BRUNO DOURADO VALENTE

BUSCA POR ESTRUTURAS CAUSAIS NO CONTEXTO DE MODELOS MISTOS EM GENÉTICA QUANTITATIVA

Tese apresentada ao Programa de Pós-Graduação em Zootecnia da Escola de Veterinária da Universidade Federal de Minas Gerais como requisito parcial para obtenção do grau de Doutor em Zootecnia.

Área de concentração: Genética e Melhoramento Animal

Orientador: Martinho de Almeida e Silva

Belo Horizonte
UFMG – Escola de Veterinária
2010

Aplica teu coração à disciplina e teus ouvidos às palavras do conhecimento. (*Provérbios 23:12*)

Dedico este trabalho à minha esposa e aos
meus pais e irmãos, pelo que representam

Agradeço...

A Deus, pois "nEle vivemos, nos movemos e existimos" (*Atos 17:28*), de modo que não há conquistas fora dEle.

Aos meus pais e irmãos, pois nada realiza aquele que não nasceu e não recebeu proteção e incentivo, mas não há quem seja capaz de provê-los por si próprio.

A Ana Angélica, pois todo trabalho se torna mais fácil quando a maior fonte de motivação está para além daquele que trabalha.

Ao prof. Martinho, pela orientação, por tudo ensinado, pelo incentivo, pelo exemplo de produtividade, integridade e mérito.

Ao prof. Guilherme Rosa, por me aceitar como orientado na University of Wisconsin, pelos ensinamentos e disponibilidade; e por tornar, sob muitos aspectos, esta tese possível.

A Gustavo de los Campos e ao prof Daniel Gianola, pelas indispensáveis contribuições para a presente tese.

Aos meus "irmãos científicos" do curso de Pós-Graduação em Zootecnia - UFMG, especialmente Gal, Luciana, Raphael Rocha, Ricardo e Vívian; e a todos aqueles que passaram pelo grupo de estudos sobre Genética e Melhoramento Animal.

Aos meus queridos vizinhos e companheiros de Madison - WI, especialmente Annaiza, Fabyano e Sandra, que estiveram próximos do início ao fim.

Às instituições Escola de Veterinária – UFMG, University of Wisconsin - Madison, CAPES e CNPq, pela oportunidade de realização deste estudo.

A todos aqueles que de alguma forma contribuíram com esta tese.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Modelos de equações estruturais (MEEs) podem ser utilizados para estudar relacionamentos de recursividade e *feedback* em análise multivariada. O número de estruturas causais recursivas distintas que podem ser utilizadas para ajustar tais modelos pode ser muito grande, mesmo no estudo de um pequeno conjunto de características. Em aplicações recentes de MEEs no contexto de modelos mistos em genética quantitativa, as estruturas causais foram pré-selecionadas utilizando apenas conhecimento biológico *a priori*. Desta forma, a ampla gama de estruturas causais possíveis não tem sido adequadamente explorada. Como alternativa, espaços de estruturas causais podem ser explorados por meio de algoritmos que, orientados por evidências nos dados, podem buscar por estruturas causais que são compatíveis com a distribuição conjunta das variáveis estudadas. Entretanto, a busca não pode ser realizada diretamente na distribuição conjunta dos fenótipos, uma vez que esta se apresenta potencialmente confundida por covariâncias genéticas. Na presente tese, o autor propõe buscar por estruturas causais recursivas entre fenótipos utilizando o algoritmo IC após ajustar os dados para efeitos genéticos. Para isso, ajusta-se um modelo multicaracterísticas padrão para se obter a matriz de covariância fenotípica condicionalmente aos efeitos genéticos não-observáveis, que é posteriormente submetida ao algoritmo IC.

Palavras-chave: busca por estrutura causal, confundimento genético, modelos de equações estruturais, modelos mistos, sistemas biológicos

ABSTRACT

Structural Equation Models (SEM) can be used to study recursive and simultaneous relationships in multivariate analyses. Nonetheless, the number of different recursive causal structures that can be used for fitting a SEM to multivariate data can be huge, even when only a few traits are considered. 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, using data driven evidence, can search for structures that are compatible with the joint distribution of the variables under study. However, the search cannot be performed directly on the joint distribution of the phenotypes as it is possibly confounded by genetic covariance among traits. In this thesis, we propose to search for recursive causal structures among phenotypes using the IC algorithm after adjusting the data for genetic effects. A standard multiple trait model is fitted to obtain a posterior covariance matrix of phenotypes conditional to unobservable additive genetic effects, which is then used as input for the IC algorithm.

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

CAPÍTULO 1

1 INTRODUÇÃO

Número cada vez maior de características vem sendo incorporado ao critério de seleção utilizado por programas de melhoramento de animais de diferentes espécies. Como exemplo que se aplica a diversas espécies, características não-produtivas vêm apresentando crescente importância na definição do critério de seleção, o que causa impacto positivo na saúde, na eficiência reprodutiva e no bem estar dos animais. Como consequência, cresce o número de características a serem estudadas no processo de avaliação genética. Todas estas características apresentam componentes genéticos e ambientais que influem em sua expressão fenotípica e no relacionamento entre elas.

Existem diferentes estratégias para analisar características múltiplas. Cada uma delas se diferencia quanto à qualidade de predição, utilização das informações disponíveis e adaptação a sistemas multivariados complexos. Uma metodologia possível na avaliação de um conjunto de características é a utilização de modelo animal unicaracterística para cada uma das características do conjunto. A maior vantagem apresentada por esta estratégia é a facilidade de implementação. Em comparação com as demais metodologias, o número de parâmetros envolvidos no sistema de equações é menor. Esta vantagem é importante, principalmente quando se tem grande número de animais e/ou de características envolvidas na avaliação (LASSEN et al., 2007). Porém, esta abordagem não considera covariâncias entre características, o que significa que as predições são feitas para cada característica ignorando-se as informações das demais.

A utilização de modelos multicaracterísticas é alternativa para avaliar duas ou mais características. Esta alternativa é extensão do modelo utilizado para a obtenção de BLUP's de valores genéticos para uma característica (HENDERSON, 1976). Este modelo permite a avaliação simultânea de duas ou mais características, considerando as covariâncias genéticas e residuais entre elas. A utilização desta abordagem pode causar impacto positivo na qualidade de predição. Modelos multicaracterísticas permitem que as informações de uma característica sejam consideradas na avaliação de outra característica, o que aumenta a acurácia das avaliações (WIGGANS e GODDARD, 1997). Mesmo pequeno aumento na acurácia das predições em razão da incorporação de informações de outras características pode resultar em grande efeito econômico na avaliação genética de grandes populações (POLLAK et al., 1984).

Modelos multicaracterísticas consideram que toda fonte de associação entre características é representada por associações lineares simétricas entre seus respectivos efeitos aleatórios (e.g. covariância genética aditiva direta e covariância residual). Modelos de equações estruturais (MEEs), por sua vez, permitem representar cenários complexos nos quais uma característica pode apresentar efeito causal sobre outra. Este tipo de relacionamento entre características é comum em sistemas biológicos, mas não pode ser expresso por modelos multicaracterísticas clássicos. MEEs foram desenvolvidos por WRIGHT (1921) e HAAVELMO (1943), e adaptados para o contexto de modelos mistos em genética quantitativa por GIANOLA e SORENSEN (2004). Estes autores reconhecem um dos maiores desafios para a aplicação de MEEs na representação de um conjunto de características: a escolha de uma única estrutura causal em um espaço tipicamente vasto de estruturas causais. Em aplicações recentes deste modelo, os autores utilizam crenças *a priori* na escolha das estruturas causais, o que representa exploração inadequada do vasto espaço de estruturas.

Por outro lado, existem diversos algoritmos que permitem selecionar estruturas causais para um MEE, explorando o conceito de *d-separação*. No entanto, a utilização direta destes algoritmos no contexto de modelos mistos, no qual efeitos correlacionados não-observados são considerados, resulta em confundimento na busca proposta. O objetivo do trabalho aqui apresentado consiste em propor uma metodologia que permita selecionar estruturas causais no contexto de modelos mistos em genética quantitativa. A presente tese se estrutura da seguinte maneira: o **CAPÍTULO 1** apresenta uma revisão de literatura a respeito de MEEs e algoritmos de busca por estruturas causais; o **CAPÍTULO 2** corresponde à tradução de um artigo publicado (VALENTE et al., 2010) no qual são apresentadas a metodologia mencionada, a aplicação da metodologia em dados simulados e a

discussão dos resultados. Em seguida, são apresentadas as **CONSIDERAÇÕES FINAIS**. Ao final da tese, apresentam-se, na seção **APÊNDICE**, detalhes a respeito do amostrador de Gibbs utilizado para obter a distribuição *a posteriori* dos parâmetros dos MEEs apresentados no **CAPÍTULO 2**.

O presente capítulo se dedica à revisão de literatura e se divide em dois tópicos principais: **MODELOS DE EQUAÇÕES ESTRUTURAIS** e **BUSCA POR ESTRUTURAS CAUSAIS**. O primeiro tópico se estrutura da seguinte maneira: a seção **Modelos de equações estruturais (MEEs)** descreve os modelos em questão de maneira generalizada e para o contexto de modelos mistos, além de ilustrar como a redução de um MEE o transforma em um modelo multicaracterísticas equivalente; **Interpretação dos parâmetros** demonstra modificações na interpretação paramétrica de acordo com o modelo utilizado (MEE ou modelo multicaracterísticas); **Implementação** ilustra as etapas necessárias no ajuste de MEEs, e **Aplicações** descreve estudos recentes utilizando modelos mistos em genética quantitativa nos quais diferentes extensões dos MEEs foram aplicadas. No segundo tópico, a seção **Terminologia** expõe alguns conceitos necessários para a apresentação das metodologias desenvolvidas para seleção de estruturas causais, como utilizados por aqueles que desenvolveram tais metodogias (PEARL, 2000; SPIRTES et al., 2000). As bases da conexão entre estruturas causais e distribuições conjuntas são descritas em **Propriedades de Markov**. A seção **Minimalidade e estabilidade (*faithfulness*)** apresenta critérios para preferência de estruturas utilizados na construção dos algoritmos de busca. As premissas das metodologias apresentadas são discutidas em **Premissas**. A seção **Algoritmo IC** descreve, discute e exemplifica a implementação de um destes algoritmos. Finalmente, a seção **Considerações** conclui esta revisão.

2 REVISÃO DE LITERATURA

2.1 Modelos de equações estruturais

Em um programa de melhoramento genético, o objetivo de seleção geralmente envolve diversas características correlacionadas. O critério para seleção de indivíduos a serem utilizados como reprodutores na próxima geração é geralmente uma função de méritos genéticos preditos para várias características de interesse econômico. Desta forma, modelos multicaracterísticas têm grande aplicação na área de Melhoramento Animal.

No contexto de tais modelos, a correlação entre características é tipicamente representada por associações lineares simétricas entre efeitos aleatórios considerados para cada característica, como efeito genético aditivo direto ou efeitos de ambiente permanente e temporário. Estas associações são representadas por componentes de covariância (VARONA et al., 2007). Uma alternativa para representar um conjunto de características são os modelos de equações estruturais (WRIGHT, 1921 e HAAVELMO, 1943), ou MEEs. Tais modelos constituem uma extensão do modelo multicaracterísticas padrão, na qual uma característica pode ser considerada como função de outras que pertencem ao conjunto de características estudadas, o que proporciona a representação de uma rede funcional entre elas. O propósito do desenvolvimento de tais modelos foi combinar informações qualitativas de causa e efeito com informações provenientes de dados para fornecer uma estimativa quantitativa da relação de causa e efeito entre as variáveis de interesse (PEARL, 2000).

Apesar de Sewall Wright, um dos pesquisadores mais influentes da história da genética quantitativa, ser o pioneiro da modelagem de equações estruturais por meio da análise de trilha, este tipo de modelo foi ignorado por pesquisadores em biologia no século XX (SHIPLEY, 2002). Em outras áreas, tais como economia e ciências sociais (GOLDBERGER, 1972; DUNCAN, 1975), os MEEs ganharam importância. Na área de genética quantitativa, pouca atenção foi dedicada a estes modelos antes de GIANOLA e SORENSEN (2004).

Modelos de equações estruturais permitem representar relações de recursividade (efeito de uma variável resposta em outra) e de *feedback* (ou simultaneidade) entre variáveis resposta. Tais relações são mais complexas do que aquelas representadas em modelos multicaracterísticas, porém são

comuns em sistemas biológicos (GIANOLA e SORENSEN, 2004). Um exemplo clássico de efeitos recursivos e de *feedback*, proposto por TURNER e STEVENS (1959), é apresentado na FIGURA 1.1. O cenário descrito envolve três variáveis: concentração de CO_2 no ar (A), concentração de CO_2 nos pulmões (B) e profundidade do movimento de respiração (C). Como demonstrado na figura, A influencia B, B influencia C, que por sua vez influencia B. A intensidade da modificação que uma variável causa em outra é ponderada por coeficientes λ. Ignorando a biologia real da situação, considere que λ_{BA} e λ_{CB} são positivos. O aumento da valor de A causa aumento do valor de B, que por sua vez causa aumento no valor de C. Dependendo do sinal de λ_{BC}, o aumento do valor de C pode levar à queda ou ao aumento do valor de B, o que por sua vez influencia C novamente. A relação entre B e C tende a um equilíbrio ou ao colapso, e modificações nos valores destas duas variáveis não causam modificações no valor de A (GIANOLA e SORENSEN, 2004). Desta forma, a complexa relação entre estas três variáveis não pode ser adequadamente representada por associações lineares simétricas como componentes de covariância em modelos multicaracterísticas clássicos.

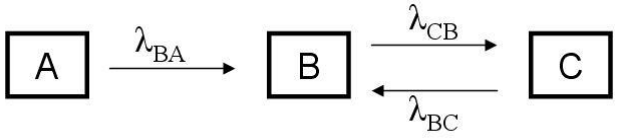


FIGURA 1.1 – Modelo de Turner e Stevens (1959) para descrever o relacionamento entre concentração de CO_2 no ar (A), concentração de CO_2 nos pulmões (B) e profundidade do movimento de respiração (C). O coeficiente λ_{ij} pondera a modificação do valor da variável *i* relativa ao valor da variável *j* (adaptado de GIANOLA e SORENSEN, 2004).

2.1.1 Modelos de equações estruturais (MEEs)

Os MEEs podem ser representados em sua forma geral como um sistema de equações em que cada equação é representada por:

$$y_j = f\left(\mathbf{y}_{pj}, e_j\right) \qquad [1]$$

em que y_j é a variável dependente da equação, y_{pj} são variáveis dentre aquelas consideradas como "variáveis dependentes" do modelo, com índices diferentes de *j*, que influenciam y_j (denominados "pais" de y_j) e e_j representa o resíduo aleatório associado a y_j (PEARL, 2000). O modelo [1] é uma generalização não-linear e não-paramétrica de MEEs lineares:

$$y_j = \sum_{k \in pj} \lambda_{jk} y_k + e_j$$

em que *pj* compreende o conjunto de "variáveis dependentes" que são pais de y_j, e λ_{jk} é o coeficiente estrutural, e corresponde à modificação de valor esperada em y_j com respeito à variável y_k (PEARL, 2000).

Observa-se que para a representação de MEEs, é necessário definir *a priori*, para cada variável resposta *j* do conjunto estudado, quais das variáveis remanescentes serão consideradas como pais de *j*. Esta estrutura de associações causais pode ser representada por um gráfico contendo variáveis conectadas por setas. Este gráfico direcionado é denominado estrutura causal (PEARL, 2000). Considere como exemplo um MEE simples representando a associação entre três variáveis: y_1, y_2 e y_3. Ao definir que y_1 influencia y_2 recursivamente, e que y_2 influencia y_3 de maneira semelhante, a

estrutura causal recursiva acíclica descrita pode ser representada graficamente por $y_1 \rightarrow y_2 \rightarrow y_3$, e o sistema de equações pode ser representado por:

$$y_1 = e_1$$
$$y_2 = \lambda_{21} y_1 + e_2$$
$$y_3 = \lambda_{32} y_2 + e_3$$

GIANOLA e SORENSEN (2004) apresentaram os MEEs sob o contexto de genética quantitativa. Segundo estes autores, um sistema de equações de duas características distintas, cujas observações do indivíduo *i* são representadas por y_{i1} e y_{i2}, pode ser descrito da seguinte forma:

$$y_{i1} = \lambda_{12} y_{i2} + \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + u_{i1} + e_{i1}$$
$$y_{i2} = \lambda_{21} y_{i1} + \mathbf{x}'_{i2}\boldsymbol{\beta}_2 + u_{i2} + e_{i2}$$

em que $\boldsymbol{\beta}_1$ ($\boldsymbol{\beta}_2$) são vetores de efeitos fixos para a característica 1 (2), $\mathbf{x}'_{i1}$ ($\mathbf{x}'_{i2}$) é o vetor conhecido de incidência dos efeitos fixos $\boldsymbol{\beta}_1$ ($\boldsymbol{\beta}_2$) na observação, u_{i1} (u_{i2}) e e_{i1} (e_{i2}) são, respectivamente, efeitos genéticos aditivos e resíduos do modelo. O parâmetro λ_{12} é a mudança no valor de y_{i1} em função do valor de y_{i2} e λ_{21} é a mudança no valor de y_{i2} em função do valor de y_{i1}. No caso descrito, se λ_{12} e λ_{21} são diferentes de zero, existe *feedback* entre as características. Nesta situação, cada uma das duas características influencia diretamente a outra e, indiretamente, a si própria. Já a recursividade ocorre quando apenas um dos dois coeficientes é diferente de zero. O modelo também poderia ser descrito, omitindo-se os efeitos fixos, como no diagrama da FIGURA 1.2.

Considerando $\mathbf{y}_i$ como o vetor de observações de *t* diferentes características para o animal *i*, MEEs podem ser representados como:

$$\mathbf{y}_i = \boldsymbol{\Lambda}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i \qquad [2]$$

em que $\mathbf{u}_i$ e $\mathbf{e}_i$ são vetores de valores genéticos aditivos e resíduos atribuídos a $\mathbf{y}_i$, $\boldsymbol{\beta}$ é o vetor que contém efeitos fixos para todas as características, $\mathbf{X}_i$ é a matriz de incidência dos efeitos contidos em $\boldsymbol{\beta}$ no vetor $\mathbf{y}_i$ e $\boldsymbol{\Lambda}$ é uma matriz quadrada com ordem *t* que contém os coeficientes estruturais nas entradas fora da diagonal principal. A estrutura causal definida *a priori* indica quais entradas de $\boldsymbol{\Lambda}$ são consideradas como parâmetros livres e quais são obrigatoriamente iguais a zero.

A seguinte distribuição conjunta é considerada para $\mathbf{u}_i$ e $\mathbf{e}_i$:

$$\begin{bmatrix} \mathbf{u}_i \\ \mathbf{e}_i \end{bmatrix} \sim N\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G}_0 & 0 \\ 0 & \boldsymbol{\Psi}_0 \end{bmatrix} \right\},$$

em que $\mathbf{G}_0$ e $\boldsymbol{\Psi}_0$ são, respectivamente, matrizes de covariância genética aditiva direta e residual.

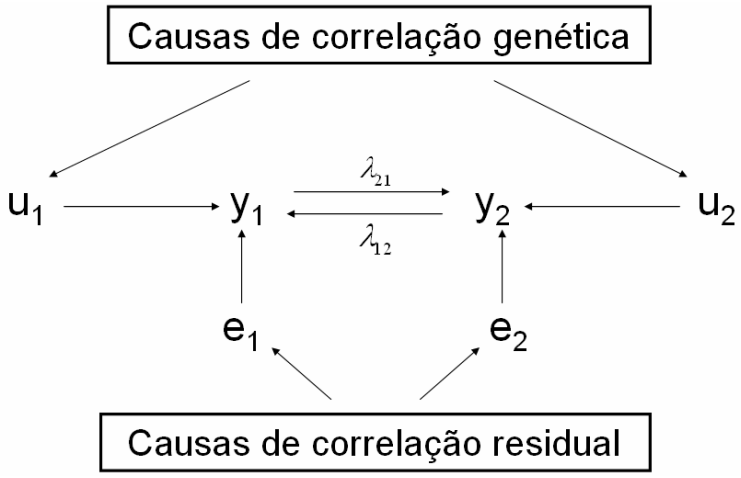


FIGURA 1.2 – Modelo considerando simultaneidade entre características: y_1 e y_2 são observações das características 1 e 2, u_1 e u_2 são efeitos genéticos, e_1 e e_2 são resíduos atribuídos às características 1 e 2. Seta unidirecional indica que a variável da base da seta influencia a variável da ponta da seta. Setas opostas representam simultaneidade entre variáveis. O parâmetro λ_{ij} é a mudança na variável *i* relativa ao valor da variável *j* (adaptado de GIANOLA e SORENSEN (2004)).

Com base no MEE [2], deriva-se o seguinte "modelo reduzido":

$$\left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i$$
$$\mathbf{y}_i = \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{X}_i\boldsymbol{\beta} + \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{u}_i + \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{e}_i . \qquad [3]$$

A distribuição do vetor $\mathbf{y}_i$ condicional aos parâmetros de local $\boldsymbol{\beta}$, $\mathbf{u}_i$ e $\boldsymbol{\Lambda}$ e à matriz de covariância residual $\boldsymbol{\Psi}_0$ é representada por:

$$\mathbf{y}_i \mid \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Psi}_0 \sim \mathbf{N}\left[\left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i\right), \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Psi}_0\left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{\prime -1}\right].$$

O modelo para *n* indivíduos é descrito por

$$\mathbf{y} = \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad [4]$$
$$\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim N\left\{\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{G}_0 \otimes \mathbf{A} & 0\\ 0 & \boldsymbol{\Psi}_0 \otimes \mathbf{I}_n\end{bmatrix}\right\}$$

em que $\mathbf{y}$, $\mathbf{u}$ e $\mathbf{e}$ são vetores de observações, efeitos genéticos aditivos e resíduos do modelo ordenados por característica e indivíduo dentro de característica, enquanto $\mathbf{X}$ e $\mathbf{Z}$ são matrizes de incidência dos efeitos em $\boldsymbol{\beta}$ e $\mathbf{u}$ no vetor $\mathbf{y}$. O modelo [4] pode ser reescrito como:

$$\left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

de modo que o modelo reduzido se torna

$$\mathbf{y} = \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{X}\boldsymbol{\beta} + \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{Z}\mathbf{u} + \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{e}.$$

Por consequência,

$$p(\mathbf{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Psi}_0) \sim N\left\{\left[\mathbf{I}_{tn} - (\boldsymbol{\Lambda} \otimes \mathbf{I}_n)\right]^{-1} (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}), \right.$$
$$\left. \left[\mathbf{I}_{tn} - (\boldsymbol{\Lambda} \otimes \mathbf{I}_n)\right]^{-1} \boldsymbol{\Psi} \left[\mathbf{I}_{tn} - (\boldsymbol{\Lambda} \otimes \mathbf{I}_n)\right]^{'-1} \right\},$$

em que $\boldsymbol{\Psi} = \boldsymbol{\Psi}_0 \otimes \mathbf{I}_n$.

Considerando [3], observa-se que no modelo reduzido, o sistema é resolvido para as "variáveis resposta". Desta forma, o modelo resultante desta transformação corresponde ao modelo multicaracterísticas padrão:

$$\mathbf{y}_i = (\mathbf{I}_t - \boldsymbol{\Lambda})^{-1} \mathbf{X}_i \boldsymbol{\beta} + (\mathbf{I}_t - \boldsymbol{\Lambda})^{-1} \mathbf{u}_i + (\mathbf{I}_t - \boldsymbol{\Lambda})^{-1} \mathbf{e}_i$$
$$\mathbf{y}_i = \boldsymbol{\mu}_1^* + \mathbf{u}_i^* + \mathbf{e}_i^*$$

em que $\boldsymbol{\mu}_l^*$, $\mathbf{u}_i^*$ e $\mathbf{e}_i^*$ são respectivamente vetores de efeitos fixos, efeitos genéticos aditivos e resíduos de um modelo que não considera elos funcionais entre variáveis resposta. Adicionalmente, pode-se representar a distribuição conjunta dos efeitos aleatórios do modelo multicaracterísticas padrão como:

$$\begin{bmatrix} \mathbf{u}_i^* \\ \mathbf{e}_i^* \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_0^* & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_0^* \end{bmatrix} \right)$$
$$\mathbf{G}_0^* = (\mathbf{I}_t - \boldsymbol{\Lambda})^{-1} \mathbf{G}_0 (\mathbf{I}_t - \boldsymbol{\Lambda})^{'-1}$$
$$\mathbf{R}_0^* = (\mathbf{I}_t - \boldsymbol{\Lambda})^{-1} \boldsymbol{\Psi}_0 (\mathbf{I}_t - \boldsymbol{\Lambda})^{'-1}$$

Desta forma, MEEs podem ser descritos como reparametrizações do modelo multicaracterísticas simples. As duas formas apresentadas são equivalentes, uma vez que geram a mesma distribuição para as variáveis resposta.

2.1.2 Interpretação dos parâmetros

Parâmetros de locação como efeitos genéticos aditivos e parâmetros de dispersão como componentes de variância e covariância genética são considerados tanto no modelo de equações estruturais quanto no modelo multicaracterísticas padrão. Porém, a interpretação paramétrica se modifica de acordo com o modelo. Considere, como exemplo, dados provenientes de um modelo amostral bicaracterística recursivo, no qual uma característica *A* tem efeito causal na característica *B*. Sob modelo recursivo, a correlação genética representa a associação linear entre dois efeitos genéticos aditivos não-observáveis, cada um afetando diretamente uma característica específica. Porém, esta não seria a única fonte de correlação genética caso o modelo multicaracterísticas fosse utilizado para representar o sistema multivariado, uma vez que existe no modelo amostral utilizado uma associação indireta entre o efeito genético de *A* e o fenótipo *B*, mediada pelo fenótipo *A* (o efeito genético de *A* tem influência causal no fenótipo de *A*, que por sua vez é pai de *B* na estrutura causal descrita). A correlação genética sob modelo recursivo não considera este efeito indireto. No entanto, a correlação genética no modelo multicaracterísticas representa toda a associação de origem genética, independentemente de esta ser considerada direta ou indireta no contexto recursivo. Por este motivo,

torna-se concebível que a correlação genética sob modelo multicaracterísticas seja diferente de zero mesmo se os valores genéticos aditivos forem independentes no contexto recursivo.

Consequências da existência de relação recursiva ou simultânea entre fenótipos foram apresentadas por GIANOLA e SORENSEN (2004) e ilustram a diferença de interpretação dos parâmetros sob os dois diferentes modelos mencionados no parágrafo anterior. Considere como exemplo simples um modelo recursivo para duas características:

$$y_{i1} = \mathbf{x}_i'\boldsymbol{\beta}_1 + u_{i1} + e_{i1}$$
$$y_{i2} = \lambda_{21} y_{i1} + \mathbf{x}_i'\boldsymbol{\beta}_2 + u_{i2} + e_{i2},$$

de modo que

$$y_{i2} = \lambda_{21}\left(\mathbf{x}_i'\boldsymbol{\beta}_1 + u_{i1} + e_{i1}\right) + \mathbf{x}_i'\boldsymbol{\beta}_2 + u_{i2} + e_{i2}$$
$$y_{i2} = \mathbf{x}_i'\left(\boldsymbol{\beta}_2 + \lambda_{21}\boldsymbol{\beta}_1\right) + \left(u_{i2} + \lambda_{21}u_{i1}\right) + \left(e_{i2} + \lambda_{21}e_{i1}\right)$$
$$y_{i2} = \mathbf{x}_i'\boldsymbol{\beta}_2^* + u_{i2}^* + e_{i2}^*$$

Observa-se que u_{i2} representa o efeito genético da combinação linear $y_{i2} - \lambda_{21} y_{i1}$. Adicionalmente, observa-se que u_{i1} só poderia ser interpretado como o valor genético aditivo que influencia apenas y_{i1} se o coeficiente estrutural λ_{21} fosse igual a zero. As variâncias dos efeitos aleatórios na primeira equação não se modificam após a redução do sistema de equações. Por outro lado, as variâncias destes efeitos para a segunda equação podem ser representadas como:

$$\text{var}\left(u_{i2}^*\right) = \sigma_{u2}^2 + \lambda_{21}^2\sigma_{u1}^2 + 2\lambda_{21}\sigma_{u1,u2}$$
$$\text{var}\left(e_{i2}^*\right) = \sigma_{e2}^2 + \lambda_{21}^2\sigma_{e1}^2 + 2\lambda_{21}\sigma_{e1,e2}$$

Sob modelo multicaracterísticas padrão, a covariância genética se torna:

$$\text{cov}\left(u_{i1}, u_{i2}^*\right) = \text{cov}\left(u_{i1}, u_{i2} + \lambda_{21}u_{i1}\right) = \sigma_{u1,u2} + \lambda_{21}\sigma_{u1}^2,$$

de modo que a correlação genética pode ser construída como

$$cor\left(u_{i1}, u_{i2}^*\right) = \frac{\text{cov}\left(u_{i1}, u_{i2}^*\right)}{\sqrt{\text{var}\left(u_{i1}\right)\text{var}\left(u_{i2}^*\right)}} = \frac{\sigma_{u1,u2} + \lambda_{21}\sigma_{u1}^2}{\sqrt{\sigma_{u1}^2\left(\sigma_{u2}^2 + \lambda_{21}^2\sigma_{u1}^2 + 2\lambda_{21}\sigma_{u1,u2}\right)}}.$$

Se $\sigma_{u1,u2} = 0$,

$$cor\left(u_{i1}, u_{i2}^*\right) = \frac{\lambda_{21}\sigma_{u1}^2}{\sqrt{\sigma_{u1}^2\left(\sigma_{u2}^2 + \lambda_{21}^2\sigma_{u1}^2\right)}} = \frac{\lambda_{21}}{\sqrt{\left(\frac{\sigma_{u2}^2}{\sigma_{u1}^2} + \lambda_{21}^2\right)}},$$

de modo que a correlação genética sob modelo multicaracterísticas padrão é diferente de zero, com sinal determinado pelo coeficiente estrutural λ_{21}. Este resultado confirma que se torna concebível que a correlação genética sob modelo multicaracterísticas seja diferente de zero mesmo se os valores genéticos aditivos forem independentes no contexto recursivo, e ilustra a diferença de interpretação dos parâmetros de cada modelo.

Para cenários que apresentam relacionamentos mais complexos (como *feedback*) entre pares de características, ou que apresentam maior número de características, as consequências na

interpretação paramétrica são análogas às descritas para o modelo recursivo acima, mas suas representações algébricas se tornam mais complexas (GIANOLA e SORENSEN, 2004).

2.1.3 Implementação

2.1.3.1 Seleção da estrutura causal

Para ajustar MEEs a observações que pertencem a um conjunto de características, é necessário definir *a priori* a estrutura causal entre as variáveis estudadas. Como mencionado anteriormente, esta estrutura pode ser representada por um gráfico direcionado ou pela simples definição de quais serão as entradas consideradas como parâmetros livres na matriz $\mathbf{\Lambda}$. Considerando que o valor dos elementos fora da diagonal podem variar livremente ou serem definidos como 0, existem potencialmente $t(t-1)$ coeficientes estruturais e $2^{t(t-1)}$ estruturas que podem representar o relacionamento causal entre t características.

A seleção da estrutura causal a ser utilizada se torna um desafio ao ajuste de um MEE, devido ao aumento explosivo do número de possíveis estruturas na medida em que cresce o número de características estudadas, como demonstrado na TABELA 1.1. O grande número de hipóteses causais ocorre mesmo em situações restritas e com pequeno conjunto de características. Por exemplo, para uma estrutura recursiva acíclica entre apenas cinco características, existem cerca de 59000 alternativas (SHIPLEY, 1997).

Desta forma, a comparação exaustiva se torna impossível pela utilização de critérios como o AIC (AKAIKE, 1974) ou o BIC (SCHWARTZ, 1978). Nas aplicações recentes de MEEs no contexto de modelo misto em genética quantitativa, utilizaram-se crenças *a priori* a respeito da estrutura causal do sistema estudado para que fossem selecionadas uma estrutura ou um pequeno grupo de estruturas, estas comparadas por critérios semelhantes aos supracitados (DE LOS CAMPOS et al., 2006a,b; WU et al., 2007; KONIG et al., 2008; DE MATURANA et al., 2009).

TABELA 1.1 – Número de possíveis estruturas causais diferentes que podem ser propostas para t variáveis resposta

t	Número de estruturas
2	4
3	64
4	4096
5	1048576
6	1073741824

2.1.3.2 Identificabilidade dos parâmetros

Como demonstrado, o modelo reduzido [3] é equivalente ao MEE [2], uma vez que ambos produzem a mesma densidade de probabilidade:

$$N\left(\left(\mathbf{I}_t-\mathbf{\Lambda}\right)^{-1}\mathbf{X}_i\boldsymbol{\beta}+\left(\mathbf{I}_t-\mathbf{\Lambda}\right)^{-1}\mathbf{u}_i,\left(\mathbf{I}_t-\mathbf{\Lambda}\right)^{-1}\mathbf{\Psi}_0\left(\mathbf{I}_t-\mathbf{\Lambda}\right)^{'-1}\right)=N\left(\boldsymbol{\mu}_t^*+\mathbf{u}_i^*,\mathbf{R}_0^*\right). \qquad [5]$$

Porém, é evidente que os dois modelos não apresentam uma correspondência entre os seus parâmetros do tipo "um a um". MEEs apresentam coeficientes estruturais (presentes na matriz $\mathbf{\Lambda}$), além dos parâmetros de locação e de dispersão análogos àqueles do modelo reduzido.

O modelo multicaracterísticas padrão é identificável, uma vez que modificações nos valores dos parâmetros resultariam necessariamente em modificações na densidade de probabilidade dos dados por ele gerados. Como consequência, a inferência com base na função de verossimilhança torna-se possível para todos os parâmetros do modelo. Entretanto, a existência de parâmetros adicionais nos MEEs faz com que este modelo seja sub-identificável na função de verossimilhança, uma vez que mais de uma combinação de valores de parâmetros resultam em uma mesma densidade de probabilidade ([5] é válida para infinitas combinações de valores de parâmetros no lado esquerdo da igualdade). Existem várias maneiras de verificar a identificabilidade dos parâmetros de um modelo. Na análise Bayesiana de MEEs, é possível obter evidências empíricas da não-identificabilidade de parâmetros pela observação do comportamento da cadeia de Markov que representa amostras da distribuição *a posteriori* dos parâmetros de modelo. WU et al. (2009) simularam dados por meio de um modelo recursivo bivariado com resíduos independentes. Posteriormente, os autores tentaram utilizar métodos de MCMC (*Markov Chain Monte Carlo*) para simular a distribuição *a posteriori* dos parâmetros, permitindo porém amostragem livre para todas as entradas da matriz de covariâcia residual. Como demonstrado na FIGURA 1.3, os comportamentos das amostras da covariânca residual e do coeficiente estrutural apresentaram forte concordância, algo típico para modelos com parâmetros não-identificáveis.

Como consequência da sub-identificabilidade do MEE apresentado, torna-se necesário aplicar restrições aos parâmetros do modelo para realizar inferências a respeito destes. Como exemplo, uma restrição suficiente para um modelo recursivo acíclico é considerar as covariâncias residuais iguais a 0 (VARONA et al., 2007).

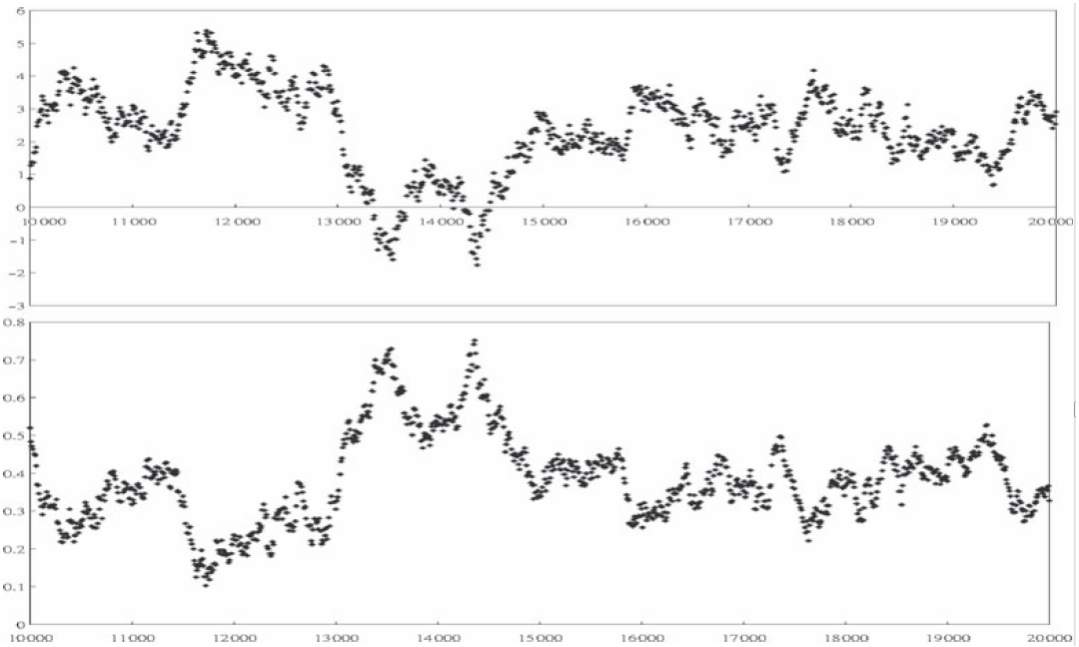


FIGURA 1.3 – Trajetórias de cadeias de Markov para coeficiente estrutural (acima) e covariância residual (abaixo) obtidos de um modelo recursivo bivariado sem restrições necessárias para atingir identificabilidade dos parâmetros (adaptado de WU et al., 2009).

2.1.3.3 Inferência

Após a definição da estrutura causal e da imposição de restrições adequadas, torna-se possível ajustar o modelo e realizar inferências. Como ocorre tipicamente no ajuste de modelos mistos, os melhores preditores lineares não viesados (BLUPs) dos efeitos genéticos aditivos são obtidos condicionalmente a outros parâmetros do modelo. Na prática estes parâmetros não são conhecidos, de

modo que são substituídos por estimadores na expressão do preditor de interesse. Uma crítica a esta abordagem é que a incerteza a respeito dos parâmetros desconhecidos não é considerada na inferência dos preditores (SORENSEN e GIANOLA, 2002). Uma alternativa que não apresenta esta limitação é a análise Bayesiana, na qual inferências a respeito de cada parâmetro é feita com base nas suas distribuições *a posteriori* marginais (SORENSEN e GIANOLA, 2002; GELMAN et al., 2004).

O método numérico de escolha para obter amostras que representem a distribuição *a posteriori* dos parâmetros depende do modelo e da distribuição *a priori* utilizados. Considere, como exemplo, um MEE como em [4], com estrutura causal recursiva acíclica e matriz de covariância residual diagonal. Para ajuste deste modelo, seria razoável propor como distribuição conjunta *a priori* dos parâmetros:

$$p(\boldsymbol{\Lambda},\boldsymbol{\beta},\mathbf{u},\mathbf{G}_0,\boldsymbol{\Psi}_0) = p(\boldsymbol{\Lambda})\,p(\boldsymbol{\beta})\,p(\mathbf{u}\,|\,\mathbf{G}_0)\,p(\mathbf{G}_0)\prod_{j=1}^{t} p(\psi_j)$$

$$\propto N(\mathbf{u}\,|\,0,\mathbf{G}_0\otimes\mathbf{A})IW(\mathbf{G}_0\,|\,\upsilon_{\mathbf{G}},\mathbf{G}_0^{\bullet})\prod_{j=1}^{t} Inv\text{-}\chi^2(\psi_j\,|\,\upsilon_{\psi},s^2),$$

em que a distribuição *a priori* de $\boldsymbol{\beta}$ e das entradas de $\boldsymbol{\Lambda}$ são uniformes não limitadas, $N(\mathbf{u}\,|\,\mathbf{0},\mathbf{G}_0\otimes\mathbf{A})$ é uma distribuição normal multivariada com média $\mathbf{0}$ e matriz de covariância igual a $\mathbf{G}_0\otimes\mathbf{A}$, $IW(\mathbf{G}_0\,|\,\upsilon_{\mathbf{G}},\mathbf{G}_0^{\bullet})$ é uma distribuição de Wishart inversa com $\upsilon_{\mathbf{G}}$ graus de liberdade e matriz de escala $\mathbf{G}_0^{\bullet}$, $Inv\text{-}\chi^2(\psi_j\,|\,\upsilon_{\psi},s^2)$ é uma distribuição qui-quadrado invertida com parâmetro de escala, com υ_{ψ} graus de liberdade e escala s^2, e ψ_j é a variância residual associado à característica *j*. Adicionalmente, υ_G, $\mathrm{G}_0^{\bullet}$, υ_{ψ} e s^2 são considerados como hiperparâmetros conhecidos da distribuição *a priori*.

A densidade conjunta *a posteriori* é obtida por:

$$p(\boldsymbol{\Lambda},\boldsymbol{\beta},\mathbf{u},\mathbf{G}_0,\boldsymbol{\Psi}_0\,|\,\mathbf{y}) \propto p(\mathbf{y}\,|\,\boldsymbol{\Lambda},\boldsymbol{\beta},\mathbf{u},\boldsymbol{\Psi}_0)\,p(\mathbf{u}\,|\,\mathbf{G}_0)\,p(\mathbf{G}_0)\prod_{j=1}^{t} p(\psi_j).$$

Esta distribuição conjunta dos parâmetros não apresenta forma conhecida, mas as condicionais completas destes parâmetros apresentam (os efeitos "fixos", valores genéticos e coeficientes estruturais apresentam distribuição normal multivariada, a matriz de covariância genética apresenta distribuição de Wishart inversa e as variâncias residuais apresentam distribuição de qui-quadrado invertida com parâmetro de escala). Desta forma, é possível amostrar MCMCs que representam a distribuição *a posteriori* dos parâmetros por intermédio do amostrador de Gibbs (GEMAN e GEMAN, 1984).

Entretanto, o amostrador de Gibbs não pode ser utilizado para obtenção da distribuição *a posteriori* em todas as situações. GIANOLA e SORENSEN (2004) descrevem a realização de inferência Bayesiana para parâmetros de um MEE sem restrições paramétricas ou estrutura casual pré-definida, atribuindo distribuições *a priori* próprias para todos os parâmetros. Nesta situação, os autores não obtiveram densidade condicional completa de forma conhecia para os coeficientes estruturais, e propõem a utilização de algoritmo Metropolis (GELMAN et al., 2004) para obter amostras da distribuição deste parâmetro.

2.1.4 Aplicações

A utilização de MEEs no contexto de modelos mistos em genética quantitativa foi introduzida por GIANOLA e SORENSEN (2004). Desde então, muitos autores utilizaram tais modelos para

estudar sistemas multicaracterísticas. MEEs foram utilizadas na avaliação bicaracterística de produção de leite e contagem de células somáticas (CCS) em caprinos de leite (DE LOS CAMPOS et al., 2006a) e em bovinos de leite (DE LOS CAMPOS et al., 2006b). Segundo os autores, a metodologia utilizada produziu resultados que indicam que a associação negativa entre as duas características tem como causa mais importante o efeito negativo da enfermidade na produção de leite e não um efeito de diluição das células somáticas no leite, quando a produção é maior. Uma limitação destes estudos foi o fato de que a análise foi implementada por intermédio do programa LISREL, o qual não considera informações de pedigree.

O relacionamento entre CCS e produção de leite também foi estudado por WU et al. (2007). Porém, o modelo utilizado por estes autores foi uma extensão daquele apresentado por GIANOLA e SORENSEN (2004). O modelo proposto assume heterogeneidade de relações de recursividade e *feedback* entre diferentes extratos do banco de dados utilizado na análise, o que também resulta em diferentes estimadores de componentes de variância para cada extrato. Para esta análise, foi utilizado o programa SirBayes. Os resultados obtidos indicaram efeitos de maior magnitude da CCS sobre a produção de leite, e efeitos de menor magnitude para o sentido inverso. A população de animais avaliados foram extratificadas em dois grupos: alta produção de leite e baixa produção de leite. O efeito de CCS sobre a produção de leite foi de maior magnitude no grupo de alta produção, quando comparado com o mesmo efeito no grupo de baixa produção.

VARONA et al. (2007) utilizaram um modelo recursivo para investigar o relacionamento entre tamanho de leitegada e peso médio de leitões em suínos Landrace e Yorkshire. Na estrutura causal proposta pelos autores, o peso médio dos leitões é influenciado pelo tamanho de leitegada. Os resultados apontam para ausência de efeitos recursivos em suínos Landrace e presença de recursividade apenas entre resíduos para suínos Yorkshire.

Outra extensão dos modelos apresentados por GIANOLA e SORENSEN (2004) foi proposta por WU et al. (2008), para estudar o relacionamento de recursividade/*feedback* entre características lineares e de limiar. Para tal, os autores utilizam modelagem hierárquica Bayesiana. Foi proposta uma variável contínua subjascente que resultava em uma classe fenotípica de acordo com a região do espaço amostral na qual se encontra o valor desta variável contínua. As possíveis regiões são delimititadas por limiares fixos. O modelo foi utlizado para estudar o relacionamento entre presença de mastite clínica e produção de leite em diferentes períodos da lactação. Os resultados obtidos indicam que dentro de um mesmo período, a presença de mastite clínica causa efeito negativo na produção de leite, enquanto o efeito da produção de leite em um período na presença de mastite clínica no período subsequente é de baixa magnitude. Adicionalmente, estes efeitos diminuem quando o animal está em periodo mais avançado da lactação.

Um modelo recursivo entre intervalo de gestação, dificuldadede parto e mortalidade perinatal foi proposto por LÓPEZ DE MATURANA et al. (2008), combinando análise de população heterogênea (WU et al., 2007) e análise de características lineares e de limiar (WU et al., 2008). Os resultados obtidos indicam um intervalo de gestação ótimo (274 dias) com respeito à dificuldade de parto e mortalidade perinatal.

2.2 Busca por estruturas causais

Como apresentado no tópico anterior, para ajustar MEEs é necessário definir *a priori*, para cada variável resposta *i* do conjunto estudado, quais das variáveis remanescentes serão consideradas como pais de *i*. Esta informação é denominada estrutura causal, e pode ser representada graficamente por diagramas. De acordo com a representação matricial de um MEE misto, a estrutura causal define quais entradas da matriz $\boldsymbol{\Lambda}$ serão consideradas como parâmetros livres em $\mathbf{y}_i = \boldsymbol{\Lambda}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i$.

A seleção da estrutura causal a ser utilizada se torna um desafio ao ajuste de um MEE devido ao aumento explosivo do número de possíveis estruturas na medida em que cresce o número de características estudadas, como demonstrado na Tabela 1.1. O grande número de hipóteses causais ocorre mesmo em situações restritas e com pequeno conjunto de características. A comparação exaustiva de modelos se torna inexequível pela utilização de critérios como o AIC (AKAIKE, 1974)

ou o BIC (SCHWARTZ, 1978). Nas aplicações recentes de MEEs no contexto de modelo misto em genética quantitativa, utilizaram-se crenças *a priori* a respeito do sistema estudado para que fossem selecionadas uma estrutura ou um pequeno grupo de estruturas, estas comparadas por critérios semelhantes aos supracitados.

Entretanto, existem algoritmos desenvolvidos por pesquisadores das áreas de inteligência artificial e filosofia da matemática que são capazes de explorar espaços de hipóteses causais e buscar estruturas que são compatíveis com a distribuição conjunta apresentada pelas variáveis estudadas. A seguir, serão apresentadas as bases teóricas, premissas e implementação de um destes algoritmos.

2.2.1 Terminologia

Estruturas causais podem ser representadas por diagramas ou gráficos. Estes consistem em conjuntos de vértices (representando as variáveis) conectadas por linhas (*edges*) que representam conexões diretas simétricas quando não possuem extremidade em seta (linha não-direcionada ou *undirected edge*) ou associações causais quando apresentam seta em uma extremidade (linha direcionada ou *directed edge*). Em $A \rightarrow C \leftarrow B$, os vértices A e B (pais, ou *parents*) são causas diretas do vértice C (filho, ou *child*). Em $A \rightarrow B \rightarrow C$, A é considerado causa indireta de C, relação esta que é mediada por B. Dois vértices são denominados adjacentes caso sejam conectados por uma linha direcionada ou não-direcionada. Se todas as linhas em um gráfico são direcionadas, este gráfico é considerado direcionado. Gráficos acíclicos direcionados (GADs) são gráficos direcionados que não possuem ciclos causais.

Em uma estrutura causal, uma trilha (*path*) é uma sequência de vértices conectados por linhas. A sequência de vértices desta trilha pode respeitar o sentido das setas (trilha direta ou *direct path,* e.g. $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$) ou não (trilha indireta ou *undirect path,* e.g. $A \rightarrow B \rightarrow C \leftarrow D \rightarrow E$). Em uma trilha indireta, podem existir vértices nos quais setas apresentam convergência, como C em $A \rightarrow C \leftarrow B$. Estes vértices são chamados *colliders*. Uma trilha se apresenta ativa quando carrega informação ou dependência entre os vértices de seus extremos. Desta forma, incondicionalmente, toda trilha é ativa, a não ser que apresente ao menos um *collider*. Tais vértices bloqueiam o fluxo de dependência. Desta forma, A é marginalmente dependente de B nas trajetórias $A \rightarrow C \rightarrow B$, $A \leftarrow C \leftarrow B$ e $A \leftarrow C \rightarrow B$. O mesmo não ocorre em $A \rightarrow C \leftarrow B$. A capacidade de um vértice para transmitir ou bloquear fluxos de dependência se inverte quando a trilha é analisada *condicionalmente* a este. Como consequência, condicionalmente a C, A e B são dependentes em $A \rightarrow C \leftarrow B$ e independentes em $A \rightarrow C \rightarrow B$, $A \leftarrow C \leftarrow B$ e $A \leftarrow C \rightarrow B$. A independência entre dois vértices pertencentes a um gráfico, condicionalmente a um subconjunto dos vértices remanescentes, é chamada d-separação (*d-separation*). Como definição formal, considerando dois vértices A e B em um GAD, eles são considerados d-separados condicionalmente a um subconjunto $\mathbf{S}$ de vértices remanescentes se não existe trilha que permita fluxo de dependência entre A e B no gráfico (i.e., não existe trilha entre A e B no GAD de modo que todos os *colliders* ou seus descendentes estão em $\mathbf{S}$ e nenhum não-*collider* está em $\mathbf{S}$).

Sob algumas premissas, d-separações na estrutura causal são refletidas como independências condicionais estatísticas na densidade conjunta das variáveis estudadas, o que vai ser explorado na tentativa de selecionar estruturas causais a partir desta densidade conjunta. Desta forma, assume-se que um modelo causal impõe algumas marcas na densidade dos dados observados, e tenta-se recuperar a estrutura causal deste modelo a partir destas marcas. Esta tentativa assume uma conexão entre estrutura causal e distribuição de probabilidade. Em seguida, serão descritos conceitos nos quais esta conexão se baseia e como estes podem ser utilizados para propor premissas da seleção de estruturas causais acíclicas.

2.2.2 Propriedades de Markov

Uma estrutura causal carrega em si um conjunto de dependências e independências condicionais entre os vértices que a constituem. Oberva-se que qualquer vértice do gráfico, condicional a seus pais, é independente de todos os vértices que não são seus descendentes. Adicionalmente, uma densidade conjunta é considerada compatível com uma estrutura causal gráfica se a primeira pode ser decomposta como em uma rede Bayesiana (PEARL, 1988), na forma de um produto envolvendo a probabilidade de cada variável condicional aos seus pais de acordo com a estrutura causal (Compatibilidade de Markov). Nesse caso, para um conjunto de variáveis **V**, em que **pais**(V) é o conjunto dos pais de cada variável V em **V** (SPIRTES et al., 2000):

$$p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p\left(V \mid \mathbf{pais}(V)\right)$$

Esta compatibilidade indica se uma estrutura causal é capaz de gerar determinada distribuição de probabilidade (e.g., a estrutura $A \leftarrow C \rightarrow B$ não é capaz de gerar uma distribuição na qual A é dependente de B condicional a C). Uma maneira de caracterizar o conjunto de distribuições compatíveis com um GAD é, com base neste, observar o conjunto de independências condicionais que a distribuição deve satisfazer. Estas são obtidas no gráfico que representa a estrutura causal pelo critério da d-separação (PEARL, 2000).

Como mencionado na seção anterior, os métodos de busca por estruturas causais têm base na conexão entre tais estruturas e densidades de probabilidade. Uma condição necessária para esta conexão é denominada Condição Causal de Markov (PEARL, 2000), na qual o modelo causal induz uma distribuição que satisfaz a compatibilidade de Markov, ou seja, condicionalmente aos pais na estrutura causal, cada variável na distribuição deve ser independente das variáveis não descendentes. Esta condição é consequência de duas premissas: o compromisso em incluir no modelo todas as causas de duas ou mais variáveis-resposta estudadas (suficiência causal) e a premissa de que não há associação entre pares de variáveis sem causalidade: ou uma variável causa a outra, ou elas apresentam uma causa comum (*Reichenbach`s common cause assumption*).

Considerando o processo amostral com base em um MEE com determinada estrutura causal, isto seria o mesmo que considerar resíduos independentes para cada variável resposta. Resíduos representam o efeito dos pais da variável estudada que não estão no modelo. Desta forma, covariância residual entre duas variáveis estudadas significa a existência de um pai comum entre elas. Por outo lado, sob suficiência causal, não há fonte de covarância residual, e uma estrutura diagonal é imposta à matriz de covariância.

Outra propriedade importante que será mencionada posteriormente é a equivalência observacional de dois GADs diferentes, em que cada distribuição compatível com um GAD é também compatível com outro GAD equivalente. Se duas estruturas causais podem responder pela mesma descrição estatística multivariada, então a evidência estatística simplemente não pode fazer distinção entre as duas estruturas. A equivalência estatística é o limite dos métodos de seleção de estruturas causais com base em dados observacionais (SPIRTES et al., 2000). Estruturas equivalentes possuem as mesmas adjacências entre vértices e os mesmos *colliders*. Desta forma, $A \rightarrow C \rightarrow B$, $A \leftarrow C \leftarrow B$ e $A \leftarrow C \rightarrow B$ são equivalentes por induzirem o mesmo padrão de distribuição conjunta. Por outro lado, qualquer modificação em $A \rightarrow C \leftarrow B$ leva a uma estrutura não equivalente (PEARL, 2000).

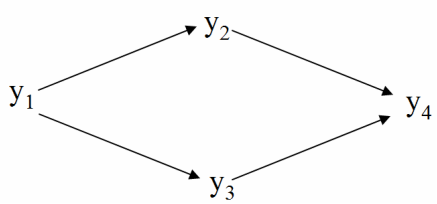
2.2.3 Minimalidade e estabilidade (*faithfulness*)

Um grande número de estruturas causais pode, em princípio, se ajustar a uma dada distribuição. Para um conjunto de variáveis em suficiência causal, uma estrutura acíclica extremamente complexa pode, com uma parametrização finamente ajustada, mimetizar a distribuição gerada por várias outras estruturas. Torna-se necessário definir o que seria um critério para preferência de modelo. Seguindo as normas geralmente utilizadas na indução científica, torna-se interessante

desprezar teorias complexas quando existe uma teoria mais simples e igualmente consistente com as evidências (PEARL, 2000).

Desta forma, uma estrutura passa a ser preferida em relação a outra quando esta última, sob determinada parametrização, tem capacidade de mimetizar a distribuição gerada por um modelo sob a primeira estrutura, mas o contrário não é verdadeiro. A estrutura preferida tem menor poder expressivo ou flexibilidade (mas não necessariamente menos parâmetros), evitando o sobreajuste. O poder expressivo de uma estrutura é dado pelas independências nela presentes, de maneira que a simples topologia pode ser utilizada para seleção, sem se preocupar com a parametrização. Parece razoável então buscar por estruturas consistentes com a distribuição conjunta observada (que pode ser utilizada por modelos amostrais e gerar tal distribuição, com todas as suas dependências) e que tenham o menor poder expressivo possível (i.e., que sejam mínimas).

Não há, porém, garantias de que a estrutura subjacente que gera os dados não apresente uma parametrização "patológica" que resulte em independências que não refletem d-separações. É possível que, dada uma estrutura causal, a distribuição conjunta de variáveis apresente independências condicionais além daquelas que seguem logicamente da Condição Causal de Markov. Considere a seguinte estrutura, utilizada como exemplo por SPIRTES et al. (2000):



Esta estrutura causal resulta no seguinte MEE:

$$y_1 = e_1$$
$$y_2 = \lambda_{21} y_1 + e_2$$
$$y_3 = \lambda_{31} y_1 + e_3$$
$$y_4 = \lambda_{42} y_2 + \lambda_{43} y_3 + e_4.$$

Algumas independências condicionais estatísticas são esperadas por consequência da Condição Causal de Markov (e.g., y_2 e y_3 condicional a y_1, ou y_1 e y_4 condicional a y_3 e y_2). Porém, é possível construir um modelo em que y_1 e y_4 sejam marginalmente independentes, o que não reflete uma d-separação, já que existem duas trilhas ativas entre os dois vértices no gráfico. Para tal, bastaria fazer com que $\lambda_{21}\lambda_{42} = -\lambda_{31}\lambda_{43}$, de modo que as duas trilhas se cancelem.

Tal situação ocorre raramente na prática, e necessitaria de ajuste fino dos parâmetros de modo a fazer com que ocorra assintoticamente. Para desprezar esta situação, parte-se da premissa de que a distribuição conjunta é estável, ou crível (*faithful*) em relação ao modelo causal que a gera. A restrição de que todas as independências condicionais na distribuição conjunta resultante de um modelo causal são estáveis tem como consequência a impossibilidade de se destruir independências estatísticas condicionais pela simples modificação dos valores dos parâmetros (no exemplo dado, basta modificar os valores do parâmetro de modo a tornar $\lambda_{21}\lambda_{42} = -\lambda_{31}\lambda_{43}$ falsa para destruir a independência marginal entre y_1 e y_4). Se uma independência condicional estatística ocorre para qualquer parametrização do modelo causal, esta independência é estável, como ocorre no caso das independências estatísticas que refletem d-separações em um modelo sob Condição Causal de Markov.

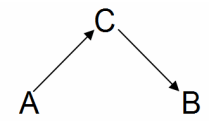
PEARL (2000) apresenta uma analogia interessante que permite observar a relação próxima entre minimalidade e estabilidade. Para tal, considera-se uma figura na qual é possível observar uma cadeira. É necessário decidir-se entre duas teorias:

T1 – O objeto na figura é uma cadeira

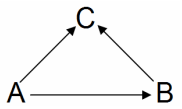
T2 – O objeto na figura pode ser uma cadeira ou pode ser duas cadeiras posicionadas de tal maneira que uma esconde a outra.

A preferência por T1 pode ser justificada baseando-se na minimalidade ou na estabilidade. Sob minimalidade, escolhe-se T1 porque o conjunto de cenas possíveis admitindo-se T1 é um subconjunto das cenas possíveis admitindo-se T2. Pelo menor poder expressivo de T1, escolhe-se esta teoria em detrimento de T2 para evitar sobreajuste, a não ser que existam evidências favoráveis a T2. Por outro lado, a escolha de T1 pode-se basear na estabilidade, por considerar que o alinhamento perfeito entre dois objetos na figura é improvável e instável em relação a pequenas modificações no ângulo de visão.

Um exemplo adicional seria uma análise hipotética de três variáveis *A*, *B* e *C*. Na distribuição conjunta, cada par é dependente marginalmente e condicionalmente á variável restante, exceto o par *A* e *B* dado *C*. Uma estrutura compatível seria:



Tal distribuição também poderia ser gerada pela estrutura:



se a parametrização do MEE fosse tal que $\lambda_{CA}\lambda_{CB} = -\lambda_{BA}$, pois condicionalmente a *C*, duas trilhas entre *A* e *B* se tornam ativas e, com a parametrização dada, uma trilha anula a outra. Entretanto, a primeira estrutura é preferida sob argumentos análogos aos utilizados para a escolha de hipóteses ao observar a figura de uma cadeira.

2.2.4 Premissas

Na próxima seção, será apresentado o algoritmo IC (*Inductive Causation Algorithm* - VERMA e PEARL, 1990; PEARL, 2000) para seleção de uma classe de estruturas causais equivalentes que seja compatível com o padrão de independências condicionais observados na distribuição conjunta das variáveis estudadas. Este algoritmo busca por estruturas causais para o modelo $\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{e}_i$. De acordo com SPIRTES et al. (2000), a seleção da classe de estruturas equivalentes é proposta sob as seguintes premissas (na realidade, os autores apresentam as premissas para o algoritmo SGS, um algoritmo que equivale ao IC, ambos desenvolvidos paralelamente (PEARL 2009; e SPIRTES 2009, comunicação pessoal) :

1 – O conjunto de variáveis estudadas é causalmente suficiente.
2 – Cada unidade na população apresenta as mesmas relações causais entre variáveis.
3 – A distribuição das variáveis observadas é crível em relação a um GAD.
4 – As decisões estatísticas exigidas pelo algoritmo são corretas para a população.

O que motiva a segunda premissa é o fato de que a conexão entre estruturas causais e distribuições pode se tornar problemática em situações em que se tem uma distribuição conjunta composta de uma mistura de distribuições, cada uma delas relativa a uma subpopulação com uma estrutura causal específica (cada uma delas compatível com uma esrutura causal diferente). A distribuicão resultante não seria, nesta situação, fonte confiável de informações a respeito da estrutura

causal (SPIRTES, 2000). A quarta premissa provém de que não há garantias de recuperação da classe correta de estruturas causais se as decisões estatísticas a respeito de independências condionais com base em uma amostra não correspondem ao que ocorre na população. Erros podem ocorrer, por exemplo, por efeito amostral. Mais detalhes a respeito das decisões estatísticas serão descritos adiante.

As premisas são fortes, mas não são mais fortes do que aquelas geralmente aplicadas na utilização de MEEs no contexto de modelos mistos em genética quantitativa, ou mesmo em qualquer tentativa de inferência causal em outras áreas, utilizando outros modelos (SPIRTES et al., 2000). A primeira premissa é, provavelmente, a mais forte. A consequência prática desta premissa é impor uma distribuição independente para os resíduos do MEE (em outras palavras, impor uma estrutura diagonal para a matriz de covariância residual). Na realidade, a imposição de tal estrutura é comum na aplicação recente de MEEs, sob a justificativa de atingir a identificabilidade dos parâmetros (DE LOS CAMPOS et al., 2006a; DE MATURANA et al., 2009 e HERINGSTAD et al., 2009). Uma premissa forte adicional nas aplicações de MEEs é o conhecimento prévio a respeito da estrutura causal. Esta premissa pode ser evitada se utilizarmos o algoritmo IC para explorar o espaço de estruturas causais sem modificar a construção feita para os resíduos do modelo.

2.2.5 Algoritmo IC

O algoritmo IC é utilizado para selecionar estruturas causais (ou uma classe de estruturas observacionalmente indistiguíveis) a partir das associações observadas entre características. O algoritmo tem base em uma série de perguntas a respeito da independência condicional estatística entre variáveis e na premissa de que estas independências são reflexos de d-separações na estrutura causal subjacente. Os dados de entrada do algoritmo são os elementos de um matriz de correlação, dos quais dependências condicionais podem ser avaliadas. A saída do algoritmo é um gráfico parcialmente direcionado (gráfico que contém linhas direcionadas e não direcionadas) que representa uma classe de estruturas causais compatíveis com as independências condicionais obtidas. Esta classe é geralmente uma grande restrição do espaço de hipóteses causais em relação ao espaço inicial. Desta forma, o algoritmo busca construir um conjunto de GADs que satisfazem um conjunto dado de d-separações, se tais GADs existem (SPIRTES et al., 2000).

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.

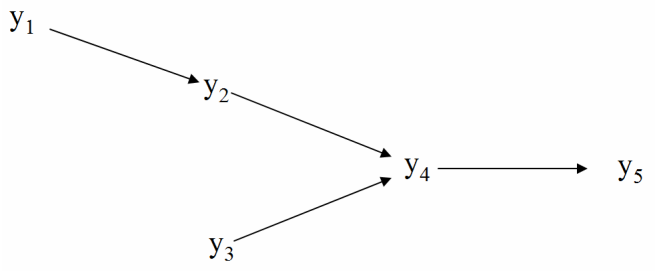
O objetivo do primeiro passo do algoritmo é obter um gráfico que especifique pares de características que são diretamente conectadas, mas sem especificar direção causal. Vértices adjacentes em um gráfico não são d-separados condicionalmente a qualquer conjunto de vértices remanescentes. A consequência observacional é que estas variáveis são estatisticamente dependentes condicionalmente a qualquer conjunto de variáveis restantes.

O objetivo do segundo passo do algoritmo é orientar linhas, por intermédio da busca por vértices no gráfico no qual setas convergem de ambos os sentidos em uma trilha (*colliders*). Estruturas internas dos gráficos compostas por um *collider* que sofre influência causal de dois vértices que não

são conectados são chamadas *unshielded colliders* (SPIRTES et al., 2000), como em $A \rightarrow C \leftarrow B$. Em tal estrutura, os pais são d-separados condicionalmente a pelo menos um conjunto de variáveis restantes no gráfico completo, mas não se o vértice *C* pertence a este conjunto. Condicionalmente a *C*, a trilha entre *A* e B por intermédio de *C* permite o fluxo de dependência, não ocorrendo d-separação. A consequência observacional é que *A* e *B* nunca são estatisticamente independentes condicionalmente a qualquer conjunto de variáveis que contenha *C*.

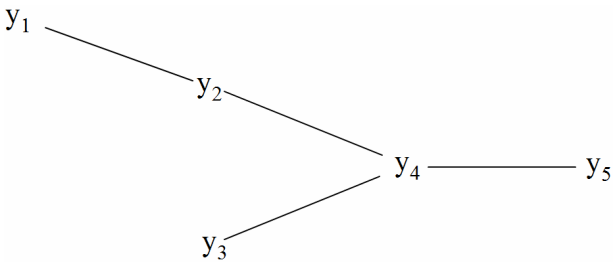
O terceiro passo consiste em orientar linhas adicionais de maneira que isso não resulte em um novo *collider* ou em um ciclo. Desta forma, se três variáveis em um gráfico semidirecionado hipotético resultante do passo 2 se apresentam conectados como em $A \rightarrow B - C$, a linha entre *B* e *C* deve ser orientada em direção a *C*, uma vez que a direção contrária resulta em um *collider* em *B* que não foi detectado no passo anterior. O mesmo deve ser feito para uma linha caso uma das direções resulte em um ciclo.

Considere, para exemplo de aplicação do algoritmo, a distribuição gerada pela seguinte estrutura:



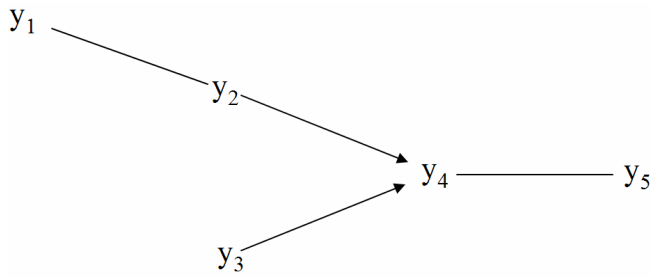
em que cada variável sofre influência de resíduos independentes com distribuição normal. Considere ainda que os parâmetros desta distribuição são conhecidos sem erros.

A matriz de covariância submetida ao primeiro passo do algoritmo resultaria no seguinte gráfico não direcionado:

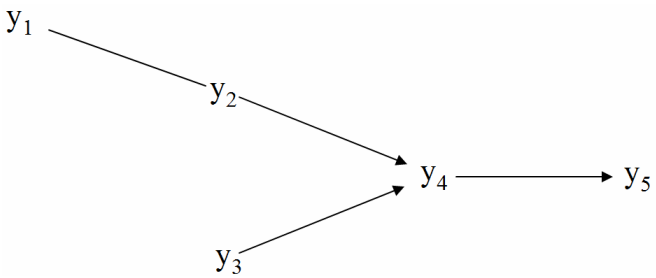


Uma vez que as variáveis desconectadas seriam independentes condicionalmente a algum conjunto de variáveis (e.g., espera-se que y_2 e y_3 sejam marginalmente independentes, e que y_2 e y_5 sejam independentes dado y_4).

As trilhas $y_1 - y_2 - y_4$, $y_2 - y_4 - y_3$, $y_2 - y_4 - y_5$, e $y_3 - y_4 - y_5$ seriam consideradas como *unshielded colliders* candidatas pelos segundo passo do algoritmo. Porém, para o exemplo descrito, espera-se que os extremos destas trilhas sejam probabilisticamente independentes dado a variável restante, com exceção feita a $y_2 - y_4 - y_3$. Desta forma, y_4 seria reconhecida como *collider*, e o segundo passo do algoritmo resulta em:



A seguir, observa-se que duas linhas restaram sem sofrer direcionamento. A linha entre os vértices y_1 e y_2 pode ser direcionada para ambos os sentidos sem criar um novo *collider*. Desta forma, a direção não pode ser escolhida com base na distribuição dos dados. Entretanto, a linha entre y_4 e y_5 deve ser direcionada para y_5, resultando em:



A direção da linha entre y_1 e y_2 pode ser escolhida com base em outras informações, como crença *a priori* (e.g., se y_1 ocorre antes de y_2 é natural direcionar a linha para y_2). Desta forma, o usuário pode ter algum conhecimente prévio que ajude a restringir ainda mais a busca. A utilização de informações *a priori* para direcionar uma linha pode resultar no direcionamento de linhas adicionais (evitando ciclos e novos *colliders*). O algoritmo também pode ser modificado para que informações *a priori* dominem em relação aos resultados do algoritmo, impondo ou proibindo a existência de uma linha, ou sua direção (SHIPLEY, 2002, SPIRTES et al., 2000).

As perguntas do algoritmo exigem que pares de variáveis sejam declarados como condicionalmente dependentes ou independentes. Esta decisão é baseada na correlação parcial entre as variáveis. No exemplo descrito, as correlações entre as variáveis estudadas são conhecidas sem erros. Porém, este não é o caso na aplicação prática do algoritmo, em que as correlações são estimadas de uma amostra finita. Neste caso, a incerteza a respeito destes parâmetros devem ser consideradas na decisão. Em abordagem frequentista, as decisões podem ser feitas pelo teste da hipótese nula de que a correlação parcial é igual a zero (mais detalhes são fornecidos em SHIPLEY, 2002 e SPIRTES et al., 2000). Para abordagem Bayesiana, estas decisões podem ser feitas utilizando interalos HPD (*Highest Posterior Density*) obtidas para cada correlação. No caso, se o intervalo contém zero, a correlação é declarada como nula. É evidente que para um mesmo conteúdo de intervalo HPD, ou mesmo nível de significância para o teste de hipótese, a eficiência em detectar correlações de baixa magnitude diminui na medida em que a incerteza a respeito da correlação aumenta (e.g., quando o número de informações é menor). Como o algoritmo tem propósito de exploração, e não há uma preferência *a priori* entre declarar erroneamente uma correlação diferente de zero como nula e declarar uma correlação nula como não-nula, torna-se razoável diminuir o conteúdo de HPD ou aumentar o nível de significância do teste de hipótese utilizados nas decisões estatísticas para cenários em que há maior incerteza a respeito das correlações parciais estimadas.

A quarta premissa é evidentemente necessária para garantir a recuperação de uma estrutura causal crível em relação à densidade conjunta populacional, mas erros estatísticos não resultam necessariamente em erros na busca. Tal situação ocore quando existem mais de um conjunto de vértices que d-separam um par específico de vértices, mas nem todos refletem independência

estatística na distribuição de probabilidade obtida de uma amostra. Neste caso, o gráfico não direcionado selecionado seria o mesmo, pois a linha entre um par de vértices é descartada na presença de ao menos uma correlação parcial declarada como nula. Esta propriedade confere certa estabilidade ao primeiro passo do algoritmo na presença de pequenos erros nos dados de entrada. Por sua vez, o restante do algoritmo é menos estável, pois algum erro na lista de d-separações baseada nas decisões estatísticas ou mesmo no gráfico não direcionado proveniente do primeiro passo podem produzir erros maiores nos resultados. Isto ocorre porque a orientação de linhas decorrente da detecção de *colliders* geralmente resulta na orientação subsequente de várias outras linhas no gráfico (SPIRTES et al., 2000).

2.2.6 Considerações

Como demonstrado, sob algumas premissas é possível realizar seleção de estruturas causais com base na distribuição conjunta de probabilidades. Tal seleção vai contra a idéia de que não se pode apreender conhecimento causal com base em correlações obtidas de dados não-experimentais.

Vários outros algoritmos mais sofisticados foram desenvolvidos para a mesma finalidade, apresentando vantagens como maior eficiência computacional, ou capacidade de trabalhar em situações em que não se assume suficiência causal. Entretanto, o algoritmo IC é interessante para estudos que envolvem um número pequeno de características (como ocorre tipicamente em modelos multicaracterísticas no contexto de genética quantitativa) e se destaca pela sua simplicidade de implementação.

Como mencionado na seção anterior, o algoritmo como apresentado realiza busca de estruturas causais para MEEs do tipo $\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{e}_i$. Desta forma, não se considera a existência de variáveis correlacionadas não-observáveis influenciando as características estudadas. Mesmo sob imposição de matriz residual diagonal, este não é o contexto típico para modelos mistos utilizados em genética quantitativa, como em $\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i$. Desta forma, a aplicação de tais algoritmos neste contexto deve ser realizada mediante algum procedimento que recupere a condição causal de Markov.

3 REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H., 1973 Information theory and an extension of the maximum likelihood principle, pp. 267–291 in *2nd International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki. Publishing House of the Hungarian Academy of Sciences, Budapest.

DE LOS CAMPOS, G., D. GIANOLA , P. BOETTCHER e P. MORONI, 2006a A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. J. Anim. Sci. **84:** 2934-2941.

DE LOS CAMPOS, G., D. GIANOLA e B. HERINGSTAD, 2006b A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows. J. Dairy Sci. **89:** 4445-4455.

DUNCAN, O. D., 1975 *Introduction to Structural Equation Models*. New York: Academic Press.

GELMAN, A., J. B. CARLIN, H. S. STERN and D. B. RUBIN, 2004 *Bayesian Data Analysis.* Chapman & Hall, London.

GEMAN, S. e D. GEMAN, 1984 Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6:** 721–741.

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

GOLDBERGER**,** A. S., 1972 Structural equation methods in the social sciences. Econometrica **40:** 979–1001.

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

HENDERSON, C. R., 1976 Multiple Trait Sire Evaluation Using the Relationship Matrix. J. Dairy Sci., 59:769–774.

HERINGSTAD, B., X.-L. WU e D. GIANOLA, 2009 Inferring relationships between health and fertility in Norwegian red cows using recursive models. J. Dairy Sci. **92**:1778–1784

KONIG, S., X. L. WU, D. GIANOLA, B. HERINGSTAD e H. SIMIANER, 2008 Exploration of relationships between claw disorders and milk yield in Holstein cows via recursive linear and threshold models. J. Dairy Sci. **91:**395–406.

LASSEN, J., M. K. SORENSEN, P. MADSEN e P. DUCROCQ, 2007 An approximate multitrait model for genetic evaluation in dairy cattle with a robust estimation of genetic trends. Genet. Sel. Evol., **39:**353–367.

DE MATURANA, E. L., X. L. WU, D. GIANOLA, K. A. WEIGEL e G. J. M. ROSA, 2009 Exploring biological relationships between calving traits in primiparous cattle with a Bayesian recursive model. Genetics **181:**277–287.

PEARL, J., 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffman, San Mateo, CA.

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

POLLAK, E. J., J. van der WERF e R. L. QUAAS, 1984 Selection Bias and Multiple Trait Evaluation. J. Dairy Sci. **67:**1590-1595.

SCHWARZ, G., 1978 Estimating the dimension of a model. Ann. Stat.**6:** 461–464.

SHIPLEY, B., 1997 Exploratory path analysis with applications in ecology and evolution. Am. Nat. **149:** 1113-1138.

SHIPLEY, B., 2002 *Cause and Correlation in Biology*. Cambridge University Press, Cambridge/London/New York.

SORENSEN, D. e D. GIANOLA, 2002 *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.

SPIRTES, P., C. GLYMOUR e R. SCHEINES, 2000 *Causation, Prediction and Search*, Ed. 2. MIT Press, Cambridge, MA.

TURNER, M. E. e C. E. STEVENS, 1959 The regression analysis of causal paths. Biometrics **15:** 236–258.

VALENTE, B.D., ROSA, G.J.M., de los CAMPOS, G., GIANOLA, D. e SILVA, M.A. Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models, Genetics **185:**633–644., doi:10.1534/genetics.109.112979, 2010.

VARONA, L., D. SORENSEN e R. THOMPSON, 2007 Analysis of litter size and average litter weight in pigs using recursive model. Genetics **177:** 1791-1799.

VERMA, T. e J. PEARL, 1990 Equivalence and synthesis of causal models. Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence (July, Cambridge, MA), 220-227. Reprinted in Uncertainty in Artificial Intelligence, **6:** 255:268, Elsevier, Amsterdam.

WIGGANS, G. R. e M. E. GODDARD, 1997 A Computationally Feasible Test Day Model for Genetic Evaluation of Yield Traits in the United States. J Dairy Sci., **80:**1795–1800.

WRIGHT, S., 1921 Correlation and causation. J. Agric. Res. **201:** 557–585.

WU, X.-L., B. HERINGSTAD, Y. M. CHANG, G. DE LOS CAMPOS e D. GIANOLA, 2007 Inferring relationships between somatic cell score and milk yield using simultaneous and recursive models. J. Dairy Sci. **90:** 3508-3521.

WU, X.-L., B. HERINGSTAD e D. GIANOLA, 2008 Exploration of lagged relationships between mastitis and milk yield in dairy cows using a Bayesian structural equation Gaussian-threshold model. Genet. Sel. Evol. **40:** 333–357.

WU, X.-L., B. HERINGSTAD e D. GIANOLA, 2010 Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. J. Anim. Breed. Genet. **127:** 3–15.

CAPÍTULO 2
BUSCA POR ESTRUTURAS CAUSAIS NO CONTEXTO DE MODELOS MISTOS EM GENÉTICA QUANTITATIVA

RESUMO

Modelos de equações estruturais (MEEs) podem ser utilizados para estudar relacionamentos de recursividade e *feedback* em análise multivariada. O número de estruturas causais recursivas distintas que podem ser utilizadas para ajustar tais modelos pode ser muito grande, mesmo no estudo de um pequeno conjunto de características. Em aplicações recentes de MEEs no contexto de modelos mistos em genética quantitativa, as estruturas causais foram pré-selecionadas utilizando apenas conhecimento biológico *a priori*. Desta forma, a ampla gama de estruturas causais possíveis não tem sido adequadamente explorada. Como alternativa, espaços de estruturas causais podem ser explorados por meio de algoritmos que, orientados por evidências nos dados, podem buscar por estruturas causais que são compatíveis com a distribuição conjunta das variáveis estudadas. Entretanto, a busca não pode ser realizada diretamente na distribuição conjunta dos fenótipos, uma vez que esta se apresenta potencialmente confundida por covariâncias genéticas. Na presente tese, o autor propõe buscar por estruturas causais recursivas entre fenótipos utilizando o algoritmo IC (*Inductive Causation*) após ajustar os dados para efeitos genéticos. Para isso, ajusta-se um modelo multicaracterísticas padrão, por meio de métodos Bayesianos, para se obter a matriz de covariância fenotípica condicionalmente aos efeitos genéticos não-observáveis, que é posteriormente submetida ao algoritmo IC. Como exemplo ilustrativo, a metodologia proposta foi aplicada em dados simulados para características múltiplas mensuradas em um conjunto de linhagens endogâmicas.

Palavras-chave: busca por estrutura causal, confundimento genético, modelos de equações estruturais, modelos mistos, sistemas biológicos

ABSTRACT

Structural Equation Models (SEM) can be used to study recursive and simultaneous relationships in multivariate analyses. Nonetheless, the number of different recursive causal structures that can be used for fitting a SEM to multivariate data can be huge, even when only a few traits are considered. 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, using data driven evidence, can search for structures that are compatible with the joint distribution of the variables under study. However, the search cannot be performed directly on the joint distribution of the phenotypes as it is possibly confounded by genetic covariance among traits. In this thesis, we propose to search for recursive causal structures among phenotypes using the Inductive Causation (IC) algorithm after adjusting the data for genetic effects. A standard multiple trait model is fitted using Bayesian methods to obtain a posterior covariance matrix of phenotypes conditional to unobservable additive genetic effects, which is then used as input for the IC algorithm. As an illustrative example, the proposed methodology was applied to simulated data related to multiple traits measured on a set of inbred lines.

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

1 INTRODUÇÃO

Em diversos sistemas biológicos, fenótipos de diferentes características exercem efeitos mútuos que podem ser estudados por meio de modelos estatísticos recursivos ou simultâneos. Como exemplo, alta produção em vacas de leite aumenta a susceptibilidade a certas doenças e, em sentido oposto, a presença de doenças pode afetar a produção de leite. De modo semelhante, o transcriptoma pode ser uma função do estado reprodutivo em mamíferos, enquanto este pode depender de outras variáveis fisiológicas. Conhecimento a respeito das redes que descrevem tais relações pode ser utilizado para predizer o comportamento de sistemas biológicos que envolvem características complexas relacionadas, por exemplo, a doenças, crescimento e reprodução. Modelos de Equações Estruturais, ou MEEs (WRIGHT, 1921; HAAVELMO 1943), são utilizados para estudar relacionamentos recursivos e de *feedback* entre fenótipos que pertencem a um sistema multivariado, como aqueles estudados por modelos multicarcterísticas na área de genética quantitativa. MEEs permitem uma interpretação do relacionamento entre características que é diferenciada em relação à oferecida por modelos multicaracterísticas padrão (MMC). Neste modelo, todos os relacionamentos entre características são representados por associações lineares simétricas entre variáveis aleatórias, ou seja, são representados por covariâncias. Por sua vez, em MEEs, uma característica pode ser considerada como preditor de outra característica, o que resulta em um elo funcional (causal) entre elas.

Para ajustar MEEs, é necessário definir a estrutura causal, que pode ser representada por um gráfico direcionado no qual cada ligação entre variáveis corresponde a um relacionamento funcional direto (PEARL, 2000). Para *k* variáveis resposta, e considerando todos os possíveis relacionamentos recursivos e de *feedback* entre pares de características, existem $k(k-1)$ possíveis coeficientes estruturais. Mesmo considerando apenas estruturas acíclicas, o número de possíveis estruturas causais cresce explosivamente na medida em que cresce o número de características estudadas (SHIPLEY, 2002). Por este motivo, a escolha da estrutura causal se torna um desafio.

Modelos de equações estruturais adaptados para o contexto de genética quantitativa foram descritos por GIANOLA e SORENSEN (2004). Posteriormente, muitos autores utilizaram estes modelos no contexto mencionado, definindo a estrutura causal com base em crenças *a priori* a respeito do sistema biológico estudado (DE LOS CAMPOS et al., 2006a,b; WU et al., 2007; KONIG et al., 2008; DE MATURANA et al., 2009; WU et al. 2010). Tipicamente, uma estrutura ou um pequeno conjunto de estruturas é pré-selecionado e os modelos construídos com base nas estruturas deste conjunto são ajustados e comparados por meio de algum critério, como por exemplo, AIC (AKAIKE, 1974) ou BIC (SCHWARTZ, 1978). Porém, esta abordagem pode ser ineficiente porque o conjunto de estruturas causais possíveis testadas é diminuto.

Como alternativa, algoritmos que utilizam o conceito de d-separação podem ser empregados para explorar o espaço de hipóteses causais de modo a selecionar uma estrutura causal (ou uma classe de estruturas observacionalmente equivalentes) que é capaz de gerar o padrão observado de independências condicionais probabilísticas entre variáveis. Algoritmos de busca por estruturas causais utilizando informações genômicas foram estudados, por exemplo, por SCHADT et al. (2005), LI et al. (2006), CHEN et al. (2007), LIU et al. (2008), CHAIBUB NETO et al. (2008), e ATEN et al. (2008). Porém, tais algoritmos ainda não foram aplicados para recuperar estruturas causais no contexto de modelos mistos em genética quantitativa, nos quais apenas informações fenotípicas e de parentesco são disponíveis.

Algoritmos que utilizam testes de d-separação buscam por estruturas causais que são compatíveis com o padrão de independências condicionais observados nos dados. Porém, no contexto de modelos mistos, covariâncias genéticas confundem a busca, pois são fontes de covariâncias fenotípicas que não são impostas pelas relações recursivas entre características. Por consequência, realizar a busca por estruturas causais diretamente sobre os dados observados pode levar a resultados de baixa qualidade. Este artigo contribui com a literatura sobre MEEs por apresentar metodologia que permite buscar por estruturas causais recursivas no contexto de modelos mistos em genética quantitativa. A metodologia proposta explora propriedades específicas dos modelos mistos e utiliza o

algoritmo IC (*Inductive Causation*, VERMA e PEARL, 1990; PEARL, 2000) em conjunto com análise de dados Bayesiana. O artigo se estrutura da seguinte maneira: **METODOLOGIA** fornece uma sucinta revisão de literatura sobre MEEs para genética quantitativa, descreve o algoritmo IC e como implementá-lo no contexto de modelos mistos, de modo que a busca pode ser realizada após considerar correlações induzidas por efeitos genéticos. Na seção **EXEMPLO**, a metodologia proposta é aplicada em dados simulados pertencendo a cinco características amostradas de um modelo causal recursivo para uma população hipotética com efeitos correlacionados de linhagem endogâmica. Finalmente, na seção **DISCUSSÃO** as principais características e desafios associados à metodologia proposta são destacados.

2 METODOLOGIA

2.1 Modelos de equações estruturais (MEEs)

De acordo com GIANOLA e SORENSEN (2004), um MEE com determinada estrutura causal recursiva e efeitos genéticos aditivos aleatórios pode ser escrito da seguinte maneira:

$$\mathbf{y}_i = \boldsymbol{\Lambda}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i, \qquad [1]$$

em que $\mathbf{y}_i$ é um vetor ($t\times1$) de fenótipos registrados para o indivíduo *i* (e.g. um animal ou planta); $\boldsymbol{\Lambda}$ é uma matriz ($t \times t$) com zeros na diagonal e coeficientes estruturais nas demais entradas; $\mathbf{X}_i\boldsymbol{\beta}$ representa uma regressão linear sobre covariáveis exógenas, em que a matriz $\mathbf{X}_i$ contém as covariáveis e $\boldsymbol{\beta}$ é um vetor de "efeitos fixos"; $\mathbf{u}_i$ é um vetor ($t\times1$) de efeitos genéticos aditivos aleatórios e $\mathbf{e}_i$ é o vetor de resíduos do modelo de mesma dimensão, ambos associados ao indivíduo *i*.

A seguinte distribuição conjunta é assumida para $\mathbf{u}_i$ e $\mathbf{e}_i$: $\begin{bmatrix}\mathbf{u}_i\\ \mathbf{e}_i\end{bmatrix} \sim N\left\{\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G}_0 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}_0\end{bmatrix}\right\}$, na qual $\mathbf{G}_0$ e $\boldsymbol{\Psi}_0$ são as matrizes de covariância genética aditiva e residual, respectivamente. No modelo [1], a estrutura causal define a escolha de quais entradas fora da diagonal de $\boldsymbol{\Lambda}$ são parâmetros livres e quais entradas são iguais a zero.

Com base em [1], o "modelo reduzido" é representado como:

$$\mathbf{y}_i = \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{X}_i\boldsymbol{\beta} + \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{u}_i + \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{e}_i. \qquad [2]$$

A distribuição do vetor $\mathbf{y}_i$ condicional aos parâmetros de local $\boldsymbol{\beta}$, $\mathbf{u}_i$ e $\boldsymbol{\Lambda}$, e a matriz de covariância residual $\boldsymbol{\Psi}_0$ é representada por:

$$\mathbf{y}_i \mid \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Psi}_0 \sim \mathbf{N}\left[\left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i\right), \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Psi}_0\left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{'-1}\right]. \qquad [3]$$

O modelo para *n* indivíduos é descrito como:

$$\mathbf{y} = \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \qquad [4]$$

e a distribuição conjunta dos vetores **u** e **e** é:

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim N\left\{\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G}_0 \otimes \mathbf{A} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}_0 \otimes \mathbf{I}_n\end{bmatrix}\right\}, \quad [5]$$

em que $\mathbf{y}$, $\mathbf{u}$ e $\mathbf{e}$ são, respectivamente, vetores de fenótipos, efeitos genético aditivos e resíduos do modelo, ordenados por característica e indivíduos dentro de característica; $\mathbf{X}$ e $\mathbf{Z}$ são matrizes de incidência dos efeitos em $\boldsymbol{\beta}$ e $\mathbf{u}$ sobre $\mathbf{y}$, e $\mathbf{A}$ é a matriz de numeradores dos coeficientes de parentesco de Wright,. O modelo [4] pode ser reescrito como:

$$\left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad [6]$$

de modo que o modelo reduzido se torna:

$$\mathbf{y} = \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{X}\boldsymbol{\beta} + \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{Z}\mathbf{u} + \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{e}. \quad [7]$$

Desta forma,

$$p\left(\mathbf{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Psi}_0\right) \sim N\left\{\left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\right), \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\boldsymbol{\Psi}\left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{'-1}\right\}, \quad [8]$$

em que $\boldsymbol{\Psi} = \boldsymbol{\Psi}_0 \otimes \mathbf{I}_n$.

2.2 Seleção de estruturas causais recursivas

Como mencionado, em grande parte das aplicações anteriores de MEEs em genética quantitativa, estruturas causais pré-selecionadas foram utilizadas. Uma alternativa é implementar algoritmos que buscam por estruturas causais. Esta seção descreve um algoritmo que realiza tal busca para o modelo $\mathbf{y}_i = \boldsymbol{\Lambda}\mathbf{y}_i + \mathbf{e}_i$. A aplicação do algoritmo no contexto de modelos mistos é apresentada na próxima seção.

Estruturas causais recursivas são representadas por Gráficos Acíclicos Direcionados (GADs), que constituem em conjuntos de variáveis conectadas por linhas direcionadas (setas) que representam relacionamentos causais diretos. Uma trilha na estrutura causal é uma sequência de variáveis conectadas, desconsiderando-se a direção das setas que as conectam. Incondicionalmente, trilhas permitem fluxo de dependência entre as variáveis que se encontram nos seus extremos, a menos que exista um *collider* (variável para a qual setas de ambas as direções convergem, como *C* em $A \rightarrow C \leftarrow B$) na trilha. *Colliders* bloqueiam o fluxo de dependência na trilha, o que faz com que *A* e *B* sejam independentes na estrutura dada. Colocar-se condicionalmente a uma variável que não está nos extremos de uma trilha bloqueia o fluxo de dependência se esta variável não é um *collider* (e.g., condicionalmente a *C* em $A \rightarrow C \rightarrow B$, $A \leftarrow C \leftarrow B$ ou $A \leftarrow C \rightarrow B$, *A* e *B* se tornam independentes), ou permite o fluxo de dependência se esta variável é um *collider*. Considerando duas variáveis *A* e *B* em um GAD, elas são consideradas d-separadas condicionalmente a um subconjunto **S** de variáveis se não há trilhas que permitam fluxo de dependência entre *A* e *B* (i.e., não há trilhas entre *A* e *B* em um GAD tais que todos os *colliders* ou seus descendentes estejam em **S** e nenhum não-*collider* esteja em **S**). Sob algumas premissas (que serão discutidas posteriormente), d-separações na estrutura causal são refletidas como independências condicionais na distribuição de probabilidade conjunta dos dados. Isto pode ser explorado para recuperar uma estrutura causal, ou uma classe de estruturas equivalentes (estruturas causais que resultam em distribuições de probabilidade com as mesmas independências condicionais) a partir da distribuição conjunta dos dados (PEARL, 2000; SPIRTES et al., 2000).

O algoritmo IC (*Inductive Causation*; VERMA e PEARL, 1990; PEARL, 2000) pode ser utilizado para recuperar estruturas causais recursivas subjacentes (ou uma classe de estruturas observacionalmente equivalentes) a partir das associações observadas entre características. A busca tem base em uma série de perguntas a respeito de independências condicionais entre variáveis e na premissa de que tais independências são reflexos de d-separações no GAD subjacente. A informação de entrada do algoritmo é uma matriz de correlação a partir da qual dependências marginais e condicionais podem ser avaliadas. A saída do algoritmo é um gráfico parcialmente orientado que representa uma classe de estruturas causais equivalentes, o que geralmente consiste em importante restrição do espaço inicial de hipóteses causais possíveis. Gráficos parcialmente orientados são gráficos com linhas direcionadas e não direcionadas. Os últimos devem ser interpretados como relacionamentos diretos simétricos entre pares de variáveis, uma vez que a direção do relacionamento causal não está especificada.

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***. Variáveis conectadas em ***U*** são chamadas de variáveis adjacentes.

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.

O objetivo do primeiro passo do algoritmo é obter um gráfico que especifique pares de características que são diretamente conectadas, mas sem especificar direção causal. Vértices adjacentes em um gráfico não são d-separados condicionalmente a qualquer conjunto de vértices remanescentes. A consequência observacional é que estas variáveis não são estatisticamente independentes condicionalmente a qualquer conjunto de variáveis restantes.

O objetivo do segundo passo do algoritmo é orientar linhas, por intermédio da busca por vértices no gráfico no qual setas convergem de ambos os sentidos em uma trilha (*colliders*). Estruturas internas dos gráficos compostas por um *collider* que sofre influência causal de dois vértices não conectados entre si são chamadas *unshielded colliders* (SPIRTES et al., 2000), como em $A \rightarrow C \leftarrow B$. Em tal estrutura, as variáveis *A* e *B* são denominadas pais de *C*, e esta variável é chamada filha de *A* e *B*. Pais não-adjacentes de um *collider* são d-separados condicionalmente a pelo menos um conjunto de variáveis, mas não são d-separados se o *collider* pertence a este conjunto. A conseqüência observacional desta propriedade é a dependência probabilística entre pais não adjacentes condicionalmente a qualquer possível conjunto de variáveis que contenha o filho comum.

O terceiro passo consiste em orientar linhas adicionais de maneira que isso não resulte em um novo *collider* ou em um ciclo. Desta forma, se três variáveis em um gráfico semidirecionado hipotético resultante do passo 2 se apresentam conectados como em $A \rightarrow B - C$, a linha entre *B* e *C* deve ser orientada em direção a *C*, uma vez que a direção contrária resulta em um *collider* em *B* que não foi detectado no passo anterior. O mesmo deve ser feito para uma linha caso uma das direções resulte em um ciclo. Conhecimento prévio a respeito do sistema estudado podem ser incorporados no algoritmo, o que resulta em restrição adicional da saída pela proibição ou imposição da presença de uma linha ou por proporcionar orientações adicionais de linhas (SHIPLEY, 2002; SPIRTES et al., 2000).

A busca realizada pelo algoritmo IC assume conexão entre gráficos causais e distribuições de probabilidade geradas por modelos que os contêm. Esta conexão se estabelece com base na premissa da ausência de variáveis não observadas que exerçam influência causal sobre duas ou mais características consideradas no estudo (i.e., premissa de suficiência causal). Sem este premissa, a conexão entre d-separações e independências condicionais estatísticas pode se perder. Considere como exemplo um conjunto de variáveis observadas **O** e duas variáveis *A* e *B* que não estão conectadas na estrutura causal subjacente às variáveis em **O**. Se *A* e *B* têm uma causa comum latente, espera-se então que eles sejam dependentes condicionalmente a qualquer subconjunto de variáveis em **O**, o que resulta na inclusão equivocada de uma linha entre o par de variáveis mencionado no gráfico parcialmente direcionado selecionado pelo algoritmo IC.

No MEE apresentado em [1], os resíduos $\mathbf{e}_i$ representam os efeitos das causas desconhecidas dos fenótipos das características consideradas. A premissa de suficiência causal indica que a busca deveria ser feita em um conjunto de características que inclui cada causa comum de duas ou mais características que pertencem a este conjunto. Desta forma, sob esta premissa, os resíduos dos MEEs são construídos como independentes, o que por sua vez tem sido uma premissa em aplicações recentes de tais modelos em genética quantitativa (DE LOS CAMPOS et al., 2006a; DE MATURANA et al., 2009; HERINGSTAD et al., 2009).

As perguntas que constituem o algoritmo IC exigem a declaração de pares de variáveis como condicionalmente dependentes ou condicionalmente independentes. Esta decisão tem como base correlações parciais estimadas de uma amostra. Por este motivo, a incerteza a respeito das correlações parciais deve ser considerada no processo de decisão. Sob abordagem frequentista, estas decisões estatísticas podem ser feitas com teste da hipótese de nulidade correspondente a uma correlação parcial igual a zero. Sob abordagem Bayesiana, estas decisões podem ser feitas utilizando intervalos HPD (*Highest Posterior Density*) para as correlações parciais.

2.3 Busca de estruturas causais no contexto de modelos mistos

Na metodologia descrita na seção anterior, os resíduos dos modelos são considerados como independentes e os efeitos recursivos são utilizados para modelar (interpretar) padrões de covariabilidade entre variáveis observadas. Entretanto, em MEEs mistos, os padrões de covariabilidade entre fenótipos podem ser explicados tanto por ligações causais entre características quanto por razões de origem genética. Em outras palavras, efeitos genéticos aleatórios correlacionados podem ser fontes de confundimento se a seleção de estruturas causais tem base na distribuição conjunta dos fenótipos, mesmo se os resíduos são considerados como independentes entre si.

Na FIGURA 2.1 são ilustrados cenários nos quais há recursividade entre fenótipos 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). A conexão entre a estrutura causal entre fenótipos e sua distribuição conjunta não ocorre em um cenário no qual efeitos genéticos são variáveis ocultas não controladas. Neste caso, y_1 e y_2 não são marginalmente independentes na FIGURA 2.1a, por causa da covariância entre u_1 e u_2. Pela mesma razão, aqueles fenótipos não são independentes condicionalmente a y_3 nas FIGURAS 2.1b, 2.1c e 2.1d.

Todavia, associações genéticas aditivas entre indivíduos podem ser exploradas como uma maneira de "controlar" o confundimento. Isto pode ser feito, por exemplo, se informações individuais de parentesco ou de marcadores estão disponíveis. Nesta abordagem, d-separações são refletidas como independências condicionais na distribuição conjunta dos fenótipos, após considerar os efeitos genéticos aditivos (i.e., na distribuição dos fenótipos condicional aos efeitos genéticos). Na FIGURA 2.1a, y_1 e y_2 são independentes dados os efeitos genéticos aditivos. Na FIGURAS 2.1b, 2.1c e 2.1d, as mesmas variáveis observáveis são independentes dados os efeitos genéticos aditivos e o fenótipo y_3. Um MEE que considera efeitos genéticos aditivos diretos pode ser representado como em [2], o que implica em

.

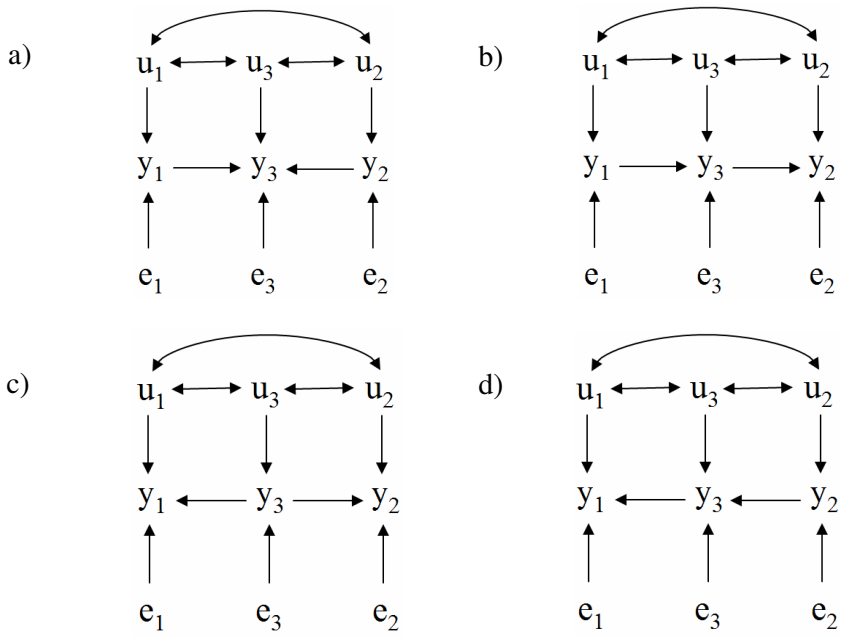


FIGURA 2.1 – Estruturas causais para três variáveis observadas (y_1, y_2 e y_3), com resíduos independentes (e_1, e_2 e e_3) e efeitos genéticos aditivos correlacionados (u_1, u_2 e u_3). Efeitos genéticos aditivos não-observáveis são fontes de confundimento se não estão considerados no modelo, e a distribuição conjunta das variáveis observadas não representariam adequadamente as independências condicionais esperadas com base na estrutura causal entre características.

$$Var(\mathbf{y}_i) = (\mathbf{I}_t - \boldsymbol{\Lambda})^{-1}\mathbf{G}_0(\mathbf{I}_t - \boldsymbol{\Lambda})^{'-1} + (\mathbf{I}_t - \boldsymbol{\Lambda})^{-1}\boldsymbol{\Psi}_0(\mathbf{I}_t - \boldsymbol{\Lambda})^{'-1}$$

Observa-se que $(\mathbf{I}_t - \boldsymbol{\Lambda})^{-1}\mathbf{G}_0(\mathbf{I}_t - \boldsymbol{\Lambda})^{'-1}$ e $(\mathbf{I}_t - \boldsymbol{\Lambda})^{-1}\boldsymbol{\Psi}_0(\mathbf{I}_t - \boldsymbol{\Lambda})^{'-1}$ são matrizes de covariância genética aditiva ($\mathbf{G}_0^*$) e residual ($\mathbf{R}_0^*$) obtidas de um modelo misto multicaracterísticas padrão que considera covariâncias entre efeitos genéticos aditivos e resíduos de diferentes características, mas não inclui relacionamento causal entre fenótipos (GIANOLA e SORENSEN; 2004, VARONA et al., 2007). A matriz de covariância de $\mathbf{y}_i$ pode ser reescrita como $Var(\mathbf{y}_i) = \mathbf{G}_0^* + \mathbf{R}_0^*$, e a matriz de covariância condicional aos efeitos genéticos aditivos pode ser representada como $Var(\mathbf{y}_i \mid \mathbf{u}_i) = (\mathbf{I}_t - \boldsymbol{\Lambda})^{-1}\boldsymbol{\Psi}_0(\mathbf{I}_t - \boldsymbol{\Lambda})^{'-1} = \mathbf{R}_0^*$. Desta forma, estimativas de $\mathbf{R}_0^*$ podem ser utilizadas para selecionar a estrutura causal entre os fenótipos. Esta matriz de covariância pode ser inferida por intermédio de métodos frequentistas ou Bayesianos. Sob análise Bayesiana, amostras da distribuição *a posteriori* de $\mathbf{R}_0^*$ podem ser obtidas, e estas amostras podem ser utilizadas para avaliar a incerteza associada a esta matriz, considerando simultaneamente a incerteza a respeito de todos os demais parâmetros incluídos no MMC reduzido. A seguir, descreve-se metodologia proposta para buscar estruturas causais no contexto de modelos mistos, utilizando amostras da distribuição *a posteriori* de $\mathbf{R}_0^*$ como informação de entrada do algoritmo IC:

1 – Ajustar MMC e obter amostras da distribuição *a posteriori* de $\mathbf{R}_0^*$.

2 – Aplicar o algoritmo IC sobre as amostras *a posteriori* de $\mathbf{R}_0^*$ para tomas as decisões estatísticas exigidas. Especificamente, para cada pergunta acerca da independência entre as variáveis *A* e *B* dado um conjunto de variáveis **S** e, implicitamente, os efeitos genéticos:

A – Obtenha a distribuição *a posteriori* da correlação parcial residual $\rho_{A,B|\mathbf{S}}$. Tais correlações parciais são funções de $\mathbf{R}_0^*$. Desta forma, suas distribuições *a posteriori* podem ser obtidas pelo cômputo da correlação correspondente para cada amostra utilizada para representar a distribuição *a posteriori* de $\mathbf{R}_0^*$.

B – Obtenha o intervalo HPD 95% para a distribuição *a posteriori* de $\rho_{A,B|\mathbf{S}}$.

C – Se o intervalo HPD contém 0, declare $\rho_{A,B|\mathbf{S}}$ como nulo. Caso contrário, declare *A* e *B* como condicionalmente dependentes.

3 – Ajustar MEE utilizando a estrutura causal selecionada (ou um membro da classe de estruturas observacionalmente equivalentes recuperadas pelo algoritmo IC).

3 EXEMPLO

Esta seção ilustra os conceitos previamente apresentados pela aplicação da metodologia proposta em dados simulados. O processo de geração dos dados e os modelos utilizados para inferência são descritos, e em seguida os resultados são apresentados e discutidos. A análise foi realizada pela utilização de programa escrito na linguagem R (R DEVELOPMENT CORE TEAM, 2008), o qual pode ser disponibilizado pelos autores mediante solicitação.

3.1 Processo de geração dos dados

Observações para 1800 indivíduos foram geradas a partir de um modelo recursivo com cinco características, com base na estrutura causal acíclica utilizada por SHIPLEY (1997) (FIGURA 2.2), com resíduos independentes. Adicionalmente, assumiu-se que as variáveis observadas sofrem influência de efeitos genéticos aditivos correlacionados simulados para 300 linhagens endogâmicas, com 6 indivíduos por linhagens.

O MEE utilizado para gerar os dados pode ser representado por:

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

em que y_{ijk} e e_{ijk} são fenótipos e efeitos residuais para a característiva *j* (*j=1,..,5*) observada no indivíduo *i* que pretence à linhagem *k*; μ_j é o valor médio da característica *j*; u_{jk} é o efeito genético aditivo atribuído à linhagem *k* para a característica *j*, e $\lambda_{jj'}$ é o valor modificado na característica *j* com respeito a *j'*.

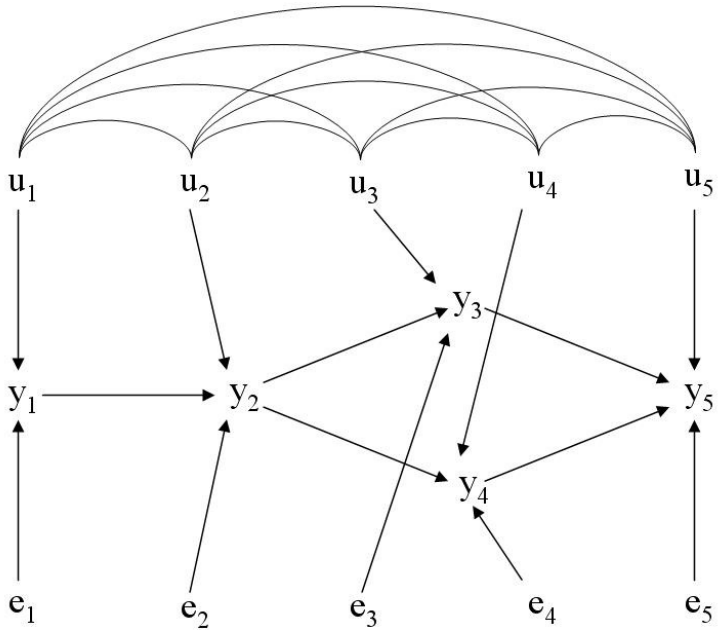


FIGURA 2.2 – Estrutura causal do modelo do qual os dados foram simulados; y_j é uma observação registrada para a característica *j*, u_j é o efeito genético aditivo que contribui para a caraterística *j* e e_j é o resíduo do modelo associado à característica *j*. Arcos conectando *u*'s representam correlações genéticas. A estrutura causal é adaptada de SHIPLEY (1997).

Efeitos genéticos aditivos atribuídos a 300 linhagens foram simulados como 50 grupos não relacionados de seis linhas derivadas de irmão completos. Desta forma, a matriz do numerador dos coeficientes de parentesco de Wright (**A**) entre linhas foi construída como bloco diagonal, na qual cada bloco consiste de uma matriz 6x6 em que as entradas fora de diagonal são iguais a 0,5 e os elementos da diagonal são iguais a 1. Vetores de efeitos genéticos aditivos diretos foram amostrados de $\mathbf{u} \sim N(0, \mathbf{G}_0 \otimes \mathbf{A})$ e $\mathbf{e} \sim N(0, \mathbf{\Psi}_0 \otimes \mathbf{I}_n)$, respectivamente.

Os valores dos parâmetros do modelo utilizados na simulação foram escolhidos arbitrariamente:

$$\mathbf{G}_0 = \begin{bmatrix} 100.000 & 47.373 & 20.283 & -38.839 & 9.773 \\ & 100.000 & 31.993 & -46.357 & -49.791 \\ & & 100.000 & 60.625 & -14.557 \\ & \text{simétrica} & & 100.000 & 6.490 \\ & & & & 100.000 \end{bmatrix},$$

$$\mathbf{\Psi}_0 = \begin{bmatrix} 200.000 & 0 & 0 & 0 & 0 \\ & 200.000 & 0 & 0 & 0 \\ & & 200.000 & 0 & 0 \\ & \text{simétrica} & & 200.000 & 0 \\ & & & & 200.000 \end{bmatrix},$$

$$\boldsymbol{\mu} = \begin{bmatrix} 100 \\ 110 \\ 90 \\ 180 \\ 50 \end{bmatrix}, \text{ e } \boldsymbol{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.35 & 0 & 0 & 0 \\ 0 & -0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & -0.4 & 0 \end{bmatrix}.$$

3.2 Inferências

As inferências se realizaram com base em um MEE com resíduos independentes como dado por $\boldsymbol{\Psi}_0$. Um modelo totalmente recursivo, em que todas as entradas abaixo da diagonal principal de $\boldsymbol{\Lambda}$ são parâmetros livres, foi utilizado para obter a distribuição *a posteriori* de $\mathbf{R}_0^*$. Posteriormente, um MEE foi ajustado com base na estrutura causal selecionada. A seguinte distribuição *a priori* conjunta foi assumida para parâmetros de local e de dispersão do modelo [7]:

$$p(\boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Psi}_0) = p(\boldsymbol{\Lambda})\, p(\boldsymbol{\beta})\, p(\mathbf{u} \mid \mathbf{G}_0)\, p(\mathbf{G}_0) \prod_{j=1}^{t} p(\psi_j)$$

$$\propto N(\mathbf{u} \mid 0, \mathbf{G}_0 \otimes \mathbf{A}) IW(\mathbf{G}_0 \mid \upsilon_{\mathbf{G}}, \mathbf{G}_0^{\bullet}) \prod_{j=1}^{t} Inv\text{-}\chi^2(\psi_j \mid \upsilon_{\psi}, s^2),$$

em que $N(\mathbf{u} \mid \mathbf{0}, \mathbf{G}_0 \otimes \mathbf{A})$ é uma normal multivariada centrada em **0** e com matriz de covariância $\mathbf{G}_0 \otimes \mathbf{A}$, $IW(\mathbf{G}_0 \mid \upsilon_{\mathbf{G}}, \mathbf{G}_0^{\bullet})$ é uma distribuição Wishart inversa com $\upsilon_{\mathbf{G}}$ graus de liberdade e matriz de escala $\mathbf{G}_0^{\bullet}$, $Inv\text{-}\chi^2(\psi_j \mid \upsilon_{\psi}, s^2)$ é uma distribuição de qui-quadrado invertida com parâmetro de escala, com υ_{ψ} graus de liberdade e parâmetro de escala s^2, e ψ_j é a variância dos resíduos do modelo para a característica *j*. Distribuições uniformes sem limites foram atribuídas para cada entrada de $\boldsymbol{\Lambda}$ considerada como parâmetro livre e para $\boldsymbol{\beta}$. Finalmente, υ_G, $\mathrm{G}_0^{\bullet}$, υ_{ψ} e s^2 foram considerados como hiperparâmetros conhecidos das distribuições *a priori*.

Deste modo, a distribuição *a posteriori* de todos os parâmetros do modelo é:

$$p(\boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Psi}_0 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Psi}_0)\, p(\mathbf{u} \mid \mathbf{G}_0)\, p(\mathbf{G}_0) \prod_{j=1}^{t} p(\psi_j).$$

Uma vez que a distribuição resultante não apresenta forma conhecida, foi utilizado um amostrador de Gibbs (GEMAN e GEMAN, 1984) para dela obter amostras, com base em distribuições condicionais completas. As derivações realizadas para obter estas distribuições são de uso corrente em análise Bayesiana com base em modelos lineares (e.g., SORENSEN e GIANOLA, 2002; GIANOLA e SORENSEN, 2004). *Descrição detalhada das condicionais completas usadas no amostrador de Gibbs é dada no* APÊNDICE *desta tese.*

Uma única cadeia de 40.000 iterações foi obtida para cada modelo. As primeiras 4000 iterações foram descartadas (*burn-in*), o que foi uma conduta conservadora em relação ao descarte de 312 iterações indicado pelo método de RAFTERY e LEWIS (1992), implementado no pacote BOA em R (SMITH, 2007). As 36000 iterações remanescentes foram consideradas como amostras da distribuição *a posteriori* dos parâmetros.

3.3 Modelo completamente recursivo

Com o objetivo de selecionar estruturas causais a serem utilizadas no ajuste de MEEs, uma matriz de covariância residual não-estruturada inferida pelo ajuste de um MMC aos dados é usada para buscar por padrões de independências condicionais que guiam a seleção de estruturas causais. Para estimar esta matriz, os dados foram analisados por intermédio de um modelo completamente recursivo, com matriz de covariância genética aditiva não-estruturada e matriz de covariância residual diagonal. Este modelo é equivalente em verossimilhança ao MMC (VARONA et al., 2007), de modo que ambos produzem matrizes de covariância similares, mas sob diferentes parametrizações. Deste modo, a distribuição *a posteriori* de uma matriz de covariância residual não estruturada de um MMC pode ser obtido pelo ajuste de um MEE completamente recursivo e pela transformação dos parâmetros amostrados em cada iteração. O modelo completamente recursivo foi reparametrizado da seguinte maneira:

$$\begin{aligned}\mathbf{y}_i &= \left(\mathbf{I}-\boldsymbol{\Lambda}_{fr}\right)^{-1}\boldsymbol{\mu}_l + \left(\mathbf{I}-\boldsymbol{\Lambda}_{fr}\right)^{-1}\mathbf{u}_k + \left(\mathbf{I}-\boldsymbol{\Lambda}_{fr}\right)^{-1}\mathbf{e}_i \\ &= \boldsymbol{\mu}_l^* + \mathbf{u}_k^* + \mathbf{e}_i^*,\end{aligned}$$

com:

$$var\left(\mathbf{e}_i^*\right) = \mathbf{R}_0^* = \left(\mathbf{I}-\boldsymbol{\Lambda}_{fr}\right)^{-1}\boldsymbol{\Psi}_{fr}\left(\mathbf{I}-\boldsymbol{\Lambda}_{fr}\right)^{\prime-1},$$

em que $\boldsymbol{\Lambda}_{fr}$ é uma matriz com coeficientes estruturais de um modelo completamente recursivo, $\mathbf{R}_0^*$ é um matriz de covariâncias não-estruturada (i.e., aquela de um MMC), e $\boldsymbol{\Psi}_{fr}$ é a matriz diagonal de variância residuais associada a um modelo completamente recursivo.

Os seguintes valores foram atribuídos aos hiperparâmetros no ajuste do modelo completamente recursivo: $s_{fr}^2 = 250$ e $\upsilon_{\boldsymbol{\Psi}fr} = 3$ para cada entrada da diagonal de $\boldsymbol{\Psi}_{fr}$, $\upsilon_{\mathbf{G}} = 7$ e

$$\mathbf{G}_0^{\bullet} = \begin{bmatrix} 150 & 0 & 0 & 0 & 0 \\ 0 & 150 & 0 & 0 & 0 \\ 0 & 0 & 150 & 0 & 0 \\ 0 & 0 & 0 & 150 & 0 \\ 0 & 0 & 0 & 0 & 150 \end{bmatrix}.$$

3.4 Inferência da estrutura causal

A distribuição *a posteriori* de $\mathbf{R}_0^*$ foi utilizada para selecionar estruturas causais recursivas entre fenótipos, por intermédio do algoritmo IC. Correlações parciais foram consideradas nulas se o valor 0 se apresentava dentro dos intervalos HPD 95% correspondentes. A correlação parcial entra características *A* e *B* condicional a um conjunto de características **S** é representada por $\rho_{A,B|\mathbf{S}}$. Estas correlações parciais são condicionais não apenas aos fenótipos em **S**, mas também aos efeitos genéticos aditivos. Estes serão subsequentemente omitidos.

3.5 Modelos de equações estruturais sob a estrutura causal selecionada

De acordo com a estrutura causal escolhida, dentre aquelas na classe de estruturas equivalentes selecionados pelo algoritmo IC, entradas apropriadas de $\mathbf{\Lambda}$ foram tratados como parâmetros livres de valor desconhecido para ajustar o MEE correspondente aos dados simulados. Valores dos hiperparâmetros usados para as distribuições *a priori* foram os mesmos utilizados para ajustar o modelo completamente recursivo.

3.6 Resultados

A média *a posteriori* da matriz $\mathbf{R}_0^*$, obtida pela transformação de $\mathbf{\Psi}_{fr}$ foi

$$\begin{bmatrix} 195.588 & 91.008 & 34.821 & -45.990 & 49.147 \\ & 243.162 & 86.383 & -119.874 & 106.481 \\ & & 233.320 & -36.905 & 192.070 \\ & \text{simétrica} & & 269.527 & -131.465 \\ & & & & 411.489 \end{bmatrix}.$$

A aplicação do primeiro passo do algoritmo IC em amostras da distribuição *a posteriori* de $\mathbf{R}_0^*$ resultou no gráfico não-direcionado ilustrado na FIGURA 2.3a. Pares de variáveis conectadas por linhas não apresentaram correlações de parciais nulas, seja qual fosse o conjunto de características remanescentes considerado em **S** (como ilustrado na FIGURA 2.4 para o par $[y_1,y_2]$, e nas FIGURAS SUPLEMENTARES 2.S1 a 2.S4 para os demais pares conectados), para os quais os intervalos HPD dos parâmetros não apresentaram o valor 0. Por outro lado, correlações parciais nulas foram encontradas para os pares remanescentes, como demonstrado na FIGURA 2.5.

O segundo passo do algoritmo IC resultou no gráfico parcialmente orientado apresentado na FIGURA 2.3b. Os conjuntos de três variáveis constituídas de pares de variáveis desconectadas com uma variável adjacente em comum no gráfico não-direcionado fornecido pelo primeiro passo foram $y_1 - y_2 - y_3$, $y_1 - y_2 - y_4$, $y_3 - y_2 - y_4$, $y_2 - y_3 - y_5$, $y_2 - y_4 - y_5$, e $y_3 - y_5 - y_4$. Todos eles foram declarados como não sendo *unshielded colliders* pelo algoritmo, com exceção de $y_3 - y_5 - y_4$. O algoritmo direcionou as linhas em direção a y_5, pois y_3 e y_4 não apresentaram correlações parciais nulas quando y_5 foi parte do grupo de características para os quais a correlação se colocava condicionalmente, como demonstrado na FIGURA 2.6.

O terceiro passo do algoritmo IC não modificou o gráfico parcialmente direcionado resultante do passo anterior, pois não houve direcionamento de linhas adicionais a ser realizado somente com base na distribuição dos resíduos.

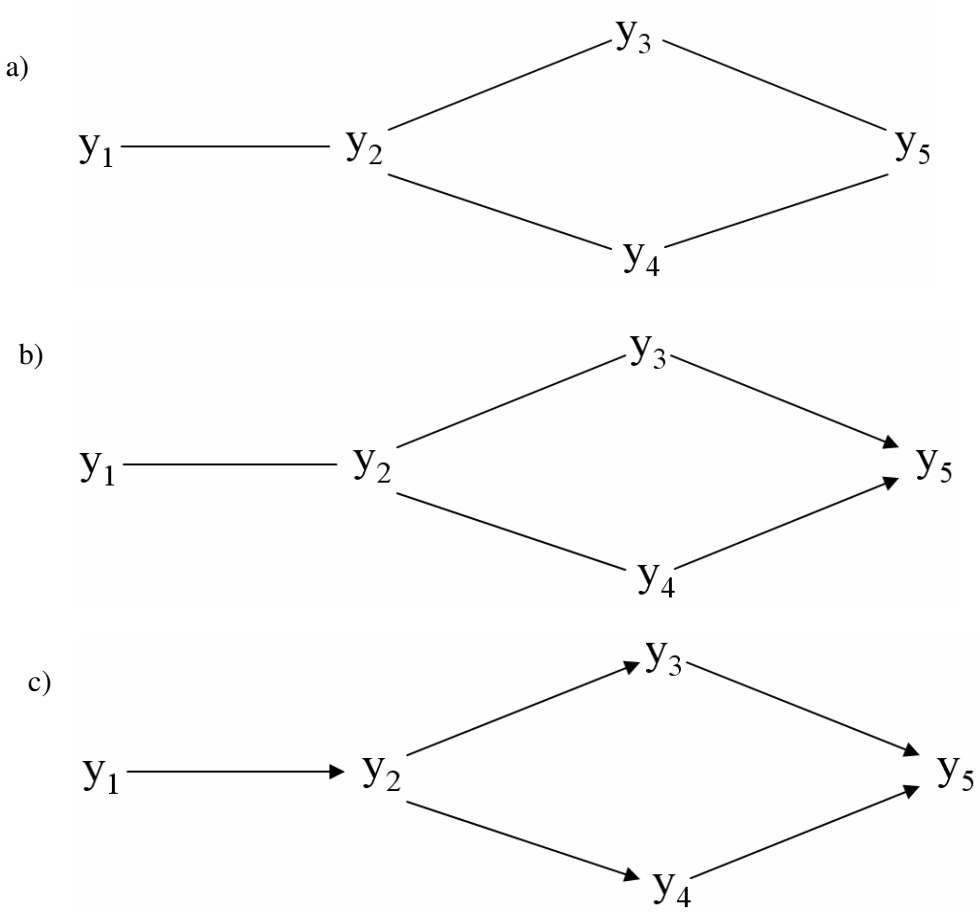


FIGURA 2.3 – Gráfico acíclico não-direcionado (a) resultante do passo 1 do algoritmo IC, e gráfico parcialmente orientado (b) recuperado pelo algoritmo IC. Conhecimento *a priori* da direção da conexão entre y_1 e y_2 (em direção a y_2) leva à escolha da estrutura (c) a partir da classe de estruturas representada por (b).

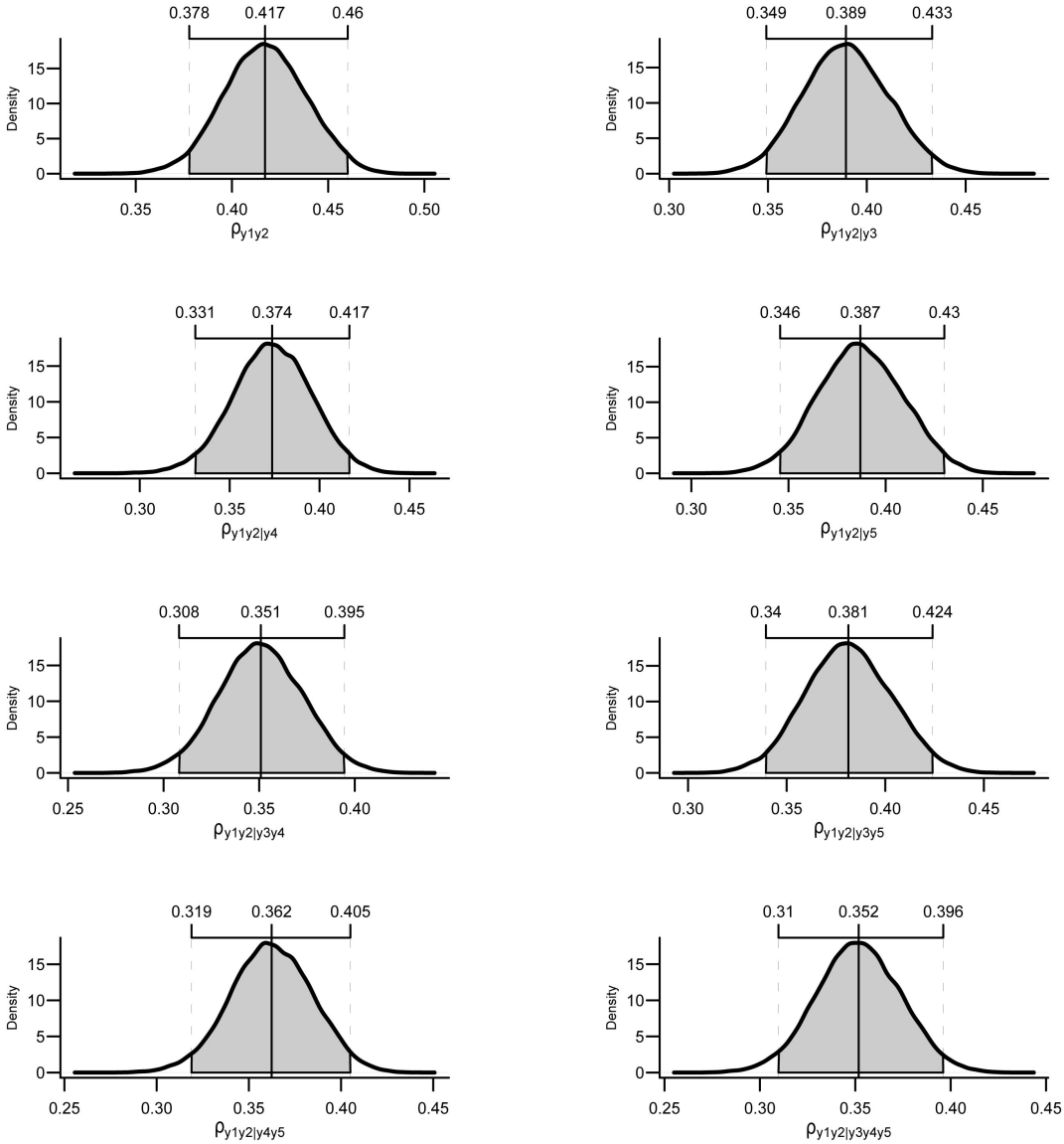


FIGURA 2.4 – Distribuições *a posteriori* e intervalos HPD de correlações totais e parciais entre y_1 e y_2, condicionalmente a cada conjunto possível de características remanescentes. A ausência de correlações parciais nulas levam o primeiro passo do algoritmo IC a inserir uma linha entre y_1 e y_2.

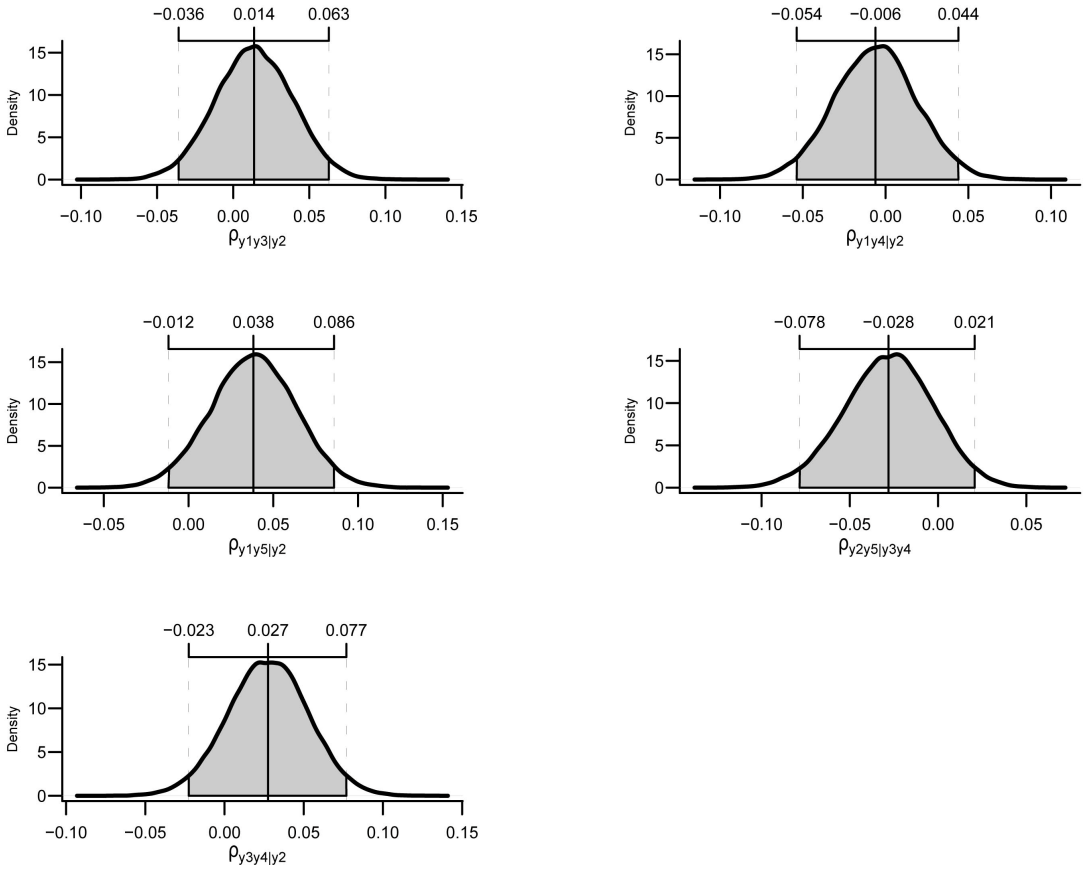


FIGURA 2.5 – Distribuições *a posteriori* e intervalos HPD de correlações parciais nulas que levam o algoritmo IC à remoção de linhas entre os pares de variáveis [y_1, y_3], [y_1, y_4], [y_1, y_5], [y_2, y_5], e [y_3, y_4].

A saída do algoritmo IC é um gráfico acíclico parcialmente orientado que representa uma classe de estruturas equivalentes, cada uma capaz de produzir o padrão de independências condicionais em $\mathbf{R}_0^*$. A estrutura causal verdadeira é um membro da classe selecionada, que contém apenas quatro estruturas alternativas (SHIPLEY, 2002). Entretanto, informações tais como precedência no tempo ou outras fontes de conhecimento *a priori* podem ser usadas para orientação adicional de linhas. Considere como exemplo o par de características [y_1,y_2]. A existência de uma linha entre o par é reconhecida pelo algoritmo IC independentemente de qualquer informação *a priori*, uma vez que a correlação entre elas, condicional a qualquer conjunto de variáveis remanescentes, foi sempre declarada como diferente de zero (FIGURA 2.4). Porém, o algoritmo não é capaz de reconhecer a direção desta conexão, já que o par não é parte de um *unshielded collider* e, na estrutura obtida pelo segundo passo, a orientação poderia ser feita em ambas as direções sem criar novos *colliders* ou ciclos. Por outro lado, se a precedência temporal de y_1 em relação a y_2 é conhecida, esta informação *a priori* não é o suficiente para impor ou proibir uma linha, mas é suficiente para orientar uma linha previamente detectada pelo algoritmo. Caso esta informação seja utilizada para direcionar a linha em direção a y_2, todas as linhas remanescentes seriam orientadas como na FIGURA 2.3c. Dado $y_1 \rightarrow y_2$, as linhas $y_2 - y_3$ e $y_2 - y_4$ devem ser orientadas em direção a y_3 e y_4, respectivamente (SHIPLEY, 2002). Qualquer outra configuração resultaria em *colliders* em y_2, o que representa uma estrutura causal que não é compatível com $\mathbf{R}_0^*$.

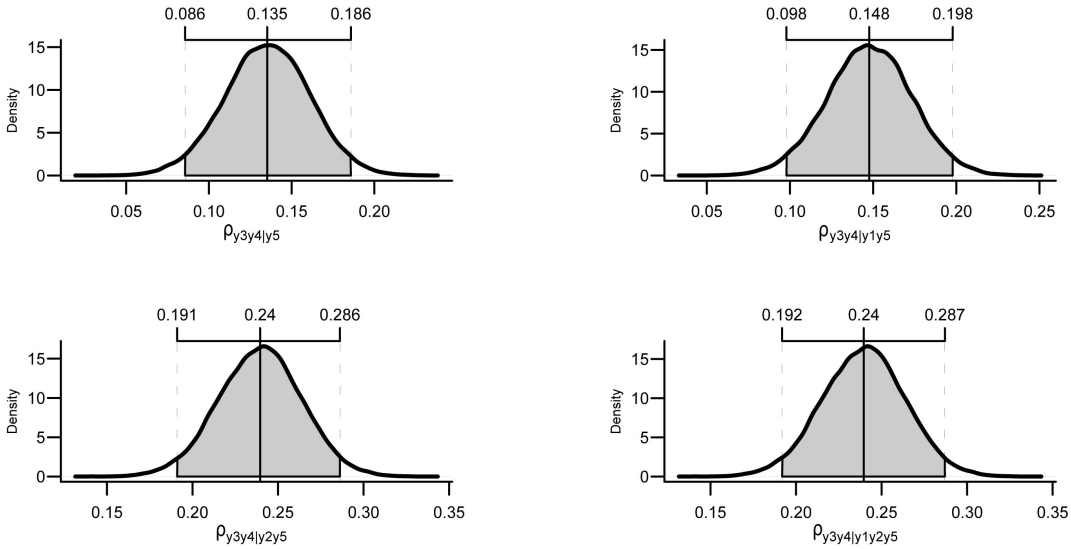


FIGURE 2.6 – Distribuições *a posteriori* e intervalos HPD da correlação parcial entre y_3 e y_4 dado cada subconjunto possível de variáveis remanescentes que incluem y_5. A ausência de correlações parciais nulas levam o algoritmo IC a declarar a trilha $y_3 - y_5 - y_4$ como *unshielded collider*, como apresentado na FIGURA 2.3b.

A TABELA SUPLEMENTAR 2.S.1 apresenta médias *a posteriori* e intervalos HPD 95% dos parâmetros de dispersão e coeficientes estruturais obtidos de um MEE ajustado condicionalmente à estrutura causal escolhida. Os valores paramétricos usados para simular os dados se encontraram dentro dos intervalos HPD mencionados previamente, sendo a covariância genética entre as características y_2 e y_5 a única exceção. Como exemplos adicionais, a abordagem proposta foi usada para estudar dados simulados de dois modelos baseados em estruturas causais diferentes. Os resultados obtidos são apresentados em MATERIAL SUPLEMENTAR.

4 DISCUSSÃO

O artigo de GIANOLA e SORENSEN (2004) gerou interesse na utilização de MEEs na análise de características quantitativas. Os referidos autores observaram o número imenso de possíveis estruturas causais a serem utilizadas, mesmo em estudos que envolviam conjuntos pequenos de características. WU et al. (2007) afirmaram que conhecimento *a priori* a respeito do sistema estudado poderia ser utilizado para reduzir o número de estruturas causais sob consideração, e este é o procedimento adotado na maioria das aplicações de MEEs na área de melhoramento animal. Porém, a qualidade desta estratégia depende da qualidade da teoria pré-existente que levou ao conhecimento *a priori* utilizado (SHIPLEY, 2002). Tipicamente, a escolha de uma estrutura ou de um pequeno conjunto de estruturas é feita com base em crenças *a priori*. Nos casos em que um pequeno conjunto de estruturas é escolhido, modelos ajustados com base nestas estruturas podem ser comparados por critérios tais como AIC ou BIC. Entretanto, as estruturas dos modelos comparados constituem um pequeno subconjunto de todas as estruturas possíveis, de modo que podem existir outros modelos que se ajustam tão bem, ou até melhor que aqueles pré-selecionados. É concebível realizar uma busca exaustiva, com base em critérios de comparação de modelos, o que exigiria o ajustamento de modelos utilizando cada uma das possíveis estruturas causais. Na maioria das situações, o número de possíveis estruturas causais é muito grande (de acordo com SHIPLEY (1997), existem aproximadamente 59000 estruturas causais acíclicas possíveis envolvendo apenas 5 características), o que faz com que esta abordagem não seja exequível.

TABELA 2.S.1 - Estimadores de Monte Carlo de médias *a posteriori* e intervalo HPD 95% de parâmetros provenientes do ajuste de MEE baseado na estrutura causal ilustrada na FIGURA 2.3c

Parâmetro[a]	Valores paramétricos	Média *a posteriori*	Intervalo HPD 95%
λ_{21}	0.5	0.4657	[0.4163 , 0.5141]
λ_{32}	0.35	0.3523	[0.3053 , 0.3981]
λ_{42}	-0.5	-0.4904	[-0.5332 , -0.4421]
λ_{53}	0.8	0.7601	[0.7129 , 0.8054]
λ_{54}	-0.4	-0.3853	[-0.4296 , -0.3418]
$\sigma^2_{g_1}$	100	105.7031	[82.9147 , 131.3272]
$\sigma_{g_1g_2}$	47.373	51.674	[32.1285 , 72.4094]
$\sigma_{g_1g_3}$	20.283	23.8453	[5.2715 , 42.5954]
$\sigma_{g_1g_4}$	-38.839	-33.6449	[-51.8859 , -15.6661]
$\sigma_{g_1g_5}$	9.773	12.4781	[-4.3035 , 29.4382]
$\sigma^2_{g_2}$	100	121.9420	[94.1627 , 149.4072]
$\sigma_{g_2g_3}$	31.993	19.2609	[-1.1349 , 40.1997]
$\sigma_{g_2g_4}$	-46.357	-49.8418	[-70.8852 , -29.9068]
$\sigma_{g_2g_5}$	-49.791	-30.6819	[-49.3386 , -12.3893]
$\sigma^2_{g_3}$	100	113.3788	[88.1318 , 139.4358]
$\sigma_{g_3g_4}$	60.625	51.7164	[32.8779 , 70.6185]
$\sigma_{g_3g_5}$	-14.557	-6.9528	[-25.1680 , 11.2717]
$\sigma^2_{g_4}$	100	103.173	[79.0714 , 128.3733]
$\sigma_{g_4g_5}$	6.49	2.6872	[-15.0006 , 20.0789]
$\sigma^2_{g_5}$	100	86.2032	[65.6027 , 107.5035]
ψ_1	200	195.4859	[181.7607 , 209.111]
ψ_2	200	200.5170	[186.682, 215.0271]
ψ_3	200	201.9552	[187.3972 , 216.0624]
ψ_4	200	209.3863	[194.8677 , 223.9679]
ψ_5	200	213.5152	[198.1786 , 228.4133]

[a] $\lambda_{jj'}$ é a modificação esperada da variável y_j com respeito à variável $y_{j'}$; $\sigma^2_{g_j}$ e $\sigma_{g_jg_{j'}}$ são, respectivamente, variância genética aditiva da característica y_j e covariância genética aditiva entre características y_j e $y_{j'}$; ψ_j é a variância residual da característica y_j.

A utilização do algoritmo IC resulta em vantagem em relação à abordagem normalmente empregada, uma vez que permite explorar o espaço de estruturas causais sem ter crenças *a priori* como base única. O algoritmo busca por uma classe de estruturas causais acíclicas equivalentes compatíveis com o padrão de independências condicionais observado. Esta busca não tem base nos critérios de seleção de modelos supracitados, mas espera-se que modelos com base nas estruturas selecionadas também apresentariam melhores *scores* para estes critérios. Esta expectativa ocorre porque o algoritmo fornece estruturas que melhor se ajustam ao padrão dado de independências condicionais e que são mínimas, i.e., elas resultam em MEEs com menor poder expressivo, ou menor flexibilidade, ao se ajustar à matriz de covariância (PEARL, 2000). Crenças *a priori* ainda podem ser usadas para escolher as estruturas mais interessantes dentro de uma classe de estruturas selecionadas. Dependendo do contexto, o algoritmo pode ser modificado para que alguns conhecimentos *a priori* dominem sobre alguns aspectos da saída do algoritmo padrão, impondo ou proibindo a presença de uma linha ou a direção de uma seta na estrutura causal (SPIRTES et al., 2000).

As premissas adotadas na utilização do algoritmo IC não são mais fortes do que aquelas consideradas nas aplicações recentes de MEEs em genética quantitativa. Nestas aplicações, além da imposição de estrutura às matrizes de covariância de variáveis aleatórias (geralmente são consideradas diagonais), as próprias estruturas causais são assumidas como conhecidas *a priori*. No presente artigo, é feita imposição de estrutura diagonal para a matriz de covariâncias residuais do MEE, como em DE LOS CAMPOS et al., 2006a; DE MATURANA et al., 2009 e HERINGSTAD et al., 2009. Dentro deste cenário, busca-se uma estrutura causal que é compatível com a distribuição conjunta dos dados.

Entretanto, o algoritmo IC não pode ser aplicado diretamente sobre a distribuição conjunta dos fenótipos, pois os efeitos genéticos confundem a busca. Para os dados simulados, a aplicação do algoritmo IC sobre a matriz de covariâncias não-condicionais entre fenótipos resultou em saída incorreta no passo 1: além das linhas não-direcionadas reconhecidas no exemplo descrito, o algoritmo também conectou os pares $[y_2,y_5]$ e $[y_1,y_5]$. Outro erro na busca foi a declaração da trilha $y_3 - y_2 - y_4$ como um *unshielded collider* ($y_3 \rightarrow y_2 \leftarrow y_4$) no passo 2. Uma importante contribuição do presente artigo foi propor uma metodologia que permita buscar por estruturas causais no contexto de modelos mistos aplicados em genética quantitativa. A metodologia proposta explora o fato de que é possível obter a distribuição dos fenótipos condicionalmente aos efeitos genéticos não-observados por intermédio do "controle" destes efeitos. Isto é realizado pelo ajuste de um MMC que considera tais efeitos.

No exemplo aqui ilustrado, todas as decisões estatísticas foram corretas e, por conseqüência, todas as d-separações foram corretamente detectadas. O problema se torna mais desafiador quando a força dos relacionamentos causais é menor. Após realizar simulação similar, porém reduzindo em 50% os valores paramétricos atribuídos aos coeficientes estruturais (resultados não apresentados), o mesmo gráfico não-direcionado foi obtido no primeiro passo do algoritmo. Entretanto, o segundo passo não conseguiu detectar o *unshielded collider* $y_3 \rightarrow y_5 \leftarrow y_4$, pois o intervalo HPD de $\rho_{y3y4|y5}$ foi [-0.003 , 0.098]. Como consequência, a variável y_5 foi declarada incorretamente como *non-collider*. Como esperado, a qualidade das decisões estatísticas decresce na medida em que as distribuições *a posteriori* das correlações parciais se tornam menos precisas (e.g., menor conjunto de dados ou sub-identificabilidade dos parâmetros do modelo). Cenários menos informativos podem levar a uma maior probabilidade de ignorar conexões entre variáveis, ou de adicionar linhas incorretamente, além de direcionar ou deixar de direcionar linhas de maneira inadequada. Após utilizar a abordagem proposta em dados simulados de acordo com a descrição na seção "**Processo de geração dos dados**", mas considerando apenas uma repetição para cada uma das 300 linhagens endogâmicas, a saída foi semelhante àquela obtida quando os coeficientes estruturais foram reduzidos em 50%. No entanto, a saída esperada foi obtida quando dois ou mais indivíduos foram simulados para cada linhagem.

Em cenários nos quais as distribuições *a posteriori* das correlações parciais são menos precisas, se as decisões estatísticas que constituem o algoritmo IC são tomadas com base em intervalos HPD de maior conteúdo (o que em conjunto com maior incerteza sobre parâmetros leva a intervalos maiores), o algoritmo perde eficiência na detecção de correlações parciais de menor magnitude. Uma vez que não há preferência entre proteção contra falhas em detectar correlações parciais diferentes de zero ou contra a declaração de uma correlação parcial de valor paramétrico igual a zero como não-nula, torna-se razoável diminuir o conteúdo do intervalo HPD usado para tomar as decisões quando as

distribuições mencionadas são pouco precisas (SHIPLEY, 2002). Não obstante, erros nas decisões estatísticas não implicam necessariamente em erros nas inferências. Por exemplo, se existem mais de um conjunto de variáveis para os quais um determinado par de variáveis se torna condicionalmente d-separado, e se alguns deles, mas não todos, refletem correlações parciais nulas na distribuição da amostra analisada, o gráfico não-direcionado resultante seria o mesmo, pois conexões são descartadas na presença de pelo menos uma correlação parcial nula (SPIRTES et al., 2000).

Efeitos genéticos e suas correlações são incluídos tanto no MEE recursivo quanto no MMC reduzido. Porém, a interpretação destes parâmetros se modifica de acordo com o modelo utilizado. Como exemplo, considere a correlação genética aditiva entre duas características hipotéticas *A* e *B*, em que *A* causa *B*. Sob modelo recursivo, a correlação genética aditiva representa a associação linear entre duas variáveis não observadas, cada uma delas afetando diretamente uma característica específica. Entretanto, esta não seria a única fonte de correlação genética entre as características sob um modelo que não expressa recursividade, porque existe uma associação indireta entre o efeito genético aditivo de *A* e o fenótipo *B*. Este efeito é mediado pelo fenótipo *A*. Sob o modelo recursivo, a correlação genética não abrange este efeito indireto. Por outro lado, a correlação genética aditiva sob MMC abrange todas as associações relativas a este efeito, não importando se estes efeitos são diretos ou não no contexto recursivo. Por este motivo, a correlação genética aditiva sob MMC poderia ser diferente de 0 mesmo se os efeitos genéticos aditivos são independentes no contexto recursivo (GIANOLA e SORENSEN, 2004). Correlações genéticas que não são mediadas pelo relacionamento causal entre fenótipos são fontes de confundimento que não permitem que a busca por estruturas causais seja realizada com base nas informações fenotípicas apenas. O estudo aqui apresentado propõe uma busca pela estrutura causal na matriz de covariância de **y** após levar em conta os efeitos que confundem a busca.

O controle das covariâncias genéticas aditivas recupera a conexão entre a probabilidade conjunta dos fenótipos e a estrutura causal. Em casos nos quais outros efeitos aleatórios importantes podem estar causando confundimento adicional, tais como efeitos maternos, a matriz $\mathbf{R}_0^*$ deve ser inferida de um modelo que também considera estes efeitos. Adicionalmente, modelos que não abrangem completamente as fontes de confundimento também podem levar a resultados inadequados. Como exemplo, no modelo reprodutor, os efeitos genéticos herdados do pai são considerados, enquanto os efeitos genéticos herdados da mãe são omitidos. Entretanto, alelos herdados da mãe são fontes de covariância fenotípica, e seus efeitos contribuiriam na matriz de covariância residual do modelo em questão. Desta forma, o modelo reprodutor falha na remoção do confundimento proveniente dos efeitos genéticos aditivos.

A exigência computacional da abordagem proposta aumenta na medida em que o conjunto de características analisadas aumenta. Para o exemplo apresentado neste artigo, 10 pares de características foram analisadas no passo 1 do algoritmo IC ($[y_1,y_2]$, $[y_1,y_3]$, $[y_1,y_4]$, $[y_1,y_5]$, $[y_2,y_3]$, $[y_2,y_4]$, $[y_2,y_5]$, $[y_3,y_4]$, $[y_3,y_5]$ e $[y_4,y_5]$). Para cada par, a dependência foi avaliada condicionalmente a oito diferentes subconjuntos de características remanescentes, como ilustrado para $[y_1,y_2]$ na FIGURA 2.4. Se um conjunto de quatro características fosse analisado, dependências relativas a seis pares de características seriam estudadas no primeiro passo, cada uma condicionalmente a quatro diferentes subconjuntos de características. Caso seis características fossem estudadas, dependências de 15 pares de variáveis observadas precisariam ser avaliadas no primeiro passo, cada uma condicionalmente a 16 diferentes subconjuntos. Adicionalmente, aumentar o número de características ou o tamanho do banco de dados aumenta o tempo computacional necessário para ajustar o modelo completamente recursivo. O comprimento da cadeia de Monte Carlo que carrega as amostras *a posteriori* da matriz de covariâncias residuais também afeta o tempo computacional necessário para a implementação do método proposto. Isso se deve não apenas ao maior tempo necessário para amostrar a cadeia, mas também porque correlações parciais devem ser calculadas para cada amostra da matriz de covariância para obter sua distribuição *a posteriori*. A análise descrita foi realizada em estação de trabalho Dell Precision T7400 workstation, com 16 Gb de memória e processadores CPU 64-bit dual-core Intel Xeon, utilizando sistema Linux Red Hat Enterprise. Foram necessárias ~6 horas para obter amostras *a posteriori* dos parâmetros do modelo completamente recursivo, e ~15 minutos para obter uma lista das conexões e *unshielded colliders* do gráfico parcialmente direcionado selecionado.

Modelos que permitem apenas descrever o relacionamento probabilístico entre características não são suficientes para a predição do efeito de intervenções. Modelos causais, por sua vez, podem ser utilizados para inferir como as probabilidades se modificariam após uma intervenção externa (PEARL, 2000). Os conceitos descritos neste estudo têm aplicação além da genética quantitativa. Por exemplo, as técnicas de busca por estruturas causais poderiam ser usadas para predição do efeito de gerenciamento de sistemas de produção ou decisões em veterinária nos quais a covariância genética atua como confundimento.

Agradecimentos: Esta pesquisa foi financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, e pela Wisconsin Agriculture Experiment Station, University of Wisconsin – Madison, USA. Os autores também agradecem os comentários por parte de Brian Yandell e Elias Chaibub Neto a respeito de uma versão anterior deste manuscrito.

5 REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H., 1973 Information theory and an extension of the maximum likelihood principle, pp. 267–291 in *2nd International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki. Publishing House of the Hungarian Academy of Sciences, Budapest.

ATEN, J. E., T. F. FULLER, A. J. LUSIS e S. HORVATH, 2008 Using genetic markers to orient the edges in quantitative trait networks: The NEO software. BMC Syst. Biol. **2:** 34

CHAIBUB NETO, E., T. C. FERRARA, A. D. ATTIE, e B. S. YANDELL, 2008 Inferring causal phenotype networks from segregating populations. Genetics **179:** 1089-1100.

CHEN L. S., F. EMMERT-STREIB e J. D. STOREY, 2007 Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biology **8**: R219.

DE LOS CAMPOS, G., D. GIANOLA , P. BOETTCHER e P. MORONI, 2006a A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. J. Anim. Sci. **84:** 2934-2941.

DE LOS CAMPOS, G., D. GIANOLA e B. HERINGSTAD, 2006b A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows. J. Dairy Sci. **89:** 4445-4455.

GEMAN, S. e D. GEMAN, 1984 Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6:** 721–741.

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

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

HERINGSTAD, B., X.-L. WU e D. GIANOLA, 2009 Inferring relationships between health and fertility in Norwegian red cows using recursive models. J. Dairy Sci. **92**:1778–1784

KONIG, S., X. L. WU, D. GIANOLA, B. HERINGSTAD e H. SIMIANER, 2008 Exploration of relationships between claw disorders and milk yield in Holstein cows via recursive linear and threshold models. J. Dairy Sci. **91:**395–406.

LI, R., S. W. TSAIH, K. SHOCKLEY, I. M. STYLIANOU, J. WERGEDAL e B. PAIGEN, G. A. CHURCHILL, 2006 Structural model analysis of multiple quantitative traits. PLoS Genet **2**: e114.

LIU, B., A. DE LA FUENTE e I. HOESCHELE, 2008 Gene network inference via structural equation modeling in genetical genomics experiments. Genetics **178**: 1763-1776.

DE MATURANA, E. L., X. L. WU, D. GIANOLA, K. A. WEIGEL e G. J. M. ROSA, 2009 Exploring biological relationships between calving traits in primiparous cattle with a Bayesian recursive model. Genetics **181:**277–287.

PEARL, J., 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffman, San Mateo, CA.

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

R DEVELOPMENT CORE TEAM, 2008 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. http://www.R-project.org.

RAFTERY, A. E. e S. LEWIS, 1992 How Many Iterations in the Gibbs Sampler?, pp 763-773 in *Bayesian Statistics IV*, edited by. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith. Oxford University Press, Oxford.

SCHADT, E. E., J. LAMB, X. YANG, J. ZHU, S. EDWARDS *et al.*, 2005 An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. **37:** 710–717.

SCHWARZ, G., 1978 Estimating the dimension of a model. Ann. Stat.**6:** 461–464.

SHIPLEY, B., 1997 Exploratory path analysis with applications in ecology and evolution. Am. Nat. **149:** 1113-1138.

SHIPLEY, B., 2002 *Cause and Correlation in Biology*. Cambridge University Press, Cambridge/London/New York.

SMITH, B. J., 2007 BOA: An R package for MCMC output convergence assessment and posterior inference. J. Stat. Softw. **21:** 1-37.

SORENSEN, D. e D. GIANOLA, 2002 *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.

SPIRTES, P., C. GLYMOUR e R. SCHEINES, 2000 *Causation, Prediction and Search*, Ed. 2. MIT Press, Cambridge, MA.

VARONA, L., D. SORENSEN e R. THOMPSON, 2007 Analysis of litter size and average litter weight in pigs using recursive model. Genetics **177:** 1791-1799.

VERMA, T. e J. PEARL, 1990 Equivalence and synthesis of causal models. Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence (July, Cambridge, MA), 220-227. Reprinted in Uncertainty in Artificial Intelligence, **6:** 255:268, Elsevier, Amsterdam.

WRIGHT, S., 1921 Correlation and causation. J. Agric. Res. **201:** 557–585.

WU, X.-L., B. HERINGSTAD, Y. M. CHANG, G. DE LOS CAMPOS e D. GIANOLA, 2007 Inferring relationships between somatic cell score and milk yield using simultaneous and recursive models. J. Dairy Sci. **90:** 3508-3521.

WU, X.-L., B. HERINGSTAD, D. GIANOLA, 2010 Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. J. Anim. Breed. Genet. **127:** 3–15.

6 MATERIAL SUPLEMENTAR – SIMULAÇÕES ADICIONAIS

Os modelos adotados para simular os conjuntos de dados adicionais (identificados como A1 e A2) são similares àquele descrito na seção "**Processo de geração dos dados**" com exceção da estrutura causal e dos valores dos coeficientes estruturais. Cada MEE pode ser representado como $\mathbf{y}_i = \boldsymbol{\Lambda}\mathbf{y}_i + \boldsymbol{\mu} + \mathbf{u}_i + \mathbf{e}_i$, apresentação similar ao modelo [1], com termo $\mathbf{X}_i\boldsymbol{\beta}$ substituído pelo vetor $\boldsymbol{\mu}$, que contém os valores médios para cada uma das 5 características. As matrizes $\boldsymbol{\Lambda}$ usadas para gerar A1 e A2 foram:

$$\boldsymbol{\Lambda}_{A1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0.3 & 0.7 & 0 & 0 \end{bmatrix}, \text{ e}$$

$$\boldsymbol{\Lambda}_{A2} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.6 & 0 & 0 & 0 & 0 \\ 0 & -0.4 & 0 & 0 & 0 \\ 0 & 0 & -0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 \end{bmatrix},$$

de acordo com as estruturas causais ilustradas na FIGURA 2.S5. Aos parâmetros remanescentes do modelo foram atribuídos os mesmos valores apresentados na seção "**Processo de geração dos dados**". Após ajustar um modelo completamente recursivo para A1 e A2, utilizando as mesmas distribuições *a priori* descritas na seção "**Modelo completamente recursivo**", e aplicar o algoritmo IC sobre a distribuição *a posteriori* das matrizes de covariância residual reparametrizadas, os gráficos selecionados foram aqueles apresentados na FIGURA 2.S6.

As simulações adicionais apresentaram diferentes resultados no que diz respeito ao número de GADs em cada classe de estruturas selecionada. Para A1, a estrutura foi completamente direcionada no segundo passo do algoritmo IC, pois todas as linhas detectadas no primeiro passo foram declaradas como partes de *unshielded colliders* no segundo passo. Por outro lado, apesar do algoritmo ter recuperado completamente o gráfico não-direcionado relativo a A2, não foi possível direcionar uma única linha sequer. Assumindo resíduos independentes, qualquer MEE que apresente as mesmas conexões e ausência de *colliders* resultaria no mesmo conjunto de independências condicionais observados na matriz de covariância residual obtido do MMC. Cada linha pode ser convertida em setas que apontam para qualquer lado em diferentes estruturas dentro da classe selecionada, de modo que todas as conexões se mantiveram como não-direcionadas.

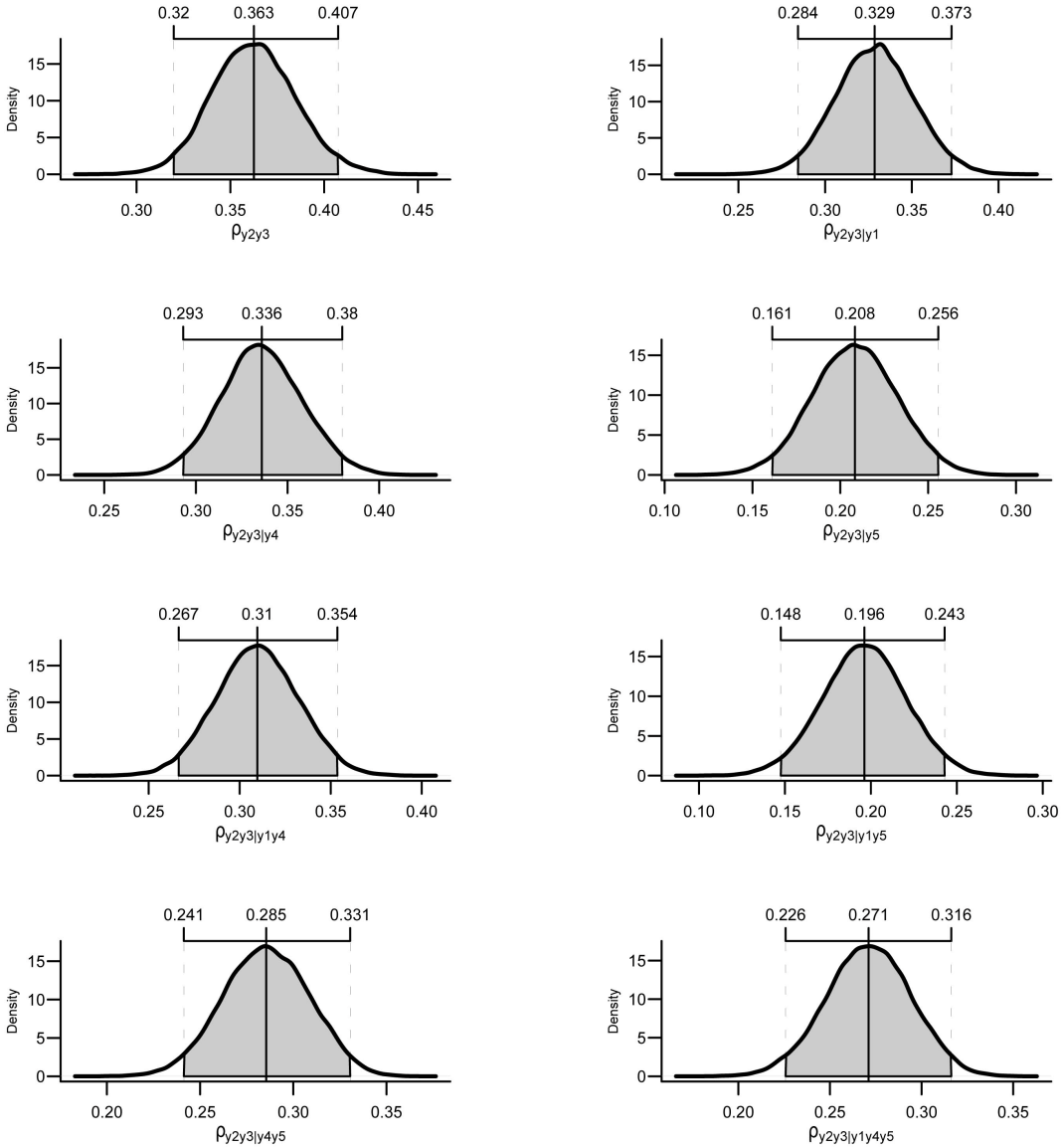


FIGURA 2.S1 – Distribuições *a posteriori* e intervalos HPD de correlações totais e parciais entre y_2 e y_3, condicionalmente a cada conjunto possível de características remanescentes. A ausência de correlações parciais nulas levam o primeiro passo do algoritmo IC a inserir uma linha entre y_2 e y_3.

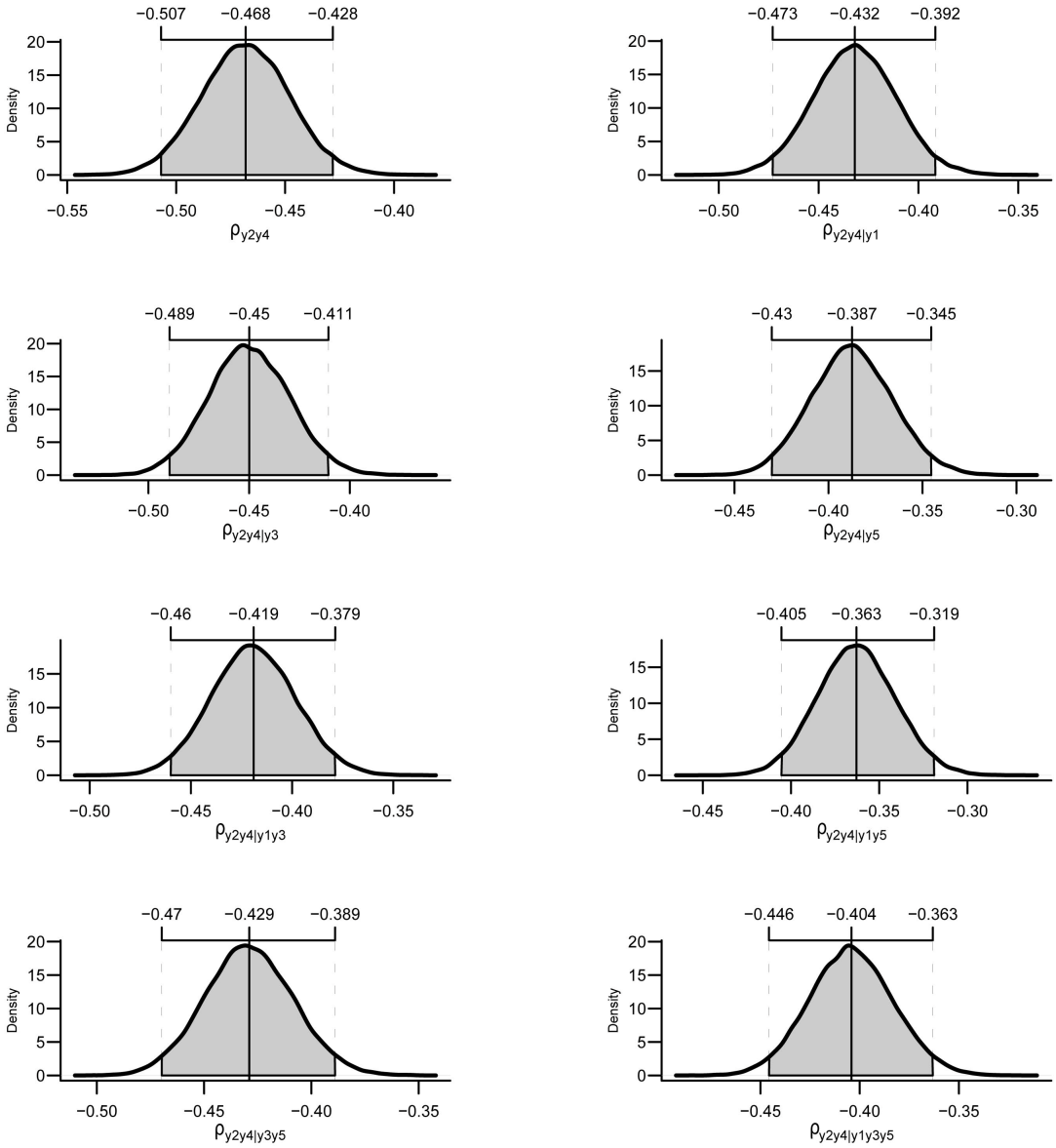


FIGURA 2.S2 – Distribuições *a posteriori* e intervalos HPD de correlações totais e parciais entre y_2 e y_4, condicionalmente a cada conjunto possível de características remanescentes. A ausência de correlações parciais nulas levam o primeiro passo do algoritmo IC a inserir uma linha entre y_2 e y_4.

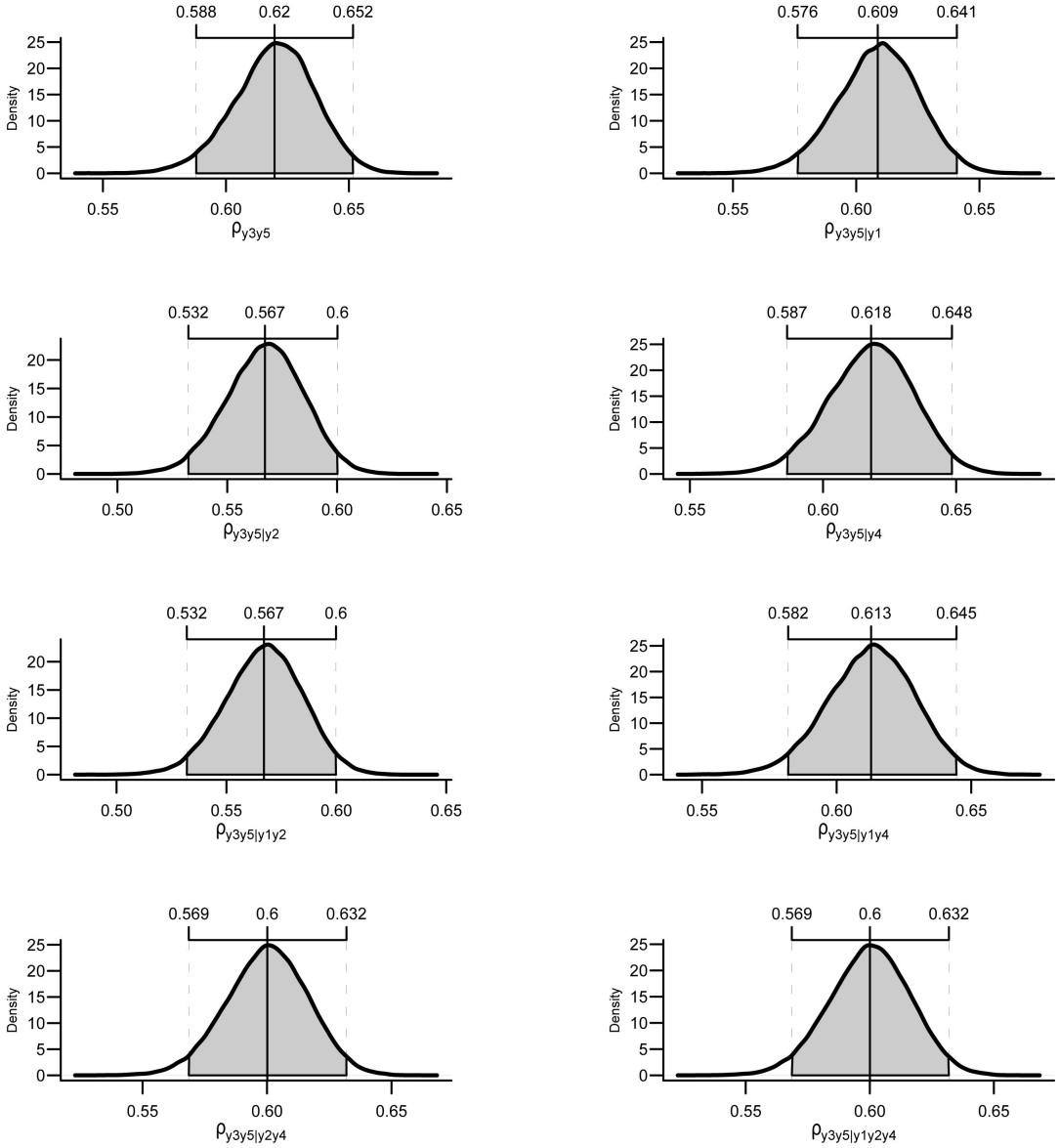


FIGURA 2.S3 – Distribuições *a posteriori* e intervalos HPD de correlações totais e parciais entre y_3 e y_5, condicionalmente a cada conjunto possível de características remanescentes. A ausência de correlações parciais nulas levam o primeiro passo do algoritmo IC a inserir uma linha entre y_3 e y_5.

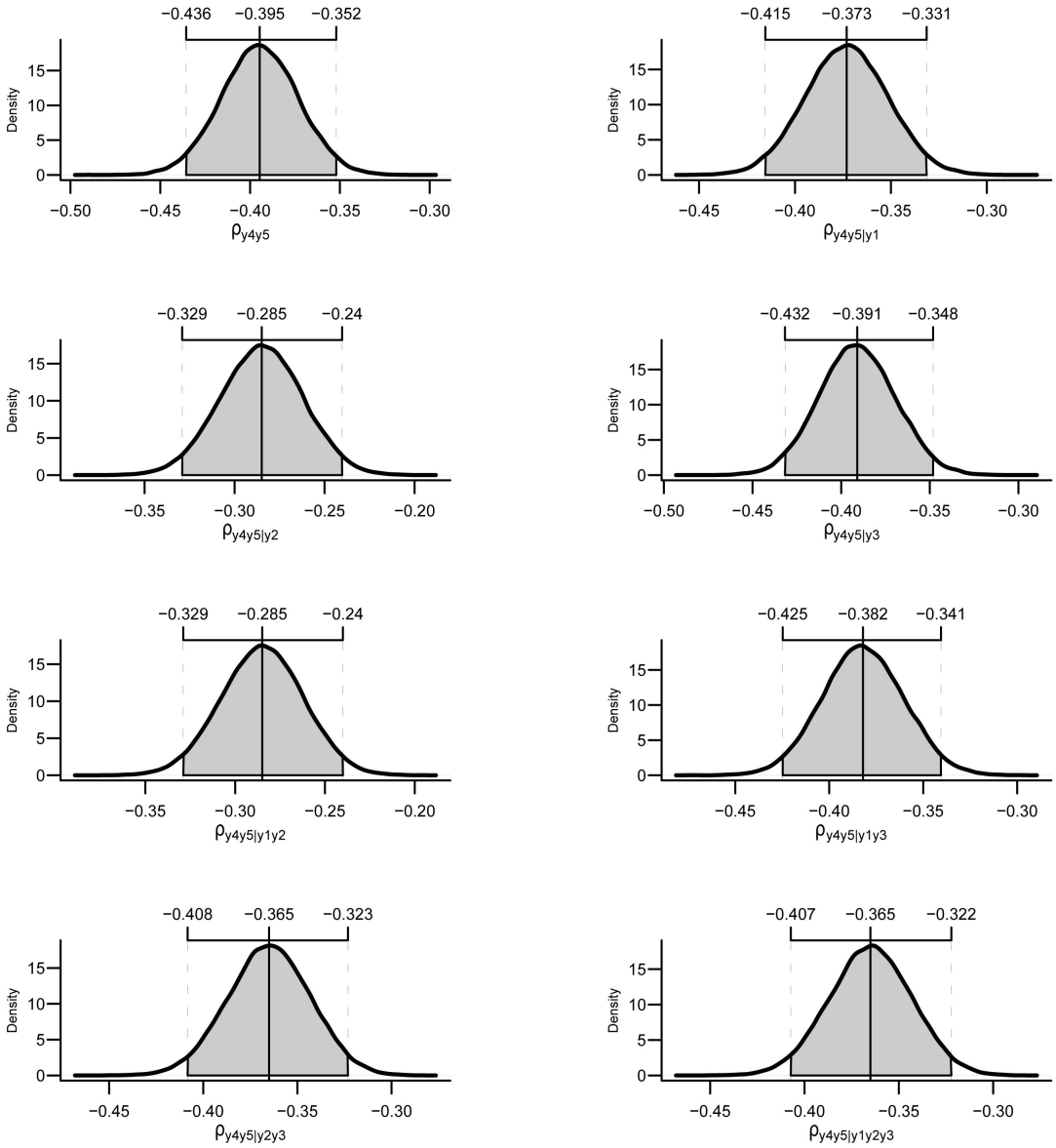
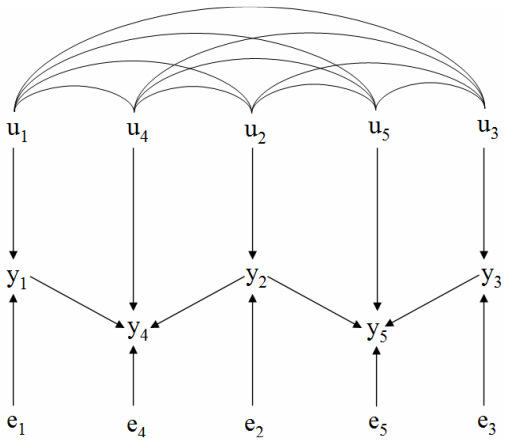


FIGURA 2.S4 – Distribuições *a posteriori* e intervalos HPD de correlações totais e parciais entre y_4 e y_5, condicionalmente a cada conjunto possível de características remanescentes. A ausência de correlações parciais nulas levam o primeiro passo do algoritmo IC a inserir uma linha entre y_4 e y_5.

a)



b)

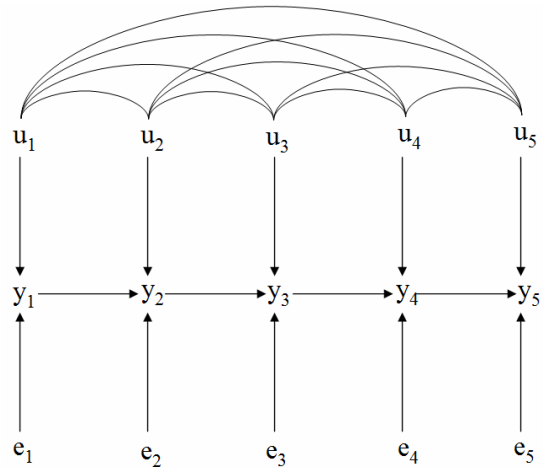


FIGURA 2.S5 – Estruturas causais dos modelos utilizados para amostrar os conjuntos de dados A1 (a) e A2 (b); y_j é uma observação registrada para a característica *j*, u_j é o efeito genético aditivo que contribui para a caraterística *j* e e_j é o resíduo do modelo associado à característica *j*. Arcos conectando *u*'s representam correlações genéticas.

a)

b)

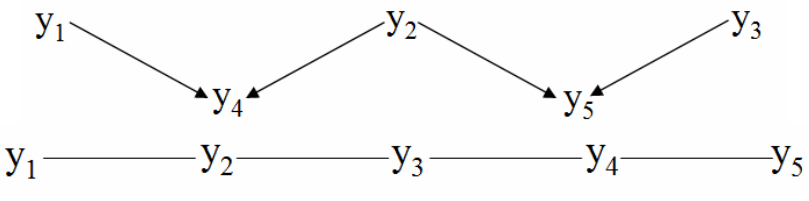


FIGURA 2.S6 – Gráfico acíclico direcionado (a) e gráfico não direcionado (b) provenientes da aplicação do algoritmo IC aos conjuntos de dados A1 e A2, respectivamente.

CONSIDERAÇÕES FINAIS

O desenvolvimento de algoritmos para seleção de estruturas causais é parte de um conjunto de metodologias propostas com o intuito de realizar inferência causal. Esta é uma área cuja importância é cada vez mais reconhecida. O presente trabalho inaugura o empreendimento destes esforços no contexto dos modelos vigentes nos estudos em Genética e Melhoramento Animal. Dentre os algoritmos disponíveis, o algoritmo IC foi escolhido como o mais adequado para este estudo inicial, pela sua facilidade de implementação e simplicidade de apresentação.

O estudo aqui apresentado indica a necessidade de investigação adicional sobre diversos tópicos. Dentre eles, se destacam maior conhecimento sobre as consequências da recursividade em um grupo de características consideradas em um programa de seleção, a utilização destes algoritmos no gerenciamento de sistemas de produção e a utilização de informações genômicas para aumentar a eficiência da seleção de estruturas causais. Cada um destes tópicos está sendo estudado pelo grupo que trabalhou na pesquisa aqui apresentada.

APÊNDICE

Dada uma estrutura causal específica, um MEE pode ser representado por $\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u} + \mathbf{e}_i$, como descrito no CAPÍTULO 2. As entradas específicas de $\mathbf{\Lambda}$ que se associam a parâmetros estruturais livres são condicionais à estrutura causal recursiva utilizada. Considerando *t* características observáveis para *n* animais, o modelo pode ser representado como $\mathbf{y} = (\mathbf{\Lambda} \otimes \mathbf{I}_n)\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$, tal que

$$\mathbf{y} = \left[\mathbf{I}_{tn} - (\mathbf{\Lambda} \otimes \mathbf{I}_n)\right]^{-1}\mathbf{X}\boldsymbol{\beta} + \left[\mathbf{I}_{tn} - (\mathbf{\Lambda} \otimes \mathbf{I}_n)\right]^{-1}\mathbf{Z}\mathbf{u} + \left[\mathbf{I}_{tn} - (\mathbf{\Lambda} \otimes \mathbf{I}_n)\right]^{-1}\mathbf{e}. \qquad [1]$$

e

$$p(\mathbf{y} \mid \mathbf{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{\Psi}_0) \sim N\left\{\left[\mathbf{I}_{tn} - (\mathbf{\Lambda} \otimes \mathbf{I}_n)\right]^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}),\right.$$
$$\left.\left[\mathbf{I}_{tn} - (\mathbf{\Lambda} \otimes \mathbf{I}_n)\right]^{-1}\mathbf{\Psi}\left[\mathbf{I}_{tn} - (\mathbf{\Lambda} \otimes \mathbf{I}_n)\right]^{\prime -1}\right\},$$

como anteriormente descrito.

As seguintes distribuições *a priori* foram assumidas para os parâmetros de local e de dispersão do modelo:

$$p(\mathbf{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{\Psi}_0) = p(\mathbf{\Lambda})\, p(\boldsymbol{\beta})\, p(\mathbf{u} \mid \mathbf{G}_0)\, p(\mathbf{G}_0) \prod_{j=1}^{t} p(\psi_j),$$
$$\propto N(\mathbf{u} \mid 0, \mathbf{G}_0 \otimes \mathbf{A})\, IW(\mathbf{G}_0 \mid \upsilon_{\mathbf{G}}, \mathbf{G}_0^{\bullet}) \prod_{j=1}^{t} Inv\text{-}\chi^2(\psi_j \mid \upsilon_{\psi}, s^2),$$

o que resulta na distribuicão *a posteriori*

$$p(\mathbf{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{\Psi}_0 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{\Psi}_0)\, p(\mathbf{u} \mid \mathbf{G}_0)\, p(\mathbf{G}_0) \prod_{j=1}^{t} p(\psi_j)$$

conforme anteriormente apresentado.

A distribuição resultante não apresenta forma conhecida, de modo que um amostrador de Gibbs (GEMAN e GEMAN, 1984) foi empregado para dela obter amostras, com base em distribuições condicionais completas. As derivações realizadas são de uso corrente em análise Bayesiana com base em modelos lineares (e.g., SORENSEN e GIANOLA, 2002; GIANOLA and SORENSEN, 2004).

Assumindo $\mathbf{\Lambda}$ como conhecido, a seguinte transformação pode ser realizada:

$$\mathbf{y}^* = \left(\mathbf{I}_{tn} - (\mathbf{\Lambda} \otimes \mathbf{I}_n)\right)\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

tal que, dado $\mathbf{y}^*$, o lado direito do MEE modificado é exatamente igual ao lado direito de um MMC. Desta forma, as distribuições condicionais completas *a posteriori* de $\boldsymbol{\beta}$, $\mathbf{u}$ e $\mathbf{G}_0$ são similares àquelas derivadas para estes parâmetros para MMCs, com $\mathbf{y}$ substituído por $\mathbf{y}^* = \left(\mathbf{I}_{tn} - (\mathbf{\Lambda} \otimes \mathbf{I}_n)\right)\mathbf{y}$. Isto se aplicaria à matriz de covariância residual $\mathbf{\Psi}_0$ se ela fosse considerada como não-estruturada, ao invés de diagonal. Neste caso, a distribuição *a priori* de $\mathbf{\Psi}_0$ deveria ser substituída por:

$$\boldsymbol{\Psi}_0 \mid \upsilon_{\boldsymbol{\Psi}}, \boldsymbol{\Psi}_0^{\bullet} \sim IW(\upsilon_{\boldsymbol{\Psi}}, \boldsymbol{\Psi}_0^{\bullet}),$$

em que $\upsilon_{\mathbf{R}}$ e $\boldsymbol{\Psi}_0^{\bullet}$ são hiperparâmetros.

Considerando $\mathbf{M} = \left[\sum_{j=1}^{t} {}^{+}\mathbf{X}_j \quad \sum_{j=1}^{t} {}^{+}\mathbf{Z}_j \right]$, para j=1,…,t características, e $\boldsymbol{\Omega} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1} \end{bmatrix}$, as equações de modelos mistos podem ser representadas como $\mathbf{C}\hat{\boldsymbol{\theta}} = \mathbf{t}$, em que $\boldsymbol{\theta}' = [\boldsymbol{\beta}' \quad \mathbf{u}']$, $\boldsymbol{\Psi} = \boldsymbol{\Psi}_0 \otimes \mathbf{I}_n$, $\mathbf{C} = \mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega}$ e $\mathbf{t} = \mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{y}^*$ (SORENSEN e GIANOLA, 2002).

A distribuição condicional completa *a posteriori* dos parâmetros de local pode ser escrita como:

$$\begin{aligned}
p(\boldsymbol{\beta}, \mathbf{u} \mid \boldsymbol{\Lambda}, \mathbf{G}_0, \boldsymbol{\Psi}_0, \mathbf{y}) &= p(\boldsymbol{\beta}, \mathbf{u} \mid \mathbf{G}_0, \boldsymbol{\Psi}_0, \mathbf{y}^*) \\
&= p(\boldsymbol{\theta} \mid \mathbf{G}_0, \boldsymbol{\Psi}_0, \mathbf{y}^*) \\
&\propto p(\mathbf{y}^* \mid \boldsymbol{\theta}, \boldsymbol{\Psi}_0)\, p(\mathbf{u} \mid \mathbf{G}_0) \\
&\propto exp\left[-\frac{1}{2}(\mathbf{y}^* - \mathbf{M}\boldsymbol{\theta})' \boldsymbol{\Psi}^{-1} (\mathbf{y}^* - \mathbf{M}\boldsymbol{\theta}) \right] exp\left[-\frac{1}{2}\mathbf{u}' (\mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1}) \mathbf{u} \right] \\
&\propto exp\left[-\frac{1}{2}(\mathbf{y}^* - \mathbf{M}\boldsymbol{\theta})' \boldsymbol{\Psi}^{-1} (\mathbf{y}^* - \mathbf{M}\boldsymbol{\theta}) \right] exp\left[-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\Omega}\boldsymbol{\theta} \right] \\
&\propto exp\left\{ -\frac{1}{2}\left[(\mathbf{y}^* - \mathbf{M}\boldsymbol{\theta})' \boldsymbol{\Psi}^{-1} (\mathbf{y}^* - \mathbf{M}\boldsymbol{\theta}) + \boldsymbol{\theta}'\boldsymbol{\Omega}\boldsymbol{\theta} \right] \right\} \\
&\propto exp\left\{ -\frac{1}{2}\left[\mathbf{y}^{*\prime}\boldsymbol{\Psi}^{-1}\mathbf{y}^* - 2\boldsymbol{\theta}'\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{y}^* + \boldsymbol{\theta}'\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M}\boldsymbol{\theta} + \boldsymbol{\theta}'\boldsymbol{\Omega}\boldsymbol{\theta} \right] \right\}
\end{aligned}$$

$$\propto exp\left\{ -\frac{1}{2}\left[\mathbf{y}^{*\prime}\boldsymbol{\Psi}^{-1}\mathbf{y}^* - 2\boldsymbol{\theta}'(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}'(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})\boldsymbol{\theta} \right] \right\}$$

$$\begin{aligned}
&\propto exp\left\{ -\frac{1}{2}\left[\mathbf{y}^{*\prime}\boldsymbol{\Psi}^{-1}\mathbf{y}^* - 2\boldsymbol{\theta}'(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}'(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})\boldsymbol{\theta} \right.\right. \\
&\left.\left. + \hat{\boldsymbol{\theta}}'(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}'(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})\hat{\boldsymbol{\theta}} \right] \right\} \\
&\propto exp\left\{ -\frac{1}{2}\left[\mathbf{y}^{*\prime}\boldsymbol{\Psi}^{-1}\mathbf{y}^* + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' (\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right.\right. \\
&\left.\left. - \hat{\boldsymbol{\theta}}'(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})\hat{\boldsymbol{\theta}} \right] \right\} \\
&\propto exp\left\{ -\frac{1}{2}\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' (\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right] \right\},
\end{aligned}$$

o que é reconhecida como a forma de uma distribuição normal multivariada com média $\hat{\boldsymbol{\theta}}$ e variância $(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{M} + \boldsymbol{\Omega})^{-1} = \mathbf{C}^{-1}$.

As distribuições condicionais de subconjuntos do vetor de parâmetros de local são $\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}, \mathbf{G}_0, \boldsymbol{\Psi}_0, \mathbf{y}^* \sim N(\tilde{\boldsymbol{\theta}}_i, \mathbf{C}_{i,i}^{-1})$, em que $\mathbf{C}_{i,i}\tilde{\boldsymbol{\theta}}_i = \left(\mathbf{M}'\boldsymbol{\Psi}^{-1}\mathbf{y}^*\right)_i - \mathbf{C}_{i,-i}\boldsymbol{\theta}_{-i}$. Esta distribuição é necessária para a implementação do amostrador de Gibbs em modo escalar.

Para *s* grupos de indivíduos geneticamente idênticos, a distribuição condicional completa *a posteriori* de $\mathbf{G}_0$ pode ser expressa como

$$
\begin{aligned}
p\left(\mathbf{G}_0 \mid \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Psi}_0, \mathbf{y}^*\right) &\propto p\left(\mathbf{u} \mid \mathbf{G}_0\right) p\left(\mathbf{G}_0\right) \\
&\propto \left|\mathbf{G}_0\right|^{-\frac{s}{2}} exp\left[-\frac{1}{2}\mathbf{u}'\left(\mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1}\right)\mathbf{u}\right]\left|\mathbf{G}_0\right|^{-\frac{1}{2}(\upsilon_{\mathbf{G}}+t+1)} \\
&\quad exp\left[-\frac{1}{2}tr\left(\mathbf{G}_0^{-1}\mathbf{G}_0^{\bullet -1}\right)\right] \\
&\propto \left|\mathbf{G}_0\right|^{-\frac{s}{2}} exp\left[-\frac{1}{2}tr(\mathbf{G}_0^{-1}\mathbf{S}_a)\right]\left|\mathbf{G}_0\right|^{-\frac{1}{2}(\upsilon_{\mathbf{G}}+t+1)} exp\left[-\frac{1}{2}tr\left(\mathbf{G}_0^{-1}\mathbf{G}_0^{\bullet -1}\right)\right] \\
&\propto \left|\mathbf{G}_0\right|^{-\frac{1}{2}(\upsilon_{\mathbf{G}}+s+t+1)} exp\left[-\frac{1}{2}tr\left(\mathbf{G}_0^{-1}\left(\mathbf{S}_a + \mathbf{G}_0^{\bullet -1}\right)\right)\right]
\end{aligned}
$$

O lado direito da expressão acima apresenta forma de um distribuição de Wishart inversa com parâmetros $\left(\mathbf{S}_a + \mathbf{G}_0^{\bullet -1}\right)^{-1}$ e $\upsilon_{\mathbf{G}} + s$, i. e. $\mathbf{G}_0 \mid \mathbf{b}, \mathbf{u}, \boldsymbol{\Psi}_0, \mathbf{y}^* \sim IW_t\left\{\left(\mathbf{S}_a + \mathbf{G}_0^{\bullet -1}\right)^{-1}, \upsilon_{\mathbf{G}} + s\right\}$, em que $\mathbf{S}_a = \begin{bmatrix} \mathbf{u}_1'\mathbf{A}\mathbf{u}_1 & \mathbf{u}_1'\mathbf{A}\mathbf{u}_2 & \cdots & \mathbf{u}_1'\mathbf{A}\mathbf{u}_k \\ \mathbf{u}_2'\mathbf{A}\mathbf{u}_1 & \mathbf{u}_2'\mathbf{A}\mathbf{u}_2 & \cdots & \mathbf{u}_2'\mathbf{A}\mathbf{u}_k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_k'\mathbf{A}\mathbf{u}_1 & \mathbf{u}_k'\mathbf{A}\mathbf{u}_2 & \cdots & \mathbf{u}_k'\mathbf{A}\mathbf{u}_k \end{bmatrix}$.

Para um MMC clássico, a condicional completa de $\boldsymbol{\Psi}_0$ seria:

$$
\begin{aligned}
p\left(\boldsymbol{\Psi}_0 \mid \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{y}^*\right) &\propto p\left(\mathbf{y}^* \mid \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Psi}_0\right) p\left(\boldsymbol{\Psi}_0\right) \\
&\propto IW_t\left[\left(\mathbf{S}_e + \boldsymbol{\Psi}_0^{\bullet -1}\right)^{-1}, \upsilon_{\boldsymbol{\Psi}} + n\right],
\end{aligned}
$$

em que $\mathbf{e} = \mathbf{y}^* - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}$ and $\mathbf{S}_e = \begin{bmatrix} \mathbf{e}_1'\mathbf{e}_1 & \mathbf{e}_1'\mathbf{e}_2 & \cdots & \mathbf{e}_1'\mathbf{e}_k \\ \mathbf{e}_2'\mathbf{e}_1 & \mathbf{e}_2'\mathbf{e}_2 & \cdots & \mathbf{e}_2'\mathbf{e}_k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{e}_k'\mathbf{e}_1 & \mathbf{e}_k'\mathbf{e}_2 & \cdots & \mathbf{e}_k'\mathbf{e}_k \end{bmatrix}$.

Como apresentado no capítulo 1, se $\mathbf{G}_0$ e $\boldsymbol{\Psi}_0$ não são matrizes estruturadas e se $\boldsymbol{\Lambda}$ apresenta parâmetros livres, os parâmetros do modelo não são identificáveis na função de verossimilhança (VARONA et al., 2007). Entretanto, se resíduos de diferentes características são considerados como mutuamente ortogonais, os parâmetros se tornam identificáveis para qualquer modelo recursivo. Neste cenário, um abordagem diferente deve ser seguida para derivar as condicionais completas *a posteriori* das variâncias residuais. A distribuição *a priori* de cada variância residual seria representada por $\psi_j \mid \upsilon_\psi, s^2 \sim Inv\text{-}\chi^2(\upsilon_\psi, s^2)$, em que υ_ψ e s^2 são hiperparâmetros. Como consequência, as condicionais completas *a posteriori* atribuídas a cada elemento da diagonal de $\boldsymbol{\Psi}_0$ se tornam:

$$
\begin{aligned}
p\left(\psi_j \mid \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{y}\right) &\propto p\left(\mathbf{y}_j^* \mid \boldsymbol{\beta}_j, \mathbf{u}_j, \psi_j\right) p\left(\psi_j\right) \\
&\propto \psi_j^{\frac{-n}{2}} exp\left(-\frac{\mathbf{e}_j'\mathbf{e}_j}{2\psi_j}\right) \psi_j^{-\left(\frac{\upsilon_\psi}{2}+1\right)} exp\left(-\frac{\upsilon_\psi s^2}{2\psi_j}\right) \\
&\propto \psi_j^{-\left(\frac{\upsilon_\psi+n}{2}+1\right)} exp\left(-\frac{\mathbf{e}_j'\mathbf{e}_j+\upsilon_\psi s^2}{2\psi_j}\right) \\
&\propto Inv\text{-}\chi^2\left(\upsilon_\psi+n, \frac{\mathbf{e}_j'\mathbf{e}_j+\upsilon_\psi s^2}{\upsilon_\psi+n}\right).
\end{aligned}
$$

Para a derivação da distribuição condicional completa *a posteriori* de $\boldsymbol{\Lambda}$, considerou-se que:

$$
\begin{aligned}
&\mathbf{y}^+ = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} = \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\mathbf{y} + \mathbf{e} \\
&p\left(\mathbf{y} \mid \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Psi}_0\right) = p\left(\mathbf{y}^+ \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}_0\right) = N\left[\left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\mathbf{y}, \boldsymbol{\Psi}\right].
\end{aligned}
$$

Dados $\boldsymbol{\beta}$ e $\mathbf{u}$, os $\mathbf{y}_i^+$ para diferentes indivíduos são independentes. Adicionalmente, considerando os resíduos de diferentes características como mutuamente independentes, o modelo $\mathbf{y}^+ = \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\mathbf{y} + \mathbf{e}$ com elementos desconhecidos em $\boldsymbol{\Lambda}$ pode ser resolvido separadamente para cada característica *j* (para cada linha *j* de $\boldsymbol{\Lambda}$). A função de verossimilhança pode ser reescrita como:

$$
p\left(\mathbf{y}^+ \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}_0\right) = \prod_{j=1}^{t} p\left(\mathbf{y}_j^+ \mid {}_j\boldsymbol{\Lambda}, \mathbf{y}_{-j}^+, \psi_j\right),
$$

em que ${}_j\boldsymbol{\Lambda}$ é a linha *j* de $\boldsymbol{\Lambda}$. Como consequência, as distribuições condicionais completas *a posteriori* dos elementos diferentes de zero em ${}_j\boldsymbol{\Lambda}$ podem ser derivadas de uma distribuição *a posteriori* conjunta como seria feito para coeficientes de regressão de um modelo univariado de regressão múltipla. Considerando $\boldsymbol{\zeta}_j$ como um vetor de elementos diferentes de zero em ${}_j\boldsymbol{\Lambda}$, e $\mathbf{Y}_{pj}$ como uma matriz contendo observações das características que são pais da característica *j*:

$$
\begin{aligned}
&\mathbf{y}_j^+ = \mathbf{Y}_{pj}\boldsymbol{\zeta}_j + \mathbf{e}_j \\
&p\left(\boldsymbol{\zeta}_j \mid \mathbf{y}_j^+, \psi_j\right) \propto p\left(\mathbf{y}_j^+ \mid \boldsymbol{\zeta}_j, \psi_j\right) p\left(\boldsymbol{\zeta}_j\right) \\
&\qquad \propto exp\left[-\frac{1}{2\psi_j}\left(\mathbf{y}_j^+ - \mathbf{Y}_{pj}\boldsymbol{\zeta}_j\right)'\left(\mathbf{y}_j^+ - \mathbf{Y}_{pj}\boldsymbol{\zeta}_j\right)\right] \\
&\qquad \propto exp\left[-\frac{1}{2\psi_j}\left(\mathbf{y}_j^{+\prime}\mathbf{y}_j^+ - 2\boldsymbol{\zeta}_j'\mathbf{Y}_{pj}'\mathbf{y}_j^+ + \boldsymbol{\zeta}_j'\mathbf{Y}_{pj}'\mathbf{Y}_{pj}\boldsymbol{\zeta}_j\right)\right] \\
&\qquad \propto exp\left[-\frac{1}{2\psi_j}\left(\mathbf{y}_j^{+\prime}\mathbf{y}_j^+ - 2\boldsymbol{\zeta}_j'\mathbf{Y}_{pj}'\mathbf{Y}_{pj}\hat{\boldsymbol{\zeta}}_j + \boldsymbol{\zeta}'_j\mathbf{Y}_{pj}'\mathbf{Y}_{pj}\boldsymbol{\zeta}_j\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\propto exp\left[-\frac{1}{2\psi_j}\left(\mathbf{y}_j^{+\prime}\mathbf{y}_j^{+}-2\boldsymbol{\zeta}_j^{\prime}\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\hat{\boldsymbol{\zeta}}_j+\boldsymbol{\zeta}_j^{\prime}\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\boldsymbol{\zeta}_j\right.\right.\\
&\left.\left.+\hat{\boldsymbol{\zeta}}_j^{\prime}\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\hat{\boldsymbol{\zeta}}_j-\hat{\boldsymbol{\zeta}}_j^{\prime}\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\hat{\boldsymbol{\zeta}}_j\right)\right]\\
&\propto exp\left[-\frac{1}{2\psi_j}\left(\mathbf{y}_j^{+\prime}\mathbf{y}_j^{+}+\left(\boldsymbol{\zeta}_j-\hat{\boldsymbol{\zeta}}_j\right)^{\prime}\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\left(\boldsymbol{\zeta}_j-\hat{\boldsymbol{\zeta}}_j\right)-\hat{\boldsymbol{\zeta}}_j^{\prime}\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\hat{\boldsymbol{\zeta}}_j\right)\right]\\
&\propto exp\left[-\frac{1}{2\psi_j}\left(\left(\boldsymbol{\zeta}_j-\hat{\boldsymbol{\zeta}}_j\right)^{\prime}\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\left(\boldsymbol{\zeta}_j-\hat{\boldsymbol{\zeta}}_j\right)\right)\right]\\
&\propto N\left[\hat{\boldsymbol{\zeta}}_j,\left(\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\right)^{-1}\psi_j\right],
\end{aligned}
$$

em que $\hat{\boldsymbol{\zeta}}_j=\left(\mathbf{Y}_{pj}^{\prime}\mathbf{Y}_{pj}\right)^{-1}\mathbf{Y}_{pj}^{\prime}\mathbf{y}_j^{+}$. O conjunto de características em $\mathbf{Y}_{pj}$ depende da estrutura causal utilizada para ajusta o MEE.