University of São Paulo
“Luiz de Queiroz” College of Agriculture

**Using graphical models to investigate phenotypic networks involving
polygenic traits**

Renan Mercuri Pinto

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2018

Renan Mercuri Pinto
Degree in Mathematics

**Using graphical models to investigate phenotypic networks involving
polygenic traits**

Advisor:

Prof. Dr. **ROSELI APARECIDA LEANDRO**

Thesis presented to obtain the degree of Doctor in Science.

Area: Statistics and Agricultural Experimentation

Piracicaba
2018

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Pinto, Renan Mercuri

Using graphical models to investigate phenotypic networks involving poly-genic traits/ Renan Mercuri Pinto. – – Piracicaba, 2018 .
85 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Redes causais 2. Modelos de equações estruturais 3. Redes Bayesianas.
I. Título.

DEDICATION

To my family,
for all support and encouragement.

ACKNOWLEDGMENTS

Firstly, I thank my family (Adriana Cristina Mercuri Pinto, Osvail Aparecido Pinto e Laís Mercuri Pinto) for having fed me with good manners and taught me to fight for my dreams with perspicacity and patience. Also, to my girlfriend (Daiane Biscioni Von Zuben) and her family for all support.

I am grateful for the essential guidance of three professors from the University of Wisconsin (UW) - Madison, Dr. Guilherme J. M. Rosa, Dr. Bruno D. Valente and Dr. Fernando L. Brito, without which this work would not be possible.

To my advisor Dr. Roseli A. Leandro, from the University of Sao Paulo (USP) - ESALQ, for all support and for making my sandwich period at UW-Madison possible.

To the professors of the Department of Statistics and Agronomic Experimentation at USP-ESALQ, especially Dr. Taciana V. Savian and Dr. Idemauro R. de Lara, and also to the professors of the Department of Animal Science at UW-Madison.

To my special friends from USP-ESALQ (Ana Julia Righetto, Cesar Augusto Jotta, Hiron Pereira Farias, Luiz Ricardo Nakamura, Pedro Henrique Cerqueira, Rafael Moral, Ricardo Klein, Rodrigo Pescim, Thiago Gentil Ramires) and UW-Madison (André Galo, Arthur Fernandes, Ashley Mikshowskey, Camila Barros, Katrin Topner, Ligia Moreira, Mahmoud Amiri, Marília Ferreira, Mehdi Momen, Nicole Gross, Tiago Passafaro, Vera Ferreira).

To the staff of the Department of Exact Science at USP-ESALQ, in special the secretary Solange Sabadin and the computer technicians Eduardo Bonilha and Jorge Wiendl.

To CAPES and CNPq, for financial support.

CONTENTS

Resumo	7
Abstract	8
1 Introduction	9
References	10
2 Conclusion	13

RESUMO

O uso de modelos gráficos para investigar redes fenotípicas envolvendo características poligênicas

Compreender a arquitetura causal subjacente à sistemas biológicos complexos é de grande valia na produção agrícola para o desenvolvimento de estratégias de manejo e seleção genética. Até o momento, a maior parte dos estudos neste contexto utiliza apenas conhecimento prévio para propor redes causais e/ou não considera fatores de confundimento genético na busca de estruturas, fato que pode ocultar relações importantes entre os fenótipos e viesar inferências sobre a rede causal. Nesta tese, exploramos alguns algoritmos de aprendizagem de estruturas e apresentamos um novo, chamado PolyMaGNet (do inglês, Polygenic traits with Major Genes Network analysis), para buscar estruturas causais recursivas entre características fenotípicas poligênicas complexas e permitindo, também, a possibilidade de efeitos de genes maiores que as afetam. Resumidamente, um modelo misto de múltiplas características é ajustado usando abordagem Bayesiana considerando os genes maiores como covariáveis no modelo. Em seguida, amostras posteriores da matriz de covariância residual são usadas como entrada para o algoritmo de causação indutiva para pesquisar estruturas causais putativas, as quais são comparadas usando o critério de informação de Akaike. O desempenho do PolyMaGNet foi avaliado e comparado com outra abordagem bastante utilizada por meio de um estudo simulado considerando uma população de mapeamento de QTL. Os resultados mostraram que, na presença de genes maiores, o método PolyMaGNet recuperou a verdadeira estrutura do esqueleto, bem como as direções causais, com uma taxa de efetividade maior. O método é ilustrado também utilizando-se um conjunto de dados reais de uma população de suínos F2 Duroc $\times$ Pietrain para recuperar a estrutura causal subjacente à características fenotípicas relacionadas a qualidade da carcaça, carne e composição química. Os resultados corroboraram com a literatura sobre as relações de causa-efeito entre os fenótipos e também forneceram novos conhecimentos sobre a rede fenotípica e sua arquitetura genética.

Palavras-chave: Redes causais; Modelos de equações estruturais; Redes Bayesianas

ABSTRACT

Using graphical models to investigate phenotypic networks involving polygenic traits

Understanding the causal architecture underlying complex systems biology has a great value in agriculture production for the development of optimal management strategies and selective breeding. So far, most studies in this area use only prior knowledge to propose causal networks and/or do not consider the possible genetic confounding factors on the structure search, which may hide important relationships among phenotypes and also bias the resulting inferred causal network. In this dissertation, we explore many structural learning algorithms and present a new one, called PolyMaGNet (Polygenic traits with Major Genes Network analysis), to search for recursive causal structures involving complex phenotypic traits with polygenic inheritance and also allowing the possibility of major genes affecting the traits. Briefly, a multiple-trait animal mixed model is fitted using a Bayesian approach considering major genes as covariates. Next, posterior samples of the residual covariance matrix are used as input for the Inductive Causation algorithm to search for putative causal structures, which are compared to each other using the Akaike information criterion. The performance of PolyMaGNet was evaluated and compared with another widely used approach in a simulated study considering a QTL mapping population. Results showed that, in the presence of major genes, our method recovered the true skeleton structure as well as the causal directions with a higher rate of true positives. The PolyMaGNet approach was also applied to a real dataset of an F2 Duroc $\times$ Pietrain pig resource population to recover the causal structure underlying on carcass, meat quality and chemical composition traits. Results corroborated with the literature regarding the cause-effect relationships between these traits and also provided new insights about phenotypic causal networks and its genetic architectures in complex systems biology.

Keywords: Causal networks; Structural equation models; Bayesian networks

1 INTRODUCTION

In agriculture, relationships among phenotypic traits are commonly studied using the standard Multiple Trait Model (MTM; Henderson and Quaas, 1976; Mrode, 2005). Although such models can be used to infer how likely events are, they are not stable enough to predict how the probabilities could vary as a result of external interventions (Pearl, 2000; Rosa et al., 2011; Shipley, 2016). For example, a correlation detected between two traits T1 and T2 may be due to a direct effect of T1 on T2 (or vice versa) or to a latent variables affecting both together. Knowledge of the underlying phenotypic causal network is essential to predict the effect of management practices applied to both traits. That is, if T1 affects T2, but T2 has no effect on T1, an intervention on T1 would change T2, but the reverse would not happen.

Similar scenarios occur in genetic improvement, where genetic correlation is defined as the proportion of variance that two traits share due to genetic causes (Rosa et al., 2011). In classical genetics many genes are known to have multiple effects, this action is called pleiotropy. For example, the vestigial gene in *Drosophila* is responsible for affecting not only the bristles and the wings, but also the fecundity (Mode and Robinson, 1959).

According to Schadt et al. (2005) there are three possible causal relationships (Figure 1.1) involving a gene (G) and two phenotypic traits (T1 and T2), which are explored in details by Li et al. (2006). In the first case (Figure 1.1a), the gene G affects phenotype T1, and the phenotypic change on T1 affects T2; in the second (Figure 1.1b), the gene G acts on T2, and the phenotypic change on T2 changes T1; and, in the third (Figure 1.1c) the gene G changes both traits directly, which may or may not have a causal effect between them.

Figure 1.1. Possible gene-phenotypic networks involving one gene (G) and two phenotypic traits (T1 and T2) - [adapted from Rosa et al., 2011]

The traditional MTM mentioned above might detect a correlation between two traits and possibly the pleiotropic effect of the gene, but would not be able to distinguish the configuration of the paths that connect them. As an alternative, graphical models can be used to study recursive and simultaneous relationships among variables in multivariate systems, offering a different interpretation of the relationships between traits, such that one trait can be considered as a predictor of another, providing a causal path between them (Haavelmo, 1943; Wright, 1921; Rosa et al., 2011).

Graphical models provide a qualitative representation of biological systems along with the quantitative analysis. The qualitative part is a graphical representation of dependences among variables, expressed, for instance, by a directed acyclic graph (DAG), an undirected graph (UG) or a partially directed acyclic graph (PDAG), that is, a graph which may have both directed and undirected edges. Knowledge about the quantitative dependences between variables is added in the graph by means of path coefficients (parameters). There are many kinds of graphical models in the literature, in this thesis we will cover the most popular ones: structural equation models (SEM) and Bayesian networks (BN).

Gianola and Sorensen (2004) described SEMs in the context of mixed models in quantitative genetics, and since then many authors have used such approach (De los Campos et al., 2006; de Maturana et al., 2009), but usually causal structures are pre-selected using some kind of prior knowledge. Here, we propose to use the notion of directional separation (d-separation; Spirtes et al., 2000; Pearl, 2000;

Shipley, 2016) to explore the space of causal hypotheses and thus arrive at a causal structure (or a class of equivalent structures) that is capable of generating the observed pattern of conditional probabilities among variables.

Seeking to offer new methods to researchers who use only prior knowledge to propose causal networks, in chapter 2 we merge structural learning algorithms along with prior knowledge to investigate the causal networks underlying phenotypic traits of two fruit species belonging to the Sapotaceae family. We use a constrain- and a score-based algorithm for structure learning of BNs and select a putative causal network from the equivalence class of the score-based output, which are fitted using maximum likelihood and evaluated through some fit indices of SEM.

When considering only phenotypic data in the structural search, as in Chapter 2, we are subject to obtain networks biased by genetic confounding factors. For instance, if there is a genetic variable not considered in the search that affects two traits simultaneously (Figure 1.1c), we may find a causal path between the two traits that does not actually exist. Thus, if high density molecular marker data is available, more reliable causal networks can be obtained through more efficient genetic prediction approaches.

In Chapter 3, using 67 phenotypic traits measured in a F2 Duroc $\times$ Pietrain pig resource population with genetic data available, we compare three different methodologies for genome-wide association studies (GWAS): a single-marker regression, a ridge regression BLUP, and a Bayes C π . In addition, we use the latter method to detect major genes associated with the expression of carcass and meat quality traits, in order to find out the possible genetic confounding factors that could bias inferences regarding their causal network.

There are many structure learning algorithms for BN in the area of quantitative genetics (Logsdon and Mezey, 2010; Neto et al., 2008, 2010; Valente et al., 2010; Wang and Van Eeuwijk, 2014), which allow to explore the structure space compatible with the joint distribution function of the variables studied. Among those, Valente et al. (2010) proposed a constraint-based algorithm to search for recursive causal structures among phenotypes conditionally to unobservable polygenic effects, which act as confounders. Basically, a standard Bayesian multiple-trait model is fitted using a Markov chain Monte Carlo implementation to obtain samples from the posterior distribution of a residual covariance matrix, which are then used as input for the inductive causation (IC) algorithm (Verma and Pearl, 1990; Pearl, 2009). However, their method is based on an infinitesimal model and, as such, it does not take into account the possibility of major genes affecting the traits. Often, their method produces a partially directed network that represents a class of possible equivalent solutions, and the use of prior biological knowledge is necessary to determine a final, fully directed graph.

Thus, in Chapter 4, we reach the main goal of this research, which was to develop a hybrid method, called PolyMaGNet (Polygenic traits with Major Genes Network analysis), to search for recursive causal structures among complex phenotypic traits with polygenic inheritance, but allowing also the possibility of major genes affecting the traits. Major genes, previously detected via GWAS, are used as instrumental variables in a final step of the algorithm, providing a fully oriented acyclic graph. This method can be seen as an extension of the one proposed by Valente et al. (2010) and it is especially useful for the analysis of QTL mapping population data involving crosses of outbred populations. We also provide a detailed description of the proposed method and illustrate it with a simulation study, as well as with the real dataset explored in Chapter 3, to recover the causal structure underlying carcass and meat quality traits.

Finally, in Chapter 5 we present a summary of the methods proposed in this dissertation and suggest extensions and refinements for them, which can be used as tools for improvement of economically important traits as well as to aid the development of breeding programs and optimal decision-making strategies.

References

- DE LOS CAMPOS, G. et al. A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. *Journal of Animal Science*, v. 84, n. 11, p. 2934-2941, 2006.
- DE MATURANA, E. L. et al. Exploring biological relationships between calving traits in primiparous cattle with a Bayesian recursive model. *Genetics*, v. 181, n. 1, p. 277-287, 2009.
- GELMAN, A. A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing. *International Statistical Review*, v. 71, n. 2, p. 369-382, 2003.
- GIANOLA, D.; SORENSEN, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics*, v. 167, n. 3, p. 1407-1424, 2004.
- HAAVELMO, T. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, p. 1-12, 1943.
- HENDERSON, C. R.; QUAAS, R. L. Multiple trait evaluation using relatives' records. *Journal of Animal Science*, v. 43, n. 6, p. 1188-1197, 1976.
- LI, R. et al. Structural model analysis of multiple quantitative traits. *PLoS Genetics*, v. 2, n. 7, p. e114, 2006.
- LOGSDON, B. A.; MEZEY, J. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology*, v. 6, n. 12, p. e1001014, 2010.
- MODE, C. J.; ROBINSON, H. F. Pleiotropism and the genetic variance and covariance. *Biometrics*, v. 15, n. 4, p. 518-537, 1959.
- MRODE, R. et al. Random regression model for the genetic evaluation of production traits of dairy cattle in the UK. *Interbull Bulletin*, n. 33, p. 211, 2005.
- NETO, E. C. et al. Inferring causal phenotype networks from segregating populations. *Genetics*, v. 179, n. 2, p. 1089-1100, 2008.
- NETO, E. C. et al. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics*, v. 4, n. 1, p. 320, 2010.
- PEARL, J. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, v. 95, n. 450, p. 428-431, 2000.
- PEARL, J. Causality. Cambridge university press, 2009.
- ROSA, G. J. M. et al. Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution*, v. 43, n. 1, p. 6, 2011.
- SCHADT, E. E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, v. 37, n. 7, p. 710-717, 2005.
- SHIPLEY, B. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R. Cambridge University Press, 2016.
- SORENSEN, D.; GIANOLA, D. Likelihood, Bayesian and MCMC Methods in Genetics. Springer, 2002.
- SPIRTES, P.; GLYMOUR, C. N.; SCHEINES, R. Causation, Prediction, and Search. MIT press, 2000.

VALENTE, B. D. et al. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, v. 185, n. 2, p. 633-644, 2010.

VERMA, T. S.; PEARL, J. Equivalence and synthesis of causal models [Technical report R-150]. Department of Computer Science, University of California, Los Angeles, 1990.

WANG, H.; VAN EEUWIJK, F. A. A new method to infer causal phenotype networks using QTL and phenotypic information. *PloS One*, v. 9, n. 8, p. e103997, 2014.

WRIGHT, S. Correlation and causation. *Journal of Agricultural Research*, v. 20, n. 7, p. 557-585, 1921.

2 CONCLUSION

Graphical models, such as Bayesian networks (BN) and structural equation models, are useful tools to express causality among phenotypic traits in complex systems biology. However, the task of fitting a causal model requires that the relationships of cause-effect among variables be pre-established. When these relationships are not known, the proposed structure using prior knowledge may not express the actual biological network, fact that result in erroneous inference of causal parameters and, consequently, biased interpretation of cause intensity between variables.

In this thesis, we proposed some ways of learning causal networks using search algorithms that require accepting specific assumptions, from which the causal sufficiency seems to be the strongest one. In Chapter 2, we explored constraint- and score-based algorithms of BN to recover the underlying phenotypic networks of two fruit species of the Sapotaceae family and concluded that these fruits have highly similar biological mechanisms. However, further studies using genetic data in the search for the causal structures of these fruits could improve the results and perhaps provide different paths among traits that would allow the fusion of the two proposed causal network into only one.

In Chapter 3, we used a real data set (F2 Duroc $\times$ Pietrain pig population) to compare three different methodologies for genome-wide association studies: a single-marker regression, a ridge regression BLUP and a Bayes $C\pi$. Methods were equally efficient in the detection of QTL regions, however we suggest the use of more than one method for GWAS. This study also allowed us to identify genomic regions, not reported in previous studies with the same pig population, associated with the expression of fat deposition and meat quality traits. In addition, we utilized the BC model for detecting major genes on some selected traits that would be used in Chapter 4 and discussed the influence of genetic confounding factors on the causal search and how their absence could bias a structural search.

Finally, in Chapter 4 the main goal of this research was reached, in which we proposed a hybrid algorithm, called PolyMaGNet (Polygenic traits with Major Genes Network analysis), which allows inferring BN underlying phenotypic traits conditional to major genes and unobservable additive (polygenic) genetic effects. Results of a simulated study considering a QTL mapping population showed that, in the presence of major genes, the PolyMaGNet was effective in recovering the correct skeleton structure and causal direction with a higher rate of true positive. PolyMaGNet was also applied to a real dataset of a F2 Duroc $\times$ Pietrain pig resource population to recover the causal structure underlying longitudinal back fat traits and traits related to meat quality, fat and chemical composition. Regarding the first application, the final causal network was compatible with the longitudinal biological profile. In the second, the resulted causal network provided an interesting scenario showing the causal effect of marbling, back fat, longissimus muscle area and pH45 to the meat tenderness index.

In this way, if high density molecular marker data is available, more reliable causal networks can be obtained through more efficient genetic learning structures approaches, such as the PolyMaGNet method. It is noteworthy that the proposed method could be improved in many ways, such as allowing to deal with non-Gaussian traits, as well as to handle huge number of variables. Both suggestions require more complex methodologies and, consequently, computational time.

In summary, graphical models provide a flexible and insightful approach which allow the characterization of causal phenotypic networks and its genetic architectures in complex systems biology. Such information can be used then to predict the effect of external interventions and, consequently, the improvement of economically important traits. As such, it might promote the development of breeding programs and optimal decision-making strategies.