O uso da estatística Bayesiana no melhoramento genético animal: uma breve explicação

MARCOS JUN-ITI YOKOO[1]; GUILHERME JORDÃO DE MAGALHÃES ROSA[2]; FERNANDO FLORES CARDOSO[1]; CLÁUDIO DE ULHÔA MAGNABOSCO[3]; LUCIA GALVÃO ALBUQUERQUE[4]

[1]Empresa Brasileira de Pesquisa Agropecuária, Centro de Pesquisa de Pecuária dos Campos Sul-Brasileiros - Embrapa Pecuária Sul - CPPSul, BR 153, km 603, Caixa Postal 242, CEP 96401-970, Bagé/RS, Brasil. E-mail: marcos.yokoo@embrapa.br. *Autor para correspondência
[2]Department of Animal Science, University of Wisconsin, Madison, WI 53706, USA
[3]Empresa Brasileira de Pesquisa Agropecuária, Embrapa Cerrados - CNPAF, BR 020, km 18, Caixa Postal 08223, CEP 73310-970, Planaltina/DF, Brasil
[4]Universidade Estadual Paulista - UNESP, Faculdade de Ciências Agrárias e Veterinárias, Departamento de Zootecnia, Jaboticabal/SP, Brasil. E-mail: lgalb@fcav.unesp.br

RESUMO

Este trabalho aborda o estudo de técnicas bayesianas no melhoramento genético animal, no intuito de discutir e elucidar essa abordagem frente à "estatística frequentista". Apresentam-se dois algoritmos de integração estocástica por meio da simulação de Monte Carlo em Cadeias de Markov (MCMC): o Amostrador de Gibbs e o Metropolis-Hastings. Considera-se a aplicação das mencionadas técnicas como uma alternativa aos programas de melhoramento animal, na estimação de parâmetros genéticos em ordem de solucionar problemas relacionados aos modelos mais complexos e a expressão de características de interesse econômico que não tenham distribuição normal. Assim, alternativas bayesianas constituem uma excelente estratégia para contornar essas situações, por meio de modelos com outras distribuições, além da normal, tendo mostrado eficiência e fácil implementação para estimação confiável de parâmetros. Além disso, a inferência bayesiana tem a vantagem adicional de permitir a incorporação de informações passadas (*a priori*), melhorando o processo de estimação. As abordagens propostas são explicadas e discutidas no desenvolvimento do trabalho.
Palavras-chave: amostrador de Gibbs; método de MCMC; Metropolis-Hastings; parâmetro; verossimilhança.

ABSTRACT

The use of Bayesian statistics in animal breeding: a brief explanation

This paper describes the study of Bayesian techniques in animal breeding, aiming to discuss and clarify this approach compared to "frequentist statistics". We present two algorithms for stochastic integration by using Markov Chain Monte Carlo (MCMC) simulation: Gibbs sampling and Metropolis-Hastings algorithm. The application of these techniques is considered an alternative to animal breeding programs in the estimation of genetic parameters to solve problems related to more complex models and the expression of traits of economic interest that do not have normal distribution. Therefore, Bayesian techniques become an excellent strategy to elucidate these situations, by using models with alternative distributions, which have shown efficiency and easy implementation for reliable estimation of parameters. Furthermore, Bayesian inference has the additional advantage of allowing the incorporation of previous (*a priori*)

SAP 6974
Data do envio: 27/08/2012
Data do aceite: 20/01/2013

information, improving the estimation process. The proposed approaches are explained and discussed in the development of this paper.
Keywords: Gibbs sampler; MCMC method; Metropolis-Hastings; parameter; verisimilitude.

INTRODUÇÃO

Nos programas de melhoramento genético animal, o objetivo principal é a mudança da média fenotípica do rebanho utilizando basicamente duas ferramentas, a seleção e os sistemas de acasalamentos. A seleção pode ser realizada por meio da escolha dos melhores indivíduos que serão utilizados como pais da próxima geração e para isso, é necessária que haja variabilidade genética, pois o que se busca, é o aumento da frequência dos genes favoráveis na população. No entanto, nem toda variação observada nos animais é herdável, ou seja, de origem genética aditiva. Assim, a seleção de reprodutores por meio de valores genéticos, se torna um processo complexo, necessitando de métodos estatísticos sofisticados para predição dos valores genéticos dos candidatos a reprodutores e estimação dos parâmetros genéticos.

Atualmente, a avaliação genética de animais é baseada na metodologia das equações dos modelos mistos, desenvolvida por Henderson em 1949 e apresentado formalmente em 1973 (HENDERSON, 1973). Esse método consiste basicamente na predição dos valores genéticos, tomados como aleatórios, ajustando-se as observações fenotípicas em relação aos efeitos fixos do modelo que, geralmente, são associados a efeitos ambientais, sendo que esses dados (observações) podem ser balanceados ou desbalanceados. Esse procedimento de predição de valores genéticos foi denominado de "Best Linear Unbiased Predictor" (BLUP - melhor predição linear não viciada). Assim, podem-se obter soluções para os efeitos fixos e simultaneamente para os efeitos aleatórios, ou seja, os BLUP. Contudo, para a utilização desta metodologia é necessário o conhecimento prévio dos componentes de (co)variância. Assim, como normalmente esses componentes são desconhecidos, torna-se necessário estimá-los. Portanto, metodologias têm sido desenvolvidas para tal finalidade, dentre as quais, podemos citar basicamente duas abordagens, a "frequentista" e a "bayesiana".

DESENVOLVIMENTO

Em meio à abordagem "frequentista", podemos destacar três métodos de estimação: os momentos, a função de verossimilhança e as funções quadráticas. Dentre os métodos derivados dos momentos para estimação dos componentes de (co)variânciaestão o método de Fisher (1918), análise de variância (ANOVA), e os métodos I, II e III de Henderson (1953), que em caso de dados balanceados, são análogos ao método ANOVA, nos quais os quadrados médios são igualados às suas respectivas esperanças.Os métodos dos estimadores quadráticos, não-viesado de norma mínima, MINQUE (RAO, 1971a), de variância mínima, MIVQUE (RAO, 1971b) e o iterativo de norma mínima, I-MINQUE (SEARLE, 1987) são exemplos de estimadores sobre funções quadráticas.

Ultimamente, os métodos derivados da função de verossimilhança são os mais utilizados para estimar os componentes de (co)variância. Assim, podemos citar os métodos da máxima verossimilhança – ML (HARTLEY & RAO, 1967), que se baseia na maximização do logaritmo da função densidade de probabilidade das observações, e da máxima verossimilhança restrita – REML (PATTERSON & THOMPSON, 1971). O método da REML considera a maximização da função de verossimilhança independente dos efeitos fixos, além da consideração dos graus de liberdade utilizados na estimação dos efeitos fixos. Contudo, em geral, esses métodos de estimação denominados "frequentistas", consideram muitas aproximações e fortes suposições, as quais se baseiam no teorema do limite central, e se estas suposições forem violadas, o que não é muito difícil, isso poderia gerar estimativas e predições equivocadas, as quais poderiam não ser confiáveis. Assim, quando se utiliza os métodos de estimação "frequentistas" deve-se proceder a

uma análise dos resíduos, que é um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo "frequentistas" com base nos resíduos gerados por este. A ideia básica da análise dos resíduos é que, se o modelo for apropriado, os resíduos devem refletir as propriedades impostas pelo termo de erro do mesmo. Tais suposições são que os resíduos gerados pelo modelo devem atender à homoscedasticidade (variância constante), à independência e à normalidade.

Em vista das limitações dos métodos "frequentistas", mais recentemente, inúmeras publicações (GIANOLA & FOULLEY, 1982; GIANOLA & FERNANDO, 1986; WANG et al., 1994; VAN TASSEL et al., 1995; SCHENKEL et al., 2002) vêm sendo feitas demonstrando a utilização de métodos "bayesianos" como uma poderosa ferramenta para resolução de problemas relacionados à estimação de parâmetros genéticos em melhoramento genético animal.

Inferência Bayesiana

O matemático, físico e astrônomo francês Pierre-Simon de Laplace (1749-1827) contribuiu de forma considerável para o desenvolvimento do "raciocínio bayesiano", uma vez que desenvolveu a primeira teoria da probabilidade. Laplace, em 1812, desenvolveu tambéma teoria analítica da probabilidadepor meio da inserção de um conjunto de técnicas de cálculo (funções geradoras) em um quadro de hipóteses capazes de fornecer dignidade autônoma à própria definição de probabilidade. Laplace tinha o ponto de vista de que a teoria daprobabilidade poderia ser definida como simplesmente o sensocomum reduzido a cálculo. Na prática, Stigler (1986) considera que a "escola bayesiana" foi fundada por Laplace, com a publicação de vários trabalhos entre os anos de 1774 e 1812, os quais tiveram um papel preponderante nas áreas científicas, durante o século XIX. Todavia, segundo Blasco (2001), alguns anos antes do primeiro artigo de Laplace, em 1774, o princípio Bayesiano foi expresso em um artigo científico apresentado na "Royal Society" de Londres, e atribuído a um desconhecido sacerdote, o reverendo Thomas Bayes, que nunca publicou um trabalho na área de matemática durante sua vida. Entretanto, segundo Stigler (1983), o princípio sobre o qual a inferência Bayesiana se baseia foi elaborado antes disso. Stigler (1983) atribui a base da inferência Bayesiana a Saunderson (1683-1739), um professor que possuía deficiência visual e que publicou um grande número de trabalhos em várias áreas da matemática.

Apesar da "teoria bayesiana" estar fundamentada nos trabalhos desenvolvidos nos séculos passados, somente nas últimas décadas surgiram publicações mostrando a sua aplicação na área da genética quantitativa. Ronningen (1971) e Dempfle (1977) discutiram que o BLUP poderia ser interpretado como um "preditor bayesiano", e Harville (1974) ofereceu uma "interpretação bayesiana" da REML. Posteriormente, Gianola & Foulley (1982) introduziram "métodos bayesianos" na análise de características de limiar e, em seguida, Gianola & Fernando (1986) destacaram possibilidades adicionais do uso da "estatística bayesiana"no melhoramento animal em geral.

Ao contrário da estatística clássica, ou seja, a "frequentista", a "inferência bayesiana" leva em conta o conceito de probabilidade de um evento a partir de dois tipos de conhecimentos: 1º o que se sabe sobre tal evento antes que o mesmo se verifique, e 2º eventuais informações que podem ser obtidas na sequência. Isto é, tal conceito de probabilidade de um evento a partir de dois tipos de conhecimentospossibilita calcular a probabilidade de uma hipótese fundamentando-se na probabilidade *a priori* e em eventuais novas evidências relevantes. Na prática, a maior diferença entre as duas estatísticas é que a bayesiana tenta medir o grau de incerteza que se tem sobre a ocorrência de um determinado evento do espaço amostral, utilizando distribuições de probabilidades *a priori* e a amostral (verossimilhança).

A inferência bayesiana secaracteriza por calcular uma função de densidade de probabilidade (densidade *a posteriori*) sobre todos os possíveis vetores de parâmetros (espaço dos parâmetros). Na inferência bayesiana, a incerteza sobre os parâmetros desconhecidos

associa-se uma distribuição de probabilidade (GIANOLA & FERNANDO, 1986), enquanto que, na inferência frequentista, os parâmetros são valores fixos ou constantes, aos quais não se associam a qualquer distribuição (BLASCO, 2001).

No contexto bayesiano, o objetivo é, condicionalmente aos dados observados (y), descrever a incerteza sobre o valor de algum parâmetro (θ, não observado), em termos de probabilidades ou densidades (BOX & TIAO, 1992). O parâmetro pode ser um escalar (θ) ou um vetor de parâmetros ($\theta = \theta_1, \theta_2, \ldots, \theta_p$). Por exemplo, se o parâmetro de interesse é a variância genética aditiva de uma determinada característica $X(\theta_X)$, o objetivo da inferência bayesiana é encontrar a densidade de probabilidade conjunta, f(θ,y) dessa variância em relação aos dados, considerando distribuição do vetor não observável θ_X e o vetor dos dados *y*.Uma vez obtida essa distribuição (posterior densidade de probabilidade conjunta), diferentes tipos de inferências podem ser feitas. Como exemplo,podemos calcular estatísticas descritivas como a probabilidade de θ_X estar entre 0,1 e 0,3, por meio da integração da função entre esses valores, assim como, é possível também obter estimativas do intervalo de 95% de maior densidade *a posteriori* de θ_X, entre outras estimativas.

Segundo Gianola & Fernando (1986), na estatística bayesiana, de forma geral, para se obter a distribuição *a posteriori* de um parâmetro θ, há a necessidade de derivar a distribuição de probabilidade conjunta de θ e y [f(θ , *y*)], a qual pode ser escrita como um produto de duas densidades, a distribuição *a priori*, f(θ), e a distribuição amostral, f(y|θ). Para melhor ilustrar, nas equações [1 e 2], descritas abaixo, pode-se observar a probabilidade de dois eventos acontecerem juntos:

$$P(\theta,y) = P(y|\theta) \bullet P(\theta) \qquad [1]$$

Da mesma forma, temos:

$$P(\theta,y) = P(\theta|y) \bullet P(y) \qquad [2]$$

onde P(θ) e P(y) são as probabilidades marginais de θ e y, respectivamente, e P(θ,y) é a probabilidade conjunta das duas marginais. Utilizando-se a propriedade básica de probabilidades condicionais, conhecida por Regra de Bayes ou Teorema de Bayes, condiciona-se θ ao valor conhecido de y [P(θ|y)]. Desta forma, da equação [2], temos P(θ|y) = P(θ,y) / P(y). Na sequencia, substituímos P(θ,y) pela segunda parte da equação [1], e ficamos com a equação [3], assim:

$$P(\theta|y) = P(y|\theta) \bullet P(\theta) / P(y) \qquad [3]$$

No presente exemplo, vamos escrever estas probabilidades condicionais [3] em termos denominado como função de densidade [4]:

$$f(\theta|y) = f(y|\theta) \bullet f(\theta) / f(y) \qquad [4]$$

onde $f(\theta|y)$ é a função de densidade*a posteriori* a qual incorpora o estado de incerteza do conhecimento prévio a respeito do parâmetro θ após a observação dos dados em *y*; $f(\theta)$ é a função de densidade*a priori* de θ, representando o conhecimento prévio a respeito dos elementos de θ antes da observação dos dados, refletindo a incerteza em relação aos possíveis valores de θ antes do vetor de dados *y* ser compreendido; $f(y|\theta)$ é a função de verossimilhança ou distribuição amostral, representando a contribuição de *y* para o conhecimento sobre θ; e f(*y*) é a função de distribuição marginal dos dados.

Como f(*y*) não é função de θ, então f(θ|*y*) é proporcional apenas ao produto de $f(\theta)$ por $f(y|\theta)$. Dessa forma, tem-se, então, que o Teorema Bayes pode ser utilizado para combinar a informação contida nos dados (verossimilhança) com a probabilidade *a priori*. Essa distribuição *a posteriori* de θ também pode ser denominada densidade *a posteriori*não-normalizada, e é dada por:

$$f(\theta|y) \propto f(y|\theta) \bullet f(\theta) \qquad [5]$$

onde∝ significa: proporcional a.

Na filosofia bayesiana, não há distinção entre estimação de efeitos fixos, predição de efeitos aleatórios, ou estimação de componentes de (co)variância, pois todo parâmetro do modelo é tratado como uma variável aleatória. Dessa forma, para estimar qualquer parâmetro,

tem-se que esse parâmetro é um vetor de quantidades não observáveis, seja ele um efeito fixo, ou aleatório, bem como, um componente de (co)variância. Para maiores detalhes sobre os métodos bayesianos ver, por exemplo, Box & Tiao (1992) e Bernardo & Smith (1994).

Embora os métodos bayesianos sejam teoricamente poderosos, nota-se na equação [5], que para se obter a distribuição *a posteriori* de um parâmetro qualquer (θ, por exemplo), é necessário obter a distribuição *a priori* desse parâmetro, assim como, a distribuição condicional do vetor de observações (y) dado θ. Dessa forma, para qualquer inferência a respeito de θ é necessário integrar a função f(θ|y), que é a distribuição *a posteriori* conjunta dos parâmetros, caso tenha mais de um parâmetro, em relação a todos os outros elementos que a constituem. Contudo, na maioria das vezes, a resolução analítica desta integral é, em geral, impraticável, pois essa operação é muito difícil de ser feita, tanto em análises univariadas como em análises multivariadas com muitos parâmetros a serem estimados, inviabilizando a aplicação da estatística bayesiana, neste caso. Mesmo quando a distribuição condicional conjunta *a posteriori* pode ser obtida analiticamente, em determinadas situações, sua expressão não tem a forma de uma densidade conhecida ou fácil de ser amostrada. As alternativas neste caso são métodos de aproximação numérica, que são baseados em simulação estocástica. Assim, algumas soluções para esse tipo de problema são sugeridas, como por exemplo, a utilização de métodos de Monte Carlo, mais especificamente Cadeias de Markov, que são métodos de simulação, os quais consideram as distribuições condicionais completas *a posteriori* de cada parâmetro para gerar amostras que convergem para a densidade marginal, com o aumento do tamanho dessa amostra (GELFAND & SMITH, 1990; GELFAND, 2000). Ou seja, a ideia básica é simular uma caminhada aleatória no espaço de θ que converge a uma distribuição estacionária, que é a distribuição de interesse do problema. Dentre os métodos de Monte Carlo por meio das Cadeias de Markov (MCMC, "Monte Carlo Markov Chain"), aqueles derivados do algoritmo de Metropolis-Hastings (METROPOLIS et al., 1953; HASTINGS, 1970), como por exemplo o "Amostrador de Gibbs" ou "Gibbs Sampler", têm-se mostrado bastante úteis e eficientes em vários problemas multidimensionais (GELFAND et al., 1990).

Amostrador de Gibbs

O Amostrador de Gibbs, que inicialmente foi utilizado por Geman & Geman (1984) no contexto de restauração de imagens, é um esquema iterativo de amostragem de uma cadeia de Markov, sendo um caso especial do Metropolis-Hastings, onde sempre se aceita a amostra do valor aleatório. Em melhoramento genético animal, o Amostrador de Gibbs vem sendo bastante utilizado no intuito de fornecer amostras aleatórias da distribuição *a posteriori* conjunta ou marginal por meio das distribuições condicionais completas, sem a necessidade de calcular a sua função densidade de probabilidade conjunta e a resolução de integrais.

A ideia básica do Amostrador de Gibbs é tornar uma resolução multivariada em uma sequência de resoluções univariadas, entre os quais se itera para produzir uma cadeia de Markov. No Amostrador de Gibbs a cadeia sempre se move para um novo valor, sendo que as transições de um estado para outro são feitas de acordo com as distribuições condicionais completas para a determinação da distribuição conjunta (CASELLA & GEORGE, 1992).Dada a função de máxima verossimilhança e a densidade *a priori*, calcula-se a densidade conjunta *a posteriori* do parâmetro desconhecido (BESAG et al., 1995). Assim, para obter a distribuição condicional completa de cada parâmetro basta selecionarsomente os termos da distribuição conjunta que dependem desse parâmetro (TANNER, 1996; BESAG et al., 1995), fixando-se as demais variáveis da densidade conjunta (CASELLA & GEORGE, 1992). Para melhor elucidar, o algoritmo pode ser explicado assim:

Supondo-se que a distribuição conjunta*a posteriori* de interesse seja f(θ_i|y), em que θ_i é um elemento do vetor $\theta = \theta_1, \theta_2, \ldots, \theta_p$. Para encontrar as distribuições marginais de θ_i, seria necessário integrar a distribuição condicional conjunta *a posteriori* f(θ|y) em relação aos demais

parâmetros do vetor θ, por exemplo, $f(\theta_i|y) = \iint ... \int f(\theta_1, \theta_2, ... , \theta_p|y)\, d\theta_{-i}$, sendo que $\theta_{-i} = \theta_1, ... , \theta_{i-1}, \theta_{i+1}, ... ,\theta_p$, que é o vetor θ com seu i-ésimo componente removido. Entretanto, o Amostrador de Gibbs gera amostras da densidade conjunta por meio de amostragem sequencial das distribuições condicionais completas.

A equação [6] da densidade condicional completa *a posteriori* de $\theta_1$pode ser descrita da seguinte forma:

$f(\theta_1|\theta_2, ... , \theta_p, y)$ [6]

e assim por diante, para os parâmetros $\theta_2, ... , \theta_p$. Portanto, para estimar as amostras, se começa com valores iniciais arbitrários para os parâmetros$\theta_2^{(1)}, \theta_3^{(1)}, ... , \theta_p^{(1)}$, obtendo-se $\theta_1^{(1)}$ pela simulação de uma variável aleatória da distribuição condicional, $f(\theta_1^{(1)}|\theta_2^{(1)}, \theta_3^{(1)}, ... , \theta_p^{(1)}, y)$. Note que $\theta_1^{(1)}$ significa o primeiro componente, na primeira iteração. Sequencialmente, o Amostrador de Gibbs usa $\theta_1^{(1)}, \theta_3^{(1)}, \theta_4^{(1)}, ...,\theta_p^{(1)}$para gerar um novo valor de $\theta_2^{(1)}$, usando a distribuição condicional, $f(\theta_2^{(1)}|\theta_1^{(1)}, \theta_3^{(1)}, \theta_4^{(1)}, ... ,\theta_p^{(1)}, y)$. Dessa forma, segue-se amostrando cada um dos valores do vetor θ (parâmetro$\theta_1^{(1)}, \theta_2^{(1)}, ... , \theta_p^{(1)}$), até completar-se a primeira iteração do Amostrador de Gibbs, que consiste em simular um parâmetro, condicionalmente aos valores dos outros parâmetros, até amostrar todos os parâmetros.

Terminada a primeira iteração, o Amostrador de Gibbs usa estes parâmetros amostrados ($\theta_2^{(1)}, \theta_3^{(1)}, ... ,\theta_p^{(1)}$) para gerar um novo valor de $\theta_1^{(2)}$, ou seja, um novo valor de $\theta_1^{(2)}$ em uma segunda iteração, usando a distribuição condicional, $f(\theta_1^{(2)}|\theta_2^{(1)}, ... , \theta_p^{(1)}, y)$, e assim por diante, até completar-se a segunda iteração do Amostrador de Gibbs.

Repetindo-se esse processo k vezes, pode-se demonstrar que, após um grande número de iterações, a sequência de valores gerados pelo Amostrador de Gibbs converge para uma distribuição estacionária igual a $f(\theta|y)$, sendo que cada valor de θ obtido pelo Amostrador de Gibbs após convergência, é um valor simulado da distribuição conjunta de seus elementos (GEMAN & GEMAN, 1984).

A partir do momento em que se alcança a convergência, a cadeia permanece nesta distribuição e "passeia" para sempre no seu novo "subespaço" (TIERNEY, 1994). Assim, a sequência de valores gerados das amostras aleatórias, $\theta_1, \theta_2, ... , \theta_p$, desse novo "subespaço", extraídas das condicionais, se aproxima da distribuição de equilíbrio, ou seja, são equivalentes às amostras aleatórias da distribuição conjunta e/ou marginal *a posteriori* (CASELLA & GEORGE, 1992). Dessa forma, pode-se calcular aproximações da estimativa da média, desvio-padrão, moda, entre outras, sem a necessidade de se resolver integrais, sendo que, tanto melhor será essa estimação, quanto maior for o número de amostras utilizadas.

Observa-se que o Amostrador de Gibbs possibilita a amostragem de uma densidade de probabilidade multivariada, utilizando apenas densidades de "subconjuntos" de vetores condicionais em todos os outros. A atratividade desta abordagem é que provê uma solução para o problema bayesiano de integração múltipla quando as densidades condicionais são identificáveis, de forma conhecida e passíveis de amostragem.Vale lembrar que, por se tratar de um processo de Markov, amostras sucessivas são dependentes entre si, assim, aconselha-se descartar algumas iterações intermediárias para se obter amostras independentes. Naturalmente, por se tratar de uma simulação com valores iniciais arbitrários, recomenda-se descartar as amostras iniciais (que teoricamente, são amostras da cadeia que ainda não convergiram), o que é denominado de "burn-in", sendo que esse descarte, obviamente está relacionado com a velocidade de convergência da cadeia de Gibbs.

Existem dificuldades em detectar a convergência da cadeia de Gibbs, assim, vários testes, os quais verificam a estacionalidade da cadeia, o descarte inicial, o tamanho efetivo de amostra, entre outros, são descritos na literatura (GELMAN & RUBIN, 1992; GEYER, 1992; RAFTERY & LEWIS 1992). Contudo, infelizmente, estes testes não são perfeitos (CASELLA & GEORGE, 1992).

Apesar dos testes de convergência apresentarem algumas desvantagens, uma grande vantagem da estatística bayesiana é a utilização de informações adicionais, ou seja, informações prévias do que se estuda. Essas informações podem ser utilizadas para a construção de uma distribuição *a priori* a qual, juntamente com a verossimilhança f($y|\theta$) (exemplo da equação 5), fornece uma estimativa média entre o conhecimento prévio (*a priori*) e as informações que os próprios dados nos oferecem (verossimilhança). O conhecimento *a priori* é mais importante quando as informações disponíveis são escassas ou pouco informativas. Quando se tem grande volume de dados (por exemplo, grande número de progênies por reprodutor) as informações *a priori*tendem a ser dominadas pela função de verossimilhança. Para maiores detalhes sobre o Amostrador de Gibbsver, por exemplo, Casella & George (1992), Tierney (1994), Besag et al. (1995) e Tanner (1996). Vale ressaltar que, quando a distribuição condicional completa de interesse, f($\theta_i|\theta_{-i}$, y), não tem a forma de uma densidade conhecida, não sendo possível gerar valores diretamente dessa distribuição, uma alternativa é utilizar o algoritmo de Metropolis-Hastings que é um método MCMC definido a partir dos trabalhos de Metropolis et al. (1953) e Hastings (1970).

Metropolis-Hastings

O algoritmo Metropolis-Hastings é um método de simulação estocástica que permite amostrar de qualquer função de densidade f(θ), sem que seja necessário conhecer as distribuições condicionais (CHIB & GREENBERG, 1995). Por exemplo, se na equação [6], a distribuição condicional completa de interesse, f($\theta_1|\theta_2$, ... , θ_p, y) não tem a forma de uma densidade conhecida, não sendo possível gerar valores diretamente dessa distribuição, o algoritmo de Metropolis-Hastings se faz uma opção. O método consiste em gerar valores candidatos θ_1^* de uma densidade auxiliar f(θ_1, θ_1^*) que possa ser amostrada. Esta densidade auxiliar deve possuir duas condições: 1º permitir a geração de amostras aleatórias (θ_1^*) de forma fácil e rápida e, 2º deve ser definida no mesmo domínio que a densidade alvo f(θ_1). Estes valores candidatos que foram simulados a partir desta densidade auxiliar são aceitos ou não, dependendo de certaprobabilidade. De acordo com Chib & Greenberg (1995), várias são as famílias dedensidades auxiliares f(θ_1, θ_1^*) que podem ser escolhidas para gerar valores candidatos θ_1^* (HASTINGS, 1970).

IMPLICAÇÕES PRÁTICAS

Ultimamente, devido aos avanços computacionais, ao conhecimento da obtenção da distribuição condicional conjunta *a posteriori*por meio de simulações e a compreensão dos diversos algoritmos bayesianos aplicados à produção animal, a inferência bayesiana tem sido amplamente aplicada em diversas outras áreas do melhoramento genético animal, como por exemplo: análises de QTL – "Quantitative Trait Loci" (GREEN, 1995; SATAGOPAN et al., 1996; UIMARI & HOESCHELE, 1997), análise de sobrevivência (DUCROQ & CASELLA, 1996), modelos lineares ou não lineares de hierarquia (VARONA et al., 1997), regressão aleatória (JAMROZIK & SCHAEFFER, 1997), modelos lineares generalizados (TEMPELMAN, 1998), curvas de crescimento linear e não linear (MIGNON-GASTREAU et al., 2000; FORNI et al., 2009), seleção genômica (MEUWISSEN et al., 2001), construção de mapas genéticos (ROSA et al., 2002), dentre muitos outros.Estes estudos demonstram a importância desta metodologia no progresso das técnicas de avaliações genéticas.Um exemplo prático é na avaliação de características de expressão categórica, onde a inferência bayesiana se demonstra uma excelente alternativa, principalmente quando a distribuição destes dados não tem a forma normal.Além disso, com a criação de computadores mais potentes podem-se modelarpor meio da metodologia *bayesiana*, de forma mais clara, dados com outros tipos de distribuições, diferentes da normal, como a distribuição de *Poisson*, binária, entre outras.

Outra implicação prática desta metodologia *Bayesiana* se dá em avaliações genéticas em escala genômica. Geralmente, nos arquivos de dados de campo existem muitos erros de informações de pedigree, os quais prejudicam as avaliações genéticas, causando prejuízos no progresso genético dos rebanhos. Uma alternativa para a construção da matriz de parentesco aditiva de pedigrees é usar as informações de marcadores moleculares em escala genômica para inferir parentescos. Assim, a metodologia *bayesiana*, se faz uma importante ferramenta na estimação de valores genéticos genômicos, principalmente pela utilização de modelos hierárquicos (GOMPERT & BUERKLE, 2011), onde se permite juntar diversas informações de maneiras hierárquicas ou não, e amostrar parâmetros da distribuição condicional conjunta *a posteriori* por meio de simulações.

CONCLUSÕES E PERSPECTIVAS

A metodologia abordada neste trabalho se faz uma excelente alternativa aos programas de melhoramento animal, na estimação de parâmetros genéticos e predição de valores genéticos em ordem de solucionar problemas relacionados aos modelos mais complexos e a expressão de características de interesse econômico que não tenham distribuição normal, uma vez que as suposições podem ser atendidas por meio de simulações e amostragem da distribuição condicional conjunta *a posteriori*.

REFERÊNCIAS BIBLIOGRÁFICAS

BERNARDO, J.M.; SMITH, A.F.M. **Bayesian theory**. John Wiley & Sons, Chichester, U.K. 1994.

BESAG, J.; GREEN, P.; HIGDON, D.; MENGERSEN, K. Bayesian computation and stochastic systems. **Statistical Science**, v.10, n.1, p.03-66, 1995.

BLASCO, A.The Bayesian controversy in animal breeding. **Journal of Animal Science**, v.79, p.2023-2046. 2001.

BOX, G.E.P.; TIAO, G.C. **Bayesian Inference in Statistical Analysis**. New York: J. Wiley-Interscience, 1992. 588p.

CASELLA, G.; GEORGE, E.I. Explaining the Gibbs Sampler. **The American Statistician**, v.46, n.3, p.167-174, 1992.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings Algorithm. **The American Statistician**, v.49, n.4, p.327-335, 1995.

DEMPFLE, L. Relation entre BLUP (Best Linear Unbiased Prediction) etestimateursbayesiens. **Genetic, Selection, Evolution**, v.9, p.27-32, 1977.

DUCROQ, V.; CASELLA, G.A. Bayesian analysis of mixed survival models. **Genetic, Selection, Evolution**, v.28, p.505-529, 1996.

FORNI, S.; PILES, M.; BLASCO, A.; VARONA L.; OLIVEIRA, H.N.; LÔBO, R.B.; ALBUQUERQUE, L.G. Comparison of different nonlinear functions to describe Nelore cattle growth. **Journal of Animal Science***,* v.87, p.496-506, 2009.

GELFAND, A.E.; SMITH, A.F.M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association***,* v.85, n.410, p.348-409, 1990.
GELFAND, A.E. Gibbs sampling. **Journal of the American Statistical Association**, v.95, n.452, p.1300-1304, 2000.

GELFAND, A.E.; HILLS, S.E.; RACINE-POON, A.; SMITH, A.F.M. Illustration of Bayesian inference in normal data models using Gibbs sampling. **Journal of the American Statistical Association**, v.85, n.41, p.972-985, 1990.

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

GEMAN, S.; GEMAN. D. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. **IEE Transactions on Pattern Analysis and Machine Intelligence**, v.6, p.721–741, 1984.

GEYER, C.J. Practical Markov chain Monte Carlo (with discussion). **Statistical Science**, v.7, p.473-511, 1992.

GIANOLA, D.; FERNANDO, R.L. Bayesian methods in animal breeding theory. **Journal of Animal Science**, v.63, p.217-244, 1986.

GIANOLA, D.; FOLLEY, J.L. Non linear prediction of latent genetic liability with binary expression: An empirical Bayes approach. In: WORLD CONGRESS OF GENETIC APPLIED TO LIVESTOCK PRODUCTION, 2., Madri, Espanha, 1982. **Proceedings**...Madri, v.7, p.293-303. 1982.

GOMPERT, Z.; BUERKLE, C.A.A Hierarchical Bayesian Model for Next-Generation Population Genomics. **Genetics**, v.187, p.903-917, 2011.

GREEN, P.J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, v.82, n.4, p. 711-732, 1995.

HARTLEY, H.O.; RAO, J.N.K. Maximum likelihood estimation for the mixed analysis of variance model. **Biometrika**, v.54, p.93-108, 1967.

HARVILLE, D.A. Bayesian inferences for variance components using only error contrasts. **Biometrika**, v.61, p.383-385, 1974.

HASTINGS, W.K. Monte Carlo sampling methods using Markov Chains and their applications. **Biometrika**, v.57, p.97-109, 1970.

HENDERSON, C.R. Estimation of variance and covariance components. **Biometrics**, v.17, p.226-252, 1953.

HENDERSON, C.R. Sire evaluation and genetic trends. In: ANIMAL BREEDING AND GENETIC SYMPOSIUM IN HONOR OF DR. JAY L. LUSH, 1., Champaign, IL, EUA, 1973. **Proceedings…** Champaign, IL: ASAS, p.10-41. 1973.

JAMROZIK, J.; SCHAEFFER, L.R. Estimates of genetic parameters for a test day model with random regressions for yield of first lactation Holsteins. **Journal of Dairy Science**, v.80, p.726-70, 1997.

METROPOLIS, N.; ROSENBLUTH, A.W.; ROSENBLUTH, M.N.; TELLER, A.; TELLER, H. Equations of state calculations by fast computing machines. **Journal of Chemical Physics**, v.21, p.1087-1091, 1953.

MEUWISSEN, T.H.E.; GODDARD, M.E.; HAYES, B.J. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, p.1819-1829, 2001.

MIGNON-GRASTEAU, S.; PILES, M.; VARONA, L.; ROCHAMBEAU, H.; POIVEY, J.P.; BLASCO, A.; BEAUMONT, C. Genetic analysis of growth curve parameters for male and female chickens resulting from selection based on juvenile and adult body weights simultaneously. **Journal of Animal Science**, v.78, p.2515-2524, 2000.

PATTERSON, H.D.; THOMPSON, R. Recovery of inter-block information when block size are unequal. **Biometrics**, v.58, p.545-554, 1971.

RAFTERY, A.E.; LEWIS, S.M. Comment: One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. **Statistical Science**, v.7, p.493-497, 1992.

RAO, C.R. Estimation of variance and covariance components – MINQUE theory. **Journal of Multivariate Analysis**, v.1, p.257-275, 1971a.

RAO, C.R. Minimum variance quadratic unbiased estimation of variance components. **Journal of Multivariate Analysis**, v.1, p.445-456, 1971b.

RONNINGEN, K. Some properties of the selection index derived by "Henderson's mixed model method". **ZeitschriftfürTierzüchtung und Züchtungsbiologie**, v.88, p.186-193, 1971.

ROSA, G.J.M.; YANDELL, B.S.; GIANOLA, D.A Bayesian approach for constructing genetic maps when markers are miscoded. **Genetic, Selection, Evolution**, v.34, p.353-369, 2002.

SATAGOPAN, J.M.; YANDELL, B.S.; NEWTON, M.A.; OSBORN, T.C.A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo. **Genetics**, v.144, p.805-816, 1996.

SCHENKEL, F.S.; SCHAEFFER, L.R.; BOETTCHER, P.J. Comparison between estimation of breeding values and fixed effects using Bayesian and empirical BLUP estimation under selection on parents and missing pedigree information. **Genetic, Selection, Evolution**, v.34, p.41-59, 2002.

SEARLE, S.R. **Linear models for unbalanced data**. New York: John Wiley & Sons, 1987. 536p.

STIGLER, S.M. **The History of Statistics: The Measurement of Uncertainty before 1900**. Harvard University Press, Cambridge, MA. 1986.

STIGLER, S.M. Who discovered Bayes's theorem? **American Statistician**, v.37, p.290-296, 1983.

TANNER, M.A. **Tools for statistical inference**, 3ed. Springer-Verlag, New York. 1996.

TEMPELMAN, R.J. Generalized Linear Mixed Models in Dairy Cattle Breeding. **Journal of Dairy Science**, v.81, p.1428-1444, 1998.

TIERNEY, L. Markov chains for exploring posterior distributions. **The Annals of Statistics**, v.22, p.1701-1762, 1994.

UIMARI, P.; HOESCHELE, I. Mapping-linked quantitative trait loci using Bayesian analysis and Markov Chain Monte Carlo algorithms. **Genetics**, v.146, p.735-743, 1997.

VAN TASSEL, C.P.; CASELLA, G.; POLLAK, E.J. Effects of selection on estimates of variance components using Gibbs sampling and restricted maximum likelihood. **Journal of Dairy Science**, v.78, p.678-692, 1995.

VARONA, L., MORENO, C., GARCIA-CORTE´S, L.A., ALTARRIBA, J. Multiple trait genetic analysis of underlying biological variables of production functions. **Livestock Production Science**, v.47, p.201-209, 1997.

WANG, C.S.; GIANOLA, D.; SORENSEN, D.A.; JENSEN, J.; CHRSTENSEN, A.; RUTHLETDGE, J.J. Response to Selection for Letter Size in Danish Landrace Pigs: A Bayesian Analysis. **Theory Applied Genetics**, v.88, p.220-230, 1994.