Avaliação da probabilidade de classificação incorreta em análises discriminantes para duas populações normais

Izabela R. C. Oliveira[1†], *Daniel F. Ferreira*[2]

[1] *Doutoranda em Estatística e Experimentação Agronômica, Escola Superior de Agricultura Luiz de Queiroz - ESALQ/USP.*

[2] *Professor Associado II, Departamento de Ciências Exatas, Universidade Federal de Lavras, Bolsista CNPq. E-mail: danielff@ufla.br.*

Resumo: *Este trabalho teve por objetivo avaliar o desempenho do método de Lachenbruch e Mickey (1968) com a modificação proposta por Giri (2004) utilizando métodos de simulação Monte Carlo mensurando-se as taxas de classificação incorreta e comparando-as com o método original. Em todos os casos os custos de classificação incorreta e as probabilidades a priori foram considerados iguais em ambas as populações. Para isso foram consideradas $k = 2$ populações homocedásticas normais multivariadas e custos de classificação incorreta e probabilidades a priori idênticos nas duas populações. Foram simuladas diferentes configurações populacionais utilizando-se $N = 2000$ repetições Monte Carlo. Em cada uma das simulações foi estimada a taxa de classificação incorreta total utilizando os métodos modificado e original de Lachenbruch e Mickey (1968). Em cada caso, como os parâmetros populacionais são conhecidos a probabilidade real de classificação incorreta foi determinada. Para avaliar o desempenho de ambos os estimadores foi determinado o viés e o erro quadrático médio. Ambos os métodos, original e modificado, são viesados e possuem grandes vieses com amostras pequenas, e pequenos vieses com amostras grandes. Os vieses dos dois métodos decrescem com o aumento da distância de Mahalanobis entre as duas populações. O método original é superior ao método modificado, principalmente em pequenas amostras.*

Palavras-chave: custo de classificação incorreta; normal multivariada; homocedástica; erro quadrático médio.

Abstract: *This work aimed to evaluate the performance of the Lachenbruch and Mickey (1968) method considering the modification proposed by Giri (2004) using Monte Carlo simulations and to measure and compare the misclassification rates with those obtained in the original method. Both methods original and modified are biased and have large biases in small samples and small biases with large samples. The biases of the two methods decrease with the increase of the Mahalanobis distance between the two populations. The original method is superior to the modified method, especially in small samples.*

Keywords: cost of misclassification; multivariate normal; homoscedastic; mean quadratic error.

[†] *Autor correspondente:* izabela.rco@gmail.com

1 Introdução

A análise discriminante é uma importante técnica da estatística multivariada para classificar uma nova observação p-variada $\boldsymbol{x}$ $(p \times 1)$ em uma entre k diferentes populações. Para se realizar esta classificação é necessário assumir que a forma das densidades populacionais seja conhecida. Em geral, a forma utilizada é a normal multivariada. Outro aspecto refere-se ao fato de os parâmetros dessas densidades serem conhecidos. Quando isso ocorre, as regras de classificação podem ser facilmente determinadas. Estas regras devem ser avaliadas quanto aos seus desempenhos em classificar um elemento na sua respectiva população de origem. Uma forma de avaliá-las é determinar as probabilidades de classificação incorreta das observações. Uma regra é considerada ótima, quando as probabilidades de classificação incorreta são mínimas. Neste trabalho tem-se particular interesse no caso de $k = 2$ populações.

Supondo que $\boldsymbol{x}$ seja uma realização de uma variável aleatória p-dimensional $\boldsymbol{X}$ com distribuição normal multivariada, para a qual tem-se a intenção de realizar a classificação em uma entre duas populações $f_1(\boldsymbol{x})$ e $f_2(\boldsymbol{x})$, sendo

$$f_i(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \tag{1}$$

para $i = 1, 2$, em que $\boldsymbol{\mu}_i$ é o vetor de médias da i-ésima população e $\boldsymbol{\Sigma}$ é a matriz de covariâncias positiva definida comum às duas populações. A princípio, consideram-se os parâmetros $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}$ conhecidos.

De acordo com a regra que minimiza o custo médio de classificação incorreta, deve-se classificar $\boldsymbol{x}$ em 1 se

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left[\frac{C(1|2)}{C(2|1)}\right]\left(\frac{p_2}{p_1}\right).$$

Se as densidades $f_1(\boldsymbol{x})$ e $f_2(\boldsymbol{x})$ forem substituídas pela densidade normal correspondente, tem-se:

$$\exp\left\{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\} \geq \left[\frac{C(1|2)}{C(2|1)}\right]\left(\frac{p_2}{p_1}\right),$$

em que $C(i|j)$ é o custo de classificação incorreta de uma observação proveniente da população j na população i, sendo $i \neq j = 1, 2$ e p_i é a probabilidade *a priori* de que uma observação seja proveniente da i-ésima população.

Como ambos os termos são positivos, pode-se tomar o logaritmo, o que preserva a ordem da desigualdade. Assim, deve-se classificar $\boldsymbol{x}$ na população 1 se

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geq \ln\left\{\left[\frac{C(1|2)}{C(2|1)}\right]\left(\frac{p_2}{p_1}\right)\right\} \tag{2}$$

e em 2, caso contrário.

No entanto, nas situações reais, embora seja razoável assumir normalidade multivariada, os parâmetros das densidades populacionais não são conhecidos. Neste caso, as informações necessárias são estimadas de uma amostra, que é chamada de amostra de treinamento. O problema é que a regra de classificação (2) pode não ser mais ótima, pois sua construção foi baseada supondo modelos normais multivariados e parâmetros conhecidos. A avaliação do desempenho da regra de classificação estimada se torna preponderante.

Vários métodos de avaliação do desempenho da regra de classificação foram propostos na literatura (Giri, 2004). Dentre eles, Lachenbruch e Mickey (1968) propuseram um método,

baseado em um procedimento que combina a técnica *jackknife* e o método das probabilidades de classificações incorretas estimadas. Para aplicar este método, deve-se omitir das $n_1 + n_2$ observações a realização p-dimensional $\boldsymbol{x}_{ij}$ referente a i-ésima população e a j-ésima unidade amostral, para $i = 1, 2$, $j = 1, 2, \ldots, n_i$, sendo n_i o tamanho da amostra extraída da i-ésima população. Deve-se estimar as médias das amostras das populações 1 e 2 e a matriz de covariâncias comum, excluindo esta observação $\boldsymbol{x}_{ij}$. Pode-se representar estas médias e a matriz de covariâncias combinada por $\bar{\boldsymbol{X}}_1^{-(ij)}$, $\bar{\boldsymbol{X}}_2^{-(ij)}$ e $\boldsymbol{S}_p$. Para esta observação calcula-se o valor y_{ij} por

$$y_{ij} = \left(\bar{\boldsymbol{X}}_1^{-(ij)} - \bar{\boldsymbol{X}}_2^{-(ij)}\right)^\top \boldsymbol{S}_p^{-1}\boldsymbol{x}_{ij} - \frac{1}{2}\left(\bar{\boldsymbol{X}}_1^{-(ij)} - \bar{\boldsymbol{X}}_2^{-(ij)}\right)^\top \boldsymbol{S}_p^{-1}\left(\bar{\boldsymbol{X}}_1^{-(ij)} + \bar{\boldsymbol{X}}_2^{-(ij)}\right). \quad (3)$$

Repete-se esse processo para todos os valores de i e de j, omitindo somente a observação $\boldsymbol{x}_{ij}$ em cada etapa e determina-se y_{ij} pela expressão (3). Assim, tem-se as amostras y_{11}, y_{12}, $\ldots$, y_{1n_1}, da população π_1 e y_{21}, y_{22}, $\ldots$, y_{2n_2}, da população π_2. Determinam-se as médias e as variâncias de cada amostra por

$$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \text{e} \qquad S_i^2 = \frac{1}{n_i - 1}\left[\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{n_i} y_{ij}\right)^2}{n_i}\right], \quad (4)$$

para $i = 1, 2$.

Lachenbruch e Mickey (1968) propuseram estimar as probabilidades $P(2|1)$ e $P(1|2)$ por

$$\hat{P}(2|1) = \Phi\left(-\frac{\bar{y}_{1.}}{S_1}\right) \qquad \text{e} \qquad \hat{P}(1|2) = \Phi\left(\frac{\bar{y}_{2.}}{S_2}\right) \quad (5)$$

e a taxa de erro aparente por

$$TEA = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_1}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_2}\right). \quad (6)$$

Giri (2004) afirma que seria interessante investigar o desempenho deste método para estimar as taxas de erro apresentadas nas expressões (5) e (6) se os estimadores S_1 e S_2 dos desvios padrões de y_{ij} forem substituídos por um estimador comum. A idéia de se realizar tal substituição é baseada no fato de as populações serem homocedásticas.

Este trabalho teve por objetivo avaliar o desempenho do método de Lachenbruch e Mickey (1968) com a modificação proposta por Giri (2004) utilizando métodos de simulação Monte Carlo mensurando-se as taxas de classificação incorreta e as comparando com o método original. Em todos os casos os custos de classificação incorreta e as probabilidades *a priori* foram considerados iguais em ambas as populações.

2 Metodologia

Para avaliar o desempenho da nova regra de classificação modificada de Lachenbruch e Mickey (1968), considerando a utilização de uma variância comum nas expressões (5) e (6) dada por

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}, \tag{7}$$

foi proposto a realização de simulações Monte Carlo.

Para isso foram consideradas $k = 2$ populações homocedásticas normais dadas por (1) e custos de classificação incorreta e probabilidades *a priori* idênticos nas duas populações. A média da população 1 foi fixada como um vetor nulo $\boldsymbol{\mu}_1 = \mathbf{0}$, sem perda de generalidade. A matriz de covariâncias populacional $\boldsymbol{\Sigma}$ teve uma estrutura equicorrelacionada dada por

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}, \tag{8}$$

em que $\sigma^2 = 1$, sem perda de generalidade e $\rho = 0$, $0,5$ e $0,9$, emulando situações de ausência de correlação, média correlação e alta correlação entre as variáveis.

O parâmetro $\boldsymbol{\mu}_2$ foi fixado em função da distância entre as médias populacionais dada por $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, que foi considerado igual a 0, 2, 4, 8, 16 e 32. Foram considerados tamanhos amostrais da população 1 iguais a $n_1 = 10$, 20, 50, 100 e da população 2 iguais a $n_2 = 10$, 20, 50, 100 combinados fatorialmente.

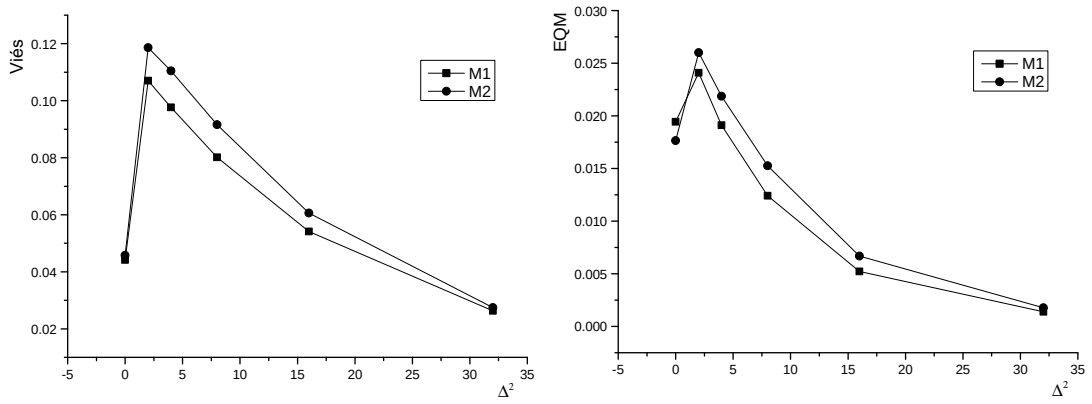
Cada uma das combinações das configurações consideradas foi simulada utilizando-se $N = 2000$ repetições Monte Carlo. Em cada uma das simulações foi estimada a taxa de classificação incorreta total utilizando os métodos modificado e original de Lachenbruch e Mickey (1968). Em cada caso, como os parâmetros populacionais são conhecidos, a probabilidade real de classificação incorreta foi determinada. Para avaliar o desempenho de ambos os estimadores foi determinado o viés e o erro quadrático médio.

3 Resultados e Discussão

Nesse trabalho as rotinas computacionais foram implementadas no programa **R**. Alguns resultados, considerando diferentes tamanhos de amostras e correlações entre as variáveis, são apresentados na seqüência. Em todos os casos foi fixado o número de variáveis em $p = 5$. Na Figura 1 são apresentados os vieses (a) e os erros quadráticos médios (EQM's) (b) para a situação de $n_1 = n_2 = 10$ e $\rho = 0$ em função de Δ^2, que é a distância de Mahalanobis entre as médias das populações 1 e 2. Quando as populações são idênticas, os valores de vieses e EQM's são idênticos nos dois métodos, exceto pelo erro de Monte Carlo. Quando Δ^2 é maior que 0, os vieses e EQM's dos dois métodos são distintos, sendo que o método modificado (M2) apresenta quase sempre pior desempenho do que o método original (M1). Quando $\Delta^2 = 32$, os valores de vieses e EQM's tendem a se igualar. Embora M1 seja superior, as diferenças tanto nos vieses quanto nos EQM's não são expressivas.

Os valores paramétricos das probabilidades totais de classificação incorreta são $0,5$; $0,2398$; $0,1587$; $0,0786$; $0,0228$ e $0,0024$ para $\Delta^2 = 0$, 2, 4, 8, 16 e 32, respectivamente. Os vieses de ambos os métodos são muito grandes quando os valores de Δ^2 são pequenos, e decrescem com o aumento de Δ^2. No caso particular de $\Delta^2 = 2$, os vieses atingiram a magnitude de aproximadamente $0,12$, para um valor paramétrico de $0,2398$, o que representa cerca de 50% do valor real. Em todos os casos o valor real foi superestimado. Estimamos os desvios padrões

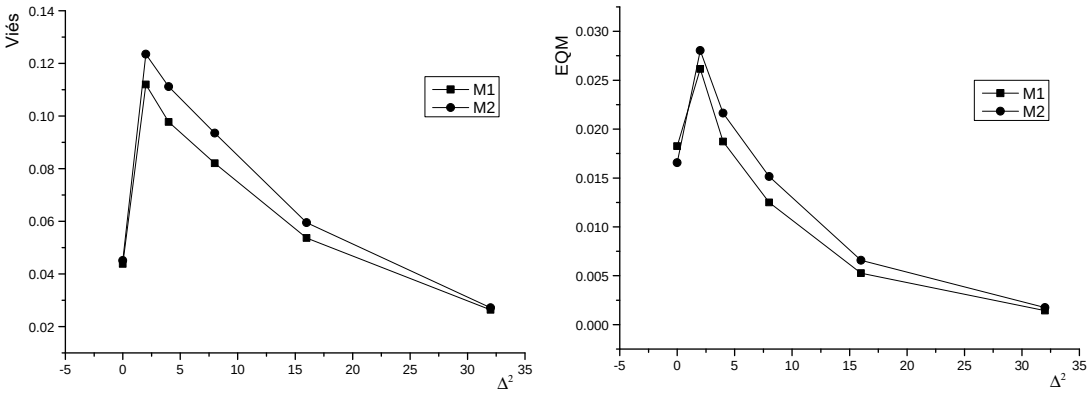
das estimativas das probabilidades totais de classificação incorreta nas 2000 simulações Monte Carlo realizadas para os dois métodos. Elas foram praticamente idênticas, o que permite inferir que a diferença observada entre os EQM's dos dois métodos é devida ao viés e não a precisão.



(a) (b)

Figura 1: Valores de viés (a) e EQM (b) dos métodos original (M1) e modificado (M2) para determinar as taxas de classificação incorretas considerando $n_1 = n_2 = 10$ e $\rho = 0$ em função de Δ^2.

Na Figura 2 são apresentados os vieses (a) e os erros quadráticos médios (EQM's) (b) para a situação de $n_1 = n_2 = 10$ e $\rho = 0,9$ em função de Δ^2. Pode-se observar o mesmo padrão de comportamento retratado na Figura 1, com ausência de correlação entre as variáveis. Então, pode-se concluir que a presença de correlação não interfere no desempenho relativo dos dois métodos.

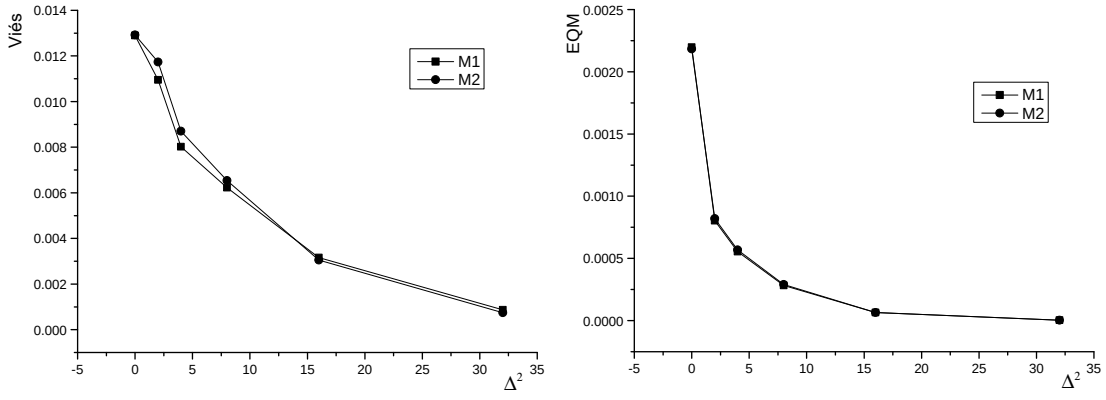


(a) (b)

Figura 2: Valores de viés (a) e EQM (b) dos métodos original (M1) e modificado (M2) para determinar as taxas de classificação incorretas considerando $n_1 = n_2 = 10$ e $\rho = 0,9$ em função de Δ^2.

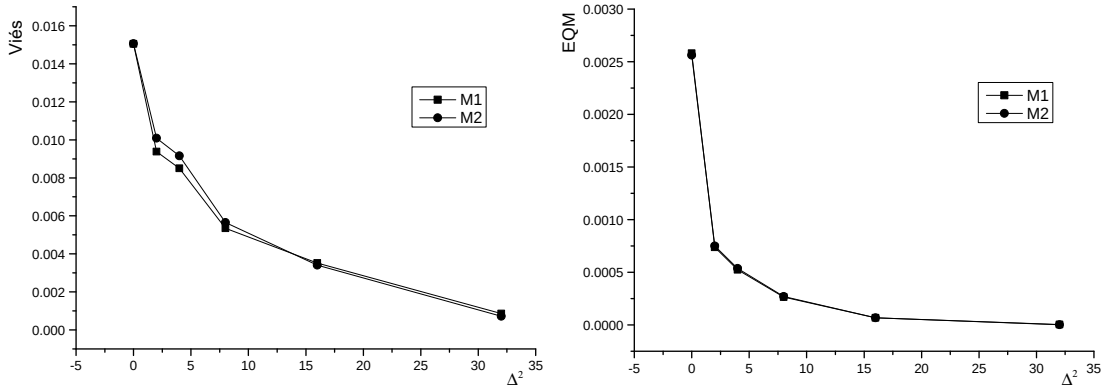
Na Figura 3 são apresentados os vieses (a) e os erros quadráticos médios (EQM's) (b) para a situação de $n_1 = n_2 = 100$ e $\rho = 0$. O que se observa é um redução drástica da magnitude dos vieses observados para $n_1 = n_2 = 10$ e um tendência forte de os métodos apresentarem desempenhos bastantes similares. Apesar disso, para os valores intermediários de Δ^2, entre 0 e 16, ainda ocorreu uma pequena vantagem para o método original em relação ao viés. Para os EQM's não foram observadas diferenças entre os métodos.

Para o caso de $n_1 = n_2 = 100$ e $\rho = 0,9$ (Figura 4) observou-se o mesmo padrão de comportamento em relação ao desempenho dos métodos considerando $\rho = 0$ (Figura 3).



(a) (b)

Figura 3: Valores de viés (a) e EQM (b) dos métodos original (M1) e modificado (M2) para determinar as taxas de classificação incorretas considerando $n_1 = n_2 = 100$ e $\rho = 0$ em função de Δ^2.



(a) (b)

Figura 4: Valores de viés (a) e EQM (b) dos métodos original (M1) e modificado (M2) para determinar as taxas de classificação incorretas considerando $n_1 = n_2 = 100$ e $\rho = 0,9$ em função de Δ^2.

4 Conclusões

Ambos os métodos, original e modificado, são viesados e possuem grandes vieses com amostras pequenas, e pequenos vieses com amostras grandes. Os vieses dos dois métodos decrescem com o aumento da distância de Mahalanobis entre as duas populações. O método original é superior ao método modificado, principalmente em pequenas amostras.

Referências

GIRI, N. C. *Multivariate statistical analysis.* 2th. ed. New York: Marcel Dekker, 2004. 558p.

LACHENBRUCH, P. A.; MICKEY, M. R. Estimation of error rates in discriminant analysis. *Technometrics*, v. 10, n. 1, p. 1–11, 1968.