Application of Bayesian causal inference and structural equation model to animal breeding

Keiichi Inoue

National Livestock Breeding Center, Nishigo, Fukushima, Japan

Correspondence
Keiichi Inoue, National Livestock Breeding Center, Nishigo, Fukushima 961-8511, Japan. Email: k1inoue@nlbc.go.jp

Abstract
Optimized breeding goals and management practices for the improvement of target traits requires knowledge regarding any potential functional relationships between them. Fitting a structural equation model (SEM) allows for inferences about the magnitude of causal effects between traits to be made. In recent years, an adaptation of SEM was proposed in the context of quantitative genetics and mixed models. Several studies have since applied the SEM in the context of animal breeding. However, fitting the SEM requires choosing a causal structure with prior biological or temporal knowledge. The inductive causation (IC) algorithm can be used to recover an underlying causal structure from observed associations between traits. The results of the papers, which are introduced in this review, showed that using the IC algorithm to infer a causal structure is a helpful tool for detecting a causal structure without proper prior knowledge or with uncertain relationships between traits. The reports also presented that fitting the SEM could infer the effects of interventions, which are not given by correlations. Hence, information from the SEM provides more insights into and suggestions on breeding strategy than that from a multiple-trait model, which is the conventional model used for multitrait analysis.

KEYWORDS
animal breeding system, Bayesian causal inference, external intervention, inductive causation algorithm, structural equation model

1 | INTRODUCTION

In animal breeding strategies, there are two types of breeding goal to ensure the ability of animal to fit the demands of producers, consumers, and markets. One is the fundamental or sustainable breeding goal, which is not dependent on market trends and focuses on properties such as reproduction, feed efficiency, and longevity. The other is the flexible breeding goal that reflects temporal market or economical demands and focuses on properties such as the qualities of milk and meat products. When we are confronted with the need to improve a new trait because of changing market demands, such as fatty acid composition, which is a meat quality trait and is related to the sweet aroma (Sakuma et al., 2012) and flavor of beef (Suzuki, Yokota, Shioura, Shimazu, & Iida, 2013) in Japanese Black cattle, we add the new trait to conventional selection traits and improve multiple traits simultaneously. In such a case, the most important thing is to confirm the genetic correlations between the traits by estimating the genetic parameters fitted by the multiple-trait model (MTM; Henderson & Quass, 1976) to prevent negative effects on the conventional economical traits. However, there are not only
'correlations' but there could also be 'causal relationships' between the traits. For example, in the case of the relationship between gestation length (GL) and calving difficulty (CD), GL affects CD, because a longer GL causes a bigger infant or harmonization. In contrast, the opposite relationship is not possible; CD never affects GL under biological conditions.

When causal relationships exist between traits, fitting a structural equation model (SEM) allows for inferences of the magnitude of the causal effects between traits (Haavelmo, 1943; Wright, 1921). SEMs have been applied in various fields, such as economics, psychology, and sociology. In recent years, Gianola and Solensen (2004) proposed adaptations of the SEM in the context of quantitative genetics and mixed-effect models. Ever since their proposal, several studies have applied the SEM in the context of animal breeding in dairy goats (de los Campos, Gianola, Boettcher, & Moroni, 2006), dairy cattle (de los Campos, Gianola, & Heringstad, 2006; Heringstad, Wu, & Gianola, 2009; Konig, Wu, Gianola, Heringstad, & Simianer, 2008; Wu, Heringstad, Chang, de los Campos, & Gianola, 2007; Wu, Heringstad, & Gianola, 2008; etc.), pigs (Ibanez-Escriche, Lopez de Maturana, Noguera, & Varona, 2010; Varona, Sorensen, & Thompson, 2007), and beef cattle (Inoue, Hosono, & Tanimoto, 2017; Inoue et al., 2016).

Fitting the SEM requires choosing a causal structure with prior biological or temporal knowledge. However, the inductive causation (IC) algorithm (Pearl, 2000; Verma & Pearl, 1990), which is one of the Bayesian causal inference, can be used to recover an underlying causal structure from observed associations between traits. The IC algorithm can output the causal relationships with a directed acyclic graph (DAG), which is a set of variables connected by directed edges, or a partially oriented graph. Searching for the causal structure in the algorithm is based on a conditional independence (Dawid, 1980) between variables.

Valente, Rosa, de los Campos, Gianola, and Silva (2010) adapted the IC algorithm to a mixed-model context and showed that applying this method to the posterior distribution of the residual (co)variance matrix of a standard MTM recovered the expected network in simulated data. They also presented an application of the methodology to real field data of European quail. After their work, several groups have tried to explore the causal structure from the real field data.

This review will first describe an overview of the IC algorithm and SEM, and then introduce some papers that apply these techniques to the context of animal breeding.

2 | THE IC ALGORITHM

The IC algorithm, which was proposed by Pearl (2000) and Verma and Pearl (1990), allows to search for causal networks. The IC algorithm performs a series of statistical decisions based on partial correlations between traits and consists of the following three steps (an example is shown in Figure 1):

2.1 | Step 1

If all partial correlations of two traits conditional on each possible set of the other traits are different from 0, the two traits are connected by an undirected edge (e.g., \( y_1 \rightarrow y_2 \)). For example, when there are four target traits for inferring the causal structure (\( y_1, y_2, y_3, \) and \( y_4 \)), as shown on the left graph of Step 1 in Figure 1, if all partial correlations \( \rho_{y_1y_2|y_3}, \rho_{y_1y_2|y_4}, \) and \( \rho_{y_1y_2|y_3y_4} \) are different from 0 (significant), the two traits are connected by an undirected edge (\( y_1 \leftarrow y_2 \)). The graph with undirected edges is finally generated by iterating the same process for the remaining possible set of traits (the right graph of Step 1 in Figure 1).

2.2 | Step 2

In the undirected graph obtained by step 1, if partial correlations of two nonadjacent traits (e.g., \( y_1 \) and \( y_3 \)) with a common adjacent trait (e.g., \( y_2 \) in \( y_1 \rightarrow y_2 \leftarrow y_3 \)) are dependent conditional on any possible set including the adjacent trait (\( y_2 \)), that is, all partial correlations \( \rho_{y_1y_3}, \rho_{y_1y_3|y_2}, \rho_{y_1y_3|y_4}, \) and \( \rho_{y_1y_3|y_2y_4} \) are not significantly different from 0, arrowheads pointing to the common adjacent trait (\( y_2 \)) are added (\( y_1 \rightarrow y_2 \leftarrow y_3 \) in the graph of Step 2 in Figure 1). This kind

FIGURE 1 An example of the steps for the inductive causation algorithm. \( y_1 \) to \( y_4 \) represent traits, \( \rightarrow \) and \( \leftarrow \) represent undirected and directed edges, respectively.
of ‘V’ structure is called an ‘unshielded collider.’ The same process is repeated for the remaining possible set of traits.

2.3 | Step 3

In the partially oriented graph obtained from step 2, as many undirected edges as possible are oriented without creating a new unshielded collider or cycle. In the case of the example in Figure 1, y2 and y4 are connected by an undirected edge. If an arrowhead pointing to y2 from y4 is added to the undirected edge (y2 ← y4), new unshielded colliders are created (y1 → y2 ← y4 and y3 → y2 ← y4); therefore, a new directed edge from y4 to y2 cannot be added. In contrast, arrowheads pointing to y4 from y2 can be added to the undirected edge (y2 → y4) because no new unshielded colliders are created. Finally, the graph of the causal structure is completed, shown as the graph in Step 3 of Figure 1. Additionally, an example of a cyclic graph is as follows: y1 → y2 → y3 → y1.

Statistical decisions regarding declaring partial correlations as null or not null are based on highest posterior density (HPD) intervals. If the interval contained the value 0, the correlation is declared null or not null are based on highest posterior density (HPD) intervals. Outputs may differ according to the probability content used for the decisions.

In the context of the mixed-model analysis, after fitting the MTM to the traits of interest, the IC algorithm can be applied to samples from the posterior distribution of the residual (co)variance matrix obtained from the multiple-trait analysis because the residual (co) variances are considered to be information from the joint distributions of the traits conditional on genetic effects, which corrects the search for confounding because of such effects when they are correlated (Valente et al., 2010). The program ICPS written in R (R Development Core Team, 2009), which can carry out the IC algorithm analysis, is provided by Valente and Rosa (2013).

3 | STRUCTURAL EQUATION MODEL

In the case of existing causal relationships between traits, as mentioned in the previous section, fitting a SEM (Haavelmo, 1943; Wright, 1921) can infer the causal coefficients between the traits. A SEM with a causal structure and random additive genetic effects, which was first proposed by Gianola and Solensen (2004), has been reviewed (Rosa et al., 2011; Wu, Heringstad, & Gianola, 2010) and the model with t traits can be written as follows:

$$Y_i = \Lambda Y_i + X_i \beta + u_i + e_i$$

where $Y_i$ is a t x 1 vector of phenotypic records on animal $i$; $\Lambda$ is a t x t matrix of structural coefficients, which is filled by 0 except for the off-diagonal elements that correspond to the causal structure; $\beta$ is a vector of fixed effects; $u_i$ is a vector of random additive genetic effects; $e_i$ is a vector of random residual effects; and $X_i$ is a known incidence matrix. The joint distribution of vectors $u_i$ and $e_i$ in the equation is as follows:

$$\begin{pmatrix} u_i \\ e_i \end{pmatrix} \sim N(0, \begin{pmatrix} G_0 & 0 \\ 0 & R_* \end{pmatrix})$$

where $G_0$ is the additive genetic (co)variance matrix and $R_*$ is the diagonal residual variance matrix.

By reducing the above SEM for $y$, the model is transformed, as indicated below (Gianola & Solensen, 2004; Varona et al., 2007):

$$y_1 = (I - \Lambda)^{-1} X_\beta + (I - \Lambda)^{-1} u_1 + e_1^*$$

where $u_1^* = (I - \Lambda)^{-1} X_\beta + (I - \Lambda)^{-1} u_1$ and $e_1^* = (I - \Lambda)^{-1} e_1$. In addition, the joint distribution of $u_i$ and $e_i^*$ is as follows:

$$\begin{pmatrix} u_i \\ e_i^* \end{pmatrix} \sim N(0, \begin{pmatrix} G_0^* & 0 \\ 0 & R_*^* \end{pmatrix})$$

with $G_0^* = (I - \Lambda)^{-1} G_0 (I - \Lambda)^{-1} r^{-1}$, $R_*^* = (I - \Lambda)^{-1} R_* (I - \Lambda)^{-1} r^{-1}$. Here, $u_i^*$, $e_i^*$, $G_0^*$, and $R_*^*$ are the similar vectors as described above ($G_0^*$ is the additive genetic (co)variance matrix and $R_*^*$ is the residual (co)variance matrix). Like this, the SEM can be transformed into a MTM, which ignores causal relationships between traits. Therefore, the likelihood between MTM and SEM is equivalent (Varona et al., 2007). However, genetic effects estimated from the two models have different meaning. The MTM only infers ‘overall’ genetic effects, which include all ‘direct’ and ‘indirect’ genetic effects on each trait, as illustrated later. Here, the indirect genetic effects are those that are mediated by other phenotype traits and the direct genetic effects are those that are not. In contrast, the SEM can infer the direct genetic effects, which are separated from the causal effects (corresponding to the indirect genetic effects; Valente, Rosa, Gianola, Wu, & Weigel, 2013). For instance, assuming the causal structure between two traits shown in Figure 2a, the SEM can be represented as follows:

$$y_1 = u_1 + e_1$$

$$y_2 = \lambda_{21} y_1 + u_2 + e_2$$

where $y_1$ and $y_2$ are the phenotypes; $u_1$ and $u_2$ are the additive genetic effects which affect directly on $y_1$ and $y_2$, respectively (direct genetic effects), but $u_1$ is also the indirect genetic effect on $y_2$: $e_1$ and $e_2$ are the residual effects; and $\lambda_{21}$ is the causal effect from $y_1$ to $y_2$. In addition, the equation described above can be transformed as follows:

$$y_1 = u_1 + e_1 = u_1^* + e_1^*$$

$$y_2 = \lambda_{21}(u_1 + e_1) + u_2 + e_2 = (\lambda_{21} u_1 + u_2) + (\lambda_{21} e_1 + e_2) = u_2^* + e_2^*$$
4 | INTERVENTIONS IN THE SEM

As shown in the previous section, the SEM can infer direct and indirect genetic effects separately but the MTM cannot. Therefore, a breeding strategy based only on estimated genetic parameters obtained from the MTM would cause a delay in achieving the breeding goal if causal relationships and external 'interventions,' which would block indirect genetic effects, exist among the traits (Valente et al., 2013). An example of the intervention for \( y_1 \) in Figure 2a is shown in Figure 2b. The value of \( y_1 \) is fixed. The equation, which represents the intervention shown in Figure 2b, is as follows:

\[
\begin{align*}
  y_1 &= c \\
  y_2 &= \lambda_{21}c + u_2 + e_2
\end{align*}
\]

where \( u_2 \) and \( e_2 \) are the additive genetic and residual effects in the MTM, respectively. This equation shows that the genetic effects in the SEM affect not only the corresponding phenotypes directly but also the other phenotypes indirectly via the causal structure. In addition, it suggests that the genetic effects in the MTM represent the overall effects, which also include the direct and indirect effects via the causal effects (\( u_2 = \lambda_{21}u_1 + u_2 \)).

5 | APPLYING THE IC ALGORITHM AND SEM FOR ANIMAL BREEDING

Ever since Gianola and Solensen (2004) proposed adaptations of the SEM in the context of quantitative genetics and mixed-effect models, several studies have applied the SEM to different species and traits in the context of animal breeding. de los Campos, Gianola, Boettcher, et al. (2006) and de los Campos, Gianola, and Heringstad (2006) studied the relationship between somatic cell score (SCS) and milk yield (MY) in dairy cows and dairy goats. This work is the first application of the SEM to field data. Similarly, in dairy cattle, there are many reports exploring the relationships between SCS or clinical mastitis and MY in Norwegian Red cows (Wu et al., 2007, 2008); between claw disorders and MY in Holstein cows (Konig et al., 2008); between health and fertility traits in Norwegian Red cows (Heringstad et al., 2009); among GL, CD, and SB in US Holstein cows (Lopez de Maturana et al., 2009, 2010); and between SCS and MY in Canadian Holstein cows (Jamrozik, Bohmanova, & Schaeffer, 2010; Jamrozik & Schaeffer, 2010). Some of the authors proposed extensions for the SEM, such as accounting for population heterogeneity (Wu et al., 2007), the Gaussian-threshold Bayesian hierarchical model (Wu et al., 2008), the heterogeneous causal model (Wu et al., 2010), and the random regression model (Jamrozik et al., 2010; Jamrozik & Schaeffer, 2010). In swine, Varona et al. (2007) used a recursive model to analyze the relationship between litter size and average litter weight in Landrace and Yorkshire pigs and found likelihood equivalence between MTM and the SEM. Ibanez-Escriche et al. (2010) developed a change-point recursive model.
using change-point detection techniques (Carlin, Gelfand, & Smith, 1992; Chib, 1998) and fitted the model to the litter size and number of SBs in Large White pigs. However, these studies assumed that causal structures were known a priori.

Recently, some studies fitted SEMs based on a data-driven causal structure search using the IC algorithm. The IC algorithm fitted to the posterior distribution of residual (co)variance matrices that are considered phenotypic correlations conditional on additive genetic correlations between traits; this approach was proposed by Valente et al. (2010). They used simulated data and showed that this approach recovered the correct causal network. Then, Valente, Rosa, Silva, Teixeira, and Torres (2011) presented the first application of such a methodology to a real dataset by exploring the space of causal structures among five productive and reproductive traits in European quail. From the application of the IC algorithm with different HPD intervals, three undirected graphs were obtained, but all had no directed edge or unsheilded collider. To construct the structures among the five traits and fit SEMs to the structures, the authors oriented the edges of the graphs according to temporal information as a prior knowledge.

Bouwman, Valente, Janss, Bovenhuis, and Rosa (2014) explored causal networks between 14 bovine milk fatty acids, which are highly correlated traits, by applying the IC algorithm. They retrieved undirected graphs using the IC algorithm with different HPD intervals, and then oriented the undirected graph obtained with the 95% HPD interval according to prior biological knowledge of de novo synthesis to fit the SEM: C4:0 → C6:0 → C8:0 → C10:0 → C12:0. The authors showed that inferred structural coefficients ranged from 0.85 to 1.05 and the SEM could be more plausible than the MTM by the SEM, the genetic variances for downstream traits in the causal structure because maternal genetic effects might not be removed completely from the residual variance components in the model.

Recently, techniques for inferring causal relationships by using the IC algorithm and SEM have applied to genome-wide association studies (GWAS). Momen et al. (2018) applied SEM to GWAS on chickens, including causal relationships among breast meat (BM), body weight (BW), and hen-house production (HHP), to compare the results obtained from both an SEM-based GWAS (SEM-GWAS) and a traditional multitrait association analysis (MTM-GWAS). They set a causal structure, BM → BW and HHP and BW → HHP, found by the IC algorithm and fitted the SEM to the structure. As a result, the application of SEM-GWAS could separate SNP effects into direct, indirect, and total effects. The authors concluded that the SEM-GWAS delivered a more comprehensive understanding of SNP effects than did the MTM-GWAS.

6 | CONCLUSIONS

The results of the papers that are introduced in this review showed that the application of the IC algorithm to infer a causal structure is one of the most helpful tools for detecting an underlying causal structure without proper prior knowledge or with uncertain relationships between traits. The reports also indicated that visualization of the causal structure among traits on a graph and fitting the SEM to the causal structure could infer the effects of interventions, which
are not given by correlations. Hence, information from the SEM provides more insights into and suggestions on breeding strategies than does that from a MTM, which is a conventional model usually used for multiple-trait analysis. These new procedures could be utilized in the field of animal breeding.

ORCID
Keiichi Inoue https://orcid.org/0000-0001-5353-8558

REFERENCES
Bouwman, A. C., Valente, B. D., Janss, L. L. G., Bovenhuis, H., & Rosa, G. J. M. (2014). Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. Genetics Selection Evolution, 46, 2. https://doi.org/10.1186/1297-9686-46-2
Carlin, B. P., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. Applied Statistics, 41, 389–405. https://doi.org/10.2307/2347570
Chib, S. (1998). Estimation and comparison of multiple-change point models. Journal of Econometrics, 86, 221–241. https://doi.org/10.1016/s0304-4076(97)00115-2
Dawid, A. P. (1980). Conditional independence for statistical operations. Annals of Statistics, 8, 598–617. https://doi.org/10.1214/aos/1176345011
de los Campos, G., Gianola, D., Boettcher, P., & Moroni, P. (2006). A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. Journal of Animal Science, 84, 2934–2941. https://doi.org/10.2527/jas.2006-016
de los Campos, G., Gianola, D., & Heringstad, B. (2006). A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows. Journal of Dairy Science, 89, 4445–4455. https://doi.org/10.3168/jds.S0022-0302(06)72493-6
Gianola, D., & Sorensen, D. (2004). Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. Genetics, 167, 1407–1424. https://doi.org/10.1534/genetics.103.025734
Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. Econometrica, 11, 1–12. https://doi.org/10.2307/1905714
Henderson, C. R., & Quaas, R. L. (1976). Multiple trait evaluation using relative’s records. Journal of Animal Science, 43, 1188–1197. https://doi.org/10.2527/jas1976.436118sx
Heringstad, B., Wu, X. L., & Gianola, D. (2009). Inferring relationships between health and fertility in Norwegian Red cows using recursive models. Journal of Dairy Science, 92, 1778–1784. https://doi.org/10.3168/jds.2008-1535
Ibanez-Escriche, N., Lopez de Maturana, E., Noguera, J. L., & Varona, L. (2010). An application of change-point recursive models to the relationship between litter size and number of stillborns in pigs. Journal of Animal Science, 88, 3493–3503. https://doi.org/10.2527/jas.2009-2557
Inoue, K., Hosono, M., & Tanimoto, Y. (2017). Inferring causal structures and comparing the causal effects among calving difficulty, gestation length and calf size in Japanese Black cattle. Animal, 11, 2120–2128. https://doi.org/10.1017/S1751731117000957
Inoue, K., Valente, B. D., Shoji, N., Honda, T., Oyama, K., & Rosa, G. J. M. (2016). Inferring phenotypic causal structures among meat quality traits and the application of a structural equation model in Japanese Black cattle. Journal of Animal Science, 94, 4133–4142. https://doi.org/10.2527/jas.2016-0554
Jamrozik, J., Bohmanova, J., & Schaeffer, L. R. (2010). Relationships between milk yield and somatic cell score in Canadian Holsteins from simultaneous and recursive random regression models.
Valente, B. D., Rosa, G. J. M., Silva, M. A., Teixeira, R. B., & Torres, R. A. (2011). Searching for phenotypic causal networks involving complex traits: An application to European quail. *Genetics Selection Evolution, 43*, 37. https://doi.org/10.1186/1297-9686-43-37

Varona, L., Sorensen, D., & Thompson, R. (2007). Analysis of litter size and average litter weight in pigs using a recursive model. *Genetics, 177*, 1791-1799. https://doi.org/10.1534/genetics.107.077818

Verma, T., & Pearl, J. (1990). Equivalence and synthesis of causal models. In Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence (pp. 220-227). Cambridge, MA: Elsevier. https://doi.org/10.5555/647233.719736

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557-585.

Wu, X.-L., Heringstad, B., Chang, Y.-M., de los Campos, G., & Gianola, D. (2007). Inferring relationships between somatic cell score and milk yield using simultaneous and recursive models. *Journal of Dairy Science, 90*, 3508-3521. https://doi.org/10.3168/jds.2006-762

How to cite this article: Inoue K. Application of Bayesian causal inference and structural equation model to animal breeding. *Anim Sci J*. 2020;91:e13359. https://doi.org/10.1111/asj.13359