Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models**

B.D. Valente[*†], G.J.M. Rosa[†], G. de los Campos[‡],
D. Gianola[†‡§] and M.A. Silva[*]

Introduction

Structural Equation Models or SEM for short (Wright (1921); Haavelmo (1943)), are used to study recursive and simultaneous relationships among phenotypes in multivariate systems such as multiple trait models in quantitative genetics. Fitting SEM requires defining a causal structure among the variables studied. The number of possible causal structures grows markedly with the number of traits. Therefore, choosing of a causal structure can become cumbersome as the number of traits increases.

Structural Equation Models for quantitative genetics settings were described by Gianola and Sorensen (2004). Typically, one model or a limited set of models (i.e., causal structures) is pre-selected, and members of the set are fitted and compared based on some model comparison criteria. However, the utility of this approach may be limited because the set of possible causal structures considered is narrow. As an alternative, the notion of d-separation (Pearl (2000)) can be used to explore the space of causal hypotheses so as to arrive to a causal structure (or a class of observationally equivalent causal structures) that is capable of generating the observed pattern of conditional probabilistic independencies between variables. However, in a mixed models context, genetic covariances act as confounders because they are a source of phenotypic covariance that is not due to recursive relationships among traits. This article presents a methodology that allows searching for recursive causal structures in the context of mixed models for genetic analysis of quantitative traits. The proposed methodology exploits features of the mixed model and uses the Inductive Causation (IC) algorithm (Pearl (2000)) coupled with Bayesian data analysis.

Methodology

Structural Equation Model (SEM). Following Gianola and Sorensen (2004), a SEM with random additive genetic effects can be written as:

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

where $\mathbf{y}_i$ is a (t×1) vector of phenotypic records on the i^{th} subject; $\boldsymbol{\Lambda}$ is a (t× t) matrix with zeroes in the diagonal and with structural coefficients in the off-diagonal, the matrix $\mathbf{X}_i$ contains exogenous covariates, $\boldsymbol{\beta}$ is a vector of 'fixed effects', $\mathbf{u}_i$ is a (t×1) vector of

*Department of Animal Sciences, Federal University of Minas Gerais, 30123-970, Brazil
†Department of Dairy Science,
‡Department of Animal Sciences, and
§Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin USA 53706
** This research was funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil.

random additive genetic effects, and $\mathbf{e}_i$ is a vector of model residuals of the same dimension, both associated with the i^{th} subject. The following joint distribution is assumed for $\mathbf{u}_i$ and $\mathbf{e}_i$: $\begin{bmatrix} \mathbf{u}_i \\ \mathbf{e}_i \end{bmatrix} \sim N\left\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_0 \end{bmatrix} \right\}$, where $\mathbf{G}_0$ and $\mathbf{\Psi}_0$ are the additive genetic and residual covariance matrices, respectively. In model [1], the causal structure is defined by choosing which of the off-diagonal entries of $\mathbf{\Lambda}$ are free parameters, and which ones are set to zero.

Recovering recursive causal structures. A recursive causal structure is represented by a Directed Acyclic Graph (DAG), which is a set of variables connected by directed edges (arrows) representing direct causal relationships. A path in the causal structure is a sequence of connected variables, regardless of the direction of the arrows that connect them. Unconditionally, paths allow flows of dependence between variables in their extremes, unless there is a collider (variable with arrows pointing at them from opposite directions, like *c* in $a \rightarrow c \leftarrow b$) in the path. Colliders block the flow of dependency in a path, which makes *a* and *b* independent in the structure above. Conditioning on a variable that is not in the extremes of the path blocks the flow of dependency if this variable is a non-collider (e.g., conditioning on *c* in $a \rightarrow c \rightarrow b$, $a \leftarrow c \leftarrow b$ or $a \leftarrow c \rightarrow b$ makes *a* and *b* independent), or allows the flow of dependency if this variable is a collider. Considering two variables *a* and *b* in a DAG, they are said to be d-separated conditionally on a subset **S** of variables if there are no paths that allow flows of dependency between *a* and *b* (i.e., no paths between *a* and *b* in a DAG such that all the colliders or its descendants are in **S** and no non-colliders are in **S**). Under some assumptions, d-separations in the causal structure are reflected as conditional independencies in the joint probability distribution of the data. This can be explored to recover a causal structure or a class of equivalent causal structures (causal structures that result in joint probability distributions with the same conditional independence relationships) from the joint distribution of the data (Pearl (2000); Spirtes et al. (2000)).

For the model $\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{e}_i$, the IC algorithm (Pearl (2000)) can be used to recover an underlying DAG structure (or a class of observationally equivalent structures) from observed associations between traits. The search is based on queries about conditional independencies between variables and on the assumption that such independencies reflect d-separations in the underlying DAG. The input of the algorithm is a correlation matrix between observable variables, from which marginal and conditional dependencies can be evaluated. The output is a partially oriented graph representing a class of equivalent causal structures. Considering a set *V* of random variables, the IC algorithm can be described by the following steps:

1 – For each pair of variables *a* and *b* in *V*, search for a set of variables S_{ab} such that *a* is independent of *b* given S_{ab}. If *a* and *b* are dependent for every possible conditioning set, connect *a* and *b* with an undirected edge. This step results in an undirected graph *U*. Connected variables in *U* are called adjacent.
2 – For each pair of non-adjacent variables *a* and *b* with a common adjacent variable *c* in *U* (i.e., $a - c - b$), search for a set S_{ab} that contains *c* such that *a* is independent of *b* given

S_{ab}. If this set does not exist, then add arrowheads pointing at c ($a \rightarrow c \leftarrow b$). If this set exists, then continue.
3 – In the resulting partially oriented graph, orient as many undirected edges as possible in such a way that it does not result in new colliders or in cycles.

The search performed by the IC algorithm relies on the connection between causal graphs and the probability distributions that they generate. This connection is established based on the assumption that there are no hidden variables that affect more than one of the variables considered in the search (i.e., causal sufficiency assumption). In the SEM presented in [1], residuals $\mathbf{e}_i$ account for the effects of the remaining unknown causes of the phenotypic traits considered. The assumption of causal sufficiency implies that the search should be performed on a set of variables that includes every common cause of two or more of these variables. Therefore, under this assumption, residuals in SEM are constructed as independent. A diagonal residual covariance matrix is also imposed in recent applications of such models in quantitative genetics, which additionally assume causal structures as known (e.g. de los Campos et al. (2006); de Maturana et al. (2009); Heringstad et al. (2009)).

Causal structure search within a mixed models context. In the formulation described in the previous section, model residuals are regarded as independent, and recursive effects are used to model (interpret) patterns of co-variability between observable variables. However, in quantitative genetics mixed models, the patterns of co-variability between phenotypes may be due either to causal links between traits or to genetic reasons. In other words, correlated genetic effects can act as confounders if one tries to select a causal structure based on the joint distribution of the phenotypes, even if residuals are assumed as independent. Take as an example the scenarios depicted in Figure 1, where there are recursive relationships among the phenotypes y_1, y_2 and y_3, with uncorrelated residuals (e_1, e_2 and e_3) and correlated additive genetic effects (u_1, u_2 and u_3). The connection between the causal structure among phenotypes and their joint probability distribution does not hold in a model where genetic effects are uncontrolled hidden variables. In this case, y_1 and y_2 are not marginally independent in Figure 1a because of the covariance between u_1 and u_2. For the same reason, they are not independent given the phenotype y_3 in Figure 1b.

Nonetheless, additive relationships between individuals give a mean for "controlling" for this confounder. This can be done, for example, if there is pedigree or marker information on subjects. In this approach, d-separations are reflected as conditional independencies on the distribution of phenotypes after taking into account the additive genetic effects (i.e., the distribution of the phenotypes conditionally on the genetic effects). In Figure 1a, y_1 and y_2 are independent given the additive genetic effects. In Figures 1b, the same observed variables are independent given the additive genetic effects and phenotype y_3. Therefore, estimates of $Var(\mathbf{y}_i \mid \mathbf{u}_i) = \mathbf{R}_0$ can be used to select a causal structure among phenotypes. This matrix can be inferred from fitting a standard multiple trait model (MTM). In a Bayesian framework one draws samples from the posterior distribution of $\mathbf{R}_0$ and these samples may be used to obtain measures of uncertainty about this matrix, while accounting for uncertainty of all other parameters included in the MTM. Next, we describe how to search for a causal

structure in a mixed models context, using samples from the posterior distribution of $\mathbf{R}_0$ as input to the IC algorithm:

1 – Fit a MTM and draw samples from the posterior distribution of $\mathbf{R}_0$.

2 – Apply the IC algorithm to the posterior samples of $\mathbf{R}_0$ to make the statistical decisions required. Specifically, for each query about the statistical independence between variables *a* and *b* given a set of variables *S* and, implicitly, the genetic effects:

2.1 – Obtain the posterior distribution of residual partial correlation $\rho_{a,b|S}$. These partial correlations are functions of $\mathbf{R}_0$. Therefore their posterior distribution can be obtained by computing the correlation at each sample drawn from the posterior distribution of $\mathbf{R}_0$.

2.2 – Compute the 95% highest posterior density (HPD) interval for the posterior distribution of $\rho_{a,b|S}$.

2.3 – If the HPD interval contains 0, declare $\rho_{a,b|S}$ as null. Otherwise, declare *a* and *b* as conditionally dependent.

3 – Fit a SEM using the selected causal structure (or one member within the class of observationally equivalent structures retrieved by the IC algorithm) as 'TRUE' causal structure.

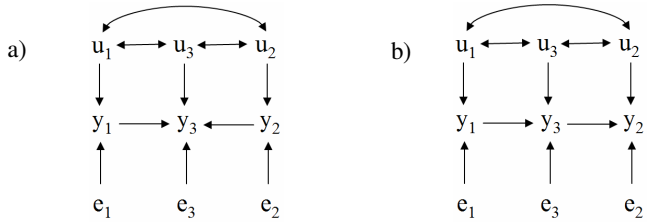

Figure 1: Causal structures for three observed variables (y_1, y_2 and y_3),with independent residuals (e_1, e_2 and e_3) and correlated additive genetic effects (u_1, u_2 and u_3).

References

de los Campos, G., Gianola, D., Boettcher, P. *et al.* (2006). *J. Anim. Sci.,* 84: 2934-2941.

Gianola, D., and Sorensen D., (2004). *Genetics,* 167: 1407-1424.

Haavelmo, T. (1943). *Econometrica,* 11: 1–12.

Heringstad, B., Wu, X.L., and Gianola, D., (2009). *J. Dairy Sci.,* 92:1778–1784.

de Maturana, E. L., Wu, X.L., Gianola, D. *et al.* (2009). *Genetics,* 181:277–287.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction and Search*.

Wright, S. (1921). *J. Agric. Res.,* 201: 557–585.