



**Simpósio de Métodos
Numéricos em Engenharia**

25 a 27 de outubro, 2017

Simulação de variáveis aleatórias beta correlacionadas via algoritmo NORTA

Ricardo Rasmussen Petterle

Programa de Pós-Graduação em Engenharia de Produção
Universidade Federal do Paraná (UFPR)

Curitiba, Brasil

E-mail: ricpetterle@gmail.com

Wagner Hugo Bonat

Departamento de Estatística/Laboratório de Estatística e Geoinformação
Universidade Federal do Paraná (UFPR)

Curitiba, Brasil

E-mail: wbonat@ufpr.br

Cassius Tadeu Scarpin

Departamento de Administração Geral e Aplicada
Universidade Federal do Paraná (UFPR)

Curitiba, Brasil

E-mail: cassiusts@gmail.com

Resumo—O objetivo deste artigo é avaliar o desempenho do algoritmo NORTA (*NOR*mal *TO* Anything) para simular vetores aleatórios cujas distribuições marginais sejam beta. A construção e simulação de distribuições beta multivariadas é difícil devido ao fato que o suporte das distribuições marginais é o intervalo unitário. Tal restrição impõe um relacionamento entre a média e a variância da distribuição que por sua vez induz restrições não triviais nos possíveis valores das correlações entre os componentes do vetor aleatório. Além disso, tais restrições dependem da média e da variância da distribuição marginal especificada. Entender tais restrições é fundamental para avaliar o comportamento e aplicabilidade do algoritmo de simulação. Os resultados desta avaliação mostram que o espaço paramétrico da correlação é bastante reduzido quando se tem baixos valores para os parâmetros de dispersão associados com altos/baixos valores das médias marginais. Estes resultados são importantes para o futuro delineamento de estudos de simulação usando o algoritmo NORTA para a avaliação de estimadores para os parâmetros do modelo beta multivariado.

Palavras-chave—Simulação estocástica; intervalo unitário; distribuição beta multivariada.

I. INTRODUÇÃO

Fenômenos aleatórios ocorrem em diversas áreas de pesquisa. Esses fenômenos, também chamados de variáveis aleatórias, ge-

ralmente são descritos por uma distribuição de probabilidades. Quando a variável aleatória é contínua e limitada ao intervalo unitário a distribuição de probabilidades beta é uma escolha usual. Variáveis com essas características são usadas em inúmeras aplicações, sobretudo, em modelos de regressão. Exemplos podem ser vistos na avaliação da distância genética entre vírus [1], em testes psicométricos [2], no estudo do Índice de Qualidade de Vida dos Trabalhadores da Indústria Brasileira - IQVT [3], dentre outras aplicações.

Para o caso de duas ou mais variáveis aleatórias, pode-se pensar numa distribuição de probabilidades bivariada ou multivariada, além da possível correlação existente entre elas. Porém, um dos problemas é especificar a distribuição conjunta do vetor aleatório [4].

Para simular dados de vetores aleatórios correlacionados vários métodos foram propostos. O algoritmo NORTA é um dos métodos mais usados para gerar vetores aleatórios correlacionados. O método foi descrito por [5] e usado em estudos de simulação por [6]. Adicionalmente, o algoritmo NORTA foi combinado com o método de redes neurais artificiais para gerar vetores aleatórios correlacionados com distribuições marginais de qualquer tipo, quando se tem problemas para resolver equações não lineares pelos métodos convencionais [7]. Mais recentemente [8] adaptou o algoritmo NORTA para

simulação de modelos marginais multivariados, no contexto de modelos lineares generalizados multivariados com estrutura de covariância (*multivariate covariance generalized linear models - McGLMs*), para análise de dados genéticos.

Estudos de simulação são necessários, pois permitem investigar o comportamento das variáveis aleatórias quando se tem controle sobre o mecanismo gerador dos dados. Além disso, pode-se estudar propriedades dos estimadores, tais como viés, consistência e taxa de cobertura.

Para o caso da distribuição beta bivariada, o principal problema é identificar qual é o valor máximo e mínimo que a matriz de correlação pode assumir em função das médias marginais. A distribuição beta, na sua forma reparametrizada [9], é modelada por dois parâmetros: um de média e outro de dispersão denotados, respectivamente, por μ e ϕ . Essa distribuição não é ortogonal, como é o caso da distribuição gaussiana. Assim, seus parâmetros são definidos por $p = \mu\phi$ e $q = (1 - \mu)\phi$.

Nesse contexto, o principal objetivo deste artigo é estudar o comportamento do algoritmo NORTA para simulação de variáveis aleatórias beta bivariada.

O artigo está organizado da forma que segue. A Seção 2 apresenta uma breve descrição da distribuição de probabilidade beta e suas propriedades. A Seção 3 apresenta o algoritmo NORTA. A Seção 4 apresenta os principais resultados do estudo de simulação. Por fim, a Seção 5 discute os principais resultados e apresenta sugestões para futuros trabalhos.

II. DISTRIBUIÇÃO DE PROBABILIDADE BETA

A distribuição de probabilidade beta é um modelo probabilístico para variáveis aleatórias contínuas e restritas ao intervalo $(0, 1)$. Essa distribuição possui dois parâmetros, ambos positivos, que controlam seu formato. Sua função densidade de probabilidade é expressa por:

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1}, \quad p > 0, \quad q > 0, \quad (1)$$

em que $0 < y < 1$ e $\Gamma(\cdot)$ denota a função gama. Sua média e variância são denotadas, respectivamente, por:

$$E(y) = \frac{p}{p+q} \quad \text{e} \quad \text{Var}(y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

Em modelos de regressão é de interesse relacionar a média de uma variável resposta y em função de uma ou mais covariáveis. Assim, [9] propuseram uma reparametrização da densidade beta, na qual pode-se modelar diretamente sua resposta média em função de covariáveis, $\mu = p/(p+q)$. Além disso, incluiu-se o parâmetro $\phi = p+q$, sendo este um parâmetro de precisão ou dispersão. Reescrevendo, tem-se: $p = \mu\phi$ e $q = (1 - \mu)\phi$. Consequentemente, sua densidade fica da seguinte forma:

$$f(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad (2)$$

em que $0 < \mu, y < 1$ e $\phi > 0$. Logo, a média e a variância da densidade beta na sua forma reparametrizada são:

$$E(y) = \mu \quad \text{e} \quad \text{Var}(y) = \frac{\mu(1-\mu)}{1+\phi}.$$

Para ilustrar a densidade beta nessa reparametrização, a Figura 1 apresenta a densidade beta para diferentes valores de μ combinados com dois valores do parâmetro de dispersão, $\phi = (3 \text{ e } 80)$. Desta figura, observa-se que quanto menor o valor de ϕ maior é a dispersão (gráfico A). Porém, para valores mais altos do parâmetro de dispersão (gráfico B) a densidade beta fica mais concentrada em torno da média.

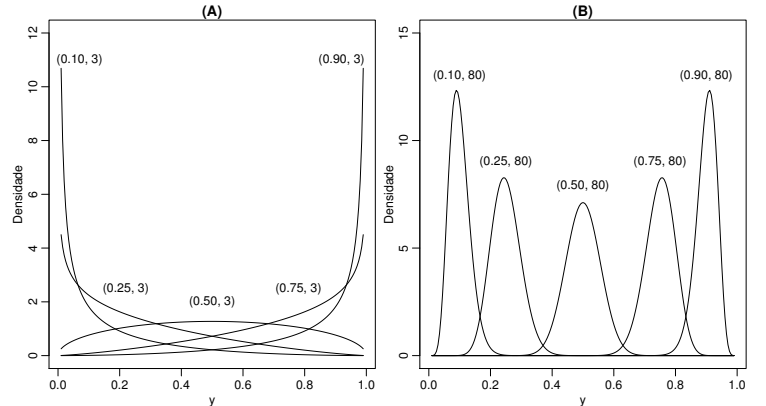


Figura 1: Função densidade de probabilidade beta para valores de μ iguais a 0,10, 0,25, 0,50, 0,75, 0,90, combinados com ϕ iguais a 3 e 80.

Fonte: Elaborada pelos autores

Nota: Figura 1 A: $\phi = 3$ e Figura 1 B: $\phi = 80$

III. ESTUDO DE SIMULAÇÃO

O algoritmo NORTA é um dos métodos mais populares para simulação de vetores aleatórios correlacionados. O método funciona como um processo em dois passos. Primeiro é gerado um vetor aleatório normal multivariado $\mathbf{Z}$ de dimensão $p \times 1$. Logo então, esse vetor é transformado num vetor uniforme multivariado $\mathbf{U}$, que é novamente transformado no vetor $\mathbf{Y}$ que tem distribuição NORTA (*NORmal To Anything*), em que cada elemento do vetor tem distribuição marginal arbitrária desejada. Logo, sua representação é dada por:

$$\mathbf{Y} = \left[F_{Y_1}^{-1}(\Phi[\mathbf{Z}_1]), F_{Y_2}^{-1}(\Phi[\mathbf{Z}_2]), \dots, F_{Y_p}^{-1}(\Phi[\mathbf{Z}_p]) \right]^\top, \quad (3)$$

onde $\Phi[\cdot]$ é a função de distribuição acumulada da distribuição normal padrão aplicada a cada elemento do vetor $\mathbf{Z}$ e $F_{Y_i}^{-1}(u) \equiv \inf\{y : F_{Y_i}(y) \geq u\}$ é a inversa da função de distribuição acumulada.

A matriz de correlação de $\mathbf{Z}$ determina diretamente a matriz de correlação de $\mathbf{Y}$, desde que

$$\rho_Y(l, l') = \text{Corr}(Y_l, Y_{l'}) = \text{Corr}(F_{Y_l}^{-1}(\Phi[\mathbf{Z}_l]), F_{Y_{l'}}^{-1}(\Phi[\mathbf{Z}_{l'}])),$$

para todo $l \neq l'$. A correlação é definida por:

$$\text{Corr}(Y_l, Y_{l'}) = \frac{E(Y_l, Y_{l'}) - E(Y_l)E(Y_{l'})}{\sqrt{\text{Var}(Y_l)\text{Var}(Y_{l'})}}, \quad (4)$$

onde as quantidades marginais $E(Y_l)$, $E(Y_{l'})$, $\text{Var}(Y_l)$ e $\text{Var}(Y_{l'})$ são definidas por F_{Y_l} e $F_{Y_{l'}}$. Vale destacar que $(Z_l, Z_{l'})$ têm distribuição normal padrão bivariada com correlação $\text{Corr}(z_l, z_{l'}) = \rho_Z(l, l')$ e que a quantidade $E(Y_l, Y_{l'})$ (4) é calculada por:

$$\begin{aligned} E(Y_l, Y_{l'}) &= E(F_{Y_l}^{-1}(\Phi(Z_l))F_{Y_{l'}}^{-1}(\Phi(Z_{l'}))) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{Y_l}^{-1}(\Phi(z_l))F_{Y_{l'}}^{-1}(\Phi(z_{l'}))\varphi_{\rho_Z(l, l')}(z_l, z_{l'})dz_l dz_{l'}, \end{aligned} \quad (5)$$

em que $\varphi_{\rho_Z(l, l')}$ denota a função densidade de probabilidade de uma distribuição normal padrão bivariada com correlação $\rho_Z(l, l')$.

Cabe ressaltar, que nem sempre a integral (5) vai ter solução devido à relação média/variação da distribuição beta. Isso impacta diretamente no espaço paramétrico da correlação (4) que também depende da especificação das médias marginais.

Todo estudo de simulação foi conduzido no *software* estatístico R [10] com auxílio do pacote NORTARA [11].

IV. RESULTADOS

Nesta seção apresentamos os principais resultados do estudo de simulação. Os resultados são ilustrados por meio de gráficos com as regiões das correlações mínima e máxima.

A função `valid_input_cormat()` do pacote NORTARA retorna os valores mínimo (ρ_L) e máximo (ρ_U) que a matriz de correlação pode assumir em função das distribuições marginais. Por exemplo, no caso da distribuição beta, quando $\mu_1 = 0,495$ e $\mu_2 = 0,851$ com $\phi = 0,00001$ (fixo) tem-se $\rho_L = -0,421$ e $\rho_U = 0,413$. Por outro lado, para os mesmos valores das médias marginais, porém com $\phi = 9$ (fixo), tem-se $\rho_L = -0,956$ e $\rho_U = 0,954$. Isso mostra como o parâmetro de dispersão afeta diretamente a média das distribuições marginais e, conseqüentemente, os limites da matriz de correlação.

Com o objetivo de tornar essa ideia mais geral, construiu-se as Figuras 2 e 3. A primeira figura mostra os limites mínimos da correlação, enquanto a segunda figura apresenta os limites máximos. Simulou-se dados para duas variáveis aleatórias beta Y_1 e Y_2 fixando-se diferentes valores para os parâmetros de dispersão $\phi = (0,00001, 0,25, 0,666, 1,5, 4 \text{ e } 9)$.

De acordo com os resultados apresentados na Figura 2, quando se tem valores baixos para ϕ a correlação mínima obtida fica muito restrita. Na medida em que o parâmetro de dispersão vai aumentando, correlações mais fortes são observadas nas regiões mais escuras do gráfico. O mesmo é observado para os valores máximos da correlação (Figura3). Dessa forma, observou-se que altos valores do parâmetro de dispersão combinados com altos/baixos valores das médias marginais produzem correlações mais fortes.

De maneira geral, pode-se ver que correlações mais fortes se

concentram no centro dos gráficos, enquanto que correlações mais fracas aparecem nas bordas.

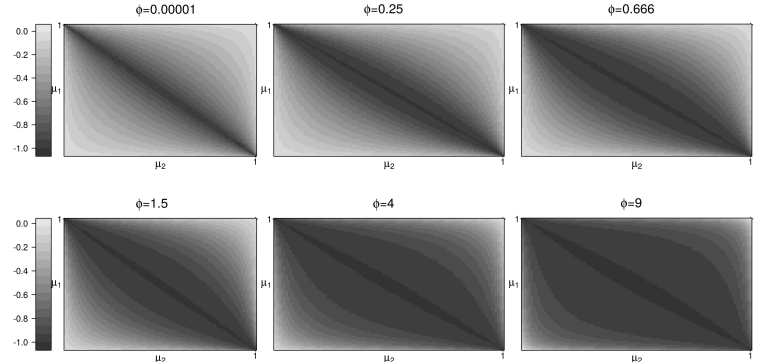


Figura 2: Valores mínimos da correlação entre duas variáveis aleatórias beta em função das suas médias marginais para diferentes valores de ϕ .

Fonte: Elaborada pelos autores

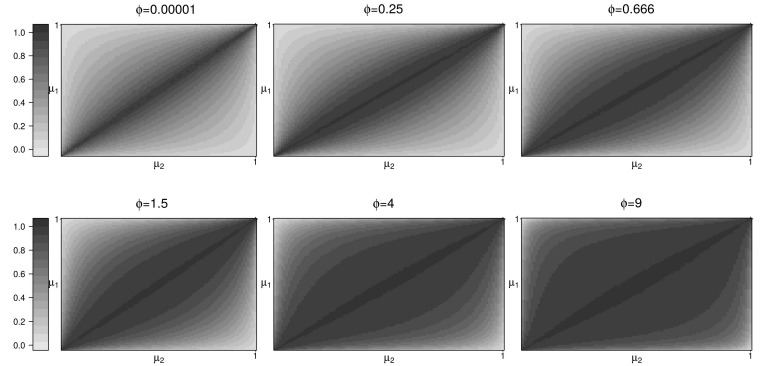


Figura 3: Valores máximos da correlação entre duas variáveis aleatórias beta em função das suas médias marginais para diferentes valores de ϕ .

Fonte: Elaborada pelos autores

V. DISCUSSÃO

Este artigo avaliou o desempenho do algoritmo NORTA na simulação de variáveis aleatórias beta correlacionadas. A simulação de distribuições beta multivariadas é uma tarefa desafiadora devido a certas restrições. Primeiro, o suporte das distribuições marginais é limitado ao intervalo $(0, 1)$. Segundo, na distribuição beta existe uma relação entre média e variância. Terceiro, dependendo da especificação das distribuições marginais pode ser difícil simular vetores aleatórios dessa distribuição. Em especial, quando se tem baixos valores do parâmetro de dispersão.

Os resultados do estudo de simulação mostraram que o espaço paramétrico da correlação ficou reduzido, quando se tem baixos valores para os parâmetros de dispersão associados com altos/baixos valores das médias marginais. Diante dos resultados obtidos, sugere-se como futuros trabalhos usar o algoritmo NORTA em estudos

de simulação para avaliar o comportamento e propriedades de estimadores para os parâmetros do modelo de regressão beta multivariado.

Houve a tentativa de simular vetores aleatórios da distribuição de probabilidades simplex, que é uma alternativa à distribuição beta, porém os resultados não foram satisfatórios. Nesse caso, o algoritmo NORTA exigiu um excessivo tempo computacional e não alcançou o objetivo proposto.

REFERÊNCIAS

- [1] A.J. Branscum, W.O. Johnson, M.C. Thurmond. Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian and New Zealand Journal of Statistics*, v.49, n.3, p.287-301, 2007.
- [2] M.J. Smithson, J. Verkuilen. A better lemon squeezer? Maximum likelihood regression with beta-distributed dependent variables. *Psychological Methods*, v.11, n.1, p.54-71, 2006.
- [3] W.H. Bonat, P.J. Ribeiro Jr, W.M. Zeviani. Regression models with response on the unit interval: Specification, estimation and comparison, *Biometric Brazilian Journal*, v.30, n.4, p.415-431, 2012.
- [4] R.A.P. Dias. Simulação estocástica de variáveis aleatórias Poisson correlacionadas: aplicação ao controle populacional do percevejo (*Euschistus heros Fabricius*) da soja (*Glycine max L.*), PhD thesis, Universidade de São Paulo, 2014.
- [5] M.C. Cario, B.L. Nelson. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix, *Technical report*, Citeseer, 1997.
- [6] H. Chen. Initialization for norta: Generation of random vectors with specified marginals and correlations, *INFORMS Journal on Computing*, v.13, n.4, p.312-331, 2001.
- [7] S.T.A. Niaki, B. Abbasi. Generating correlation matrices for normal random vectors in NORTA algorithm using artificial neural networks, *Journal of Uncertain Systems*, v.2, n.3, p.192-201, 2008.
- [8] W.H. Bonat. Modelling Mixed Types of Outcomes in Additive Genetic Models. *International Journal of Biostatistics*, v.x, p.1-16, 2017.
- [9] S. Ferrari, F. Cribari-Neto. Beta regression for modelling rates and proportions, *Journal of Applied Statistics*, v.31, n.7, p.799-815, 2004.
- [10] R Development Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [11] P. Su. NORTARA: Generation of Multivariate Data with Arbitrary Marginals, R package version 1.0.0, 2014.