FÁBIO MATHIAS CORRÊA

BIBLIOTECA R PARA A ANÁLISE BAYESIANA DE DADOS CATEGORIZADOS USANDO MODELOS MISTOS DE LIMIAR

LAVRAS - MG

2013

FÁBIO MATHIAS CORRÊA

BIBLIOTECA R PARA A ANÁLISE BAYESIANA DE DADOS CATEGORIZADOS USANDO MODELOS MISTOS DE LIMIAR

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração Estatística e Experimentação Agropecuária, para a obtenção do título de doutor.

Orientador
Dr. Júlio Sílvio de Sousa Bueno Filho

LAVRAS - MG

2012

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Corrêa, Fábio Mathias.
Biblioteca R para a análise bayesiana de dados categorizados usando modelos mistos de limiar / Fábio Mathias Corrêa. – Lavras : UFLA, 2013.
66 p.: il.

Tese (doutorado) – Universidade Federal de Lavras, 2012.
Orientador: Júlio Sílvio de Sousa Bueno Filho.
Bibliografia.

1. Análise threshold. 2. Biblioteca R. 3. MCCM. I. Universidade Federal de Lavras. II. Título.

CDD – 519.7

FÁBIO MATHIAS CORRÊA

BIBLIOTECA R PARA A ANÁLISE BAYESIANA DE DADOS CATEGORIZADOS USANDO MODELOS MISTOS DE LIMIAR

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração Estatística e Experimentação Agropecuária para a obtenção do título de doutor.

APROVADA em 14 de Novembro de 2012.

Dr. Edwin Moises Marcos Ortega	Esalq-USP
Dr. Eric Batista Ferreira	UNIFAL-MG
Dr. Renato Ribeiro de Lima	UFLA
Dra. Thelma Sáfadi	UFLA

Dr. Júlio Sílvio de Sousa Bueno Filho

Orientador

LAVRAS - MG
2012

AGRADECIMENTOS

À minha companheira e esposa Raquel, que se sempre esteve ao meu lado e que carrega em seu ventre o mais novo membro da família. Meu muito obrigado!

À minha filha Maria Clara Mello Corrêa (MCMC), que me enche de alegria e que deu um novo significado a sigla MCMC;

Ao filhão João Gabriel que está, ainda, no ventre de sua mãe, mas já nos enche de alegrias;

Ao Prof. Julio Sílvio de Sousa Bueno Filho, pela orientação, amizade e paciência durante estes anos;

Aos amigos Ivan e Fernanda, pela amizade e convivência;

Aos amigos Manoel, Luciana e Dona Abigail por, toda amizade e companheirismo;

Aos amigos Brou, Wyzy, Rose e Deyse, pelo convívio e por todos os momentos bons que passamos;

Ao grupo de discussão R_br, por toda a colaboração;

Aos professores do DEX-UFLA pelo auxílio na minha formação;

Ao Prof. Daniel Furtado Ferreira, pelas boas conversas que tivemos;

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas;

Ao CNPq, pelo apoio financeiro;

E a todos aqueles que, de forma direta ou indireta, contribuíram para meu êxito pessoal e profissional.

RESUMO

Nesta tese avaliamos os algoritmos para a análise bayesiana de modelos mistos em dados categorizados ordinais, bem como a sua implementação na biblioteca Bayesthresh para o ambiente de programação R. O pacote Bayesthresh apresenta uma estrutura flexível para inserção de modelos mistos e utiliza-se do processos de Monte Carlo via Cadeias Markov (MCCM) para obtenção das aproximações numéricas das distribuições a posteriori para os parâmetros do modelo. O estudo sobre a eficiência dos algoritmos implementados no pacote avaliou o tempo de processamento, a dependência das cadeias MCCM geradas e efeitos de sensibilidade a especificação das distribuições "a priori"para as componentes da variância. Adicionalmente foram calculados os erros (viés e Erro Quadrático Médio - EQM) das estimativas a posteriori obtidas para os efeitos fixos, efeitos aleatórios e componentes da variância. Um exemplo é apresentado de um experimento com variedades de tomateiro cujo objetivo é a seleção para a resistência à requeima causada pelo fungo Phytophthora infestans. Os algoritmos descritos por Nandram e Chen (1996) e sua modificação para a introdução da distribuição t para o traço latente foram os mais rápidos e precisos, porém, para experimentos simulados com correlação intraclasse igual a 0.8 (valor alto na prática), estes algoritmos tenderam a superestimar esta correlação. Na ilustração são apresentadas estimativas úteis para a seleção de variedades bem como o equivalente bayesiano de um teste para decidir se o traço latente tem distribuição gaussiana, não encontrando evidências em contrário. Em artigo adicional explora-se a utilização da ferramenta construída com exemplos de possibilidades de análise. Nesta ilustração foi analisado um experimento sensorial de conservas de banana desidratadas sob diferentes concentrações de açúcar, descrito por Silva (2008).

Palavras-chave: Análise bayesiana. Análise threshold. Biblioteca R. Modelos mistos. MCCM

ABSTRACT

In this thesis we evaluate algorithms for Bayesian analysis of ordinal categorical data as well as their implementation in the library Bayesthresh for the R statistical programing environment. Bayesthresh package has a flexible structure to insert mixed models and uses Markov Chain Monte Carlo (MCMC) sampling to approximate posterior distributions for model parameters. Simulation study of the efficiency of the algorithms considered processing time, dependency in the MCMC sampling chains and sensitivity to prior specifications for the variance components. Mean Squared Error (MSE) and average Bias in the marginal posterior distributions where also evaluated. An example is discussed from a breeding experiment for resistance to late blight (Phytophthora infestans) in varieties of tomato. Algorithms described by Nandran and Chen (1996) and derived algorithms were the faster and more accurate, although slightly overestimating intraclass correlation (ρ) in experiments with higher parametric values ($\rho = 0.8$). In the example useful estimates for plant breeding are presented, as Bayes factor test to decide on latent trait having Gaussian distribution. Gaussian distribution was as likely as Student's t distribution. An additional paper explores the uses of the modeling tool with an example from sensory analysis of dehydrated banana candy recipes with different sugar content, described by Silva (2008).

Keywords: Bayesian analysis. Mixed models. R package. Threshold models. MCMC

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

Variáveis discretas são classificadas como ordinais quando os seus níveis podem ser ordenados de forma a representar algum tipo de mensuração, como por exemplo, a escala de dureza de um mineral, que varia de 1 a 10, escalas diagramáticas utilizadas para avaliação da severidade de doenças de plantas, notas de escore corporal de animais e a escala hedônica utilizada em análise sensorial de alimentos. Durante o século passado, a forma mais comum de análise prática deste tipo de variável foi a aproximação normal após transformações que preservassem grosseiramente sua escala. Recentemente, diversas metodologias de análises para dados categorizados ordinais são propostas, tanto no âmbito da inferência Bayesiana com os modelos hierárquicos, quanto no âmbito da inferência fiducial com os modelos generalizados mistos. É crescente a percepção de que os modelos com efeitos aleatórios têm melhores propriedades explicativas em experimentos planejados.

Generalizações de efeitos aleatórios para variáveis não Gaussianas apresentam complicações matemáticas quanto ao processo de integração utilizado para estimar seus efeitos. Uma forma de contornar o problema de integração é o uso dos métodos de Monte Carlo via Cadeias Markov (MCMC). O uso destes algoritmos em análises de rotina é dificultado pela pouca disponibilidade de softwares que possuem os mesmos implementados. O pacote MCMCglmm (HADFIELD, 2010) disponível para o software R (R DEVELOPMENT CORE TEAM, 2012) possui o algoritmo proposto por Cowles (1996) para uso. Porém, não foi realizado nenhum estudo que avalie a eficiência dos algoritmos propostos por Albert e Chib (1993), Sorensen et al. (1995), Cowles (1996) e Nandran e Chen (1996) e as extensões proposta por Silva e Bueno-Filho (2010).

Visando facilitar o uso desta metodologia, objetivou-se com a presente tese de doutoramento avaliar a eficiência dos algoritmos existentes para análises de modelos mistos em dados categorizados e implementar um pacote para uso no ambiente de programação estatística R que apresente uma estrutura flexível para a inserção de modelos mistos.

2 REFERENCIAL TEÓRICO

Variáveis ordinais representam uma escala de grandeza, sendo expressa por valores inteiros, de forma que o resultado final seja a ordenação dos elementos avaliados (STEVENS, 1968). Este tipo de variável pode ser encontrada em diversas áreas do conhecimento (CORRÊA; BUENO-FILHO; CARMO, 2009; GOB; MCCOLLIN; RAMALHOTO, 2007; PIEPHO; KALKA, 2003; SORENSEN et al., 1995; STEVENS, 1968; TABOR, 1954), porém a atribuição da variável ao elemento avaliado é realizada de forma subjetiva, principalmente pela dificuldade em se quantificar determinadas características com o uso da medição.

As análises de variáveis ordinais, em sua grande maioria, não consideram a observação feita por Stevens (1968) sobre a invariância de que uma variável ordinal, que pode ser transformada desde que a sua informação empírica seja preservada, pois a maior parte das análises consideram a variável ordinal como uma variável contínua. Diante dos avanços computacionais ocorridos nos últimos 30 anos, diversas formas de análise para variáveis ordinais foram propostas, dentre elas o uso de modelos generalizados e generalizados mistos (BRESLOW; CLAYTON, 1993; MCCULLOCH; SEARLE, 2001), modelos hierárquicos bayesianos (BROWNE; DRAPER, 2006a), modelos thresholds baseados na verossimilhaça (BROCKHOFF; CHRISTENSEN, 2010; PIEPHO; KALKA, 2003) e modelos thresholds bayesianos (ALBERT; CHIB, 1993; COWLES, 1996; KIZILKAYA et al., 2003; NANDRAN; CHEN, 1996; POON; WANG, 2012; SILVA; BUENO-FILHO, 2010).

Dentre as diversas metodologias de análise, os modelos thresholds bayesianos são considerados extremamente flexíveis e utilizados em diversas áreas de pesquisa (ALBERT; CHIB, 1993; COWLES, 1996; NANDRAN; CHEN, 1996; KIZILKAYA et al., 2003; SILVA; BUENO-FILHO, 2010). Este tipo de modelo pode ser especificado de forma que uma resposta é observada em uma dada categoria se o valor desta variável está entre os limites que definem tal categoria (SILVA; BUENO-FILHO, 2010).

Apesar das recomendações para o uso de modelos thresholds bayesianos na análise de dados categorizados ordinais, problemas de convergência nos parâmetros thresholds são observados, tornando a análise extremamente lenta (ALBERT;

CHIB, 1993). Visando contornar este problema, alguns autores sugerem modificações nos algoritmos de forma a torná-los mais eficazes no processo de convergência dos parâmetros thresholds (COWLES, 1996; KIZILKAYA et al., 2003; NANDRAN; CHEN, 1996; POON; WANG, 2012; SILVA; BUENO-FILHO, 2010).

Albert e Chib (1993) foram os primeiros autores a utilizarem um modelo threshold para análise de dados binários ou com múltiplas categorias utilizando inferência Bayesiana, porém, devido a complicações no processo de convergência, modificações no algoritmo original foram propostas (COWLES, 1996; KIZILKAYA et al., 2003; NANDRAN; CHEN, 1996; SORENSEN et al., 1995).

Inicialmente, Albert e Chib (1993) propuseram a amostragem de Gibbs para a obtenção da distribuição marginal dos parâmetros do modelo utilizando a distribuição t-Student acumulada. Sorensen et al. (1995) adaptaram o modelo proposto por Albert e Chib (1993), porém, utilizando a distribuição normal acumulada como função de ligação. Um problema evidenciado por Sorensen et al. (1995) neste algoritmo é a forte autocorrelação entre as estimativas geradas no processo de amostragem de Gibbs, comprometendo o tempo de processamento devido a necessidade de obtenção de longas cadeias. Na tentativa de amenizar a forte autocorrelação, Cowles (1996) propôs um algoritmo que a atualização dos parâmetros de limiar e da variável latente é feita de forma a aceitar ou rejeitar um vetor completo com todos estes parâmetros. Neste caso, utiliza-se a amostragem de Gibbs com um passo de Metropolis-Hastings para o vetor em questão. Apesar do algoritmo proposto por Cowles (1996) apresentar melhores propriedades de convergência que os algoritmos propostos por Albert e Chib (1993) e Sorensen et al. (1995), a variância da distribuição geradora de candidatos para a variável latente é de difícil obtenção.

Devido à dificuldade de obtenção da variância para a distribuição geradora de candidatos para os parâmetros de limiar e a fim de melhorar o processo de amostragem, com a redução da autocorrelação entre as amostras, Nandran e Chen (1996) propuseram um algoritmo para amostrar parâmetros associados aos de limiar limitados entre 0 e 1, tendo como função geradora de candidatos para tais parâmetros a distribuição de Dirichlet e a distribuição Gaussiana para a variável latente.

Muitos autores dedicam-se à melhoria dos algoritmos para análise de modelos thresholds bayesianos, porém, eles não são utilizados em análises de rotina

devido a pouca disponibilidade de softwares que contemplem tais algoritmos. Os poucos softwares que disponibilizam a análise de modelos thresholds bayesianos podem ser enumerados, sendo o WinBUGS (SPIEGELHALTER; THOMAS; LUNN, 2003), MLwiN (BROWNE, 2011), glmmBUGS (BROWN, 2010) e o MCMCglmm (HADFIELD, 2010), porém, nenhum destes softwares possuem a implementação do algoritmo descrito por Nandran e Chen (1996) para análise de modelos thresholds mistos bayesianos ou dos demais algoritmos descritos de forma a facilitar a análise para o usuário final.

3 CONSIDERACÕES GERAIS

As generalizações teóricas propostas no presente trabalho se mostraram eficazes na melhoria do processo de convergência dos parâmetros thresholds para análise de dados categorizados ordinais, sendo que nos estudos de simulação realizados e no exemplo de aplicação não houve diferenças entre a distribuição Gaussiana acumulada ou da t-Student acumulada como função de ligação.

O software desenvolvido para a análise de modelos thresholds mistos bayesianos implementado para uso em software R versão 2.15.1 ou versão superior se mostoru uma ferramenta de uso fácil e acessível, podendo ser utilizado em análises de rotina. O presente software, denominado Bayesthresh, já está disponível no site http://cran.r-project.org/web/packages/Bayesthresh/index.html

O pacote apresentado nesta tese é passvel de atualizações e o seu uso é de inteira responsabilidade do usuário final.

Os capítulos seguintes desta tese constituem dois artigos que resumem o estudo da análise bayesiana de modelos threshold por nós implementado. No primeiro artigo comparamos os algoritmos em um estudo de simulação para tamahos experimentais representativos de potenciais usos da metodologia. No segundo artigo apresentamos uma biblioteca R para a análise bayesiana de modelos mistos threshold.

REFERÊNCIAS

ALBERT, J. H.; CHIB, S. Bayesian analysis of binary and polychotomous response data. **Journal of the American Statistical Association**, v. 88, n. 442, p. 669–679, 1993.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, v. 88, n. 421, p. 9–25, 1993.

BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. Thurstonian models for sensory discrimination tests as generalized linear models. **Food Quality and Preference**, v. 21, n. 3, p. 330–338, 2010.

BROWN, P. **glmmBUGS: Generalised Linear Mixed Models and Spatial Models with WinBUGS, BRugs, or OpenBUGS**. [S.l.], 2010. R package version 1.9. Disponível em: <http://CRAN.R-project.org/package=glmmBUGS>.

BROWNE, W. J. **MCMC estimation in MLwiN**. Bristol, UK, 2011. Version 2.24. Disponível em: <http://www.bristol.ac.uk/cmm/software/mlwin/>.

BROWNE, W. J.; DRAPER, D. A comparison of bayesian and likelihood-based methods for fitting multilevel models. **Bayesian analysis**, v. 1, n. 3, p. 473–514, 2006.

BROWNE, W. J.; DRAPER, D. A comparison of bayesian and likelihood-based methods for fitting multilevel models. **Bayesian Analysis**, v. 1, n. 3, p. 473–514, 2006.

CORREA, F. M.; BUENO-FILHO, J. S. de S. **Bayesthresh: A package for categorical data analysis using Bayesian inference**. [S.l.], 2012.

CORRÊA, F. M.; BUENO-FILHO, J. S. S.; CARMO, M. G. F. Comparison of the three diagrammatic key for the quantification of late blight in tomato leaves. **Plant Pathology**, v. 58, n. 6, p. 1128–1133, 2009.

COWLES, M. K. Accelerating monte carlo markov chain convergence for cumulative link generalized linear models. **Statistics and Computing**, v. 6, n. 2, p. 101–111, 1996.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian data analysis**. [S.l.]: London: Chapman and Hall, 2003. 668 p.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, n. 4, p. 457–511, 1992.

GOB, R.; MCCOLLIN, C.; RAMALHOTO, M. F. Ordinal methodology in the analysis of likert scales. **Quality and Quantity**, v. 41, n. 5, p. 601–626, 2007.

HADFIELD, J. D. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. **Journal of Statistical Software**, v. 33, n. 2, p. 1–22, 2010. Disponível em: <http://www.jstatsoft.org/v33/i02/>.

HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Applied Statistics**, v. 32, n. 1, p. 69–83, 1976.

JEFFREYS, H. **Theory of probability**. [S.l.]: Oxford: Claredon Press, 1961. 470 p.

KIZILKAYA, K.; CARNIER, P.; ALBERA, A.; BITTANTE, G.; TEMPELMAN, R. J. Cumulative t-link threshold models for genetic analysis of calving ease scores. **Statistics and Computing**, v. 35, n. 1, p. 489–512, 2003.

MCCULLOCH, C. E.; SEARLE, S. R. **Generalized, linear, and mixed models**. [S.l.]: Wiley series in probability and statistics, 2001. 325 p.

MCCULLOGH, C. E.; SEARLE, S. R. **Generalized, Linear and Mixed Models**. [S.l.]: Wiley, New York, 2001.

NANDRAN, B.; CHEN, M. Reparameterizing the generalized linear model to accelerate gibbs sample convergence. **Journal of Statistical Computation and Simulation**, v. 54, n. 1, p. 129–144, 1996.

PIEPHO, H.-P.; KALKA, E. Threshold models with fixed and random effects for ordered categorical data. **Food Quality and Preference**, v. 14, n. 1, p. 343–357, 2003.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for mcmc. **R News**, v. 6, n. 1, p. 7–11, 2006. Disponível em: <http://CRAN.R-project.org/doc/Rnews/>.

POON, W.-Y.; WANG, H.-B. Latent variable models with ordinal categorical covariates. **Statistics and Computing**, v. 22, n. 5, p. 1135–1154, 2012.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2012. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

RAFTERY, A. E.; LEWIS, S. M. One long run with diagnostics: Implementation strategies for markov chain monte carlo. **Statistical Science**, v. 7, n. 4, p. 493–497, 1992.

SILVA, J. W. **Algoritmos para modelos de limiar utilizando as distribuições acumuladas Normal e t de Student**. Tese (Tese de Doutorado) — Universidade Federal de Lavras, 2008.

SILVA, J. W.; BUENO-FILHO, J. S. S. Um algoritmo para modelos de limiar usando as distribuições acumuladas normal e "t"de student. **Revista Brasileira de Matemática e Estatística**, v. 28, n. 3, p. 59–83, 2010.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. [S.l.]: Springer-Verlag New York, 2002. 740 p.

SORENSEN, D. A.; ANDERSEN, S.; GIANOLA, D.; KORSGAARD, I. Bayesian inference in threshold models using gibbs sampling. **Genetic Selection Evoution**, v. 27, n. 1, p. 229–249, 1995.

SPIEGELHALTER, D. J.; THOMAS, A.; LUNN, B. N. G. d. **WinBUGS User Manual**. Cambridge, UK, 2003. Version 1.4. Disponível em: <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>.

STEVENS, S. S. Measurement, statistics, and the schemapiric view. **Science**, v. 161, n. 3844, p. 849–856, 1968.

STRANDÉN, I.; GIANOLA, D. Attenuating effects of preferential treatment with student-t mixed linear models: a simulation study. **Genetics selection evolution**, v. 30, n. 1, p. 25–42, 1998.

TABOR, D. Mohs's hardness scale - a physical interpretation. **Proceedings of the Physical Society. Section B**, v. 67, n. 3, p. 249–257, 1954.

THERNEAU, T.; ATKINSON, E.; SINNWELL, J.; MATSUMOTO, M.; SCHAID, D.; MCDONNELL, S. **kinship2: Pedigree functions**. [S.l.], 2011. R package version 1.3.3. Disponível em: <http://CRAN.R-project.org/package=kinship2>.

WILKINSON, G. N.; ROGERS, C. E. Symbolic description of factorial models for analysis of variance. **Applied Statistics**, v. 22, n. 3, p. 392–399, 1973.

SEGUNDA PARTE

ARTIGO 1 Algoritmos para análise Bayesiana de modelos de limiar mistos com efeitos aleatórios

RESUMO

Neste estudo, foram avaliados os algoritmos descritos por Albert e Chib (1993) (AC), Cowles (1996) (MC) e Nandram e Chen (1996) (NC) para análise bayesiana de dados categorizados ordinais. Foram também avaliadas as modificações destes algoritmos propostas por Silva (2008) (NCG, NCt) e Silva e Bueno Filho (2010) (ACt). Os algoritmos foram avaliados em diferentes situações experimentais em dois delineamentos diferentes (um sistema triplo de Steiner com 7 tratamentos e um látice simples 10x10). Os experimentos foram simulados supondo quatro valores diferentes para a correlação intraclasse. Os algoritmos foram avaliados quanto a dependência da cadeia gerada pelo processo de amostragem, o total de iterações para convergência dos parâmetros do modelo, o erro quadrático médio (EQM) dos efeitos aleatórios, efeitos fixos e da correlação intraclasse estimada. Tanto em experimentos cujo o interesse está nas estimativas dos efeitos fixos quanto aleatórios, os algoritmos NCG e NCt apresentam estimativas com menor EQM, além da vantagem de apresentar menor dependência entre as amostras MCCM e por conseguinte, a rápida convergência dos parâmetros. Em delineamento com um grande número de efeitos aleatórios, os algoritmos descritos NC e NCt superestimaram a correlação intraclasse quando o valor simulado era alto (0.8). Isto no entanto não acarretou maior viés ou maior EQM do que o encontrado nos demais delineamentos. Um exemplo é apresentado com a análise de dados do melhoramento do tomateiro para a resistência à requeima (Phytophtora infestans). Aplicando o fator de Bayes a este experimento, não houve indícios para trocar a distribuição Gaussiana pela distribuição t-Student para o traço latente. Os algoritmos são consistentes quanto ao ranqueamento dos efeitos das variedades e podem ser utilizados para a seleção de genótipos resistentes com base no traço latente.

Palavras-chave: Algoritmos MCCM. Análise bayesiana. Modelos de limiar.

ABSTRACT

In this paper we evaluate algorithms for Bayesian analysis of ordinal categorical data described by Albert and Chib (1993) (AC), Cowles (1996) (MC) and Nandram e Chen (1996) (NC) as well as the modifications proposed by Silva (2008) (MCt, NCt) and Silva and Bueno Filho (2010) (ACt). Behaviour of the algorithms were studied under different experimental situations under simulation, such as two different designs (a Steiner Triple System with 7 treatments and a Simple Lattice Square 10x10) and four values for the intraclass correlation of random effects. Algorithms were evaluated about dependency in the resulting Markov Chain of the sampling process, the number of iterations until convergence diagnostics the mean squared error (MSE) and the average Bias for the parameters estimated in the posterior chain. NC and NCt resulted in faster and accurate algorithms, showing also smaller dependency in the sampling process. For larger designs and higher values of intraclass correlation (0.8) NC and NCt overestimated these correlation. Nevertheless this brings no problem of inflating MSE or average bias. An example from tomato breeding to select varieties with higher resistance to late Blight (Phytophtora infestans). Bayes factor applied to this example brought no evidence of Student-t distribution being better than Gaussian distribution to model liability. Algorithms are consistent as to the ranking of varieties and could be a useful tool to selection in plant breeding.

Keywords: Bayesian analysis. MCMC algorithms. Threshold models

1 INTRODUÇÃO

Dados categorizados ordinais são amplamente utilizados em diversas áreas de pesquisa, sendo em grande maioria, oriundos de medidas subjetivas. Diversos exemplos podem ser encontrados na literatura, como por exemplo em estudos de opinião (GOB; MCCOLLIN; RAMALHOTO, 2007), ciência dos alimentos com o uso da escala hedônica para avaliação de atributos sensoriais Piepho e Kalka (2003), na genética quantitativa animal Sorensen et al. (1995), na fitopatologia com as escalas diagramáticas para quantificação de doenças (CORRÊA; BUENO-FILHO; CARMO, 2009) e nas ciências físicas com a escala de dureza de Mohs (TABOR, 1954).

Alguns métodos são propostos para análises de dados categorizados, e entre eles, estão os modelos de limiar. Modelos de limiar utilizam uma variável latente com distribuição contínua de modo que a resposta observada em uma determinada categoria esteja entre os limites que definem tais categorias (ALBERT; CHIB, 1993; KIZILKAYA et al., 2003; MCCULLOGH; SEARLE, 2001; SORENSEN et al., 1995; PIEPHO; KALKA, 2003). Implmentações Bayesianas para os modelos de limiar foram propostas inicialmente por Albert e Chib (1993) para análise de modelos com efeitos fixos utilizando a amostragem de Gibbs para obter as distribuições marginais dos parâmetros com a distribuição Gaussiana para a variável latente. Entretanto, a forte autocorrelação entre as amostras para a obtenção das marginais dos parâmetros dificulta o uso deste algoritmo devido ao tamanho da amostra gerada durante o processo de Gibbs, fato observado por Sorensen et al. (1995) ao extenderem o algoritmo para uso em modelos de efeitos mistos. Uma extensão deste algoritmo também foi proposta por Silva e Bueno-Filho (2010) ao utilizar a distribuição t-Student como variável latente em modelos de efeitos mistos.

Para acelerar o processo de convergência, Cowles (1996) propôs um algoritmo que utiliza amostragem de Gibbs com um passo do algoritmo de Metropolis-Hastings, de forma a aceitar ou rejeitar todo o vetor de parâmetros thresholds. Nandran e Chen (1996) para acelerar o processo de amostragem, propuseram uma reparametrização para os parâmetros de limiar considerando a distribuição Dirichlet como geradora de candidatos para os paâmetros thresholds, ficando os mesmos limitados entre 0 e 1, o que não ocorre com os algoritmos propostos por Albert e Chib (1993) e Cowles (1996), pois seus parâmetros thresholds ficam limitados entre 0 e ∞. Silva (2008) iniciou a extensão do algoritmo proposto por Nandran e Chen (1996) para análise de modelos mistos, utilizando a distribuição Gaussiana e a t-Student para a variável latente, mas não há estudos que comparem a eficiência dos algoritmos propostos por Albert e Chib (1993)(ACG e ACt), Cowles (1996)

(MCG e MCt) e Nandran e Chen (1996) (NCG e NCt) para análise de modelos mistos. Portanto, o presente trabalho tem como objetivos apresentar as extensões do algoritmo proposto por Nandran e Chen (1996), iniciadas por Silva (2008) e concluí-las, assim como comparar a eficiência destes algoritmos com as extensões propostas por Silva e Bueno-Filho (2010) e Sorensen et al. (1995) que extendem o algoritmo de Albert e Chib (1993) para análise de modelos mistos e também com os algoritmos propostos por Cowles (1996), via processo de simulação e analisar um conjunto de dados referente a seleção 66 famílias de tomate resistentes a *Phytophthora infestans*.

2 EXTENSÕES PARA MODELOS MISTOS DO ALGORITMO DE Nandran e Chen (1996)

Iremos apresentar as extensões para modelos mistos do algoritmo proposto por Nandran e Chen (1996) utilizando a distribuição Gaussiana e t-Student para variável categorizadas com 3 ou mais categorias.

2.1 NCGaussian

O vetor de parâmetros thresholds, definido por γ, é o vetor que divide a reta real em K intervalos disjuntos, da seguinte forma, $(\gamma_0,\gamma_1); [\gamma_1,\gamma_2); \ldots ; [\gamma_{k-1},\gamma_k)$ com $\gamma_0 = -\infty$ e $\gamma_K = +\infty$. NC propuseram a distribuição de Dirichlet como geradora de candidatos para os parâmetros thresholds e adotaram $\gamma_1 = 0$, desta forma, cada γ_k^* será a distância entre γ_k e γ_1 dado pela transformação $\gamma_k^* = \gamma_k - \gamma_1$. Assim, a variável latente é obtida por meio da expressão $L_i^* = L_i - \gamma_i, i = 1,...,n$ e então, $Y_i = k$ se $\gamma_{k-1} - \gamma_1 \leq L_i^* \leq \gamma_k - \gamma_1$. A partir da reparametrização proposta por Nandran e Chen (1996), temos δ como uma variável auxiliar, γ_k^{**} o novo vetor dos parâmetros thresholds, θ^* como o novo vetor de efeitos fixos e aleatórios e L^* sendo a variável latente reparametrizada (2.1).

$$\begin{aligned} &\delta = 1/\gamma_{K-1}^*, \\ &\gamma_k^{**} = \delta_{\gamma_k^*},\ k = 0,1,2,...,K, \\ &\theta^* = \delta\theta \text{ e } L^* = \delta L^* \end{aligned} \tag{2.1}$$

Com esta reparametrização, para problemas com três categorias, não existirá nenhum parâmetro de limiar a ser estimado. O jacobiano da transformação de δ é dados por $[\delta^2]^{-\frac{1}{2}(n+m+K)}$, que representa o erro do traço latente, em que n é o número de observações (comprimento do vetor y), m o número de variáveis explicativas e K o número de classes observadas em y. Assumindo uma priori gamma inversa para δ^2 e considerando uma distribuição acumulada Gaussiana (ϕ) como função de ligação, temos a distribuição conjunta a posteriori, dada por:

$$\begin{aligned} p\left(\theta^*,\gamma_k^{**},\delta^2,L^*|y\right) \propto &\left[\prod_{i=1}^{n} \Phi\left(L_i^*, w_i\theta^*,\delta^2\right) I_{[\gamma_k^{**},\gamma_{k+1}^{**})}(L_i^{**})\right] \\ &\times \Phi(\theta^*;\mathbf{0},\delta^2 V)(\delta^2)^{-k/2} p(\sigma_u^2) p(\delta^2) \end{aligned} \tag{2.2}$$

O (σ_u^2) representa os componentes da variância dos efeitos aleatórios. Para o caso de um modelo misto, V é uma matriz na dimensão dos efeitos fixos mais efeitos aleatórios, sendo os efeitos fixos igual a zero e para os efeitos aleatórios, ela é a matriz **A**, que pode ser uma matriz I_{mxm}, ou no caso de genealogia conhecida, na genética, pode ser a matriz de parentesco (HENDERSON, 1976) ou ainda alguma medida de similaridade genética (obtida com marcadores moleculares, por exemplo).

A distribuição condicional completa para θ^* é dada por:

$$\theta^*|L^*, \delta^2, Y, \sigma_u^2 \sim N\left(B^{-1}W'L^*, \delta^2 B^{-1}\right) \tag{2.3}$$

com $W = [X|Z]$, sendo X e Z as matrizes de efeitos fixos e aleatórios respectivamente. Em (2.3), B é o vetor de efeitos fixos, que é dado por $B = \delta^2 V^{-1} + W'W$. Para $p(\delta^2)$ em (2.2), nós temos uma distribuição gamma inversa como priori.

$$p(\delta^2) \propto (\delta^2)^{-(c+1)} exp\left\{-\frac{d}{\delta^2}\right\} \tag{2.4}$$

em que c e d são hiperparâmetros desta priori, então, a distribuição condicional completa a posteriori para δ^2 será:

$$\delta^2|\theta^*, Y, \sigma_u^2 \sim GI(a_\delta, b_\delta) \tag{2.5}$$

sendo, a_δ o parâmetro de forma e b_δ o parâmetro de escala, conforme (2.6) e (2.7)

$$a_\delta = \frac{n + m + K + 2c}{2} \tag{2.6}$$

e

$$b_\delta = \frac{(L^* - W\theta^*)'(L^* - W\theta^*) + \theta^{*'} V^{-1}\theta^* + 2d}{2} \tag{2.7}$$

Para a variância dos efeitos aleatórios (σ_u^2), a priori adotada é uma gamma inversa, conforme (2.8).

$$P(\sigma_u^2) = (\sigma_u^2)^{-(a+1)} exp\left(\frac{-b}{\sigma_u^2}\right) \tag{2.8}$$

A distribuição condicional completa para (σ_u^2) é dada por:

$$p(\sigma_u^2|\theta^*, L^*, Y) = (\sigma_u^2)^{-(\frac{q}{2}+a+1)} exp\left\{\frac{-1}{2\sigma_u^2}(u'u + 2b)\right\} \tag{2.9}$$

Em (2.9), temos o núcleo de uma distribuição gamma inversa (2.10)

$$\sigma_u^2|\theta^*,L^*,y \sim GI\left(\frac{q+2a}{2},\frac{u^{'}u+2b}{2}\right) \tag{2.10}$$

Com a reparametrização em (2.1) a distribuição condicional para $(\gamma^{**}|\theta^{**},\delta^2,Y)$ fica sendo:

$$\begin{aligned}\pi(\gamma^{**}|\theta^{**},\delta^2,Y) \propto & \prod_{Y_i=2}\left[\Phi\left(\frac{\gamma_2^{**}-w_i^{'}\theta^{**}}{\delta}\right)-\Phi\left(\frac{-w_i^{'}\theta^{**}}{\delta}\right)\right] \\ & \times\prod_{Y_i=3}\left[\Phi\left(\frac{\gamma_3^{**}-w_i^{'}\theta^{**}}{\delta}\right)-\Phi\left(\frac{\gamma_2^{**}-w_i^{'}\theta^{**}}{\delta}\right)\right]\cdots \\ & \times\prod_{Y_i=K-1}\left[\Phi\left(\frac{1-w_i^{'}\theta^{**}}{\delta}\right)-\Phi\left(\frac{\gamma_{K-2}^{**}-w_i^{'}\theta^{**}}{\delta}\right)\right]\end{aligned} \tag{2.11}$$

e para $(L^*|\gamma^{**},\theta^*,\delta^2,Y)$, sua condicional a posteriori completa é dada por:

$$L^*|\gamma^{**},\theta^*,\delta^2,Y=k \sim N(W\theta^*,\delta^2). \tag{2.12}$$

Para a amostragem de γ^{**} é utilizado um passo do algoritmo Metropolis-Hastings, pois este parâmetro não tem uma distribuição fechada para que se possa utilizar o método de Gibbs. Conforme a reparametrização em (2.1), temos que $\gamma_1^{**}=0$ $\gamma_{k-1}^{**}=1$, e desta forma, construímos um vetor auxiliar **p**, definido em (2.13)

$$\mathbf{p}_{k-1}=\gamma_k^{**}-\gamma_{k-1}^{**},\ k=2,...,K-1; \tag{2.13}$$

em que

$$p=(p_1,p_2,\ldots,p_{k-2})^{'},\ p_k\geq 0,\ k=1,2,\ldots,K-2 \text{ e } \sum_{k=1}^{K-2}p_k=1 \tag{2.14}$$

De acordo com o Nandran e Chen (1996), pelo teorema do valor médio, nós tem-se:

$$\Phi\left(\frac{\gamma_k^{**}-w_i^{'}\theta^{**}}{\delta}\right)-\Phi\left(\frac{\gamma_{k-1}^{**}-w_i^{'}\theta^{**}}{\delta}\right)=\frac{1}{\delta}\Phi\left(\frac{\xi_{k-1}-w_i^{'}\theta^{**}}{\delta}\right)p_{k-1} \tag{2.15}$$

em que $\xi_{k-1} \in (\gamma_k^{**}; \gamma_{k-1}^{**}), k = 2,3,...,K-1,$ e $\Phi(.)$ é a função densidade acumulada normal padrão. Utilizando o resultado de (2.15) tem-se:

$$\pi(\gamma^{**}|\theta^*,\delta^2,Y) \propto h1(\xi)h2(p) \tag{2.16}$$

e na expressão anterior, $h1$ e $h2$ são respectivamente:

$$h_1(\xi) = \prod_{k=1}^{K-2} \prod_{i=1}^{n_k} \Phi\left(\frac{\xi_k - w'\theta^*}{\delta}\right)$$

e

$$h_2(\mathbf{p}) = \prod_{k-1}^{K-2} P_k^{n_k+1}$$

Pode-se observar que a distribuição condicional de γ^{**} é dada pelo produto de $h_1 \times h_2$, em que h_2 é o núcleo de uma distribuição Dirichlet com parâmetros $n = (n_2 + 1, ...,n_{k-1} + 1)'$ e h_2 não depende de θ^{**} e nem de δ. Com os novos valores de $\mathbf{p}$ gerados a partir de h_2 a construção dos novos valores de γ_k^{**} é como se segue:

$$\gamma_{kj}^{**} = \sum_{i=1}^{k-1} p_{i,j}, k = 2,...,K-2. \tag{2.17}$$

A probabilidade de aceitação do novo vetor

$$\gamma_j^{**} = \left(\gamma_{2,j}^{**},...,\gamma_{K-2,j}^{**}\right) \tag{2.18}$$

é o mín(1, α), em que,

$$\alpha = w(\gamma^{**(j)}, \mathbf{p}^j)/w(\gamma^{**(j-1)}, \mathbf{p}^{(j-1)}), \tag{2.19}$$

e j representa a j-ésima iteração do algoritmo. Em (2.19), a forma geral da distribuição de $w(\gamma^{**}, \mathbf{p})$ é dada por $w(\gamma^{**},\mathbf{p}) = P(\gamma^{**}|\theta^*,\delta^2)/P(\mathbf{p}|n,\theta^*,\delta^2)$. A expressão de $P(\gamma^{**}|\theta^*,\delta^2)$ é dada em (2.11) e $P(\mathbf{p}|n, \theta^*,\delta^2)$ é a distribuição Dirichlet,

$$p(\mathbf{p}) = \frac{1}{Z(n)} \prod_{k=1}^{K-2} p_k^{n_{(k+1)}-1} \tag{2.20}$$

em que $p_1,...,p_{K-2} \geq 0; \sum_{k=1}^{K-2} p_k = 1$ e $n_2,...,n_{K-1} > 0$. Na expressão acima,

$Z(n)$ é a constante normalizadora dada por

$$Z(n) = \frac{\prod_{k=1}^{K-2} \Gamma_{(n_{k+1})}}{\Gamma\left(\sum_{k=1}^{K-2} n_{k+1}\right)} \tag{2.21}$$

Como o interesse é na distribuição conjunta de L^* e γ^{**}, a amostragem é realizada para **p** e, a partir dos elementos deste vetor constrói-se γ_k^{**}, e caso o novo vetor de parâmetros thresholds seja aceito atualiza-se os valores de L^*. Caso contrário continua com a amostra anterior da variável latente.

2.2 NCt

Utilizando a distribuição t-Student acumulada como função de ligação e a parametrização apresentada em (2.1) , a verossimilhança é dada por:

$$P(Y_i = k|\theta^*,\delta,v,\gamma^{**}) = F_v\left(\frac{\gamma_k^{**} - w_i^{'}\theta^*}{\delta}\right) - F_v\left(\frac{\gamma_{k-1}^{**} - w_i^{'}\theta^*}{\delta}\right), k = 1,2,...,K \tag{2.22}$$

em que F_v é a função de distirbuição acumulada t-Student com v graus de liberdade. Por facilidades algébricas para a obtenção das condicionais completas a posteriori, a distribuição t-Student é escrita em dois estágios (2.23) e (2.24), conforme Sorensen e Gianola (2002), sendo um misto de distribuição Gaussiana com uma distribuição gama inversa para os parâmetros de variância.

$$\lambda_i|v \sim Gamma(\frac{v}{2},\frac{v}{2}) \tag{2.23}$$

$$L_i^*|\theta^*,\delta^2,\lambda_i \sim N\left(w_i^{'}\theta^*, \frac{\delta^2}{\lambda_i}\right) \tag{2.24}$$

Desta forma, o modelo em (2.22) pode ser reescrito como:

$$P(Y_i = k|\theta^*,\delta_i,v,\gamma^{**}) = F_v\left(\frac{\gamma_k^{**} - w_i^{'}}{\frac{\delta}{\sqrt{\lambda_i}}}\theta^*\right) - F_v\left(\frac{\gamma_{k-1}^{**} - w_i^{'}}{\frac{\delta}{\sqrt{\lambda_i}}}\theta^*\right) k = 1,2,...,K \tag{2.25}$$

A priori para a distribuição de v, é conforme Kizilkaya et al. (2003), sendo $p(v) = 1/(1+v)^2$. Assumindo priori vaga para θ^* a posteriori conjunta para todos os parâmetros é:

$$\begin{aligned} p(\theta^*,\gamma_k^{**},L^{**},\lambda|y) \propto (\delta^2)^{-\frac{k}{2}} & \left[\prod_{i=1}^{n} \phi\left(L_i^*,w_i\theta^*,\frac{\delta^2}{\lambda_i}\right) I_{[\gamma_k^{**},\gamma_{k+1}^{**})}(L_k^*)\right] \\ & \times \left[\prod_{i=1}^{n} \lambda_i^{(\frac{v}{2})-1} exp\left(-\frac{\lambda_i}{2}v\right)\right] \Phi(\theta^*,\mathbf{0},\delta^2 V)\frac{1}{(1+v)^2} \\ & \times p(\sigma_u^2)p(\delta^2)p(v) \end{aligned} \tag{2.26}$$

em que $\lambda = \{\lambda_i\}_{i=1}^n$, V foi definido em (2.2). A priori para δ^2 é como em (2.4) e para σ_u^2 é uma gamma inversa, conforme (2.8)

A seguir são apresentadas as condicionais completas a *posteriori* para todos os parâmetros observados. Sendo a distribuição condicional completa para θ^* indicada abaixo.

$$\theta^*|L^*,\delta^2,v,\lambda \sim N(M^{-1}W^{'}R^{-1}L^*,\delta^2 M^{-1}) \tag{2.27}$$

sendo $M = W^{'}R^{-1}W + \delta^2 V^{-1}$

O parâmetro λ_i tem distribuição condicional conforme expressão a seguir Kizilkaya et al. (2003).

$$p(\lambda_i|\lambda_{-i},L^*,\theta^*,v) \propto \lambda_i^{\left(\frac{v+1}{2}\right)-1} exp\left(-\frac{\lambda_i}{2}\left(\left(L_i^* - w_i^{'}\theta^*\right)^2 + v\right)\right) \tag{2.28}$$

em que λ_{-1}, denota todos os elementos de λ exceto λ_i. A distribuição acima é proporcional a uma distribuição gamma com parâmetros $(v+1)/2$ e $(L_i^* - w_i^{'}\theta^*)^2 + v$, isto é,

$$\lambda_i|\lambda_{-i},L_i^*,\theta^*,v,\delta^2 \sim Gamma\left(\frac{v+1}{2},\frac{1}{2}\left(\left(L_i^{**} - w_i^{'}\theta^*\right)^2 + v\right)\right) \tag{2.29}$$

e

$$L_i^*|\lambda_i,\theta^*,v,\delta^2 \sim N\left(w_i^{'}\theta^*,\frac{\delta^2}{\lambda_i}\right) \tag{2.30}$$

A distribuição condicional para v não tem forma fechada, portanto, sua amostragem é realizada conforme Kizilkaya et al. (2003), utilizando o algoritmo Metropolis-Hastings para sua amostragem da sua distribuição (2.31).

$$p(v|\theta^*,L^*,\lambda,\delta^2) \propto \left(\frac{(\frac{v}{2})^{(v/2)}}{\Gamma(v/2)}\right)^n \left(\prod_{i=1}^{n} \lambda_i^{\frac{v}{2}-1} exp\left(\frac{v}{2}\lambda_i\right)\right) \frac{1}{(1+v)^2} \tag{2.31}$$

A distribuição condicional para δ^2 tem a mesma forma como em (2.5). Como para o caso do modelo normal, a distribuição conjunta de L^{**} e γ^{**} pode ser escrita por meio do produto das condicionais $(\gamma^{**}|\theta^*,\delta^2,\lambda,Y)$ e $(L^*|\gamma^{**},\theta^*,\delta^2,\lambda,Y)$. A distribuição completa para a variável latente é como em (2.30). Para γ^{**} a distribuição condicional completa é dada por:

$$\begin{aligned} p(\gamma^{**}|\delta^2,\lambda,\theta^*,Y) \propto & \prod_{Y_i=2}\left[\Phi\left(\frac{\gamma_2^{**}-w_{,}'\theta^{**}}{\frac{\delta}{\sqrt{\gamma_i}}}\right)-\Phi\left(\frac{-w_i'\theta^{**}}{\frac{\delta}{\sqrt{\lambda_i}}}\right)\right] \\ & \times\prod_{Y_i=3}\left[\Phi\left(\frac{\gamma_3^{**}-w_{,}'\theta^{**}}{\frac{\delta}{\sqrt{\lambda_i}}}\right)-\Phi\left(\frac{\gamma_2^{**}-w_i'\theta^{**}}{\frac{\delta}{\sqrt{\lambda_i}}}\right)\right] \\ & \cdots\prod_{Y_i=K-1}\left[\Phi\left(\frac{1-w_{,}'\theta^{**}}{\frac{\delta}{\sqrt{\lambda_i}}}\right)-\Phi\left(\frac{\gamma_{K-2}^{**}-w_i'\theta^{**}}{\frac{\delta}{\sqrt{\lambda_i}}}\right)\right] \end{aligned} \tag{2.32}$$

Da mesma forma que em (2.15), de acordo com o teorema do valor médio, temos:

$$\Phi\left(\frac{\gamma_k^{**}-w_i'\theta^*}{\frac{\delta}{\sqrt{\lambda_i}}}\right)-\Phi\left(\frac{\gamma_{k-1}^{**}-w_i'\theta^*}{\frac{\delta}{\sqrt{\lambda_i}}}\right)=\frac{1}{\frac{\delta}{\sqrt{\lambda_i}}}\Phi\left(\frac{\xi_{k-1}-x_i'\theta^*}{\frac{\delta}{\sqrt{\lambda_i}}}\right)p_{k-1} \tag{2.33}$$

em que $\xi_{k-1}\in(\gamma_k^{**};\gamma_{k-1}^{**}), k=2,...,K-1$, e $\Phi(.)$ é a função densidade normal padrão. E apartir de (2.33) temos:

$$\pi(\gamma^{**}|\theta^*,\delta^2,Y)\propto h_3(\xi)h_4(p) \tag{2.34}$$

sendo h_3 e h_4 de (2.34), respectivamente:

$$h_3(\xi)=\prod_{k=1}^{K-2}\prod_{i=1}^{n_k}\left(\frac{\xi_k-w_i'\theta^{**}}{\frac{\delta}{\sqrt{\lambda_i}}}\right)$$

$$h_4(p)=\prod_{k=1}^{K-2}p_k^{n_k+1} \tag{2.35}$$

O procedimento para gerar os candidatos para a variável latente com distribuição t-Student, a construção dos *thresholds* e a probabilidade de aceitação em (2.19) são idênticos ao modelo normal, com exceção do parâmetro de escala, que para a distribuição t-Student é $\frac{\delta}{\sqrt{\lambda_i}}$.

3 ESTUDO DE SIMULAÇÃO

O estudo de simulação foi conduzido para avaliar o erro quadrático médio das estimativas dos efeitos fixos, aleatórios e correlação intraclasse. Também foram avaliados a correlação entre os valores preditos e observados, o tempo de processamento, o grau de autocorrelação da cadeia simulada e o descarte inicial. Todos os algoritmos foram implementados para uso em software R, com o uso do pacote Bayesthresh. Foram considerados dois delineamentos experimentais em blocos incompletos: um delineamento pequeno representado pelo Sistema Triplo de Steiner (STS) com sete tratamentos e um delineamento grande parcialmente balanceado representado por um Látice Quadrado Simples (LQS) com 100 tratamentos. A descrição dos delineamentos será apresentada nas próximas subseções.

3.1 Sistema Triplo de Steiner (STS)

O sistema triplo de Steiner foi constituído de sete tratamentos, duas repetições e sete blocos de tamanho três. O modelo linear utilizado para simular os efeitos e as realizações da variável observada foi:

$$u = \mu + Xb + Za + \epsilon \tag{3.36}$$

sendo u o vetor da realizações da variável observada, μ é a média geral, sendo 5; b é o vetor dos efeitos fixos, com $b \sim N(0,1)$ e X é a matriz de delineamento dos efeitos fixos; a é o vetor dos efeitos aleatórios, simulado de uma distribuição normal, então $a \sim N(0,\sigma_a^2)$ e Z é a matriz de delineamento dos efeitos aleatórios. No caso dos efeitos aleatórios, foram adotados quatro valores para σ_a^2, sendo (0.1, 0.2, 0.5 e 0.8). Para o erro amostral (ϵ) foi considerado que $\epsilon \sim N(0,\sigma_e^2)$, onde, $\sigma_e^2 = (0.9, 0.8, 0.5 \text{ e } 0.2)$. A partir das combinações de σ_a^2 e σ_e^2, nós obtemos quatro valores paramétricos para a correlação intraclasse $\rho = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$, quais sejam 0.1, 0.2, 0.5 e 0.8.

O vetor u gerado a partir de (3.36) é um vetor de uma variável aleatória contínua e sua transição para variável categórica foi realizada a partir de quatro valores para os quantis de u, obtendo-se assim cinco categorias, indo de 1 a 5. Foram adotados três distribuições para a variável categorizada, simétrica, assimétrica e uniforme. Para a distribuição simétrica foram considerados os quantis 0.0001, 0.15, 0.50 e 0.85, no caso assimétrico, os quantis 0.4, 0.5, 0.75 e 0.90

e para o caso da distribuição uniforme, temos 0.017, 0.28, 0.525 e 0.775, então $\eta_i \in k,\ k = 1,2,\ldots,5$ se:

$$\eta_i = \begin{cases} 1,\ u_i \leq Q_1 \\ 2,\ Q_1 < u_i \leq Q_2 \\ 3,\ Q_2 < u_i \leq Q_3 \\ 4,\ Q_3 < u_i \leq Q_4 \\ 5,\ u_i > Q_4 \end{cases} \tag{3.37}$$

A figura (1) ilustra a distribuição da variável resposta categorizada.

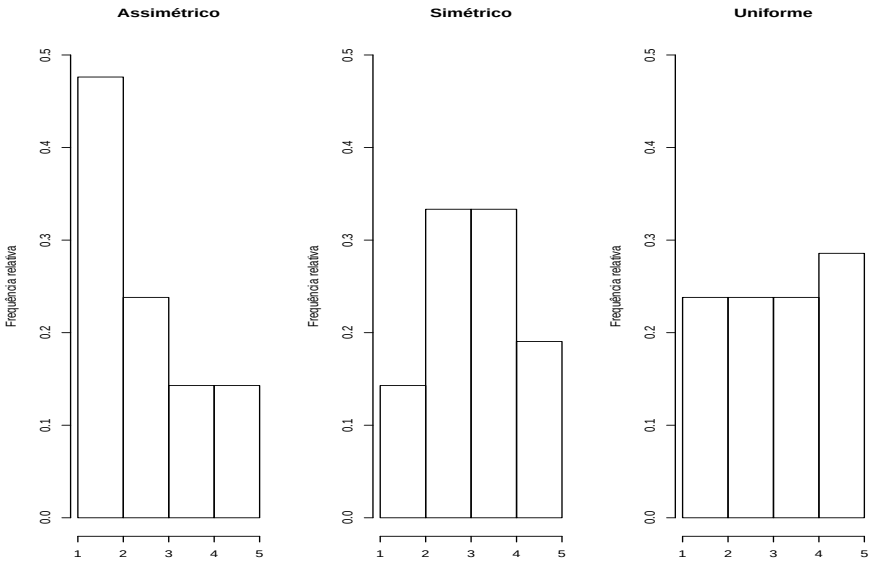

Figura 1 Distribuição da variável resposta nas diferentes configurações de experimentos, no Sistema Triplo de Steiner

Por fim, temos 4 diferentes valores para ρ combinados com 3 distribuições para a variável resposta, totalizando 12 situações experimentais diferentes. As situações experimentais foram analisadas considerando duas diferentes prioris para a variância dos efeitos aleatórios, sendo uma gamma inversa, $GI \sim (3,\ 5)$(priori menos informava) e $GI \sim (8,\ 5)$(priori mais informativa). Para o caso dos modelos NCG e NCt, cuja a priori para σ_e^2 é uma gama inversa, foi adotado $GI \sim (20,\ 5)$.

Os experimentos simulados foram analisados pelo modelo apresentado em (3.38).

$$\eta_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk} \tag{3.38}$$

onde η_{ijk} é o vetor associado as observações do tratamento i no bloco j; μ é uma constante; τ_i é o efeito do i^{th} tratamento, assumido como fixo; β_j é o efeito do j^{th} bloco, assumido como aleatório; ϵ_{ijk} é o erro experimental. Para cada situação experimental, foram simulados 1000 experimentos.

3.2 Látice quadrado (LQ)

Um delineamento látice quadrado simples 10x10, constituído por com 2 repetições, 10 blocos e 100 tratamentos foi utilizado nesta simulação. O modelo descrito para a geração das observações foi descrito em (3.36). O vetor de resposta simulado foi categorizado em 9 classes utilizando os quantis 0.005, 0.075, 0.185, 0.325, 0.50, 0.675, 0.825 e 0.925, sendo considerada uma distribuição assimétrica, com as categorias indo de 1 a 9. Para os efeitos fixos e aleatórios, foi utilizado o mesmo procedimento descrito na seção (3.1), totalizando 4 situações experimentais. O modelo de análise foi descrito em (3.38).

Assim como o STS, o LQ foi analisado considerando duas prioris distintas para os componentes da variância, sendo uma $GI \sim (3,\ 5)$ (Priori menos informativa) e $GI \sim (10,\ 2)$ (priori mais informativa). A priori para a variância residual foi a mesma utilizada na análise do experimento STS, sendo $GI\ (20,\ 5)$, para os algoritmos NCG e NCt. Os demais algoritmos apresentam variância residual igual a 1. Para o LQ foram simulados 500 experimentos para cada situação experimental.

3.1 Procedimento de simulação

O procedimento de simulação foi iniciado com a obtenção de uma cadeia inicial de 4000 amostras, sendo avaliada pelo teste de Raftery e Lewis (RAFTERY; LEWIS, 1992) para avaliar o descarte inicial ("burn in") e a dependência da cadeia ("jump"). A partir do diagnóstico de Raftery e Lewis foram geradas duas cadeias independentes e aplicado o teste de Gelman e Rubin (GELMAN; RUBIN, 1992) para avaliar a convergência dos parâmetros. O procedimento de análise foi realizado com o auxílio do pacote Bayesthresh (CORREA; BUENO-FILHO, 2012) e o diagnóstico da cadeia foi realizado com o auxílio do pacote coda (PLUMMER et al., 2006), ambos implementados para uso no software R Development Core Team (2012).

4 RESULTADOS E DISCUSSÕES

Os algoritmos NCG e NCt que utilizam a reparametrização proposta por Nandran e Chen (1996) que possuem como geradora de candidatos para os parâmetros thresholds a distribuição de Dirichlet, apresentaram menor EQM para as estimativas dos efeitos fixos, quando comparada aos demais algoritmos, independente do delineamento utilizado (Figura 2).

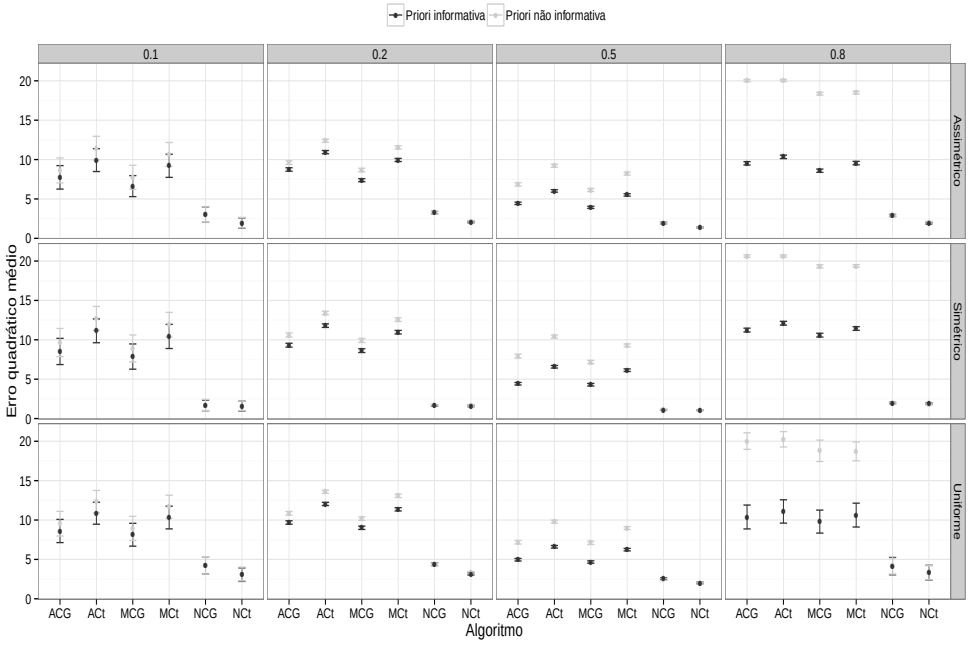

Figura 2 Erro quadrático médio para a média a posteriori dos efeitos fixos no delineamento STS

Não foi observado efeito da priori para os componentes da variância no EQM dos efeitos fixos para os algoritmos NCG e NCt. Para os demais algoritmos, quanto maior a correlação intraclasse, maior foi a diferença entre o EQM da priori informativa e da priori não informativa, sendo a priori não informativa a que apresentou maior EQM (Figura 2). A função de ligação não afetou a posteriori dos efeitos fixos, sendo que a distribuição Gaussiana ou a t-Student apresentaram comportamento semelhante quanto ao EQM, seja considerando uma distribuição simétrica, assimétrica ou uniforme.

Comportamento semelhante ao encontrado para o EQM dos efeitos fixos, foi observado para os algoritmos NCG e NCt para o EQM dos efeitos aleatórios,

ou seja, independente da priori utilizada e da distribuição da variável resposta, o EQM foi semelhante. Já os algoritmos ACG, ACt, MCG e MCt, quando utilizamos a priori informativa apresentaram EQM semelhante ao obtido com os algoritmos NCG e NCt (Figura 3).

Apesar de Strandén e Gianola (1998) sugerirem que a distribuição t-Student pode ser mais flexível que a distribuição Gaussiana e sugerem que em casos de assimetria na distribuição da variável resposta, a t-Student pode ser mais robusta, este efeito não foi observado, pois a distribuição t-Student apresentou comportamento semelhante a distribuição Gaussiana. Kizilkaya et al. (2003) alertou sobre a necessidade de maiores estudos com o uso da distribuição t-Student em modelos thresholds, afim de verificar suas propriedades diante dos resultados apresentados por Strandén e Gianola (1998). Apesar do menor EQM para os algoritmos NCG e

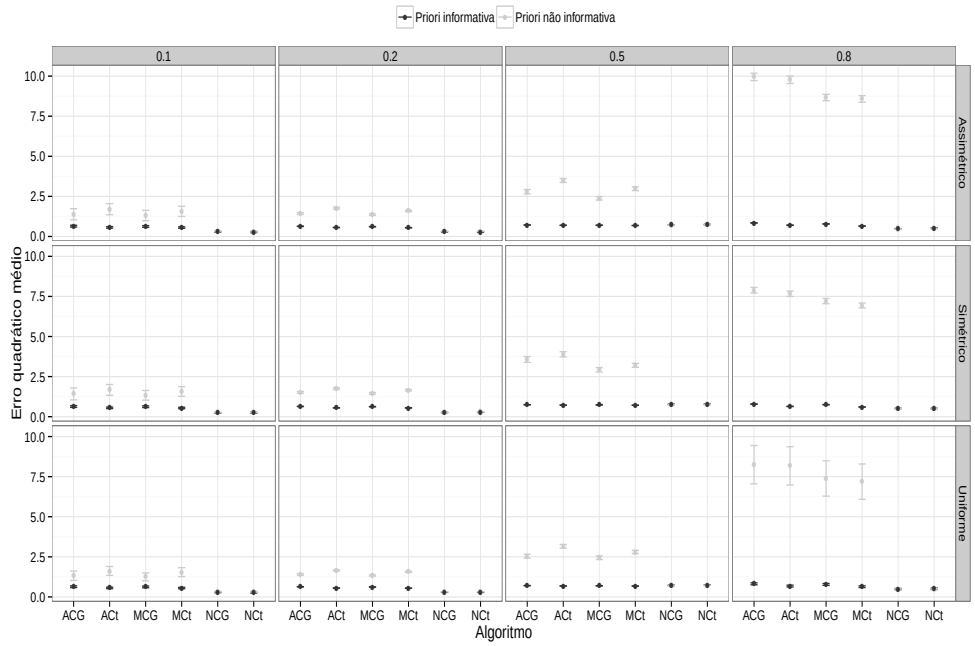

Figura 3 Erro quadrático médio para a média a posteriori dos efeitos aleatórios no delineamento STS

NCt para a posteriori dos efeitos fixos e aleatórios, os mesmos apresentaram maior sensibilidade a priori usada para os componentes de variância, sendo a priori informativa $(GI \sim (3,\ 5))$ mais acurada apenas nas combinações em que a correlação intraclasse era de 0.1 e 0.2, com tendência a superestimar os valores de ρ para os

maiores valores da correlação intraclasse (Figura 4). Para os casos em que a correlação intraclasse era igual ou superior a 0.5, a priori menos informativa apresentou menor EQM para ρ. Os demais algoritmos, quando utilizamos a priori informativa apresentaram menor EQM para ρ, mas, quando avaliadas com $\rho = 0.8$, o EQM destes algoritmos (ACG, ACt, MCG e MCt) também se elevou, porém, em menor valor que os algoritmos NCG e NCt. A correlação entre os valores preditos e ob-

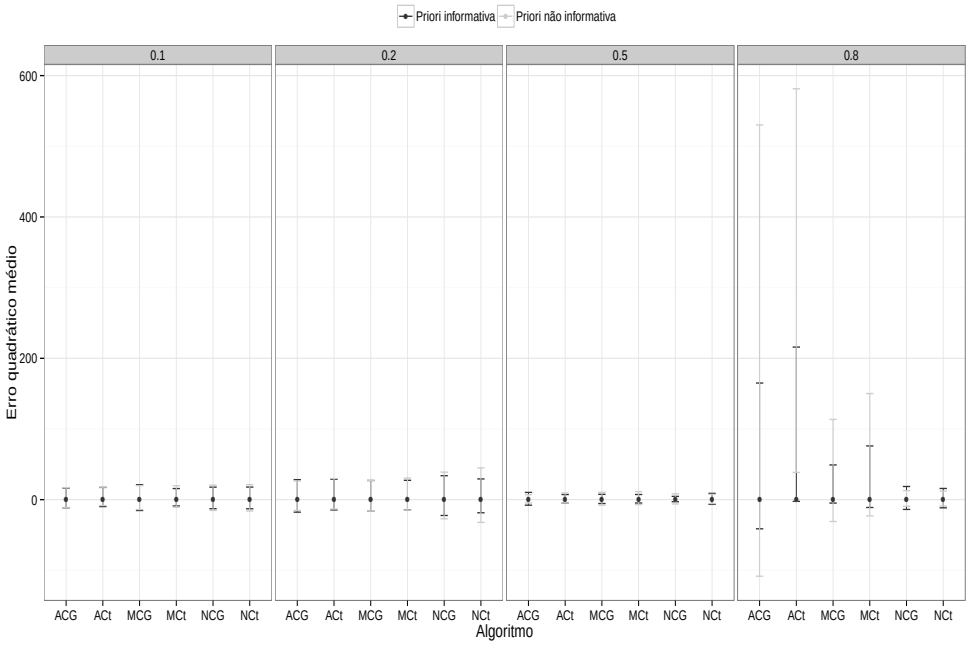

Figura 4 Erro quadrático médio para a correlação intraclasse no delineamento LQ

servados foi aproximadamente de 0.99 para a correlação intraclasse de 0.8 e para valores de $\rho < 0.8$ a correlação variou entre 0.75 e 0.90.

Não evidenciamos diferenças entre o uso da distribuição t-Student e da distribuição Gaussiana como função de ligação. As maiores diferenças foram observadas entre os algoritmos, indicando que a reparametrização proposta por Nandran e Chen (1996) com o uso da distribuição Dirichlet como geradora de candidatos de γ^* se apresentou mais flexível diante das diferentes situações experimentais simuladas, com menor EQM para as estimativas dos efeitos fixos e aleatórios, apesar de apresentar maior sensibilidade para a priori dos componentes da variância, quando a correlação intraclasse a ser estimada é considerada alta. O menor EQM para

os algoritmos NCG e NCt podem ser evidenciados na Figura (5), onde pode ser observado que o algoritmo NCG amostra uma cadeia com menor dispersão.

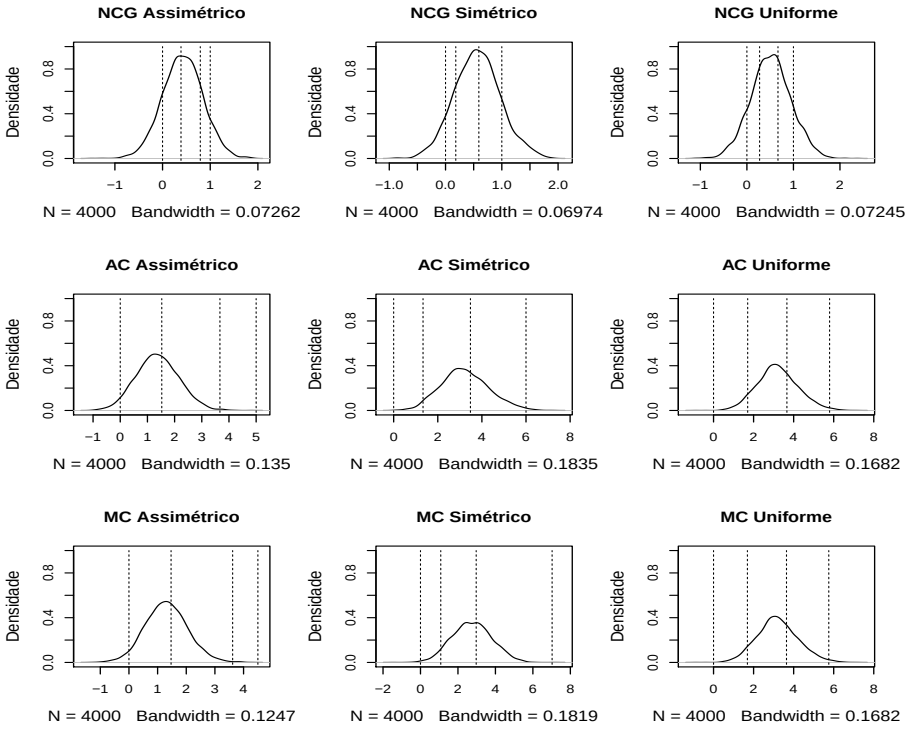

Figura 5 Densidade a posteriori do vetor λ para um exemplo simulado no delineamento STS com $\rho = 0.2$, com burn=0, jump=1 e tamanho final da cadeia igual a 4000 com distribuição simétrica, assimétrica e uniforme para a variável resposta. As linhas pontilhadas representam as médias a posteriori das posições dos thresholds

A menor dispersão da apresentada na cadeia gerada com os algoritmos utilizando a reparametrização de Nandran e Chen (1996), não apenas reduz o EQM dos efeitos fixos e aleatórios, como também, acelera o processo de convergência da cadeia. Nandran e Chen (1996) ao compararem seu algoritmo utilizando a distribuição acumulada Gaussiana como função de ligação, com os algoritmos de Albert e Chib (1993) e Cowles (1996) observaram uma melhoria no processo de convergência. Para o caso dos modelos mistos, também foi evidenciado uma redução no número de iterações para que a convergência ocorra, apesar de não haver diferença entre o uso da distribuição Gaussiana ou t-Student no processo

Tabela 1 Média do tempo de processamento, burn-in, jump e total de iterações para os seis algoritmos avaliados nos delineamentos STS e LQS, pelo teste de Raftery e Lewis.

Algoritmo	Tempo(min.)	Burning	Jump	Total de iterações
		STS		
ACG	3.57	78.36	17.59	67032.08
ACt	4.12	89.23	22.21	83281.63
MCG	1.97	61.87	15.38	57742.86
MCt	4.33	75.05	18.72	70173.76
NCG	0.91	23.79	6.65	25255.95
NCt	1.06	23.48	6.84	25959.31
		LQS		
ACG	23.74	60.74	15.17	56869.24
ACt	41.40	94.38	20.34	76290.67
MCG	11.39	27.00	8.09	30317.97
MCt	39.20	47.25	13.76	51562.65
NCG	9.74	14.74	4.57	17150.79
NCt	16.20	16.43	5.15	19226.35

de convergência, pelo teste de Raftery e Lewis. A tabela 1 apresenta o tempo de processamento, o descarte inicial, o salto entre as iterações e o total de iterações.

Para os componentes da variância, a priori informativa acelera o processo de convergência dos algoritmos (Tabela 2), mas não afeta o desempenho dos algoritmos de forma a melhorar seu desempenho global, ou seja, os algoritmos NCG e NCt apresentaram maior velocidade de convergência que os demais algoritmos independente da priori utilizada. O fato de um delineamento apresentar um maior número de efeitos aleatórios a serem estimados, como o LQS, em relação a um delineamento com poucos efeitos aleatórios, STS, não aumentou a dependência da cadeia gerada durante o processo de amostragem, ou seja, independente do número de efeitos aleatórios a serem estimados, os algoritmos NCG e NCt demonstraram performance superior aos demais algoritmos avaliados.

O algoritmo NCG, além de reduzir a dependência das amostras, promoveu a maior convergência dos parâmetros estimados, em ambos delineamentos utilizados. Kizilkaya et al. (2003) ao sugerir que a amostragem do vetor de parâmetros thresholds fosse realizada a partir da aceitação ou não do vetor, reduziu a independência entre as amostras geradas, mas não promoveu uma alta convergência dos parâmetros nas situações simuladas (tabela 3). Uma das vantagens nos algoritmos NCG e NCt é a reparametrização proposta por Nandran e Chen (1996) para

Tabela 2 Média do tempo de processamento, burn-in, jump e total de iterações para as duas prioris utilizadas na estimativa dos componentes das variâncias para os delineamentos STS e LQS, pelo teste de Raftery e Lewis.

Priori	Tempo (min.)	Burn-in	Jump	Total de iterações
	STS			
Informativa	2.35	52.58	13.48	50485.61
Não informativa	2.97	64.68	15.65	59309.06
	LQS			
Informativa	19.08	36.49	10.57	39591.30
Não informativa	28.08	50.36	11.79	44214.49

Tabela 3 Taxa de convergência das amostras para os delineamentos em STS e LQS, nos seis algoritmos avaliados, pelo teste de Gelman e Rubin.

Algoritmo	Delineamento	
	STS	Látice
ACG	0.77	0.81
ACt	0.74	0.75
MCG	0.61	0.65
MCt	0.50	0.42
NCG	0.89	0.96
NCt	0.72	0.96

a distribuição geradora de candidatos para os parâmetros thresholds, que tem por característica definir os thresholds no intervalo $[0,1]$. A distribuição de Dirichlet, proposta por Nandran e Chen (1996) é uma distribuição conjugada da multinomial, o que caracterizou ser uma distribuição mais adequada que a distribuição normal para a amostragem dos thresholds.

A acelaração no processo de convergência promovido pelos algoritmos NCG e NCt pode ser melhor entendido observando a figura (6), que ilustra a trajetória da cadeia do segundo threshold para o delineamento STS, apresentado na figura 5. Podemos observar que o algoritmo adaptado de Nandran e Chen (1996) necessita de poucas interações para obter uma cadeia estacionária e independente.

Os resultados obtidos no processo de simulação corroboram com os encontrados por Nandran e Chen (1996) quanto a melhoria no processo de convergência com o uso da distribuição de Dirichlet para amostragem dos thresholds e a sua reparametrização. As simulação também demonstraram que os algoritmos NCG e NCt podem ser utilizados em diversas situações experimentais, para análise

de modelos mistos, obtendo estimativas precisas, tanto para a posteriori dos efeitos fixos, aleatórios, componentes da variância e valores preditivos.

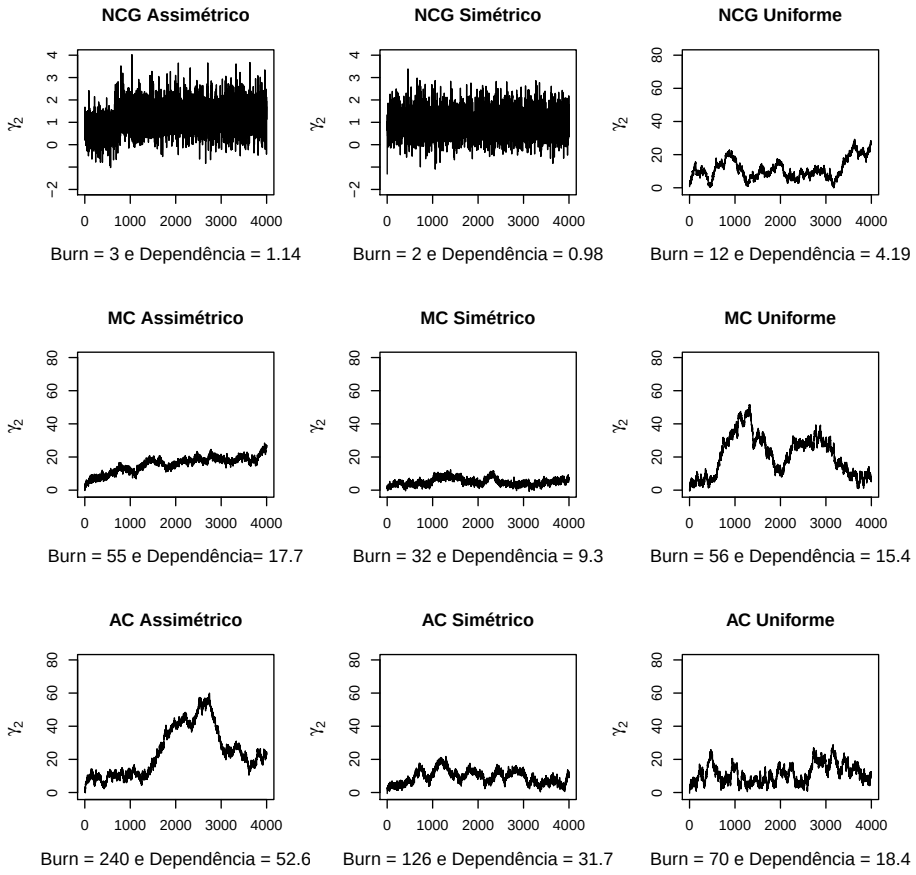

Figura 6 Trajetória da cadeia para um exemplo simulado no delineamento STS com $\rho = 0.2$, com burn=0, jump=0 e tamanho final da cadeia igual a 4000 para o threshold 2 com distribuição simétrica, assimétrica e uniforme para a variável resposta

5 UM EXEMPLO DE APLICAÇÃO

Com o presente experimento objetivou-se avaliar a resistência de 66 famílias de tomateiro à requeima. O delineamento utilizado foi de blocos incompletos, sendo cada bloco constituído por 33 plantas, sendo 10 repetições por família, totalizando 20 blocos. As avaliações da severidade da doença foram realizadas semanalmente durante 54 dias. A escala de avaliação da severidade da doença possuia 6 categorias, sendo que cada categoria representa 1%, 5%, 10%, 16%, 32% e 50% de severidade da doença (CORRÊA; BUENO-FILHO; CARMO, 2009). O modelo adotado para análise do experimento foi parcela subdividida no tempo, onde as parcelas são os blocos incompletos casualizados e as subparcelas são as observações no tempo dentro de cada planta. A análise inicial para os 6 algoritmos utilizados no presente estudo, se deu com uma cadeia inicial de 4000 iterações e analisada pelo teste de Raftery e Lewis para obtenção do tamanho de amostra ideal. A partir da amostra estimada pelo teste de Raftery e Lewis foram geradas duas cadeias para cada algoritmo e aplicado o teste de Gelman e Rubin para avaliar a convergência das amostras. O fator de Bayes Jeffreys (1961) foi aplicado entre os modelos com distribuição acumulada Normal e t-Student para verificar evidências sobre a superioridade de um modelo em relação ao outro na estimação dos parâmetros. Para o cálculo da estimativa do fator de Bayes, foi gerada uma cadeia para cada modelo com um descarte inicial de 100 iterações, salto de 10 iterações e um número efetivo de iterações de 4000. O fator de Bayes foi obtido conforme Gelman et al. (2003) utilizando a estimativa média para a log-verossimilhança da distribuição a posteriori. As análises do experimento foram realizadas utilizando um computador com processador Core(TM)i7-2600 de 3.4GHz com 16 Gb de memória RAM. Na tabela (4) são apresentados os parâmetros estimados e os parâmetros utilizados no processo de amostragem, como o descarte inicial (Burn), o salto (Jump) e o número de iterações efetivas (Iter).

Podemos observar que os algoritmos propostos por Kizilkaya et al. (2003) apresentaram os menores tempo de processamento e o algoritmo ACt apresentou o maior tempo de processamento, mas, para os parâmetros de descarte inicial, salto e número efetivo de iterações não houve grandes diferenças entre os algoritmos avaliados. Apesar do processo de simulação indicar um menor tempo de processamento para os algoritmos NCG e NCt, no presente experimento apenas o algoritmo ACt apresentou um tempo muito superior aos demais, sendo uma diferença de aproximadamente 93 vezes a mais no tempo de processamento. Apesar das menores estimativas apresentadas pelos algoritmos NCG e NCt para a variância genotípica, a correlação intraclasse estimada para os genótipos foi maior nestes algoritmos (tabela 4). Os resultados das estimativas da correlação intraclasse para os

Tabela 4 Parâmetros do processo de iteração após aplicação do teste de Raftery e Lewis para a obtenção da amostra ideal e parâmetros dos modelos a partir da amostra ideal.

	ACG	ACt	MCG	MCt	NCG	NCt
LogVeros.	-4813.46	-5072.754	-4302.46.16	-4701.44	-7095.68	-6801.88
σ^2_{gen}	0.137	0.127	0.198	0.141	0.059	0.059
σ^2_{bl}	0.028	0.052	0.031	0.029	0.036	0.026
σ^2_{res}	1.00	1.00	1.00	1.00	0.16	0.11
ρ_{gen}	0.12	0.11	0.16	0.12	0.26	0.34
Burn	688	698	333	31	6	194
Jump	37	45	96	1	9	66
Tempo (min.)	2.17	192.68	1.68	41.31	2.60	194.76

genótipos, não apresentaram valores semelhantes entre todos os algoritmos, apesar da priori utilizada para os componentes da variância ter sido a mesma priori informativa usada nas simulações.

A estimativa do fator de Bayes foi aproximadamente igual a 1 entre as comparações dos modelos com distribuição normal acumulada e os modelos com a distribuição t-Student acumulada, evidenciando que ambas distribuições apresentaram desempenho semelhante.

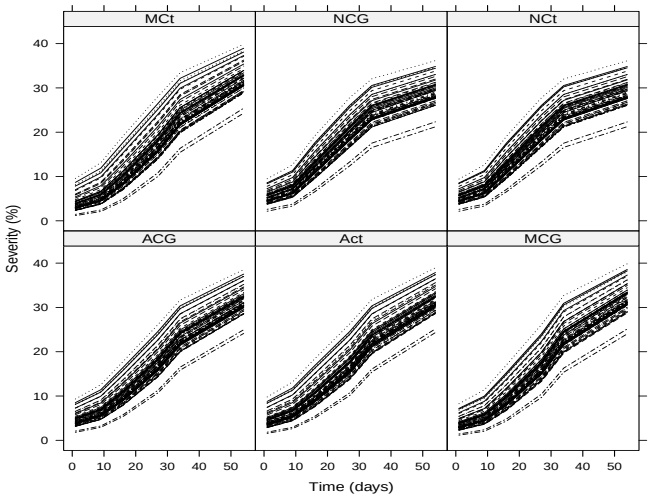

Figura 7 Valores preditos para os 66 gentótipos de tomate quanto a resistência da requeima nos 6 algoritmos utilizados para análise

Apesar das diferenças entre os tempos de processamento e parâmetros para convergência dos algoritmos, não houve diferenças no ordenamento dos genótipos quanto a resistência da doença, indicando que a predição entre os algoritmos foi semelhante. A figura (7) apresenta a predição dos 66 gentótipos quanto a resistência a doença nos algoritmos utilizados para análise. E a figura (8) apresenta a predição e o intervalo HPD para o genótipo mais resistente (genótipo 62), para um genótipo com resistência moderada (genótipo 10) e um genótipo suscetível à doença (genótipo 2)

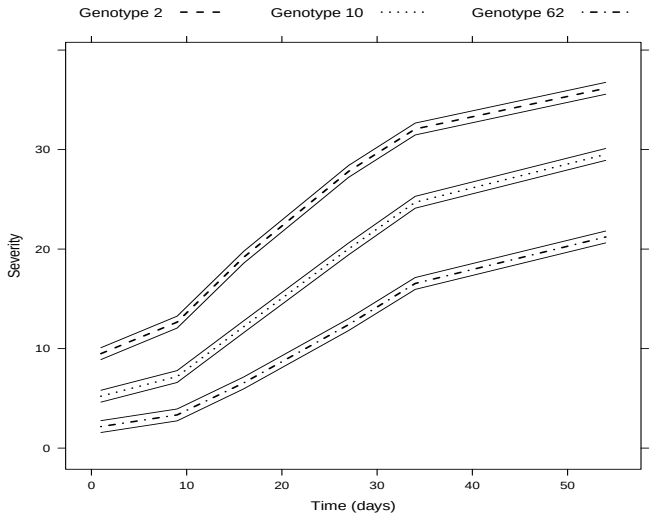

Figura 8 Valores preditos para 3 genótipos com diferentes níveis de resistência a doença. A linha contínua indica o intervalo HPD em relação ao valor predito para a severidade

6 CONCLUSÃO

A proposta de extensão do algoritmo de Nandran e Chen (1996) para análise de modelos mistos é uma alternativa mais rápida que os algoritmos apresentados por Albert e Chib (1993) e Kizilkaya et al. (2003), nas diversas situações experimentais simuladas. Os resultados observados nas simulações e no exemplo de aplicação não evidenciaram diferenças entre o uso da distribuição Gaussiana ou t-Student, indicando que a distribuição Gaussiana, mesmo em experimentos com um número pequeno de amostras pode ser utilizada.

REFERÊNCIAS

ALBERT, J. H.; CHIB, S. Bayesian analysis of binary and polychotomous response data. **Journal of the American Statistical Association**, v. 88, n. 442, p. 669–679, 1993.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, v. 88, n. 421, p. 9–25, 1993.

BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. Thurstonian models for sensory discrimination tests as generalized linear models. **Food Quality and Preference**, v. 21, n. 3, p. 330–338, 2010.

BROWN, P. **glmmBUGS: Generalised Linear Mixed Models and Spatial Models with WinBUGS, BRugs, or OpenBUGS**. [S.l.], 2010. R package version 1.9. Disponível em: <http://CRAN.R-project.org/package=glmmBUGS>.

BROWNE, W. J. **MCMC estimation in MLwiN**. Bristol, UK, 2011. Version 2.24. Disponível em: <http://www.bristol.ac.uk/cmm/software/mlwin/>.

BROWNE, W. J.; DRAPER, D. A comparison of bayesian and likelihood-based methods for fitting multilevel models. **Bayesian analysis**, v. 1, n. 3, p. 473–514, 2006.

BROWNE, W. J.; DRAPER, D. A comparison of bayesian and likelihood-based methods for fitting multilevel models. **Bayesian Analysis**, v. 1, n. 3, p. 473–514, 2006.

CORREA, F. M.; BUENO-FILHO, J. S. de S. **Bayesthresh: A package for categorical data analysis using Bayesian inference**. [S.l.], 2012.

CORRÊA, F. M.; BUENO-FILHO, J. S. S.; CARMO, M. G. F. Comparison of the three diagrammatic key for the quantification of late blight in tomato leaves. **Plant Pathology**, v. 58, n. 6, p. 1128–1133, 2009.

COWLES, M. K. Accelerating monte carlo markov chain convergence for cumulative link generalized linear models. **Statistics and Computing**, v. 6, n. 2, p. 101–111, 1996.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian data analysis**. [S.l.]: London: Chapman and Hall, 2003. 668 p.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, n. 4, p. 457–511, 1992.

GOB, R.; MCCOLLIN, C.; RAMALHOTO, M. F. Ordinal methodology in the analysis of likert scales. **Quality and Quantity**, v. 41, n. 5, p. 601–626, 2007.

HADFIELD, J. D. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. **Journal of Statistical Software**, v. 33, n. 2, p. 1–22, 2010. Disponível em: <http://www.jstatsoft.org/v33/i02/>.

HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Applied Statistics**, v. 32, n. 1, p. 69–83, 1976.

JEFFREYS, H. **Theory of probability**. [S.l.]: Oxford: Claredon Press, 1961. 470 p.

KIZILKAYA, K.; CARNIER, P.; ALBERA, A.; BITTANTE, G.; TEMPELMAN, R. J. Cumulative t-link threshold models for genetic analysis of calving ease scores. **Statistics and Computing**, v. 35, n. 1, p. 489–512, 2003.

MCCULLOCH, C. E.; SEARLE, S. R. **Generalized, linear, and mixed models**. [S.l.]: Wiley series in probability and statistics, 2001. 325 p.

MCCULLOGH, C. E.; SEARLE, S. R. **Generalized, Linear and Mixed Models**. [S.l.]: Wiley, New York, 2001.

NANDRAN, B.; CHEN, M. Reparameterizing the generalized linear model to accelerate gibbs sample convergence. **Journal of Statistical Computation and Simulation**, v. 54, n. 1, p. 129–144, 1996.

PIEPHO, H.-P.; KALKA, E. Threshold models with fixed and random effects for ordered categorical data. **Food Quality and Preference**, v. 14, n. 1, p. 343–357, 2003.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for mcmc. **R News**, v. 6, n. 1, p. 7–11, 2006. Disponível em: <http://CRAN.R-project.org/doc/Rnews/>.

POON, W.-Y.; WANG, H.-B. Latent variable models with ordinal categorical covariates. **Statistics and Computing**, v. 22, n. 5, p. 1135–1154, 2012.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2012. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

RAFTERY, A. E.; LEWIS, S. M. One long run with diagnostics: Implementation strategies for markov chain monte carlo. **Statistical Science**, v. 7, n. 4, p. 493–497, 1992.

SILVA, J. W. **Algoritmos para modelos de limiar utilizando as distribuições acumuladas Normal e t de Student**. Tese (Tese de Doutorado) — Universidade Federal de Lavras, 2008.

SILVA, J. W.; BUENO-FILHO, J. S. S. Um algoritmo para modelos de limiar usando as distribuições acumuladas normal e "t"de student. **Revista Brasileira de Matemática e Estatística**, v. 28, n. 3, p. 59–83, 2010.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. [S.l.]: Springer-Verlag New York, 2002. 740 p.

SORENSEN, D. A.; ANDERSEN, S.; GIANOLA, D.; KORSGAARD, I. Bayesian inference in threshold models using gibbs sampling. **Genetic Selection Evoution**, v. 27, n. 1, p. 229–249, 1995.

SPIEGELHALTER, D. J.; THOMAS, A.; LUNN, B. N. G. d. **WinBUGS User Manual**. Cambridge, UK, 2003. Version 1.4. Disponível em: <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>.

STEVENS, S. S. Measurement, statistics, and the schemapiric view. **Science**, v. 161, n. 3844, p. 849–856, 1968.

STRANDÉN, I.; GIANOLA, D. Attenuating effects of preferential treatment with student-t mixed linear models: a simulation study. **Genetics selection evolution**, v. 30, n. 1, p. 25–42, 1998.

TABOR, D. Mohs's hardness scale - a physical interpretation. **Proceedings of the Physical Society. Section B**, v. 67, n. 3, p. 249–257, 1954.

THERNEAU, T.; ATKINSON, E.; SINNWELL, J.; MATSUMOTO, M.; SCHAID, D.; MCDONNELL, S. **kinship2: Pedigree functions**. [S.l.], 2011. R package version 1.3.3. Disponível em: <http://CRAN.R-project.org/package=kinship2>.

WILKINSON, G. N.; ROGERS, C. E. Symbolic description of factorial models for analysis of variance. **Applied Statistics**, v. 22, n. 3, p. 392–399, 1973.

ARTIGO 2 Bayesthresh: library para análise de dados categorizados via inferência Bayesiana

RESUMO

A biblioteca Bayesthresh desenvolvida para o ambiente de programação R consiste na implementação dos algoritmos descritos por Albert e Chib (1993) (AC), Cowles (1996) (MC) e Nandram e Chen (1996) (NC) para análise bayesiana de dados categorizados ordinais. Foram também implementadas modificações deste algoritmos como as propostas por Silva (2008) (NCG, NCt) e Silva e Bueno Filho (2010) (ACt). A biblioteca apresenta uma estrutura flexível para análise de modelos lineares em que se possa considerar efeitos aleatórios (assumindo uma distribuição a priori comum). É possível especificar os hiperparâmetros das prioris para as componentes da variância, sendo que para os algoritmos NCG e NCt é também preciso especificar a priori para a variância do erro no traço latente. Estruturas de matrizes de variância para os efeitos aleatórios podem também ser incluídas. Algumas possibiidades de análise são apresentadas em um exemplo. Utilizamos nesta ilustração um experimento da análise sensorial de conservas de banana desidratadas sob diferentes concentrações de açúcar, descrito em Silva (2008).

Palavras-chave: Modelos mistos. Análise bayesiana. Análise de limiar. Biblioteca R.

ABSTRACT

Bayesthresh is a library developed for the R statistical programing environment. It implements algorithms to Bayesian analysis of ordinal categorical data described by Albert and Chib (1993) (AC), Cowles (1996) (MC) and Nandram and Chen (1996) (NC). Modifications of those algorithms as described by Silva (2008) (NCG, NCt) and Silva and Bueno Filho (2010) (ACt) were also implemented. The library brings a flexible structure for analysing linear mixed models assuming common prior distributions for random effects. It is also possible to specify hyperparameters for variance components. For the NC and NCt algorithms, hyperparameters for error variance in liability scale are also needed. Covariance structures for random effects can also be modeled. Some analytical possibilities are presented in an example from sensory analysis of dehydrated sweet banana produced with different sugar content, described by Silva (2008).

Keywords: Bayesian analysis. Mixed models. Threshold models. R package.

1 INTRODUÇÃO

Em diversas áreas do conhecimento há presença de dados categorizados, como por exemplo, na ciência dos alimentos com o uso da escala hedônica para avaliação de atributos sensoriais (PIEPHO; KALKA, 2003), na genética quantitativa animal (SORENSEN et al., 1995), na fitopatologia com as escalas diagramáticas para quantificação de doenças (CORRÊA; BUENO-FILHO; CARMO, 2009), entre outras. Diversos métodos de análises são propostos para tais variáveis (ALBERT; CHIB, 1993; KIZILKAYA et al., 2003; MCCULLOGH; SEARLE, 2001; PIEPHO; KALKA, 2003; SORENSEN et al., 1995). Dentre as diversas metodologias propostas, temos os métodos de Monte Carlo via Cadeias Markov (MCMC), que fornecem uma estratégia para marginalizar os efeitos aleatórios, sendo considerados robustos (BROWNE; DRAPER, 2006b). Uma alternativa de análise utlizando os métodos MCMC é o uso de modelos thresholds mistos, na qual uma variável latente, com distribuição contínua é especificada de forma que a resposta é observada em uma dada categoria se o valor desta variável está entre os limites que definem tal categoria (ALBERT; CHIB, 1993). A literatura disponibiliza diversos algoritmos para análise de modelos thresholds mistos, podendo esses algoritmos serem divididos em três grupos (tabela 5). O primeiro grupo utiliza a amostragem de Gibbs a partir da distribuição a posteriori conjunta para gerar os parâmetros thresholds, o segundo e o terceiro grupos utilizam o algoritmo Metropolis-Hastings (MH), sendo que um grupo utiliza a distribuição Gaussiana como geradora de candidatos e o outro grupo utiliza a distribuição Dirichlet como geradora de candidatos para os parâmetros thresholds.

Tabela 5 Algoritmos encontrados na literatura para análise de modelos thresholds mistos e seus respectivos processos de amostragem para obtenção dos parâmetros thresholds.

Amostragem	Variável latente	Autores	Algoritmo
Gibbs	Gaussiana	Sorensen et al. (1995)	ACG
	t-Student	Silva e Bueno-Filho (2010)	ACt
MH+Gaussiana	Gaussiana	Kizilkaya et al. (2003)	MCG
	t-Student	Kizilkaya et al. (2003)	MCt
MH+Dirichlet	Gaussiana	Silva (2008)	NCG
	t-Student	Silva (2008)	NCt

Apesar da existência de diversos algoritmos para análise de modelos thresholds mistos, problemas como a presença de forte aucorrelação das amostras ger-

adas durante o processo MCMC, afeta o desempenho dos algoritmos, necessitando de longas cadeias para que ocorra a convergência dos parâmetros estimados (ALBERT; CHIB, 1993; KIZILKAYA et al., 2003). Atualmente, temos diversos softwares que realizam a análise de modelos thresholds, WinBUGS (SPIEGELHALTER; THOMAS; LUNN, 2003), MLwiN (BROWNE, 2011), glmmBUGS (BROWN, 2010) e MCMCglmm (HADFIELD, 2010), porém, nenhum destes possui a implementação dos algoritmos adaptados de Nandran e Chen (1996) para modelos thresholds mistos e os demais algoritmos da tabela (5) em um único pacote. Os algoritmos de Nandran e Chen (1996) adaptados para modelos mistos surgem como uma alternativa para o problema de autocorrelação das amostras, resultando em menor tempo de processamento, além da flexibilidade quanto a variância do erro no traço latente (σ_ϵ^2), que tem como priori a distribuição gamma inversa. O pacote Bayesthresh está disponível em Comprehensiva R Archive Network em http://CRAN.R-project.org/package=Bayesthresh.

Neste artigo nós faremos um breve resumo sobre os algoritmos adaptados de Nandran e Chen (1996) para modelos mistos quanto a estratégia de estimação. Poucos resultados são apresentados e informações adicionais podem ser obtidas consultando as referências da tabela (5). Temos como principal objetivo apresentar um pacote para análise de modelos thresholds mistos utilizando inferência Bayesiana que seja rápido e de uso fácil para an'alises em diversas áreas do conhecimento.

2 O SOFTWARE

A ilustração do uso do software será realizada com um conjunto de dados de análise sensorial apresentado no artigo de Silva e Bueno-Filho (2010), cujo o interesse era de avaliar diferenças sensoriais em amostras de bananas desidratadas em 3 concentrações diferentes de sacarose, sendo 30%, 40% e 50%. A escala ordinal apresentava 9 valores, que variavam de 1 a 9. Para o presente experimento, 36 consumidores entre crianças e adultos foram consultados.

```
R> library(Bayesthresh)
```

2.1 Notação

A inserção dos termos no modelo pode ser realizada com fatores e variáveis numéricas nos efeitos fixos e aleatórios (2.39). Não há limites para o números de efeitos aleatórios a serem incluídos no modelo, sendo que cada efeito aleatório resulta em um componente da variância. A variável resposta fica a esquerda do operador $\sim$ e os demais termos separados por $'+'$. Os efeitos aleatórios são especificados do lado direito de uma barra vertical 'l'. A biblioteca segue a notação Wilkinson e Rogers (1973). Efeitos aninhados podem ser inseridos com o operador '\' tanto nos efeitos fixos como nos efeitos aleatórios e interações com o operador '*', interações entre efeitos aleatórios devem ser indicados com ':'.

$$Y \sim fixed_1 + \cdots + fixed_n + (1|random_1) + \cdots + (1|random_n)$$

2.2 Argumentos

A função Bayesthresh é provida dos algoritmos descritos na tabela (5), com interesse nas estimativas dos efeitos aleatórios e componentes da variância, sendo possível a inserção de estruturas de variâncias para os efeitos aleatórios. Para o exemplo utilizado, os níveis de sacarose serão analisados como fatores.

```
R> model <- Bayesthresh(cor ~ Sacarose + (1|Consumer),
+ data = sensory, Write=TRUE,
+ algor = list(algorithm = 'NC', link='Gaussian'),
+ burn = 50, jump = 5, ef.iter = 4000)
```

Nas próximas seções iremos descrever os principais argumentos da função Bayesthresh, como a inserção de uma matriz de variâncias (A) para os efeitos aleatórios, os parâmetros das prioris para os componentes da variância (σ^2_u) e variância residual (σ^2_ϵ), os algoritmos a serem utilizados e os parâmetros do processo de amostragem.

2.3 Estrutura da matriz dos efeitos aleatórios (A)

A inserção de uma matriz de variâncias para os efeitos aleatórios não é realizada de forma direta. A matriz deve ser construída com o auxílio da função $kinship()$ que está presente no pacote kinship2 (THERNEAU et al., 2011), que retorna a matrix descrita por Henderson (1976), que representa a correlação estrutural entre os efeitos aleatórios, sendo por padrão $A = I_{m \sim n}$

2.4 Prioris e algoritmos

Temos como parâmetros para as posterioris de σ_u^2 e σ_ϵ^2 uma distribuição gamma inversa, que deve ser especificada no argumento $priors$, sendo definido como uma $lista$. A $lista$ deve conter os argumentos ru e su que indicam respectivamente, os parâmetros de forma e escala para os componentes da variância. Os argumentos dre e dse são os parâmetros da gamma inversa para o erro no traço latente, que deve ser especificado apenas para os algoritmos NCG e NCt.

```
R> priors = list(ru = 10, su = 2, dre = 20, dse = 5)
```

É importante salientar que os algoritmos ACG, ACt, MCG, MCt apresentam variância do erro no traço latente igual a 1, por isto, não há necessidade de especificar a variância do erro no traço latente (SORENSEN; GIANOLA, 2002) . Por padrão, temos como parâmetros para σ_u^2 e σ_ϵ^2, uma $GI \sim (10{,}2)$ e $GI \sim (20{,}5)$, respectivamente, cuja a frequência no processo de amostragem é apresentado na figura (9). A especificação do algoritmo utilizado e da distribuição da variável latente é feita com o algoritmo $algor$, que é uma $lista$. O argumento $algor$ deve conter os objetos $algorithm$ e $link$, sendo o primeiro a indicação do algoritmo e o segundo a especificação da distribuição da variável latente. A tabela (6) apresenta as especificações dos algoritmos e suas prioris.

```
R> algor = list(algorithm='NC', link='Gaussian')
```

2.5 Parâmetros do processo de amostragem

Os parâmetros do processo de amostragem são definidos pelos argumentos $burn$ que indica o descarte inicial da cadeia, o $jump$ que representa o salto entre as iterações e o $ef.iter$ que indica o número efetivo de iterações após o descarte inicial e o salto entre as iterações. Por padrão, o valor inicial para o $burn$, $jump$ e $ef.iter$ é 25, 5 e 4000, respectivamente.

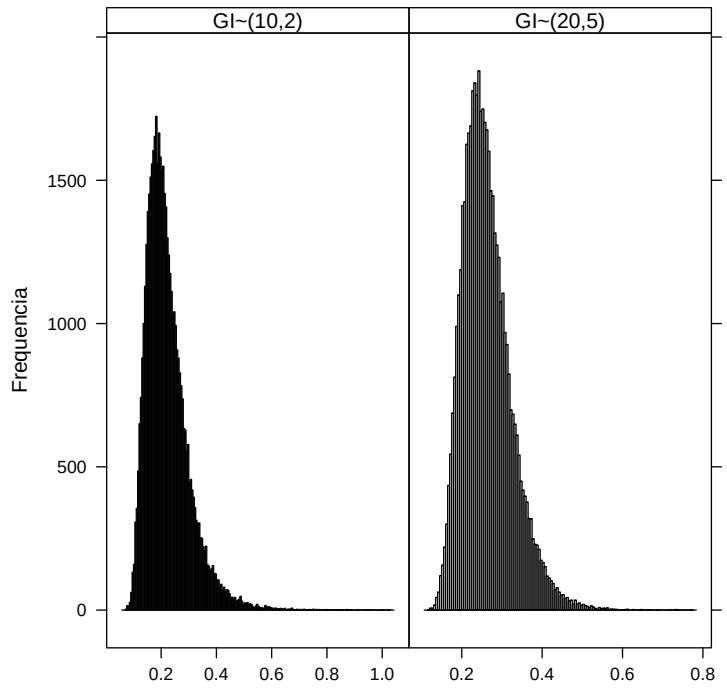

Figura 9 Frequência das prioris utilizadas como padrão no processo de amostragem

Tabela 6 Identificação dos algoritmos e dos parâmetros das prioris utilizadas para o processo de análise.

Algorithm	Link	σ_u^2	σ_ϵ^2
ACG	Gaussian	$GI \sim (ru,su)$	1
ACt	t	$GI \sim (ru,su)$	1
MCG	Gaussian	$GI \sim (ru,su)$	1
MCt	t	$GI \sim (ru,su)$	1
NCG	Gaussian	$GI \sim (ru,su)$	$GI \sim (dre,dse)$
NCt	t	$GI \sim (ru,su)$	$GI \sim (dre,dse)$

3 EXEMPLO DE ANÁLISE

O conjunto de dados utilizado por Silva e Bueno-Filho (2010) será utilizado para ilustrar o uso do pacote. O experimento teve como objetivo avaliar amostras de banana desidratada com três concentrações diferentes de açúcar, sendo 30%, 40% e 50%. Foi utilizado uma escala ordinal de 9 categorias que variava de 1 a 9. Para este exemplo, 36 consumidores entre crianças e adultos foram consultados.

Para ilustrar as diferenças entre os algoritmos implementados, iremos considerar os algoritmos AC, MC e NC utilizando a distribuição $Gaussiana$ para a variável latente. Uma análise inicial para determinar o descarte inicial (burn) e a dependência da cadeia (jump) será realizada com 4000 iterações totais e aplicado o teste de Raftery e Lewis (RAFTERY; LEWIS, 1992).

```
R> # Algoritmo padr\~ao (NAMDRAM;CHEN, 1996)
R> modelNC<- Bayesthresh(cor ~ Sacarose+(1|Consumer),
+   Write=TRUE, burn = 0, jump = 1)
R> # Algoritmo de Albert e Chib (1993)
R> modelAC<- Bayesthresh(cor ~ Sacarose+(1|Consumer),
+   Write=TRUE,
+   algor=list(algorithm='AC', link='Gaussian'),
+   burn = 0, jump = 1)
R> # Algoritmo de Cowles (1996)
R> modelMC<- Bayesthresh(cor ~ Sacarose+(1|Consumer),
+   Write=TRUE,
+   algor=list(algorithm='MC', link='Gaussian'),
+   burn = 0, jump = 1)
```

3.1 Saídas

O sumário das função $Bayesthresh$ apresenta a Deviance do modelo, a verossimilhaça marginal, a média a posteriori dos componentes da variância e efeitos fixos, os parâmetros do processo de amostragem e o tempo de processamento, em segundos. O desvio-padrão das médias a posteriori também são listadas. A cadeia do processo de amostragem só será armazenada caso o argumento $Write$ seja definido como TRUE, $Write = TRUE$, caso contrário a cadeia não será armazenada.

```
R> summary(modelNC)
```

```
Threshold model with algorithm NC and link Gaussian
Formula: cor ~ Sacarose + (1 | Consumer)

Deviance: 312.7847

Marginal Log-likelihood:
  Post. mean Post.std.dev
   -156.3923     4.262392

Random effects:
          Post.variance Post.std.dev
Consumer      0.2110930   0.07024418
Residuals     0.4627787   0.07458932

Fixed effects:
              Estimate  Std. Dev
(Intercept)  1.0296664 0.1620369
Sacarose40  -0.1940975 0.1833103
Sacarose50  -0.1669788 0.1805000

Iteraction Control:
Burn = 0 , Jump = 1 , Iteraction = 4000
Time elapsed 14.466 seconds
```

A manipulação das cadeias geradas durante o processo de amostragem pode ser realizada com o auxílio do pacote $coda$ Plummer et al. (2006), que possui diversas funções para manipular cadeias geradas por processos MCMC. Para ilustrar as diferenças entre as cadeias geradas para os thresholds, nos diferentes algoritmos utilizados para análise, iremos destacar o threshold 4 (Figura 2).

A cadeia armazenada por ser extraída com a função $MCMCsample$, que retorna uma lista com 3 elementos, sendo um que armazena o vetor da posteriori dos parâmetros estimados, outro elemento que armazena a posteriori dos componentes da variância e o terceiro que armazena os parâmetros dos thresholds amostrados.

```
R> chainAC <- MCMCsample(modelAC)
R> chainMC <- MCMCsample(modelMC)
R> chainNC <- MCMCsample(modelNC)
```

As diferenças no comportamento das cadeias dos parâmetros amostrados, pode ser observada no teste de Raftery e Lewis para diagnóstico da cadeia. Aplicando o teste, na cadeia a posteriori da análise considerando o algoritmo NC, e lembrnado que os thresholds, deste algoritmo, ficam limitados entre 0 e 1, temos:

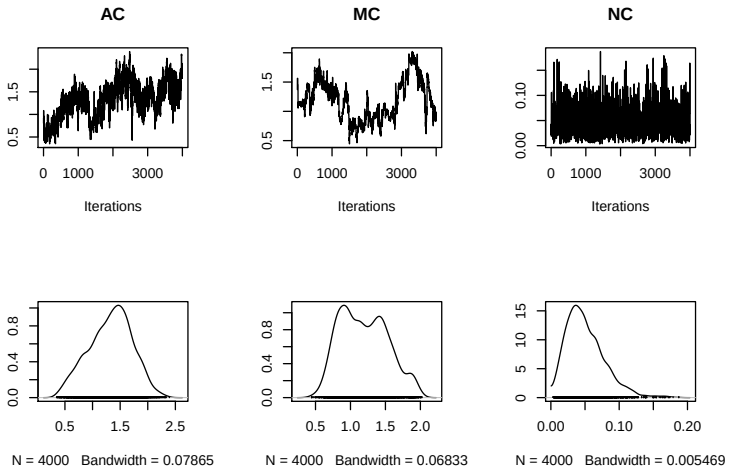

Figura 10 Traço e densidade da cadeia amostrada dos algoritmos AC, MC e NC para o threshold 4

```
R> round(chainNC[[3]][1:3,],3)
     [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7] [,8]
[1,]    0 0.014 0.021 0.022 0.082 0.263 0.455    1
[2,]    0 0.018 0.025 0.026 0.060 0.206 0.349    1
[3,]    0 0.002 0.002 0.032 0.103 0.171 0.404    1
```

Aplicando o teste na cadeia dos thresholds, temos:

```
R> raftery.diag(mcmc(
+    chainNC[[3]][,-c(1,ncol(chainNC[[3]]))]))
Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

 Burn-in  Total Lower bound  Dependence
 (M)      (N)   (Nmin)       factor (I)
 4        5167  3746         1.38
 5        5797  3746         1.55
 5        5747  3746         1.53
 4        4753  3746         1.27
 4        4753  3746         1.27
 4        5277  3746         1.41
```

Após o teste de Raftery e Lewis, obtemos duas cadeias válidas para avaliar a convergência dos parâmetros com o teste de Gelman e Rubin Gelman et al. (2003).

```
R> chainNC1<- Bayesthresh(cor ~ (1|Consumer)+Sacarose,
+     data=sensory,
+     Write=TRUE, burn=5, jump=2, ef.iter=4000)
R> chainNC2<- Bayesthresh(cor ~ (1|Consumer)+Sacarose,
+    data=sensory,
+    Write=TRUE, burn=5, jump=2, ef.iter=4000)
```

Aplicando o teste de Gelman e Rubin para testar convergência.

```
R> chainTheta <- gelman.diag(mcmc.list(mcmc(
+  MCMCsample(chainNC1)[[1]]),
+  mcmc(MCMCsample(chainNC2)[[1]])))
R> chainVari <- gelman.diag(mcmc.list(mcmc(
+  MCMCsample(chainNC1)[[2]]),
+  mcmc(MCMCsample(chainNC2)[[2]])))
```

É importante lembrar que os algoritmos NCG e NCt possuem os thresholds limitados entre 0 e 1, por isso a necessidade de excluí-los do teste de diagnóstico de convergência de Gelman e Rubin.

```
R> chainThresh <- gelman.diag(mcmc.list(mcmc(
+    MCMCsample(chainNC1)[[3]][,-c(1,8)]),
+    mcmc(MCMCsample(chainNC2)[[3]][,-c(1,8)])))
R> chainThresh
Potential scale reduction factors:

     Point est. Upper C.I.
[1,]          1       1.00
[2,]          1       1.00
[3,]          1       1.00
[4,]          1       1.01
[5,]          1       1.00
[6,]          1       1.00

Multivariate psrf

1
```

3.2 Efeitos fixos e aleatórios

A média a psoteriori dos efeitos fixos pode ser obtida com o comando $coef$, ou seu sinônimo $coefficients$. Intervalos de credibilidade HPD para a posteriori dos efeitos fixos podem ser obtidos com o argumento $HPDinterval = TRUE$, que por padrão é $FALSE$.

```
R> coef(chainNC2, HPDinterval=TRUE)
$Coefficients
             Estimate  Std. Dev
(Intercept)  1.0310335 0.1548962
Sacarose40  -0.2009396 0.1800662
Sacarose50  -0.1692107 0.1791730

$HPD.interval
                lower     upper
(Intercept)  0.7436855 1.3452706
Sacarose40  -0.5513032 0.1524372
Sacarose50  -0.5299764 0.1812917
attr(,"Probability")
[1] 0.95
```

Um sumário da média a posteriori dos efeitos aleatórios pode ser obtido com a função $random.effects$ e o comando $plot$ apresenta a média a posteriori e os intervalos de credibilidade pivotal e HPD, sendo o padrão os intervalos pivotal para as médias a posteriori (Figura 3). Para plotar intervalos de credibilidade HPD, o argumento $interval$ deve ser definido como $'hpd'$, por padrão é $'confidence'$.

```
R> aleat <- random.effects(chainNC2, HPDinterval=TRUE)
```

Os valores preditos podem ser obtidos pela função $predict$.

```
R> predict(chainNC2)
```

3.3 Fator de Bayes

Comparações entre 2 modelos podem ser realizadas com o uso do Fator de Bayes, conforme (GELMAN et al., 2003), o qual de forma geral é dado por:

$$B_{ij} = \frac{p(y|M_i)}{p(y|M_j)} = \frac{\int p(y|\theta_i,M_i)p_i(\theta_i|M_i)d\theta}{\int p(y|\theta_j,M_j)p_j(\theta_j|M_j)d\theta}$$

```
R> plot(aleat, interval="hpd", main="HPD interval")
```

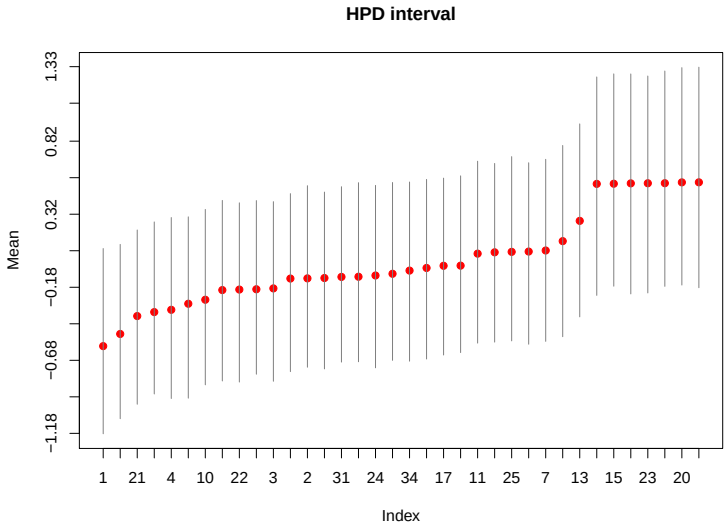

Figura 11 Média a posteriori dos efeitos aleatórios para os Consumidores e os intervalos de credibilidade HPD

em que $p(y|M_i)$ e $p(y|M_j)$ é a probabilidade marginal do respectivo modelo. A obtenção da probabilidade marginal é realizada com a amostra a $posteriori$, na escala log. O estimador da marginal dos dados é a média aritmética dos valores de verossimilhança obtidos a cada ciclo do processo MCMC.
Para verificar se há evidências de que um modelo apenas com a média geral apresenta a mesma explicação que um modelo com a fonte de variaçõSacarose, nós utilizamos o fator de Bayes conforme a sintaxe abaixo. Sendo que a média geral é indicada por 1.

```
R> model2 <- Bayesthresh(cor ~ -1 + (1|Consumer),
+           data = sensory,
+           burn = 5, jump = 2, ef.iter = 4000)
```

Aplicando o fator de Bayes nos modelos 1 e 2, temos:

```
R> Bayes.factor(chainNC1,model2, inter=TRUE)
Bayes factor for comparison two models


Model 1: cor ~ (1 | Consumer) + Sacarose
Model 2: cor ~ -1 + (1 | Consumer)
```

```
              Bayes factor
model1/model2     1.028316

 Scale for interpretation of the Bayes factor
------------------------------------------
B_ij                Evidence in favor of M_1
------------------------------------------
  <1                 negative (favor of M_2)
1 to 3                            doubtfull
3 to 10                         substantial
10 to 30                             strong
30 to 100                       very strong
 >100                              decisive
------------------------------------------
Jeffreys(1961)
```

Caso fossem detectadas diferenças entre os níveis de Sacarose, seria mais adequado considerar tais níveis como variável contínua e ajustar um modelo de regressão.

O sumário da função Bayesthresh, tem por padrão o argumento $inter = TRUE$, que retorna um quadro auxiliar para interpretação do resultado do fator de Bayes. Para suprimir o quadro, basta especificar $inter = FALSE$

4 CONSIDERAÇÕES FINAIS

Este artigo apresenta uma biblioteca em R para análise de modelos thresholds mistos utilizando o processo de Monte Carlo via Cadeias Markov de aproximações numéricas para a distribuição a posteriori dos parâmetros do modelo. Um aspecto importante do pacote é a inclusão de dois algoritmos cuja a variância do erro do traço latente é parâmetro do modelo, podendo ser amostrada com o uso de uma distribuição a priori gamma inversa. Além disso, permite a inclusão de uma matriz para estruturar os efeitos aleatórios. O pacote apresenta diversos algoritmos com diferentes características para análises de dados categorizados de forma simples e rápida. Informações teóricas sobre a implementação dos algoritmos, podem ser obtidas com maiores detalhes nas referências apresentadas na tabela (5).

REFERÊNCIAS

ALBERT, J. H.; CHIB, S. Bayesian analysis of binary and polychotomous response data. **Journal of the American Statistical Association**, v. 88, n. 442, p. 669–679, 1993.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, v. 88, n. 421, p. 9–25, 1993.

BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. Thurstonian models for sensory discrimination tests as generalized linear models. **Food Quality and Preference**, v. 21, n. 3, p. 330–338, 2010.

BROWN, P. **glmmBUGS: Generalised Linear Mixed Models and Spatial Models with WinBUGS, BRugs, or OpenBUGS**. [S.l.], 2010. R package version 1.9. Disponível em: <http://CRAN.R-project.org/package=glmmBUGS>.

BROWNE, W. J. **MCMC estimation in MLwiN**. Bristol, UK, 2011. Version 2.24. Disponível em: <http://www.bristol.ac.uk/cmm/software/mlwin/>.

BROWNE, W. J.; DRAPER, D. A comparison of bayesian and likelihood-based methods for fitting multilevel models. **Bayesian analysis**, v. 1, n. 3, p. 473–514, 2006.

BROWNE, W. J.; DRAPER, D. A comparison of bayesian and likelihood-based methods for fitting multilevel models. **Bayesian Analysis**, v. 1, n. 3, p. 473–514, 2006.

CORREA, F. M.; BUENO-FILHO, J. S. de S. **Bayesthresh: A package for categorical data analysis using Bayesian inference**. [S.l.], 2012.

CORRÊA, F. M.; BUENO-FILHO, J. S. S.; CARMO, M. G. F. Comparison of the three diagrammatic key for the quantification of late blight in tomato leaves. **Plant Pathology**, v. 58, n. 6, p. 1128–1133, 2009.

COWLES, M. K. Accelerating monte carlo markov chain convergence for cumulative link generalized linear models. **Statistics and Computing**, v. 6, n. 2, p. 101–111, 1996.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian data analysis**. [S.l.]: London: Chapman and Hall, 2003. 668 p.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, n. 4, p. 457–511, 1992.

GOB, R.; MCCOLLIN, C.; RAMALHOTO, M. F. Ordinal methodology in the analysis of likert scales. **Quality and Quantity**, v. 41, n. 5, p. 601–626, 2007.

HADFIELD, J. D. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. **Journal of Statistical Software**, v. 33, n. 2, p. 1–22, 2010. Disponível em: <http://www.jstatsoft.org/v33/i02/>.

HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Applied Statistics**, v. 32, n. 1, p. 69–83, 1976.

JEFFREYS, H. **Theory of probability**. [S.l.]: Oxford: Claredon Press, 1961. 470 p.

KIZILKAYA, K.; CARNIER, P.; ALBERA, A.; BITTANTE, G.; TEMPELMAN, R. J. Cumulative t-link threshold models for genetic analysis of calving ease scores. **Statistics and Computing**, v. 35, n. 1, p. 489–512, 2003.

MCCULLOCH, C. E.; SEARLE, S. R. **Generalized, linear, and mixed models**. [S.l.]: Wiley series in probability and statistics, 2001. 325 p.

MCCULLOGH, C. E.; SEARLE, S. R. **Generalized, Linear and Mixed Models**. [S.l.]: Wiley, New York, 2001.

NANDRAN, B.; CHEN, M. Reparameterizing the generalized linear model to accelerate gibbs sample convergence. **Journal of Statistical Computation and Simulation**, v. 54, n. 1, p. 129–144, 1996.

PIEPHO, H.-P.; KALKA, E. Threshold models with fixed and random effects for ordered categorical data. **Food Quality and Preference**, v. 14, n. 1, p. 343–357, 2003.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for mcmc. **R News**, v. 6, n. 1, p. 7–11, 2006. Disponível em: <http://CRAN.R-project.org/doc/Rnews/>.

POON, W.-Y.; WANG, H.-B. Latent variable models with ordinal categorical covariates. **Statistics and Computing**, v. 22, n. 5, p. 1135–1154, 2012.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2012. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

RAFTERY, A. E.; LEWIS, S. M. One long run with diagnostics: Implementation strategies for markov chain monte carlo. **Statistical Science**, v. 7, n. 4, p. 493–497, 1992.

SILVA, J. W. **Algoritmos para modelos de limiar utilizando as distribuições acumuladas Normal e t de Student**. Tese (Tese de Doutorado) — Universidade Federal de Lavras, 2008.

SILVA, J. W.; BUENO-FILHO, J. S. S. Um algoritmo para modelos de limiar usando as distribuições acumuladas normal e "t"de student. **Revista Brasileira de Matemática e Estatística**, v. 28, n. 3, p. 59–83, 2010.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. [S.l.]: Springer-Verlag New York, 2002. 740 p.

SORENSEN, D. A.; ANDERSEN, S.; GIANOLA, D.; KORSGAARD, I. Bayesian inference in threshold models using gibbs sampling. **Genetic Selection Evoution**, v. 27, n. 1, p. 229–249, 1995.

SPIEGELHALTER, D. J.; THOMAS, A.; LUNN, B. N. G. d. **WinBUGS User Manual**. Cambridge, UK, 2003. Version 1.4. Disponível em: <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>.

STEVENS, S. S. Measurement, statistics, and the schemapiric view. **Science**, v. 161, n. 3844, p. 849–856, 1968.

STRANDÉN, I.; GIANOLA, D. Attenuating effects of preferential treatment with student-t mixed linear models: a simulation study. **Genetics selection evolution**, v. 30, n. 1, p. 25–42, 1998.

TABOR, D. Mohs's hardness scale - a physical interpretation. **Proceedings of the Physical Society. Section B**, v. 67, n. 3, p. 249–257, 1954.

THERNEAU, T.; ATKINSON, E.; SINNWELL, J.; MATSUMOTO, M.; SCHAID, D.; MCDONNELL, S. **kinship2: Pedigree functions**. [S.l.], 2011. R package version 1.3.3. Disponível em: <http://CRAN.R-project.org/package=kinship2>.

WILKINSON, G. N.; ROGERS, C. E. Symbolic description of factorial models for analysis of variance. **Applied Statistics**, v. 22, n. 3, p. 392–399, 1973.