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: metodologia

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.br
[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: Um grande número de estruturas causais recursivas pode ser utilizado para o ajuste de Modelos de Equações Estruturais (MEEs). Em aplicações recentes de MEEs no contexto de modelos mistos em genética quantitativa, estruturas causais foram selecionadas com base somente em conhecimento *a priori* a respeito do sistema biológico estudado. Desta forma, o vasto espaço de possíveis estruturas causais não tem sido explorado adequadamente. Como alternativa, algoritmos de busca podem ser utilizados para selecionar estruturas que são compatíveis com a densidade de probabilidade conjunta das variáveis estudadas. Entretanto, no contexto de genética quantitativa, tais algoritmos não podem ser utilizados diretamente sobre a densidade conjunta dos fenótipos, uma vez que covariâncias genéticas atuam como fonte de confundimento. Neste trabalho, propomos a busca por estruturas causais recursivas entre fenótipos utilizando o algoritmo IC (*Inductive Causation*), aplicado à distribuição conjunta dos fenótipos condicionalmente aos efeitos genéticos aditivos.

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: methodology

Abstract: The number of different recursive causal structures that can be used for fitting a structural equation model (SEM) to multivariate data can be huge. In recent applications of SEM in mixed model quantitative genetics settings, causal structures were pre-selected based on prior biological knowledge alone. Therefore, the wide range of possible causal structures has not been properly explored. Alternatively, causal structure spaces can be explored using algorithms which can search for structures that are compatible with the joint distribution of the traits. However, the search cannot be performed directly on the joint distribution of the phenotypes as it is possibly confounded by genetic covariances. In this paper we propose to search for recursive causal structures among phenotypes by applying the Inductive Causation (IC) algorithm to the joint distribution of phenotypes conditionally on the additive genetic effects.

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

Introdução

Sistemas biológicos apresentam interações complexas entre fenótipos, tais como relações de feedback ou de recursividade entre substratos e enzimas em sistemas bioquímicos. Modelos de Equações Estruturais ou MEEs (Wright (1921); Haavelmo (1943)) são utilizados para estudar cenários nos quais diferentes características apresentam tais relações. Estes 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 o relacionamento entre as características estudadas. Tal escolha pode ser realizada, por exemplo, utilizando-se informação *a priori* sobre o sistema em questão (estratégia utilizada em aplicações recentes dos MEEs). Alternativamete, o

espaço de estruturas causais pode ser explorado por meio de algoritmos de busca por estruturas capazes de gerar a densidade conjunta das características estudadas. Um obstáculo para utilização de tais algoritmos no contexto de modelos mistos é a presença de covariâncias genéticas que confundem a busca. O objetivo deste trabalho é propor metodologia que permita a busca por estruturas causais recursivas na presença de confundimento genético.

Material e Métodos

Segundo Gianola e Sorensen (2004), MEEs no contexto de modelos mistos podem ser representados por $\mathbf{y}_i = \mathbf{\Lambda y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i$. Esta expressão representa um modelo semelhante ao modelo multicaracterísticas clássico para representação de um vetor de fenótipos do indivíduo *i*, com termo adicional $\mathbf{\Lambda y}_i$. Neste termo, $\mathbf{\Lambda}$ é uma matriz com zeros na diagonal e coeficientes estruturais ou zeros nos elementos fora da diagonal principal, de acordo com a estrutura causal entre características.

Uma estrutura causal recursiva pode ser representada por um gráfico direcionado acíclico (GDA), composto por um conjunto de variáveis conectadas por setas, cujos sentidos representam direções dos relacionamentos causais entre pares de variáveis. Em uma sequência de variáveis conectadas (ou trilha), variáveis que não estão no extremo desta trilha permitem fluxo de dependência, a não ser que setas desta trilha apresentem convergência nesta variável (como *C* em $A \rightarrow C \leftarrow B$). Neste caso, a variável *C* é denominada *collider* e bloqueia o fluxo de dependência entre *A* e *B*. Condicionalmente a variáveis que não estão nos extremos da trilha, a capacidade de bloquear ou permitir fluxo de dependência se inverte. Em um GDA, duas variáveis são ditas d-separadas condicionalmente a um grupo **S** de variáveis remanescentes se, condicionalmente a **S**, não existe trilha que permita fluxo de dependências entre estas duas variáveis. Sob algumas premissas, d-separações na estrutura causal de um MEE impõem independências condicionais na densidade conjunta das variáveis estudadas, o que pode ser explorado para reconstrução de estruturas causais recursivas capazes de gerar tal densidade (Pearl (2000); Spirtes et al. (2000)). Com base em uma matriz de correlações entre as variáveis estudadas, o algorimo IC (*Inductive Causation*; Pearl, 2000) fornece ao usuário uma estrutura causal (ou uma classe de estruturas causais equivalentes) compatível com aquela matriz. Para um conjunto **V** de variáveis aleatórias, o algoritmo IC consiste nos seguintes passos:

1 – Para cada par de variáveis *A* e *B* em **V**, procure por um conjunto de variáveis $\mathbf{S}_{AB}$ de modo que *A* seja independente de *B* condicionalmente a $\mathbf{S}_{AB}$. Se *A* e *B* são dependentes condicionalmente a qualquer um dos possíveis grupos de variáveis remanescentes, conecte *A* e *B* com uma linha não-direcionada. Esta etapa do algoritmo tem como resultado o gráfico não direcionado ***U***.

2 – Para cada par de variáveis não adjacentes *A* e *B* com uma variável adjacente em comum *C* em ***U*** (i.e., $A - C - B$), procure por um conjunto de variáveis $\mathbf{S}_{AB}$ que contém *C* de modo que *A* seja independente de *B* dado $\mathbf{S}_{AB}$. Se tal conjunto não existe, oriente as linhas da estrutura estudada em direção a *C* ($A \rightarrow C \leftarrow B$). Caso o conjunto exista, continue.

3 – No gráfico parcialmente direcionado resultante da etapa anterior, oriente ao máximo as linhas restantes, de maneira que não apareçam ciclos ou *colliders* além daqueles previamente identificados.

A conexão entre estrutura causal e densidade conjunta depende de uma premissa de resíduos independentes. Esta premissa implica em considerar estrutura diagonal para a matriz de covariância residual, o que é considerada em aplicações recentes de MEE em genética quantitativa, que também consideram a estrutura causal como conhecida *a priori*.

No contexto de modelos mistos o algoritmo descrito não pode ser aplicado diretamente sobre a densidade conjunta das características observadas, uma vez que, mesmo considerando resíduos independentes, existem efeitos genéticos não observados que são causa extra de covariância fenotípica. Considere os exemplos hipotéticos ilustrados na Figura 1, em que há uma estrutura causal recursiva entre y_1, y_2 e y_3; e_1, e_2 e e_2 são resíduos independentes e u_1, u_2 e u_3 são efeitos genéticos correlacionados. A conexão entre densidade conjunta e estrutura causal se perde na presença de efeitos genéticos correlacionados e não controlados. Na Figura 1a, y_1 e y_2 não são independentes devido à correlação entre

u_1 e u_2. Pelo mesmo motivo, y_1 e y_2 não são independentes condicionalmente a y_3 na Figura 1b. Entretanto, o relacionamento genético aditivo entre indivíduos permite "controlar" os efeitos genéticos e se colocar condicionalmente a eles, o que recupera a conexão entre estrutura causal e densidade conjunta. Condicionalmente aos efeitos genéticos, y_1 e y_2 são independentes marginalmente na Figura 1a e condicionalmente a y_3 na Figura 1b. Desta forma, $Var(\mathbf{y}_i \mid \mathbf{u}_i) = \mathbf{R}_0$ pode ser utilizada para seleção de estruturas causais no contexto de modelos mistos. Estimativas desta matriz podem ser obtidas pela matriz de variância residual de um modelo multicaracterísticas simples. A seguir, propomos metodologia para busca de estruturas causais no contexto de modelos mistos, utilizando análise Bayesiana, o que permite obter amostras *a posteriori* de $\mathbf{R}_0$ para decisões estatísticas:

1 – Ajustar modelos multicaracterísticas e obter amostra da distributição *a posteriori* de $\mathbf{R}_0$.

2 – Aplicar algoritmo IC utilizando a amostra de $\mathbf{R}_0$ para as decisões estatísticas. Especificamente, para cada pergunta envolvendo a independência das variáveis *A* e *B* condicionalmente a um conjunto de variáveis **S** e, implicitamente, aos efeitos genéticos:

2.1 – Obter distribuição *a posteriori* da correlação parcial $\rho_{A,B|\mathbf{S}}$. Esta é uma função de $\mathbf{R}_0$ e sua distribuição pode ser obtida calculando a correlação para cada amostra de $\mathbf{R}_0$.

2.2 – Computar o intervalo HPD (*Highest Posterior Density*) 95% para a distribuição *a posteriori* de $\rho_{A,B|\mathbf{S}}$.

2.3 – Se o intervalo mencionado contém o valor 0, declarar $\rho_{A,B|\mathbf{S}}$ como nulo. Caso contrário, declarar *A* e *B* como condicionalmente dependentes.

3 – Ajustar MEE utilizando a estrutura causal selecionada (ou uma das estruturas causais equivalentes da classe selecionada).

Mais detalhes a respeito da metodologia proposta são fornecidos em Valente et al. (2010).

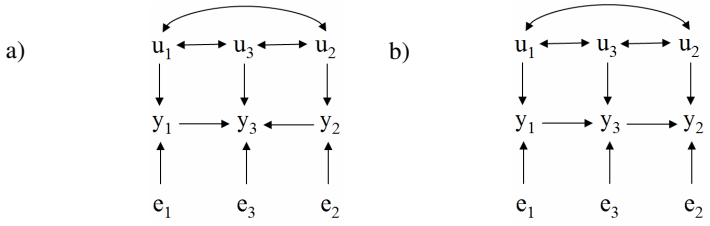

Figura 1 Estruturas causais para três variáveis observadas (y_1, y_2 e y_3), com resíduos independentes (e_1, e_2 e e_3) e efeitos genéticos aditivos correlacionados (u_1, u_2 e u_3).

Literatura citada

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

HAAVELMO, T., The statistical implications of a system of simultaneous equations. **Econometrica**, v.11**,** p.1–12, 1943.

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

SPIRTES, P.; GLYMOUR, C.; SCHEINES, R. **Causation, Prediction and Search**. 2. ed. MIT Press, Cambridge, MA, 2000. 543p.

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, 2010.

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