COMUNICAÇÃO

COMPARAÇÃO DE MATRIZES DE COVARIÂNCIAS DE POPULAÇÕES NORMAIS DEPENDENTES: UM ESTUDO DE CASO

Comparing the covariance matrices of dependent normal populations: a case study

Marcelo Angelo Cirillo[1], Daniel Furtado Ferreira[2], Thelma Sáfadi[3]

RESUMO

Inferências sobre comparações de matrizes de covariâncias em populações normais dependentes são usualmente obtidas considerando testes assintóticos baseados na maximização de funções de verossimilhanças. Entretanto, se o número de populações e/ou de variáveis consideradas é excessivo pode-se ter problemas na convergência dos métodos numéricos utilizados para obtenção dos estimadores de máxima verossimilhança. Face a esse problema, objetivou-se, neste trabalho, ilustrar por meio de um conjunto de dados reais, a aplicação de um teste para comparar matrizes de covariâncias de populações correlacionadas, usando uma estatística baseada na razão de variâncias generalizadas, cuja distribuição empírica foi obtida por meio da técnica *bootstrap*.

Termos para indexação: *Bootstrap*, covariâncias, variâncias generalizadas.

ABSTRACT

Inferences about dependent normal populations are usually obtained considering asymptotic tests based on the maximization likelihood functions. However, if the number of populations and/or variables considered are too high one way have convergence problems with the numerical methods used to obtain the maximum likelihood estimators. This work aimed to illustrate, using a real data set, the application of a test to compare covariance matrices of correlated populations using a statistic based on generalized variances ratio, whose empirical distribution was obtained via *bootstrap* methods.

Index terms: *Bootstrap*, covariances, generalized variances.

(Recebido em 6 de junho de 2006 e aprovado em 27 de fevereiro de 2008)

Entre as suposições exigidas para a realização da inferência estatística paramétrica, a suposição de que as amostras sejam independentes e provenientes de populações cujas distribuições são conhecidas, é primordial. No contexto multivariado, essas suposições também são exigidas para a realização de testes estatísticos. Assim, caso o objetivo do pesquisador seja o de comparar matrizes de covariâncias, os testes comumente encontrados na literatura, como os de Bartlett (O'BRIEN, 1992) e Levene (LEVENE, 1960) poderão ser utilizados. Entretanto, poderá haver casos em que as respostas são provenientes de populações cujas amostras são correlacionadas, levando a violação da suposição de independência exigida para os referidos testes. Se considerarmos a dependência entre as populações, tem-se que, em geral, os testes propostos na literatura, são baseados na razão de verossimilhanças (JIANG et al., 1999) e limitados em relação ao número de variáveis e de populações, uma vez que suas estatísticas são baseadas em aproximações assintóticas da distribuição qui-quadrado (JIANG et al., 1999).

Uma outra questão que deve ser ressaltada é a complexidade de obter expressões analíticas para tais testes, além do mais, problemas numéricos para maximização das verossimilhanças considerado um maior número de variáveis e populações, comumente encontrado. Mediante esses problemas, como alternativa surge o uso de técnicas computacionais, das quais os métodos de computação intensiva, como técnicas de *bootstrap* são de grande importância, nas mais variadas situações reais (MANLY, 1997).

Objetivando ilustrar a aplicação de um teste baseado na razão de variâncias generalizadas (CIRILLO, 2006), sendo que sua distribuição empírica foi gerada por meio da técnica *bootstrap*, utilizou-se um conjunto de dados reais cedido pela UNIFAL - Universidade Federal de Alfenas. Esse conjunto de dados trata-se de um estudo

[1]Doutor em Estatística e Experimentação Agropecuária – Departamento de Ciências Exatas/DEX – Universidade Federal de Lavras/UFLA – Cx. P. 3037 – Lavras, MG – marcufla@gmail.com – Bolsista FAPEMIG

[2]PhD em Estatística e Experimentação Agropecuária, Professor Adjunto – Departamento de Ciências Exatas/DEX – Universidade Federal de Lavras/UFLA – Cx. P. 3037 – Lavras, MG – danielff@ufla.br

[3]PhD em Estatística, Professora Adjunta – Departamento de Ciências Exatas/DEX – Universidade Federal de Lavras/UFLA – Cx. P. 3037 – Lavras, MG – safadi@ufla.br

referente à análise do efeito do exercício de força sobre o estresse oxidativo no plasma de mulheres na terceira idade, realizado no período de 01/09 a 31/10/2004, no laboratório de Bioquímica Clínica do Departamento de Análises Clínicas e Toxicológicas da UNIFAL. Compararam-se as matrizes de covariâncias representadas por duas populações distintas dadas pelas situações em que as mulheres foram submetidas à avaliação antes e após quatro semanas de exercícios físicos.

A estatística do teste foi determinada em função da razão de variâncias generalizadas representada pela razão dos determinantes (λ_1).

$$\lambda_1 = \frac{\max_j \left(\left|S_{jj}\right|\right)}{\min_j \left(\left|S_{jj}\right|\right)}, \qquad (1)$$

em que, S_{jj} correspondeu à matriz de somas de quadrados e produtos amostral da j-ésima população (j = 1, 2) .

Convém salientar que a geração das matrizes S_{jj} foi feita considerando a matriz dos desvios de cada observação, em relação à média X_d . O uso dessa matriz foi necessário para que não houvesse a influência de possíveis médias diferentes entre as 2 populações consideradas no estimador das referidas matrizes de covariâncias. Assim, para evitar que essas matrizes considerando as p-variáveis de uma determinada população fosse afetada por esse efeito, optou-se por fazer toda a inferência, considerando as observações X_d . A matriz S_{jj} foi estimada por:

$$S_{jj} = \left(X_d\right)^t Q X_d \qquad (2)$$

em que a matriz de projeção é dada por $Q = I-(\underset{\sim}{1}\underset{\sim}{1}^t)/N.$, sendo $\underset{\sim}{1}$ um vetor composto de 1.

A imposição da hipótese H_0: $S_{11}=S_{22}$ versus $S_{11} \neq S_{22}$ foi feita por meio da técnica *bootstrap* (CIRILLO, 2006). A aplicação do algoritmo se deu na amostra aleatória X_d, da qual estimou-se $S_{jj.}$ Em cada reamostragem, obteve-se uma nova amostra denominada por X_{db}, em que foi estimada a matriz de somas de quadrados e produtos, porém denominada por S_b^* em que o índice b representou a b-ésima (b=1,...,1000) reamostragem . Em cada simulação foram computados os valores $l_{1(b)}$ e confrontados com os valores de l_1 obtidos na amostra original (expressão 1). O valor-p foi determinado como proporção dos valores de $l_{1(b),}$ obtidos por meio da distribuição empírica originada pelo método *bootstrap* que superaram os respectivos valores da estatística proveniente da amostra original (expressão 1). Convém salientar que, nesse procedimento, evitaram-se todas as restrições dos métodos numéricos de maximização da função de verossimilhança, o que computacionalmente representou uma grande contribuição. Conforme mencionado anteriormente, as amostras foram avaliadas antes e após a realização de um exercício físico em quatro semanas, definindo-se, portanto, amostras de duas populações (k = 2). Em cada população foram mensurados 3 variáveis (p=3), respectivamente definida como proteínas T (g/dl); albumina (g/dl) e peróxido (nmol/g de proteína). Preliminarmente à comparação dessas matrizes, realizou-se o teste de normalidade multivariada de Royston (1983) das kp = 6 variáveis, uma vez que, a construção dos testes se deu considerando as populações normalmente distribuídas e dependentes, Trata-se de uma situação de dados pareados, portanto caracterizando uma dependência entre os valores amostrais, assim, as duas situações avaliadas representadas pela realização do exercício físico antes e após quatro semanas, no contexto desse trabalho, caracterizaram as populações, nas quais desejou-se comparar as matrizes de covariâncias.

O valor da estatística do teste de normalidade, representada pelo valor de W = 0,925. O valor-p, obtido foi 0,6615, sendo a hipótese de normalidade multivariada não rejeitada em um nível de significância de 5%.

Em relação à comparação das matrizes de covariâncias, as hipóteses estatísticas a serem investigadas nesse trabalho foram definidas por:

$$\begin{aligned} H_0 &: \Sigma_{antes} = \Sigma_{depois} \\ H_1 &: \Sigma_{antes} \neq \Sigma_{depois} \end{aligned} \qquad (3)$$

Seguindo uma análise exploratória dos dados amostrais, estimou-se a matriz de correlação para cada população por:

$$R_A = \begin{bmatrix} 1 & & \\ 0,046 & 1 & \\ -0,271 & -0,032 & 1 \end{bmatrix}; \; R_D = \begin{bmatrix} 1 & & \\ 0,420 & 1 & \\ -0,581 & 0,148 & 1 \end{bmatrix}. \qquad (4)$$

Com base nos resultados exploratórios, referentes às matrizes de correlações estimadas para cada população R_A e R_D pode-se verificar que as variáveis utilizadas nesse estudo, representadas pelas quantidades de proteínas T (g/dl); albumina (g/dl) e peróxido (nmol/g de proteína) observadas antes e após quatro semanas de exercícios físicos apresentaram uma baixa correlação.

A importância desse resultado, embora sendo exploratório, é dada em dois aspectos, e o primeiro a ser

ressaltado é a questão de que o teste proposto nesse trabalho irá comparar matrizes de covariâncias populacionais, cujas amostras resultaram em matrizes de correlação livre do efeito da multicolinearidade.

Um outro aspecto sugere que a escolha dessas variáveis foi feita de forma adequada, pois todas as variáveis são relevantes, no sentido de que nenhuma delas apresentou uma forte correlação. Caso uma forte correlação entre duas variáveis fosse identificada, possivelmente uma delas poderia ser excluída da análise.

Sem perder o foco de comparar as matrizes de covariâncias dependentes, procede-se com a execução do teste considerando X como a matriz dos dados. Dessa forma, tem-se o vetor de médias $\bar{\underset{\sim}{X}}$ calculado por:

$$\bar{\underset{\sim}{X}} = \begin{bmatrix} 6,53125 \\ 3,99375 \\ 64,2262 \\ 6,7375 \\ 3,76875 \\ 65,5731 \end{bmatrix} \quad (5)$$

Para cada observação, os desvios em relação à média de cada variável foram estimados, originando-se, assim a matriz dos desvios X_d dada por:

$$X_d = \begin{bmatrix} 6,0-6,53125 & 3,7-3,99375 & \cdots & 72,3-65,5731 \\ 6,0-6,53125 & 3,5-3,99375 & \cdots & 77,0-65,5731 \\ \vdots & \vdots & \vdots & \vdots \\ 6,6-6,53125 & 4,1-3,99375 & \cdots & 65,57-65,5731 \end{bmatrix} \quad (6)$$

Com base na matriz X_d calculou-se então a matriz de somas de quadrados e produtos determinada por S.

$$S = \begin{bmatrix} 2,394 & 0,143 & -26,351 & 1,611 & 0,605 & -26,039 \\ 0,143 & 4,009 & -4,015 & 0,053 & 1,436 & 0,844 \\ -26,351 & -4,015 & 3952,389 & -24,445 & -2,348 & 582,980 \\ 1,611 & 0,053 & -27,445 & 2,317 & 0,698 & -33,568 \\ 0,605 & 1,436 & -2,348 & 0,698 & 1,194 & 6,149 \\ -26,039 & 0,844 & 582,980 & -33,568 & 6,149 & 1442,669 \end{bmatrix} \quad (7)$$

As matrizes indicadas pelos blocos diagonais de S (3 x 3) representaram as estimativas das matrizes de somas de quadrados e produtos de cada população dados por:

$$S_{Antes} = \begin{bmatrix} 2,394 & 0,143 & -26,351 \\ 0,143 & 4,009 & -4,015 \\ -26,351 & -4,015 & 3952,389 \end{bmatrix}$$

$$S_{Depois} = \begin{bmatrix} 2,317 & 0,698 & -33,568 \\ 0,698 & 1,194 & 6,149 \\ -33,568 & 6,149 & 1442,669 \end{bmatrix} \quad (8)$$

Calculando as estatística do teste baseado na razão dos determinantes (9).

$$\lambda_1 = \frac{Det\left(S_{Antes}\right)}{Det\left(S_{Depois}\right)} = \frac{35069,412}{1566,8342} = 22,382 \quad (9)$$

Se for realizada uma avaliação preliminar do resultado obtido por meio da estatística l_1, pode-se supor que as matrizes de covariâncias populacionais sejam heterogêneas, pois se o valor de l_1 estivesse próximo de 1 ter se ia um forte indicativo que essas matrizes seriam homogêneas. Com o objetivo de realmente comprovar essa hipótese, dado um nível de significância fixado em 5%, procedeu-se com a obtenção do valor-p.

O algoritmo utilizado para impor a hipótese H_0 foi aplicado à matriz X_d. Para cada reamostragem, calculou-se novamente a matriz de somas de quadrados e produtos S_b, com b = 1, ..., B = 1000. Conseqüentemente, obteve-se 1000 valores para os critérios baseado na razão dos determinantes $l_{1(b)}$. A determinação do valor-p para a tomada de decisão em relação à rejeição ou não da hipótese H_0 foi realizada computacionalmente. Para isso, inicialmente definiu-se uma variável indicadora Z (10) por:

$$Z_i = \begin{cases} 1 \text{ se } \lambda_{1(b)} > \lambda_1 \\ 0 \text{ c.c.} \end{cases} \quad (10)$$

para b = 1, 2, ..., B.

Assim, obteve-se o valor-p dado por:

$$\text{valor} - p = \frac{\sum_{i=1}^{B} Z_i}{B} \quad (11)$$

Para o caso específico desse exemplo, obteve-se o valor-p de 0,014. Assim, considerando o nível de significância fixado em 5%, concluiu-se que o teste de razão de variâncias generalizadas baseada na razão dos

determinantes foi significativo ($p<0,014$). Concluiu-se que há evidências estatísticas ($p<0,05$) de que a covariância do efeito do exercício de força sobre o estresse oxidativo no plasma de mulheres, na terceira idade obtidos após quatro semanas de exercícios seja diferente da covariância da população inicialmente avaliada, ou seja antes do prazo de quatro semanas. Um outro importante resultado é que, caso seja feito testes como por exemplo T^2 de Hotteling (JOHNSON & WICHERN, 1998) para comparar performances do vetor de médias das populações classificadas em antes e depois deve-se considerar a heterogeneidade das matrizes de covariâncias ou testes alternativos, como os testes *bootstrap*.

REFERÊNCIAS BIBLIOGRÁFICAS

CIRILLO, M. A. **Propostas de testes multivariados para comparar matrizes de covariâncias de populações normais dependentes**. 2006. 111 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2006.

JIANG, G.; SARKAR, K. S.; HSUAN, F. A likelihood ratio test and its modifications for the homogeneity of the covariance matrices of dependent multivariate normals. **Journal of Statistical Planning and Inference**, [S.l.], v. 81, p. 95-111, 1999.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 4. ed. New Jersey: Prentice Hall, 1998.

LEVENE, H. **Contributions to probability and statistics**: essays in Honor of Harold Hotelling. [S.l.]: Stanford University, 1960.

MANLY, B. F. J. **Randomization, bootstrap and Monte Carlo methods in biology**. 2. ed. New Zealand: University of Otago, 1997. 356 p.

O'BRIEN, C. O. P. Robust procedures for testing equality of covariance matrices. **Biometrics**, Washington, v. 48, p. 819-827, 1992.

ROYSTON, J. P. Some techniques for assessing multivariate normality basead on the Shapiro-Wilk W. **Applied Statistics Journal of the Royal Statistical Society – Series C**, London, v. 32, n. 2, p. 121-133, 1983.