Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Avaliação Monte Carlo do teste para comparação de duas matrizes de covariâncias normais na presença de correlação

Vanessa Siqueira Peres da Silva [1] [2]
Daniel Furtado Ferreira [1]

1 Introdução

É comum em determinadas situações experimentais o pesquisador estar interessado em comparar matrizes de variância e covariância de duas populações. Se as duas amostras são independentes nenhuma covariância entre os dados é esperada. Por outro lado, os dados podem ser pareados em situações em que um grupo de variáveis é mensurado antes e após a realização de um determinado tratamento. Se o interesse focar na comparação das matrizes de covariância das populações, não é possível, por exemplo, aplicar o teste de Bartlett (1937). A razão para isso é que de alguma forma esses dados são correlacionados e o teste possui baixo poder em detectar possíveis diferenças na matriz de covariância. Para o caso em que apenas uma variável é mensurada em cada situação, pré (X) e pós (Y) tratamento, Morgan (1939) e Pitman (1939) propuseram um teste t exato baseado na correlação entre as variáveis normais X e Y e na correlação de duas novas variáveis que são combinações lineares de X e Y. O teste de Morgan (1939) e Pitman (1939) se torna bem mais poderoso do que as alternativas existentes para dados não correlacionados à medida que a correlação entre X e Y tende para 1 ou para -1. Esse teste considera, no entanto, apenas a situação de $q = 1$ população e $p = 1$ variável. Pitman (1939) e Morgan (1939) propuseram pela primeira vez o teste da razão de verossimilhança para a igualdade das variâncias, em uma população normal bivariada, ou seja, $(p = 1, q = 2)$ com correlação desconhecida. Desde então, muitos pesquisadores têm explorado ainda mais este problema. Para lidar com o problema de comparar matrizes de covariâncias de distribuições normais dependentes este trabalho foi proposto procurando generalizar o teste de Morgan (1939) e Pitman (1939) para o caso multivariado, considerando a situação de $q = 2$ populações.

2 Material e métodos

2.1 Teste de comparação de covariâncias na presença de correlação

Seja para isso a variável multidimensional $\boldsymbol{X}$ de dimensão $(2p \times 1)$ suposta normal multivariada com vetor média $\boldsymbol{\mu}$ $(2p \times 1)$ e covariância $\boldsymbol{\Sigma}$ $(2p \times 2p)$, $\boldsymbol{X} \sim N_{2p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Nesse trabalho somente o caso de $q = 2$ populações relacionadas é considerado, resultando na dimensão

[1] DEX - UFLA. e-mail: *dexvanessa@gmail.com*
[2] Agradecimento à FAPEMIG pelo apoio financeiro.

$qp = 2p$ das matrizes. Seja considerada uma partição de $\boldsymbol{X}$ em $q = 2$ grupos de p variáveis, quais sejam, $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ ambos de dimensões $(p \times 1)$. Igualmente, $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ devem ser particionadas de forma correspondente. Os grupos 1 e 2 são formados de tal forma que as p variáveis dos vetores $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ correspondam as mensurações antes e após à aplicação de algum tipo de tratamento. O principal objetivo é testar a hipótese especificada em (1):

$$H_0 : \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} \quad \text{vs} \quad H_1 : \boldsymbol{\Sigma}_{11} \neq \boldsymbol{\Sigma}_{22} \tag{1}$$

quando $\boldsymbol{\Sigma}_{12} \neq \mathbf{0}$. Os testes para o caso em que $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ são bastantes conhecidos na literatura (BARTLETT, 1937). Para desenvolver o teste especificamente para $q = 2$ populações, sejam definidas as variáveis U e V como combinações lineares das variáveis de $\boldsymbol{X}$. Para isso, seja o vetor não nulo $\boldsymbol{a}\,(p \times 1)$ de coeficientes reais que estabelecem as combinações lineares de $\boldsymbol{X}$. Sejam os vetores $\boldsymbol{b}^\top = (\boldsymbol{a}^\top, \boldsymbol{a}^\top)$ e $\boldsymbol{c}^\top = (\boldsymbol{a}^\top, -\boldsymbol{a}^\top)$ de dimensão $(1 \times 2p)$. Portanto sejam $U = \boldsymbol{b}^\top \boldsymbol{X} = \boldsymbol{a}^\top \boldsymbol{X}_1 + \boldsymbol{a}^\top \boldsymbol{X}_2 = U_1 + U_2$ e $V = \boldsymbol{c}^\top \boldsymbol{X} = \boldsymbol{a}^\top \boldsymbol{X}_1 - \boldsymbol{a}^\top \boldsymbol{X}_2 = U_1 - U_2$. Como U e V são combinações lineares de variáveis normais $\boldsymbol{X}$, então, $U \sim N_1(\mu_U, \sigma_U^2)$ e $V \sim N_1(\mu_V, \sigma_V^2)$ (ANDERSON, 1978). Sendo que:

$$\begin{aligned} \mu_U &= \boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2 \quad \text{e} \quad \sigma_U^2 = \boldsymbol{a}^\top (\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{22} + 2\boldsymbol{\Sigma}_{12})\boldsymbol{a}; \\ \mu_V &= \boldsymbol{a}^\top \boldsymbol{\mu}_1 - \boldsymbol{a}^\top \boldsymbol{\mu}_2 \quad \text{e} \quad \sigma_V^2 = \boldsymbol{a}^\top (\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{22} - 2\boldsymbol{\Sigma}_{12})\boldsymbol{a}. \end{aligned}$$

Para desenvolver o teste sugerido, inicialmente é obtida a correlação entre essas combinações lineares U e V (ρ_{UV}).

$$\rho_{UV} = \frac{\boldsymbol{a}^\top (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{22})\boldsymbol{a}}{\sqrt{\boldsymbol{a}^\top (\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{22} + 2\boldsymbol{\Sigma}_{12})\boldsymbol{a}}\sqrt{\boldsymbol{a}^\top (\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{22} - 2\boldsymbol{\Sigma}_{12})\boldsymbol{a}}}. \tag{2}$$

O estimador de ρ_{UV} é dado por

$$r_{UV} = \frac{\boldsymbol{a}^\top (\boldsymbol{S}_{11} - \boldsymbol{S}_{22})\boldsymbol{a}}{\sqrt{\boldsymbol{a}^\top (\boldsymbol{S}_{11} + \boldsymbol{S}_{22} + 2\boldsymbol{S}_{12})\boldsymbol{a}}\sqrt{\boldsymbol{a}^\top (\boldsymbol{S}_{11} + \boldsymbol{S}_{22} - 2\boldsymbol{S}_{12})\boldsymbol{a}}}. \tag{3}$$

Como $U \sim N(\mu_U, 1)$ e $V \sim N(\mu_V, 1)$, então a estatística pivô

$$t = \frac{r_{UV}\sqrt{n-2}}{\sqrt{1 - r_{UV}^2}} \tag{4}$$

tem distribuição t de Student com $\nu = n - 2$ graus de liberdade, se $\rho_{UV} = 0$. A inspeção da expressão (2) permite que se conclua que ρ_{UV} será nula se e somente se $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}$. Portanto, testar a hipótese $H_0 : \rho_{UV} = 0$ é equivalente a testar a hipótese $H_0 : \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}$. Quando, no entanto, $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, o teste de Bartlett (1937) pode ser utilizado para testar essa hipótese. Para definir qual é o valor do vetor $\boldsymbol{a}$ que torna o teste (4) mais poderoso foi considerado a razão de

formas quadráticas F, em que $\boldsymbol{S}_{11}$ e $\boldsymbol{S}_{22}$ são matrizes positivas definidas, dada por:

$$F = \frac{\boldsymbol{a}^\top \boldsymbol{S}_{11}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{S}_{22}\boldsymbol{a}}. \tag{5}$$

O próximo passo é expressar r_{UV} de (3) em função de F de (5).

$$r_{UV} = \frac{F-1}{\sqrt{(F+1)^2 - 4r_{U_1U_2}^2 F}}. \tag{6}$$

A substituição de (6) em (4) permite que t de Student seja expresso em função de F e de $r_{U_1U_2}$. O resultado final obtido é dado por:

$$t = \frac{(F-1)\sqrt{n-2}}{2\sqrt{F(1-r_{U_1U_2}^2)}}. \tag{7}$$

Essa estatística segue a distribuição t de Student com $\nu = n-2$ graus de liberdade, dado que $\boldsymbol{a}$ é conhecido. A partir desse instante é possível pensar em uma alternativa para obter o valor de $\boldsymbol{a}$ para o cálculo de t de (7). Poderia-se pensar em uma solução, maximizando o valor de F.

2.1.1 Maximizar F

Tomando-se o valor máximo de F em relação ao vetor $\boldsymbol{a}$, o qual é obtido derivando F de (5) em relação à $\boldsymbol{a}$ e igualando a derivada a zero, $\partial F/\partial \boldsymbol{a} = \boldsymbol{0}$. O resultado obtido é dado pelo sistema de equações homogêneas (8).

$$(\boldsymbol{S}_{11} - \hat{F}\boldsymbol{S}_{22})\boldsymbol{a} = \boldsymbol{0}. \tag{8}$$

Esse sistema de equações homogêneas pode ser resolvido (Bock, 1975) fazendo a transformação linear $\boldsymbol{a} = (\boldsymbol{\Gamma}^\top)^{-1}\boldsymbol{z}$, em que $\boldsymbol{\Gamma}$ é o fator de Cholesky de $\boldsymbol{S}_{22}$, tal que $\boldsymbol{S}_{22} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$. Pré-multiplicando o sistema resultante por $\boldsymbol{\Gamma}^{-1}$, obtém-se:

$$(\boldsymbol{H} - \hat{F}\boldsymbol{I})\boldsymbol{z} = \boldsymbol{0} \tag{9}$$

em que $\boldsymbol{H} = \boldsymbol{\Gamma}^{-1}\boldsymbol{S}_{11}(\boldsymbol{\Gamma}^\top)^{-1}$. Portanto o máximo é obtido no ponto correspondente ao maior autovalor de $\boldsymbol{H}$, diga-se, $\hat{F} = \lambda_1$ e de seu autovetor normalizado correspondente $\boldsymbol{z}$. O valor de $\hat{F}$ não é alterado pela transformação singular, mas o valor de $\boldsymbol{a}$ deve ser recuperado. Com os valores de $\boldsymbol{a}$ e de $\hat{F}$ que correspondem ao máximo da expressão (5) é possível obter a correlação $r_{U_1U_2}$ utilizando (10) e finalmente aplicar o teste (7).

$$r_{U_1U_2} = \frac{\boldsymbol{a}^\top \boldsymbol{S}_{12}\boldsymbol{a}}{\sqrt{(\boldsymbol{a}^\top \boldsymbol{S}_{11}\boldsymbol{a})}\sqrt{(\boldsymbol{a}^\top \boldsymbol{S}_{22}\boldsymbol{a})}}. \tag{10}$$

3 Resultados e discussões

A simulação Monte Carlo realizada neste trabalho teve como intuito avaliar as taxas de erro tipo I do teste t proposto, com o objetivo de compará-lo com os testes apresentados por Jiang e Sarkar (1998) (W_2 e W_5) e Jiang et. al, (1999) (LRT, LRT_1, LRT_2 e LRT_3). Em cada simulação foi aplicado o teste t em um nível nominal de significância (α) fixado em 5%, para verificar se a hipótese nula $H_0 : \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}$ quando $\boldsymbol{\Sigma}_{12} \neq \mathbf{0}$ foi ou não rejeitada. Em todos os casos, $N = 10000$ amostras foram simuladas e a proporção de rejeição da hipótese nula H_0 foi computada para o teste ao longo de todas as simulações Monte Carlo. Os valores da taxa de erro tipo I empírica foram comparados com o valor nominal fixado em cada um dos casos. O teste t foi aplicado a cada uma dessas amostras, em cada uma das configurações formadas pela combinação dos valores de n e $\boldsymbol{\Sigma}$, considerando o nível de significância $\alpha = 0{,}05$. As simulações foram realizadas no software R. Foram simuladas amostras aleatórias normais multivariadas de dimensão $(2p \times 1)$ com vetor média $\boldsymbol{\mu}$ $(2p \times 1)$ e covariância $\boldsymbol{\Sigma}$ $(2p \times 2p)$, $\boldsymbol{X} \sim N_{2p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Nessa simulação e na apresentada por Jiang et al. (1999) foram considerados os mesmos valores para os parâmetros como no estudo de Jiang e Sarkar (1998), tais como: $n = 10, 15, 20, 15, 50, 75, 100$,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_4, \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1{,}0000 & -1{,}0000 & 0{,}7071 & -0{,}7071 \\ -1{,}0000 & 5{,}0000 & -0{,}7071 & 0{,}7071 \\ 0{,}7071 & -0{,}7071 & 1{,}0000 & -1{,}0000 \\ -0{,}7071 & 0{,}7071 & -1{,}0000 & 5{,}0000 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 1{,}0 & 0{,}1 & 0{,}2 & 0{,}3 \\ 0{,}1 & 1{,}0 & 0{,}4 & 0{,}5 \\ 0{,}2 & 0{,}4 & 1{,}0 & 0{,}1 \\ 0{,}3 & 0{,}5 & 0{,}1 & 1{,}0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_4 = \begin{bmatrix} 1{,}0000 & -1{,}0000 & 0{,}7071 & -1{,}4142 \\ -1{,}0000 & 5{,}0000 & -0{,}7071 & 1{,}4142 \\ 0{,}7071 & -0{,}7071 & 1{,}0000 & -1{,}0000 \\ -1{,}4142 & 1{,}4142 & -1{,}0000 & 5{,}0000 \end{bmatrix},$$

para estimar os níveis de significância. Pode-se observar, na tabela 1, que os testes LRT_3 e W_2 se destacam por manter a taxa de erro tipo I empírica próxima do nível nominal fixado, mesmo quando $N = 10$.

4 Conclusões

Os testes LRT, LRT_1, LRT_2, W_5 e t são considerados liberais. E, quando $N \geq 20$ os testes LRT_2 e W_5 são recomendados, sendo que o teste LRT_2 tem uma ligeira vantagem sobre o teste W_5.

Tabela 1: Taxas de erro tipo I de sete testes de igualdade de matrizes de covariâncias: teste LRT, teste LRT_1, teste LRT_2, teste LRT_3, teste W_2, teste W_5, teste t, considerando diferentes tamanhos amostrais (n), diferentes matrizes de covariância $(\boldsymbol{\Sigma})$ e valor nominal de significância de 5%.

$\boldsymbol{\Sigma}$	n	LRT[a]	LRT_1[a]	LRT_2[a]	LRT_3[a]	W_2[b]	W_5[b]	t
$\boldsymbol{\Sigma}_1$	10	14,00	10,80	7,50	3,80	5,30	7,20	11,54
	15	9,20	7,40	5,70	3,80	4,20	6,10	10,46
	20	8,10	6,90	5,70	4,20	4,00	5,90	9,80
	25	7,30	6,30	5,50	4,30	4,00	5,70	8,90
	50	6,10	5,70	5,30	4,80	4,20	5,20	8,87
	75	5,60	5,30	5,10	4,70	4,50	5,10	7,92
	100	5,40	5,30	5,10	4,80	4,70	5,20	8,01
$\boldsymbol{\Sigma}_2$	10	13,50	10,20	7,10	3,70	3,70	7,40	9,50
	15	9,70	7,90	6,10	4,00	3,30	6,40	8,50
	20	8,30	7,20	6,00	4,50	3,30	6,00	8,24
	25	7,40	6,50	5,50	4,40	3,30	5,70	7,97
	50	6,20	5,80	5,30	4,80	4,00	5,40	7,11
	75	5,70	5,40	5,20	4,80	4,30	5,30	6,91
	100	5,50	5,30	5,10	4,80	4,50	5,20	6,87
$\boldsymbol{\Sigma}_3$	10	13,80	10,50	7,20	3,90	5,00	7,50	9,63
	15	9,50	7,80	6,00	4,00	4,20	6,50	8,62
	20	8,30	7,10	5,80	4,40	3,80	5,90	8,10
	25	7,30	6,40	5,50	4,40	4,00	5,90	8,08
	50	6,10	5,70	5,30	4,80	4,30	5,40	6,92
	75	5,90	5,60	5,30	5,00	4,50	5,30	7,16
	100	5,50	5,40	5,20	4,90	4,60	5,10	6,83
$\boldsymbol{\Sigma}_4$	10	13,20	10,30	7,00	3,30	4,00	7,50	8,95
	15	9,60	7,80	6,10	4,10	3,40	6,50	7,88
	20	8,30	7,00	5,90	4,30	3,40	6,00	7,62
	25	7,50	6,60	5,60	4,50	3,50	5,70	7,68
	50	6,10	5,70	5,30	4,80	4,10	5,40	6,79
	75	5,60	5,40	5,10	4,80	4,50	5,40	6,10
	100	5,50	5,30	5,10	4,90	4,60	5,20	6,17

[a] Taxas de erro tipo I dos testes LRT, LRT_1, LRT_2 e LRT_3 apresentados por Jiang, et. al (1999).
[b] Taxas de erro tipo I dos testes W_2 e W_5 apresentados por Jiang e Sarkar (1998).

Referências

[1] ANDERSON, T. W. Maximum Likelihood Estimation for vector autoregressive moving average models. **Technical Report**. Califórnia, n. 35, July 1978.

[2] BARTLETT, M. S. Properties of sufficiency and statistical tests. **Proc. R. Soc. A**. London, v. 160, p. 268-282, 1937.

[3] BOCK, R. D. **Multivariate statistical methods in behavioral research.** New York: McGraw-Hill, 1975. 658 p.

[4] JIANG, G.; SARKAR, S. K. Some asymptotic tests for the equality of the covariance matrices of two dependent bivariate normals.**Biom. J.**, v. 40, p. 205-225, 1998.

[5] JIANG, G.; SARKAR, S. K.; 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.

[6] MORGAN, W. A. A test for the significance of the difference between the two variances in a sample from a normal bivariate population. **Biometrika**, v. 31, p. 13-19, 1939.

[7] PITMAN, E. J. G. A note on normal correlation. **Biometrika**, v. 31, p. 9-12, 1939.