VIII Simpósio Brasileiro de Melhoramento Animal
Maringá, PR – 01 e 02 de julho de 2010
Melhoramento Animal no Brasil: UMA VISÃO CRÍTICA

Busca por estruturas causais no contexto de modelos mistos multivariados em genética quantitativa: exemplo simulado

Bruno Dourado Valente[12], Guilherme Jordão de Magalhães Rosa[23], Gustavo de los Campos[4], Daniel Gianola[234], Martinho de Almeida e Silva[5]

[1]Doutorando do Programa de Pós-Graduação em Zootecnia – UFMG/Belo Horizonte. Bolsista da CAPES e CNPq (doutorado sandwich). e-mail: bvalente66@yahoo.com
[2]Department of Dairy Science,
[3]Department of Biostatistics and Medical Informatics, e
[4]Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin USA
[5]Departamento de Zootecnia – UFMG/Belo Horizonte

Resumo: O objetivo deste trabalho é ilustrar, por meio de um exemplo simulado, a utilização do algoritmo IC para a realização de busca por estruturas causais recursivas no contexto de modelos mistos em genética quantitativa, no qual efeitos genéticos correlacionados são possíveis fontes de confundimento. Para isso, o modelo multicaracterísticas padrão é ajustado utilizando-se métodos Bayesianos para obtenção da distribuição *a posteriori* da matriz de covariância dos fenótipos, condicionalmente aos efeitos genéticos aditivos não observáveis. Esta é utilizada como informação de entrada no algoritmo IC. A metodologia proposta é aplicada a dados simulados para múltiplas características mensuradas em um conjunto de linhagens endogâmicas.

Palavras–chave: busca por estruturas causais, confundimento genético, metodologia, modelos de equações estruturais, modelos mistos, sistemas biológicos

Searching for recursive causal structures in multivariate quantitative genetics mixed models: simulation study

Abstract: The goal of this article is to illustrate, with a simulated example, the application of the IC algorithm to search for recursive causal structures in a quantitative genetics mixed models scenario, where correlated genetic effects may act as confounders. The methodology first fits a standard multiple trait model using Bayesian methods to obtain the posterior distribution of the covariance matrix of phenotypes conditional to unobservable additive genetic effects, which is then used as input for the IC algorithm. The proposed method was applied to simulated data related to multiple traits measured on a set of inbred lines.

Keywords: causal structure search, genetic confounders, methodology, mixed models, structural equation models, systems biology

Introdução

Modelos de Equações Estruturais ou MEEs (Wright, 1921) são utilizados para estudar cenários nos quais diferentes características apresentam relação de recursividade ou *feedback* entre si. Tais modelos foram adaptados para o contexto de modelos mistos por Gianola e Sorensen (2004). O ajuste de MEEs exige escolher, dentro de um espaço tipicamente vasto de hipóteses causais, uma estrutura causal que descreve a relação entre as características estudadas. Em um trabalho apresentado neste Simpósio, Valente et al. (2010a) discutem metodologia para, com base na matriz de covariância entre as características condicionalmente aos efeitos genéticos e no algoritmo IC (*Inductive Causation*; Pearl, 2000) buscar por estruturas causais recursivas no contexto de modelo mistos em genética quantitativa. O objetivo deste trabalho é ilustrar a aplicação desta metodologia por meio de um exemplo simulado.

Material e Métodos

Observações para 1800 indivíduos foram amostradas de um MEE de acordo com estrutura causal utilizada em Shipley (1997). Adicionalmente, as características sofrem influência de efeitos genéticos aditivos correlacionados, como demonstrado na Figura 1. Considerou-se que os indivíduos pertenciam a

300 linhagens endogâmicas (seis indivíduos em cada linhagem), para as quais efeitos genéticos específicos foram amostrados. O MEE do qual os dados foram simulados pode ser representado como:

$$\begin{cases} y_{i1k} = \mu_1 + u_{1k} + e_{i1k} \\ y_{i2k} = \mu_2 + \lambda_{21} y_{i1k} + u_{2k} + e_{i2k} \\ y_{i3k} = \mu_3 + \lambda_{32} y_{i2k} + u_{3k} + e_{i3k} \\ y_{i4k} = \mu_4 + \lambda_{42} y_{i2k} + u_{4k} + e_{i4k} \\ y_{i5k} = \mu_5 + \lambda_{53} y_{i3k} + \lambda_{54} y_{i4k} + u_{5k} + e_{i5k} \end{cases},$$

na qual y_{ijk} e e_{ijk} são fenótipo e resíduo para a característica j (j = 1, ..., 5), atribuídos ao indivíduo i que pertence à linha endogâmica k, μ_j é a média da característica j, $\lambda_{jj'}$ é a modificação na característica j com respeito a j` e u_{jk} é o efeito genético aditivo da linha endogâmica k para a característica j. Os efeitos genéticos aditivos correlacionados das 300 linhagens foram simulados como pertencendo a 50 grupos não aparentados de seis irmão completos. Por sua vez, os resíduos de diferentes características foram amostrados independentemente. Valores dos parâmetros de local e de dispersão utilizados na simulação são fornecidos em Valente et al. (2010b).

Subseqüentemente, a metodologia proposta para busca de estrutura causal recursiva foi empregada. Amostras *a posteriori* da matriz $Var(\mathbf{y}_i \mid \mathbf{u}_i) = \mathbf{R}_0$ foram obtidas pelo ajuste de modelo multicaracterísticas via análise Bayesiana e amostragem Gibbs. Estas amostras foram utilizadas para obtenção da distribuição das correlações parciais necessárias para decisões estatísticas do algoritmo IC.

Resultados e Discussão

O gráfico semidirecionado obtido pelo algoritmo IC é apresentado na Figura 2. Algumas linhas entre variáveis não foram direcionadas, uma vez que cada uma delas pode apresentar qualquer direção sem necessariamente resultar em um ciclo ou em um *collider*. Desta forma, o gráfico semidirecionado descrito representa um conjunto de estruturas causais acíclicas que não podem ser discriminadas com base nas observações, mas que constitui geralmente em importante restrição do espaço de estruturas possíveis. No presente exemplo, a estrutura utilizada na simulação dos dados faz parte da classe de estruturas selecionada pelo algoritmo IC. Informações *a priori* podem tornar o resultado da busca ainda mais específico. Como exemplo, a informação hipotética de que a variável y_1 precede y_2 temporalmente levaria a direcionar a linha entre as duas variáveis em direção a y_2, o que por sua vez direcionaria as linhas restantes, resultando na estrutura utilizada na simulação.

Existem cerca de 59000 possíveis estruturas causais recursivas que podem ser utilizadas no ajuste de MEE para o estudo das 5 características simuladas (Shipley, 1997). A estratégia utilizada na aplicação recente de MEEs é a escolha de uma ou um pequeno conjunto de estruturas com base em informação *a priori* a respeito do relacionamento causal entre as características estudadas. A metodologia proposta permite explorar de maneira mais apropriada o espaço de estruturas causais. A utilização de $Var(\mathbf{y}_i \mid \mathbf{u}_i) = \mathbf{R}_0$ como entrada para o algoritmo IC permite realizar a busca em cenários nos quais correlações genéticas podem confundir esta busca.

Na simulação descrita não ocorreram erros nas decisões estatísticas, mas o risco de erros aumenta em situações em que a força das relações causais é menor. Adicionalmente, a qualidade das decisões estatísticas é menor em situações menos informativas, uma vez que as distribuições das correlações parciais se tornam menos precisas.

O esforço computacional necessário para se implementar esta metodologia aumenta com o número de características avaliadas, uma vez que o número de pares de características estudadas pelo algoritmo IC se torna maior, bem como o número de possíveis subconjuntos de características remanescentes para cada par. Outro efeito do aumento do número de características estudadas é tornar mais lento o ajuste de modelos multicaracterísticas. O número de observações utilizadas na análise também influencia o tempo necessário para ajuste deste modelo. Finalmente, o tamanho da cadeia que representa a distribuição de $\mathbf{R}_0$ também é importante na definição da carga de trabalho computacional necessária, não só por influenciar o tempo para obtenção da cadeia, mas também porque correlações parciais utilizadas nas decisões estatísticas do algoritmo IC são calculadas para cada amostra da cadeia.

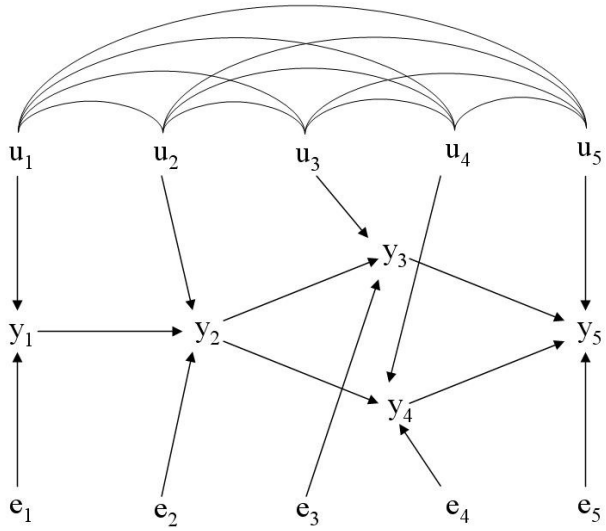

Figura 1 Estrutura causal do MEE utilizado para simular dados: y_j é observação da característica j, u_j é o efeito genético aditivo da característica j e e_j é o resíduo do modelos associado a y_j. Arcos conectando u`s representam correlações. A estrutura é adaptada de Shipley (1997).

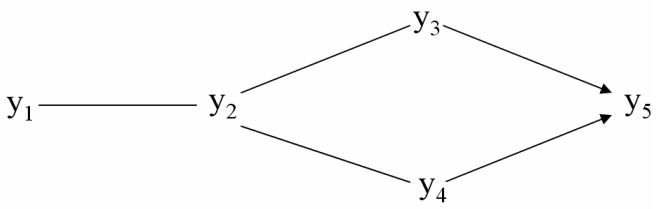

Figura 2 Gráfico semi-direcionado obtido pelo algoritmo IC.

Literatura citada

GIANOLA, D.; SORENSEN D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. **Genetics**, v.167**,** p.1407-1424, 2004.

PEARL, J. **Causality: Models, Reasoning and Inference**. Cambridge University Press, Cambridge, UK, 384p, 2000.

SHIPLEY, B. Exploratory path analysis with applications in ecology and evolution. **The American Naturalist**, v.149, p.1113-1138, 1997.

VALENTE, B.D.; ROSA, G.J.M.; de los CAMPOS, G.; GIANOLA, D.; SILVA, M.A. Busca por estruturas causais no contexto de modelos mistos multivariados em genética quantitativa: metodologia, in: VIII SIMPÓSIO BARSILEIRO DE MELHORAMENTO ANIMAL, 2010a, Maringá, PR. **Anais...** Maringá, PR, 2010a.

VALENTE, B.D.; ROSA, G.J.M.; de los CAMPOS, G.; GIANOLA, D.; SILVA, M.A. Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models, **Genetics**, (aceito para publicação), doi:10.1534/genetics.109.112979, 2010b.

WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, v.201**,** p.557–585, 1921.