41ª Reunião Anual da Sociedade Brasileira de Zootecnia
19 de Julho a 22 de Julho de 2004 - Campo Grande, MS

CONVERGÊNCIA PARA ESTIMATIVAS DE COMPONENTES DE (CO)VARIÂNCIAS OBTIDAS POR MEIO DO AMOSTRADOR DE GIBBS "1"

AUTORES

SANDRA MARIA SIMONELLI "2", ELIAS NUNES MARTINS "3", EDUARDO SHIGUERO SAKAGUTI "3", ELIANE GASPARINO "3", LUIS OTÁVIO C. SILVA "4", DANIEL PEROTTO "5"

[1] parte da tese de doutorado em Zootecnia do primeiro autor.
[2] professora do CESUMAR E CIES
[3] professores do curso de Zootecnia da Universidade Estadual de Maringá - Av. Colombo, 5790. Cep. 87020-900. Maringá-PR
[4] pesquisador da Embrapa - Gado de Corte - MS
[5] pesquisador do IAPAR - Ponta Grossa – PR

RESUMO

Foram estimados componentes de (co)variâncias pelo o amostrador de Gibbs. Foram utilizados diferentes intervalos de retirada de uma cadeia de 530.000 iterações, com período de descarte inicial de 5.000 iterações. O conjunto final de dados conteve 68.345 informações de bovinos da raça Nelore criados em cinco regiões do estado do Mato Grosso do Sul, provenientes Embrapa-Gado de Corte. Foi estudado o peso aos 205 dias de idade(P205) considerada em cada região como características distintas. Foi assumido para os efeitos fixos distribuição uniforme e para os componentes de variância genética direta e materna, permanente de ambiente e residual distribuição de Wishart Invertida. Os componentes de (co)variâncias foram obtidos pelo MTGSAM e foram calculadas médias e desvios padrão para os valores gerados em cada um dos intervalos. A verificação da convergência foi realizada pela biblioteca CODA ("Convergence Diagnosis and Output Analysis"). O componente de covariância entre o efeito genético direto da região 5 e materno da região 2 foi o único que não obteve a convergência. O comportamento das médias posteriores foi semelhante em todos os intervalos. Pôde-se concluir que uma cadeia não atingir a convergência, depois de um grande número de iterações, não implicou em modificações nas estimativas.

PALAVRAS-CHAVE

bovinos de corte, amostrador de gibbs, critério de convergência, monte carlo, cadeias de markov

TITLE

CONVERGENCE FOR ESTIMATIONS OF COMPONENTS OF (CO)VARIANCES OBTAINED THROUGH THE GIBBS SAMPLER

ABSTRACT

Components of covariances were estimated using Gibbs sampler to study the convergence of the Markov chains in different thinning intervals of a chain of 530.000 interactions with burn-in period of 5.000 iterations. The data had 68.345 records from the Embrapa – Beef Cattle of the five regions of the state of Mato Grosso do Sul. It was the weight at the weaning period, adjusted at 205 days old considered in each region as distinct traits. For the fixed effects flat prior distribution was assumed and for the components of permanent of environment, maternal, and direct genetic random and residual variance, distribution of Inverted Wishart was assumed. The components of (co)variances were obtained through the MTGSAM program and the average and standard deviations for the generated values in each one of the intervals were calculated. The verification of the convergence of the distributions of the chains generated by the Gibbs Sampler was carried out by the library CODA (Convergence Diagnosis and Output Analysis). The component of covariance between the direct genetic effects from the region 5 and maternal from the region 2 was the only one that didn't obtain the convergence with any thinning intervals. The behaviour of the subsequent averages in the differents intervals was similar enough between themselves. However, a chain don't reach such convergence, after a big number of interactions, doesn't implicate in modifications in the estimates.

KEYWORDS

beef cattle, convergence criteria, gibbs sampler, monte carlo method, markov chains.

INTRODUÇÃO

Os métodos de Monte Carlo via Cadeias de Markov (MCMC) têm sido bastante utilizados em simulações a partir de distribuições posteriores aproximadas. A idéia do MCMC é criar cadeias aleatórias, ou processos de Markov, que tenham uma distribuição posterior estacionária resultando em uma amostra aproximada da função de verossimilhança (Qian, et al., 2003). Há vários caminhos para gerar cadeias aleatórias por MCMC. No entanto, (Geman e Geman, 1984) todos esses caminhos requerem o uso do algoritmo de Metropolis Hastings, especialmente o amostrador de Gibbs (Metropolis et al., 1953; Hastings, 1970). O Amostrador de Gibbs é uma técnica iterativa de Monte Carlo utilizada quando uma integração de alta dimensão é requerida. Um dos problemas das cadeias de Markov é indicar quando a convergência é alcançada. Para resolver esse problema, um número de diferentes diagnósticos tem sido propostos. A origem teórica, as suposições, o número de cadeias que são necessárias, os amostradores para os quais o diagnóstico é aplicado e a interpretabilidade são alguns critérios para a escolha do diagnóstico. Atualmente, vários diagnósticos de convergência têm sido utilizados, dentre eles o de Heideberger e Welch (1983) estão entre os mais populares. O objetivo deste trabalho foi estudar a convergência das cadeias de Markov utilizando o amostrador de Gibbs em diferentes intervalos de retirada e verificar quais suas implicações dentro do melhoramento genético animal.

MATERIAL E MÉTODOS

Os dados utilizados foram provenientes da Associação Brasileira de Criadores de Zebu (ABCZ) e estão sobre a responsabilidade pela Embrapa – Gado de Corte. Foi obtido um conjunto final de 68.345 informações de animais da raça Nelore criados em cinco regiões do estado do Mato Grosso do Sul sendo elas região do Alto Taquari (1); regiões Bodoquena e Dourados (2); região de Campo Grande (3); região do Pantanal (4) e Paranaíba Três Lagoas (5). O número de observações em cada região é mostrado na Tabela 1. A característica estudada neste trabalho foi o peso à desmama, ajustado para 205 dias de idade (P205). Somente foram mantidos grupos contemporâneos com 15 observações ou mais. Os componentes de co(variância) foram obtidos por meio do programa MTGSAM - Multiple Trait Gibbs Sampling in Animal Models desenvolvido por Van Tassel e Van Vleck (1995). A mesma característica em cada região foi considerada como sendo características distintas utilizando-se um modelo que incluiu como efeitos fixos o grupo contemporâneo (GC) o sexo dos animais e a covariável idade da mãe ao parto e os efeitos aleatórios genéticos diretos e maternos, permanentes de ambiente e residual. Para os efeitos fixos foi assumida distribuição uniforme e para os componentes de variância genética direta e materna, permanente de ambiente e residual foi assumido distribuição de Wishart Invertida *(IW).* A função densidade de probabilidade conjunta dos parâmetros e a informação dos “priors” foram escritas como o produto das distribuições dos priors e a função de verossimilhança (Van Tassel e Van Vleck, 1996). O número de iterações descartadas para retirar o efeito dos valores iniciais e para que as amostras pudessem ser consideradas amostras da distribuição posterior foi de 5.000. Foram adotados intervalos de retirada de 40, 200, 400, 600, 800, 1.000 e 1200 iterações sendo consideradas amostras de uma cadeia de 530.000 iterações e foi verificada a convergência da distribuição das cadeias geradas em cada um destes intervalos. Foram calculadas as médias e desvios padrão para os valores gerados em cada um dos intervalos e comparações entre as estimativas foram feitas.

A verificação da convergência das distribuições das cadeias geradas pelo Gibbs Sampler foi realizada pela biblioteca CODA (“Convergence Diagnosis and Output Analysis”) versão 0.4, desenvolvido por Cowles et al. (1995), e o método adotado neste trabalho foi o de Heidelberg e Welch (1983).

RESULTADOS E DISCUSSÃO

Primeiramente foi aplicado o diagnóstico de convergência de Heidelberg e Welch (1983) para os componentes de (co)variâncias de todas as regiões. Foi observado que para os componentes de variâncias, todas as cadeias geradas atingiram a convergência. No entanto, para alguns componentes de covariância, tanto entre os efeitos genéticos diretos, entre os genéticos maternos

e entre os efeitos genéticos diretos e maternos, nem todas as cadeias geradas pelo Amostrador de Gibbs atingiram a convergência, quando o intervalo de retirada foi de 40 iterações. Na Tabela 2 são mostrados os componentes de (co) variâncias cujas cadeias não atingiram a convergência e suas respectivas médias posteriores. Para efeito de comparação foram apresentados alguns componentes de (co)variância cujas cadeias geradas pelo amostrador de Gibbs atingiram a convergência. Observa-se (Tabela 2) a convergência para os diferentes intervalos de retirada variou entre os componentes de (co)variâncias. O componente de covariância entre o efeito genético direto da região 5 e materno da região 2 foi o único que não obteve a convergência com nenhum intervalo de retirada não atingindo assim distribuição estacionária. Quanto aos comportamentos das médias posteriores nos diferentes intervalos, observa-se (Tabela 2) que foram bastante semelhantes entre si, tanto para as cadeias que obtiveram como para as cadeias que não obtiveram a convergência. Além disso, ao consideramos os desvios padrão posteriores do componente de variância, as médias posteriores são abrangidas em todos os intervalos. Pode-se dizer então, quanto à convergência, que o comportamento dos componentes, em relação às suas médias e desvios padrão, não se alterou bruscamente. Isso é de suma importância dentro do melhoramento genético, pois pode-se dizer que depois de um grande número de iterações, o fato da cadeia, para um determinado componente, não obter a convergência não implica, necessariamente, na mudança expressiva de suas estimativas posteriores. Os procedimentos bayesianos por meio do amostrador de Gibbs permitem, além das estimativas pontuais, determinar intervalos de credibilidade para distribuição posterior dos componentes de (co)variância. Um intervalo de credibilidade, segundo Casella e George (1992), pode ser definido pela região de alta densidade (RAD) posterior do parâmetro quando a distribuição for simétrica. A RAD é a região que contém (1 - α)100% da probabilidade posterior, em que α é o nível de significância. Quando a densidade posterior for assimétrica, a RAD e o intervalo de credibilidade podem diferir substancialmente. As estimativas dos componentes serão tanto mais precisas quanto menores forem seus intervalos de credibilidade. Neste trabalho observou-se que os intervalos de credibilidade e as RADS foram bastante próximos nos diferentes intervalos de retirada para todos os componentes, independentemente da ocorrência ou não da convergência, mostrando a simetria das distribuições posteriores. O que pôde-se notar é que o padrão de comportamento para os componentes que convergiram foi o mesmo que para os componentes que convergiram, mostrando que a precisão das estimativas se mantém depois de um grande número de iterações, mesmo sem ter atingido a convergência.

CONCLUSÕES

Não houve grande modificação nas médias e desvios padrão posteriores dos componentes de (co)variâncias quando houve ou não a convergência nos diferentes intervalos de retirada. Uma cadeia não atingir a convergência, depois de um grande número de iterações, não implica em modificações nas estimativas encontradas não implicando em modificações nos parâmetros genéticos.

REFERÊNCIAS BIBLIOGRÁFICAS

1.CASELLA, G.; GEORGE, E. Explaining the Gibbs Sampler. The American Statistical, v.46, p.167-174, 1992

2.COWLES, M. K.; BEST, N.; VINES, K. Convergence Diagnostics and Output Analysis. MRC Biostatistics Unit, UK. Version 0.40. 1995.

3.HEIDELBERGER, P.; WELCH, P. D. Simulation run length control in the presence of an initial transient. Operations Research, v.31, p.1109-11144, 1983.

4.QIAN, S.S.; STOW, C.A.; BORSUK, M. E. On Monte Carlo methods for Bayesian inference. Ecological modeling. v.159, p.269-277,2003.

5.VAN TASSEL, C. P. The use of Gibbs sampling for variance component estimation with simulated and weaning weight data using animal and maternal effects models. Ithaca, NY: Cornell University, 1994. 120p. PhD Thesis (Animal breeding)- Cornell University, 1994.

6.VAN TASSEL, C. P.; VAN VLECK, D. L. A set of FORTRAN programs to apply Gibbs sampling to animal models for variance component estimation [DRAFT]. In: A manual for use of MTGSAM. U. S Departament of Agriculture: Agricultural Research Service.1995.

TABELA 1 - Distribuição do conjunto final de dados de peso à desmama (P205), por região do estado do Mato Grosso do Sul.

Região	**P205**	
	Nº	**%**
1 – Alto Taquari (1)	3256	5
2 – Bodoquena e Dourados (2)	31566	46
3 – Campo Grande (3)	5928	9
4 – Pantanal (4)	8477	12
5 – Paranaíba e Três lagoas (5)	19118	28
Total	68345	100

TABELA 2 - Estimativas da convergência, médias (acima) e desvios - padrão (abaixo) posteriores dos componentes de (co)variância para o P205 com intervalos de retirada de 40, 200, 400, 600, 800, 1000 e 1200 iterações geradas pelo amostrador de Gibbs.

Componentes de (co)variâncias	Intervalos de retirada						
	40	200	400	600	800	1000	1200
S^{2}_{a1}	387,11* 20,10	387,35* 19,99	387,36* 20,42	387,37* 20,98	386,43* 21,03	387,59* 19,60	387,23* 19,4
S_{a1a2}	15,33 * 21,43	15,37* 21,39	15,18* 21,4	15,22* 20,96	14,59* 21,31	14,87* 22,16	14,64* 22,20
S_{a1a3}	15,75nc 34,55	15,77nc 34,45	15,71* 34,50	15,33* 34,64	15,39* 34,49	14,34* 34,25	16,11* 35,63
S_{a1m1}	-180,42 * 19,84	-180,5* 19,77	-180,32* 20,11	-180,5* 20,26	-180,36* 20,03	-181,2* 20,06	-180,19* 19,90
S_{a1m3}	-5,39nc 25,10	-5,40nc 25,16	-5,10* 25,13	-5,14* 24,97	-5,07* 24,67	-5,43* 24,68	-5,39* 25,35
S_{a5m2}	-12,78 nc 12,90	-12,8nc 12,90	-12,80nc 13,02	-12,8nc 13,06	-12,52nc 13,0	-12,4nc 12,58	-13,08nc 13,09
S_{a5m3}	-4,44* 20,37	-4,44* 20,37	-4,39* 20,68	-4,2* 20,07	-4,73* 20,22	-4,79* 20,01	-4,48* 20,53
S_{m2m4}	6,09* 11,63	6,09* 11,63	6,05* 11,46	6,15* 11,60	6,28* 11,74	6,0 11,46	6,02 11,82
S_{m2m5}	9,73 nc 8,73	9,73nc 8,73	9,83nc 8,77	9,15* 8,76	9,54* 8,60	9,58* 8,53	9,89* 8,61

S^{2}_{a1} = variância para o efeito genético direto da região 1; S_{a1a2} = covariância entre os efeitos genéticos diretos das regiões 1 e 2; S_{a1a3} = covariância entre os efeitos genéticos diretos das regiões 1 e 3; S_{a1m1} = covariância entre os efeitos genéticos direto da região 1 e materno da região 3; S_{a5m2} = covariância entre os efeitos genéticos direto da região 5 e materno da região 2; S_{a5m3} = covariância entre os efeitos genéticos direto da região 5 e materno da região 3; S_{m2m4} = covariância entre os efeitos genéticos maternos das regiões 2 e 4; S_{m2m5} = covariância entre os efeitos genéticos maternos das regiões 2 e 5;

* - Cadeias que atingiram a convergência nc - Cadeias que não atingiram a convergência