigma}_1^{-1} (x_{jk} - \bar{x}_1) + \log|\hat{\Sigma}_{1(l)}| + (x_{jk} - \bar{x}_{1(l)})' \hat{\Sigma}_{1(l)}^{-1} (x_{jk} - \bar{x}_{1(l)})$$

O termo $R_j^2(x_{jk})$ pode ser escrito como $\left[\log \hat{f}_1(x) - \log \hat{f}_{1(l)}(x)\right]^2$, devido ao cancelamento da estimativa da densidade para o grupo 2, que não é afetado. Sendo assim, é mais razoável propor medidas de diagnóstico para o grupo da observação eliminada como

$$LOQRG_i = \frac{1}{n_1}\sum_{k=1}^{n_1}\left[\log \hat{f}_1(\mathbf{x}_{1k}) - \log \hat{f}_{1(i)}(\mathbf{x}_{1k})\right]^2 = \frac{1}{n_1}\sum_k R_i^2(\mathbf{x}_{1k}) \qquad \text{em que}$$

$$2R_i(\mathbf{x}_{1k}) = -\log\left|\hat{\Sigma}_1\right| - (\mathbf{x}_{1k}-\bar{\mathbf{x}}_1)'\hat{\Sigma}_1^{-1}(\mathbf{x}_{1k}-\bar{\mathbf{x}}_1) + \log\left|\hat{\Sigma}_{1(i)}\right| + (\mathbf{x}_{1k}-\bar{\mathbf{x}}_{1(i)})'\hat{\Sigma}_{1(i)}^{-1}(\mathbf{x}_{1k}-\bar{\mathbf{x}}_{1(i)}).$$

5. Observações influentes em análise discriminante para g grupos

5.1 Observações influentes na probabilidade estimada de uma observação pertencer a determinado grupo

Nesta seção são apresentadas as medidas que detectam observações influentes em análise discriminante múltipla, ou seja, quando há mais que duas populações.

A estimativa da probabilidade de uma observação de origem desconhecida pertencer a um certo grupo tem papel importante em análise discriminante. Fung (1995-b) sugeriu quatro medidas de diagnóstico para avaliar o quanto estas estimativas podem ser afetadas pelas observações.

Seja $\mathbf{x}_{jk}$, $j = 1, 2, \ldots, g$, $k = 1, \ldots, n_j$ uma amostra de tamanho $n = n_1 + n_2 + \ldots + n_g$ selecionada aleatoriamente de g populações, de modo que $\mathbf{x}_{jk}$ tem distribuição normal p-variada com média μ_j e matriz de covariância Σ.

A média amostral do grupo j é denotada por $\bar{\mathbf{x}}_j$ e a estimativa da matriz de covariância é $\hat{\Sigma} = \sum_j\sum_k (\mathbf{x}_{jk}-\bar{\mathbf{x}}_j)(\mathbf{x}_{jk}-\bar{\mathbf{x}}_j)'/(n-g)$.

Se q_j é a probabilidade *a priori* de que uma observação $\mathbf{x}_0$ de origem desconhecida venha da população j, $j = 1,2, \ldots, g$, a estimativa da probabilidade de que $\mathbf{x}_0$ pertença ao grupo j é:

$$\hat{p}_j(\mathbf{x}_0) = \frac{q_j\hat{f}_j(\mathbf{x}_0)}{\sum_j q_j\hat{f}_j(\mathbf{x}_0)} \qquad j = 1,\cdots,g \quad \text{ou seja}$$

$$\hat{p}_j(\mathbf{x}_0) = \frac{q_j \exp[-(\mathbf{x}_0-\bar{\mathbf{x}}_j)'\hat{\Sigma}^{-1}(\mathbf{x}_0-\bar{\mathbf{x}}_j)/2]}{\sum_j q_j \exp[-(\mathbf{x}_0-\bar{\mathbf{x}}_j)'\hat{\Sigma}^{-1}(\mathbf{x}_0-\bar{\mathbf{x}}_j)/2]} \qquad j = 1, 2, \cdots, g. \qquad (5.1)$$

Essa quantidade é muito importante uma vez que, quando os custos de classificação incorreta são iguais, a regra de classificação é obtida alocando-se $\mathbf{x}_0$ no grupo j tal que $\hat{p}_j(\mathbf{x}_0)$ é máxima.

Quando os custos não forem iguais, aloca-se $\mathbf{x}_0$ na população π_j com $\sum_{\substack{i=1\\ i\neq j}}^{g} q_i p_i(\mathbf{x}) c(j/i)$ mínima.

Para determinar elementos influentes no cálculo de $\hat{p}_j(\mathbf{x})$ será usado novamente o método da omissão de observações. Se $\mathbf{x}_{ri}$, i-ésima observação do grupo r, $r = 1, 2, ..., g$, $i = 1, 2, ..., n_r$, é a observação eliminada da amostra, a média do r-ésimo grupo fica:

$\bar{\mathbf{x}}_{r(ri)} = \sum_{\substack{j=1 \\ j \neq i}}^{n_r} \frac{\mathbf{x}_{rj}}{n_r - 1}$ e as demais médias não são afetadas. A matriz de covariância resultante é denotada como $\hat{\Sigma}_{(ri)}$.

Após a retirada de $\mathbf{x}_{ri}$, a probabilidade estimada, $\hat{p}_j(\mathbf{x}_0)$, é calculada através da expressão (5.1) usando as estimativas para a amostra reduzida. Indicando esta estimativa por $\hat{p}_{j(ri)}(\mathbf{x})$, a influência da observação retirada é calculada como $\hat{p}_j(\mathbf{x}_0) - \hat{p}_{j(ri)}(\mathbf{x}_0)$ e com base nela, Fung (1995-b) propõe as seguintes medidas de diagnóstico:

a) diferença absoluta das probabilidades estimadas

$$DAP_{ri} = \sum_{j=1}^{g} \sum_{k=1}^{n_j} \frac{\left| \hat{p}_j(\mathbf{x}_{jk}) - \hat{p}_{j(ri)}(\mathbf{x}_{jk}) \right|}{n}, \quad r = 1, 2, ..., g, \quad i = 1, 2, ..., n_r$$

b) diferença quadrática das probabilidades estimadas

$$DQP_{ri} = \sum_{j=1}^{g} \sum_{k=1}^{n_j} \frac{\left[\hat{p}_j(\mathbf{x}_{jk}) - \hat{p}_{j(ri)}(\mathbf{x}_{jk}) \right]^2}{n}, \quad r = 1, 2, ..., g, \quad i = 1, 2, ..., n_r$$

c) diferença absoluta dos logaritmos das probabilidades estimadas

$$DALP_{ri} = \sum_{j=1}^{g} \sum_{k=1}^{n_j} \frac{\left| \log \hat{p}_j(\mathbf{x}_{jk}) - \log \hat{p}_{j(ri)}(\mathbf{x}_{jk}) \right|}{n}, \quad r = 1, 2, ..., g, \quad i = 1, 2, ..., n_r$$

d) diferença quadrática dos logaritmos das probabilidades estimadas

$$DQLP_{ri} = \sum_{j=1}^{g} \sum_{k=1}^{n_j} \frac{\left[\log \hat{p}_j(\mathbf{x}_{jk}) - \log \hat{p}_{j(ri)}(\mathbf{x}_{jk}) \right]^2}{n}, \quad r = 1, 2, ..., g, \quad i = 1, 2, ..., n_r.$$

Na definição destas medidas, foi usado o estimador não viciado para Σ. No entanto, Fung (1995-b) verificou que, se usadas outras estimativas como a de máxima verossimilhança ou estimativas Bayesianas, os resultados seriam similares.

Os diagnósticos de influência propostos medem a influência total, pois são definidos sobre todas as observações da amostra, mas pode-se também usar medidas para um grupo específico, por exemplo,

$$DQLPG_{j(ri)} = \sum_{k=1}^{n_j} \frac{\left[\log \hat{p}_j(\mathbf{x}_{jk}) - \log \hat{p}_{j(ri)}(\mathbf{x}_{jk}) \right]^2}{n_j},$$

que mede o efeito da eliminação da observação $\mathbf{x}_{ri}$ no logaritmo da probabilidade estimada para os elementos do grupo j, $j = 1, ..., g$. As outras medidas também podem ser definidas desse modo.

Em um estudo de simulação, Fung (1995-b) observou que as medidas quadráticas *DQP* e *DQLP* fornecem resultados similares entre si e melhores do que as medidas absolutas *DAP* e *DALP*, e as medidas *DALP* e *DQLP*, que utilizam logaritmos, são superiores. Sendo assim, a

medida *DQLP* seria preferível dentre as quatro medidas propostas. Foram usados também diferentes conjuntos de probabilidades *a priori*, e os resultados foram próximos, sendo que a análise com as probabilidades *a priori* iguais se aproximou mais do conjunto de probabilidades *a priori* real, o que indica o uso de probabilidades *a priori* iguais quando não se tem informações sobre as probabilidades *a priori* verdadeiras.

5.2 Medidas baseadas na probabilidade de classificação incorreta

Prosseguindo sua pesquisa em análise discriminante múltipla, Fung (1996-c) calculou a probabilidade de classificação incorreta, sob a seguinte regra de classificação:

Um vetor aleatório x_0, de origem desconhecida, é alocado em π_j se $x_0 \in R_j$, sendo que

$$R_j = \bigcap_{\substack{k=1 \\ k \neq j}}^{g} [2(\mu_k - \mu_j)'\Sigma^{-1}x_0 + \mu_j'\Sigma^{-1}\mu_j - \mu_k'\Sigma^{-1}\mu_k < 0] \qquad (5.2)$$

que é uma forma alternativa de definir a regra discriminante de Fisher.

A probabilidade de classificação incorreta para $x_0 \in \pi_j$, $j = 1, 2, ..., g$, sob a regra de alocação (5.2) é

$$PCI_j = 1 - G(b_{j1}, \cdots, b_{j,j-1}, b_{j,j+1}, \cdots, b_{jg})$$

em que $b_{jk} = \frac{\Delta_{jk}}{2}$, $\Delta^2_{jk} = (\mu_j - \mu_k)'\Sigma^{-1}(\mu_j - \mu_k)$ é a distância de Mahalanobis entre as populações j e k, para $k = 1, 2, ..., g$ e $k \neq j$

$G(b_{j1}, \cdots, b_{j,j-1}, b_{j,j+1}, \cdots, b_{jg})$ é a função distribuição de uma normal multivariada, de modo que

$$G(b_{j1}, \cdots, b_{j,j-1}, b_{j,j+1}, \cdots, b_{jg}) = P\left[\bigcap_{k \neq j} z_k < b_{jk}\right],$$

$z_k \sim N(0, 1)$ e $\text{cov}(z_k, z_l) = \frac{(\Delta^2_{jk} + \Delta^2_{jl} - \Delta^2_{kl})}{2\Delta_{jk}\Delta_{jl}}$, $k \neq j$, $l \neq j$.

Note-se que a probabilidade de classificação incorreta é função somente das distâncias entre as médias populacionais.

Uma vez que o critério mais comum para estudar a influência de uma observação é o método da omissão, seja x_{ri} a observação i do grupo r eliminada da amostra. Sob essa amostra reduzida obtém-se nova regra de classificação bem como diferentes estimativas para os parâmetros e para a probabilidade de classificação incorreta que fica $PCIE_{(ri)} = \sum_{j=1}^{g} \frac{PCIE_{j(ri)}}{g}$

em que $PCIE_{j(ri)}$ é a probabilidade de classificação incorreta estimada para um elemento do grupo j eliminando-se a observação i, do grupo r. O efeito dessa observação na estimativa da probabilidade de classificação incorreta é avaliado pela diferença $PCIE - PCIE_{j(ri)}$.

Finalizando, Fung (1996-c) obtém a função de influência para a probabilidade de classificação incorreta.

Se a r-ésima distribuição é perturbada, $r = 1, 2, \ldots, g$, então a função de influência para Δ^2_{jk} é dada por

$$I(\mathbf{x};\Delta^2_{jk}) = \begin{cases} w_r\Delta^2_{jk} - w_r\phi^2_{jk} & j \neq r, k \neq r \\ w_r\Delta^2_{jk} + 2\phi_{jk} - w_r\phi^2_{jk} & j = r \\ w_r\Delta^2_{jk} - 2\phi_{jk} - w_r\phi^2_{jk} & k = r \end{cases} \qquad (5.3)$$

em que $\phi_{jk} = (\mu_j - \mu_k)\Sigma^{-1}\mathbf{z}$ e $\mathbf{z} = \mathbf{x} - \mu_r$.

O autor apresenta a função de influência para a probabilidade de classificação incorreta para um elemento da população j com a r-ésima distribuição perturbada, $I(\mathbf{x}; PCI_j)$, escrevendo-a a partir de $I(\mathbf{x};\Delta^2_{jk})$

$$I(\mathbf{x}; PCI_j) = -\sum_{\substack{k=1 \\ k\neq j}}^{g} \frac{\delta\, G(b_{j1},\cdots,b_{j,j-1},b_{j,j+1},\cdots,b_{jg})}{\delta\, b_{jk}} \cdot \frac{I(\mathbf{x};\Delta^2_{jk})}{4\Delta_{jk}}$$

em que $I(\mathbf{x};\Delta^2_{jk})$ é dado em (5.3) e $\dfrac{\delta\, G(\cdot)}{\delta\, b_{jk}}$ é a derivada parcial de

$$G(\cdot) = P\left(\bigcap_{\substack{l=1 \\ l\neq j \\ l\neq k}} z_l < b_{jl}\right), \text{ em que } z_l \sim N(0,1) \text{ e } \operatorname{cov}(z_l, z_k) = \frac{\left(\Delta^2_{jl} + \Delta^2_{jk} - \Delta^2_{lk}\right)}{2\Delta_{jl}\Delta_{jk}}, \quad l \neq j, k \neq j.$$

6. Aplicação das medidas em um conjunto de dados

O objetivo desta seção é o da aplicação das medidas apresentadas a um conjunto de dados reais. Para esse fim foi utilizado um subconjunto do conjunto de dados obtido no Centro de Estatística Aplicada (CEA) do Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP). Trata-se dos dados referentes ao Relatório de Análise Estatística (RAE-CEA-9415) sobre o projeto: Biologia de uma comunidade de marsupiais e roedores, em floresta atlântica de montanha, no Parque Estadual da Serra do Tabuleiro, Santa Catarina, Brasil, (Leite, Singer e Lourenço (1994)).

Em cada animal capturado, foram medidas 26 características. Na análise discriminante feita no projeto, foram testados 8 conjuntos de funções discriminantes com suas respectivas variáveis. O modelo adotado foi o que apresentou melhor relação custo/benefício pois utilizava apenas 4 medidas, com 1% de animais classificados incorretamente, e se baseia na informação de 88 animais, 13 do grupo 1, de marsupiais, e 75 do grupo 2, de roedores. As quatro variáveis são comprimento total (mm), comprimento da cauda (mm), comprimento do pé (mm) e comprimento do pelo do dorso (mm).

Assumindo que os dados têm distribuição normal multivariada, a análise de diagnóstico feita em Reigada (2005) iniciou-se com o teste de homogeneidade de matrizes de covariâncias

(Morrison (1976, pg 252)) para determinar a medida mais adequada ao conjunto de dados. Toda a programação foi feita no software S-plus.

Uma vez que o teste apontou para matrizes de covariância diferentes, utilizou-se primeiramente a medida LOQRG, para análise discriminante quadrática:

$$LOQRG_i = \frac{1}{n_1}\sum_{k=1}^{n_1}\left[\log \hat{f}_1(\mathbf{x}_{1k}) - \log \hat{f}_{1(i)}(\mathbf{x}_{1k})\right]^2 = \frac{1}{n_1}\sum_{k} R_i^2(\mathbf{x}_{1k})$$

em que

$$2R_i(\mathbf{x}_{1k}) = -\log\left|\hat{\Sigma}_1\right| - (\mathbf{x}_{1k} - \bar{\mathbf{x}}_1)'\hat{\Sigma}_1^{-1}(\mathbf{x}_{1k} - \bar{\mathbf{x}}_1) + \log\left|\hat{\Sigma}_{1(i)}\right| + (\mathbf{x}_{1k} - \bar{\mathbf{x}}_{1(i)})'\hat{\Sigma}_{1(i)}^{-1}(\mathbf{x}_{1k} - \bar{\mathbf{x}}_{1(i)}).$$

Tal medida apontou as observações de números 3, 8 e 13 do grupo 1 e 6, 9, 10, 20, 21, 29, 41, 67 e 73 do grupo 2 como as mais influentes no logaritmo da razão das probabilidades estimadas, conforme Figura 6.1.

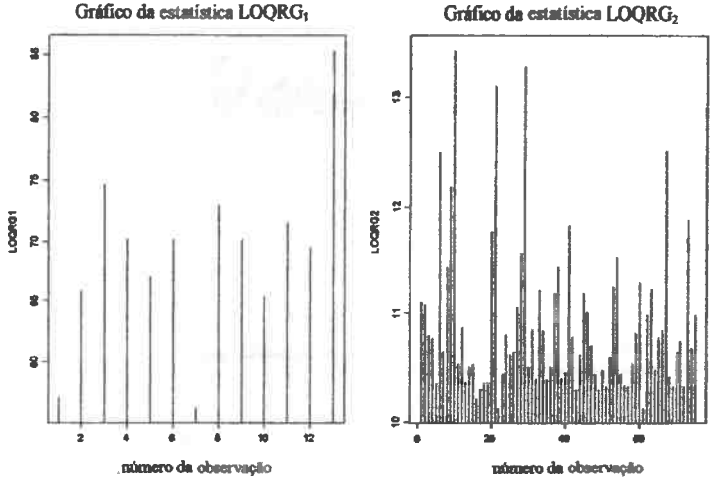

Figura 6.1 Gráficos da estatística LOQRG.

Apesar do teste de homogeneidade de matrizes de covariâncias indicar covariâncias diferentes, para efeito de ilustração, foi verificado também o desempenho das demais medidas.

As Figuras 6.2 a 6.6 são os gráficos dos valores destas medidas de diagnóstico para ambos os grupos.

- $LORAM_i = n^{-1}\sum_{j=1}^{2}\sum_{k=1}^{n_j}\ln\left\{\frac{o(\mathbf{x}_{jk})}{o_{(i)}(\mathbf{x}_{jk})}\right\}$ **medida Bayesiana**

em que $\ln[o(\mathbf{y})] = \ln\frac{q}{(1-q)} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$,

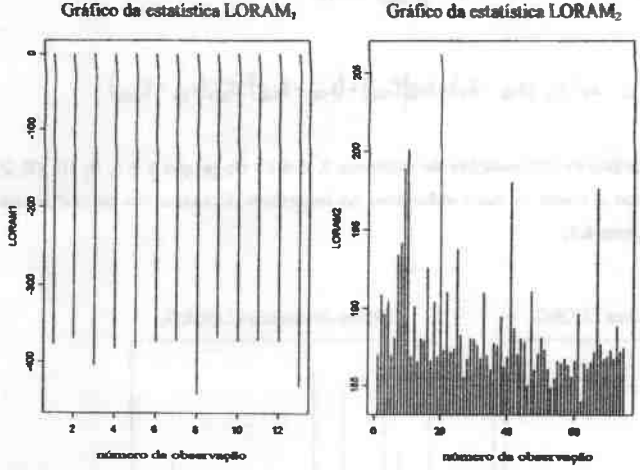

Figura 6.2 *Gráficos da estatística LORAM* **para o grupo 1 e grupo 2.**

- $I(\mathbf{x};\Delta^2) = w_1\Delta^2 + 2\phi - w_1\phi^2$,

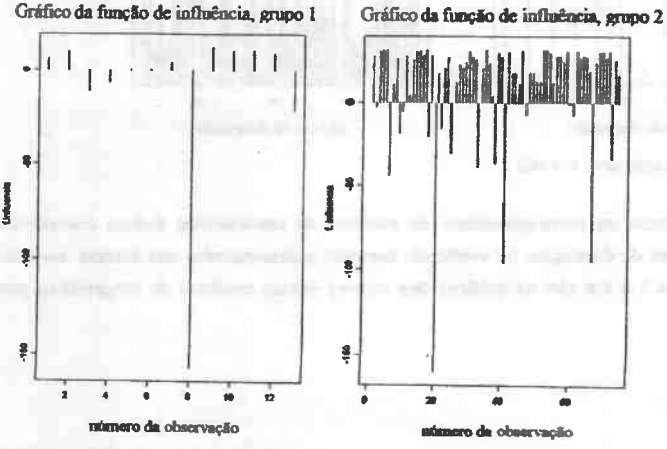

Figura 6.3 – Gráfico da função de influência

- $\hat{I}_m\left(\mathbf{x};\Delta^2\right)=w_1\left(\hat{\boldsymbol{\phi}}-w_1^{-1}\right)^2$

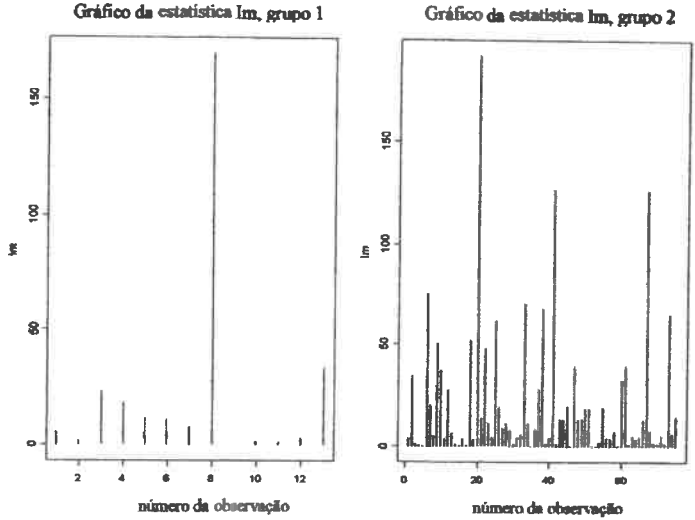

Figura 6.4 – Gráfico da estatística $\hat{I}_m\left(\mathbf{x};\Delta^2\right)$.

- $DPCI_i \cong \dfrac{\Phi\left(-\frac{1}{2}D\right)}{4D(n_1-1)^2}\left[\left(1-w_1\hat{\phi}_i\right)^2\left(d_i^2-\hat{\phi}_i^2/D^2\right)+\dfrac{1}{4}\hat{\phi}_i^2\right]$.

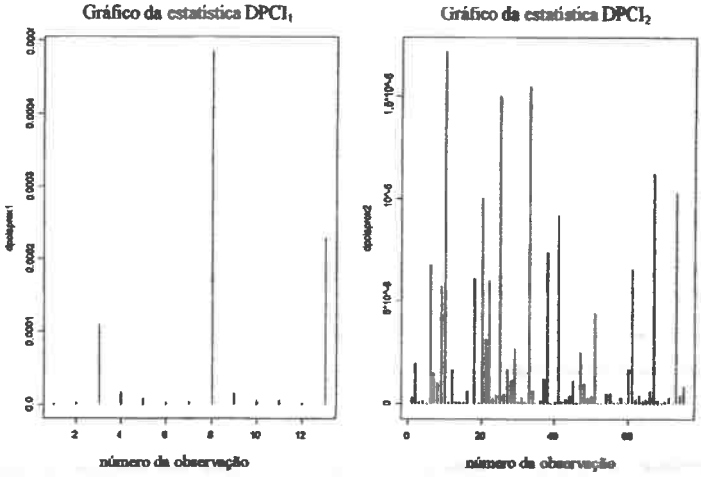

Figura 6.5 – Gráfico da estatística DPCI para o grupo 1 e grupo 2.

- $$\hat{\Delta}^2_{1(1)} = \left(\frac{n_1+n_2-1}{n_1+n_2}\right)\left[\hat{\Delta}^2 - \frac{1}{w_1(n_1-1)} + \frac{\left\{c(x_{11}) - w_1^{-1}\right\}^2}{w_1^{-1}(n_1-1) - \hat{\alpha}_1^2(x_{11})}\right].$$

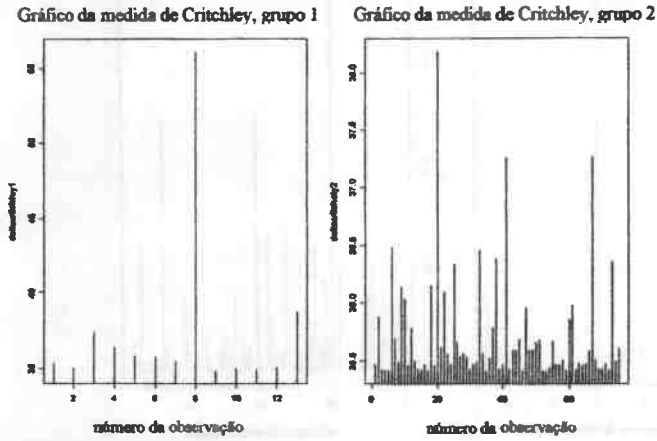

Figura 6.6 – Gráfico da medida $\Delta^2_{1(1)}$.

- $$E(DLO_i)^2 = (w_1B_1 + w_2B_2)^2 + V \ .$$

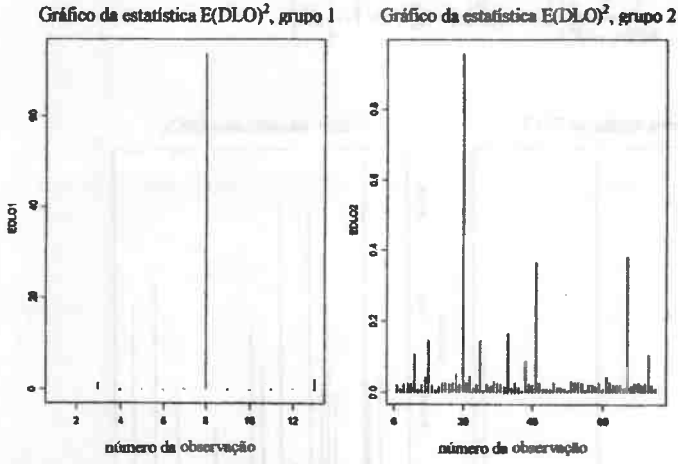

Figura 6.7 – Gráfico da estatística $E(DLO)^2$.

Observando os gráficos, selecionou-se as observações que apresentaram os maiores valores para as medidas em relação às demais, dentro de cada grupo.

Nota-se que, para o grupo 1, as observações de número 8 e 13 foram apontadas por todas as medidas, sendo que a observação de número 3 deixou de ser apontada apenas pela estatística $I(x;\Delta^2)$. Já para o grupo 2, as observações 20 e 67 foram indicadas em todas as medidas e a observação 73 deixou de ser indicada apenas pela medida $I(x;\Delta^2)$. No grupo 1, destacam-se as observações 3, 8 e 13, e no grupo 2, as observações 6, 20, 25, 33, 38, 41, 67 e 73.

Complementando a análise gráfica, são apresentados a seguir dois gráficos do tipo *qq plot*, que também apontam as observações discrepantes de ambos os grupos. Apresenta-se na Figura 6.8 os valores da estatística $\frac{\hat{\phi}_i}{D}$ contra os quantis da distribuição normal padrão.

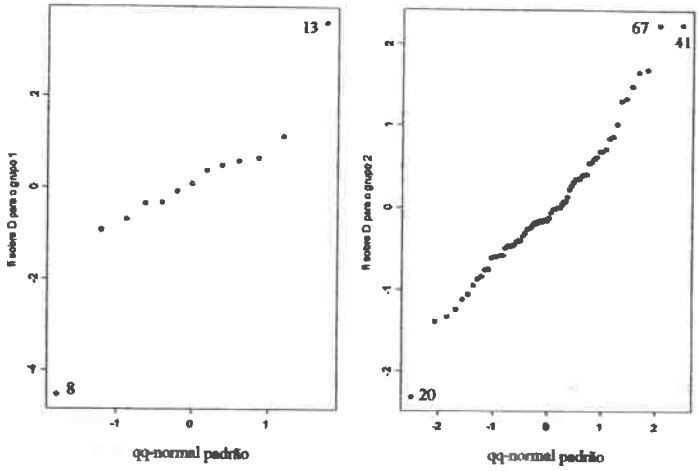

Figura 6.8 – Gráfico da estatística $\frac{\hat{\phi}_i}{D}$ contra os quantis da distribuição normal padrão.

No gráfico da Figura 6.9 observa-se os valores da estatística d_i^2 contra os quantis da distribuição quiquadrado.

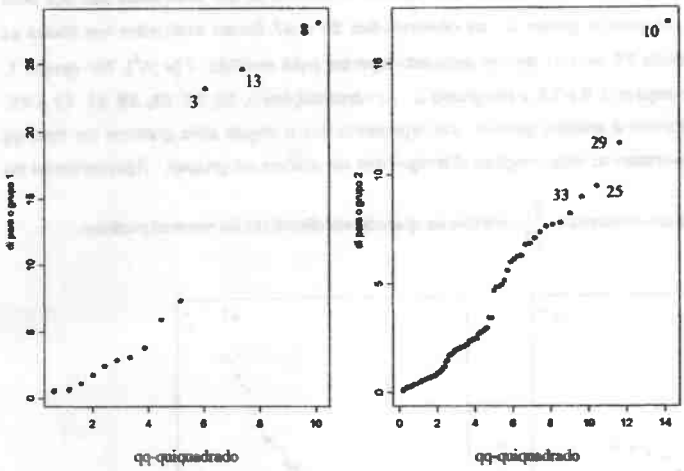

Figura 6.9– Gráfico da estatística d_i^2 contra quantis da distribuição quiquadrado.

Finalizando o estudo, foram feitas análises discriminantes linear e quadrática para o conjunto de dados, através do software Minitab.

A tabela A-1, apresentada no Anexo 4, indica as observações que se mostraram mais influentes e os correspondentes valores de cada medida de influência.

Na análise discriminante quadrática para o conjunto completo dos dados, a proporção de classificação correta foi de 0,989, sendo apontada apenas a observação 8 do grupo 1, classificada incorretamente, com probabilidade 0,33 de ser do grupo 1 e 0,67 de pertencer ao grupo 2. Na análise discriminante linear, a proporção correta de classificação foi de 0,989, apontando também a observação do grupo 1 como classificada incorretamente.

Procedeu-se à análise discriminante na ausência dos pontos 8 do grupo 1 e 20, 41 e 67 do grupo 2, por serem os mais discrepantes.

Eliminando-se somente a observação 8 do grupo 1 tanto na análise discriminante quadrática quanto na análise discriminante linear, obteve-se uma proporção de 100% de classificação correta. A distância de Manalanobis, que anteriormente era 35,8264, passou para 56,6779, fato que, na análise discriminante linear implica numa menor probabilidade de classificação incorreta. Com a retirada da observação 20 do grupo 2, a proporção de classificação correta é 0,989 tanto na análise linear quanto na quadrática, (observação 8 do grupo 1 classificada incorretamente) . A distância de Mahalanobis passou para 37,42 , valor este que difere pouco da distância D^2 do conjunto de dados completo, que é de 35,8264.

Excluindo-se a observação 41 do grupo 2, verificam-se as mesmas proporções de classificação correta obtidas quando da retirada da observação 20 do grupo 2, o mesmo ocorrendo na ausência da observação 67 do grupo 2.

Excluída a observação 41 do grupo 2, a distância de Mahalanobis é de 38,0738 e excluída a observação 67 do grupo 2, essa distância é de 38,0813.

Uma vez que a observação 10 do grupo 2 foi apontada pela medida *LOORG* (quadrática) e apontada também por algumas medidas (Bayesiana, *DPCI*), foi feita a análise discriminante quadrática e linear sem esse ponto. Observa-se que a distância de Mahalanobis nessa situação é de 36,5536 que é bem próxima da distância de Mahalanobis para o conjunto de dados completo, e a única observação classificada incorretamente é a de número 8 do grupo 1.

Observou-se também os elementos 21 e 29 do grupo 2, apontados pela medida *LOORG* como mais influentes. Na análise discriminante linear realizada na ausência desses pontos, obteve-se distância de Mahalanobis de 35,9158 e 35,4021, respectivamente, valores estes bem próximos de 35,8264 que é a distância de Mahalanobis para o conjunto completo dos dados. As funções lineares discriminantes também foram próximas àquelas para o conjunto completo, e tanto na análise discriminante quadrática, quanto na análise discriminante linear é apontada apenas a observação 8 como classificada incorretamente.

Nota-se que os três pontos indicados pela estatística *LOORG* como os mais influentes, casos 10 , 21 e 29 do grupo 2, são aqueles que apresentam altos valores para d_i^2, no entanto procedendo-se à análise discriminante na ausência desses pontos, as alterações não são muito significativas.

Examinando a observação 8 do grupo 1, verifica-se que apresenta comprimento total 155 mm, comprimento da cauda 57mm, comprimento do pé 17mm e comprimento do pelo 4 mm. Se excluído esse caso do grupo 1 , os valores mínimos para essas variáveis seriam 250mm, 148mm, 20mm e 8 mm respectivamente. Portanto, todos os valores das variáveis da observação 8 do grupo 1 estão abaixo desses valores mínimos, identificando-a claramente como um *outlier*.

No grupo 2, os valores mínimos são de 130mm, 46mm, 17mm e 5 mm, sendo que a observação 8, a princípio, estaria melhor classificada nesse grupo. Sugere-se então que se verifique as outras variáveis medidas nesse elemento a fim de explicar o porquê de sua inclusão no grupo 1, e questionar se essa classificação é realmente correta, uma vez que sua permanência no grupo 1 causa grande impacto não só na distância de Mahalanobis como também na probabilidade de classificação incorreta.

Deste exemplo, pode-se concluir que as medidas de diagnóstico apresentadas têm o potencial de apontar observações atípicas, que podem influir na análise. Desta forma, o pesquisador deve examinar com mais cautela essas observações, verificando a necessidade ou não da exclusão desses casos. Por outro lado, a escolha das medidas de discrepância depende do objetivo inferencial e também das características do problema apresentado.

Considerando o estudo de todas as medidas apresentadas no presente trabalho, sugere-se o uso da medida $\hat{\Delta}^2_{(i)}$ de Critchley e Vitiello (1991):

$$\hat{\Delta}^2_{1(i)} = \left(\frac{n_1+n_2-1}{n_1+n_2}\right)\left[\hat{\Delta}^2 - \frac{1}{w_1(n_1-1)} + \frac{\left\{c(\mathbf{x}_{11}) - w_1^{-1}\right\}^2}{w_1^{-1}(n_1-1) - \hat{\alpha}_1^2(\mathbf{x}_{11})}\right],$$

uma vez que ela fornece diretamente a distância de Mahalanobis quando da eliminação da observação, e, consequentemente, pode-se avaliar também a alteração da probabilidade de classificação incorreta, visto que $PCI = \Phi\left(-\frac{\Delta}{2}\right)$.

7. Considerações finais

Técnicas de diagnóstico são exaustivamente estudadas em regressão, mas pouco exploradas em análise discriminante. Neste trabalho, procurou-se descrever as técnicas de diagnóstico em análise discriminante, devido à sua pouca utilização, apesar da extensa literatura existente.

Concluindo, sugere-se como primeiro passo da análise examinar aquelas observações identificadas como mais influentes, pois algumas delas podem se tratar de erros, devendo ser corrigidas ou descartadas. Para as outras observações influentes que permanecerem no conjunto de dados, é importante verificar como as inferências de interesse seriam alteradas se esses casos não tivessem sido observados. Sendo assim, o pesquisador poderá decidir sobre a permanência ou não desses *outliers* em seu estudo, e talvez até partir para uma análise mais aprofundada dessas observações.

Todos os cálculos das medidas apresentadas estão descritos com detalhes em Reigada(2005).

Apêndice

A-1 Resultado

Se X tem distribuição Normal p-variada de média μ e matriz de covariância Σ, $X \sim N_p(\mu, \Sigma)$, então qualquer combinação linear das variáveis $a'X = a_1X_1 + a_2X_2 + \cdots + a_pX_p$ é distribuída como normal univariada de média $a'\mu$ e variância $a'\Sigma a$. Também, se $a'X$ é distribuída como $N(a'\mu, a'\Sigma a)$ para todo a, então X tem distribuição $N_p(\mu, \Sigma)$.

A-2 Resultado

Se $X \sim N_p(\mu, \Sigma)$ e $Z = X - \mu \sim N_p(0, \Sigma)$, devido ao Resultado A-1,

$\phi = l'Z \sim N(0, l'\Sigma l)$. Dessa forma, $\frac{l'z}{\sqrt{l'\Sigma l}} \sim N(0,1)$ e $\frac{\phi^2}{l'\Sigma l} \sim \chi_1^2$.

Assim, $E\left(\frac{\phi^2}{l'\Sigma l}\right) = 1$, Mediana $\left(\frac{\phi^2}{l'\Sigma l}\right) = 0{,}455$, $P\left(\frac{\phi^2}{l'\Sigma l} \le 0{,}455\right) = 0{,}5$ e

$P\left(\phi^2 \le 0{,}455\, l'\Sigma l\right) = 0{,}5$.

Então, $E(\phi^2) = l'\Sigma l$, Mediana $(\phi^2) = 0{,}455\, l'\Sigma l$, ou seja,

a média de ϕ^2 é maior que a mediana e portanto, ϕ^2 tem assimetria positiva.

Como conseqüência, $-w_1\Delta^2\phi_p^2$ tem assimetria negativa e como $w_1\Delta^2$ é constante e $2\Delta\phi_p$ é simétrica, então $I(x, \Delta^2)$ terá assimetria negativa.

A-3 Distribuição quiquadrado não central

Se y tem distribuição normal p-variada com média μ e matriz de covariância Σ, $y \sim N_p(\mu, \Sigma)$, então $y'\Sigma^{-1}y$ tem distribuição quiquadrado não central com p graus de liberdade e parâmetro de não centralidade $\mu'\Sigma^{-1}\mu$.

A-4

Tabela A.1 - Observações mais influentes e respectivos valores das medidas de diagnóstico.

Casos	Estatísticas								
Grupo 1	*LOQRG*	*LORAM*	$\hat{I}(x,\Delta^2)$	$\hat{Im}(x,\Delta^2)$	*DPCI*	$\hat{\Delta}^2_{1(i)}$	*EDLO*	$\hat{\phi}_i$	d_i^2
3	74,74	-403,40	-10,67	2,27E+01	1,09E-04	37,49	1,57	-5,64	23,17
8	73,06	-442,90	-157,23	1,69E+02	4,86E-04	56,13	73,70	-27,08	27,93
9	70,22	-367,70	12,06	2,83E-04	1,48E-04	34,87	0,18	6,73	7,33
13	85,32	-432,30	-21,29	3,34E+01	2,29E-04	38,81	2,30	21,8	24,53
Grupo 2									
6	12,50	188,09	-44,35	76,06	6,75E-06	36,49	0,11	-8,27	4,66
8	11,45	194,18	26,20	5,52	1,02E-06	35,49	0,03	3,72	5,97
9	12,19	198,89	-18,75	50,46	5,70E-06	36,14	0,05	8,87	5,58
10	13,41	200,08	-5,65	37,37	1,72E-06	36,03	0,15	7,8	16,85
20	11,77	206,22	-159,73	191,45	1,00E-05	38,19	0,96	-13,82	6,82
21	13,09	187,30	17,68	14,03	3,12E-06	35,62	0,00	5,23	7,76
25	10,61	193,78	-30,54	62,25	1,50E-05	36,34	0,14	-7,37	9,46
29	13,28	188,11	23,72	7,99	2,68E-06	35,53	0,03	-1,89	11,39
33	11,22	191,02	-38,91	70,62	1,55E-05	36,46	0,17	-7,93	8,95
38	11,46	189,52	-36,51	68,22	7,35E-06	36,39	0,09	10,12	6,08
41	11,84	198,06	-95,32	127,03	9,21E-06	37,26	0,37	13,38	7,05
67	12,52	197,65	-94,66	126,37	1,13E-05	37,26	0,38	13,35	7,59
73	11,89	188,87	-33,64	65,35	1,03E-05	36,37	0,11	9,93	7,68

8. Referências bibliográficas

CAMPBELL, N. A. (1978). *The Influence Function as an Aid in Outlier Detection in Discriminant Analysis.* Applied Statistics, 27 nr 3: 251-258.

CRITCHLEY, F. ; VITIELLO, C. (1991). *The Influence of Observations Misclassification Probability Estimates in Linear Discriminant Analysis.* Biometrika, 78: 677-690.

FUNG, W. K. (1992). *Some Diagnostic Measures in Discriminant Analysis.* Statistics and Probability Letters, 13: 279-285.

FUNG, W. K. (1995-a). *Diagnostics in Linear Discriminant Analysis.* Journal of the American Statistical Association, 90 nr 43: 952-956.

FUNG, W. K. (1995-b). *Detecting Influential Observations for Estimated Probabilities in Multiple Discriminant Analysis.* Computational Statistics & Data analysis, 20: 557-568.

FUNG, W. K. (1996-a). *Diagnosing Influential Analysis.* Biometrics, 52: 1235-1241.

FUNG, W. K. (1996-b). *The Influence of Observations for Local Log-odds in Linear Discriminant Analysis.* Communications in Statistics – Theory and Methods, 25: 257-268.

FUNG, W. K. (1996-c). *The Influence of Observations on Missclassification Probability in Multiple Discriminant Analysis.* Communications in Statistics – Theory and Methods, 25(8), 1917-1930.

FUNG, W. K. (1998). *On the Equivalence of Two Diagnostic Measures in Discriminant Analysis.* Communications in Statistics – Theory and Methods, 27(8): 1915-1922.

GEISSER, S. (1964). *Posterior Odds for Multivariate Normal Classifications.* Journal of the Royal Statistical Society, Ser B, 26: 69-76.

GEISSER, S. (1966). *" Predictive Discrimination" in Proceedings of the International Symposium on Multivariate Analysis.* New York: Academic Press, pg 149-163.

HAMPEL, F. R. (1974). *The Influence Curve and its Role in Robust Estimation.* Journal of the American Statistical Association, 69: 383-393.

HEALY, M. J. R. (1968). *Multivariate Normal Plotting.* Applied Statistics 17 : 157-161.

JOHNSON, W. (1987). *The Detection of Influential Observations for Allocation, Separation, and the Determination of Probabilities in a Bayesian Framework.* Journal of Business & Economic Statistics 5: 369-381.

KULLBACK, S.; LEIBLER, R. A. (1951). *On Information and Sufficiency.* The Annals of Statistics, 22: 79-86.

LEITE, J.G.; SINGER, J. M. e LOURENÇO, F.C. (1994). Relatório de análise Estatística sobre o projeto: Biologia de uma comunidade de marsupiais e roedores, em Floresta Atlântica de montanha, no Parque Estadual da Serra do Tabuleiro, Santa Catarina, Brasil. São Paulo, IME - USP, 33 p (RAE-CEA-94P11)

MORRISON, D. F. (1976). *Multivariate Statistical Methods* (2nd ed) New York McGraw Hill.

REIGADA, S. M. B. (2005). Diagnóstico em Análise Discriminante. Dissertação de Mestrado. IME - USP, SP.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2005-01 - DE SOUZA BORGES,W., GUSTAVO ESTEVES, L., WECHSLER, S. Process Parameters Estimation in The Taguchi On-Line Quality Monitoring Procedure For Attributes. 2005.19p. (RT-MAE-2005-01)

2005-02 - DOS ANJOS,U.,KOLEV, N. Copulas with Given Nonoverlapping Multivariate Marginals. 2005.09p. (RT-MAE-2005-02)

2005-03 - DOS ANJOS,U.,KOLEV, N. Representation of Bivariate Copulas via Local Measure of Dependence. 2005.15p. (RT-MAE-2005-03)

2005-04 - BUENO, V. C., CARMO, I. M. A constructive example for active redundancy allocation in a k-out-of-n:F system under dependence conditions. 2005. 10p. (RT-MAE-2005-04)

2005-05 - BUENO, V. C., MENEZES, J.E. Component importance in a modulated Markov system. 2005. 11p. (RT-MAE-2005-05)

2005-06 - BASAN, J. L., BRANCO, M.D´E., BOLFARINE, H. A skew item response model. 2005. 20p. (RT-MAE-2005-06)

2005-07 - KOLEV, N., MENDES, B. V. M., ANJOS, U. Copulas: a Review and recent developments. 2005. 46p. (RT-MAE-2005-07)

2005-08 - VENEZUELA, M. K., BOTTER, D. A., SANDOVAL, M. C. Diagnostic techniques in generalized estimating equations. 2005. 16p. (RT-MAE-2005-08)

2005-09 - **BOLFARINE, H., LACHOS, V.H.** Skew-probit measurement error models. 2005. 10p. (RT-MAE-2005-09)

2005-10 - **SALINAS, V.H., ROMEO, J.S., PENA, A.** On Bayesian estimation of a survival curve: Comparative study and examples. 2005. 14p. (RT-MAE-2005-10)

2005-11 - **LACHOS, V.H., BOLFARINE, H., VILCA, F., GALEA, M.** Estimation and influence diagnostics for structural comparative calibration models under the skew-normal distributions. 2005. 24p. (RT-MAE-2005-11)

2005-12 - **PEREIRA, B.B., PEREIRA, C.A.B.** A likelihood to diagnostic tests in clinical medicine. 2005. 23p. (RT-MAE-2005-12)

2005-13 - **BUENO, V.C.** A note on component reliability importance through parallel improvement. 2005. 9p. (RT-MAE-2005-13)

2005-14 - **SILVA, G.T., SEN, P.K., LIMA, A. C. P.** Estimation of the mean quality adjusted survival using a multistate model for the sojourn times. 2005. 9p. (RT-MAE-2005-14)

2005-15 - **PEREIRA, C.A., STERN, J.M., WECHSLER, S.** Can a significance test be genuinely Bayesian? 2005. 27p. (RT-MAE-2005-15)

2005-16 - **CYSNEIROS, A.H.M.A., FERARRI, S.L.P.** An improved likelihood ratio test for varying dispersion in exponential family nonlinear models. 2005. 12p. (RT-MAE-2005-16)

2005-17 - **FERRARI, S.L.P., SILVA, M.F., CRIBARI-NETO, F.** Adjusted profile likelihoods for the Weibull shape parameter. 2005. 18p. (RT- MAE2005-17)

2005-18 - **BRANCO, M. D., RODRIGUES, C.B.** Bayesian Inference for the skewness parameter on skew normal distribution. 2005. 21p. (RT- MAE2005-18)

2005-19 - AZEVEDO, C. L. N., NOBRE, J. S., O princípio da equivariância: conceitos e aplicações. 2005. 32p. (RT-MAE-2005-19)

2005-20 - BUENO, V.da C., A Series representation of a coherent system. 2005. 15p. (RT-MAE-2005-20)

2005-21 - BUENO, V.da C., Hot standby redundancy allocation in a k-out-of-n:f systems of dependent components. 2005. 14p. (RT-MAE-2005-21)

2005-22 - BUENO, V.da C., CARMO, I.M. do Active redundancy allocation for a k-out-of-n:F system under a point process transform. 2005. 14p. (RT-MAE-2005-22)

2005-23 - MONTENEGRO, L.C.; BOLFARINE, H., LACHOS, V.H. Inference and influence diagnostics in the grubbs model under the skew-normal distribution. 2005. 27p. (RT-MAE-2005-23)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05311-970 - São Paulo, Brasil