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

Structural equation models applied to quantitative genetics

Pedro Henrique Ramos Cerqueira

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

Piracicaba
2015

Pedro Henrique Ramos Cerqueira
Degree in Statistics

Structural equation models applied to quantitative genetics

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:

Prof. Dr. **ROSELI APARECIDA LEANDRO**

Co-Advisor:

Prof. Dr. **GUILHERME JORDÃO DE MAGALHÃES ROSA**

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

Piracicaba
2015

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

Cerqueira, Pedro Henrique Ramos

Structural equation models applied to quantitative genetics / Pedro Henrique Ramos
Cerqueira. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - -
Piracicaba, 2015.

95 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz".

1. Inferência bayesiana 2. Modelos de equações estruturais 3. Genética quantitativa
4. Regressão polinomial 5. Modelos lineares mistos 6. Amostrador de Gibbs 7. Modelos
mistos multi característicos 8. Gado leiteiro da raça Holandesa I. Título

CDD 636.214
C411s

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

DEDICATION

Firstly to God
Without him nothing would be possible

To my Mom,
Juraci Ramos Cerqueira (in memorian) .

To my grandma,
Adelaide Ramos Cerqueira.

To my aunt,
Roseli Ramos Cerqueira.

To my lovely wife,
Camila Rodrigues Gonçalves Cerqueira .

To them,
I lovingly dedicate this work
for all the support along my journey.

ACKNOWLEDGMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis, especially to my grandma Adelaide Ramos Cerqueira, my aunt Roseli Ramos Cerqueira, my father Sérgio de Souza, for their love and supporting during all my life.

An especial thanks to my wife Camila Rodrigues Gonçalves Cerqueira, for supporting me during all academic formation, being my partner and helping me in my most important decision.

To my all family in-law, especially my father in-law Lourivalter (Seu Valter), mother in-law Maria Helena, sister in-law Isabella and grandparents in-law Joaquim and Yolanda, for all the support, and also, being so kind and lovely during those years.

To my advisor, Prof. Dr. Roseli Aparecida Leandro, for the continuous support of my all academic pathway, for her patience, motivation, enthusiasm and immense knowledge.

To my co-advisor Prof. Dr. Guilherme Jordão de Magalhães Rosa, for receiving me on my internship period and for being so helpful during my stay in Madison Wisconsin giving me all support that was needed.

To Bruno and Francisco Peñagaricano (Pancho), from University of Wisconsin, for their friendship, scientific contribution, intellectual input and especially for participating in the process of Doctorate.

To my especial friends from ESALQ, Mariana Ragassi Urbano, Rodrigo Rosseto Pescim, Luiz Ricardo Nakamura (Zan), Ana Julia Righetto, Thiago Gentil (Bem), Iuri Ferreira (Pipi) and Rafael Maia, who were always willing to help, also, for providing me a more enjoyable life in the last four years and for all discussion regarding ours carriers and “science” during the daily coffee breaks.

To my especial colleagues and friends from University of Wisconsin, Vivian Felipe, Tom Murphy, Llibertat, Ferran, Claudia, Rodrigo Pacheco, Renato Ribeiro and his family for all their support friendship helping me to have a more pleasure life during my stay in Madison.

To Prof. Dr. Taciana Villela Salvian, Prof. Dr. Cristian Marcelo Villegas Lobos and Prof. Dr. Clarice Garcia Borges Demétrio at ESALQ/USP and Prof. Dave L. Thomas at UW, for their valuable guidance and friendship. To Prof. Kent Weigel for conceding the data set for the analysis.

To my friends from the Department of Exact Science at ESALQ/USP, Cássio Dessotti, Guilherme Biz, Lucas Cunha, Edilan Quaresma, Ezequiel Lopez, Maurício Lordello, Djair (Djavan), Everton da Rocha, Thiago Oliveira, Ricardo Klein, Maria Cristina Martins, Rafael Moral, Tiago Santana, Marcus Gurgel, Everton de Toledo Hanser and Altemir.

To all my family, especially, my uncles Cid and Sérgio and their spouses Carmen and Michele, my cousins Rodrigo, Felipe and Gustavson and to my godmother Irene.

To the staff of the Department of Exact Science at ESALQ/USP, the secretaries Solange de Assis Paes Sabadin, Mayara Segatto, Rosni Pinto and Luciane Brajão, in special to the computer technicians Jorge Alexandre Wiendl and Eduardo Bonilha for always helping with techniques issues.

This work was supported by CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil.

Luke: I don't believe it,

Master Yoda: That is why you fail

Master yoda, in "The Empire Strikes Back".

SUMMARY

RESUMO	11
ABSTRACT	13
1 INTRODUCTION	15
References	17
2 GENERAL OVERVIEW	23
Abstract	23
2.1 Introduction	23
2.2 Linear Mixed models	24
2.2.1 General description	25
2.2.2 Parameter estimation	25
2.2.3 Restricted maximum likelihood	26
2.2.4 Bayesian inference: an overview	29
2.2.5 Linear mixed models in quantitative genetics	32
2.3 Causal inference	35
References	37
3 ADVANTAGES OF USING HIGHER POLYNOMIAL INSTEAD OF LINEAR RELATIONSHIP BETWEEN TRAITS IN STRUCTURAL EQUATION MOD- ELS IN QUANTITATIVE GENETICS: A SIMULATION STUDY	43
Abstract	43
3.1 Introduction	43
3.2 Material and Methods	45
3.2.1 Methods	45
3.2.2 Simulation data	48
3.2.3 Analysis	49
3.2.4 Results of fitting polynomial SEM	50
3.2.5 Results of fitting SEM with linear effects	57
3.2.6 Discussion	62
3.3 Conclusion	63
References	64
4 USING POLYNOMIAL STRUCTURAL EQUATION MODELS TO ESTIMATE THE EFFECTS RELATED TO CALVING IN PRIMIPAROUS HOLSTEIN CATTLE	67
Abstract	67
4.1 Introduction	67
4.2 Material and Methods	69
4.2.1 Data	69
4.2.2 Methods	71

4.2.3 Estimation and computation	74
4.3 Results and Discussion	75
4.3.1 Results of Multiple trait mixed models (fixed effects)	75
4.3.2 Results of linear structural models (fixed effects)	77
4.3.3 Results of second-degree polynomial structural models (fixed effects)	79
4.3.4 Results of third-degree polynomial structural models (fixed effects)	82
4.3.5 Results without causal relationships between GL and CD and second-degree polynomial structural models (fixed Effects)	84
4.3.6 Results of variances-covariances components	87
4.3.7 Discussion	88
4.4 Conclusion	89
References	89
5 CONCLUSION	95
5.1 Prospective Works	95

RESUMO

Modelos de equações estruturais aplicados à genética quantitativa

Modelos causais têm sido muitos utilizados em estudos em diferentes áreas de conhecimento, a fim de compreender as associações ou relações causais entre variáveis. Durante as últimas décadas, o uso desses modelos têm crescido muito, especialmente estudos relacionados à sistemas biológicos, uma vez que compreender as relações entre características são essenciais para prever quais são as consequências de intervenções em tais sistemas. Análise do grafo (AG) e os modelos de equações estruturais (MEE) são utilizados como ferramentas para explorar essas relações. Enquanto AG nos permite buscar por estruturas causais, que representam qualitativamente como as variáveis são causalmente conectadas, ajustando o MEE com uma estrutura causal conhecida nos permite inferir a magnitude dos efeitos causais. Os MEE também podem ser vistos como modelos de regressão múltipla em que uma variável resposta pode ser vista como explanatória para uma outra característica. Estudos utilizando MEE em genética quantitativa visam estudar os efeitos genéticos diretos e indiretos associados aos indivíduos por meio de informações relacionadas aos indivíduos, além das características observadas, como por exemplo o parentesco entre eles. Neste contexto, é tipicamente adotada a suposição que as características observadas são relacionadas linearmente. No entanto, para alguns cenários, relações não lineares são observadas, o que torna as suposições mencionadas inadequadas. Para superar essa limitação, este trabalho propõe o uso de modelos de equações estruturais de efeitos polinomiais mistos, de segundo grau ou superior, para modelar relações não lineares. Neste trabalho foram desenvolvidos dois estudos, um de simulação e uma aplicação a dados reais. O primeiro estudo envolveu a simulação de 50 conjuntos de dados, com uma estrutura causal completamente recursiva, envolvendo 3 características, em que foram permitidas relações causais lineares e não lineares entre as mesmas. O segundo estudo envolveu a análise de características relacionadas ao gado leiteiro da raça Holandesa, foram utilizadas relações entre os seguintes fenótipos: dificuldade de parto, duração da gestação e a proporção de morte perinatal. Nós comparamos o modelo misto de múltiplas características com os modelos de equações estruturais polinomiais, com diferentes graus polinomiais, a fim de verificar os benefícios do MEE polinomial de segundo grau ou superior. Para algumas situações a suposição inapropriada de linearidade resulta em previsões pobres das variâncias e covariâncias genéticas diretas, indiretas e totais, seja por superestimar, subestimar, ou mesmo atribuir sinais opostos as covariâncias. Portanto, verificamos que a inclusão de um grau de polinômio aumenta o poder de expressão do MEE.

Palavras-chave: Inferência bayesiana; Modelos de equações estruturais; Genética quantitativa; Regressão polinomial; Modelos lineares mistos; Amostrador de Gibbs; Modelos mistos multi característicos; Gado leiteiro da raça Holandesa

ABSTRACT

Structural equation models applied to quantitative genetics

Causal models have been used in different areas of knowledge in order to comprehend the causal associations between variables. Over the past decades, the amount of studies using these models have been growing a lot, especially those related to biological systems where studying and learning causal relationships among traits are essential for predicting the consequences of interventions in such system. Graph analysis (GA) and structural equation modeling (SEM) are tools used to explore such associations. While GA allows searching causal structures that express qualitatively how variables are causally connected, fitting SEM with a known causal structure allows to infer the magnitude of causal effects. Also SEM can be viewed as multiple regression models in which response variables can be explanatory variables for others. In quantitative genetics studies, SEM aimed to study the direct and indirect genetic effects associated to individuals through information related to them, beyond the observed characteristics, such as the kinship relations. In those studies typically the assumptions of linear relationships among traits are made. However, in some scenarios, nonlinear relationships can be observed, which make unsuitable the mentioned assumptions. To overcome this limitation, this paper proposes to use a mixed effects polynomial structural equation model, second or superior degree, to model those nonlinear relationships. Two studies were developed, a simulation and an application to real data. The first study involved simulation of 50 data sets, with a fully recursive causal structure involving three characteristics in which linear and nonlinear causal relations between them were allowed. The second study involved the analysis of traits related to dairy cows of the Holstein breed. Phenotypic relationships between traits were calving difficulty, gestation length and also the proportion of perinatal death. We compare the model of multiple traits and polynomials structural equations models, under different polynomials degrees in order to assess the benefits of the SEM polynomial of second or higher degree. For some situations the inappropriate assumption of linearity results in poor predictions of the direct, indirect and total of the genetic variances and covariance, either overestimating, underestimating, or even assign opposite signs to covariances. Therefore, we conclude that the inclusion of a polynomial degree increases the SEM expressive power.

Keywords: Bayesian inference; Structural equation models; Quantitative genetics; Polynomial regression; Linear mixed models; Gibbs sampler; Multiple trait mixed models; Holstein dairy cattle

1 INTRODUCTION

Studies related to genetics are extremely important and they have been increasing during the past decades, in numbers and size of records, specially in areas such as medical sciences and agronomy. The computational advancement provides new techniques for acquisition and analysis of genetic data. In medical sciences, some techniques have been used for instance to detect genes associated to diseases, and in agriculture studies have been developed in order to improve the animal and plant breeding. Molecular technologies such as SAGE, microarray and RNA-seq, have been used to identify those genes associations, allowing the discovery of complex network of biochemical processes related to living organisms, common diseases in humans, gene discovery and structure determination (SCHADT et al., 2005; HUGHES et al., 2000; KARP et al., 2003).

In medical sciences, studies related to the most common problems such as heart diseases, osteoporosis, diabetes and cancer, have been developed using genetic and environmental information. Such diseases are typically deemed as results of complex interactions of multiple genes and environmental aspects (LI et al., 2006; CHAIBUB NETO et al., 2008). Genetics associations between disease may be due to common genetic factors, or may occur due to physiological interactions. For instance, cardiovascular disease can be related to a plethora of traits, such as blood pressure, insulin levels, triglyceride and cholesterol levels (STOLL et al., 2001; NADEAU et al., 2003; SINICROPE; SARGENT, 2012).

In agronomy, specifically in animal breeding, there is an interest on studying the relationship between phenotypic traits such as growth, milk production and diseases. Generally, those studies involve others covariates, such as age, season and year, that can be common (the same) for all traits or specifics for some of them, also it is often incorporated information concerning to herds and especially family using some kinship information, for instance pedigree.

Most of genetics studies are based on traditional probabilistic models, where the response variables (i.e. traits of interest) are associated to covariates (ROSA et al., 2011). Those models are efficient in verifying how likely the occurrence of the characteristic of interest is. However, they are not efficient in predicting how the probability of a particular event can be affected by external interventions (ROSA et al., 2011; PEARL, 2009).

In studies where several traits are evaluated, the correlation between a pair of them can be estimated although such correlation does not sufficiently indicate a direct causal relationship with a specific direction among them (ROSA et al., 2011; PEARL, 2009). This concept that correlations can happen without a relationship among traits or direct causal effect, when an external factor is common for the observed traits, leads to an important statement “Correlation does not imply causation”. Pearl (2009) exemplifies complexities and how causation and correlation are related using simple examples such as the rooster crow and the sunrise. Another example is related to number of people using umbrella and car accidents, these variables are correlated to

each other, however their relationship is given by a common effect, the rain. It is also possible to find cases where one can not find a common factor among variables, some examples of spurious correlations (e.g. Per capita consumption of mozzarella cheese with civil engineering doctorates awards) can be found on <http://www.tylervigen.com/spurious-correlations>.

All those connections among traits can be seen as a network, or phenotypic network, that explains which traits modify each other, or if a trait modify more than one trait, those modifications can also be called as the effects caused by an specific trait. The effects can be positive or negative, but they do not necessarily present the same direction, for example: suppose two traits y_1 and y_2 , where y_1 influences y_2 positively, while y_2 may have a positive, negative or may not have any effect on y_1 . Rosa et al. (2011) exemplifies this influence using the fact that a high production in dairy cows can increase the chances of a certain type of disease, however, the occurrence of this disease can affect production negatively.

To study how traits are causally related and what is the magnitude of each pairwise relationship, causal inference models, or simply causal models, have been used (ROSA et al., 2011; FOX, 2008). The number of studies involving the inference of causal relationships has been increasing during the last two decades, specially in economy and social sciences, as for instance Fox (2008), Duncan et al. (1968), Ferron and Hess (2007) and Greene (2011). In genetics, those models have been used in order to discover, comprehend and explain gene networks for example Chaibub Neto (2010), Chaibub Neto et al. (2010), Rosa and Valente (2014), Gianola and Sorensen (2004), Shadt et al. (2005), Xiong et al. (2004) and others.

Two different approaches are used to investigate how variables are causally connected and to infer the magnitude of their causal relationships: the graph analysis (GA) and the structural equation models (SEM). The GA methods allow performing a search for causal structures and help visualize those relations, i.e, they express qualitatively how the variables are causally connected (LIU et al., 2008; LI et al., 2006, CERQUEIRA, et al., 2014). Some procedures to recover the causal relationships can be found in literature, e.g inductive causation (IC) and Peter-Clarck (PC) algorithms (GLYMOUR et al., 1986; PEARL, 2009; HARRIS; DRTON, 2013; ROSA et al., 2011; VALENTE et al., 2010).

Once the relationships are inferred, the structural equation model (SEM) can be applied to quantify the magnitude of the effects the variables exert on each other. These models can account for either simultaneous or recursive relations (i.e. cyclic or acyclic relations) among variables in multivariate systems. In the equations assigned to those systems, a response variable of some equations can be considered as a predictor variable in the equations for another response variables (ROSA et al., 2011; FOX, 2008; FOX, 2006; DUNCAN et al., 1968). Some specific software and packages have been developed implement SEM analyses, such as TETRAD and LISREL. The software TETRAD allows to simulate data using a prior structure and also to estimate the coefficients related to that structure as well, using maximum likelihood and other procedures (more information can be found in TETRAD manual). The LISREL is a software developed by Jöreskog and Sörbom (1974) that uses the full information maximum likelihood

as a procedure to estimate the coefficient based on latent structural models. Also packages such as `sem` and `lavaan` were developed for the open software R (FOX, 2008; FOX, 2006).

In animal breeding studies, several traits are typically measured and modelled. Generally, these studies are performed using observational data, and also information related to environmental factors, e.g farms and cities, and genetic distance, e.g pedigree, are used for predicting individual genetic effects and estimating genetic parameters. The most common approach for those situations is the linear mixed model, where fixed and random effects are modeled at the same time (LAIRD; WARE, 1982; PINHEIRO; BATES, 2000; MRODE, 2005).

The amount of studies using SEM in quantitative have increased after Gianola and Sorensen (2004), they were the forerunners proposing mixed effects structural equation models, under which the authors adapted structural equation models to the mixed models context (VALENTE et al., 2011; VALENTE et al., 2013).

Studies involving SEM in quantitative genetics generally are developed under the assumption of linear causal relationships among traits, which is not realistic in some situations (MATURANA et al., 2009; GONZÁLES-RODRIGUÉZ et al., 2014; VARONA et al., 2014; KÖNIG et al., 2008). In order to show the underlying inference mistakes of using this commonly adopted assumption, this work propose an approach using structural equation model with higher order polynomials.

Chapter 2 introduces a general overview of important aspects to the issue here tackled. Chapter 3 shows a simulation study to compare the structural equation models using standard linear and the second degree polynomial approaches. In Chapter 4, an application related to calving traits in Holstein cows is used to illustrate how estimates change under different assumptions of polynomials degrees. Chapter 5 present a general conclusion and prospective works.

References

CERQUEIRA, P.H.R.; VALENTE, B.; ROSA, G.J.M; LEANDRO, R.A. Second degree polynomial structural equation modeling using animal model: A simulation study In: INTERNATIONAL BIOMETRIC CONFERENCE, 27., 2014, Florence, **Abstracts...** Florence: IBS, 2014.

CHAIBUB NETO, E.. **Causal inference methods in statistical genetics**. Madison: University of Wisconsin, 2010. 140 p.

CHAIBUB NETO, E; FERARRA, T.C.; ATTIE, A.D.; YANDELL, B.S.. Inferring causal phenotype networks from segregating populations. **Genetics**, Baltimore, v. 179, p. 1089-1100, 2008.

CHAIBUB NETO, E.; KELLER, M.P; ATTIE, A.D.; YANDELL, B.S.. Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. **Annals Applied Statistic**, Cleveland, v. 4, p. 320-339, 2010.

DUNCAN, O.D.; HALLER, A.O.; PORTES, A.. Peer influences on aspirations: A reinterpretation. **American Journal of Sociology**, Chicago, v. 74, p. 119-137, 1968.

FERRON, J.M.; HESS, M.R. Estimation in SEM: A concrete example. **Journal of Educational and Behavioral Statistics**, Washington, v. 32, p. 110-120, 2007.

FOX, J.. **An introduction to structural equation modeling**, curso para programa de Computação Científica, FIOCRUZ: Rio de Janeiro, Brasil , 2008. p. 138.

FOX, J.. Structural equation modeling with the sem package in R. **Structural equation modeling**, Hillsdale, v. 13, p. 465-486, 2006.

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

GLYMOUR, C; SCHEINES, R.; SPIRTES, P.; KELLY, K.. **Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling**. Pittisburgh, Academic Press, 1986. 412 p.

GONZÁLEZ-RODRÍGUEZ, A.; MOURESAN, E.F.; ALTARRIBA, J.; MORENI, C.; VARONA, L.. Non-linear recursive models for growth traits in the Pirenaica beef cattle breed. **Animal : an international journal of animal bioscience**, Cambridge, v. 8, p. 904-911, 2014.

GREENE, W.H.. **Econometric analysis**, 7 ed. New York, Macmillan. 2011. 1232 p.

HARRIS, N.; DRTON, M. PC Algorithm for Nonparanormal Graphical Models **Journal of Machine Learning Research**, Cambridge, v. 14, p. 3365-3383, 2013.

HUGHES, T.R.; MARTON, M.J.; JONES, A.R.; ROBERTS, C.J.; STOUGHTON, R.; A, C.D.; BENNETT, H.A.; COFFEY, E.; DAI, H.; HE, Y.D.; KIDD, M.J.; KING, A.M.; MEYER, M.R.; SLADE, D.; LUM, P.Y.; STEPANIANTS, S.B.; SHOEMAKER, D.D.; GACHOTTE, D.; CHAKRABURTTY, K.; SIMON, J.; BARD, M.; FRIEND, S.H.. Functional discovery via a compendium of expression profiles. **Cell**, Cambridge, v. 2, p. 109-146, 2000.

JÖRESKPG, K.G.; SÖRBOM, D. **LISREL III Computer software**, Chicago, IL: Scientific Software International, Inc. 1974.

KARP, C.L.; GRUPE, A.; SCHADT, E.; EWART, S.L.; KEANE-MOORE, M.; CUOMO, P.J.; KÄHL, J.; Larry WAHL, L.; KUPERMAN, D.; GERMER, S.; AUD, D.; PELTZ, G.; WILLS-KARP, M.. Identification of complement factor 5 as a susceptibility locus experimental allergic asthma. **Nature Immunology**, New York, v. 1, p. 221-226, 2003.

KÖNIG, S.; WU, X. L.; GIANOLA, D.; HERINGSTAD, B.; SIMIANER, H.. Exploration of relationships between claw disorders and milk yield in Holstein cows via recursive linear and threshold models. **Journal of Dairy Science**, Champaign, v. 91, p. 395-406, 2008.

LAIRD, N.M.; WARE, J.H. Random-Effects Models for longitudinal Data. **Biometrics**, Alexandria, v. 38, n. 4, p. 963-974, 1982.

LI, R.; TSAIH, SHING-WERN; STYLIANOU, I. M.; WERGEDAL, J.; PAIGEN, B.; CHURCHILL, G.A.. Structural model analysis of multiple quantitative traits. **PLoS Genetics**, San Francisco, v. 2, p. 1046-1057, 2006.

LIU, B.; FUENTE, A. de la; HOESCHELE, I.. Gene network inference via structural equation modeling in genetical genomics experiments. **Genetics**, Baltimore, v. 178, p. 1763-1776, 2008.

MRODE, R. A.. **Linear Models for the Prediction of Animal Breeding Values**, 2 ed. Wallingford, Oxon, UK: CAB International, 2005. 344 p.

NADEAU, J. H.; BURRAGE, L. C.; RESTIVO, J.; PAO, Yoh-Han, CHURCHILL, G.; HOIT, B.D.. Pleiotropy, homeostasis, and functional networks based on assays of cardiovascular traits in genetically randomized populations. **Genome Research**, Cold Spring Harbor, v. 13 p. 2082-2091, 2003.

PEARL, J. **Causality Models, Reasoning and inference**. 2 ed. Cambridge, RU: Cambridge University Press, 2009. 484 p.

PINHEIRO, J. C.; BATES, D. M.. **Mixed-Effects Models in S and S-PLAS**. New York: Springer, 2000. 528 p.

R. Development Core Team. R Foundation for Statistical Computing. **R 2.15.2**: A language and environment for statistical computing, Vienna, 2012. Available: <http://www.r-project.org/>. Access em: 23 nov. 2012

ROSA, G.J.M.; VALENTE, B.D. Structural Equation Models for Studying Causal Phenotype Networks in Quantitative Genetics in SINOQUETE, C.; MOURAD, R.. **Probabilistic Graphical Models for Genetics, Genomics and Postgenomics**, Oxford University Press, Oxford, 2014. 480 p.

ROSA, G.J.M.; VALENTE, B.D.; de lo CAMPOS, G.; WU, X.L.; GIANOLA, D.; SILVA, M.A.. Inferring causal phenotype networks using structural equation models. **Genetics, Selection, Evolution**, London, v. 2, p. 1046-1057, 2011.

SCHADT, E.E.; LAMB, J.; YANG, X.; ZHU, J.; EDWARDS, S.; GUHATHAKURTA, D.; SIEBTS, S.K.; MONKS, S.; REITMAN, M.; ZHANG, C.; LUM, P.Y.; LEONARDSON, A.; THIERINGER, R.; METZGER, J.M.; YANG, L.; CASTLE, J.; ZHU, H.; KASH, S.F.; DRAKE, T.A.; SACHS, A.; LUSIS, A.J. An integrative genomics approach to infer causal associations between gene expression and disease. **Nature Genetics**, New York, v. 37, p. 710-717, 2005.

SINICROPE, F.A.; SARGENT, D.J. Molecular Pathways: Microsatellite Instability in Colorectal Cancer: Prognostic, Predictive and Therapeutic Implications. **Clinical Cancer Research**, Denville, v. 18, p. 1506-1512, 2012.

STOLL, M.; COWLEY JUNIOR, .A.W.; TONELLATO, P.J.; GREENE A.S.; KALDUNSKI, M.L.; ROMAN, R.J.; DUMAS, P.; SCHORK, N.J.; WANG, Z.; JACOB, H.J.. A Genomic-Systems biology map for cardiovascular function. **Science**, Washington, v. 294, p. 1723-1726, 2001.

The TETRAD Project Causal Models and Statistical Data. **tetrad**. Available: <http://www.phil.cmu.edu/projects/tetrad/current.html>. Access in: 23 nov. 2012

VALENTE, B.D.; ROSA, G.J.M.; CAMPOS, G. de los; GIANOLA, D.; SILVA, M.A..
Searching for recursive causal structures in multivariate quantitative genetics mixed models. **Genetics**, Baltimore, v. 185, p. 633-644, 2010.

VALENTE, B.D.; ROSA, G.J.M.; SILVA, M.A.; TEIXEIRA, R.B.; TORRES, R.A.. Searching for phenotypic causal networks involving complex traits: an application to European quails. **Genetics, Selection, Evolution**, London, v. 43, p. 37-48, 2011.

VALENTE, B.D.; ROSA, G.J.M.; GIANOLA, D.; WU, X-L; WEIGEL, K.. Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? **Genetics**, Baltimore, v. 194, p. 561-572, 2013.

VARONA, L.; SORENSEN, D.. Joint Analysis of Binomial And Continuous Traits with a Recursive Model: A Case Study Using Mortality and Litter Size of Pigs. **Genetics**, Baltimore, v. 196, p. 643-651, 2014.

VIGEN, T. **Spurious Correlations**. Available in <<http://tylervigen.com/spurious-correlations>>. Access: 2 jun. 2014 .

XIONG, M.; Li, J.; FANG, X.. Identification of genetic networks. **Genetics**, Baltimore, v. 166, p. 1037-1052, 2004.

2 GENERAL OVERVIEW

Abstract

Multivariate linear mixed models have been widely used in quantitative genetics to estimate the fixed effects and covariance components, as well as to predict random effects. Some environmental and genetic effects are typically treated as random effects related to each trait while other factors are treated as having fixed effects. However, standard mixed effects models traits do not take into account that traits may be causally related to each other. Nonetheless, several studies have verified that traits are connected because of some causal relationship. Causal effect models have been used as a tool to account for such relationships. The application of causal modeling involves two fundamentals aspects: graph analysis, to infer how traits are causally related, and structural equation modeling, to infer the magnitude of those effects. Restricted maximum likelihood and Bayesian inference approaches are typically applied for inferences. In order to have a better comprehension of aspects that will be covered in Chapter 3 and Chapter 4, a general overview associated with those concepts is presented in this chapter.

Keywords: Causal Models; Graph Models; Structural Models; Restricted Maximum Likelihood Linear Mixed Models; Quantitative Genetics; Bayesian Inference

2.1 Introduction

Studies related to quantitative genetics involve predicting genetic effects (random effects) and estimating effects related to covariates representing environmental factors (typically assumed to be fixed effects). Generally, the joint inference of those effects involves using single-trait or multiple-trait linear mixed models (MRODE, 2005; HENDERSON, 1975; ROSA; VALENTE, 2014; VALENTE et al., 2013; VALENTE et al., 2015; VALENTE; ROSA, 2013).

Although they are not expressed in the mentioned standard models, phenotypic traits can have mutually causal effects. Suppose that interventions can be made to increase milk yield. High milk production may increase the liability to certain diseases, and conversely, the incidence of a disease may affect yield negatively (ROSA et al., 2011). In order to describe causal relationships and to predict the behavior of complex systems (e.g., biological pathways underlying complex traits related to diseases, growth, and reproduction), as well as possible consequences of external interventions, knowledge of phenotype networks is crucial (ROSA, et al. 2011; ROSA; VALENTE, 2014).

To obtain information about causal relationships, the structural equation model (SEM) can be used as an alternative to traditional animal breeding models (ROSA et al., 2011; WRIGHT, 1921; HAAVELMO, 1943). This model allows one to study recursive and simultaneous rela-

tionships among traits in multivariate systems. According to Rosa et al. (2011), Pearl (2009) and Shipley (2004), SEM can be used as a general model to estimate and verify the relationships between traits. Before fitting SEM, it is necessary to define qualitatively how traits are causally connected, whether by using prior knowledge or alternative data-driven search procedures. This information forms the “causal structure”, and it is generally represented by a directed graph.

This chapter reviews some concepts that are important to the development of this thesis. In Section 2.1, an overview related to linear mixed models is presented, and Section 2.1.1 contains a general description. Section 2.1.2 introduces estimations methods and Section 2.1.3 presents some applications of mixed models in quantitative genetics. Finally, Section 2.2 presents a review of causal models.

2.2 Linear Mixed models

Studies in quantitative genetics usually take into account information related to environmental factors, e.g., farms and cities, and genetic information, e.g., pedigree, for predicting individual genetic effects and estimating genetic parameters. The most common approach used in these situations is a linear mixed model, where fixed and random effects are modeled jointly (LAIRD; WARE, 1982; PINHEIRO; BATES, 2000; MRODE, 2005)

An advantage of using mixed models is that they provide a flexible and powerful tool for the analysis of data. Observations are grouped over some average effects with random deviations from it, such that some dependency among observations in the same group is accounted for, which can be common in many diverse areas, such as agriculture, biology, economics and genetics. Examples of clustered data are longitudinal or family member studies. The latter is common in genetic studies (PINHEIRO; BATES, 2000; ROSA; VALENTE, 2014).

Given their flexibility, mixed models have been used to make inferences about environmental effects, genetic parameters and variance components. They are also able to handle complex pedigrees, unequal family sizes, overlapping generations, sex-limited traits, assortative mating, and natural or artificial selection (MRODE, 2005; HENDERSON, 1975; ROSA; VALENTE, 2014).

2.2.1 General description

According to Laird (1982), Pinheiro and Bates (2000), Rosa and Valente (2014) a mixed linear model can be presented as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad (2.1)$$

where $\mathbf{y}$ is the response vector, with dimension $(n \times 1)$, $\boldsymbol{\beta}$ is a vector of fixed parameters with dimension $(p \times 1)$, $\mathbf{b}$ is a vector of unknown random effects with dimension $(q \times 1)$, $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices with dimension $(n \times p)$ and $(n \times q)$ related to $\boldsymbol{\beta}$ and $\mathbf{b}$, respectively, and $\boldsymbol{\varepsilon}$ is the vector of residual terms with dimension $(n \times 1)$. Usually, it is assumed that $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ are independent and normally distributed with mean zero and variance-covariance matrices equal to $\mathbf{G}$ and $\mathbf{R}$, respectively.

One of the main goals of linear mixed model applications in animal and plant breeding is to predict random effects, especially the genetic merit, or breeding values. The predictions are given by the conditional expectation of $\mathbf{b}$ given the data, $E(\mathbf{b}|\mathbf{y})$. The joint distribution of $\mathbf{y}$ and $\mathbf{b}$ is a multivariate normal such as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{b} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{V} & \mathbf{Z}\mathbf{G} \\ \mathbf{G}\mathbf{Z}' & \mathbf{G} \end{bmatrix} \right) \quad (2.2)$$

where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, and following multivariate normal distributions properties, $E(\mathbf{b}|\mathbf{y})$ is given by

$$\begin{aligned} E(\mathbf{b}|\mathbf{y}) &= E(\mathbf{b}) + \text{Cov}(\mathbf{b}, \mathbf{y}') \text{Var}^{-1}(\mathbf{y})(\mathbf{y} - E(\mathbf{y})), \\ &= \text{Cov}(\mathbf{b}, \mathbf{y}') \text{Var}^{-1}(\mathbf{y})(\mathbf{y} - E(\mathbf{y})), \\ &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \quad (2.3)$$

When $\mathbf{R} = \sigma^2\mathbf{I}$ and $\mathbf{Z} = 0$, the mixed model in equation 2.1, is reduced to a standard linear model, where the residual terms are assumed independent and there are no other random effects.

2.2.2 Parameter estimation

Assuming that $\mathbf{G}$ and $\mathbf{R}$ are known and $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ are normally distributed, the density of the distribution of $\mathbf{y}$ is given by

$$f(\mathbf{y}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2} |\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}|^{1/2}} \exp \left\{ \frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))] \right\}, \quad (2.4)$$

where $\boldsymbol{\theta}$ is the vector of parameters $(\mathbf{b}, \boldsymbol{\beta}, \mathbf{G})$ and the joint probability density function, $f(\mathbf{y}, \mathbf{b}) = f(\mathbf{y}|\mathbf{b})f(\mathbf{b})$ is given by

$$\begin{aligned} f(\mathbf{y}, \mathbf{b}) &= \frac{1}{(2\pi)^{n/2} |\mathbf{R}|^{1/2}} \exp \left\{ \frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})' \mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})] \right\} \\ &\times \frac{1}{(2\pi)^{q/2} |\mathbf{G}|^{1/2}} \exp \left\{ \frac{1}{2} \mathbf{b}' \mathbf{G}^{-1} \mathbf{b} \right\}. \end{aligned} \quad (2.5)$$

The logarithm of equation 2.5 is given by

$$\begin{aligned}\ell(\mathbf{y}, \mathbf{b}) &= \frac{1}{2}2n \log(2\pi) - \frac{1}{2}(\log |\mathbf{R}| + \log |\mathbf{G}|) \\ &- \frac{1}{2}(\mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - 2\mathbf{y}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} + 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} \\ &+ \boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{b}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} + \mathbf{b}'\mathbf{G}^{-1}\mathbf{b})\end{aligned}\quad (2.6)$$

Deriving equation 2.6 in $\boldsymbol{\beta}$ and $\mathbf{b}$ and equating to 0 yields

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}.\quad (2.7)$$

From equation 2.7 it possible to obtain the best linear unbiased predictor (BLUP) of $\mathbf{b}$, given by

$$\hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),\quad (2.8)$$

and it is also possible to obtain the best linear unbiased estimator (BLUE) of $\hat{\boldsymbol{\beta}}$, given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \{\mathbf{X}'[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}]\mathbf{X}\}^{-1} \\ &\times \mathbf{X}'[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}]\mathbf{y}\end{aligned}\quad (2.9)$$

It is possible to verify by some algebraic manipulation that the estimates of equations 2.8 and 2.9 are the same as the estimates presented in equation 2.3.

2.2.3 Restricted maximum likelihood

The restricted maximum likelihood (REML) method, developed in 1971 by Patterson and Thompson under the assumptions of normal distribution, has been widely used to estimate variance components in mixed models, because it takes into account the degrees of freedom involved in estimating the fixed parameters, providing a less biased estimate than the maximum likelihood (ML) estimate (PATTERSON; THOMPSON, 1971; HARVILLE, 1977; GILMOUR; THOMPSON; CULLIS, 1995). The REML maximizes the joint likelihood function of all contrasts of $\mathbf{y}^* = \mathbf{L}'\mathbf{y}$, where $\mathbf{L}$ is a full rank matrix with $n - \text{rank}(\mathbf{X})$ columns, and its columns are orthogonal to the column space of $\mathbf{X}$, i.e $\mathbf{L}'\mathbf{X} = \mathbf{0}$. Thus the REML is a method that maximize the part of the maximum likelihood function that is invariant to fixed effects. Let $\mathbf{L} = [\mathbf{L}_1\mathbf{L}_2]$, where

$$\mathbf{L}_1'\mathbf{X} = \mathbf{I}_p \text{ and } \mathbf{L}_2'\mathbf{X} = \mathbf{0},$$

and let $\mathbf{y}_j^* = \mathbf{L}_j'\mathbf{y}$, $j = 1, 2$ then $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$ can be rewritten as

$$\begin{aligned}\mathbf{y}_1^* &= \mathbf{L}_1' \mathbf{y} \\ &= \mathbf{I}_p \boldsymbol{\beta} + \mathbf{L}_1' \mathbf{Z} \mathbf{b} + \mathbf{L}_1' \boldsymbol{\varepsilon},\end{aligned}$$

and

$$\begin{aligned}\mathbf{y}_2^* &= \mathbf{L}_2' \mathbf{y} \\ &= \mathbf{L}_2' \mathbf{Z} \mathbf{b} + \mathbf{L}_2' \boldsymbol{\varepsilon},\end{aligned}$$

consequently,

$$\mathbb{E}(\mathbf{y}^*) = \mathbb{E} \left[\begin{pmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{pmatrix} \right] = \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} \quad (2.10)$$

and

$$\text{Var}(\mathbf{y}^*) = \text{Var} \left[\begin{pmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{pmatrix} \right] = \begin{pmatrix} \mathbf{L}_1' \mathbf{V} \mathbf{L}_1 & \mathbf{L}_1' \mathbf{V} \mathbf{L}_2 \\ \mathbf{L}_2' \mathbf{V} \mathbf{L}_1 & \mathbf{L}_2' \mathbf{V} \mathbf{L}_2 \end{pmatrix}, \quad (2.11)$$

thus,

$$\begin{pmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{pmatrix} \sim N \left[\begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{L}_1' \mathbf{V} \mathbf{L}_1 & \mathbf{L}_1' \mathbf{V} \mathbf{L}_2 \\ \mathbf{L}_2' \mathbf{V} \mathbf{L}_1 & \mathbf{L}_2' \mathbf{V} \mathbf{L}_2 \end{pmatrix} \right] \quad (2.12)$$

The complete distribution $\mathbf{L}' \mathbf{y}$ can be split into a conditional distribution $\mathbf{y}_1^* | \mathbf{y}_2^*$ and a marginal distribution of $\mathbf{Y}_2^*$ to estimate $\boldsymbol{\beta}$ and the variance components, respectively.

Assuming that $\mathbf{k}' = (\gamma', \phi')$ is the vector of variance components related to $\mathbf{b}$ and $\boldsymbol{\varepsilon}$, respectively, their likelihood logarithms are given by

$$\begin{aligned}\ell_R &= -\frac{1}{2} [\log \det(\mathbf{L}_2' \mathbf{V}^{-1} \mathbf{L}_2) + \mathbf{y}_2'^* (\mathbf{L}_2' \mathbf{V}^{-1} \mathbf{L}_2)^{-1} \mathbf{y}_2^*] \\ &\quad - \frac{1}{2} \log \det(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}) + \log \det \mathbf{V} + \mathbf{y}' \mathbf{P} \mathbf{y},\end{aligned} \quad (2.13)$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1}, \quad (2.14)$$

and

$$\mathbf{y}' \mathbf{P} \mathbf{y} = (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}). \quad (2.15)$$

The REML estimates for k_l where $\mathbf{k} = (k_1, k_2, \dots, k_L)$ can be obtained by solving for k the score function given by

$$U(k_l) = \frac{\partial l_R}{\partial k_l} = -\frac{1}{2} \left[\text{tr} \left(\mathbf{P} \frac{\partial \mathbf{V}}{\partial k_l} \right) - \mathbf{y}' \mathbf{P} \frac{\partial \mathbf{V}}{\partial k_l} \mathbf{P} \mathbf{y} \right], \quad (2.16)$$

and the elements of the observed and expected information matrix are give by

$$\begin{aligned} -\frac{\partial^2 l_R}{\partial k_l \partial k_k} &= \frac{1}{2} \text{tr} \left(\mathbf{P} \frac{\partial^2 \mathbf{V}}{\partial k_l \partial k_k} \right) - \frac{1}{2} \text{tr} \left(\mathbf{P} \frac{\partial \mathbf{V}}{\partial k_l} \mathbf{P} \frac{\partial \mathbf{V}}{\partial k_k} \right) \\ &+ \mathbf{y}' \mathbf{P} \frac{\partial \mathbf{V}}{\partial k_l} \mathbf{P} \frac{\partial \mathbf{V}}{\partial k_k} \mathbf{P} \mathbf{y} - \frac{1}{2} \mathbf{y}' \mathbf{P} \frac{\partial^2 \mathbf{V}}{\partial k_l \partial k_k} \mathbf{P} \mathbf{y}, \end{aligned} \quad (2.17)$$

and,

$$\mathbb{E} \left(-\frac{\partial^2 l_R}{\partial k_l \partial k_k} \right) = \frac{1}{2} \text{tr} \left(\mathbf{P} \frac{\partial \mathbf{V}}{\partial k_l} \mathbf{P} \frac{\partial \mathbf{V}}{\partial k_k} \right). \quad (2.18)$$

However, to solve $U(k_k) = 0$, it is necessary to use an iterative algorithm. Thus, given an initial starting value $\mathbf{k}^{(0)}$, a new value $\mathbf{k}^{(1)}$ is obtained using the Fisher score algorithm, given by

$$\mathbf{k}^{(1)} = \mathbf{k}^{(0)} + \mathbf{I}(\mathbf{k}^{(0)}, \mathbf{k}^{(0)})^{-1} \mathbf{U}(\mathbf{k}^{(0)}) \quad (2.19)$$

where $\mathbf{U}(\mathbf{k}^{(0)})$ is the score vector presented in equation 2.16 and $\mathbf{I}(\mathbf{k}^{(0)}, \mathbf{k}^{(0)})$ represents the expected information matrix of $\mathbf{k}$ presented in equation 2.18, evaluated in $\mathbf{k}^{(0)}$.

When using big datasets or data with high dimension, the evaluation of the traces in equations 2.17 and 2.18 can be unfeasible or computationally intensive. For those cases Gilmour, Thompson and Cullis (1995) propose the average information (AI) algorithm which has convergence properties like the Fisher score algorithm, although avoiding the high computational effort.

The AI algorithm essentially works with a modified form of the expected information matrix. Instead of $\mathbf{I}(k_l, k_k)$ it is used $\mathbf{I}_A(k_l, k_k)$ where

$$\mathbf{I}_A(k_l, k_k) = \frac{1}{2} \mathbf{y}' \mathbf{P} \frac{\partial \mathbf{V}}{\partial k_l} \mathbf{P} \frac{\partial \mathbf{V}}{\partial k_k} \mathbf{P} \mathbf{y}. \quad (2.20)$$

Even though other algorithms exist for maximization, such as the EM algorithm, the details regarding them are not shown in this work, since only the AI is used in Chapter 4.

2.2.4 Bayesian inference: an overview

Bayesian inference allows modeling and expressing the parameters uncertainty using prior information, via known probability distributions. The Bayesian approach combines two sources of information: One is the data evidence, which is expressed by the likelihood function, and the other is information from prior knowledge of the model's unknown quantities, represented by the prior distribution. Updating the prior knowledge after considering the data evidence from the likelihood function involves combining both types of information according to Bayes theorem, obtaining 2.21.

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \quad (2.21)$$

where $P(y|\theta)$ represents the likelihood, $P(\theta)$ represents the prior distribution, and $P(\theta|y)$ is the so-called posterior distribution. $P(y)$ for the continuous and discrete cases can be constructed as in 2.22 and 2.23, respectively.

$$P(y) = \int_{\theta} P(y, \theta) d\theta = \int_{\theta} P(\theta|y)P(\theta) d\theta \quad (2.22)$$

and

$$P(y) = \sum_{\theta} P(y, \theta) = \sum_{\theta} P(\theta|y)P(\theta). \quad (2.23)$$

The function $P(y)$ is independent of θ , and therefore can be considered a constant for the posterior distribution. For this reason, equation 2.21 can be expressed as

$$P(\theta|y) \propto P(y|\theta)P(\theta). \quad (2.24)$$

According to Box and Tiao (1992), the symbol of proportionality, represented by $\propto$, holds because the information from $p(y)$ does not contribute to the parameter's posterior distribution.

The prior distribution expresses the uncertainty regarding the parameters before observing the data. Such information can be obtained by asking a specialist (or a researcher), or from results provided by previous studies (BOX; TIAO, 1992; PAULINO; TURKMAN; MURTEIRA, 2003).

The prior distribution can be considered informative or non-informative. The former applies when there is a sharper, more definite previous knowledge about the parameter, while latter is applied when there is no precise prior parametric information.

2.2.4.1 Conjugate prior distribution

Among the prior distributions widely used, one that should be stressed is the conjugate prior distribution. For a given model, providing a conjugate prior distribution to Bayes theorem results in a posterior distribution from the same family. Actually, the kernel is the same when we use conjugate priors.

Let $\mathbf{y}' = (y_1, y_2, \dots, y_n)$ be a vector of i.i.d. observational random variables of the exponential family. The joint distribution is given by

$$f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^n \exp\{a(\boldsymbol{\theta})b(y_i) + c(\boldsymbol{\theta}) + d(y_i)\}, \quad (2.25)$$

where $\boldsymbol{\theta}$ represent the vector of parameters, and the likelihood function can be represented by

$$L(\boldsymbol{\theta}|\mathbf{y}) \propto \exp \left\{ a(\boldsymbol{\theta}) \sum_{i=1}^n b(y_i) + nc(\boldsymbol{\theta}) \right\} \quad (2.26)$$

where $a(\boldsymbol{\theta})$ and $c(\boldsymbol{\theta})$ are real functions of $\boldsymbol{\theta}$, $b(y_i)$ and $d(y_i)$ are real functions of $\mathbf{y}$. Assuming that the conjugate prior distribution for $\boldsymbol{\theta}$ is given by

$$P(\boldsymbol{\theta}; k_1, k_2) \propto \exp\{k_1 a(\boldsymbol{\theta}) + k_2 c(\boldsymbol{\theta})\} \quad (2.27)$$

then the posterior distribution is of the same family, given by

$$P(\boldsymbol{\theta}|\mathbf{y}) \propto \exp \left\{ a(\boldsymbol{\theta}) \left[\sum_{i=1}^n b(y_i) + k_1 \right] + c(\boldsymbol{\theta})[n + k_2] \right\}. \quad (2.28)$$

According to Gamerman and Lopes (2006) the conjugate prior distributions are very important and useful, although they, should be used carefully, since in some scenarios they may not be able to suitably represent the prior parameter knowledge.

2.2.4.2 Non-informative priors

Non-informative priors are used when there is no precise knowledge about the parameters or when the accounting for prior information is considered as unimportant. In this situations, the data information are the most important for the posterior distribution, i.e this distribution will be more similar to the likelihood (GELMAN et al., 2000).

Assuming that any parameter can occur with equally chance a method to assign non-informative prior distribution (e.g, the uniform distribution, $P(\boldsymbol{\theta}) = \frac{1}{b-a}$, which can be seen as $P(\boldsymbol{\theta}) \propto k$, since there is no parameter dependence). Nevertheless, using such priors sometimes can result in difficulties, such as $p(\boldsymbol{\theta})$ might improper, i.e, $\int P(\boldsymbol{\theta})d\boldsymbol{\theta} \rightarrow \infty$. To avoid these difficulties, a class of invariant non-informative prior distributions based on the Fisher's information was proposed by Jeffreys (1961).

2.2.4.3 Computational methods

To infer any parameter in θ it is necessary to integrate the joint posterior distribution in relation to all the remaining parameters. In other words, the aim is to obtain the marginal posterior distribution for each parameter (PAULINO et al., 2003; BOX and TIAO, 1992). It is usually complicated to obtain the marginal distribution using an analytic approaches, due to the complexity of the joint distributions or the highly dimensional structure of the vector of parameters. Therefore, it is necessary, for example, to apply computational methods based on Markov chain process to obtain a sample of the posterior distribution from which samples of the posterior marginal distribution can be easily obtained.

2.2.4.4 Markov chains

A Markov chain is a stochastic process where any step of a chain ϕ_t is obtained conditionally on the information present in the previous step, ϕ_{t-1} , and the data, so each step does not take into account all the historical information in the chain (GAMERMAN; LOPES, 2006). In this process, the first iterations are influenced by the choice of the initial value ϕ_1 ; hence, this information should be discarded in order to eliminate this dependence (burn-in). Furthermore, it is possible to observe dependency among the iterations, expressed as an auto-correlation among them. To mitigate the auto-correlation, one might consider only equally spaced data points, i.e., values sampled each k iterations (thinning).

The main idea of the Markov Chain Monte Carlo (MCMC) process is to obtain a sample of the parameters joint distribution via an iterative process. Each updating cycle generates values, which are considered random samples from the joint probability distribution. The most popular sampling algorithms used in Bayesian inference are the Gibbs Sampler and the Metropolis-Hastings algorithm.

2.2.4.5 Gibbs sampler

Geman and Geman (1984) described the Gibbs sampling algorithm in the context of image restoration, and since then many studies have been carried out in a wide range of research areas (GELFAND; SMITH, 1990; GELFAND, 2000). In the Bayesian inference context, this allows one to generate samples of a joint posterior distribution $p(\theta|\mathbf{y})$, using the fully conditional distributions for each of the parameters $p(\theta_i|\theta_{-i}, \mathbf{y})$. However, for this method to be feasible, the complete conditional posterior distributions should have closed form, i.e., known distributions (CASELLA; GEORGE, 1992; GELFAND, 2000).

According to Gamerman and Lopes (2006), the general Gibbs sampling procedure consists of the following steps

1. Define the parameters initial values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)})$;
2. Sample interactively from θ_i to complete a transition of $\theta^{(0)}$ to $\theta^{(1)}$

$$\begin{aligned}
& p(\theta_1^{(1)} | \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_p^{(0)}, \mathbf{y}) \\
& p(\theta_2^{(1)} | \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_p^{(0)}, \mathbf{y}) \\
& p(\theta_3^{(1)} | \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_p^{(0)}, \mathbf{y}) \\
& \vdots \\
& p(\theta_p^{(1)} | \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{p-1}^{(1)}, \mathbf{y})
\end{aligned}$$

3. Repeat the second step exhaustively, until the number of k samples for each parameters is achieved, i.e. when $\boldsymbol{\theta}^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \dots, \theta_p^{(k)})$ is sampled.

The set of k values represents a sample of the joint posterior distribution from $\boldsymbol{\theta}$, which expresses the vector of p parameters. Using the information from the samples it is possible to obtain the point estimates such as the posterior mean, median and mode, as well as achieve interval estimates, such as the High Posterior Density (HPD) and the Credible Interval (CI).

2.2.4.6 Metropolis-Hastings algorithm

In some situations neither the joint posterior distribution nor the fully conditional posterior distribution has closed form. As these features forbids using the Gibbs sampling, one alternative method to obtain posterior samples is the Metropolis-Hastings algorithm. The central idea is sampling a value from a candidate (proposal) or auxiliary distribution and then accepting or rejecting it with a specific probability (METROPOLIS et al., 1953; HASTINGS, 1970).

The Metropolis-Hastings algorithm can be structured in steps as follows:

1. Initialize the iteration counter at $t = 0$ and attribute initial value $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)})$;
2. Generate a value of θ^c from the proposed distribution $q(.|\theta)$;
3. Calculated the acceptance probability $\alpha(\theta_1, \theta^c)$

$$\alpha(\theta_1, \theta^c) = \min \left(1, \frac{p(\theta^c | \theta_2, \theta_3, \dots, \theta_p) q(\theta_1 | \theta^c)}{p(\theta_1 | \theta_2, \theta_3, \dots, \theta_p) q(\theta^c | \theta_1)} \right);$$
4. Generate a random value u from a uniform distribution $U(0, 1)$
5. If $u < \alpha$, then accept the candidate value and upgrade $\theta_1^{(t+1)} = \theta^c$. Otherwise, reject the value and upgrade $\theta_1^{(t+1)} = \theta_1^{(t)}$;
6. Modify t to $t + 1$ and start the step 2 until convergence is achieved.

2.2.5 Linear mixed models in quantitative genetics

The linear mixed model is very often used in the context of quantitative genetics and animal breeding, where it can be applied with different approaches. Among different applications are the “sire model”, “animal model” and others (ROSA; VALENTE, 2014). The simplest cases

involve only one phenotypic trait and a single observation per subject, in which case the animal model can be represented as in equation 2.1. The vector $\mathbf{y}$ represent the observations and $\boldsymbol{\varepsilon}$ contains the residual effects, which are commonly assumed independent across animals.

The aforementioned distribution for the residuals indicates that the residual covariance structure can be expressed as $\mathbf{R} = \mathbf{I}\sigma_\varepsilon^2$, where $\mathbf{I}$ is an identity matrix, and σ_ε^2 is the residual variance. Environmental factors, such as herds, age, year, sex, season of birth and others, that can affect the phenotypes $\mathbf{y}$ are in most cases assumed as fixed effects, and are represented by the vector $\boldsymbol{\beta}$. The $(q \times 1)$ vector of random effects $\mathbf{b}$ represents the breeding values not only for the n animals with known phenotypes but also for the remaining animals in the pedigree, in which case q will be bigger than n (ROSA; VALENTE, 2014).

The covariance among the breeding values is represented by $\mathbf{G}$. The additive genetic covariance between relatives i and i' is given by $2w_{ii'}\sigma_a^2$, where $w_{ii'}$ is the coefficient of coancestry between individuals i and i' and σ_a^2 is the additive genetic variance in the base population (WRIGHT, 1921). In the animal model the matrix $\mathbf{G}$ is considered as $\mathbf{A}\sigma_a^2$, where $\mathbf{A}$ is known as a additive genetic relationship matrix, having elements given by $a_{ii'} = 2w_{ii'}$. Then replacing $\mathbf{G}^{-1} = \mathbf{A}^{-1}\sigma_a^{-2}$ and $\mathbf{R}^{-1} = \mathbf{I}\sigma_\varepsilon^{-2}$, in equation (2.7), the mixed multivariate equation (MME) can be reduced to

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \quad (2.29)$$

and then

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \quad (2.30)$$

where $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_a^2} = \frac{1-h^2}{h^2}$, and the quantity h^2 represent the heritability of the trait, i.e. the proportion of the total phenotypic variance that is due to additive genetic effects. For this model, the heritability is computed as $h^2 = \frac{\sigma_a^2}{\sigma_\varepsilon^2 + \sigma_a^2}$. The matrix $\mathbf{A}^{-1}$ can be directly constructed from the pedigree information, and therefore inverting the typically large $\mathbf{A}$ is not required (HENDERSON et al., 1959; HENDERSON; QUAAS, 1976; ROSA; VALENTE, 2014)

Under a Bayesian approach, commonly the joint posterior distribution for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{b}, \sigma_a^2, \sigma_\varepsilon^2)$, when using the animal model is given by

$$P(\boldsymbol{\beta}, \mathbf{b}, \sigma_a^2, \sigma_\varepsilon^2) \propto P(\boldsymbol{\beta})P(\mathbf{b}|\sigma_a^2)P(\sigma_a^2)P(\sigma_\varepsilon^2) \quad (2.31)$$

where $P(\boldsymbol{\beta})$ is a uniform distribution, $P(\mathbf{b}|\sigma_a^2)$ is a normal distribution $N(\mathbf{0}, \mathbf{A}\sigma_a^2)$, $P(\sigma_a^2)$ and $P(\sigma_\varepsilon^2)$ are scaled inverse chi-squared distributions, $\text{Inv} - \chi^2(v_\bullet, S_\bullet^2)$ where $v_\bullet$ and $S_\bullet^2$ are the specific degrees of freedom and scale parameters. The joint posterior is given by

$$P(\boldsymbol{\beta}, \mathbf{b}, \sigma_a^2, \sigma_\varepsilon^2 | \mathbf{y}) \propto P(\mathbf{y} | \boldsymbol{\theta}, \mathbf{b}, \sigma_\varepsilon^2)P(\boldsymbol{\beta})P(\mathbf{b}|\sigma_a^2)P(\sigma_a^2)P(\sigma_\varepsilon^2). \quad (2.32)$$

It is easy to verify that the joint posterior distribution in 2.32 do not have a closed form, but the full conditionals do. In this case it is possible to see that $\mathbf{P}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma_a^2, \sigma_\varepsilon^2)$ and $\mathbf{P}(\mathbf{b}|\mathbf{y}, \boldsymbol{\beta}, \sigma_a^2, \sigma_\varepsilon^2)$ are normal multivariate distributions, $\mathbf{P}(\sigma_a^2|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \sigma_\varepsilon^2)$ and $p(\sigma_\varepsilon^2|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \sigma_a^2)$ are both scaled inverse chi-square.

In quantitative genetics it is common to observe more than one trait per subject, a situation in which the animal model can be extended to a multiple trait model (ROSA and VALENTE, 2014; HENDERSON and QUAAS, 1976; SCHAEFFER, 1984). Suppose an example where k traits are observed for each subjected, so that 2.1 can be rewritten as

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{Z}_j \mathbf{b}_j + \boldsymbol{\varepsilon}_j, \quad (2.33)$$

where $\mathbf{y}_j, \mathbf{X}_j, \boldsymbol{\beta}_j, \mathbf{Z}_j, \mathbf{b}_j$ and $\boldsymbol{\varepsilon}_j$ follows the same definitions that were used before and the index j represent the trait ($j = 1, 2, \dots, k$). The mixed linear model that jointly accounts for the k traits is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad (2.34)$$

where $\mathbf{y} = [\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_k]$, $\boldsymbol{\beta} = [\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \dots, \boldsymbol{\beta}'_k]$, $\mathbf{b} = [\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_k]$ and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}'_1, \boldsymbol{\varepsilon}'_2, \dots, \boldsymbol{\varepsilon}'_k]$. The incidence matrices in this situation are given by

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{Z}_k \end{bmatrix}$$

and

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_k \end{bmatrix}.$$

In addition, it is assumed that the variance of $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ are

$$\text{Var} \begin{bmatrix} \mathbf{b} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_0 \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{E} \otimes \mathbf{I} \end{bmatrix},$$

where

$$\mathbf{G}_0 = \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_2} & \dots & \sigma_{a_1 a_k} \\ \sigma_{a_1 a_2} & \sigma_{a_2}^2 & \dots & \sigma_{a_2 a_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{a_1 a_k} & \sigma_{a_2 a_k} & \dots & \sigma_{a_k}^2 \end{bmatrix},$$

and

$$\mathbf{E} = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \sigma_{\varepsilon_1\varepsilon_2} & \cdots & \sigma_{\varepsilon_1\varepsilon_k} \\ \sigma_{\varepsilon_1\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \cdots & \sigma_{\varepsilon_2\varepsilon_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\varepsilon_1\varepsilon_k} & \sigma_{\varepsilon_2\varepsilon_k} & \cdots & \sigma_{\varepsilon_k}^2 \end{bmatrix},$$

represent the genetic matrix and the residual variance covariance matrices, respectively, where $\mathbf{A}$ and $\mathbf{I}$ are the relationship matrix and an identity matrix, respectively.

It is possible to express graphically the multiple trait animal model (MTAM) presented in equation (2.34). For example, in Rosa and Valente (2014) the authors show a graph as an example taking into account three traits, which can be seen in Figure 2.1.

Figure 2.1 – Multiple trait animal model expressed using a graphical method

It is possible to express the MME for multiple traits as

$$\begin{bmatrix} \mathbf{X}'(\mathbf{E}^{-1} \otimes \mathbf{I})\mathbf{X} & \mathbf{X}'(\mathbf{E}^{-1} \otimes \mathbf{I})\mathbf{Z} \\ \mathbf{Z}'(\mathbf{E}^{-1} \otimes \mathbf{I})\mathbf{X} & \mathbf{Z}'(\mathbf{E}^{-1} \otimes \mathbf{I})\mathbf{Z} + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'(\mathbf{E}^{-1} \otimes \mathbf{I})\mathbf{y} \\ \mathbf{Z}'(\mathbf{E}^{-1} \otimes \mathbf{I})\mathbf{y} \end{bmatrix}, \quad (2.35)$$

using equation (2.7) as reference, the BLUP and BLUE of $\boldsymbol{\beta}$ and $\mathbf{b}$ can be obtained by solving the MME for the above multiple traits above.

Under Bayesian inference context, the MTAM is very similar to the single trait model that were presented previously. The difference comes from the prior distributions assigned to the dispersion parameters, where instead of scaled inverse chi-squared distributions, Wishart distributions are applied instead.

Some studies involve repeated measurements related to the same trait, or traits with maternal effects. In these situations, extensions or variations of the animal models are applied, for the contexts of either single or multiple traits, as can be seen in (SORENSEN; GIANOLA, 2002; ROSA; VALENTE, 2014; HENDERSON, 1984; MRODE, 2005).

2.3 Causal inference

The description of relationships between many variables requires using the concept of cause and effect, as for instance in Valente et al (2010) and Valente et al (2011), where the authors show

a method to search for causal structures and an application to European quails. Some of those relationships can be easily verified (and therefore easily accepted) using simple situations, as in Pearl (2009), like the relationship between switch position and the working status of a lamp or the fact of using a hose and having a wet roof.

However, other causal relationships can be harder to accept and to verify, such as the efficiency of a weight gain treatment, teaching techniques, pest control via pesticides, smoking and cancer, among many examples (FOX, 2006; FOX, 2008; FERRON and HESS, 2007; GREENE, 2011; DUNCAN; HALLER; PORTES, 1968). Also, for some situations the assumption of causal relations should be analyzed carefully, since external factors can exert some influence on more than one variable at the same time, leading to spurious causal relationship, such as the number of people using umbrellas and car accidents, in which case rainfall easily can be seen as a common factor for both variables.

In this sense to verify if causal statements can be considered true, it is possible to conduct an experiment and randomize the treatments of interest, but in some situations randomization it is not possible, for various reasons as for example

- Operational Issues: Due to financial or logistical reasons, e.g., an experiment is too expensive or there is not enough time available to run the randomization properly.
- Ethical Problems: A common problem related to research developed in medical science, where the randomization cannot be done, for example when some treatments can be too risky for the patient (e.g., forcing a group of patients to smoke or eat a high-calorie diet)
- Observational Variables: Variables that can be observed, but not externally defined, in which cases the randomization is impossible to perform, for instance gender, number of offspring and gestation length

These situation cited above can often occur and even though several studies concerning on explaining causal relationship among variables (e.g traits), usually are developed using traditionally probabilistic models, that relates response variables to covariates (ROSA et al., 2011). In order explain these causal connections, causal inference methods were developed.

During the past fifty years this method has been widely used in different areas of human science, such as in studies related to economics, sociology and psychology. In economics sometimes there is interest in studying the market behavior in response to interventions, which usually are not performed randomly. In sociology and psychology causal studies have been used as a tool to explain many aspects of human behavior (BOLLEN, 1989; DUNCAN; HALLER; PORTES, 1968; HAAVELMO, 1943; FOX, 2008).

In quantitative genetics where several traits are measured simultaneously and randomization cannot be performed, the application of this method has been increasing, especially after Gianola and Sorensen (2004). However, the causal inference technique should be carefully applied because statistics interpreted as representing a causal effect might be only a conditional

association among the variables that do not reflect an effect, leading to spurious inferences. The well-known caveat is “Correlation does not imply causation” (PEARL, 2009; ROSA et al., 2011; ROSA; VALENTE, 2014; VALENTE et al. 2013).

Many other examples of common effects can be given besides the umbrella use and car accident case. For instance, a correlation exists between the shoe size of primary school students and mathematics ability, but this does not mean children with big feet are better at calculating. The common cause is age: older students are more advanced in math and have bigger feet. For this reason, in studies related to causal modeling two aspects are extremely important to verify

- i) How are the effects: The factors that are involved and how they are related to each other;
- ii) How much: The magnitude of causal relationships among the factors.

In order to investigate the relationship structure between variables, graph analyses (GA) can be used. This method allows to searching for how variables are related to each other, as well as the direction of the relationships. To estimate the magnitude of those relationships, one can use structural equation models (SEM). These models are used to estimate how much each variable affects others. When there is no information regarding the causal connections, both methods should be used jointly, however when the researcher has an idea about the relationships, fitting SEM allows inferring the magnitude of causal relationships.

Some algorithms have been developed to find causal structures, such as Peter Clark (PC), inductive causation (IC) and others. All of them are based on information about dependence, independence, conditional independence and conditional dependence to assume there is a relation among variables. The first method is already available in the open software R.

To estimate the effects, specific software, such as TETRAD and LISREL, and packages for the open software R, such as sem and lavaan, have been developed. TETRAD allows to simulating data using a prior structure and estimating the coefficients related to structure as well, using maximum likelihood and other procedures (more information can be found in TETRAD manual). LISREL is a software developed by Jöreskog and Sörbom (1974) that uses the full information maximum likelihood as a procedure to estimate the coefficient based on latent structural models (FOX, 2008; FOX, 2006).

References

BOLLEN, K.A.. **Structural Equations with latent Variables**. New York: John Wiley Sons Inc. 1995. 528 p.

BOX, G.E.; TIAO, G.C. **Bayesian inference in statistical analysis**. New York: Wiley, 1992. 588 p.

CASELLA, G.; GEORGE, E. I. Explaining the Gibbs Sampler. **The American Statistician**, New York, v. 46, p. 167-174, 1992.

CHAIBUB NETO, E.C; FERARRA, T.C.; ATTIE, A.D.; YANDELL, B.S.. Inferring causal phenotype networks from segregating populations. **Genetics**, Baltimore, v. 179, p. 1089-1100, 2008.

CHAIBUB NETO, E.; KELLER, M.P; ATTIE, A.D.; YANDELL, B.S.. Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. **Annals Applied Statistic**, Cleveland, v. 4, p. 320-339, 2010.

DUNCAN, O.D.; HALLER, A.O.; PORTES, A.. Peer influences on aspirations: A reinterpretation. **American Journal of Sociology**, Chicago, v. 74, p. 119-137, 1968.

FERRON, J.M.; HESS, M.R. Estimation in SEM: A concrete example. **Journal of Educational and Behavioral Statistics**, Washington, v. 32, p. 110-120, 2007.

FOX, J.. **An introduction to structural equation modeling**, curso para programa de Computação Científica, FIOCRUZ: Rio de Janeiro, Brasil , 2008. 138p.

FOX, J.. Structural equation modeling with the sem package in R. **Structural equation modeling**, Hillsdale, v. 13, p. 465-486, 2006.

GAMERMAN, D.; LOPES, H. F. **Markov Chain Monte Carlo**: stochastic simulation for Bayes inference. London: Chapman Hall, 2006. 323 p.

GELFAND, A.E. Gibbs Sampling. **Journal of the American Statistical Association**, Alexandria, v. 95, n. 452, p. 1300-1304, 2000.

GELFAND, A.E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, Alexandria , v. 85, n. 410, p. 348-409, June, 1990.

GELMAN, A.; CARLIN, J. B.; STER, H. S.; RUBIN, D.B. **Bayesian data analysis**. Boca Raton: Chapman Hall/CRC, 2000. 526 p.

GEMAN, S.; GEMAN, D. Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. New York. v. 6, p. 721-741, 1984.

GILMOUR, A.R.; THOMPSON, R.; CULLIS, B.R. Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models. **Biometrics**, Arlington, v. 51, n. 4, p. 1440-1450, 1995.

GREENE, W.H.. **Econometric analysis**, 7 ed. New York: Macmillan. 2011. 1232 p.

HAAVELMO, T.. The statistical implications of a system of simultaneous equations. **Econometrica**, London, v. 11, 1943.

HARVILLE, D.A. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. **Journal of the American Statistical Association**, Alexandria, v. 72, n. 358, p. 320-338, Jun 1977.

HASTINGS, W. K. Monte Carlo Sampling methods using Markov chains and their application. **Biometrika**, London, v. 57, n. 1, p. 97-109, 1970.

HENDERSON, C.R. **Applications of Linear Models in Animal Breeding**. Ontario: University of Guelph, 1984. 462 p.

HENDERSON, C.R.. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, Alexandria, v. 31, p. 423-447, 1975.

HENDERSON, C.R. and QUAAS, R.L.. Multiple trait evaluation using relatives records. **Journal of Animal Science**, Champaign, v. 43, p. 1188-1197, 1976.

HENDERSON, C.R.; KEMPTHORNE, O.; SEARLE, S.R.; VON KROSIGK, C.N. Estimation of environmental and genetic trends from records subject to culling. **Biometrics**, Alexandria, v. 15, p. 192-218, 1959.

JEFREYS, H. **Theory of probability**. Oxford: Clarendon Press, 1961. 447 p.

JÖRESKOG, K.G.; SÖRBOM, D. **LISREL 8.8 for Windows [Computer software]**. Skokie, IL: Scientific Software International, Inc.

LAIRD, N.M.; WARE, J.H.. Random effects models for longitudinal data. **Biometrics**, Alexandria, v.38, p. 963-974, 1982.

METROPOLIS, N.; ROSEMBLUT, A. W.; ROSEMBLUT, M.N. ; TELLER, A. H.; TELLER , E. Equation of state calculations by fast computing machines. **Journal of Chemical Physics**, New York, v. 21, p. 1987-1092, 1953

MRODE, R. A.. **Linear Models for the Prediction of Animal Breeding Values**, 2 ed. Wallingford, Oxon, UK: CAB International, 2005. 344 p.

PAULINO, C.D.; TURKMAN, M. A.; MURTEIRA, B. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 446 p.

PATTERSON, H.D.; THOMPSON, R. Recovery of inter-block information when blocks sizes are unequal. **Biometrika**, Oxford, v. 58, n. 3, p. 545-554, 1971.

PEARL, J. **Causality Models, Reasoning and inference**. 2 ed. Cambridge, RU: Cambridge University, 2009. 484 p.

PINHEIRO, J.C.; BATES, D.M..**Mixed-Effects Models in S and S-PLAS**. New York: Springer, 2000. 528 p.

R. Development Core Team. R Foundation for Statistical Computing. **R 2.15.2**: A language and environment for statistical computing, Vienna, 2012. Available in <<http://www.r-project.org/>>. Acesso em: 23 nov. 2012

ROSA, G. J. M.; VALENDE, B. D.. Structural Equation Models for Studying Causal Phenotype Networks in Quantitative Genetics IN: SINOQUETE, C.; MOURAD, R.. **Probabilistic Graphical Models for Genetics, Genomics and Postgenomics**, Oxford University Press, 2014. 480 p.

ROSA, G.J.M.; VALENTE, B.D.; CAMPOS, G. de los; WU, X.L.; GIANOLA, D.; SILVA, M.A.. Inferring causal phenotype networks using structural equation models. **Genetics Selection Evolution**, London, v. 43, p. 1046-1057, 2011.

SCHAEFFER, L.R. Sire and cow evaluation under multiple trait models. **Journal of Dairy Science**, Champaing, v.67, p.1567-1580, 1984.

SHIPLEY, B.. **Cause and Correlation in Biology**. Cambridge, RU: Cambridge University Press, 2004. 336 p.

SORENSEN, D.; GIANOLA, D. Likelihood. **Bayesian and MCMC methods in quantitative genetics**. New York: Springer, 2002. 740 p.

The TETRAD Project Causal Models and Statistical Data. **tetrad**. Available in
<<http://www.phil.cmu.edu/projects/tetrad/current.html>>. Access in: 23 nov. 2012

VALENTE, B.D.; ROSA, G.J.M. Mixed effects structural equation models and phenotypic causal networks. **Methods in molecular biology**, Totowa, v. 1019, 449-564, 2013.

VALENTE, B.D.; ROSA, G.J.M.; CAMPOS, G. de los; GIANOLA, D.; SILVA, M.A.. Searching for recursive causal structures in multivariate quantitative genetics mixed models. **Genetics**, Baltimore, v. 185, p. 633-644, 2010.

VALENTE, B.D.; ROSA, G.J.M.; GIANOLA, D.; WU, X-L; WEIGEL, K.. Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? **Genetics**, Baltimore, v. 194, p. 561-572, 2013.

VALENTE, B.D.; MOROTA, G.; PEÑAGARICANO, F., GIANOLA, D.; WEIGEL, K.; ROSA, G.J.M.. The causal meaning of genomic predictors and how it affects construction and comparison of genome-enabled selection models. **Genetics**, Baltimore, v. 200, 2015.

VALENTE, B.D.; ROSA, G.J.M.; SILVA, M.A.; TEIXEIRA, R.B.; TORRES, R.A.. Searching for phenotypic causal networks involving complex traits: an application to European quails. **Genetics, Selection, Evolution**, London, v. 43, p. 37-48, 2011.

WRIGHT, S.. Systems of mating. i. the biometric relations between parents and offspring. **Genetics**, Baltimore, v. 6, p. 111-123, 1921.

3 ADVANTAGES OF USING HIGHER POLYNOMIAL POLYNOMIAL INSTEAD OF LINEAR RELATIONSHIP BETWEEN TRAITS IN STRUCTURAL EQUATION MODELS IN QUANTITATIVE GENETICS: A SIMULATION STUDY

Abstract

The concept of causality has been used in many studies in different areas of knowledge and its application has been growing during the past decade. Scientific investigation in biology generally involves learning causal relationships among variables, which are important in order to predict the consequences of interventions in the system. In quantitative genetics, so far, the most common approach to solve problems involving causality is assuming linear relationships between traits. However, such formulations are not suitable for scenarios where evidences of non linear relationships between traits can be observed. To overcome this limitation, this work proposes the use of polynomial mixed-effects structural equation model. In order to verify the advantages of the polynomial approach, a simulation study with 50 data sets was performed. Data was sampled from a fully recursive causal structure involving 3 traits, considering linear and non linear relationships between traits with one exogenous covariates associated to each endogenous trait. The Gibbs sampler was used to obtain posterior samples of model parameters and other unknowns. The standard linear model and a model accounting for second degree polynomials were compared. The results show that the inclusion of an extra polynomial degree enhances the SEM expressive power. For some situations the inappropriate assumption of linearity results in poor predictions of overall genetic effects, either by overestimating, underestimating or even suggesting an opposite directions for them. The results also shows that there is no loss when using polynomial approach, because when the relationships were assumed linear the model estimate values equal o zero for the quadratic term.

Keywords: Structural models; causal inference; quantitative genetics; polynomial regression; Bayesian inference, Gibbs sampling

3.1 Introduction

Human beings typically try to comprehend the causal relationships among measurable variables in many different contexts. From as simple and obvious as using a hose and wetting the floor, to a more complex as the fact that smoking can be a cause of cancer or some more subjective examples as how intelligence affects career success, where the response can-not be measured (PEARL, 2009).

In some scenarios, controlled randomized experiments (e.g management practices for improving beef cattle body weight or maize yield) can be applied to infer causal effects. In that case, the observed difference among response variables of subjects assigned to different levels of treatment could be attributed to the treatment effect (HOYER et al., 2008). However, randomization cannot be applied to some other studies due to many reasons, e.g. ethical or financial issues. Methods and models for causal inference on the basis of observational data

were developed to tackle limitations in such cases .

There are two fundamental aspects to be learned in the application of causal models to study the relationships among a set of variables , and there are two distinct methods for learning each aspect: Graph analysis (GA) and structural equation modeling (SEM). While GA involves searching for causal structures that qualitatively represent how variables are causally connected, fitting SEM with a known causal structure allows inferring the magnitude of causal relationships. The latter involves a model that can be expressed as a multiple-trait regression model where some response variables may be considered as covariates in the right hand side of equations for other response variables (ROSA et al., 2011; LEE; ZHU, 2000; VALENTE et al., 2013; LEE; TANG, 2006).

During the past decades the number of studies involving causal models has been increased in many fields of scientific investigation. In biology, due to the complexity networks that express the relationship among traits, the concept of causality can present some extra difficulties (HOYER et al., 2008; ROSA et al., 2011; BLAIR, 2012; VALENTE et al., 2010; MATURANA et al., 2009; CHAIBUB NETO, 2010; CHAIBUB NETO et al., 2010; VALENTE et al., 2011; CHAIBUB NETO et al., 2008).

Scientific investigation in biology frequently involves learning causal relationships among traits, in these systems can be easily seen situations where phenotypic traits may exert effects among themselves (i.e. calf liability and gestation length) (ROSA et al., 2011; VALENTE et al., 2010, MATURANA et al., 2009). This investigation is important in order to predict the consequences of interventions in the system. In this context, methods for causal inference become important tools. Notwithstanding , these structures of relationship take in account a complex functional network that should be analyzed carefully, specifically in quantitative genetics applications where individuals are genetically related due to familial relationships, such dependencies should also be modeled in SEM (CHAIBUB NETO, 2008; ROSA et al., 2011; VALENTE et al., 2010; VALENTE et al., 2013; VALENTE et al., 2011; WU; HERINGSTAD; GIANOLA, 2010; MATURANA et al., 2009; LIU; FUENTE; HOESCHELE, 2008; CHAIBUB NETO, 2010; GONZÁLEZ-RODRÍGUEZ et al., 2014; XIONG; FANG, 2004; LI et al., 2006). To solve situations using the genetic covariance, Gianola and Sorensen (2004) adapted the causal models to the mixed models context (VALENTE et al., 2013).

However, so far, most studies in SEM applied to quantitative genetics have assumed that relationships between traits are linear, which may be unrealistic for some cases as for example (MATURANA et al., 2008; GONZÁLEZ-RODRÍGUEZ et al., 2014; VARONA; SORENSEN, 2014; KÖNIG et al., 2008).

Our goal is to propose a mixed effects SEM with polynomial relationship among traits, as well as proposing inference methods for such model. We used simulated data to evaluate this method and compared it to standard approaches to fit linear SEM. The intention is to illustrative how structural equation modeling can benefit from exploring more flexible functions in some situations.

The section Material and Methods explains the parameters estimation, data simulation and analysis procedure. The section Results shows and discusses inferences when using polynomial and linear SEM. The section Conclusion contain final remarks on the proposed methodology and its advantages.

3.2 Material and Methods

In this section we propose methods to estimate the parameters and present how data was simulated.

3.2.1 Methods

Recently many applications of structural equation models in quantitative genetics, generally, assuming that the causal relationships as linear (GIANOLA; SORENSE, 2004; VALENTE et al., 2010; VALENTE et al., 2013; ROSA et al., 2011). Such assumptions are not always realistic, requiring different approaches to circumvent the non-linearity (GONZÁLEZ-RODRÍGUEZ et al., 2014 and MATURANA et al., 2009). A general structural equation model can be written as

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}_1 + \mathbf{e}_1 \quad (3.1)$$

$$\mathbf{y}_2 = f_{12}(\mathbf{y}_1) + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}_2 + \mathbf{e}_2 \quad (3.2)$$

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$\mathbf{y}_J = f_{1J}(\mathbf{y}_1) + f_{2J}(\mathbf{y}_2) + \dots + f_{J'J}(\mathbf{y}_{J'}) + \mathbf{X}_J\boldsymbol{\beta}_J + \mathbf{u}_J + \mathbf{e}_J; \quad (3.3)$$

where, $\mathbf{y}_j$ represents the vector ($1 \times n$) of observed values for the j^{th} trait and $j = 1, 2, \dots, J$, the function $f_{j'j}(\mathbf{y}_{j'})$ represent the function that express the causal relationships between trait j' and j , the vector $\boldsymbol{\beta}_j = [\beta_{0j} \ \beta_{1j} \dots \beta_{kj}]^\top$ represent the fixed effects of exogenous variables for the j^{th} trait, $\mathbf{X}_j = [\mathbf{1} \dots \mathbf{x}_{kj}]$ represent the incidence matrix of fixed effects on $\mathbf{y}_j$, $\mathbf{u}_j$ represent the vector of direct random genetic effects for the trait j and $\mathbf{e}_j$ are independently and normally distributed residuals terms.

Under causal linear assumptions, the links $f_{j'j}(\mathbf{y}_{j'})$ are given by $\mathbf{y}_{j'} \times \lambda_{j'j}$. In this work, we propose using a polynomial function as a link between traits to modeling non-linear relationships. For this case, the function $f_{j'j}(\mathbf{y}_{j'})$ can be rewritten as: $\sum_{d=1}^D \lambda_{j'j_d} \mathbf{y}_{j'}^d$, where D represents the polynomial degree chosen and d represent the polynomial index. The equations 3.1, 3.2 and 3.3 can be combined and written as:

$$\mathbf{y}_i = \mathbf{F}_y(\mathbf{y}_i) + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u} + \mathbf{e}_i, \quad (3.4)$$

where $\mathbf{y}_i$ represents a vector of t traits for the i^{th} individual. $\mathbf{F}_y(\mathbf{y}_i)$ represents a vector of polynomial functions, $\mathbf{X}_i$ represents the incidence matrix for the effects of the vector $\boldsymbol{\beta}$ on $\mathbf{y}_i$, $\mathbf{u}_i$ and $\mathbf{e}_i$ are vectors with direct genetic effects and model residuals distributed as

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

where $\mathbf{G}_0$ and $\mathbf{\Psi}_0$ are respectively direct genetic and residual covariance matrices. Assuming that $\mathbf{\Lambda}_D$ represents a partitioned matrix $[\mathbf{\Lambda}_1 \mathbf{\Lambda}_2 \dots \mathbf{\Lambda}_d]$, in which each partition contain a $k \times k$ matrix where the diagonal contains only zero and off-diagonals represent the causal links related to each polynomial degree, $\mathbf{F}_y(\mathbf{y}_i)$ can be seen as $\mathbf{\Lambda}_1 \mathbf{y}_i + \mathbf{\Lambda}_2 \mathbf{y}_i^2 + \dots + \mathbf{\Lambda}_d \mathbf{y}_i^d = \mathbf{\Lambda}_D \mathbf{y}_{iD}$. The conditional distribution of $\mathbf{y}_i$ given $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{\Lambda}_d$ and $\mathbf{\Psi}_0$ is:

$$\mathbf{y}_i | \mathbf{\Lambda}_d, \boldsymbol{\beta}, \mathbf{u}_i, \mathbf{\Psi}_0 \sim N[\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u} + \mathbf{\Lambda}_D \mathbf{y}_{iD}, \mathbf{\Psi}_0], \quad (3.5)$$

and the model for n individuals

$$\mathbf{y} = (\mathbf{\Lambda}_d \otimes \mathbf{I}_n) \mathbf{y}_d + \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u} + \mathbf{e}, \quad (3.6)$$

where $\mathbf{y}_D = [\mathbf{y}^1 \mathbf{y}^2 \dots \mathbf{y}^d]$. The conditional distribution is given by:

$$\mathbf{y} | \mathbf{\Lambda}_D, \boldsymbol{\beta}, \mathbf{u}, \mathbf{\Psi}_0 \sim N[\mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u} + (\mathbf{\Lambda}_D \otimes \mathbf{I}_n) \mathbf{y}_D, \mathbf{\Psi}_0 \otimes \mathbf{I}_n]. \quad (3.7)$$

Parameters were inferred via Bayesian inference. Assuming that the joint prior distribution for the parameters is given by

$$\pi(\mathbf{\Lambda}_D, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{\Psi}_0) = \pi(\mathbf{\Lambda}_D) \pi(\boldsymbol{\beta}) \pi(\mathbf{u} | \mathbf{G}_0) \pi(\mathbf{G}_0) \pi(\mathbf{\Psi}_0), \quad (3.8)$$

where $\pi(\mathbf{\Lambda}_D)$ and $\pi(\boldsymbol{\beta})$ are considered uniform distributions, $\pi(\mathbf{u} | \mathbf{G}_0)$ represents a Normal distribution parameterized as $N(0, \mathbf{G}_0 \otimes \mathbf{A})$, $\pi(\mathbf{G}_0)$ is a Inverse Wishart $IW(v_G, \mathbf{G}_0^\bullet)$, where v_G and $\mathbf{G}_0^\bullet$ are the degree of freedom and scale parameter, respectively. Assuming independence among ψ_j , $\pi(\mathbf{\Psi}_0)$ can be expressed as $\pi(\mathbf{\Psi}_0) = \prod_{j=1}^t \pi(\psi_j)$ and $\pi(\psi_j)$ are considered a scaled inverted chi-square $\text{Inv-}\chi^2(v_\psi, S^2)$ where v_ψ and S^2 are the degrees of freedom and scale, respectively, and ψ_j is the residual variance for the trait j . Therefore, the joint prior distribution can be seen as:

$$\pi(\mathbf{\Lambda}_D, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{\Psi}_0) \propto N(0, \mathbf{G}_0 \otimes \mathbf{A}) IW(v_G, \mathbf{G}_0^\bullet) \prod_{j=1}^t \text{Inv-}\chi^2(v_\psi, S^2). \quad (3.9)$$

The joint posterior distribution for the parameters is given by:

$$\pi(\mathbf{\Lambda}_D, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{\Psi}_0 | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{\Lambda}_D, \boldsymbol{\beta}, \mathbf{u}, \mathbf{\Psi}_0) \pi(\mathbf{u} | \mathbf{G}_0) \pi(\mathbf{G}_0) \prod_{j=1}^t \pi(\psi_j). \quad (3.10)$$

It is easy to see that the joint posterior distribution in 3.10 does not have a closed form, but the full conditional distributions do. To estimate the parameters, the Gibbs sampler was used. Assuming that $\mathbf{\Lambda}_D$ are known and let $\mathbf{y}^* = \mathbf{y} - (\mathbf{\Lambda}_D \otimes \mathbf{I}_n)(\mathbf{y}_D)$, the equation 3.6 can be rewritten as:

$$\mathbf{y}^* = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u} + \mathbf{e}, \quad (3.11)$$

and

$$p(\mathbf{y}^*|\boldsymbol{\beta}, \mathbf{u}, \Psi_0) \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, (\boldsymbol{\Lambda}_D \otimes \mathbf{I}_n)\Psi_0). \quad (3.12)$$

Following Sorensen and GIANOLA (2002), let $\mathbf{M} = \begin{bmatrix} \sum_{j=1}^t \mathbf{X}_j & \sum_{j=1}^t \mathbf{Z}_j \end{bmatrix}$,
 $\Omega = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1} \end{bmatrix}$, where $\boldsymbol{\theta}^\top = [\boldsymbol{\beta}^\top \ \mathbf{u}^\top]$, $\Psi = \Psi_0 \otimes \mathbf{I}_n$, $\mathbf{C} = \mathbf{M}^\top \Psi^{-1} \mathbf{M} + \Omega$,

$\mathbf{t} = \mathbf{M}^\top \Psi^{-1} \mathbf{y}^*$ and $\mathbf{C}\boldsymbol{\theta} = \mathbf{t}$. The full conditional distribution for $\boldsymbol{\beta}$ and $\mathbf{u}$ is given by:

$$\begin{aligned} p(\boldsymbol{\beta}, \mathbf{u}|\boldsymbol{\Lambda}_D, \mathbf{G}_0, \Psi_0, \mathbf{y}) &= p(\boldsymbol{\theta}|\mathbf{G}_0, \Psi_0, \mathbf{y}^*) \\ &\propto p(\mathbf{y}^*|\boldsymbol{\theta}, \Psi_0)p(\mathbf{u}|\mathbf{G}_0) \\ &\propto \exp \left\{ -\frac{1}{2} [(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top (\mathbf{M}^\top \Psi^{-1} \mathbf{M} + \Omega)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})] \right\}, \end{aligned} \quad (3.13)$$

it can be seen that the equation (3.13) is a the kernel of a normal distribution with mean equal to $\hat{\boldsymbol{\theta}}$ and covariance matrix $(\mathbf{M}^\top \Psi^{-1} \mathbf{M} + \Omega)^{-1} = \mathbf{C}^{-1}$, $p(\boldsymbol{\theta}|\mathbf{G}_0, \Psi_0, \mathbf{y}^*) \propto N(\hat{\boldsymbol{\theta}}, (\mathbf{M}^\top \Psi^{-1} \mathbf{M} + \Omega)^{-1})$. The posterior distribution for the subset of parameters is $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \mathbf{G}_0, \Psi_0, \mathbf{y}^*) \sim N(\tilde{\boldsymbol{\theta}}_i, \mathbf{C}_{ii}^{-1})$, where i is the i^{th} partition of $\boldsymbol{\theta}$.

The conditional distribution of the additive genetic covariance matrix for S groups of subjects genetically identical in the SEM is given by:

$$\begin{aligned} p(\mathbf{G}_0|\boldsymbol{\beta}, \mathbf{u}, \Psi_0, \mathbf{y}^*) &\propto p(\mathbf{u}|\mathbf{G}_0)p(\mathbf{G}_0) \\ &\propto |\mathbf{G}_0|^{-\frac{1}{2}(v_g + S + t + 1)} \exp \left\{ -\frac{1}{2} \text{tr} (\mathbf{G}_0^{-1} (S_A + \mathbf{G}_0^{\bullet-1})) \right\}, \end{aligned} \quad (3.14)$$

which is the kernel of a Inverse Wishart with $(v_g + S)$ and $(S_A + \mathbf{G}_0^{\bullet-1})$ as degrees of freedom

and scale, respectively, and $S_A = \begin{bmatrix} \mathbf{u}'_1 \mathbf{A} \mathbf{u}_1 & \mathbf{u}'_1 \mathbf{A} \mathbf{u}_2 & \dots & \mathbf{u}'_1 \mathbf{A} \mathbf{u}_j \\ \mathbf{u}'_2 \mathbf{A} \mathbf{u}_1 & \mathbf{u}'_2 \mathbf{A} \mathbf{u}_2 & \dots & \mathbf{u}'_2 \mathbf{A} \mathbf{u}_j \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}'_j \mathbf{A} \mathbf{u}_1 & \mathbf{u}'_j \mathbf{A} \mathbf{u}_2 & \dots & \mathbf{u}'_j \mathbf{A} \mathbf{u}_j \end{bmatrix}$.

For each trait the conditional distribution for the residual variance is given by:

$$\begin{aligned} p(\psi_j|\boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{y}^*) &\propto p(\mathbf{y}^*|\boldsymbol{\beta}, \mathbf{u}, \psi_j)p(\psi_j) \\ &\propto \psi_j^{-\frac{n + v_\psi}{2} - 1} \exp \left\{ -\frac{\mathbf{e}_j^\top \mathbf{e}_j + v_\psi S^2}{2\psi_j} \right\}, \end{aligned} \quad (3.15)$$

which refers to a scaled inverse chi-square where $(v_\psi + n)$ and $\left(-\frac{\mathbf{e}^\top \mathbf{e} + v_\psi S^2}{v_\psi + n} \right)$ represent the degrees of freedom and scale, respectively. Let $\mathbf{y}^+ = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}$ and then $p(\mathbf{y}|\boldsymbol{\Lambda}_d, \boldsymbol{\beta}, \mathbf{u}, \Psi_0) = p(\mathbf{y}^+|\boldsymbol{\Lambda}_d, \Psi_0) = N(\boldsymbol{\Lambda}_d \mathbf{y}_d, \Psi)$.

Given β and $\mathbf{u}$, the y^+ for different individuals are independent. Furthermore, assuming that the residuals of different traits are independent as well, the model for each trait can be written as:

$$\mathbf{y}_j^+ = \mathbf{F}_{y_j}(\mathbf{y}_p) + \mathbf{e}_j, \quad (3.16)$$

where $\mathbf{F}_{y_j}(\mathbf{y}_p)$ is $\sum_{d=1}^D \mathbf{y}_{pjd} \lambda_{jd}$, where p represents the parents of the trait j , in others words, the traits that affect y_j , D represents the polynomial degree, $\mathbf{y}_{pjd}$ is a matrix $[\mathbf{y}_{pj}^d]$ where $\mathbf{y}_{pj}$ is a matrix that contain information about the parents of the trait in each column, and λ_{jd} is a vector of loading effects for each parent at polynomial order d . Then the conditional distribution for the j^{th} trait is given by:

$$\begin{aligned} p(\lambda_j | \mathbf{y}_j^+, \psi_j) &\propto p(\mathbf{y}_j^+ | \lambda_j, \psi_j) p(\lambda_j) \\ &\propto \exp \left\{ -\frac{1}{2\psi_j} \left[(\lambda_j - \hat{\lambda}_j)^\top \mathbf{y}_{pj}^\top \mathbf{y}_{pj} (\lambda_j - \hat{\lambda}_j) \right] \right\}, \end{aligned} \quad (3.17)$$

which is the kernel of a normal distribution with mean $\hat{\lambda}_j = (\mathbf{y}_{pj}^\top \mathbf{y}_{pj})^{-1} \mathbf{y}_{pj}^\top \mathbf{y}_j^+$ and variance $(\mathbf{y}_{pj}^\top \mathbf{y}_{pj})^{-1} \psi_j$, where λ_j express the effects for all polynomial degrees, In this work linear and second degree polynomials were assumed as causal effects.

3.2.2 Simulation data

The causal model for the simulation presented a fully recursive causal structure involving 3 traits (endogenous variables $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$) and 3 exogenous covariates ($\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$) respectively assigned to the endogenous traits. All of these variables were simulated for 1,800 subjects, and genetics effects were simulated for 300 inbred lines, each one with 6 individuals. Traits were simulated as having linear and quadratic relationships among themselves. A total of 50 samples were drawn following the structural equations below.

$$\begin{cases} \mathbf{y}_1 = \mathbf{X}_1 \beta_1 + \mathbf{u}_1 + \mathbf{e}_1 \\ \mathbf{y}_2 = f_{12}(\mathbf{y}_1) + \mathbf{X}_2 \beta_2 + \mathbf{u}_2 + \mathbf{e}_2 \\ \mathbf{y}_3 = f_{13}(\mathbf{y}_1) + f_{23}(\mathbf{y}_2) + \mathbf{X}_3 \beta_3 + \mathbf{u}_3 + \mathbf{e}_3 \end{cases},$$

where, $\mathbf{y}_j$ represents the j^{th} trait, $f_{j'j}(\mathbf{y}_{j'})$ represents the function that expresses the causal relationships between trait j' and j , β_j represents the vector of fixed effects for the j^{th} trait $\beta_j = [\beta_{0j} \ \beta_{1j}]$, $\mathbf{X}_j$ are the incidence matrices where $\mathbf{X}_j = [\mathbf{1} \ \dots \ \mathbf{x}_j]$, $\mathbf{u}_j$ represents the direct random genetic effects for the trait j , and $\mathbf{e}_j$ residuals, which are normally distributed and independent for different traits. For this work data were simulated based on both linear and non linear links between traits under the same structure, with the purpose of verifying if fitting SEM with flexible functions returns good inferences even when causal relationships are simulated as linear. The simulated structure can be seen as directed acyclic graph (DAG) in Figure 3.1

Figure 3.1 – DAG representing the structure of the simulation, expressing relationship between traits (y_1, y_2, y_3), genetics effects (u_1, u_2, u_3) and covariates (X_1, X_2, X_3)

3.2.3 Analysis

Two approaches were used, the first approach involved fitting a SEM, where all causal relationships were estimated using second degree polynomial functions among traits. The second approach consisted in fitting a SEM where all causal relationship were assumed linear relationships among traits. While performing the simulation study, the estimators of the effects obtained were compared to the true values. Under the linear SEM approach, the quadratic terms were not estimated and for these terms the error were not evaluated.

Parameter inferences were based on Gibbs sampling (GEMAN, 1984), for each data set their posterior mean were calculated, and for all procedures the open software R was used. A total of 550,000 samples were simulated where a total of 15,000 were discarded as burnin.

Autocorrelations between samples were computed, even after the thinning process. To solve the problem with the autocorrelations the equivalent number of independent samples was computed using the effective sample size (KASS et al., 1998). The procedure to estimate the effective sample size it is given by

$$ESS = \frac{N_c}{1 + 2 \sum_{k=1}^{\infty} \rho_k}, \quad (3.18)$$

where, N_c is the number of simulated sample, $1 + 2 \sum_{k=1}^{\infty} \rho_k$ represent the autocorrelation time and ρ_k represent the auto correlation function (ACF) for the k^{th} lag.

To test if the extended model performed better than the standard model in fitting the simulated data, we used the mean of posterior means (MPM), the mean of Monte Carlo Error (MMCE) (KOEHLER; BRWON; HANEUSE, 2004), the mean of estimate parameter deviation (MEPD) and the mean of mean square error (MMSQE).

The MPM is given by

$$MPM_k = \frac{1}{N_s} \sum_{i=1}^{N_s} PM_{ki} \quad (3.19)$$

where k represent the estimate parameter and PM_{ki} represent the posterior mean for the k^{th} parameter and for the i^{th} simulated data set, the MMCE is given by

$$MMCE(\hat{\varphi}_k) = \frac{1}{N_s} \sum_{i=1}^{N_s} \sqrt{var(\hat{\varphi}_{k_i})}, \quad (3.20)$$

the MEPD is given by:

$$MEPD = \frac{1}{N_s} \sum_{n_s=1}^{N_s} (p - \hat{p}_{n_s}), \quad (3.21)$$

and the MMSQE is given by:

$$MMSQE = \frac{1}{N_s} \sum_{n_s=1}^{N_s} (p - \hat{p}_{n_s})^2, \quad (3.22)$$

where N represents the total of observation in the data set and N_s represents the number of data sets analyzed, p represent the true value of the parameter and $\hat{p}$ represent the estimate value of the parameter. For all comparison methods, the smallest value are assigned to the best model.

3.2.4 Results of fitting polynomial SEM

Table 3.1 presents the mean of the posterior mean and mean of the Monte Carlo error for the causal parameter under the polynomial approach. Results suggested that the estimation for the causal parameters were accurate. An important results is that when the true value of the causal parameter is zero, the inferred values tended to be zero as well. Furthermore, results indicated small levels of Monte Carlo Error.

Table 3.1 – Parameter true value, mean of posterior (MPM) mean and mean of Monte Carlo error (MMCE) for causal effects using polynomial SEM

Parameter	True value	MPM	MMCE
λ_{12_1}	-1.850	-1.8498	0.0041
λ_{12_2}	0.000	0.0000	0.0000
λ_{13_1}	15.550	15.5501	0.0033
λ_{13_2}	0.000	0.0000	0.0000
λ_{23_1}	140.850	140.8383	0.0009
λ_{23_2}	-0.155	-0.1550	0.0000

Figure 3.2 presents the credible interval for each one of the structural parameter for the polynomial approach. The credible interval contained the true value for all parameters. Furthermore,

inferences of the quadratic coefficient were around zero where the structural relationship were linear (the true value were considered zero).

Figure 3.2 – Credible interval for the structural parameters representing the linear (a) and quadratic (b) effects between traits 1 and 2; linear (c) and quadratic (d) effects between traits 1 and 3; linear (e) and quadratic (f) effects between traits 2 and 3. The flat blue line represents the true value of the parameters

Table 3.2 presents the mean of the posterior mean and mean of the Monte Carlo error for the fixed effects of the exogenous covariates. Results show that the estimations were accurate. However, it is possible to see that the MCE obtained for β_{02} and β_{03} were considerably higher than the others due to problems related to identifiability.

Table 3.2 – Parameter true value, mean of posterior mean and Monte Carlo error for fixed effects using polynomial SEM

Parameter	True value	MPM	MCE
β_{01}	100.0	100.4929	0.0217
β_{02}	10.0	10.5917	0.4909
β_{03}	100.0	102.5722	0.7558
β_{11}	12.0	11.9517	0.0020
β_{12}	7.0	6.9956	0.0006
β_{13}	30.8	30.8001	0.0001

Table 3.3 presents the mean of the posterior mean and mean of the Monte Carlo error for the genetics (co)variances of direct effects. when compared the MPM the values close to true values, even though such inferences were less accurate than the one obtained for the fixed and structural parameters, they still can be considered accurate. For the MMCE the values were considered small.

Table 3.3 – Parameter true value, mean of posterior mean and Monte Carlo error for genetic covariance matrix using polynomial SEM

Parameter	True value	MPM	MCE
$\sigma_{g_1}^2$	150.000	145.911	0.096
$\sigma_{g_1g_2}$	87.232	85.634	0.084
$\sigma_{g_1g_3}$	30.590	28.873	0.094
$\sigma_{g_2}^2$	130.000	134.255	0.107
$\sigma_{g_2g_3}$	-51.002	-53.024	0.095
$\sigma_{g_3}^2$	150.000	152.074	0.103

Table 3.4 presents the mean of the posterior mean and mean of the Monte Carlo error for the residuals for the polynomial approach. Results suggested that the inferences for the residual variance were accurate, once the MPM values were close to true values. Also the MMCE for all parameters were considered small.

Table 3.4 – Parameter true value, mean of posterior mean and Monte Carlo error for residual using Polynomial SEM

Parameter	True Value	MPM	MCE
$\sigma_{\varepsilon_1}^2$	450.000	451.0170	0.0712
$\sigma_{\varepsilon_2}^2$	450.000	456.3898	0.0878
$\sigma_{\varepsilon_3}^2$	450.000	452.8508	0.0977

Table 3.5 – Mean of estimate parameter deviation (MEPD) and the variance of mean estimate parameter deviation (MMSQE) for the polynomial approach

Parameter	MEPD	MMSQE
λ_{12_1}	0.0028	0.0065
λ_{12_2}	0.0000	0.0000
λ_{13_1}	-0.0038	0.0100
λ_{13_2}	0.0000	0.0000
λ_{23_1}	0.0110	0.0013
λ_{23_2}	0.0000	0.0000
β_{01}	-0.1658	4.8738
β_{11}	0.0217	0.0342
β_{02}	-0.7451	123.7960
β_{12}	0.0046	0.0020
β_{03}	-2.3228	220.3624
β_{13}	0.0008	0.0003
$\sigma_{g_1}^2$	4.1400	318.7808
$\sigma_{g_1 g_2}$	2.3959	182.8982
$\sigma_{g_1 g_3}$	1.8017	177.7561
$\sigma_{g_2}^2$	-2.6491	408.7518
$\sigma_{g_2 g_3}$	1.7887	189.6544
$\sigma_{g_3}^2$	-1.5003	533.3991
σ_{ε_1}	-0.1761	206.6767
σ_{ε_2}	-6.4260	224.3231
σ_{ε_3}	-3.5813	225.1995

Table 3.5 presents the mean estimate posterior deviation, representing the difference between the true value and the polynomial approach estimate, and the mean of square mean error of the parameter, all of them computed for the polynomial approach. Even though the MEPD

in absolute values were for all parameters smaller than 10, for the random effects can be seen that the MMSQE are greater than the fixed and causal effects, except for the intercept related to the traits y_2 and y_3 .

Figure 3.3 – Credible interval for the fixed effects representing the intercept (a) and the covariate x_1 (b) for the trait 1; intercept (c) and covariate x_2 (d) for the trait 2; intercept (e) and covariate x_3 (f) for the traits 3. The flat blue line represents the true value of the parameters

Figure 3.3 presents the credible interval for each one of the fixed effects for the polynomial approach. It can be seen that for the intercept the intervals are higher than the ones for the covariates. Additionally it is possible to verify that few situations for the parameters, β_{01} , β_{11}

and β_{13} the credible interval does not contain the true values. For the parameters β_{02} and β_{03} the credible interval were greater than others and also contain the zero value. As mentioned, such level of posterior uncertainty could be due to problems related to the identifiability, once mistakes occurred in parameter estimate for a trait can be carried for those traits that are related to them.

Figure 3.4 – Credible interval for genetics effects representing the variance (a), (b) and (c), for the traits 1, 2 and 3, respectively; covariance (d) between traits 1 and 2; covariance (e) between traits 1 and 3; covariance (f) between traits 2 and 3. The flat blue line represents the true value of the parameters

Figure 3.4 presents the credible interval for each one of the direct (co)variances of genetic

effects for the polynomial approach. It can be seen that the credible interval does not contain the true value for all situations. Furthermore, the credible interval for the covariance $\sigma_{g_1g_3}$ in some situations contained the zero value.

Figure 3.5 presents the credible interval for the residuals associated to all traits under the polynomial approach. For all estimates only for one case the credible interval do not contain the true value.

Figure 3.5 – Credible interval for the error representing the σ_{ϵ_1} (a) the error for the trait 1; σ_{ϵ_2} (b) the error for the trait 2; σ_{ϵ_3} (c) the error for the trait 3. The blue line represents the true value of the parameters

3.2.5 Results of fitting SEM with linear effects

Table 3.6 presents the mean of the posterior mean and mean of the Monte Carlo error for the causal parameter for the linear SEM. The estimation for the causal parameters were accurate for the first two λ_{12_1} and λ_{13_1} . However λ_{23_1} was poorly inferred, even though the MMCE, was relatively small, this lack of precision may be explained by the reason of the quadratic effect also be modeled by the linear effect.

Table 3.6 – Parameter true value, mean of posterior mean and mean of Monte Carlo error for causal effects using standard linear SEM

Parameter	True value	MPM	MMCE
λ_{12_1}	-1.850	-1.8493	0.0001
λ_{12_2}	0.000	-	-
λ_{13_1}	15.550	15.6631	0.0489
λ_{13_2}	0.000	-	-
λ_{23_1}	140.850	-18.1381	0.0196
λ_{23_2}	-0.155	-	-

Table 3.7 presents the mean of the posterior mean and mean of the Monte Carlo error for the fixed effects for the linear approach. Inferences were generally accurate, except for the intercept β_{03} . Posterior distributions for this parameter were far from the true value and resulted in the highest MCE, this values can be seen as consequence of the mistakes occurred on previews estimate of λ_{23_1}

Table 3.7 – Parameter true value, mean of posterior mean and Monte Carlo error for fixed effects using standard linear SEM

Parameter	True value	MPM	MCE
β_{01}	100.00	100.1937	0.0186
β_{02}	10.00	9.8503	0.0630
β_{03}	100.00	39159.4819	21.0205
β_{11}	12.00	11.9755	0.0013
β_{12}	7.00	6.9990	0.0004
β_{13}	30.80	30.4693	0.0129

Figure 3.6 presents the credible interval for each one of the causal parameter for the linear approach. The credible interval for the causal parameters, λ_{12_1} and λ_{13_1} consistently contain the true value, however for λ_{23_1} inferences were too far from the true value, such that none of credible interval contains the true parameter value, as mentioned previously, this huge difference may be a result of modeling the quadratic and linear effect only using the linear effect.

Figure 3.6 – Credible interval for the structural parameters representing the linear (a) effect between traits 1 and 2; linear (b) effect between traits 1 and 3; linear (c) effect between traits 2 and 3. The flat blue line represents the true value of the parameters

Table 3.8 – Parameter true value, mean of posterior mean and Monte Carlo error for genetic covariance matrix using standard linear SEM

Parameter	True value	MPM	MCE
$\sigma_{g_1}^2$	150.000	151.5480	0.1599
$\sigma_{g_1g_2}$	87.232	84.5224	0.1208
$\sigma_{g_1g_3}$	30.590	0.0788	3.5752
$\sigma_{g_2}^2$	130.000	136.6579	0.1547
$\sigma_{g_2g_3}$	-51.002	-0.3897	3.2739
$\sigma_{g_3}^2$	150.000	345.1650	21.7534

Table 3.8 presents the mean of the posterior mean and mean of the Monte Carlo error for genetic(co)variances obtained with for the linear approach. Inferences were accurate for the parameters $\sigma_{g_1}^2$, $\sigma_{g_1g_2}$ and $\sigma_{g_2}^2$. On the other hand for the parameters $\sigma_{g_3}^2$, $\sigma_{g_1g_3}$ and $\sigma_{g_2g_3}$ the estimate value was far from the true value, and their MCE were the higher than one. Remarkably, inferences for the covariance $\sigma_{g_1g_3}$ not only presented a different value from the true, but also presented the opposite direction.

Figure 3.7 presents credible interval for the fixed effects for the linear approach. Posterior distributions of β_{01} , β_{02} , β_{11} , β_{12} and β_{13} contain the true value, however for β_{03} inferences were too far from the true value and none of credible interval contains it, also as consequence of problems for the estimate of λ_{23_1} .

Figure 3.7 – Credible interval for the fixed effects representing the intercept (a) and the covariate x_1 (b) for the trait 1; intercept (c) and covariate x_2 (d) for the trait 2; intercept (e) and covariate x_3 (f) for the traits 3. The flat blue line represents the true value of the parameters

Table 3.9 presents the mean of the posterior mean and mean of the Monte Carlo error for residuals associated to each trait. Inferences were accurate for the parameters ψ_1 and ψ_2 . How-

ever, inferences for the parameter ψ_3 were far from the true value. Additionally, the MCE were the highest for all parameters in this work. This results clearly shows that estimation problems can be found under the linear assumption when such assumption does not hold.

Table 3.9 – Parameter true value, mean of posterior mean and Monte Carlo error for residual using Linear SEM

Parameter	True Value	MPM	MMCE
$\sigma_{\varepsilon_1}^2$	450.000	449.2234	0.1228
$\sigma_{\varepsilon_2}^2$	450.000	452.6058	0.1237
$\sigma_{\varepsilon_3}^2$	450.000	6419135.5227	1652.5363

Table 3.10 – Mean of estimate parameter deviation (MEPD) and the mean of mean square error (MMSQE) for the linear approach

Parameter	MEPD	VMEPD
λ_{12_1}	-0.0007	0.0002
λ_{13_1}	0.1131	4.4550
λ_{23_1}	-158.9881	25279.1059
β_{01}	-0.1937	5.1308
β_{11}	0.0245	0.0314
β_{02}	0.1497	44.5509
β_{22}	0.0010	0.0017
β_{03}	-39059.4819	1.527156e+09
β_{33}	0.3307	2.3452
$\sigma_{g_1}^2$	-1.5480	309.9445
$\sigma_{g_1g_2}$	2.7096	187.0382
$\sigma_{g_1g_3}$	30.5112	1393.4941
$\sigma_{g_2}^2$	-6.6579	418.1242
$\sigma_{g_2g_3}$	-50.6123	2928.7626
$\sigma_{g_3}^2$	-195.1650	40746.4954
$\sigma_{\varepsilon_1}^2$	0.7766	205.1541
$\sigma_{\varepsilon_2}^2$	-2.6058	197.8305
$\sigma_{\varepsilon_3}^2$	6418685.5227	4.171439e+13

Table 3.10 presents the mean of the estimate parameter deviation, and the variance of the parameter deviation, based on the results from fitting SEM with linear effects among traits. Is possible to verify that in comparison of the Table 3.5 the deviations associate to genetics parameters and the residuals for Table 3.10 were greater than, also is possible to verify the same behavior for the MMSQE.

Figure 3.8 presents the credible interval for the genetics covariances based on the linear approach. The credible interval contains the true value for all the parameter. However posterior uncertainty was too high as indicated by the excessively wide credible intervals. Credible intervals for the covariances $\sigma_{g_1g_3}$ and $\sigma_{g_2g_3}$ contained negative and positive values. The genetic variance for the trait y_3 , $\sigma_{g_3}^2$, present a huge range, which explains the values of the MMCE presented in Table 3.8.

Figure 3.8 – Credible interval for genetics effects representing the variance (a), (b) and (c), for the traits 1, 2 and 3, respectively; covariance (d) between traits 1 and 2; covariance (e) between traits 1 and 3; covariance (f) between traits 2 and 3. The flat blue line represents the true value of the parameters

Figure 3.9 presents the credible interval for the residual variance of linear SEM. The true value for the parameters σ_{ε_1} was not in the credible interval only for two samples. Reflecting results reported in Table 3.9, credible intervals for the parameter σ_{ε_3} were far from the true value.

Figure 3.9 – Credible interval for the residuals, under linear approach, representing the σ_{ε_1} (a) the error for the trait 1; σ_{ε_2} (b) the error for the trait 2; σ_{ε_3} (c) the error for the trait 3. The blue line represents the true value of the parameters

3.2.6 Discussion

Comparing both modeling approaches resulted in substantially different inferences, which could translate to severe changes in the actions and decisions that depend of these results.

For both approaches (pSEM and sSEM), inferences for the structural effects were consistent when relationships were truly linear, i.e. λ_{121} and λ_{131} . However inferences for the linear relationship λ_{231} using sSEM were inconsistent with the true value, given that the function imposed for this causal relationship was not flexible enough for expressing the true relationship. On the other hand, when pSEM was used the results were consistent with the true values. An important aspect of the results is that, for situations where only linear relationships were assigned, the pSEM tended to assign vanishing magnitudes for the quadratic term, indicating no relevant loss from including more flexibility to the causal function.

Performances for the inferences for the fixed effects of the exogenous covariates (Tables 3.2 and 3.7; Figures 3.3 and 3.7) differed depending on the model used. Inferences of β_{01} and β_{02} on the basis of sSEM were more stable and befitting with the true value than pSEM. However, for β_{03} results for sSEM were very poor, as also indicated by the high MCE (27.4374). Even though pSEM inferences for these parameters were not too accurate, results were generally more suitable than from sSEM with smaller (almost thirty times) MCE (0.7558). For the others covariates, both models were coherent with the true value and presents equivalents MCE.

The pSEM performed better for the inferences of genetic covariances (Tables 3.3 and 3.8; Figures 3.4 and 3.8), which are the pivotal inferences in many applications of mixed effects SEM. Even though the estimated values of σ_{g1}^2 and σ_{g1g2} for sSEM were close to the true value and the MMCE obtained was small, inferences for the other parameters were too far from the true value. Inferences for σ_{g1g3} presented themselves in opposite direction. The parameters σ_{g3}^2 and σ_{g2g3} were overestimated and underestimated, respectively. Also for the genetic variance related to y_3 the MMCE were higher in the sSEM than in pSEM. These results are extremely important, as they can lead to erroneous decisions.

Inferences for residuals variances (Tables 3.4 and 3.9; Figures 3.5 and 3.9) from both models were accurate for the first two traits. However, for the third trait the results for the sSEM were heavily overestimated and the MMCE was too high, this can be seen as a result of the a “chain” effect, where poor estimates for other parameters reflects in the following parameter.

Deviations (Table 3.5 and 3.10) for both models, specially for sSEM, were considerably small for parameters associated to y_1 and y_2 . However, for y_3 , which was assumed affected by polynomial functions of other traits, the sSEM presents deviations values far greater than the pSEM. For instance, β_{03} equals to -39058.66 and -2.3989, for sSEM and pSEM, respectively. Another important issue to consider was the computational time: for both models the time to generate the 550,000 chains were equivalent, approximately 30 hours.

3.3 Conclusion

The polynomial approach resulted in equivalent or better inferences. When the relationships between traits are in fact linear, the quadratic effect estimate value was close to 0. Overall results from pSEM were more accurate.

One explanation for the lack of consistence for results of sSEM is that the inference of some parameters depend on the inference quality of other parameters. If parameters related to one trait are overestimated (or underestimate) this error is carried on the estimates of parameters of subsequent parameter and traits.

Those results show that the inclusion of an extra polynomial degree enhances the SEM expressive power, which is especially relevant when causal relationships among traits are indeed nonlinear. In such situations, similarly to linear SEM, recovering the direct, indirect and overall effects for causal and fixed are straightforward. However recovering indirect and overall effects

for the genetic effects are not trivial, once those inferences can present extra effort and require more attention. For this reason, the interpretation of genetic parameters such as heritability and genetics associations from such extended models are more complex than based on the sSEM.

References

BLAIR, R.H; KLIEBENSTEIN, D.J; CHURCHILL, G.A. What Can Causal Networks Tell Us about Metabolic Pathways? **PLoS Computational Biology**, San Francisco, v. 8, e1002458, 2012.

CHAIBUB NETO, E.. **Causal Inference Methods in Statistical Genetics**. Madison: University of Wisconsin, 2010. 140 p.

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

CHAIBUB NETO, E.; KELLER, M.P.; Keller, ATTIE, A.D.; YANDELL, B.S.. 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**, Cleveland, v. 4, p. 320-339, 2010.

GEMAN, S.; GEMAN, D. Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. New York. v. 6, p. 721-741, 1984.

GIANOLA, D.; SORENSEN, D.. Quantitative Genetic Models for Describing Simultaneous and Recursive Relationships Between Phenotypes. **Genetics**, Baltimore, v. 167, p. 1407-1424, 2004.

GONZÁLEZ-RODRÍGUEZ, A.; MOURESAN, E.F.; ALTARRIBA, J.; MORENI, C.; VARONA, L.. Non-linear recursive models for growth traits in the Pirenaica beef cattle breed. **Animal : an International Journal of Animal Bioscience**, Cambridge, v. 8, p. 904-911, 2014.

HOYER, P.O.; SHIMIZU, S.; KERMINEN, A.J.; PALVIAINEN, M.. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. **International Journal of Approximate Reasoning**, North-Holland, v. 49, p. 362-378, 2008.

KASS, R.E.; CARLIN, B.P.; GELMAN, A.; NEAL, R.. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. **The American Statistician**, Washington, DC, v. 52, p. 93-100, 1998.

KOEHLER, E.; BRWON, E.; HANEUSE, S.J.P.A.. On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. **The American Statistician**, Washington, DC, v. 63, p. 155-162, 2009.

KÖNIG, S.; WU, X. L.; GIANOLA, D.; HERINGSTAD, B.; SIMIANER, H.. Exploration of relationships between claw disorders and milk yield in Holstein cows via recursive linear and threshold models. **Journal of Dairy Science**, Champaign, v. 91, p. 395-406, 2008.

LEE, S-Y; ZHU, H-T. Statistical analysis of nonlinear structural equation models with continuous and polytomous data. **British Journal of Mathematical and Statistical Psychology**, London, v. 53, p. 209-232, 2000.

LEE, S-Y; TANG, N-S. Bayesian analysis of nonlinear structural equation models with nonignorable missing data. **Psychometrika**, Research Triangle Park, v. 71, p. 541-564, 2006.

LI, R.; TSAIH, S-W.; SCHOCLEY, K.; STYLIANOU, I.M.; WERGEGAL, J.; PAIGEN, B.; CHURCHILL, G.A.. Structural model analysis of multiple quantitative traits. **PLoS Genetics**, San Francisco, v. 2, p. 1046-1057, 2006.

LIU, BING; FUENTE, A. de la; HOESCHELE, I.. Gene Network Inference via Structural Equation Modeling in Genetical Genomics Experiments. **Genetics**, Baltimore, v. 178, p. 1763-1776, 2008.

MATURANA E.L. de; WU, XIAO-LIN; GIANOLA, D.; WEIGEL, K.A.; ROSA, G.J.M. Exploring Biological Relationships Between Calving Traits in Primiparous Cattle with a Bayesian Recursive Model. **Genetics**, Baltimore, v. 181, p. 277-287, 2009.

PEARL, J.. **Causality Models, Reasoning and Inference**. 2 ed. Cambridge, RU: Cambridge University Press, 2009. 484 p.

R. Development Core Team. R Foundation for Statistical Computing. **R 2.15.2**: A language and environment for statistical computing, Vienna, 2012. Available in <<http://www.r-project.org/>>. Acesso em: 23 nov. 2012

ROSA, G.J.M.; VALENTE, B.D.; de lo CAMPOS, G.; WU, X-L.; GIANOLA, D.; SILVA, M.A.. Inferring causal phenotype networks using structural equation models. **Genetics Selection Evolution**, London, v. 43, p. 1046-1057, 2011.

SORENSEN, D.; GIANOLA, D. Likelihood. **Bayesian and MCMC Methods in Quantitative Genetics**. New York: Springer, 2002. 740 p.

VALENTE, B.D.; ROSA, G.J.M.; de los CAMPOS, G. GIANOLA, D.; SILVA, M.A. Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models. **Genetics**, Baltimore, v. 185, p. 633-644, 2010.

VALENTE, B.D.; ROSA, G.J.M.; GIANOLA, D., WU, Xiao-Lin; WEIGEL, K. Is structural equation modeling advantageous for the genetic improvement of multiple traits?. **Genetics**, Baltimore, v.194, p. 561-572, 2013.

VALENTE, B.D.; ROSA, G.J.M.; SILVA, M.A.; TEIXEIRA, R.B.; TORRES, R.A.. Searching for phenotypic causal networks involving complex traits: an application to European quails. **Genetics, Selection, Evolution**, London, v. 43, p. 37-48, 2011.

VARONA, L.; SORENSEN, D.. Joint Analysis of Binomial And Continuous Traits with a Recursive Model: A Case Study Using Mortality and Litter Size of Pigs. **Genetics**, Baltimore, v. 196, p. 643-651, 2014.

XIONG, M.; LI, J.; FANG, X. Identification of Genetic Networks. **Genetics**, Baltimore, v. 166, p. 1037-1052, 2004.

WU, X-L; HERINGSTAD, B.; GIANOLA, D.. Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. **Journal of Animal Breeding and Genetics**, Berlin, v. 27, p. 3-15, 2010.

4 USING POLYNOMIAL STRUCTURAL EQUATION MODELS TO ESTIMATE THE EFFECTS RELATED TO CALVING IN PRIMIPAROUS HOLSTEIN CATTLE

Abstract

Holstein is the most important dairy breed in USA, as indicated by the great income these animals provide relative to the feed costs, their genetic merit, as well as outstanding production and adaptability to a wide range of environmental conditions. Studies of birth-related traits (such as gestation length, calving difficulty and stillbirth) are pivotal for dairy breeds. Such multiple trait contexts often involve causal relationships between traits. Structural equation models (SEM) are an alternative to classic multiple trait mixed models that account for causal relationships. However, SEM in quantitative genetics typically account for linear causal relationship among traits that can be unrealistic for many scenarios. We propose a polynomial approach to overcome this problem, using data related to calving traits in Holsteins. We compare the multiple trait model and polynomial structural equation models using different polynomial degrees in order to verify the advantages of higher degree polynomials in the SEM approach. For the causal models, we assume that gestation length affected calving difficulty and stillbirth, and calving difficulty affected stillbirth. Results provided by second-degree polynomial SEM, when comparing the trait prediction, were more accurate than the linear SEM for all causal relationships. Even though the estimates for multiple trait mixed models and polynomial SEM should not be compared directly, for situations where there is interest in recovering direct associations this is useful. In the comparison between them to multiple trait mixed models, we observed that direct genetic covariances changed drastically, resulting potentially in mistaken decisions in animal selection.

Keywords: Structural models; Causal inference; Quantitative genetics; Polynomial regression; Linear mixed models; Holstein dairy cattle

4.1 Introduction

The Holstein dairy cattle breed is considered the most important for milk production in the USA. Some reasons for this are: good return over feed costs, genetic merit for economically important traits, as well as outstanding production and adaptability to a wide range of environmental conditions.

During recent decades studies involving birth-related traits (such as gestation length, calving difficulty and stillbirth) have increased (MATURANA et al., 2008; GROEN et al. 1997). The primary concerns that motivated this research are the economic importance of these traits and animal well-being considerations. However, in evaluating those traits, it should be observed that a relationship among them may occur, leading to “chain effects”.

In scenarios where genetic and environmental correlations exist between phenotypic traits, multiple trait mixed models (MTMM) have been widely used. However, when there is interest in accounting for direct functional relationships between phenotype traits, MTMM present

some limitations, even though for some situations reparameterization can be performed to fit those relationships, since those models cannot express recursive effects, something not rare in biological systems (MATURANA et al., 2008; CHAIBUB NETO, 2010; LIU; FUENTE; HOESCHELE, 2008; VALENTE et al., 2010; VALENTE et al., 2011; BLAIR; KLIEBENSTEIN; CHURCHILL, 2012).

Those phenotypic relationships can be interpreted as causal relationships. Structural equation models (SEM) have been used to estimate these causal effects (WRIGHT 1934, MATURANA et al., 2008; PEARL, 2009; CHAIBUB NETO, 2010). SEM can be seen as multiple-trait regression models where the response variables, usually on the left-hand side of the equation, can be considered to be covariates in the equations, for other response variables (ROSA et al. 2011; ROSA; VALENTE, 2014; LEE; ZU, 2000; VALENTE et al., 2013).

Many applications of SEM in quantitative genetics were published after Gianola and Sorensen (2004). They were not the first authors to apply SEM in genetics, as can be seen in Wright (1921). However, their work can be considered groundbreaking in proposing mixed effects of SEM in genetics, which is specially relevant in animal breeding (CAMPOS et al., 2006; CAMPOS; GIANOLA; HERINGSTAD, 2006; MATURANA et al., 2007; VARONA et al., 2007; WU; HERINGSTAD; GIANOLA, 2007; de MATURANA et al., 2008).

Traits related to calving difficulty and mortality in cows and calves are potentially interesting examples where causal relationships among traits can be observed and studied. Maturana et al. (2008) and Meijering (1984) present situations where the calving difficulty increases the culling risk as well as veterinary and labor costs. Those difficulties can be caused by incompatibility between the calf's size and dam's pelvic area (BARRIER; HASKELL, 2011; PHILIPSSON; STEINBOCK, 2003). Furthermore, the mortality of calves decreases milk production in the next lactation, and leads to lower female fertility in the next reproductive cycle (MATURANA et al., 2007; MATURANA; UGARTE; GONZÁLEZ-RECIO, 2007; MATURANA et al., 2008; DEMATAWEWA; BERGER, 1997).

Interest in studies of dairy cattle breeding involving birth traits, such as gestation length, calving difficulty and perinatal mortality have been increasing in the past decade (FOUZ et al., 2012; MATURANA et al., 2008; GROEN et al. 1997, STEINBOCK, 2003; BARRIER, 2012, MURRAY et al., 2015). Calving can present varying difficulty. This difficulty, also known as dystocia, generally means that during the delivery process some type of assistance is required. The amount of assistance is commonly used to provide a score of difficulty (BARRIER et al., 2012; BARRIER; HASKELL; 2011, UEMATSU et al., 2013).

Stillbirth, also known perinatal mortality, can be defined as calf death before, during or within 48 hours after calving (MATURANA et al., 2008; PURFIELD et al., 2015; HINRICHS et al., 2015, MEE; BERRY; CROMIE, 2011). According to Maturana et al. (2008) and Meyer et al., (2000), calving difficulty can be an important predictor of stillbirth, although the former trait is not the only factor affecting the latter (PURFIELD et al., 2015; HINRICHS et al., 2015, JOHANSON et al., 2011, MEE; BERRY; CROMIE, 2011; UEMATSU et al., 2013).

Typically, applications of SEM to quantitative genetics assume a linear relationship among the traits, even though this assumption is unrealistic in many scenarios (MATURANA et al., 2008; GONZÁLEZ-RODRÍGUEZ et al., 2014; VARONA et al., 2014; KÖNIG et al., 2008). Non-linear SEM have been used in other areas, such as by Berndt, Hall and Hall (1974), Dijkstra and Hoenseler (2011), Jeffrey, Weiss and Hsu (2012), Lee and TANG (2006), Stephan et al. (2008) and Turner et al. (1961).

4.2 Material and Methods

4.2.1 Data

The data consisted of a total of 2,305,499 observations of Holstein cows between 2000 and 2013, recorded as part of the National Association of Animal Breeders Calving Ease Program. Three traits were evaluated: gestation length (GL), calving difficulty (CD) and stillbirth (SB). Four other variables were treated as covariates: age of cow, sex of calf, year and season of calving. After editing and quality control, data from 291,241 primiparous cows, sired by 16,403 bulls distributed in 1,374 herds, were used in the analysis.

Gestation length was evaluated as the difference between the breeding and calving date, measured in days. Following Maturana et al. (2008), Johanson et al. (2011), Fouz et al. (2013), an interval that ranges from 255 to 295 days for GL was considered, so that unrealistic values were dropped (BARRIER; HASKELL, 2011; BARRIER et al., 2012; DEMATAWEWA; BERGER, 1997; FOUZ et al., 2013; JOHANSON et al., 2011).

The trait stillbirth, which indicates if calves died within 24 or 48 hours after calving, was initially scored as follows: 1, alive; 2, dead under 24 hours after calving; and 3, dead between 24 and 48 hours after calving (MATURANA et al., 2008; PURFIELD et al., 2015; HINRICHS et al., 2015; MEE; BERRY; CROMIE, 2011). However, the values 2 and 3 were combined as a unique information in order to study the alive animal and dead animal. However, the values 2 and 3 were combined as a single data point to study the alive and dead animals. According to Meyer et al. (2001), Maturana et al. (2008), Mee, Berry and Cromie (2011) and Uematsu et al. (2013), this information is relevant because early mortality imposes significant costs on the industry.

Table 4.1 presents summary descriptions related to each trait. The proportion of labor with some difficulty (i.e. CD between 2 and 5) was 32.47%. The proportion of calves that survived 48 h after calving was almost eleven times greater than perinatal mortality. The variance coefficient for the GL trait was 2.13%, indicating small variation. It is possible to verify symmetry for this trait.

Table 4.1 – Summary statistics of calving traits

Calving Difficulty					
	1	2	3	4	5
%	67.53	17.29	10.34	2.98	2.04

Stillbirth		
	1	2
%	91.55	8.45

Gestation Length (in days)				
SD	Min	Median	Mean	Max
5.9	255.0	278.0	278.2	295.0

Table 4.2 and 4.3 present the summary of the fixed effects age, measured in days, seasons and year of calving, and sex of the calf. Age distribution is symmetric. The frequencies of season and sex are almost uniform for all categories. Only a small proportion of data was collected in the first three years (2003, 2004 and 2005), which comprise less than 5% of the labors.

Table 4.2 – Summary statistics of covariates used as fixed effects

Calf sex		
	Female	Male
%	45.63	54.36

Age (in days)				
SD	Min	Mean	Median	Max
189.1	545.0	855.6	759	1280.0

Season				
	Autumn	Winter	Spring	Summer
%	23.58	26.02	26.43	23.97

Table 4.3 – Proportion of calves born per year

	Calf year				
	2003	2004	2005	2006	2007
%	0.0087	0.0167	0.0180	0.0987	0.1282

	Calf year					
	2008	2009	2010	2011	2012	2013
%	0.1215	0.1113	0.1353	0.1388	0.1385	0.0843

Figure 4.1 presents the relationship between gestation length and proportion of birth without difficulty (green line), and the proportion of alive cattle (blue line). As it can be seen, for both traits the relationships are nonlinear. The behavior for live birth is biologically expected, since earlier calving can result in underdeveloped calves.

Figure 4.1 – Proportion of alive calves (live birth) and labor difficulty according to gestation length

4.2.2 Methods

Different models were fitted to study the system of traits: multivariate linear mixed models, also known as multiple trait mixed models in quantitative genetics, and mixed effects structural equation models with linear and polynomial effects. The effects related to herd and genetics were treated as random effects affecting all traits. The equations for each trait were treated as

the animal model, since this model carries more genetic information than others (e.g., sire, or grand-sire models) because it accounts for kinship information among all the individuals.

As fixed effects, we used the age of the dams grouped into four categories defined by the quantiles, the calving year (11 levels), the calf sex (2 levels), and the seasons (4 levels). We treated as response variables the traits GL, CD and SB. GL was rescaled as $\frac{GL}{\max(GL)}$ in order to avoid problems related to variance dimensionality.

4.2.2.1 Multiple trait mixed model (MTMM)

Assuming that each trait is a function of the same random and fixed effects, the MTMM is given by equation 4.1

$$Y = X\beta + Z_h h + Z_u u + e \quad (4.1)$$

where Y represents the vector of t traits for the n individuals with dimension $(n * t \times 1)$, X represents the matrix of incidence for the p fixed effects with dimension $(n * t \times p)$, β is a vector of fixed effects with dimensions $(p \times 1)$, Z_h represents the incidence matrix of the random effects related to herd with dimension $(t \times h)$ and Z_u represents the incidence matrix of the random effect related to the animal with dimension $(t \times n)$.

The h term contains the random effects related to herd, which are assumed to be a normally distributed, $h \sim N(0, I_n \otimes \Sigma_h)$, where I_n represent the identity matrix and Σ_h represent the variance covariance matrix for the herd.

The u term contains the random effect related to genetics, which is assumed to have a normal distribution, $u \sim N(0, A \otimes \Sigma_u)$, where A represents the kinship matrix for the cows and Σ_u represent the variance covariance matrix for the genetic effects. The e term contains the residuals, which are assumed to be normally distributed $e \sim N(0, I \otimes \Sigma_e)$, where Σ_e represents the variance and covariance matrix for the residuals. It is also assumed that $Cov(e_i, e_j) = 0$.

We can express the structure of this model using graphical notation as in Figure 4.2. The single arrows represent the directed effects and the double pointed arrows represent the covariances between the variables. The fixed and the random effects related to herd were omitted to provide a clear view.

Figure 4.2 – Multiple trait animal model, with three traits, expressed using a graphic method, where single arrows express the causal relationship and double arrows express the correlation among variables

4.2.2.2 Linear mixed effects structural equation models

A recursive model was proposed to account for biological relationships between traits, as described next. To explore such causal relationships, we assumed a known causal relationship and the structural equation model (SEM) has been used as tool to estimate these effects (CERQUEIRA et al. 2014).

We assume the same relationship between GL, CD and SB as already applied in other studies, such as Maturana et al. (2008), Meijering (1984), Barrier and Haskell (2011) and Purfield et al. (2014). It is assumed that GL affects the liabilities to CD and SB, and that the chance of CD has a causal effect on the chance of SB. These relationships can be represented with equations or graphically, as in Figure 4.3.

We assume linear, second and third degree polynomial SEM to estimate the causal relationships between traits. The model for the linear SEM (ISEM) can be seen in equation 4.2.

$$Y = (\Lambda \otimes I_n)Y + X\beta + Z_h h + Z_u u + e \quad (4.2)$$

where Y , X , β , Z_h , h , Z_u , u and e have the analogous meaning as previously defined. Λ represents the matrix of linear structural effect with dimensions $(t \times t)$, also with diagonal equal to 0. The multiple trait mixed model can be considered a particular case of the SEM when the Λ matrix contains only zeros.

Figure 4.3 presents the DAG proposed by Maturana et al. (2008). Apart from the direct effects already described, GL has an indirect effect on the liability to SB mediated by CD. The liability to SB would be affected by a change in GL through the mediating effect of the liability to CD ($GL \rightarrow CD \rightarrow SB$). The indirect effect is calculated as the product of the structural coefficients $GL \rightarrow CD \times CD \rightarrow SB$. The overall effect of GL on SB can be represented as the sum of the direct and indirect effects ($GL \rightarrow SB + GL \rightarrow CD \times CD \rightarrow SB$) (SHIPLEY, 2002).

Figure 4.3 – The causal structure, with three traits, expressed using a graphical methodology, representing that GL has a direct effect on both liabilities CD ($GL \rightarrow CD$) and SB ($GL \rightarrow SB$). CD has a directed effect on the liability SB ($CD \rightarrow SB$)

4.2.2.3 Polynomial mixed structural equation models

For a polynomial SEM (pSEM) of d degree the relationship structures between the traits can be seen in equation 4.3. Figure 4.3 still applies as the structure for the pSEM.

$$Y = (\Lambda_1 \otimes \mathbf{I}_n)Y + (\Lambda_2 \otimes \mathbf{I}_n)Y^2 + \dots + (\Lambda_d \otimes \mathbf{I}_n)Y^d + X\beta + Z_h h + Z_u u + e \quad (4.3)$$

where Y , X , β , Z_h , h , Z_u , u and e are as presented before. The index d represents the polynomial degree. Λ_k contains the coefficients of the structural effects as previously defined associated to each polynomial degree. Y^d is the matrix of observed traits, in which every element is raised to d . When $d = 1$, the model is reduced to the ISEM.

4.2.3 Estimation and computation

The estimation process was split into two stages. The first was estimating fixed effects, using the maximum likelihood process. For the second stage, the restricted maximum likelihood (REML) with the average information (AI) algorithm was used to estimate the (co)variance components. This method has been used elsewhere under the assumptions of normal distribution and was developed in 1971 by Patterson and Thompson (HARVILLE, 1977).

REML methods have been widely used, because they correct for the degree of freedom involved estimating fixed parameters, leading to less biased estimates than the maximum likelihood estimate (HARVILLE, 1977; GILMOUR; THOMPSON; CULLIS, 1995; PINHEIRO; BATES, 2000).

Two free software, R and DMU (MADSEN et al., 2010), were used for the analysis. The R software was used to edit the dataset, define the model and analyze the results. The software DMU was used to fit the model.

4.3 Results and Discussion

This section presents and compares the results obtained from using the three different models. Section 4.3.1 presents the results of fixed effects for the multiple trait mixed models; section 4.3.2 shows the results of fixed effects for the linear SEM; section 4.3.3 presents the results of fixed effects for second-degree polynomial SEM; in section 4.3.4 shows the results of fixed effects for third-degree polynomial SEM; section 4.3.5 sets out the results of fixed effects for an alternative model without a causal relationship between GL and CD and a third-degree polynomial SEM as link between GL and SB are featured, section 4.3.6 contains the results of random effects; and section 4.3.7 discusses all the results.

4.3.1 Results of Multiple trait mixed models (fixed effects)

Table 4.4 – Parameter estimates and standard errors for the fixed effects for all traits using MTMM

Characteristics	Group	Gestation Length Estimate (s.e)	Calving Difficulty Estimate (s.e)	Stillbirth Estimate (s.e)
Age	1	0.93804 (0.00023)	0.26699 (0.00791)	0.11376 (0.00307)
	2	0.94321 (0.00023)	0.24864 (0.00785)	0.10443 (0.00301)
	3	0.94460 (0.00022)	0.21444 (0.00780)	0.08230 (0.00297)
	4	0.94716 (0.00022)	0.13462 (0.00769)	0.03563 (0.00288)
Sex	Male	Reference	Reference	Reference
	Female	0.00385 (0.00007)	0.09252 (0.00152)	0.00962 (0.00102)
Year	2002	Reference	Reference	Reference
	2003	0.00224 (0.00045)	0.11006 (0.00976)	0.02192 (0.00636)
	2004	0.00109 (0.00034)	0.05441 (0.00743)	0.00001 (0.00482)
	2005	0.00076 (0.00033)	0.05675 (0.00713)	0.00308 (0.00464)
	2006	-0.00019 (0.00020)	0.04805 (0.00451)	0.00502 (0.00288)
	2007	0.00026 (0.00019)	0.03852 (0.00421)	0.00206 (0.00269)
	2008	0.00024 (0.00019)	0.02968 (0.00413)	0.00027 (0.00266)
	2009	0.00028 (0.00019)	0.01856 (0.00409)	0.00142 (0.00265)
	2010	-0.00016 (0.00018)	0.00560 (0.00391)	-0.00225 (0.00254)
	2011	-0.00014 (0.00017)	-0.00658 (0.00375)	-0.00498 (0.00245)
	2012	-0.00002 (0.00017)	0.00185 (0.00352)	-0.00566 (0.00233)
Season	Autumn	Reference	Reference	Reference
	Winter	-0.00080 (0.0001)	0.02692 (0.00228)	0.01031 (0.00152)
	Spring	-0.00338 (0.0001)	0.00098 (0.00222)	-0.01212 (0.00148)
	Summer	-0.00291 (0.0001)	-0.01095 (0.00218)	-0.01055 (0.00145)

Table 4.4 presents the results for the fixed effects related to MTMM for the traits GL, CD and SB. The covariate age presents a higher value for GL. Also it is possible to see that for all groups the estimates are almost the same, but for SB and CD, group 4 presents the smallest values. The calf sex male, the year 2002 and autumn were used as reference. For the year of 2003 for CD preset highest estimate value. The results for season indicate it as an important factor, especially winter for CD and also spring and summer for GL.

Figure 4.4 – Observed and estimated relation between the proportion of calving difficulty (a) and live birth (b) according to gestation length for the MTMM

Figure 4.4 presents the proportion of CD and SB according to gestation length, where the observed CD proportion and fitted values are represented by the blue and red lines, respectively. The dotted lines represent the confidence interval. Figure 4.4 (a) shows that the model did a good job expressing the average behavior of CD, although the performance was not so good for

SB, as can be seen in Figure 4.4 (b).

The confidence intervals for GL shorter than 270 days are larger, but it should be stressed that few observations of CD and SB were considered in the analysis for this range of GL. For SB, the values of s are also poorer (in terms of expected value and confidence interval) in the extreme values of GL (i.e., smaller than 270 and larger than 287).

4.3.2 Results of linear structural models (fixed effects)

Table 4.5 – Parameter estimates and standard errors for the fixed effects for all traits using ISEM

Characteristics	Group	Gestation Length Estimate (s.e)	Calving Difficulty Estimate (s.e)	Stillbirth Estimate (s.e)
Age	1	0.93803 (0.00023)	-0.28020 (0.03747)	0.67017 (0.02417)
	2	0.94320 (0.00022)	-0.30125 (0.03765)	0.66682 (0.02429)
	3	0.94458 (0.00022)	-0.33650 (0.03770)	0.65038 (0.02432)
	4	0.94714 (0.00022)	-0.41803 (0.03777)	0.61704 (0.02438)
Sex	Male	Reference	Reference	Reference
	Female	0.00385 (0.00007)	0.09016 (0.00154)	-0.00173 (0.00101)
Year	2002	Reference	Reference	Reference
	2003	0.00222 (0.00045)	0.11461 (0.00966)	0.00882 (0.00613)
	2004	0.00108 (0.00034)	0.06013 (0.00733)	-0.00522 (0.00463)
	2005	0.00075 (0.00033)	0.06138 (0.00706)	-0.00272 (0.00446)
	2006	-0.00020(0.00020)	0.05206 (0.00440)	-0.00054 (0.00271)
	2007	0.00025 (0.00019)	0.04220 (0.00411)	-0.00184 (0.00254)
	2008	0.00022 (0.00019)	0.03316 (0.00406)	-0.00239 (0.00252)
	2009	0.00025 (0.00019)	0.02124 (0.00404)	-0.00007 (0.00253)
	2010	-0.00020(0.00018)	0.00931 (0.00387)	-0.00190 (0.00243)
	2011	-0.00017(0.00017)	-0.00353 (0.00373)	-0.00305 (0.00236)
	2012	-0.00003(0.00017)	0.00331 (0.00351)	-0.00528 (0.00226)
Season	Autumn	Reference	Reference	Reference
	Winter	-0.00080 (0.00011)	0.02812 (0.00228)	0.00613 (0.00148)
	Spring	-0.00338 (0.00010)	0.00314 (0.00222)	-0.01451 (0.00144)
	Summer	-0.00291 (0.00010)	-0.00879 (0.00218)	-0.01072 (0.00142)

Table 4.5 present the results for all fixed effects for all traits when using ISEM. The estimates related to age for all traits seem not present differences. In comparison to MTMM, it is possible to see that for CD all the directions of age effect change from positive to negative. The sex of the calf for GL and CD lead us to believe there is an effect, however SB seems not to have any effect. The year 2003 for CD presents values almost twice those of any of the others related to CD. For season, the results are quite similar to those obtained when using MTMM.

Table 4.6 presents the magnitudes of causal relationships. All linear effects were statistically significant, as their confidence intervals do not contain the value 0. It is possible to see that GL has a positive effect on CD (λ_{12}) and a direct negative effect on SB (λ_{13}). CD presented a negative effect on SB (λ_{23}). The indirect effect of GL on SB were negative, equal to $-0,08674$ ($\lambda_{12} \times \lambda_{23}$).

Table 4.6 – Parameter estimates and standard errors and 95% confidence intervals for causal effects using ISEM

Parameter	Estimate (s.d)	2.5% C.I	97.5% C.I
λ_{12}	0.58416 (0.03907)	0.50758	0.66074
λ_{13}	-0.63526 (0.02557)	-0.68537	-0.58515
λ_{23}	-0.14850 (0.00119)	-0.14618	-0.15083

Figure 4.5 presents the proportion of CD and SB according to gestation length for the ISEM. Figure 4.5 (a)) shows that the model expresses the behavior of CD, but it is possible to verify that this relation does not express the real value, since the confidence interval does not contain the observed value. For SB, the ISEM does not recover the behavior and cannot be considered a good approach, as can be seen in Figure 4.5 (b). In comparison to the MTMM, the ISEM present worse results, so this approach cannot be considered a good one.

Figure 4.5 – Observed and estimated relation between the proportion of calving difficulty (a) and live birth (b) according to gestation length for the ISEM

4.3.3 Results of second-degree polynomial structural models (fixed effects)

Table 4.7 presents the results for all fixed effects for all traits when using the second pSEM. As can be seen, the results related to GL are the same, which was expected since the model for this trait is equal in both approaches. For the age related to CD, the result changed direction in relation the ISEM, from negative to positive, and the magnitude for all ages was at least three times the value in the MTMM. Even though having higher magnitude, the effects related to age were not significantly different than zero. The results for sex, year and season did not change significantly. For SB, the age effect changed drastically in relation to MTMM and ISEM, however for the other covariates the results were quite similar.

Table 4.7 – Parameter estimates and standard errors for the fixed effects for all traits using second degree pSEM

Characteristics	Group	Gestation Length Estimate (s.e)	Calving Difficulty Estimate (s.e)	Stillbirth Estimate (s.e)
Age	1	0.93803 (0.00023)	0.84065 (0.95415)	23.87770 (0.62242)
	2	0.94319 (0.00022)	0.81962 (0.95416)	23.87460 (0.62243)
	3	0.94457 (0.00022)	0.78438 (0.95418)	23.85840 (0.62244)
	4	0.94714 (0.00022)	0.70286 (0.95425)	23.82670 (0.62248)
Sex	Male	Reference	Reference	Reference
	Female	0.00385 (0.00007)	0.09023 (0.00154)	-0.00085 (0.00101)
Year	2002	Reference	Reference	Reference
	2003	0.00222 (0.00045)	0.11469 (0.00966)	0.00927 (0.00612)
	2004	0.00107 (0.00034)	0.06023 (0.00733)	-0.00429 (0.00462)
	2005	0.00075 (0.00033)	0.06149 (0.00706)	-0.00152 (0.00445)
	2006	-0.00020 (0.00020)	0.05204 (0.00440)	-0.00094 (0.00271)
	2007	0.00025 (0.00019)	0.04217 (0.00411)	-0.00204 (0.00253)
	2008	0.00022 (0.00019)	0.03312 (0.00406)	-0.00267 (0.00252)
	2009	0.00025 (0.00019)	0.02120 (0.00404)	-0.00055 (0.00252)
	2010	-0.00020 (0.00018)	0.00926 (0.00387)	-0.00262 (0.00242)
	2011	-0.00018 (0.00017)	-0.00357 (0.00373)	-0.00387 (0.00236)
	2012	-0.00003 (0.00017)	0.00329 (0.00351)	-0.00574 (0.00226)
Season	Autumn	Reference	Reference	Reference
	Winter	-0.00080 (0.00011)	0.02813 (0.00228)	0.00633 (0.00148)
	Spring	-0.00338 (0.00010)	0.00313 (0.00222)	-0.01479 (0.00144)
	Summer	-0.00291 (0.00010)	-0.00881 (0.00218)	-0.01139 (0.00142)

Table 4.8 – Parameter estimates, standard deviations and 95% confidence intervals for causal effects using second-degree pSEM

Parameter	Estimate (s.d)	2.5% C.I	97.5% C.I
λ_{12_1}	-1.79472 (2.02273)	-5.75920	2.16976
λ_{12_2}	1.26196 (1.07185)	-0.83883	3.36275
λ_{13_1}	-49.84520 (1.31953)	-52.43143	-47.25897
λ_{13_2}	26.07350 (0.69927)	24.70296	27.44404
λ_{23_1}	0.14819 (0.00118)	0.14586	0.15051

Table 4.8 shows the results related to the causal coefficients for the second pSEM. The linear, λ_{12_1} and quadratic effect λ_{12_2} related to the causal effects between GL and CD were considered not significant, since the confidence interval contains zero. The linear λ_{13_1} and quadratic λ_{13_2} effects were considered significant. The CD exerts a positive effect on SB λ_{23} . For the structural relationship between CD and SB, no second-degree polynomials were used,

as the variables have only two possible values. The direct effects GL on SB have two intervals, one where the effect is positive for the quadratic component and another where the effect is negative for the linear component. However, this relation is always decreasing, given that GL ranges between 0 and 1, so the indirect and overall effect the behavior is similar.

Figure 4.6 presents the proportion of CD and SB according to gestation length, where the red, blue and dotted lines are as defined previously. Figure 4.6 (a) shows that the model expressed the behavior of CD, but did not perform as well in recovering the real value as did the ISEM. For SB, the second pSEM performed well in recovering the behavior of SB. In comparison to the MTMM and ISEM, the second-degree pSEM can be considered a better approach, as indicated by Figure 4.6 (b).

Figure 4.6 – Observed and estimated relation between the proportion of calving difficulty (a) and live birth (b) according to gestation length for the second pSEM

4.3.4 Results of third-degree polynomial structural models (fixed effects)

Table 4.9 present the results for all fixed effects for all traits when using the third pSEM. As can be seen, the results related to GL were the same as presented previously. For the age related to CD, the estimate of fourth quantile was zero, the estimates for the other parameters were smaller than the MTMM and second pSEM. However, results for sex, year and season did not change significantly. For the SB the age estimate was smaller than the obtained when using the others models and estimate for the fourth quantile related to age was zero, however for the others covariates the results were quite similar.

Table 4.9 – Parameter estimate, standard error for the fixed effects for all traits using the third degree pSEM

Characteristics	Group	Gestation Length Estimate (s.e)	Calving Difficulty Estimate (s.e)	Stillbirth Estimate (s.e)
Age	1	0.93803 (0.00023)	0.13789 (0.00243)	0.05078 (0.00157)
	2	0.94319 (0.00022)	0.11683 (0.00229)	0.04773 (0.00148)
	3	0.94457 (0.00022)	0.08157 (0.00225)	0.03168 (0.00146)
	4	0.94714 (0.00022)	0.00000 (0.00000)	0.00000 (0.00000)
Sex	Male	Reference	Reference	Reference
	Female	0.00385 (0.00007)	0.09019 (0.00154)	-0.00075 (0.00101)
Year	2002	Reference	Reference	Reference
	2003	0.00222 (0.00045)	0.11469 (0.00966)	0.00936 (0.00612)
	2004	0.00107 (0.00034)	0.06021 (0.00733)	-0.00422 (0.00462)
	2005	0.00075 (0.00033)	0.06147 (0.00706)	-0.00148 (0.00445)
	2006	-0.00020 (0.00020)	0.05206 (0.00440)	-0.00090 (0.00271)
	2007	0.00025 (0.00019)	0.04218 (0.00411)	-0.00201 (0.00253)
	2008	0.00022 (0.00019)	0.03313 (0.00406)	-0.00264 (0.00252)
	2009	0.00025 (0.00019)	0.02122 (0.00404)	-0.00052 (0.00252)
	2010	-0.00020 (0.00018)	0.00929 (0.00387)	-0.00262 (0.00243)
	2011	-0.00018 (0.00017)	-0.00354 (0.00373)	-0.00386 (0.00236)
	2012	-0.00003 (0.00017)	0.00331 (0.00351)	-0.00574 (0.00226)
Season	Autumn	Reference	Reference	Reference
	Winter	-0.00080 (0.00011)	0.02813 (0.00228)	0.00630 (0.00148)
	Spring	-0.00338 (0.00010)	0.00316 (0.00222)	-0.01489 (0.00144)
	Summer	-0.00291 (0.00010)	-0.00878 (0.00218)	-0.01147 (0.00142)

Table 4.10 shows the results related to the causal coefficients for the third-degree pSEM. The linear (λ_{12_1}), quadratic (λ_{12_2}) and cubic (λ_{12_3}) coefficients related to the causal effects among GL and CD cannot be considered significant, as their confidence intervals contain zero. The linear, quadratic and cubic effects that explains the relation of GL and SB (λ_{13_1}), λ_{13_2} and λ_{13_3}), were considered significant. The CD exert a positive effect on SB (λ_{23}). The direct and indirect effects GL on SB presented two intervals, one where the effect was positive and another where

the effect was negative.

Table 4.10 – Parameter estimates, standard deviations and 95% confidence intervals for causal effects using third-degree pSEM

Parameter	Estimate (s.d)	2.5% C.I	97.5% C.I
λ_{12_1}	-0.19874 (1.02779)	-2.21318	1.81569
λ_{12_2}	0.24311 (2.17498)	-4.01977	4.50599
λ_{13_3}	0.12410 (1.15077)	-2.13136	2.37957
λ_{13_1}	26.06240 (0.67042)	24.74841	27.37639
λ_{13_2}	-54.48110 (1.41892)	-57.26213	-51.70007
λ_{13_3}	28.47620 (0.75080)	27.00466	29.94774
λ_{23_1}	0.14824 (0.00118)	0.14592	0.15056

Figure 4.7 presents the proportion of CD and SB according to gestation length, where the red, blue and dotted lines areas as defined previously. Figure 4.7 (a) shows that the model expressed the behavior of CD, however did not perform as well in recovering the real value as in the ISEM. For SB, the second pSEM was able to recover the behavior of SB well. In comparison to the MTMM and ISEM, the third pSEM can be considered a better approach. Both pSEM present similar behavior, as can be verified in Figure 4.7 (b).

Figure 4.7 – Observed and estimated relation between the proportion of calving difficulty (a) and live birth (b) according to gestation length for third pSEM

4.3.5 Results without causal relationships between GL and CD and second-degree polynomial structural models (fixed Effects)

The causal relationship between GL and CD was ignored and a second-degree polynomial causal relationship between GL and SB was assumed in order to verify if a model composed of the combination of causal and non-causal structure was capable of achieving better results than those previously presented. This model was named mixed pSEM.

Table 4.11 present the results for all fixed effects for all traits when using the mixed pSEM. As can be seen, the results related to GL remained the same as the previous models, which was expected since no modification was made in this part of the structure for all models. The results obtained related to CD were similar to MTMM, since both structures were similar for this trait.

Also, the results related to SB were similar to the second pSEM, as expected given that the structures are same.

Table 4.11 – Parameter estimates and standard errors for the fixed effects for all traits using the mixed pSEM

Characteristics	Group	Gestation Length Estimate (s.e)	Calving Difficulty Estimate (s.e)	Stillbirth Estimate (s.e)
Age	1	0.93803 (0.00023)	0.26831 (0.00782)	23.85680 (0.62240)
	2	0.9431 (0.00022)	0.25026 (0.00776)	23.85360 (0.62241)
	3	0.94458 (0.00022)	0.21584 (0.00771)	23.83740 (0.62242)
	4	0.94714 (0.00022)	0.13565 (0.00762)	23.80550 (0.62247)
Sex	Male	Reference	Reference	Reference
	Female	0.00385 (0.00007)	0.09243 (0.00153)	-0.00089 (0.00101)
Year	2002	Reference	Reference	Reference
	2003	0.00220 (0.00045)	0.11574 (0.00967)	0.00930 (0.00612)
	2004	0.00106 (0.00034)	0.06055 (0.00734)	-0.00428 (0.00462)
	2005	0.00073 (0.00033)	0.06172 (0.00706)	-0.00150 (0.00445)
	2006	-0.00021 (0.00020)	0.05184 (0.00440)	-0.00090 (0.00271)
	2007	0.00023 (0.00019)	0.04223 (0.00412)	-0.00203 (0.00253)
	2008	0.00021 (0.00019)	0.03326 (0.00406)	-0.00266 (0.00252)
	2009	0.00025 (0.00019)	0.02142 (0.00404)	-0.00055 (0.00252)
	2010	-0.00021 (0.00018)	0.00923 (0.00387)	-0.00262 (0.00242)
	2011	-0.00019 (0.00017)	-0.00363 (0.00373)	-0.00387 (0.00236)
	2012	-0.00004 (0.00017)	0.00329 (0.00352)	-0.00574 (0.00226)
Season	Autumn	Reference	Reference	Reference
	Winter	-0.00080 (0.00011)	0.02762 (0.00228)	0.00637 (0.00148)
	Spring	-0.00338 (0.00010)	0.00116 (0.00222)	-0.01470 (0.00144)
	Summer	-0.00291 (0.00010)	-0.01050 (0.00218)	-0.01132 (0.00142)

Results related to the causal relationship are presented in Table 4.12. The causal coefficients, linear (λ_{13_1}) and quadratic (λ_{13_2}), that explain the relation between GL and SB were considered significant and this relation was decreasing. The CD exerted a positive effect on SB. In others words, an intervention in CD will positively affect SB. Also it is possible to verify that the values of all coefficients for these traits are similar to those in Table 4.8.

Figure 4.8 presents the proportion of CD and SB according to gestation length, where the red, blue and dotted lines are as previous defined. Figure 4.8 (a) shows that the model expressed the behavior of CD, but this model did not do as good a job of recovering the real value as in the MTMM, even though the estimates for fixed effects were similar in both models. Figure 4.8 (b) shows that the behavior of SB in relation to GL for the mixed pSEM was similar to the second-degree pSEM, a result that was expected since the structure for this trait in both models are similar.

Table 4.12 – Parameter estimates, standard deviations and 95% confidence intervals for causal effects using mixed pSEM

Parameter	Estimate (s.d)	2.5% C.I	97.5% C.I
λ_{13_1}	-49.82630 (1.31950)	-52.41247	-47.24013
λ_{13_2}	26.07720 (0.69925)	24.70670	27.44770
λ_{23_1}	0.14758 (0.00118)	0.14526	0.14990

Figure 4.8 – Observed and estimated relation between the proportion of calving difficulty (a) and live birth (b) according to gestation length for the mixed pSEM

4.3.6 Results of variances-covariances components

Results for the variances and covariances components associated with herd and genetics, for all models, are presented in this section. All values for variance and covariance were multiplied by 100,000 to present a better view

Table 4.13 presents the results for the random effects related to herd for all models. As can be seen, the herd variance of CD was far greater than the other variances and covariances. The covariance among CD and SB, $\sigma_{CD \times SB}$, changed direction when comparing the MTMM to all SEMs. The variance for SB σ_{SB}^2 doubled in size for the SEM in comparison to MTMM. From the results related to correlation, it can be seen that the variance-covariance components to herd for GL and SB are slightly correlated and for the other traits there is no correlation.

Table 4.13 – Herd covariance and correlations estimate for GL, CD and SB for the MTMM, second pSEM, third pSEM and mixed pSEM

Parameter	MTMM	lSEM	Second pSEM	Third pSEM	Mixed pSEM
σ_{GL}^2	1.1481	1.1508	1.1551	1.1549	1.1552
$\sigma_{GL \times CD}$	2.9736	2.2238	2.1521	2.1329	2.8038
$\sigma_{GL \times SB}$	1.3106	1.6158	2.0771	2.0834	2.0571
σ_{CD}^2	5,028.4955	5,024.3453	5,023.8798	5,023.4151	5,024.9032
$\sigma_{CD \times SB}$	6.6036	-740.2688	-732.1253	-732.6795	-727.8421
σ_{SB}^2	91.2676	200.7861	204.7821	204.8242	203.7719
$\rho_{GL \times CD}$	0.0391	0.0292	0.0283	0.0800	0.0368
$\rho_{GL \times SB}$	0.1280	0.1063	0.1351	0.1355	0.1341
$\rho_{CD \times SB}$	0.0098	-0.7370	-0.7218	-0.7223	-0.7193

Table 4.14 shows the results for additive genetic dispersion parameters. The direct genetic variance of CD, σ_{CD}^2 was greater than the other variances. The genetic covariance among CD and SB was much greater than the others. Also, it is possible to see that genetic variance and covariance for the MTMM changed drastically when compared to all the SEMs, however it is important to emphasize that the variance and correlation for all SEM are associated to the direct relation. For the covariance between CD and SB, the transition from MTMM to SEM changed the direction of the association, from positive to negative. This same changed occurred for the same covariance presented in Table 4.13. Regarding genetic correlations GL and CD as well as GL and SB are slightly correlated and CD and SB are highly correlated.

Table 4.14 – Genetic covariance and correlations estimates for GL, CD and SB for the MTMM, second pSEM, third pSEM and mixed pSEM

Parameter	MTMM	ISEM	Second pSEM	Third pSEM	Mixed pSEM
σ_{GL}^2	4.1179	3.4080	3.4119	3.4119	3.5606
$\sigma_{GL \times CD}$	26.0832	9.6948	9.6055	9.4032	31.3534
$\sigma_{GL \times SB}$	-10.6111	1.7821	2.5287	2.6299	1.3535
σ_{CD}^2	5,779.2041	2,818.1863	2,818.0229	2,817.7107	2,895.3239
$\sigma_{CD \times SB}$	2,638.9852	-110.0507	-105.4473	-105.3589	-92.3975
σ_{SB}^2	1,589.8956	414.7493	411.4561	411.7004	409.8699
$\rho_{GL \times CD}$	0.1691	0.0989	0.0980	0.0959	0.3088
$\rho_{GL \times SB}$	-0.1311	0.0474	0.0675	0.0700	0.00031
$\rho_{CD \times SB}$	0.8706	-0.1018	-0.0979	-0.0978	-0.0848

4.3.7 Discussion

Coefficients related to age changes presented the highest values. Those changes can be explained due to the absence of the intercept. Regarding genetic covariances and correlations, the inclusion of the causal relationships in the equations made them change drastically. The values of the covariances for MTMM values were at least 20% greater than the covariances obtained by any structural model, and in some cases involved changes of direction (positive to negative and vice versa). Specifically for the covariance between GL and SB, the inclusion of quadratic polynomials changed the estimates by around 30%. However, the estimates of covariance did not change much with the inclusion of more polynomial degrees (around 5%).

For the herd covariances and correlations, it can be seen that the inclusion of the causal relationship caused the herdsâ covariance and correlations to change drastically, but the behavior did not follow a pattern as for the genetic covariance. The only situations where MTMM presented a greater value than any of the SEMs was for the covariance between GL and CD, or around 33%, 38% and 39%, for the linear, second- and third-degree models, respectively. For the covariance between GL and SB and CD and SB, and the variance of SB, the pSEMs presented estimate values greater than the MTMM. Also, for the covariance between CD and SB, the inclusion of polynomial approach changed the direction from positive to negative and represented, in absolute value, more than 100 times the magnitude of the estimated value of the MTMM. For the ISEM, for the covariance between GL and SB, the inclusion of extra polynomial degree changed the genetic estimate by around 30%, but the estimated covariance between the pSEMs did not change with the inclusion of more polynomial degrees.

The MTMM seems to be a good approach to study the relation of CD and GL, but to study the relationship between GL and SB it is not a good approach. The ISEM seems not efficient to study both relationship. The pSEM is a good approach to study the relationship of GL and SB, but for the relationship between GL and CD it cannot be considered efficient, sine

it presented some problems in the tails. An alternative model was tested to recover the best of both models, ignoring the relation between GL and CD. Nonetheless, the model associated with CD still presented the same shape and estimation problems. Even though the results show two possibilities for estimation, the choice of using MTMM or any SEM depends on the aim of the study and also the biological structure

4.4 Conclusion

In this study it was possible to verify the difference between the multiple trait mixed model and structural equation models, using different polynomial degrees to fit the relationships between traits.

The MTMM presented better results for CD, but was weak in explaining SB. The results for the ISEM were the worst, showing poor fit for CD and SB. For all pSEMs, the results were good to explain SB but not very accurate for CD.

Since the causal relationships between traits can be found in the literature and all structural effects were considered significant, the SEM should be used to obtain a more faithful biological model. These assumptions drastically change the significance of the relationships between, leading to mistakes in animal selection. Some changes were not only related to the magnitude but also related to direction, which can generate huge mistakes.

Taking into account that the causal relations for those traits are perfectly explained and that SB can be considered the most important trait in this study, since perinatal death reduces profit and causes problems for the next lactation, we believe the polynomial SEM should be used, especially with quadratic effects

References

- BARRIER, A.C.; HASKELL, M.J.. Calving difficulty in dairy cows has a longer effect on saleable milk yield than on estimated milk production. **Journal of Dairy Science**, Champaign, v. 94, p. 1804-1812, 2011.
- BARRIER, A.C.; HASKELL, M.J. Calving difficulty in dairy cows has a longer effect on saleable milk yield than on estimated milk production. **Journal of dairy science**, Champaign, v. 94, P. 1804-1812, 2012.
- BERNDT, E.K.; HALL, B.H.; HALL, R.E.. Estimation and Inference in Nonlinear Structural Models. **Annals of Economic and Social Measurement**, Berkley, v. 3, p. 103-116, 1974.
- BLAIR, R.H; KLIEBENSTEIN, D.J; CHURCHILL, G.A. What Can Causal Networks Tell Us about Metabolic Pathways? **PLoS computational biology**, San Francisco, v. 8, e1002458, 2012.

CAMPOS, G. de los; GIANOLA, D.; BOETTCHER, P; MORONI, P. A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. **Journal of Animal Science**, Champaign, v. 84, p. 2934-2941, 2006.

CAMPOS, G., de los; GIANOLA, D.; HERINGSTAD, B. A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows. **Journal of Dairy Science**, Champaign, v. 89, p. 4445-4455, 2006.

CERQUEIRA, P.H.R.; VALENTE, B.; ROSA, G.J.M; LEANDRO, R.A. Second degree polynomial structural equation modeling using animal model: A simulation study In: INTERNATIONAL BIOMETRIC CONFERENCE, 27., 2014, Florence, **Abstracts...** Florence: IBS, 2014.

CHAIBUB NETO, E.. **Causal Inference Methods in Statistical Genetics**. Madison: University of Wisconsin, 2010. 140p.

CHAIBUB NETO, E.; KELLER, M.P.; Keller, ATTIE, A.D.; YANDELL, B.S.. 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**, Cleveland, v. 4, p. 320-339, 2010.

DEMATAWEWA, C.M.B.; BERGER, P. J.. Effect of dystocia on yield, fertility and cow losses and an economic evaluation of dystocia scores for Holsteins. **Journal of Dairy Science**, Champaign, v. 80, p.754-761, 1997.

DIJKTRA, T.K.; HOENSELER, J.. Linear indices in nonlinear structural equation models: best fitting proper indices and other composites. **Quality Quantity**, Dordrecht, v. 45, p. 1505-1518, 2011.

FOUZ, R.; GANDOY, F.; SANJUÁN, L.; YUS, E; DIÉGUEZ, F. J.. The use of crossbreeding with beef bulls in dairy herds: effects on calving difficulty and gestation length. **Animal : an International Journal of Animal Bioscience**, Cambridge, v. 7, p. 211-215, 2013.

GIANOLA, D.; SORENSEN, D.. Quantitative Genetic Models for Describing Simultaneous and Recursive Relationships Between Phenotypes. **Genetics**, Baltimore, v. 167, p. 1407-1424, 2004.

GILMOUR, A.R.; THOMPSON, R.; CULLIS, B.R. Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models. **Biometrics**, Arlington, v. 51, n. 4, p. 1440-1450, 1995.

GONZÁLEZ-RODRÍGUEZ, A.; MOURESAN, E.F.; J. ALTARRIBA, C.M.; VARONA, L. Non-linear recursive models for growth traits in the Pirenaica beef cattle breed. **Animal : an international journal of animal bioscience**, Cambridge, v. 8, p. 904-9011, 2014.

HARVILLE, D.A. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. **Journal of the American Statistical Association**, Alexandria, v. 72, n. 358, p. 320-338, Jun 1977.

HINRICHS, D.; BENNEWITZ, J.; WELLMANN, R.; THALLER G. Estimation of ancestral inbreeding effects on stillbirth, calving ease and birthweight in German Holstein dairy cattle. **Journal of animal breeding and genetics**, Hamburg, v. 132, p. 59-67, 2015.

JEFFREY, R.H.; WEISS, B.A.; HSU, J-C. A Comparison of Methods for Estimating Quadratic Effects in Nonlinear Structural Equation Models. **Psychological methods**, Washington, DC, v. 17, p. 193-214, 2012.

JOHANSON, J.M.; BERGER, P.J.; TSURUTA, S.; MISZTAL, I.. A Bayesian threshold-linear model evaluation of perinatal mortality, dystocia, birth weight, and gestation length in a Holstein herd. **Journal of dairy science**, Champaign, v. 94, 450-560, 2011.

KÖNIG, S.; WU, X. L.; GIANOLA, D.; HERINGSTAD, B.; SIMIANER, H.. Exploration of relationships between claw disorders and milk yield in Holstein cows via recursive linear and threshold models. **Journal of Dairy Science**, Champaign, v. 91, p. 395-406, 2008.

LEE, S-Y; ZHU, H-T. Statistical analysis of nonlinear structural equation models with continuous and polytomous data. **British Journal of Mathematical and Statistical Psychology**, London, v. 53 p. 209-232, 2000.

LEE, S-Y; TANG, N-S. Bayesian analysis of nonlinear structural equation models with nonignorable missing data. **Psychometrika**, Research Triangle Park, v.71, p.541-564, 2006.

LIU, B.; FUENTE, A. de la; HOESCHELE, I.. Gene Network Inference via Structural Equation Modeling in Genetical Genomics Experiments. **Genetics**, Baltimore, v. 178, p. 1763-1776, 2008.

- MADSEN, P.; Su, G.; LAUBOURIAU, R.; CHRISTENSEN, O. F.. DMU - A package for analyzing multivariate mixed models. Page 732 in **Proc. 9th World Congress on Genetics Applied to Livestock Production (WCGALP)**, Leipzig, Germany. Gesellschaft für Tierzuchtwissenschaften e. V., Neustadt, Germany, 2010.
- MATURANA, E.L. de; WU, XIAO-LIN; GIANOLA, D.; WEIGEL, K.A.; ROSA, G.J.M. Exploring Biological Relationships Between Calving Traits in Primiparous Cattle with a Bayesian Recursive Model. **Genetics**, Baltimore, v. 181, p. 277-287, 2008.
- MATURANA, E.L. de; LEGARRA, A.; VARONA, L.; UGARTE, E.. Analysis of fertility and dystocia in Holsteins using recursive models to handle censored and categorical data. **Journal of Dairy Science**, Champaign, v. 90, p. 2012-2024, 2007.
- MATURANA, E.L. de; UGARTE, E.; GONZÁLEZ-RECIO, O.G.. Impact of calving ease on functional longevity and herd amortization costs in Basque Holsteins using survival analysis. **Journal of Dairy Science**, Champaign, v. 90, 4451-4457, 2007.
- MEE, J.F.; BERRY, D.P.; CROMIE, A.R. Risk factors for calving assistance and dystocia in pasture-based Holstein-Friesian heifers and cows in Ireland. **Veterinary journal**, London, v. 187, p. 189-194, 2011.
- MEIJERING, A.. Dystocia and stillbirth in cattle: a review of causes, relations and implications. **Livestock Production Science**, Amsterdam, v. 11, p. 143-177, 1984.
- MURRAY, C.F.; VEIRA, D.M.; NADALIN, A.L.; HAINES, D.M.; JACKSON, M.L.; PEARL, D.L.; LESLIE, K.E. The effect of dystocia on physiological and behavioral characteristics related to vitality and passive transfer of immunoglobulins in newborn Holstein calves. **Canadian Journal of Veterinary Research**, Ottawa, v. 79, p. 109-119, 2015.
- PEARL, J. **Causality Models, Reasoning and Inference**. 2 ed. Cambridge, RU: Cambridge University, 2009. 484 p.
- PHILIPSSON, J.; STEINBOCK, L... Definition of calving traits - results from Swedish research. **Interbull Bulletin**, Berlin, v.30, p.71-74, 2003.
- PINHEIRO, J.C.; BATES, D.M.. **Mixed-Effects Models in S and S-PLAS**. New York: Springer, 2000. 528 p.

PURFIELD, D.C.; BRADLEY , D. G.; KEARNEY, J. F.; BERRY, D. P. Genome-wide association study for calving traits in Holstein-Friesian dairy cattle. **Animal : an International Journal of Animal Bioscience**, Cambridge, v. 8, p. 224-235, 2014.

ROSA, G.J.M.; VALENTE, B.D.; de lo CAMPOS, G.; WU, X.L.; GIANOLA, D.; SILVA, M.A.. Inferring causal phenotype networks using structural equation models. **Genetics, Selection, Evolution**, London, v. 43, p. 1046-1057, 2011.

STEPHAN, K.E.; KASPER, L.; HARRISON, LEE M.; DAUNIZEU, J; OUDEN, H.E.M. den; BREAKSPEAR, M.; FRISTON, K.J.. Nonlinear Dynamic Causal Models for fMRI. **NeuroImage**, Orlando, v. 42, p. 649-662, 2008.

STEINBOCK, L.; NÄSHOLM, A.; BERGLUND, B.; JOHASSON, K.; PHILIPSSON, J.. Genetic effects on stillbirth and calving difficulty in Swedish Holsteins at first and second calving. **Journal of dairy science**, Champaign, v. 86, p. 2228-2235, 2003.

TURNER, M.E.; MONROE, R.J.; LUCAS JR., H.L. Generalized Asymptotic Regression and Non-Linear Path Analysis. **Biometrics**, Alexandria, v. 17, p. 120-143, 1961.

UEMATSU, M. SASAKI, Y.; KITAHARA, G.; SAMEISHIMA, H.; OSAWA T. Risk factors for stillbirth and dystocia in Japanese Black cattle. **Veterinary journal**, London, v. 198, p. 212-216, 2013

VALENTE, B.D.; ROSA, G.J.M.; de los CAMPOS, G. GIANOLA, D.; SILVA, M.A. Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models. **Genetics**, Baltimore, v. 185, p. 633-644, 2010.

VALENTE, B.D.; ROSA, G.J.M.; GIANOLA, D., WU, Xiao-Lin; WEIGEL, K. Is structural equation modeling advantageous for the genetic improvement of multiple traits?. **Genetics**, Baltimore, v. 194, p. 561-572, 2013.

VALENTE, B.D.; ROSA, G.J.M.; SILVA, M.A.; TEIXEIRA, R.B.; TORRES, R.A.. Searching for phenotypic causal networks involving complex traits: an application to European quails. **Genetics, Selection, Evolution**, London, v. 43, p. 37-48, 2011.

VARONA, L.; SORENSEN, D.. Joint Analysis of Binomial And Continuous Traits with a Recursive Model: A Case Study Using Mortality and Litter Size of Pigs. **Genetics**, Baltimore, v. 196, p. 643-651, 2014.

XIONG, M.; LI, Jun; FANG, X. Identification of Genetic Networks. **Genetics**, Baltimore, v. 166, p. 1037-1052, 2004.

WRIGHT, S. Systems of mating. i. the biometric relations between parents and offspring. **Genetics**, Baltimore, v. 6, p. 111-123, 1921.

WRIGHT, S. An analysis of variability in number of digits in an inbred strain of guinea pigs. **Genetics**, Baltimore, v. 19, p.506-536, 1934.

WU X-L; HERINGSTAD, B.; GIANOLA, D.. Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. **Journal of animal breeding and genetics**, Berlin, v. 27, p. 3-15, 2010.

5 CONCLUSION

In this work, we proposed and studied polynomials structural equation models applied to quantitative genetics in order to fit the non linear relationships among traits, since in literature usually this models are developed using linear structural equation model. Two studies, a simulation and an application to Holstein dairy cows, were developed to verify the advantages of using the polynomial approach.

In the first study we developed a simulation study using a recursive causal structure involving 3 traits and 1,800 subjects. We use a Bayesian approach to recovering the effects. As a result we obtained that the polynomial SEM approach shows to be a better solution even when those relationships are in fact linear, in the other hand when the linear SEM was used to fit a non linear relationships the results shows to be a terrible approach, leading to a huge mistakes.

In the second study we analyzed a data set related to primiparous dairy cows of Holstein breed. In this study we compare the multiple trait mixed model and structural equation models using a linear, quadratic and cubic polynomials. We assume a causal relation between gestation length, calving difficulty and perinatal dead. The results shows a huge difference when assuming multiple trait mixed model against the polynomial structural equations models. The polynomial SEM was more accurate to fit the perinatal dead.

In this sense we can conclude that the proposed model is an alternative to solve problems when the non linear relationships between traits can be observed. Even more we can conclude that using polynomials there is no loss of information.

5.1 Prospective Works

During the development of this thesis some ideas for future works emerged. In this sense this section will present some prospective works.

Realize a more intensive study related to the interpretation of the random effects when the polynomial structural equation models is used.

In this thesis we focused in recovering the effects, specially the causal and genetics, assuming a known causal structure. For this reason the development of a methodology to recovering those causal relationships when the polynomials are used is necessary.

Even though the polynomial approach present better results for some situations they might not be accurate in describing the true causal function. The use a non linear functions or splines to fit the non-linear causal relationships can be an alternative for such situations.

In the developed studies the assumption of normality was made for all situations. However, in some situations this assumptions might be violated, and for this reason a study using generalized linear mixed models can be developed to improve the analysis.