Causal not Confounded Gene Networks:
Inferring Acyclic and Non-acyclic Gene Bayesian Networks in mRNA Expression Studies using Recursive V-Structures, Genetic Variation, and Orthogonal Causal Anchor Structural Equation Models

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biomathematics

by

Jason Erik Aten

2008
The dissertation of Jason Erik Aten is approved.

Aldons J. Lusis

Janet S. Sinsheimer

Elliot M. Landaw

Kenneth L. Lange, Committee Co-chair

Steve Horvath, Committee Co-chair

University of California, Los Angeles
2008
# Table of Contents

1 Introduction: Bilayer Verification and Recursive V-structures with Verification (RVV) for recovering gene-genetic networks . . 1

2 Background and prior theory .............................................. 6
   2.1 Motivation ................................................................. 6
   2.2 Bayesian networks theory ............................................. 7
       2.2.1 The property of d-separation at colliders .................... 10
       2.2.2 Placing genetic tests within the Bayesian network framework 11
   2.3 Learning network structure from observed conditional independencies .................................................. 14
   2.4 Prior work: The PC Algorithm ........................................ 14
   2.5 Prior work: the Local Causal Discovery (LCD) Algorithm . . . . 16

3 New theory ................................................................. 19
   3.1 Novel algorithm: the RVL Algorithm for learning DAGs efficiently 19
   3.2 Novel theory: Bilayer verification and the RVV algorithm for verifying graphs in the presence of unobserved confounders . . . . . . 21
   3.3 Novel theory for computational acceleration: gaining efficiency via the Clique Partitioning Theorem ........................................ 28

4 RVV Application: Methods and details ................................. 32
   4.1 C3H/HeJ x C57BL/6J microarray data ................................ 32
   4.2 Parameters and the v-structure test ................................. 33
4.2.1 Mechanics of the v-structure test .......................... 34
4.3 Special handling for genotypes ............................... 36

5 RVV Results ......................................................... 37
5.1 Estimating α false positive rates for the v-structure test ....................... 37
5.2 Learning an aortic lesion network ................................ 46
5.3 Further utilizing networks: assigning functional roles to genes ............ 47

6 Introduction to Orthogonal Causal Anchors and Local-structure
Edge Orienting (LEO) Scores for Robust Causality Testing and
Non-acyclic Gene Network Recovery .................................. 51

7 Relationship of multi-anchor and OCA methods to prior work ................. 58

8 LEO Methods .......................................................... 61
8.1 SEM and the definition of model p-values, P(model) ......................... 61
8.2 Network Edge Orienting (NEO) ..................................... 65
8.3 Step 1: Align traits (gene expression traits and clinical traits) and
SNPs across subjects .................................................. 65
8.4 Step 2: SNP selection and assignment to nodes ............................. 66
8.5 Step 2(a): QTL analysis or user specification of loci of interest ............ 66
8.6 Step 2(b): Automated unsupervised SNP selection methods ............... 66
8.7 Step 3: Aggregate SNPs into local structural equation models
and compute Local-structure Edge Orienting scores to assess the
strength of the evidence for each edge direction .......................... 68
8.8  Step 4: For each edge with high LEO score, evaluate the fit of the underlying local SEM models .......................... 71
8.9  Step 5: Robustness analysis with respect to significance thresholds and SNP selection ................................. 72
8.10 Step 6: Repeat analysis for next A-B trait-trait edge and apply edge score thresholds to orient the network .................. 72

9  Monte Carlo methods to evaluate power and specificity of $A \rightarrow B$
edge orienting .............................................................. 74
9.1  Restricted heritability as a parameter .............................. 74
9.2  The performance evaluation model ................................ 75

10 LEO Results .................................................................... 78
10.1 Results: Edge Orienting Algorithms: Comparison for Sensitivity, Specificity, and Anchor Count Parameter Robustness ........... 78
10.1.1 Single anchor/CPA versus OCA method ....................... 78
10.1.2 False positive and power comparisons ....................... 82
10.1.3 Robustness to SNP count parameter and handling of multiple SNP traits .................................................... 91
10.1.4 Conclusions about algorithm performance ................... 102
10.2 Automated SNP selection methods: comparing Greedy, Forward, and Both-Forward-and-Greedy ................................. 103
10.3 Speciality scores for diagnosis of confounding and acceleration of network edge orienting ...................................... 105
10.3.1 BLV: Bilayer Verification Scores for detecting confounding 105
**List of Figures**

2.1 The three patterns of connection in Bayesian networks. 12

2.2 The Transmission/Disequilibrium Test checks for d-separation (conditional independence) after conditioning on Parent genotypes. 13

3.1 Illustrating bilayer verification, graphs A1, A2, B, C, D. 23

3.2 Illustrating bilayer verification, graphs E, F, G, H. 24

3.3 Illustrating bilayer verification, graphs I, J. 24

5.1 Truncated results of the RVV algorithm run on Aortic Lesions in the C3H x B6 data, male mice, Liver mRNA profiled. 47

5.2 Relationships between clinical covariates, genes and genotypes for metabolic syndrome. 48

5.3 The subgraph for assigning gene functional roles. 48

6.1 The logic of network edge orienting, and the advantage of orthogonal causal anchors. 56

6.2 Overview of Network Edge Orienting. 57

8.1 Illustrating the single anchor models used in the definition of the central LEO.NB score (equations 8.6 - 8.9) for each candidate directed edge $A \rightarrow B$. 62
8.2 Illustrating the orthogonal causal anchor (orthomarker) models used in the definition of the central LEO.NB score (equations 8.6 - 8.9) for each candidate directed edge $A \rightarrow B$. 63

9.1 Performance evaluation model used in Monte Carlo studies. 77

10.1 Statistical power advantage of the LEO.NB score from utilizing orthogonal causal anchors. ............................ 80

10.2 Statistical power advantage of the LEO.NB score from utilizing orthogonal causal anchors. ............................ 81

10.3 SNP selection robustness titration for the LEO.NB.OCA scores under a 100% confounded Null model. ............. 84

10.4 SNP selection robustness titration for the LEO.NB.CPA scores under a 100% confounded Null model. ............. 85

10.5 SNP selection robustness titration for the LEO.NB.ALL scores under a 100% confounded Null model. ............. 86

10.6 SNP selection robustness titration for the MAX.MAX scores under a 100% confounded Null model. ............. 87

10.7 The LEO.NB.OCA score is robust against the addition of extraneous noise SNPs. ................................. 93

10.8 SNP selection robustness titration for the LEO.NB.CPA scores for a 4 SNP trait. ................................. 94

10.9 SNP selection robustness titration for the LEO.NB.ALL scores for a 4 SNP trait. ................................. 95

10.10 SNP selection robustness titration for the MAX.MAX scores for a 4 SNP trait. ................................. 96
10.11 SNP selection robustness titration for the LEO.NB.OCA scores for a single SNP trait. ............................... 98
10.12 SNP selection robustness titration for the LEO.NB.CPA scores for a single SNP trait. ............................... 99
10.13 SNP Selection robustness titration for the LEO.NB.ALL scores for a single SNP trait. ............................... 100
10.14 SNP selection robustness titration of the MAX.MAX scores for a single SNP trait. ............................... 101
10.15 The impact of LOD peak shoulders on the choice of SNP aggregation methods. ............................... 107
10.16 Monte Carlo study of zero true signal and use of the BLV score for detecting false positives due to confounding. 108
10.17 Correspondence between LEO.NB and ZEO scores. 109
10.18 Orthomarker models reproduce known biology in the cholesterol biosynthesis pathway: Insig1 → Dhrc7 and Insig1 → Fdft1. ............................... 117

A.1 Box plots illustrating the decomposition of the ZEO score. 132

C.1 Covariance matrix notation used. ............................... 142
List of Tables

3.1 Analysis of the depicted graphs with respect to two specific independence properties. ........................................ 25

5.1 Simulation results addressing v-structure false positive rates in small sample sizes. ...................................... 42

5.2 V-structure false positive rate results for N = 25. ......... 43

5.3 V-structure false positive rate results for N = 50. ......... 43

5.4 V-structure false positive rate results for N = 50. ......... 43

5.5 V-structure false positive rate results for N = 100. ....... 43

5.6 V-structure false positive rate results for N = 100. ....... 44

5.7 Conditional false positive rate study, sample size of N = 25. ................................................................. 44

5.8 Conditional false positive rate study, sample size of N = 50. ................................................................. 44

5.9 I. Composite false positive probabilities for bilayer verification following equation (5.2). ................................. 45

5.10 II. Composite false positive probabilities for bilayer verification following equation (5.2). ............................ 45

5.11 III. Composite false positive probabilities for bilayer verification following equation (5.2). ............................ 45

5.12 IV. Composite false positive probabilities for bilayer verification following equation (5.2). ............................ 46

10.1 Orthomarker model (LEO.NB.OCA) recommended score thresholds: false positive and power evaluation. ........ 88

10.2 Common Pleiotropic Anchor (LEO.NB.CPA) recommended score thresholds: false positive and power evaluation. ... 89

10.3 Averaging single anchor models LEO.NB.ALL recommended score thresholds: false positive and power evaluation. ... 90

10.4 Best single anchor model 1 versus best single anchor model 2 MAX.MAX recommended score thresholds: false positive and power evaluation. .................. 90

10.5 Descriptions of the four simulation studies reported in Tables 10.1, 10.2, 10.3 and 10.4. .................. 91

10.6 Edge orienting report for the cholesterol biosynthesis pathway positive controls—models depicted in Figure 10.18. .. 112

10.7 Fourteen positive control genes, and nine novel genes that appear downstream of Insig1 using NEO analysis.. .... 113
ACKNOWLEDGMENTS

Thanks go to Fred Fox, the IGERT Bioinformatics Training grant and the Genomic Analysis Training grant committees for supporting my research. Thanks to my co-chair Steve who taught me much about the organization, implementation, and presentation of one’s work. Thanks to my co-chair Ken who taught me that if you persist on a problem long enough, you will find an angle of attack. Thanks to Jake for your support, encouragement, and games of Go. You are a true role model. Thanks to Janet for offering insight from your knowledge and experience in genetics. Thanks to Elliot for introducing me to diabetes research and for your sense of humor—it reminds me to keep all things in perspective.


Special thanks and intellectual recognition go to Judea Pearl and Eric Schadt, whose diligent work in this field has inspired my own thinking.
Vita

1969 Jason Erik Aten born in Des Moines, Iowa, USA.

1991 B.A. Bucknell University, Lewisburg, PA


1994 Research Tech, Dept. of Microbiology, University of Iowa. PCR-based site-directed mutagenesis studies of Lambda phage packaging.

1994 Research Tech, Dept. of Microbiology, University of Washington, Seattle. FIV as an HIV model; biosafety 3 protocols.


1998 B.S. Computer Science.
University of Washington, Seattle

1997 - 1998 Software Engineer, WRQ Inc., Seattle, WA.


2003  M.S. Biomathematics  
UCLA School of Medicine  
Los Angeles, California

2001 - 2004  National Science Foundation, IGERT Fellow in Bioinformatics.  
Grant DGE9987641.

Grant HG02536-04.

PUBLICATIONS


xiv


ABSTRACT OF THE DISSERTATION

Causal not Confounded Gene Networks: Inferring Acyclic and Non-acyclic Gene Bayesian Networks in mRNA Expression Studies using Recursive V-Structures, Genetic Variation, and Orthogonal Causal Anchor Structural Equation Models

by

Jason Erik Aten

Doctor of Philosophy in Biomathematics
University of California, Los Angeles, 2008
Professor Steve Horvath, Co-chair
Professor Kenneth L. Lange, Co-chair

To improve the recovery of gene-gene and marker-gene interaction networks from microarray and genetic data, we first propose a new procedure for learning Bayesian networks. This algorithm, termed Bilayer Verification, starts with a user-specified leaf node, and then searches upstream to locate portions of the biological interaction network that can be verified as un-confounded by hidden variables such as protein levels.

Estimates of the specificity of the algorithm are made through small sample simulation, and a illustrative network is learned from mouse microarray data that implicates particular liver genes in the Apoe null mouse model of diet-induced atherosclerosis.
We next extend these algorithms by exploring how multiple independent causal anchors that impact the same trait can be used to organize gene expression data into non-acyclic gene-trait causal networks. While earlier methods begin with sets of single pleiotropic QTL, we formulate a gene network recovery approach based on a synthesis of (1) Bilayer verification theory; (2) selecting orthogonal causal anchors (independent Quantitative Trait Loci (QTL) $M_A$ and $M_B$ that show asymmetric $M_A \rightarrow A \rightarrow B \leftarrow M_B$ impact on traits $A$ and $B$; abbreviated OCA); (3) Structural Equation Model comparison; and (4) forward-stepwise regression. Combining these, we introduce a family of Local-structure Edge Orienting (LEO) scoring algorithms that generate model-comparison metrics. LEO scores weigh the evidence for competing causal graphs using local models that isolate each $A \rightarrow B$ edge evaluation from its neighbors to prevent error propagation and relax the constraint of network acyclicity.

Our studies show that the OCA-based LEO scores have almost twice the detection power at comparable false positive rates compared to single QTL and common pleiotropy anchor models in the face of confounded association. Moreover if we match thresholds to obtain comparable power, the orthomarker methods obtain better false positive rates than competing methods.

We demonstrate the method by recovering multiple positive controls in the cholesterol biosynthesis pathway and implicating four novel genes as being downstream and hence co-regulated by the sterol regulatory pathway in mouse liver: $Tld1$, $Slc25a44$, $Slc23a1$, and $Qdpr$. 
CHAPTER 1

Introduction: Bilayer Verification and Recursive V-structures with Verification (RVV) for recovering gene-genetic networks

Complex diseases such as metabolic syndrome, atherosclerosis, obesity, and diabetes in which environmental and clinical covariates contribute significantly along with genetic susceptibility factors pose special challenges to biology and genetics. For one, modern Quantitative Trait Loci (QTL) mapping strategies in humans, mice and rats are limited in resolution simply due to the infrequency of crossover events. Thus despite technical advances in Single Nucleotide Polymorphism (SNP) typing such as the Illumina(GSR06) and ParAllele(WMK05) technologies that leave little doubt as to the exact cross over points at which recombination did occur, the fact remains that we often cannot fully resolve QTLs to within fewer than 100 candidate genes. This is a significant problem for the establishment of a functional genome, as well as discovering the specific roles of genes in complex diseases. Secondly, due to technical or ethical constraints, we are often unable to experimentally control clinical and environmental factors, and thus must rely upon observational and longitudinal studies alone. Fisher’s ideal of randomized experiments and ANOVA data analysis frequently cannot be ethically or practically realized in the context of complex disease.

Hence while we have the genomic parts list, for the vast majority of genes we
lack their complete functional role in the biological blueprint. For many issues of serious and pressing medical importance, we cannot undertake precise experimental studies. We therefore ask how much our analytical, algorithmic and probabilistic modeling can realize from observational data and from the semi-experimental genomic technologies that have emerged in recent years. These technologies include mRNA microarrays, antibody protein arrays, and the comparison of genomic DNA sequence across strains and species.

A potentially exciting avenue of attack is to utilize Bayesian network learning algorithms (Pea00; Nea04; Coo97; CGB95; SGS00a; Pea88) to help decode the correlated genes in microarray data and to integrate genetic marker data with microarray measures. However such learning algorithms, when applied to microarray data sets face a significant missing data problem which greatly confounds the recovery of accurate genetic-gene and gene-gene networks. This is because actual protein levels, the active entity of most genes, are typically unknown. These unknowns may actively confound the correct derivation of a gene-gene interaction networks based on microarray data.

In a synthesis of classical and new approaches, Schadt and coworkers (SLY05) brought the power of classical genetic Quantitative Trait Loci (QTL) analysis to bear on gene selection from modern microarray data, producing expression-Quantitative Trait Loci (eQTL). While successful at selecting genes for further experimental investigation, this work is limited to a comparison of a handful of three node models and does not learn gene-gene interactions.

As we further examined the consequences of using genetic data to help inform the recovery of larger gene-gene interaction networks from microarray data, we encountered the issue of confounding of such interactions by unknown variables, in particular protein levels. More generally, this analysis applies to hidden variables in any Bayesian network. In our intended application, the network describes the
genetic marker-gene, gene-gene, and gene-clinical trait influence network within a tissue.

We describe here a new verification process by which confounding influence can, in recognizable cases, be ruled out when the observations flow from a faithful probability distribution. Such edges in the Bayesian network are verified edges, and their discovery is based on two inter-connected layers of v-structures. Verified edges in a Bayesian network are associations which, modulo sampling variation and given a faithful process (see theory definitions and results), cannot be associations due only to hidden confounders.

Although common biological feedback mechanisms would initially suggest undirected or cyclic graphs, the restriction of learned networks to directed acyclic graphs (DAGs) is useful in that DAGs suggest priorities in terms of causal order and more readily suggest intervention points that may be useful in gene, siRNA, or drug therapy. In subsequent chapters we relax this constraint and explore how to utilize multiple causal anchors effectively without imposing acyclicity.

The goal of this work is to learn a model of the underlying genetic network sufficient to predict the best intervention point of an siRNA (or other specific mRNA targeting) based intervention. The criteria for evaluating impact is that the targeted gene would specifically alter the clinical trait of interest, while minimally impacting the rest of the network.

What distinguishes the Bilayer verification based learning algorithms uniquely is the ability efficiently handle large-scale microarray sized data sets, to learn directional Bayesian networks even in the presence of confounders, and to learn structure consistent with any DAG-consistent causal theory. We offer the ability to readily impose prior source/sink structure on the network, and the ability to integrate genetic marker, experimental and observational data into the learning algorithm.
We now preview the next four chapters. In chapter 2, we review the essential terminology, definitions and theory on learning Bayesian networks and discuss influential prior work that illuminates our contribution. In chapter 3, we derive a process of learning Bayesian networks by recursively finding two levels of v-structures in the data, a procedure we term bilayer verification. In chapter 4 we extend this theory and give details of the methods used in the application of the algorithms, and further describe the microarray data which is analyzed. We report the results of simulation studies as to the false-positive rate for v-structure recovery in chapter 5, and sketch the output of the algorithm when run on aortic lesion and microarray data. In concluding we offer a brief look at a further application in terms of functional role assignment for genes in the process of mediating the effect of HDL cholesterol on atherosclerosis.

In the subsequent chapters, starting with the introduction and overview in chapter 6, we continue our studies with several practical methods to evaluate local $A \rightarrow B$ candidate network edges when there are multiple competing causal anchors (SNPs) that may be affecting the network. These studies introduce the utility of orthogonal causal anchors (OCA), which are connected, exogenous and independent upstream variables which provide additional insurance that a causal edge is correctly oriented. We study several aggregation methods and algorithms for weighing the evidence from multiple anchors and multiple orthogonal causal anchors. The Network Edge Orienting (NEO) methods and the Local-structure Edge Orienting (LEO) scores on which they are based allow us to learn potential non-acyclic networks in a principled fashion. We conclude with studies that document with Monte Carlo studies the power and false positive advantages of these approaches. As an application, we reproduce the known network biology of the cholesterol biosynthesis pathway in the liver. Several novel genes are also implicated as being co-regulated with cholesterol biosynthesis genes as a result.
of these studies.
CHAPTER 2

Background and prior theory

2.1 Motivation

Models are needed which can integrate prior knowledge of molecular biology, microarray mRNA level measurements, genetic markers, DNA sequence information, and eventually proteomic measurements of multiple proteins in multiple circumstances.

Often it is the discovery of a particular interaction (model structure) that is more important than the relative strength of an already understood interaction. Put another way, once we have evidence for a new network structure, parameter estimates often take a back seat to experimental verification and exploration.

Bayesian networks marry graphs with probability theory. Graphs have been adopted for modeling because they have several distinct advantages.

1. Graphs allow clear expression of prior knowledge and assumptions, in particular knowledge of conditional independence.

2. Graphs convey how to efficiently encode joint probability dependencies.

3. Graphs give explicit models for making inference from observations, and suggest efficient ways of doing so.

4. They provide natural means of integrating data from disparate sources.
5. Directed graphs provide a language in which to model prior causal relationships and to describe the discovery of new causal relations. They have been used to put discussions of causality on a mathematical basis (Pea00).

Although common biological feedback mechanisms might initially suggest undirected or cyclic graphs, the restriction of learned networks to directed acyclic graphs (DAGs) is useful in that DAGs suggest priorities in terms of causal order and more readily suggest intervention points that may be useful in gene or siRNA therapy.

2.2 Bayesian networks theory

Bayesian networks (Pea88; Pea00; Jen01) were invented originally with the intention of encoding data dependencies. They were conceived of as a general way of efficiently encoding a large dimensional multi-variate joint probability distribution (Pea85). More recently they have been used to specify causal models and subsequently put discussions of counterfactuals and causality on a mathematical basis (Pea00). Bayesian networks do not imply a commitment to Bayesian statistical methods. Here the term Bayesian refers to the fact that Bayes’ rule (also called Bayes’ conditioning) is used to update information when computing the probability of parents given their children in the graph.

**Definition 1 Bayesian network.** A Bayesian network is a pair \((G, \theta)\) consisting of a directed acyclic graph \(G\) in which each node is a random variable, together with a set of parameters \(\theta\) that encode for each node the conditional probability distribution of that node given its parents in the DAG.

The directed edges leading into a node \(X\) indicate the necessary and sufficient inputs to determine \(X\)’s probability distribution. Hence the missing edges in
the DAG typically encode the most information about the joint distribution: information about conditional independencies.

Information about independence in Bayesian networks is determined from the graph using two principles: the Markov property, and d-separation (the d prefix is for directed; in contrast to the simpler property of separation in an undirected graph).

**Definition 2 The Markov property.** Conditional upon knowing the state of its parents, a given node is independent of its non-descendants in the graph.

Equivalently, the Markov property is formulated

\[
P(V) = \prod_{v \in V} P(v | pa(v))
\]

where \( V \) is all variables (all nodes) in the graph, and \( pa(v) \) is the set of parents for a given node \( v \). If \( pa(v) \) is the empty set, then \( v \) is termed a root or exogenous node. Nodes with parents are also called endogenous nodes ("of internal origin", meaning causes are given within the graph) in the structural equation modeling literature.

In addition to the Markov property, the d-separation criteria allows one to read more complex induced conditional independence statements from a graph. It is also the basis for the learning of Bayesian networks. Suppose we specify three disjoint sets of nodes \( X, Y, \) and \( Z \), where \( Z \) might be the empty set, but \( X \) and \( Y \) must have cardinality at least one. If \( Z \) d-separates \( X \) from \( Y \), then the graph is claiming that in the probability distribution that it represents, \( X \) is conditionally independent of \( Y \) given \( Z \). This is written symbolically as \( X \perp\!\!\!\!\!\!\!\perp Y | Z \), indicating \( P(X, Y | Z) = P(X | Z)P(Y | Z) \). The opposite, conditional dependence, is noted by \( X \nparallel Y | Z \). This notation is due to Dawid (Daw79), while Pearl (Pea88) uses
this notation and further adopts 3-tuple relation notation from predicate calculus which is also commonly seen to denote conditional independence: \( I(X, Z, Y) \), meaning that \( X \) is independent of \( Y \) given \( Z \).

The definition of d-separation is a condition on all paths in the network whose legs can be taken with or against the arrows on the directed edges. While traversing each path, three types of connection patterns may be encountered when going from a node in \( X \) to a node in \( Y \). These are illustrated in Figure 2.1, and are called serial, diverging, and converging connections. They are also known as chain, fork, and collider connections.

**Definition 3 d-separation.** In a DAG, two disjoint sets of variables \( X \) and \( Y \) are d-separated by a third set \( D \), itself disjoint from \( X \) and \( Y \), if and only if along every path from an \( X \) node to a \( Y \) node we find that each node \( Z \) encountered satisfies one of the following three criteria:

- \( Z \) is at a converging connection, and neither \( Z \) nor any of its descendants (along directed paths) are in \( D \)
- \( Z \) is at a diverging connection, and \( Z \in D \)
- \( Z \) is at a serial connection, and \( Z \in D \)

For example, considering Figure 2.1, if \( Z \in D \), then in both the serial (a) and diverging (b) connections we would have d-separation between \( X \) and \( Y \). However at the collider (c) we only have d-separation if \( Z \notin D \), and this requirement of not being in \( D \) would apply to any of \( Z \)'s descendants as well. If the two parent nodes of a collider are not directly connected by an arrow, this structure is termed an unshielded collider, or v-structure.
2.2.1 The property of d-separation at colliders

To illustrate and provide intuition for the dependence inducing effect of conditioning on a causal child, consider a recessive lethal gene, with lethal allele little a and dominant allele big A. Suppose that mother and father are both heterozygous Aa, and that we can actually observe which allele was contributed from the mother and which from the father.

<table>
<thead>
<tr>
<th></th>
<th>Dad gives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>a</td>
</tr>
<tr>
<td>Mom gives</td>
<td>A</td>
<td>AA</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>Aa</td>
</tr>
</tbody>
</table>

When we don’t condition upon survival of the offspring, the big A and the little a alleles are passed along in independent 50/50 proportions from both parents.

However conditioning upon survival induces dependence in what were previously independent alleles. For if we know the offspring survived and that one parent gave a little a, then there is no doubt that the other parent gave a big A.

<table>
<thead>
<tr>
<th></th>
<th>Dad gave</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>Mom gave</td>
<td>A</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>33%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Conditioning on child survival induces dependence between the previously independent parent alleles for that child. The d-separation rule corresponds exactly to this situation: we don’t condition on any causal children or children’s descendants if we wish to d-separate the parents.
2.2.2 Placing genetic tests within the Bayesian network framework

The Transmission Disequilibrium Test (SME93) is a commonly used genetic test that illustrates the power of the Bayesian network representation for analysis and creation of new inference procedures. Consider Figure 2.2 which illustrates the test as originally utilized for diabetes candidate gene testing by Spielman et al (SME93).

The analyst specifies a Bayesian network with causal relationships hypothesized as draw in Figure 2.2a. If the starred edge (labelled *) does not exist, then the d-separation properties of the graph indicate that it is sufficient to condition on the parental genotypes to eliminate confounding by ethnic stratification. By confirming conditional independence between the Gm locus genotype and clinical diabetes, the authors of the test were utilizing the d-separation properties of the graph to infer that Figure 2.2b, which lacks the *-edge, was the true graph. The paradigm of testing edges in a learned or partially specified graph gives us a powerful methodology for discovering, discarding, and verifying candidate edges in a given or recovered network.

Definition 4 Faithful probability distribution. A probability distribution is faithful if each d-separation and d-dependence observed in the underlying causal graph is mirrored in the observed joint probability distribution.

A fundamental assumption in structure learning algorithms such as the PC algorithm (named after its authors, Peter and Clark (Pea00) p50, (SGS00a)), and the Local Causal Discovery (LCD) (Coo97) algorithm, is that the observed data follow from a faithful probability distribution. The interested reader is referred to (Pea88; Pea00). Other names for faithfulness (SGS00a) are stability (Pea00) and DAG-isomorphism (Pea88, p128). The primary difficulty being addressed by the faithfulness assumption is that the particular probability distribution being
Figure 2.1: The three patterns of connection in Bayesian networks.

observed may have independencies that are due to an unusual coincidence of parameters rather than being a result of the structure of the edges in the graph. Such peculiar (parameter-induced) independencies will fool learning algorithms, but are expected to be uncommon.

The last essential component of the theory of Bayesian networks for interpreting the causal discovery algorithms is Verma and Pearl’s (VP90) theorem on the equivalence of models under observation alone. Theorem 1.2.8 from ([Pea00], p19) tells us that using observational data alone, the v-structures are what allow us to distinguish directionality in the undirected Markov Random Field or skeleton graph.

Theorem [Verma and Pearl 1990(VP90)] (Observational Equivalence): Two DAGs are observationally equivalent if and only if they have the same skeletons and the same sets of v-structures, where a v-structure is two con-
a) testing edge *

b) d–separation implies that edge * does not exist

Figure 2.2: The Transmission/Disequilibrium Test checks for d-separation (conditional independence) after conditioning on Parent genotypes.
verging arrows whose tail variables are not themselves connected by an arrow.

2.3 Learning network structure from observed conditional independencies

To provide context for our algorithms we survey two influential causal discovery algorithms from which we draw ideas. The PC algorithm and the LCD algorithm can handle very different sized data sets, given the time complexity of their execution. The PC algorithm will never finish on 100 variables, while the LCD will return a little structural information about many billions of variables, without necessarily tying any two micro observations together. Our algorithms, RVL and RVV fall in the middle, in that they capture more of the structure of the network than is learned by LCD, albeit less than that learned by PC, and yet will operated efficiently on large data mining problems such as those posed by microarray data sources.

2.4 Prior work: The PC Algorithm

The PC algorithm of Spirtes, Glymour, and Scheines (SGS00a; KN04) learns, to the extent possible, the global Bayesian network structure from an observed data set by testing for conditional independence between various sets of nodes. Given the results of these tests, a network pattern is constructed so that the Markov property holds and so that the d-separation claims of the resulting graph mirror the conditional independencies found in the data. There are two phases to the PC algorithm. In the first phase, an undirected graph is learned. This is known as the skeleton of the Bayesian network. In the second phase, arrowheads are added to some of the edges, when they can be inferred. The output graph
may not be fully oriented, and is called a pattern. When the pattern contains undirected edges, these indicate that the data is consistent with models in which either orientation is possible (KN04).

PC Algorithm

1. Begin with the complete undirected graph on $V$. That is, create a fully connected skeleton model in which each node is connected to every other node by an undirected edge.

2. Track the cardinality of the conditioning sets in $k$. Initialize $k \leftarrow 0$. For each pair of nodes $X$ and $Y$, instantiate a set $\text{DSEp}(X,Y)$ and initialize it to the empty set $\{\phi\}$.

3. Chisel the skeleton: For every adjacent single nodes $X$ and $Y$, remove the arc between them iff for all subsets $Z$ of size $k$ containing nodes adjacent to $X$ but not containing $Y$, we find that $X \perp Y|Z$. When $k = 0$ this is simply a marginal independence test. Add the nodes in $Z$ to $\text{DSEp}(X,Y)$.

4. If any arcs were deleted, increment $k$ and repeat the previous step. This results in $X$ and $Y$ being connected iff $X \not\perp Y|\{\text{every subset of } V\setminus\{X,Y\}\}$.

5. Orientation phase: Collider induction. For each triple of 3 individual nodes $X - Y - Z$ such that $X$ and $Y$ are connected, $Y$ and $Z$ are connected, and yet $X$ and $Z$ are not connected, orient the chain as $X \rightarrow Y \leftarrow Z$ iff $Y \not\in \text{DSEp}(X,Z)$.

6. Orientation phase: Consequences of collider induction. Iterate through all undirected arcs $Y - Z$ in the graph and:

   (a) If $X \rightarrow Y$, $Y - Z$, and $X$ is not adjacent to $Z$, then orient $Y \rightarrow Z$.

   (b) If we have $Y - Z$ and there is already a directed path from $Y$ to $Z$, then orient $Y \rightarrow Z$. Otherwise a directed cycle would be introduced, violating the DAG property. Depth-first search along the directed edges will determine if a violating path would be created.

   (c) Continue at step 6a cycling through all undirected $Y - Z$ arcs until one pass fails to direct any arcs.

The major draw back to the PC algorithm is that its execution time is exponential in the number of variables observed. Two observations will allow us
to modify the PC algorithm for speed and applicability to genetic data. First we may be solely interested in determining the impact of various sources of genetic variation on one specific sink variable. Secondly, despite a large number of correlated variables, we may specifically interested only in those which we can determine fall into the pattern of direct causal influences of our sink variable. Furthermore among those direct causes, those that appear to be under genetic control are the most interesting. We therefore proceed to develop an algorithm based on these observations which significantly speeds up the causal analysis, at the expense of missing some parts of the complete network in return for focusing on determining most relevant candidate genes for further experimental investigation. Our formulation extends a similar local structure discovery algorithm by Cooper, which we briefly review next.

2.5 Prior work: the Local Causal Discovery (LCD) Algorithm

Cooper (Coo97) proposed a polynomial-time constraint-based local causal discovery algorithm for discovering relationships in large data sets, based on his observation that information about statistical independence and dependence relationships among a set of variables can be used to constrain (sometimes significantly) the possible causal relationships among a subset of the variables (Coo97; SBM00). As Silverstein and coworkers write:

A simple example of such a constraint is that if attributes A and B are independent, then it is clear that there is no causal relationship between them. It has been shown that, under some reasonable set of assumptions about the data (to be discussed later), a whole array of valid constraints can be derived on the causal relationships between the variables.
The LCD algorithm proceeds as follows. This formulation is by Silverstein and coworkers (SBM00), however it corrects an error in their conclusion ($x$ and $y$ were interchanged at line 7 in their table 1).

**LCD Algorithm**

1. Input: A set $V$ of variables. $w$ is a given variable known to have no causes among the other variables. A data set $D$. Tests for dependence, independence, and conditional dependence. ($D$, $I$, and $CI$ respectively).
2. Output: A list of causal relationships supported by the data
3. For all variables $x \neq w$
4. If $D(x, y)$
5. For all variables $y \notin \{x, w\}$
6. If $D(x, y)$ and $D(y, w)$ and $x \perp \perp w|y$
7. output `y causes x`

The structure being tested for at line 6 is for the chain or serial connection. Directionality of information flow in chains can be distinguished when one can make the assumption that there exists $w$, a variable with no prior causes in the network. Lacking such a variable $w$, the other means of inducing directionality is by locating the unshielded collider or v-structure. This forms the basis of the directional arrow assignment in all causal discovery algorithms without prior knowledge of exogenous (no parents in the graph) variables. Both chain location (termed CCC causality by Silverstein) and v-structure location (termed CCU causality by Silverstein) (SBM00) can be used in genetic network analysis, given that we can assume that the DNA changes inherited since conception precode the transcription of the genes. Although maternal environment affects may be correlated with the mothers DNA in ways that are not present in the child’s
DNA, we neglect such situations, not because they are not expected, but simply because we lack data to make such distinctions at this point.
CHAPTER 3

New theory

3.1 Novel algorithm: the RVL Algorithm for learning DAGs efficiently

Here we describe an algorithm which recursively learns directional arrows without first learning the complete undirected skeleton. We outline an algorithm for building the network from the bottom up, starting with the desired outcome or response variable as the sink node in the network.

Other prior work in which a sink node is chosen and structural equation models are learned includes the FBD and FTC algorithms (CGB95). The RVL algorithm below uses a more stringent link addition criteria than FBD and FTC do, and learns two parents at once, rather than the single parent addition at a time. Both FBD and FTC are susceptible to incorrectly inferring the directionality of the links, even using the strength of the multiple regression coefficient $R^2$ as FTC does to try and select more direct links. However the intuition behind the FTC algorithm is useful: using the strength of $R^2$ as a proxy for how direct an influence a variable has: “variables with high $R^2$ are less likely to have latent influences, so they are preferred as dependent variables” (CGB95). While not specifically reflected in the algorithm below, it would be a simple variant to prioritize successive sink nodes by drawing them from a priority queue $Q$ in terms of the maximal $R^2$ available.
Algorithm: Recursive V-Structure Location (RVL)

1. Choose a sink node $Z$. In our case a downstream clinical trait such as aortic lesions will be used to demonstrate.

2. Compute all pairwise marginal dependencies and note the degree of each node. That is, note how many pairwise marginal dependencies involve each node (variable). For instance count the correlations of $Z$ with any other variable that has a correlation coefficient p-value less than some threshold of significance.

3. Considering the set $U$ of neighbors of $Z$, pare down the set $U$ by eliminating indirect relationships as follows. For distinct nodes $u_i \in U$ and $u_j \in U$, if $u_i \perp \perp Z|u_j$, then $u_i$ is at best an indirect cause of $Z$, and so eliminate $u_i$ from $U$ for the purposes of v-structure detection.

4. Now detect upstream causes by applying the unshielded collider (v-structure) test. For each distinct pair $(X, Y)$, where both $X$ and $Y$ are drawn from the chiselled down set $U$, test for both $X \not\perp Y|Z$ and for $X \perp \perp Y$ marginally. If both tests hold true, then add both $X$ and $Y$ as parents of $Z$. If either $X$ or $Y$ has not yet been considered as a sink, add it to a queue $Q$ of sinks to check. Repeat step 4 until all distinct pairs $(X, Y)$ have been checked for v-structure.

5. Recursively select a new sink node $Z$ from the queue $Q$ and start again at step 3.

The algorithm terminates when all sinks have been considered in step 5. Alternatively it is easy to add additional stopping criteria that halt the algorithm after a particular depth of network has been learned.
RVL does not detect all causes, but we do proceed upstream as long as we can locate natural controls in the form of v-structures. And this algorithm is applicable to much larger data sets in which a outcome of interest, such as a clinical degree of disease, is the effect to which we wish to find direct causes. However our algorithm is applicable to much larger sets, and may be compared with the LCD algorithm and adaptations (Coo97; SBM00).

3.2 Novel theory: Bilayer verification and the RVV algorithm for verifying graphs in the presence of unobserved confounders

We now formulate a theorem which will allow us to both verify direct links and detect hidden confounding influences in the learned graph. Its name derives from the fact that by examining two layers of learned network at once we can verify the edges and check for confounding at the earlier learned edge. Prior work and inspiration includes Pearl’s definitions of Potential and Genuine Cause ((Pea00), pp54–56), Cooper’s analysis (Coo97) of the "+++" pattern, and the FTC/FBD algorithms(CGB95). Once understood, this theorem suggests a natural extension to the RVL algorithm if two layers of v-structure have been learned.

Theorem 1 Bilayer Verification.

Let five observed random variables $X_1, X_2, X_3, X_4, X_5$ be drawn from a faithful probability distribution, so that independence between variables is entirely due to the underlying influence structure and not a rare cancellation of parameter values and observed frequencies. Suppose that the following pattern of dependence, characteristic of two interlocked v-structures (collider patterns), is observed.
First v-structure:

\[ X_1 \not\perp X_2 \quad (3.1) \]
\[ X_1 \not\perp X_3 \quad (3.2) \]
\[ X_2 \perp X_3 \quad (3.3) \]
\[ X_2 \not\perp X_3 \mid X_1 \quad (3.4) \]

Second v-structure:

\[ X_2 \not\perp X_4 \quad (3.5) \]
\[ X_2 \not\perp X_5 \quad (3.6) \]
\[ X_4 \perp X_5 \quad (3.7) \]
\[ X_4 \not\perp X_5 \mid X_2 \quad (3.8) \]

In the absence of confounding, such dependencies would be evidence for Figure 3.1 graph C to be the underlying causal structure. \cite{SGS00, Pea00} However even in the presence of confounding, the edge ** from X2 to X1 can be either verified, ambiguous (possibly confounded), or dismissed as surely confounded according to the following rules for edge deletion and verification:

1. if (X1 \not\perp X4|X2) and (X4 \perp X1), then delete the ** edge X2 \to X1 as it was only due to confounding.

2. if (X1 \perp X4|X2) and (X4 \not\perp X1), then mark ** edge X2 \to X1 as verified (not confounded).

\textbf{Proof}: Since X2 \not\perp X1, there exists a marginally d-connected path between X2 and X1. Similarly, X3 \not\perp X1, shows that there exists a marginally d-connected path between X3 and X1. Because X2 \perp X3 and X2 \not\perp X3|X1 are
Figure 3.1: **Illustrating bilayer verification, graphs A1, A2, B, C, D.**

together characteristic only of a collider at X1, we know that all paths from X2 to X3 that account for the conditional dependence \(X2 \perp X3|X1\) must transit X1 via two distinct directional edges both heading into X1 from either side. Hence any final edge from X2 to X1 such as ** in the figures must be oriented with arrowhead into X1.

Having used X3 and the properties of d-separation at a v-structure to establish the directionality of the final edge ** in any paths from X2 to X1, it now suffices to analyze just the left side of the v-structure in terms of confounding. The right side paths (from X3 to X1) may have additional confounders, but they will not alter the conclusion about the directionality of the ** edge. An identical analysis holds for the second layer of v-structure utilized in equations (5)-(8), and hence confounding involving X5 and X2 is also safely ignored.
Figure 3.2: Illustrating bilayer verification, graphs E, F, G, H.

Figure 3.3: Illustrating bilayer verification, graphs I, J.
| graph | $X_1 \perp X_4 | X_2$ | $X_1 \perp X_4$ |
|-------|-----------------|-----------------|
| A1    | —               | —               |
| A2    | —               | yes             |
| B     | yes             | —               |
| C     | yes             | —               |
| D     | —               | yes             |
| E     | —               | —               |
| F     | yes             | —               |
| G     | —               | —               |
| H     | —               | —               |
| I     | —               | —               |
| J     | —               | —               |

Table 3.1: Analysis of the depicted graphs with respect to two specific independence properties.

Now consider the following table (Table 3.1) in conjunction with Figures 3.1 and 3.2 and 3.3, which exhaust the possibilities for confounding of the 3 variables $X_1$, $X_2$, and $X_4$. Circles represent latent, unobserved confounders. Squares represent observed variables. The table is marked yes if the independence statement holds, and — if it does not.

Graphs B, C, and F share the $X_1 \perp X_4 | X_2$ and $X_1 \not\perp X_4$ property, and in each of these three graphs the ** edge $X_2 \rightarrow X_1$ is direct and unconfounded. Hence if this pattern of independence is observed we can confirm or mark the ** edge as verified (not confounded). Graphs A2 and D share the $X_1 \not\perp X_4 | X_2$ and $X_1 \perp X_4$ pattern, and in both cases there is no true direct cause from $X_2 \rightarrow X_1$, instead just confounding exists. Graphs I and J share the same properties in the table as the other indeterminant cases A1, E, G, and H. Moreover these properties hold for all permutations of the presence and absence of the * and ** edges of graphs I and J; this conclusion follows from a simple enumeration (not shown) and checking of the four possibilities for the * and ** edge states (present/absent).

□
More broadly, note that $X_1 \not\perp X_4$ reveals that there is some marginally d-connected path between $X_1$ and $X_4$, and $X_1 \perp X_4|X_2$ reveals that $X_2$ intercepts (is present on) all such paths. More specifically, on all such paths $X_2$ is not the location of a collider, but rather is a chain or fork connection. A fork guarantees that the path departs $X_2$ and arrives directionally at $X_1$, and any chain must be oriented to flow from $X_2$ towards $X_1$ by the directionality argument for edge ** that began the proof. Hence we know that a directed path from $X_2$ to $X_1$ exists and can denote it by the ** edge.

Additionally, in the case when the grandparent ($X_4$ above) is a genetic marker and hence is reasonably modelled as exogenous (does not have any input arrows from observed or confounding unobserved variables), then we note that (referring again to Figures 3.1 - 3.3 and Table 3.1) only graphs $C$, $D$, and $E$ apply, and yet the exact same conclusion is reached. Indeed since this case is a subset of the above analysis, the conclusions of the theorem continue to hold for any parentless grandparent variables. However we gain from the additional knowledge of no arrows directed into $X_4$. Graph $C$, $D$, and $E$ now correspond in an invertible manner to unique patterns of independence. If we encounter the pattern associated with graph $C$ or $E$ we can immediately conclude that edge ** is verifiably present when $X_1 \not\perp X_4$, even though there is also confounding present in graph $E$. This follows from examining graphs $C$, $D$, and $E$, which exhaust the possibilities when learning second layer relationships in which $X_4$ has no inputs. Notice that once we conclude that $X_1 \not\perp X_4$, we can immediately rule out graph $D$. In the remaining graphs ($C$ and $E$) we are assured of the presence of the ** edge.

This reasoning results in the following corollary to the Bilayer verification theorem.

**Corollary 1** Parentless grandparent verifies parent-child relationship. 
If a variable $M$ is known to have no parents (as with a genetic marker), then
observing $M \not \models X_2$ and $M \models X_1|X_2$ at a stage in learning recursive $v$-structures at which we've learned $X_2 \rightarrow X_1$ means that we can immediately mark $X_2 \rightarrow X_1$ as a verified edge, because the **edge must be present even if the $X_2 \rightarrow X_1$ relationship is also confounded (as in graph $E$).

The same relationship analysis that holds for $X_4 \rightarrow X_2 \rightarrow X_1$ holds by symmetry for the other grandparent learned by RVL, such as the pictured $X_5 \rightarrow X_2 \rightarrow X_1$ relationship. The Recursive V-structure Verification (RVV) Algorithm is the result of incorporating these rules into RVL as each possible grandparent relationship is learned.

RVV has been implemented in C++ and a sample network learned is pictured in Figure 5.1. The pruning rules from the Bilayer verification theorem are implemented in RVV which also adds in root (no parent) connections (e.g. QTL information) in the manner of the LCD algorithm.

Bringing together RVL and the bilayer theorem yields a new means of learning network structure that suggests when edges can be verified. We term the resulting learning algorithm RVV, for Recursive V-structures with Verification.

**Algorithm: Recursive V-Structures with Verification (RVV)**

1. Choose a sink node $Z$.
2. Compute all pairwise marginal dependencies.
3. Considering the set $U$ of neighbors of $Z$, pare down the set $U$ by eliminating indirect relationships as follows. For distinct nodes $u_i \in U$ and $u_j \in U,$ if $u_i \perp Z|u_j$, then $u_i$ is at best an indirect cause of $Z$, and so eliminate $u_i$ from $U$ for the purposes of v-structure detection. If $u_i$ survives this test and $u_i$ is a genetic marker (or a known exogenous node) then add the marker immediately as a parent of $Z$.
4. Now detect upstream causes by applying the unshielded collider (v-structure) test. For each distinct pair $(X,Y)$, where both $X$ and $Y$ are drawn from the chiselled down set $U$, test for both $X \not \models Y|Z$ and $X \not \models Y$ marginally. If both tests hold true, then add both $X$ and $Y$ as parents of $Z$. If either $X$ or $Y$ has not yet been considered as a sink, add it to a queue $Q$ of sinks to check.
5. Verification: once there are two or more layers in the network, verification becomes possible. If \( X \to Z^1 \to Z^0 \), then test for two conditional independencies: if \( X \Perp Z^0 | Z^1 \) and \( X \not\Perp Z^0 \) then mark \( Z^1 \to Z^0 \) as confirmed. If \( X \not\Perp Z^0 | Z^1 \) and \( X \Perp Z^0 \) then delete edge \( Z^1 \to Z^0 \) as it was due to confounding.

6. Root connections: if exogenous (root) variables such as genetic markers \( M \) or predicted genotypes at specified loci are in the graph, for each \( Y \) learned as a parent of \( Z \), check for \( Y \not\Perp M \) (QTL existence) and \( M \Perp Z|Y \). If both these conditions hold, then add \( M \) as a parent of \( Y \).

7. Repeat step 4 until all distinct pairs \((X,Y)\) have been checked for v-structure.

8. Recursively select a new sink node \( Z \) from the queue \( Q \) and start again at step 3.

The root connections test in step 6 examines whether \( M \to Y \to Z \). This conditional independence test can be accomplished for the gene expression microarray data as follows, given that \( Y \) and \( Z \) are continuous gene expression values, while \( M \), being a genotype in our context, takes on either two (in the case of SNP genetic marker data) or three (for microsatellite genetic marker data) discrete values. We simply regress \( Z \) on \( Y \) and compute the residuals \( \{E(Z_i|Y_i)\} \) for \( i = 1..n \) after \( Y \)'s influence on \( Z \) has been taken into account.

3.3 Novel theory for computational acceleration: gaining efficiency via the Clique Partitioning Theorem

Supposing that we have made a first pass and computed and stored in a database the marginal (undirected) correlation structure between all pairs of variables. This is an efficient \( O(n^2) \) procedure for \( n \) variables. In considering how to apply these algorithms to large data sets, one is immediately faced with the fact that a large number of three variables conditional independence tests will, at a minimum be required. Rather than attempt to compute naively all \( \binom{n}{3} \) such triples, we focus only on those which would inform the location of v-structures.

The number of conditional independence tests to be preformed when exam-
ining large data sets in either global or local causal discovery can be reduced by utilizing the following theorem. This theorem shows that if we are given a list $U$ of the simple marginal correlations of node $X$, then we can partition that list $U$ into two subsets: those to check for the unshielded collider condition, and those to check for conditional independence.

Theorem 2 Clique partitioning theorem. For a node $X$, let $U$ be the set of all nodes in the graph which have marginal dependence on $X$, so that for all $\omega \in U$, we know that $\omega \not \perp X$. Now quite possibly some subset of $\{U \cup X\}$ may form a clique $C$ containing $X$. Here a clique is meant to indicate a maximal complete subgraph in which all nodes are connected to all other nodes in the clique. Our proposition here is that the nodes in the clique are the only ones that need to be considered in conditional independence testing, in that any node not in the clique cannot hope to $d$-separate any two nodes in the clique. Formally, let $Z$ be a node in $U$ that is not in any clique $C$ that is a subset of $\{U \cup X\}$ containing $X$. Given a faithful distribution, we claim that conditioning on $Z$ cannot $d$-separate $X$ and any other node in $C$.

Proof: Without loss of generality, let $Y$ be a node in $C$, and consider an undirected path from $X$ to $Y$ that does not encounter $Z$. First we establish the existence of such a path. Are there any such paths, given that we know $X \not \perp Y$. For sake of contradiction suppose that there is no path at all, and hence $X$ and $Y$ are $d$-separated given the empty set $\phi$, that is marginally independent. However for any faithful distribution $X \not \perp Y$, hence there must be some path between $X$ and $Y$ that is unblocked marginally, that is when not conditioning on any variables. Now is $Z$ on this path? If the path is direct $X \rightarrow Y$, or $X \leftarrow Y$, then clearly no, $Z$ is not on the path, and $Z$ cannot hope to $d$-separate $X$ and $Y$. So if the clique is of size two then we are done. If not, there must be some other node in
the clique besides X and Y. Call it S, and note that since Z is not in the clique, 
S \neq Z. Now there is a path X to S to Y since all three nodes are in the same 
clique. This path does not contain Z. If the junction at S is a fork or a chain, 
then the path is unblocked by conditioning on Z, since only conditioning on S 
would block the path. However there remains the situation when the junction 
at S is a collider. Then the path would be blocked upon conditioning on Z, if 
Z were a descendent of S. However since marginally X and Y are d-connected 
(when doing no conditioning) then the path through S would be separated, and 
hence this path could not be the one accounting for the d-dependence of X and 
Y. Hence either the junction at S is not a collider, or there exists another such 
path through C which does not contain a collider. Since Z is not in C, Z is not 
on this path, and consequently conditioning on Z cannot d-separate X and Y.

The result of this theorem is a large gain in efficiency when considering which 
nodes to include in the conditional independence checks of the skeleton construc-
tion phase of the PC algorithm. Consider that any node Z that is not in a 
clique with both X and Y for any marginally dependent pair \((X, Y)\) now need 
not be checked for conditional independence \(X \perp Y | Z\). And likewise any node 
Q found in a clique which contains X and Y can be ignored for the purposes of 
v-structure detection, because it is by definition at best a shielded collider, given 
that \(Q \not\perp X, X \not\perp Y, \text{ and } Y \not\perp Q\), which all must hold for membership in the 
clique.

Exact clique detection would impose a large computation burden. However 
the following heuristic provides the majority of the benefits of the previous anal-
ysis. Given the database of marginal independencies, and an enumeration of the 
degree of marginal connectivity for each node in the network, adopt the following 
heuristic.
Clique heuristic: For a given node $X$, consider the set $U$ of marginally dependent neighboring nodes. When testing the conditional independence of $X$ and some $Y$ in $U$, take the intersection of the neighbors of $Y$ and $U$. This intersection set are the only nodes that need to be considered when testing for conditional independencies between $X$ and $Y$ for edge deletion during skeleton chiseling.

Hence during the v-structure search phases of any DAG recovery algorithm, we can eliminate some nodes in the testing for v-structures by rendering them as indirect influences via chiseling. In our initial implementations, we find that application of the Clique heuristic can reduce by up to $\sim 50\%$ the computational work of skeleton chiseling. We do not explore this optimization further in this text, but now proceed to detail our method’s application to genetic data.
CHAPTER 4

RVV Application: Methods and details

In this methods and detailed procedures chapter we first describe the BxH data in some detail. We then describe the critical parameters and detailed mechanics for the v-structure test. We conclude with comments about how this test can be specialized to handle genotype data, in particular data from an F2 intercross.

4.1 C3H/HeJ x C57BL/6J microarray data

The C3H x B6 data set used in the network recovery trials was kindly provided by collaborators and has been studied in various aspects in earlier work(SLY05; WYS06; YSW06; DSD05b; SLJ06; DSL06; DSD05a). Additional background for the experimental methods that provided the data can be see in work on the network analysis of eQTL(GDS05) and in the implication by such approaches of the Insig2(CEZ05) and the 5-lipoxygenase genes in specific diseases(MAS05).

The data set for these analysis consists of the F2 generation from an intercross of a C57BL/6J (Apoe-null) inbred strain with a C3H/HeJ (Apoe-null) inbred strain. C57BL/6J is susceptible to a variety of atherosclerosis, diabetes, and heart disease related traits to which C3H is resistant. Particularly on the Apoe-null background, the F2 mice are expected to show significant spectrum of atherosclerosis and metabolic syndrome response to being fed a high-fat diet until 5 months of age. A total of 334 mice were bred, genotyped with classical microsatellite
and newer SNP markers at high density, and had a variety of clinical covariates measured, including fat mass determination via NMR, aorta sectioning and measurement of aortic lesions, insulin, glucose, free fatty-acid levels in the blood, and cholesterol fractions (unesterified, HDL, and LDL+VLDL)(WYS06). Significantly, the liver, brain, adipose tissue and skeletal muscle tissue from each F2 was assayed via Agilent Technologies(HDS03) spotted-oligo microarray, returning some 23K mRNA levels for each of the 4 tissues. Significant differences in the gene expression profiles between the male and female F2s were observed(YSW06).

Initial genotyping was carried out at 1392 loci across the 19 mouse autosomes and the X chromosome, with an average of 70 markers per chromosome. Additional ParAllele SNP genotyping was also done at 1822 SNPs(WMK05).

4.2 Parameters and the v-structure test

Finite samples and noisy data leave open the possibility of false positives to any data based learning algorithm. To evaluate the type I error rate or probability of a false positive verification, we first describe in detail the mechanics of the procedure for locating v-structures. Specific parameters are used to control the actual implementation of the v-structure test process and subsequently to compute the probability of this test being fooled by a sampling variation. Here we outline the relevant portion of the logic used in the test and describe the process and parameters in detail. In the subsequent results section, we will present simulation studies measuring the relative importance of these parameters in the small sample sizes often encountered (sample size le 100).

There are several parameters which control the central v-structure test.

- \( \eta_1 = \text{pvalue}_\text{for}_\text{correlation} \)
• $\eta_2 = \text{pvalue}_{\text{for_uncorrelated}}$

• $\eta_3 = \text{rho}_{\text{below_for_uncorrelated}}$

• $\eta_4 = \text{induced}_{\text{conditional_dependence_jump}}$

The meaning of these parameters is as follows. Referring to Figure 2.1c, the value of $\eta_1$ gives the threshold p-value above which an parent X to child Z edge or parent Y to child Z is rejected as not significant. The value of $\eta_2$ gives the p-value threshold for the association test below which the variables X and Y are considered not independent (hence resulting voiding the v-structure). The value of $\eta_3$ is a second requirement for concluding that the parents X and Y are independent; their correlation must fall below $\eta_3$ in absolute value. Lastly the characteristic of conditional dependence is one of the uniquely defining attributes of an unshielded collider or v-structure, and the parameter $\eta_4$ gives the minimum jump in correlation that we will consider as indicated conditional dependence between the two parents X and Y. A further minor parameter, $\text{min_number_cases_in_induced_dependence}$, is simply a minimum number of cases (typically set to 20) to allow the test to proceed.

In pseudo-code the first test for v-structure is implemented as follows, given three random variables $X$, $Y$, $Z$ to be checked for v-structure as labelled in Figure 2.1c.

4.2.1 Mechanics of the v-structure test

Compute the Pearson correlation $\rho_{xy}$ and its p-value $p_{xy}$.

Compute $\rho_{xz}$ and $p_{xz}$.

Compute $\rho_{yz}$ and $p_{yz}$.

Compute $\rho_{xy|z}$ and $p_{xy|z}$. 
if ( \( p_{xy} \geq \text{pvalue\_for\_uncorrelated} \))
    
    \[ \text{AND } |\rho_{xy}| \leq \rho_{\text{below\_for\_uncorrelated}} \]
    \[ \text{AND } p_{xz} \leq \text{pvalue\_for\_correlation} \]
    \[ \text{AND } p_{yz} \leq \text{pvalue\_for\_correlation} \]
    \[ \text{AND } p_{xy|z} < \text{pvalue\_for\_correlation} \]
    \[ \text{AND } |\rho_{xy|z}| > \text{induced\_conditional\_dependence\_jump} + |\rho_{xy}| \]
    \[ \text{AND cases} \geq \text{min\_number\_cases\_in\_induced\_dependence} \]

    then
    
    {
        
        1st layer v-structure detected, for bilayer verification, repeat same check for 2nd layer v-structure.

        If 2nd layer passes, then apply the bi-layer verification rules from the theorem. e.g.:

        Let \( p_1 = \text{p-value for grandparent-grandchild | parent test} \);
        Let \( p_2 = \text{p-value for grandparent-grandchild test} \);

        if \( (p_1 < \text{pvalue\_for\_correlation AND} p_2 > \text{pvalue\_for\_uncorrelated}) \)
            then delete parent-child edge as confounded
        else if \( (p_1 > \text{pvalue\_for\_uncorrelated AND} p_2 < \text{pvalue\_for\_correlation}) \)
            then conclude: parent-child is a verified edge.

    }

35
4.3 Special handling for genotypes

Genotypes in an F2 mouse inter-cross commonly constructed for QTL analysis (SMD03) are discrete valued variables which do not have parents in the graph, as their status is fixed from conception. Hence they are readily presumed to be upstream causal factors relative to all gene expression traits in the network. When the marker is at the beginning of a chain of marker $M$ followed by two traits $X$, and $Y$, as in $M \rightarrow X \rightarrow Y$, then we check for correlation with an additive, dominant, and recessive model between the marker $M$ and the conditioned trait $Y|X$. Since both $X$ and $Y$ are continuous, conditioning is easily accomplished by computing the residuals $Y - E(Y|X)$ after a linear regression of $Y$ on $X$. Since in the graph we are checking $M \rightarrow X \rightarrow Y$, if we remove the influence of the direct parent $X$ on the variation of $Y$, then we should see no further relationship between $M$ and the residuals $Y|X$. 
CHAPTER 5

RVV Results

In this chapter we present simulation and application results from our study of the RVV algorithm. We first derive a conservative formula for estimating the false positive rate for v-structure discovery. We then present results from small sample simulations which estimate the components of this formula. We provide tables suitable for incorporation into a software implementation of these tests. We next briefly illustrate the results of the RVV when run on the BxH aortic lesion data using liver mRNA microarray data. Lastly we discuss an extension that becomes possible once such a directed network has been learned. Specifically, we demonstrate how a directed network can be used to assign functional roles to genes. We present preliminary results suggesting that the gene *Iqf1* may play an role (when expressed in liver) in mediating between HDL cholesterol levels and aortic lesions.

5.1 Estimating $\alpha$ false positive rates for the v-structure test

Given the small sample sizes involved in our study and in most microarray experiments, it seemed most useful to examine the behavior of the v-structure test with simulation of small sample sizes. Here we report simulation results obtained during varying both sample size ($N=25$ and $N=50$ were tested) as well as each of the four $\eta$ parameters described above.
In the following discussion we introduce the notation \( \simL \) and \( \simL \) to discuss the results of statistical tests on finite sample data for dependence and independence. Specifically, \( A \simL B \) means that \( A \) and \( B \) have been found statistically dependent, despite the fact that this may be due to type I (false positive) error in the statistical dependence test. Similarly, \( A \simL B \) means that \( A \) and \( B \) look independent to our tests and our sample, even though in they may in reality (in the true underlying population distribution) exhibit dependence.

To facilitate the analysis of the test, we distinguish between three parts of the computation, namely \( V_1 \), \( V_2 \), and \( V_3 \) below. By simulating under the Gaussian distribution to evaluate the false positive rate of three distinct portions of the test, we combine analysis and simulation to estimate the false positive rates for the bilayer verification process under the assumption that our starting variables \( X_1 \ldots X_5 \) are identically and independently distributed Normal(0,1). Analytically, we can express the probability of type I error as follows. It may help to follow the notation by referring to the labelling of Figure 3.1, Graph C.
\[ P(\text{false positive verification}) = P(\text{verification}|X_1, X_2, X_3, X_4, X_5 \text{ are i.i.d. } N(0, 1)) = P(X_1 \not\perp \perp X_2, (5.1)\]
\[ X_1 \not\perp \perp X_3, \]
\[ X_2 \perp X_3, \]
\[ X_2 \not\perp \perp X_3|X_1, \]
\[ X_2 \not\perp \perp X_4, \]
\[ X_2 \not\perp \perp X_5, \]
\[ X_4 \perp X_5, \]
\[ X_4 \not\perp \perp X_5|X_2, \]
\[ X_4 \not\perp \perp X_1, \]
\[ X_4 \perp X_1|X_2 \]

Given that \(X_1 \ldots X_5\) are i.i.d. \(N(0, 1)\)

To manage notation and make plain the approximation we will use, let us define \(V_1\) as shorthand for the proposition of the first four statements above, namely that the first v-structure is found.

\[ V_1 \text{ denotes } (X_1 \not\perp \perp X_2, \]
\[ X_1 \not\perp \perp X_3, \]
\[ X_2 \perp X_3, \]
\[ X_2 \not\perp \perp X_3|X_1) \]
Likewise, define $V_2$ as an indicator for the proposition of the second v-structure is found.

$$V_2 \text{ denotes } (X_2 \not\perp X_4, \quad X_2 \not\perp X_5, \quad X_4 \perp X_5, \quad X_4 \not\perp X_5|X_2)$$

Next define $V_3$ as an indicator for the proposition that we observe the verification rules holding.

$$V_3 \text{ denotes } (X_1 \not\perp X_4, \quad X_1 \perp X_4|X_2)$$

Lastly let $I_0$ be an indicator for the proposition that $X_1 \ldots X_5$ are i.i.d. $N(0, 1)$.

Then we can write equation (5.1) in simplified notation to illustrate our method.

$$P(V_1, V_2, V_3 | I_0) = P(V_1 | I_0) \cdot P(V_2 | V_1, I_0) \cdot P(V_3 | V_1, V_2, I_0)$$

$$= P(V_1 | I_0) \cdot P(V_2 | I_0) \cdot P(V_3 | V_1, V_2, I_0)$$

$$\approx P(V_1 | I_0) \cdot P(V_2 | I_0) \cdot P(V_3 | (X_4 \not\perp X_2), (X_2 \not\perp X_1), I_0)$$

$$= \{P(V_1 | I_0)\}^2 \cdot P(V_3 | (X_4 \not\perp X_2), (X_2 \not\perp X_1), I_0) \quad (5.2)$$

The third line of (5.2) is a conservative approximation for generating p-values in the sense that the actual probability of false v-structures $V_1$ and $V_2$ is smaller than the probabilities of $(X_4 \not\perp X_2)$ and $(X_2 \not\perp X_1)$ which we substitute. The
substituted terms capture the most salient feature of the relationship between the two v-structures that is subsequently tested during application of the bilayer verification rules, and the probabilities of these two terms holding under $I_0$ can be addressed in reasonable time by simulation, whereas simulations looking for both $V_1$ and $V_2$ to hold are, as the results below demonstrate, such rare events that their occurrence cannot easily be addressed by simulation. Moreover we may actually be quite interested in the type I error rate in the presence of partially acknowledged dependence between the variables, in which case this approximation is most appropriate.

The fourth line of (5.2) follows as the two terms $P(V_1|I_0)$ and $P(V_2|I_0)$ are identical under $I_0$. The values under various parameter choices for the $V_1$ portion of the test are given in Tables 5.1 through 5.6 for small sample sizes. We also obtained results by simulation for the last term, $P(V_3|(X_4 \neq X_2), (X_2 \neq X_1), I_0)$, for shorthand labelled $\alpha_{\text{inlinex}}$, and these are rendered in Tables 5.7 and 5.8.

In detail, after appropriate transformation, our microarray data can be reasonably treated as Gaussian (SLY05). To determine $P(V_1|I_0)$ we therefore simulate three independent and identically distributed Normal(0,1) variables and check for v-structure under varying parameter values. The raw counts reported in Tables 5.1, 5.2, 5.3, 5.4, 5.5, and 5.6 are the number of independent simulations of size N to locate 10 false positive v-structures. Six significant figures are reported. The resulting estimates of the false positive rate $\alpha$ for single v-structure location $P(V_1|I_0)$ under various values of the control parameters are given. A rounding to nearest half-decade is given as is a subjective indication of whether the results cluster near each other. As might be expected, the values of $\eta_1$ and N dominate the determination of the false positive rate. We can see that although $\eta_2$, $\eta_3$, and $\eta_4$ appear to influence $\alpha$ to a small extent, their settings might be more reasonably chosen to maximize power as their impact on $\alpha$ appears fairly minor. Note
### Table 5.1: Simulation results addressing v-structure false positive rates in small sample sizes.

Here $N=25$ and $\eta_1 = 0.05$. $\text{raw}$ = count of number of independent simulations of size $N$ to reach 10 false positives; $\text{var}$ = variance of the count of simulations between each false positive; $\alpha$ = observed frequency of type I error or falsely predicted v-structure.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$\eta_3$</th>
<th>$\eta_4$</th>
<th>raw</th>
<th>var</th>
<th>$\alpha$ (10/\text{raw})</th>
<th>rounded $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>218862</td>
<td>4.15E+08</td>
<td>4.57E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>216353</td>
<td>3.34E+08</td>
<td>4.62E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>146164</td>
<td>8.26E+07</td>
<td>6.84E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>153541</td>
<td>2.48E+08</td>
<td>6.51E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>191842</td>
<td>3.04E+08</td>
<td>5.21E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>289193</td>
<td>6.85E+08</td>
<td>3.46E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>169346</td>
<td>1.18E+08</td>
<td>5.91E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>130867</td>
<td>6.27E+07</td>
<td>7.64E-05</td>
<td>1.00E-04</td>
</tr>
</tbody>
</table>

that the roles of $X$, $Y$, and $Z$ as labelled in Figure 2.1c are fixed in advance of the simulation and test results reported in Tables 5.1 through 5.6, so that there is no multiple testing correction needed as when $X$ and $Y$ and $Z$ are allowed to permute roles in the v-structure.

Joining the results from Tables 5.1 - 5.4 for $P(V_1|I_0)$ together the results from Tables 5.7 and 5.8 for $P(V_3|(X4 \perp \!\!\!\perp X2), (X2 \perp \!\!\!\perp X1), I_0)$ yields the composite final simulation results as presented in Tables 5.9 - 5.12 for sample sizes $N = 25$ and $N = 50$. At $N = 100$ the simulated events corresponding to Tables 5.7 and 5.8 are so rare as to be prohibitively expensive in terms of computation time. Nonetheless the final $P($verification$|I_0)$ results do demonstrate the effects of sample size and various parameters adjustments in terms of taking the estimated false positive rate from 1e-10 to 1e-17.

To summarize these results, we draw two primary conclusions. First, we have established small sample $\alpha$ rate tables (in particular Tables 5.9 - 5.12) for reference and incorporation into computer programs which utilize v-structure based DAG recovery and wish to report false positive rates for their algorithms and tests. Second, by taking the $\eta_1$ p-value threshold as the primary parameter of
Table 5.2: **V-structure false positive rate results for N = 25.** Increasing the stringency by taking $\eta_1 = 0.01$

<table>
<thead>
<tr>
<th>N</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$\eta_3$</th>
<th>$\eta_4$</th>
<th>raw</th>
<th>var</th>
<th>$\alpha$ (10/raw)</th>
<th>rounded $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>11286400</td>
<td>9.12E+11</td>
<td>8.86E-07</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>22599200</td>
<td>1.30E+13</td>
<td>4.42E-07</td>
<td>5.00E-07</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>15264300</td>
<td>1.02E+12</td>
<td>6.55E-07</td>
<td>5.00E-07</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>12657100</td>
<td>1.33E+12</td>
<td>7.90E-07</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>12356600</td>
<td>1.04E+12</td>
<td>8.09E-07</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>19434800</td>
<td>2.17E+12</td>
<td>5.15E-07</td>
<td>5.00E-07</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>13468000</td>
<td>2.26E+12</td>
<td>7.43E-07</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>9328720</td>
<td>1.69E+11</td>
<td>1.07E-06</td>
<td>1.00E-06</td>
</tr>
</tbody>
</table>

Table 5.3: **V-structure false positive rate results for N = 50; $\eta_1 = 0.05$**

<table>
<thead>
<tr>
<th>N</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$\eta_3$</th>
<th>$\eta_4$</th>
<th>raw</th>
<th>var</th>
<th>$\alpha$ (10/raw)</th>
<th>rounded $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>171428</td>
<td>1.41E+08</td>
<td>5.83E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>181414</td>
<td>2.77E+08</td>
<td>5.51E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>101932</td>
<td>5.57E+07</td>
<td>9.81E-05</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>211450</td>
<td>6.82E+08</td>
<td>4.73E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>3761080</td>
<td>7.41E+10</td>
<td>2.66E-06</td>
<td>5.00E-06</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>4360960</td>
<td>1.03E+11</td>
<td>2.28E-06</td>
<td>5.00E-06</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>3148920</td>
<td>4.48E+10</td>
<td>3.18E-06</td>
<td>5.00E-06</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>3663490</td>
<td>4.82E+10</td>
<td>2.73E-06</td>
<td>5.00E-06</td>
</tr>
</tbody>
</table>

Table 5.4: **V-structure false positive rate results for N = 50; $\eta_1 = 0.01$**

<table>
<thead>
<tr>
<th>N</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$\eta_3$</th>
<th>$\eta_4$</th>
<th>raw</th>
<th>var</th>
<th>$\alpha$ (10/raw)</th>
<th>rounded $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>21629000</td>
<td>3.23E+12</td>
<td>4.62E-07</td>
<td>5.00E-07</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>40282500</td>
<td>3.32E+12</td>
<td>2.48E-07</td>
<td>1.00E-07</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>10390000</td>
<td>8.95E+11</td>
<td>9.62E-07</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>23242900</td>
<td>3.33E+12</td>
<td>4.30E-07</td>
<td>5.00E-07</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>24225400</td>
<td>7.45E+12</td>
<td>4.13E-07</td>
<td>5.00E-07</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>33825200</td>
<td>1.03E+13</td>
<td>2.96E-07</td>
<td>5.00E-07</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>25620500</td>
<td>3.71E+12</td>
<td>3.90E-07</td>
<td>5.00E-07</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>34328300</td>
<td>9.71E+12</td>
<td>2.91E-07</td>
<td>5.00E-07</td>
</tr>
</tbody>
</table>

Table 5.5: **V-structure false positive rate results for N = 100; $\eta_1 = 0.05$, $\eta_4 = 0.1$**

<table>
<thead>
<tr>
<th>N</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$\eta_3$</th>
<th>$\eta_4$</th>
<th>raw</th>
<th>var</th>
<th>$\alpha$ (10/raw)</th>
<th>rounded $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>9120470</td>
<td>5.69E+11</td>
<td>1.10E-06</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>100</td>
<td>0.05</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>5905130</td>
<td>4.08E+11</td>
<td>1.69E-06</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>100</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>6215150</td>
<td>5.57E+11</td>
<td>1.61E-06</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>100</td>
<td>0.05</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>4355320</td>
<td>1.15E+11</td>
<td>2.30E-06</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>N</td>
<td>( \eta_1 )</td>
<td>( \eta_2 )</td>
<td>( \eta_3 )</td>
<td>( \eta_4 )</td>
<td>raw</td>
<td>var</td>
<td>( \alpha ) (10raw)</td>
<td>rounded ( \alpha )</td>
</tr>
<tr>
<td>-----</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>---------</td>
<td>---------</td>
<td>---------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>88889800</td>
<td>7.26E+13</td>
<td>1.12E-07</td>
<td>1.00E-07</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>836504000</td>
<td>3.55E+15</td>
<td>1.20E-08</td>
<td>1.00E-08</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>110320000</td>
<td>3.54E+13</td>
<td>9.06E-08</td>
<td>1.00E-07</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>1319030000</td>
<td>8.52E+15</td>
<td>7.58E-09</td>
<td>1.00E-08</td>
</tr>
</tbody>
</table>

Table 5.6: **V-structure false positive rate results for** \( N = 100; \eta_1 = 0.01, \eta_4 = 0.1 \)

<table>
<thead>
<tr>
<th>N</th>
<th>( \eta_1 )</th>
<th>( \eta_2 )</th>
<th>( \eta_3 )</th>
<th>( \alpha_{\text{inline}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.2</td>
<td>0.3</td>
<td>3.38E-02</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.2</td>
<td>0.2</td>
<td>4.85E-03</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>6.10E-02</td>
</tr>
<tr>
<td>25</td>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
<td>8.59E-03</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.2</td>
<td>0.3</td>
<td>9.68E-03</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>1.78E-03</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.1</td>
<td>0.3</td>
<td>1.71E-02</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
<td>0.1</td>
<td>0.2</td>
<td>1.65E-03</td>
</tr>
</tbody>
</table>

Table 5.7: **Conditional false positive rate study, sample size of** \( N = 25 \). Simulation under Normal(0,1) distribution until 20 conditional false positives were obtained. The field \( \alpha_{\text{inline}} \) gives the probability of falsely concluding that \( \gamma \perp \chi \mid \pi \) and also that \( \gamma \nmid \chi \), while conditioning upon the fact that we have already falsely concluded that \( \gamma \nmid \pi \) and that \( \pi \nmid \chi \).

<table>
<thead>
<tr>
<th>N</th>
<th>( \eta_1 )</th>
<th>( \eta_2 )</th>
<th>( \eta_3 )</th>
<th>( \alpha_{\text{inline}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.2</td>
<td>0.3</td>
<td>5.28E-03</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.2</td>
<td>0.2</td>
<td>8.55E-03</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>3.57E-02</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
<td>1.19E-02</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.2</td>
<td>0.3</td>
<td>4.32E-04</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>3.12E-04</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.1</td>
<td>0.3</td>
<td>4.32E-03</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.1</td>
<td>0.2</td>
<td>8.85E-04</td>
</tr>
</tbody>
</table>

Table 5.8: **Conditional false positive rate study, sample size of** \( N = 50 \). Simulation under Normal(0,1) distribution until 20 conditional false positives were obtained. The field \( \alpha_{\text{inline}} \) gives the probability of falsely concluding that \( \gamma \perp \chi \mid \pi \) and also that \( \gamma \nmid \chi \), while conditioning upon the fact that we have already falsely concluded that \( \gamma \nmid \pi \) and that \( \pi \nmid \chi \).
| $N$ | $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $P(\text{verification}|I_0)$ |
|-----|-----------|-----------|-----------|-----------|-----------------|
| 25  | 0.05      | 0.1       | 0.2       | 0.1       | 1.79E-11        |
| 25  | 0.05      | 0.2       | 0.2       | 0.1       | 1.04E-11        |
| 25  | 0.05      | 0.1       | 0.3       | 0.1       | 2.83E-10        |
| 25  | 0.05      | 0.2       | 0.3       | 0.1       | 1.44E-10        |
| 25  | 0.05      | 0.1       | 0.2       | 0.2       | 2.33E-11        |
| 25  | 0.05      | 0.2       | 0.2       | 0.2       | 5.80E-12        |
| 25  | 0.05      | 0.1       | 0.3       | 0.2       | 2.13E-10        |
| 25  | 0.05      | 0.2       | 0.3       | 0.2       | 1.98E-10        |

Table 5.9: I. Composite false positive probabilities for bilayer verification following equation (5.2). Sample size of $N = 25$ and $\eta_1 = 0.05$. 

| $N$ | $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $P(\text{verification}|I_0)$ |
|-----|-----------|-----------|-----------|-----------|-----------------|
| 25  | 0.01      | 0.1       | 0.2       | 0.1       | 1.30E-15        |
| 25  | 0.01      | 0.2       | 0.2       | 0.1       | 3.49E-16        |
| 25  | 0.01      | 0.1       | 0.3       | 0.1       | 7.33E-15        |
| 25  | 0.01      | 0.2       | 0.3       | 0.1       | 6.04E-15        |
| 25  | 0.01      | 0.1       | 0.2       | 0.2       | 1.08E-15        |
| 25  | 0.01      | 0.2       | 0.2       | 0.2       | 4.72E-16        |
| 25  | 0.01      | 0.1       | 0.3       | 0.2       | 9.42E-15        |
| 25  | 0.01      | 0.2       | 0.3       | 0.2       | 1.11E-14        |

Table 5.10: II. Composite false positive probabilities for bilayer verification following equation (5.2). Sample size of $N = 25$ and $\eta_1 = 0.01$. 

| $N$ | $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $P(\text{verification}|I_0)$ |
|-----|-----------|-----------|-----------|-----------|-----------------|
| 50  | 0.05      | 0.1       | 0.2       | 0.1       | 4.03E-11        |
| 50  | 0.05      | 0.2       | 0.2       | 0.1       | 2.60E-11        |
| 50  | 0.05      | 0.1       | 0.3       | 0.1       | 3.44E-10        |
| 50  | 0.05      | 0.2       | 0.3       | 0.1       | 1.18E-11        |
| 50  | 0.05      | 0.1       | 0.2       | 0.2       | 8.38E-14        |
| 50  | 0.05      | 0.2       | 0.2       | 0.2       | 4.46E-14        |
| 50  | 0.05      | 0.1       | 0.3       | 0.2       | 3.60E-13        |
| 50  | 0.05      | 0.2       | 0.3       | 0.2       | 3.93E-14        |

Table 5.11: III. Composite false positive probabilities for bilayer verification following equation (5.2). Sample size of $N = 50$ and $\eta_1 = 0.05$. 

45
| $N$ | $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $P(\text{verification}|I_0)$ |
|-----|---------|---------|---------|---------|-----------------|
| 50  | 0.01    | 0.1     | 0.2     | 0.1     | 1.89E-16        |
| 50  | 0.01    | 0.2     | 0.2     | 0.1     | 1.92E-17        |
| 50  | 0.01    | 0.1     | 0.3     | 0.1     | 4.00E-15        |
| 50  | 0.01    | 0.2     | 0.3     | 0.1     | 7.99E-17        |
| 50  | 0.01    | 0.1     | 0.2     | 0.2     | 1.51E-16        |
| 50  | 0.01    | 0.2     | 0.2     | 0.2     | 2.72E-17        |
| 50  | 0.01    | 0.1     | 0.3     | 0.2     | 6.59E-16        |
| 50  | 0.01    | 0.2     | 0.3     | 0.2     | 3.66E-17        |

Table 5.12: IV. Composite false positive probabilities for bilayer verification following equation (5.2). Sample size of $N = 50$ and $\eta_1 = 0.01$.

interest and conservatively locating the maximum p-value reported over all other parameter choices in each simulation table, we draw the following conclusion. At small sample sizes in the neighborhood of $N = 25$ and $N = 50$, for a critical parameter of $\eta_1 = 0.05$, the false positive rate for verification is conservatively less that $1 \times 10^{-10}$, independent of the choices for $\eta_2$, $\eta_3$, and $\eta_4$ that were studied. Under these same conditions, for $\eta_1 = 0.01$, the probability of false positive verification is less than $1 \times 10^{-15}$. These findings offer guidance to the utilization of v-structure based causality analysis, and may find particular utility when accounting for multiple comparisons.

5.2 Learning an aortic lesion network

The RVV algorithm was run over the C3H x B6 liver microarray data given the starting point of the aortic lesion scores, and parameters ($\eta_1 = 0.05, \eta_2 = 0.3, \eta_3 = 0.2, \eta_4 = 0.1$). In the resulting network for the male mice, shown in Figure 5.1, three nodes stood as direct inputs to the aortic lesion nodes. These three direct parents had many, many more parents than could be pictured in the truncated output shown. The algorithm was halted after the two layers were learned. The data shows many very highly correlated genes along with the primary parents,
Figure 5.1: **Truncated results of the RVV algorithm run on Aortic Lesions in the C3H x B6 data, male mice, Liver mRNA profiled.**

A630026L20, Atp11c and Nola2. As a caveat, we suspect that the algorithm may be vulnerable to an order effect as to which genes are assayed first when a high degree of correlation amongst many hundreds of correlated genes is found. This will be investigated in future work.

### 5.3 Further utilizing networks: assigning functional roles to genes

Given a learned network of clinical covariates or a partial pre-specification of a causal Bayesian network based on prior knowledge of the relationships between clinical traits, it then becomes possible to assign functional roles to genes by assaying which gene, when interposed in the network, best satisfies the causal constraints of d-separation.

Figure 5.2 is a model of the processes involved in metabolic syndrome. Consider the subgraph (Figure 5.3) from the protective HDL Cholesterol via an unknown gene A to Atherosclerosis. Which gene in a given tissue known to be of import, say Liver, best d-separates Cholesterol from Aortic lesions in the observed data? Such a gene might be a candidate for investigating the protective effects of HDL cholesterol on atherosclerosis.
Figure 5.2: Relationships between clinical covariates, genes and genotypes for metabolic syndrome.

Figure 5.3: The subgraph for assigning gene functional roles.
Results for an analysis of the subgraph appear in Table 5.13, where Insulin like growth factor 1 (Igf1) is located as the gene that most screens between HDL cholesterol and aortic lesions. All genes in livers for the male mice were ranked (approximately 23K genes); only the top ranked genes are shown. We note a very encouraging sharp numeric drop in partial correlation when going from Igf1 to the closest runner up, Serpincl1. HDL cholesterol levels and aortic lesion counts were initially correlated at -0.18 (Pearson correlation p-value: 0.019 from 158 mice). We find in the males that liver Igf1 levels renders HDL and aortic lesions conditionally independent to the degree of -0.049 partial correlation, the smallest of any gene. The negative correlation is expected, as HDL cholesterol is known to have a protective effect with regards to heart disease.

Although derived from accurate experimental data, these results await secondary experimental confirmation. The principle value of these results is to suggest that the methodology here may be of value. By first inferring or describing a know clinical trait network, we in effect create roles within a known drama; we can then cast genes into those roles, investigating the microarray data at hand to see which genes might best fit which given roles. To do so we utilize the predicted conditional independence (d-separation) properties from the learned or specified graph.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Partial correlation between HDL and Aortic lesions given Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Igf1</td>
<td>-0.0494</td>
</tr>
<tr>
<td>Serpinc1</td>
<td>-0.0812</td>
</tr>
<tr>
<td>Cyp2d9</td>
<td>-0.0936</td>
</tr>
<tr>
<td>Arsa</td>
<td>-0.0966</td>
</tr>
<tr>
<td>Mbl2</td>
<td>-0.0980</td>
</tr>
<tr>
<td>Eef1a1</td>
<td>-0.0985</td>
</tr>
<tr>
<td>Ahcy</td>
<td>-0.101</td>
</tr>
<tr>
<td>CRAD-L</td>
<td>-0.103</td>
</tr>
<tr>
<td>MMT00027866</td>
<td>-0.108</td>
</tr>
<tr>
<td>2310014L17Rik</td>
<td>-0.109</td>
</tr>
<tr>
<td>2810460C24Rik</td>
<td>-0.109</td>
</tr>
<tr>
<td>Ugt1a1</td>
<td>-0.110</td>
</tr>
<tr>
<td>5430417P09Rik</td>
<td>-0.113</td>
</tr>
<tr>
<td>Omd</td>
<td>-0.113</td>
</tr>
<tr>
<td>Ftl1</td>
<td>-0.113</td>
</tr>
<tr>
<td>Treh</td>
<td>-0.113</td>
</tr>
</tbody>
</table>

Table 5.13: Functional gene assignment given a specified clinical covariate network. Ranked best candidates.
CHAPTER 6

Introduction to Orthogonal Causal Anchors and Local-structure Edge Orienting (LEO) Scores for Robust Causality Testing and Non-acyclic Gene Network Recovery.

Our studies of causal network recovery next focus on the opportunities afforded when causal anchors, exogenous variables known to be upstream of endogenous variables in the network, are available. Common genetic variations represented by genetic markers, such as single nucleotide polymorphisms (SNPs), are our primary example. We study how these anchors can be utilized both in isolation (single marker analysis) and in combination (multiple marker analysis) when multiple causal anchors can be exploited for network recovery. Two features distinguish our work from earlier studies. First we focus on the advantages of multiple marker methods and in particular on the advantages of utilizing one or more orthogonal causal anchors (see Figure 6.1). In this context we develop and evaluate several automated methods of choosing and aggregating multiple anchors and leveraging orthogonal causal anchors for causal edge orienting and confirmation. Secondly, for each given candidate network edge $A \rightarrow B$, we formulate an edge score (the Local-structure Edge Orienting or LEO score) which summarizes the evidence from multiple causal anchors without imposing the mathematical constraint of acyclicity on the final network. By relaxing the common invoked
acyclicity requirement on the underlying graph, we acknowledge the ubiquity of feedback mechanisms in biological networks. In addition we shield ourselves from the problem of error propagation when mistaken acyclicity-derived conclusions unduly influence neighboring edge evaluations.

Although formulated in mathematically general terms, our causal network recovery algorithms are primarily responses to advances in technology. Genome-wide microarray data measurements and high-throughput genotyping technologies supply a wealth of data at the tissue level for gene expression changes induced by common genetic variation. These data invite new avenues of attack on complex disease. However the utilization and combination of these data sources presents challenges in part due to the large number of (often correlated) measurements, and in part due to the underlying natural complexity of the biological interaction networks they reveal. While classical QTL interval mapping approaches have been expanded to encompass multiple markers (Zen93; Zen94; JZ95; JZ97), they remain single layer inference procedures. Meanwhile gene co-expression network analysis can pick out useful modules or sets of co-regulated genes that may be responding to multiple pathways and on multiple levels. However it is important to note that a co-expression network is an undirected network that lacks information about causal flow.

Although the study of genetic variation and its phenotypic correlates gives us a strong handle on causal information flow through biological systems, the information avalanche of genome wide SNP chips presents the challenge of how best to integrate and weigh at times conflicting genetic evidence for causal direction in gene-gene and gene-trait networks.

Here we explore various ways of integrating and weighing the genetic information to induce structure on microarray-derived gene networks. Our methods construct directional networks without acyclicity constraints, and do so in a way
that isolates the score for each edge from its neighbors so as to resist error propagation.

One appeal of the directed acyclic graph (DAG) representation of a biological network is the apparent ability to predict the downstream effects of interventions in the network. However, the insistence on acyclic graphs in many Bayesian network learning algorithms is often more a mathematical convenience than a biological reality. When more and more edges are oriented, as in the PC/PC*(SGS00a) and IC/IC*(Pea00) algorithms, from information gained from neighboring edges of potentially ambiguous orientation, a mistake in one part of the network can ripple across and cause erroneous arrows in another unrelated portion of the network. Most often these amplified mistakes are due to confusion between confounded and truly causal directional flow.

Here we design algorithms to harness the multiple simultaneous genetic variations available in F2 intercross to better our recovery truly causal gene-trait expression relationships. Inspired by Pearl’s definition of Genuine Cause(Pea00) and our earlier work on bilayer verification theory(Ate06; Ate08), the principle novel feature of the algorithms introduced here is our emphasis on exploiting orthogonal causal anchors (OCA). This is illustrated in Figure 6.1, which also summarizes earlier approaches.

Specifically we here propose and explore methods that allow the relaxation of the acyclicity constraint while still recovering directed networks. We base our edge orienting on local models which integrate evidence only from causal anchors: exogenous variables (genetic markers or SNPs) which we can be reasonably certain are acting upstream in the network of gene-gene and gene-trait interactions. We focus on the utility of orthogonal causal anchors (OCA) in particular. In the model \( M_A \rightarrow A \rightarrow B \leftarrow M_B \), the anchors \( M_B \) are orthogonal to the \( M_A \) anchors, meaning that \( M_B \) independently impacts the same downstream trait \( B \) that is
affected by anchor \( M_A \) through \( A \). The asymmetry of the OCA relationship offers us additional insurance that the causal \( A \to B \) flow is correct, for otherwise we would see the influence of \( M_B \) on \( A \).

While we and others have found that edge orienting without causal anchors can be done from observational data alone (Pea88; Pea00; SGS00a; Ate08; Ate06; Coo97; SBM00; Dav85), significant statistical power and specificity can be gained by employing the causal anchors provided by SNPs (SLY05; LTS06), instrumental variables (DS05; Smi06; TC04), and assigned treatment randomization (Fis54). Each of these offers a means of establishing causal anchors (Coo97).

Figure 6.1 briefly summarizes earlier approaches. In contrast to earlier work, we study expanded models shown in Figure 6.1(c). Specifically, we search for and fit models with the additional \( M_B \) set of causal anchors (markers) that show a pattern of association with \( B \) but not with \( A \), and are simultaneously pair-wise independent of \( M_A \). The \( M_A \) markers we call the common pleiotropic anchors (CPA).

The phrase “pleiotropic anchor” indicates that \( M \) impacts both \( A \) and \( B \) simultaneously. Similarly, the set \( M_A \) contains markers that share in the common their pleiotropic impact on both \( A \) and \( B \). These \( M_A \) markers are called common pleiotropic anchors.

The \( M_B \) markers we term the orthogonal causal anchors (OCA) to signify their independence from both \( M_A \) and \( A \). The \( M_A \to A \to B \leftarrow M_B \) model in the lower right (Fig. 6.1(c)) we call the orthomarker model. In contrast, earlier (Fig. 6.1(a) and (b)) methods may locate one or more pleiotropic QTL that appear to control the same two \( A \) and \( B \) traits, without also finding an OCA. Although CPA alone are sufficient to suggest causality, when one or more OCA is available we have secured additional evidence that the \( A \to B \) causal flow is correct. The OCA also improves our detection power in the presence of
confounded correlation, as shown in the results sections (see Figures 10.1 and 10.2). The use of OCA is a practical extension to genetic causal anchors of Bilayer verification theory (Ate06; Ate08). Since repeated applications of these kinds of tests can serve to orient many if not most of the edges in an undirected network, we call the software implementation of these algorithms the Network Edge Orienting methods. Our NEO software implements these algorithms.

Specifically, our Network Edge Orienting (NEO) algorithms provide automated methods for modeling the impacts of multiple QTLs on multiple traits and for selecting and generating directed genetic network models of gene-trait and gene-gene interactions. To orient directed network edges, we fit and contrast specific local Structural Equation Models (SEM) to score the directed edges between pairs of variables. Figure 6.2 gives an overview of the procedure. SEM based models are, by design, linear models of interacting variables.

The balance of this text is structured as follows. We first review related work as it applies to genetic and microarray edge orienting. Then we detail the methods, theory, definitions, and algorithms of our Network Edge Orienting approach and the Local-structure Edge Orienting (LEO) algorithms it utilizes. Next we present the results of the Monte Carlo studies of power and specificity. Lastly, we apply NEO to a mouse intercross data set and demonstrate that we recover established biological relationships in the well studied cholesterol biosynthesis pathway (HGB02). Our results also suggest new genes that may be co-regulated with cholesterol in the liver. These genes appear, by our analysis, to be downstream of the key sterol response mediator, Insig1 (RIK07).
Figure 6.1: The logic of network edge orienting, and the advantage of orthogonal causal anchors. To visualize the central ideas of this paper, consider first (a) single anchor edge orienting, suggested by previous authors (Coo97; SLY05; LTS06). In the upper half of (a) we can begin orienting when we see the pattern of correlation between a marker and two traits shown. The bottom half of (a) shows a successfully oriented $A \rightarrow B$ edge. In linear terms, the correlation between $M$ and $B$ was here successfully described as the product of the $M \rightarrow A$ correlation and the $M \rightarrow B$ correlation. In terms of independence ($\perp$) and dependence ($\not\perp$) statements, we found that $M \not\perp A$ and $M \not\perp B$, while $(M \perp B) | A$ and $(M \not\perp A) | B$. Considering (b), authors (SLY05) have also extended the single anchor model by searching for a set of causal anchors ($M_A$) that have common pleiotropic effects. We also build sets of causal anchors $M_A$ and $M_B$, however we employ forward step-wise regression to assemble these sets. Also in contrast to earlier work, we study expanded models shown in (c). Specifically, we search for and fit models with the additional $M_B$ set of causal anchors (markers) that show a pattern of association with $B$ but not with $A$, and are simultaneously pair-wise independent of $M_A$. The $M_A$ markers we call the common pleiotropic anchors (CPA). The $M_B$ markers we term the orthogonal causal anchors (OCA) to signify their independence from both $M_A$ and $A$. The $M_A \rightarrow A \rightarrow B \leftarrow M_B$ model in the lower right (c) figure we call the orthomarker model. In contrast, the (a) and (b) methods may locate one or more pleiotropic QTL that appear to control the same two $A$ and $B$ traits, without also finding an OCA. Although CPA alone are sufficient to suggest causality, when one or more OCA is available we have secured additional evidence that the $A \rightarrow B$ causal flow is correct. The OCA also improves our detection power in the presence of confounded correlation, as shown in Figures 10.1 and 10.2. The use of OCA is a practical extension to genetic causal anchors of Bilayer verification theory (Ate06; Ate08).
Overview of Network Edge Orienting

1) Align genetic markers and traits (clinical phenotypes and mRNA traits) across individuals 1..N

2) Specify manually genetic markers of interest, or invoke automated marker selection & assignment to trait nodes

   Automated tools:
   • greedy & forward-stepwise SNP selection;
   • marker assignment consistency principle

3) Compute Local-structure edge orienting (LEO) scores to assess the causal strength of each A-B edge
   • based on likelihoods of local Structural Equation Models
   • integrates the evidence of multiple SNPs

4) For each edge with high LEO score, evaluate the fit of the underlying local SEM models
   • fitting indices of local SEMs: RMSEA, chi-square statistics
   • BLV scores for confounding

5) Robustness analysis with regard to automatic marker selection;

6) Repeat analysis for next A-B edge

Output
• NEO spreadsheet summarizes LEO scores and provides hyperlinks to model fit logs
• graph of the directed network

Figure 6.2: Overview of Network Edge Orienting.
CHAPTER 7

Relationship of multi-anchor and OCA methods to prior work

While the specific application of path models to microarray data with genetic variation is being actively explored (SLY05; LTS06; ZWZpr; KJ06), several fields have addressed the general problem of learning directed networks, and inferring causal relationships from observed and experimental data. While this literature is extensive, important prior work that can be applied to genetic and microarray analysis exists in the fields of learning in graphical models (Jor98; LS03; SS04) and in Bayesian networks (Pea88; Nea04; SGS00b; KN04; Ste01; Coo97; SBM00). In epidemiology, this approach is closely related to Mendelian randomization, a causal inference procedure based on the random assortment of alleles from parents to offspring that occurs during meiosis (TMA05; LK03; CM01; SE04; SE03; Kat04; Kat86). Readers orienting themselves to the field may also benefit from work describing the details of Structural Equation Model (SEM) practice and from certain works on automatically generating SEM from data (Pea00; ZLL04; CGB95; Fox06; Shi00a; Shi00b).

We next briefly discuss two specific examples of more closely related work by Schadt et al. and Li et al., and contrast their methods to our approach.

Schadt et al. (SLY05) utilize a hybrid between the single anchor and the common pleiotropic anchor approach (refer to our Fig. 6.1, panels (a) and (b)). Similar to CPA, the authors utilize Efroymson’s forward-backward stepwise re-
gression procedure to build the initial genetic model for the downstream trait (Mil96). The method next applies a generalization of Jiang and Zeng’s pleiotropy-versus-close-linkage test (JZ95) in an attempt to discard pairs of cis-eQTL regulated genes that are correlated due to linkage and retain only truly pleiotropic loci. For each locus retained in the genetic model for a given trait, the LCMS (Likelihood-based Causality Model Selection) test evaluates genes for causality by comparing three of the five single anchor models (models 1, 2, and 3) shown in our Figure 8.1 for smallest AIC. Taking just the genes for which the causal model (model 1 of Figure 8.1 for a single gene A and trait B) scores best for at least two common pleiotropic markers, the candidate causal gene list is output by ranking genes according to “the amount of genetic variance of the...trait that was causally explained by variation in their transcript abundance,” which, while not defined precisely in published accounts, seems to correspond reasonably directly to comparing the probabilities of the CPA models for all final candidate genes. In contrast, we utilize orthogonal anchors and fit multiple anchor and multiple orthogonal anchors at once; we document the power and false positive advantages of this approach below.

Li et al (LTS06) start with classical QTL significance analysis and then, choosing a significant trait QTL, they utilize repeated application of the single anchor (pleiotropic QTL; see Fig. 6.1(a)) test to build up a pool of $M \rightarrow A \rightarrow B$ paths that are then assembled by hand into a final model. Although in this output, some loci may act independently on the downstream trait, this is not a characteristic that is used to validate the finding or searched for a priori. In this method, single edge removal and addition is next evaluated iteratively by SEM goodness of fit measures to optimize the medium scale multi-marker and multi-trait model. Multiple (two and three) trait models and a model with a latent factor (here Adiposity) are assembled in this piecewise fashion. This model build-
ing method is in contrast to our explicitly evaluating the fit of an orthomarker model in advance to help secure correct edge directionality before proceeding to full network evaluation. Also we expect obtain higher power in as much as we do not require a significant trait QTL up front, but rather a better scoring causal $A \rightarrow B$ model for the trait.
CHAPTER 8

LEO Methods

8.1 SEM and the definition of model p-values, P(model)

Structural equation modeling (SEM) extends multivariate linear regression models to include hierarchical relationships, latent variables, and factor analysis derived components (J70; Bol89; Loe04; Pea00; Fox06; Fox84; Shi00a). The latent variables and factor analysis or measurement model components are optional. Variables or nodes are connected by arrows denoting causal relationships. The parents of a node define a linear regression model in which the parents of a node are used to predict the child node’s response. Although variable means are also characterized in extended SEM methodology, a typically SEM analysis starts with variables centered on their means and focuses on the covariance relationships. The structural models used in this paper are illustrated in the Figures 8.1 and 8.2.
Figure 8.1: Illustrating the single anchor models used in the definition of the central LEO.NB score (equations 8.6 - 8.9) for each candidate directed edge $A \rightarrow B$. By definition, $\text{LEO.NB}(A,B) = \log_{10}\{P(\text{model 1})\}/\{\max_{i>1}\{P(\text{model } i)\}\}$ for a candidate $A \rightarrow B$ edge orientation, where the models in the definition are pictured above for single anchor LEO.NB scores, and in Figure 8.2 for multiple anchor LEO.NB scores. Residual arrows or error terms are modeled for $A$ and $B$ nodes but not pictured. A positive LEO.NB$(A,B)$ score indicates that the directional edge $A \rightarrow B$ is better than any of the competing models. The LEO.NB.ALL aggregation method weighs the average of the single anchor LEO.NB$(A,B)$ against the average of the single anchor LEO.NB$(B,A)$. The LEO.NB.CPA method uses these same single anchor models, but allows multiple markers to fill the role of $M$. In Figure 8.2 we show the orthomarker models (1 and 2) used for the LEO.NB.OCA marker aggregation method. When orthogonal causal anchors are available, they can be used to obtain higher power and lower false positive rates. While the multiple anchor models presented next include a model with a hidden (latent) variable connecting $A$ and $B$, no such model is included here because it was found to be indistinguishable from the models with a collider, such as model 4 and model 5, that are included here.
Figure 8.2: Illustrating the orthogonal causal anchor (orthomarker) models used in the definition of the central LEO.NB score (equations 8.6 - 8.9) for each candidate directed edge $A \rightarrow B$. Residual arrows or error terms are modeled for $A$ and $B$ nodes but not pictured. The top level nodes ($M_A^{(i)}$, $M_B^{(j)}$) represent markers (SNPs) or causal anchors. Multiple anchor models that include orthogonal causal anchors give added insurance that the causal flow is correct.

Following the exposition of Loehlin(Loe04) and Bollen(Bol89), given $m$ observed variables, $S_{m \times m}$ denotes the observed sample covariance matrix, and $\Sigma(\hat{\theta})_{m \times m}$ or simply $\hat{\Sigma}$ is the sample (as opposed to population) covariance matrix as predicted by the model.

Maximum likelihood fitting for the multivariate normal assumption uses the following likelihood function. For maximization purposes, it may be observed that $-\ln |S|$ and $m$ are constant and can be ignored during optimization.

$$F_{ML} = \ln |\hat{\Sigma}| - \ln |S| + trS\hat{\Sigma}^{-1} - m \quad (8.1)$$
Supposing that residual sampling variation is multivariate normal and that the model is correctly capturing the underlying process, then the maximum likelihood fit can be evaluated by the chi-squared distribution, since at the best value of the likelihood function $F_{ML}^*$, residual variation should then be the sum of squared normal errors so that:

$$X^2 \text{ defined as } = F_{ML}^*(N - 1) \quad (8.2)$$

and

$$X^2 \sim \chi^2_{(df = \frac{m(m+1)}{2} - t)} \text{ as } N \to \infty \quad (8.3)$$

Here the degrees of freedom is a function of $m$ and $t$, where $t$ is the number of free parameters (not fixed values and not functions of other free parameters). Free parameters are those not constrained by the causal structure hypothesis nor by prior knowledge and are estimated during optimization of the fit. If the error terms for the endogenous variables’ equations are thought of as independent variables, then the free parameters consist of (1) the variances and covariances of the independent (including exogenous) variables, plus (2) the regression coefficients from the endogenous variables’ equations (Ben06).

Multiplying the minimized fit function by one minus the sample size yields an (asymptotically) chi-squared distributed variable under the hypothesis of model correctness and multivariate normal sampling variation. Here a smaller chi-squared statistic is preferred and indicates that the data fit the model well. Given the maximum likelihood or two-stage least squares parameter estimates (Bol89), we take as our operational definition of model probability the p-value obtained from the right tail of the chi-squared distribution under the hypothesis that the data is consistent with the model up to sampling variation.

$$\text{Model } \chi^2 \text{ test p-value or } P(\text{model}) = \Pr(\chi^2_{df = \frac{m(m+1)}{2} - t} > X^2) \quad (8.4)$$
This operational definition of model p-value (the term model probability is used interchangeably) reflects that the higher the model p-value, the better the causal model fits the data. A small model p-value (say \( p < 0.05 \)) indicates that the causal model does not fit well. This is common in a ‘accept-support’ context (SF97; Kli05) where the null hypothesis represents the researchers belief, and it is the failure to reject the null hypothesis that supports the causal model. We utilize this definition for \( P(\text{model}) \) in the central procedure for edge orienting, which is described next.

8.2 Network Edge Orienting (NEO)

Here we introduce a multi-step method that involves aligning the data, specifying QTL or choosing automated SNP selection, computation of the LEO scores, evaluating the fit of the models for the significant edges, and potentially doing a robustness analysis for the edge. We next detail each of these steps in turn. Figure 6.2 provides a visual overview of the algorithms.

8.3 Step 1: Align traits (gene expression traits and clinical traits) and SNPs across subjects

As currently implemented the NEO methods take as input an R data frame of SNP columns and trait columns where each row represents an individual subject on which those traits and genotypes were both measured. Traits can include microarray data as well as clinical phenotypes. Each SNP and each supplied trait is taken as a node in the network, and the NEO methods evaluate and score each analyst-specified pair of \( A \rightarrow B \) traits. NEO scores both the \( A \rightarrow B \) and the \( B \rightarrow A \) directions in a single pass, using the causal anchors chosen in step 2.
8.4 Step 2: SNP selection and assignment to nodes

8.5 Step 2(a): QTL analysis or user specification of loci of interest

If previous linkage or association studies are available then manual selection of SNPs at the loci of interest may be desirable. SNPs of interest may already be readily at hand and are then supplied by the analyst in this step 2(a). The investigator in effect pre-specifies these SNPs of interest to play the roles of \( M, M_A, \) and \( M_B \) in the LEO scoring computations, and a flag is set to tell the software not to do the Step 2(b) automated SNP selection.

While earlier authors (SLY05; LTS06) implement a pleiotropy test at this point by searching for traits that share common QTL, we take the novel approach of integrating the search for pleiotropic QTL with the process of automated model building and scoring. We argue that this approach is more statistically powerful, and supply evidence to that effect in the results section.

If prior QTL analysis is not available, or if one wishes to perform a de novo search through a broad array of surveyed loci, then we propose next several automated scan and model building methods for choosing the SNPs in the edge orienting LEO models. Step 2(b) describes these automated SNP selection techniques.

8.6 Step 2(b): Automated unsupervised SNP selection methods

Automated QTL scanning and unsupervised causal model building can be useful if the analyst possesses no prior information about loci of interest. In addition,
if one suspects that multiple loci of medium impact are influencing a trait, then automated SNP selection methods may capture the synergy of the genetic variations in ways that the common practice of establishing a strict genome wide QTL significance threshold may miss. The Insig1 example in the results section provides a case to this point. We outline our automated SNP selection methods in a series of three steps.

(2.b.1) Isolated optimal selection. First each trait in the $A - B$ pair to be oriented is considered in isolation, and both a greedy (selecting the top $K$ strongest associations from the panel of SNPs) and a forward-stepwise regression approach (repeatedly removing the strongest signal SNP and then sifting through the trait’s residuals for the next strongest SNP) are employed for each of $A$ and $B$ independently. The user can control, by the adjustment of particular parameters, the maximum number of SNPs chosen, and the minimum association strength for a SNP to be selected. We call the set of markers chosen for the $A$ node the set $M_A$, and the corresponding $B$ node SNPs are $M_B$.

(2.b.2) Obtain the marker assignment consistency invariant. Second, we consider the traits $A$ and $B$ together again as a pair, and invoke the following heuristic principle.

Definition 5 The marker assignment consistency principle. Any candidates for the role of $M_A$ must be more strongly associated with $A$ than with $B$. Similarly $M_B$ markers should show stronger association with $B$ that with $A$, since it seems quite unlikely that the marginal linear association between SNP and trait can be stronger for a more distal node in the graph when compared to a more proximal node.

In this step SNPs are swapped between $M_A$ and $M_B$ if necessary to obtain consistency with the marker assignment principle. These mutually consistent sets
are then employed directly in the ALL aggregation strategy. With this invariant in place, we return to the task of optimally selecting SNPs from within the candidate pools of $M_A$ and $M_B$.

(2.b.3) Second-stage forward-stepwise modeling. Once the assignment invariant above is in force, we then do forward-stepwise modeling now using AIC stopping criteria on the set of consistent $M_A$ and $M_B$ SNPs, establishing multiple regression formulae for the relationship between $M_A$ and $A$ and likewise between $M_B$ and $B$. These forward stepwise regression formulae become the building blocks of the OCA aggregation strategies. The strongest associated single pair of independent SNPs from $M_A$ and $M_B$ are reported in the MAX aggregation strategy.

Using only Greedy SNPs may miss important secondary causal SNPs whose signals are swamped out by the tightly linked markers neighboring the primary causal signal SNP. The forward-step selected SNPs remedy this potential defect.

8.7 Step 3: Aggregate SNPs into local structural equation models and compute Local-structure Edge Orienting scores to assess the strength of the evidence for each edge direction

Four different strategies for aggregating multiple SNPs into the edge models were studied. The building blocks for these models are pictured in Figures 8.1 and 8.2. The MAX.MAX, LEO.NB.ALL, LEO.NB.CPA, and LEO.NB.OCA strategies combine the evidence from SNPs for edge direction in distinct ways. The MAX.MAX method is implemented here for contrast, in that it represents a simple baseline and readily understood model to compare against. To generate
a MAX.MAX score, the maximum single-marker probability for one edge orientation is contrasted with the maximum probability marker for the opposite orientation.

\[
\text{MAX.MAX}(A, B, M_A, M_B) = \max_i \left[ P(M_A^{(i)} \rightarrow A \rightarrow B) \right] - \max_k \left[ P(M_B^{(k)} \rightarrow B \rightarrow A) \right] \tag{8.5}
\]

In LEO.NB.ALL scoring, we take the average of the single anchor LEO.NB(A,B) scores, as computed over the all single anchor models fit to \(M_A\) and then subtract the separately computed average of the single anchor LEO.NB(B,A) scores for \(M_B\).

\[
\text{LEO.NB.ALL}(A, B, M_A, M_B) = \frac{\sum_i \text{single anchor LEO.NB}(A, B, M_A^{(i)})}{|M_A|} - \frac{\sum_k \text{single anchor LEO.NB}(A, B, M_B^{(k)})}{|M_B|} \tag{8.6}
\]

The single anchor LEO.NB scores, used in the LEO.NB.ALL, are pictured in Figure 8.1 and are defined as follows.

single anchor LEO.NB(A,B, M) =

\[
\log_{10} \max \left( \frac{P(\text{model 1: } M \rightarrow A \rightarrow B)}{P(\text{model 2: } M \rightarrow B \rightarrow A), P(\text{model 3: } A \leftarrow M \rightarrow B), P(\text{model 4: } M \rightarrow A \leftarrow B), P(\text{model 5: } M \rightarrow B \leftarrow A)} \right) \tag{8.7}
\]
In LEO.NB.CPA scoring, we compare the single anchor models shown in Figure 8.1 however in place of the single anchor $M$ we use an $M_A$ set of markers that can be curated manually but by default is assembled by forward stepwise regression.

$$LEO.NB.CPA(A, B, M_A) = \log_{10} \left( \frac{P(\text{model 1: } M_A \rightarrow A \rightarrow B) \max \left( P(\text{model 2: } M_A \rightarrow B \rightarrow A), P(\text{model 3: } A \leftarrow M_A \rightarrow B), P(\text{model 4: } M_A \rightarrow A \leftarrow B), P(\text{model 5: } M_A \rightarrow B \leftarrow A) \right)}{P(\text{model 1: } M_A \rightarrow A \rightarrow B)} \right)$$ (8.8)

In LEO.NB.OCA scoring, we compute the LEO.NB score for multiple markers (pictured in Figure 8.2 that and defined in equation 8.9), where forward-stepwise regression is used to obtain the $M_A$ and $M_B$ sets from those markers that satisfy the marker assignment consistency principle. Here $H$ is a hidden confounder. The LEO.NB.OCA score utilizes orthogonal causal anchors and asks if the $A \rightarrow B$ orthomarker model has higher probability than other models.

$$LEO.NB.OCA(A, B, M_A, M_B) = \log_{10}(y)$$

with $y = \frac{P(\text{m1: } M_A \rightarrow A \rightarrow B \leftarrow M_B)}{\max \left( P(\text{m2: } M_A \rightarrow A \leftarrow B \leftarrow M_B), P(\text{m3: } B \leftarrow M_A \rightarrow A; B \leftarrow M_B \rightarrow B)), P(\text{m4: } M_A \rightarrow A \leftarrow H \rightarrow B \leftarrow M_B) \right)}$ (8.9)

We note that in the definitions above, intra-set $M_A$ and intra-set $M_B$ covariance arrows were omitted from the model specifications for clarity, but are present in each of the multiple marker models between each pair of markers drawn
from $M_A$; and independently between pairs of markers in $M_B$. Covariances between the $M_A$ and $M_B$ members are explicitly excluded as the advantage of the independent control set is the principle feature under examination here.

Regarding more elaborate models than those pictured in Figure 8.1: while authors such as Li et al. (LTS06) illustrate and so presumably also study single anchor structural equation models that include two outward flowing paths from the SNP marker $M$ (i.e. a direct path from $M$ to each of $A$ and $B$), we prefer to test in advance whether the $M \rightarrow A$ relationship or the $M \rightarrow B$ relationship is dominant, and so to ascertain, even if the truth does contain a mixture of influences, which of the illustrated models most closely the evidence. An additional practical reason for excluding such models is that they are “just identified”–with no fixed parameters. Such models have zero degrees of freedom and so do not permit computation of $\chi^2$ fits or the corresponding goodness of fit indices and probabilities. Each of the single anchor models illustrated here (in Figure 8.1) includes one degree of freedom and hence permits probability computation and model comparison. For the orthomarker and multiple marker models of Figure 8.2, there will be some variation in the degrees of freedom; the models are compared then using their overall relative probabilities, derived from the (corresponding degree of freedom) $\chi^2$ distribution.

8.8 Step 4: For each edge with high LEO score, evaluate the fit of the underlying local SEM models

From the NEO spreadsheet output we examine the fit of the high scoring LEO.NB edges by inspecting the individual models fits for each of the models depicted in Figures 8.1 and 8.2. Ideally model 1 $A \rightarrow B$ will show high probability and low RMSEA score, while the competing models show low probability and
high RMSEA scores. A vulnerability of any probability ratio approach with the LEO.NB score is that the ratio of two small numbers may be arbitrarily inflated due to the variance of small measurements. Hence for rigor and confidence in the conclusions, it is important to inspect the model probabilities which give the probability of observing any departures from the model structure and fit parameters due to Gaussian sampling variation.

8.9 Step 5: Robustness analysis with respect to significance thresholds and SNP selection

Since the results of automated SNP selection and model building may be sensitive to which particular SNPs are included in the models, we offer the option to check the robustness of this model building by titrating the number of SNPs selected. Such analysis was used to obtain the results shown in Figures 10.11 - 10.14.

8.10 Step 6: Repeat analysis for next A-B trait-trait edge and apply edge score thresholds to orient the network

If there are multiple nodes to be learned in a network, then steps 1-5 are repeated for each pair of trait-trait nodes that we wish to orient. Then the NEO results are displayed in a spreadsheet which summarizes the scores and evidence from the local model fitting processes for each candidate edge orientation. When inspecting each edge we also provide certain specialty scores, the BLV and the ZEO scores, which can be used to evaluate specific aspects of the models (in the case of BLV), and to speed the overall computation (in the case of the ZEO) when genome-scale computations are desired. These are presented in results chapter 10.3, while in the next chapter we details discussion of the Monte Carlo performance evaluation
testbed used to evaluate, compare and contrast our algorithms.
CHAPTER 9

Monte Carlo methods to evaluate power and
specificity of $A \rightarrow B$ edge orienting

Here we provide details of the simulation testbed used for method comparison and evaluation. We first define a particularly useful parameter for our models, restricted heritability, and then outline our performance evaluation model.

9.1 Restricted heritability as a parameter

Genetic heritability describes the percentage of total variation in a sampled trait that is due to genetic variation. For the purposes of the Monte Carlo studies, we parameterize the model of Figure 9.1 using a heritability parameter we call restricted heritability. Restricted heritability or $h'$ restricts attention to the genetic influences on $A$ that do not act via $B$, and to the genetic influences on $B$ that do not act via $A$. Hence in these models, $h'_A$ is the restricted heritability from the genetic variables into $A$, excluding the genetic influences that are acting via $B$. The observed heritability for $A$ is obtained if we include those influences acting via $B$; by adding $\omega^2 h'_B$.

$$h_A = h'_A + \omega^2 h'_B \quad (9.1)$$

Restricted heritability allows us to cleanly and discretely vary one param-
eter of the Monte Carlo model at a time, without confounding heritability and the causal strength $\omega$ during our testing and evaluation.

9.2 The performance evaluation model

Figure 9.1 illustrates the performance evaluation model for studying ability of OCA based edge orienting methods to distinguish between causal $A \rightarrow B$ correlation and confounded $A \leftarrow C \rightarrow B$ correlation. Without loss of generality, the latent variable $C$ and each marker is designated unit variance and that $\sum \alpha_i^2 = h'_A$ and that $\sum \beta_i^2 = h'_B$. Hence we have $0 \leq h'_A, h'_B < 1$. We also assume that the markers into $A$ and the markers into $B$ are independent conditional upon $A$ and $B$. Under these assumptions, the variance and covariances of the standardized variables $A$, $B$, and $C$ are as follows.

\[
\begin{align*}
\text{Var}(A) & = h'_A + \gamma^2 + \omega^2 + 2\gamma^2\omega + \sigma^2_{\epsilon_A} = 1 \\
\text{Var}(B) & = h'_B + \gamma^2 + \sigma^2_{\epsilon_B} = 1 \\
\text{Cov}(A, B) & = \gamma^2 + \omega \\
\text{Cov}(A, C) & = \gamma + \omega\gamma \\
\text{Cov}(B, C) & = \gamma
\end{align*}
\]

(9.2) \hspace{2cm} (9.3) \hspace{2cm} (9.4) \hspace{2cm} (9.5) \hspace{2cm} (9.6)

Here we note that because of the standardized unit variance variables, $\rho_{AB} = \text{Cor}(A, B) = \text{Cov}(A, B)$. We derive from the above the following Monte Carlo simulation procedure for generating data from a specified model. Let $\theta_{PCS} = \omega/(\omega^2 + \gamma^2)$ be the percentage of signal that is causal out of the observed correlation ($\rho_{AB} = \omega + \gamma^2$). Given an observed A-B correlation of $\rho_{AB}$, a $\theta_{PCS}$ value, and the restricted heritabilities $h'_A$ and $h'_B$, generating simulated data sets whose variation is consistent with the performance evaluation model of Figure 9.1 can
be accomplished as follows.

\[
\begin{align*}
\omega &= \rho_{AB} \cdot \theta_{\text{PCS}} \quad (9.7) \\
\gamma^2 &= \rho_{AB} - \omega \quad (9.8) \\
\sigma^2_{\epsilon_A} &= 1 - 2\omega\gamma^2 - \gamma^2 - \omega^2 - h_A' \quad (9.9) \\
\sigma^2_{\epsilon_B} &= 1 - \gamma^2 - h_B' \quad (9.10) \\
C &\sim N(0, 1) \quad (9.11) \\
\epsilon_A &\sim N(0, \sigma^2_{\epsilon_A}) \quad (9.12) \\
\epsilon_B &\sim N(0, \sigma^2_{\epsilon_B}) \quad (9.13) \\
B &= \sum_j \beta_j M_B^{(j)} + \gamma C + \epsilon_B \quad (9.14) \\
A &= \sum_i \alpha_i M_A^{(i)} + \gamma C + \omega B + \epsilon_A \quad (9.15)
\end{align*}
\]

This method was used to evaluate the algorithms under controlled conditions, and the results of these evaluations are presented in chapter 10.
Figure 9.1: **Performance evaluation model used in Monte Carlo studies.**
When $\gamma^2 > 0$, the hidden confounding variable $C$ induces non-causal correlation between traits $A$ and $B$ without being observed. The $\omega$ parameter gives the path coefficient for the true direct causal flow from $B \rightarrow A$. The strength of the impact of the multiple causal anchors shown influencing $A$ and $B$ here is measured by the restricted heritabilities $h'_A$ and $h'_B$. The percentage of the variance of $A$ that is due to the causal anchors $\{M_A^{(i)}\}$ and not due to influence from any other variables shown is the restricted heritability $h'_A$. Restricted heritability gives the genetic-other influence ratio for $A$ while excluding any genetic causal flow from $B$, and allows us fine grained control over the simulation. Likewise the restricted heritability $h'_B$ describes the heritability of $B$ independent of $A$. Under this model, $h_B = h'_B$, but in general this may not be true.
CHAPTER 10

LEO Results

In the first results section here, we evaluate and compare our algorithms against existing methods. By necessity, these evaluations utilize Monte Carlo simulation. When we know the underlying truth or state of nature, we can then characterize the sensitivity and specificity of the competing approaches. We show the advantage of the orthogonal anchors, and then how the competing edge orienting algorithms trade off statistical power and false positive rates. We also present the results of a study on the robustness of anchor selection when the unsupervised SNP selection methods are employed. Finally, our results conclude with applications to the BxH data set. These studies recover known and novel genes in the cholesterol biosynthesis pathway.

10.1 Results: Edge Orienting Algorithms: Comparison for Sensitivity, Specificity, and Anchor Count Parameter Robustness

10.1.1 Single anchor/CPA versus OCA method

The advantage of using orthomarker models to detect causal association is illustrated by Monte Carlo study in Figures 10.1 and 10.2. At a fixed observed correlation, as the proportion of causal signal grows and the degree of confounding lessens in turn, the orthomarker based LEO.NB.OCA scores consistently show
more statistical power than the single anchor scores on the same data. As the signal to noise passes the 1:1 threshold at 50% causal signal, the power of the scores goes to 70% and then 80%, with the orthomarker scores showing a edge in power that is consistent and statistically significant at the higher heritability level (in Figure 10.2).
Figure 10.1: **Statistical power advantage of the LEO.NB score from utilizing orthogonal causal anchors.** Using Monte Carlo simulation, we determined LEO.NB significance thresholds and compare the power of orthomarker score (two markers here) versus single anchor scores. In 50 simulations of N=200 observations, the observed A-B correlation was held constant at 0.4. As the degree of confounding lessens and the true causal signal correspondingly increases, our power to detect the true interaction rises. The restricted heritabilities were set to $h_A^2 = h_B^2 = 0.2$. Bars show mean scores with whiskers showing the 95% confidence intervals for the mean rather than standard errors. At no causal signal (far left) the rate shown is the false positive rate. To ensure a fair comparison, we choose a threshold of 0.3 for the orthomarker LEO.NB,OCA score and the threshold for the single anchor LEO.NB was chosen to attain a comparable false positive rate. The thresholds here give mean false positive rates of less than 0.05. While both the single and multimarker method are successful at retrieving the signal for high proportion causal signal, the multimarker method which incorporates an unlinked marker shows a power advantage, especially when the causal signal exceeds the 50% threshold (at 1:1 signal-to-noise ratio).
Figure 10.2: **Statistical power advantage of the LEO.NB score from utilizing orthogonal causal anchors.** Using Monte Carlo simulation, we determined LEO.NB significance thresholds and compare the power of orthomarker score (two markers here) versus single anchor scores. In 50 simulations of N=200 observations, the observed A-B correlation was held constant at 0.4. As the degree of confounding lessens and the true causal signal correspondingly increases, our power to detect the true interaction rises. Here the restricted heritabilities were set to \( h_A^2 = h_B^2 = 0.6 \). Bars show mean scores with whiskers showing the 95% confidence intervals for the mean rather than standard errors. At no causal signal (far left) the rate shown is the false positive rate. To ensure a fair comparison, we choose a threshold of 0.3 for the orthomarker LEO.NB,OCA score and the threshold for the single anchor LEO.NB was chosen to attain a comparable false positive rate. The thresholds here give mean false positive rates of less than 0.05. While both the single and multimarker method are successful at retrieving the signal for high proportion causal signal, the multimarker method which incorporates an unlinked marker shows a power advantage, especially when the causal signal exceeds the 50% threshold (at 1:1 signal-to-noise ratio).
10.1.2 False positive and power comparisons

We studied by Monte Carlo simulation the tradeoff between false positive rates and power for three different scores using a single loci trait. Tables 10.1, 10.2, 10.3, and 10.4 summarize the results from this study in which we compare the LEO.NB.OCA to the LEO.NB.CPA, LEO.NB.ALL, and the MAX.MAX methods. In these simulations, a single causal signal SNP influences trait A, and likewise a single causal signal SNP influences trait B. We simulated \( h_A' = h_B' = 0.4 \) with an observed 100% casual correlation of 0.3; and \( \text{N}=200 \) for each replicate, with 300 replicates generated. The LEO.NB.OCA method shows power of 90-94% (± 0.99% SE) at a false positive rate estimate of 0.41% (± 0.078% SE) at the recommended 0.3 threshold. The improved LEO.NB.CPA method shows somewhat less power (83-86%) (± 1.5% SE) at a somewhat higher false positive rate of 2.7% (± 0.40% SE). Using a 4.53 ALL threshold which keeps the false positive rate for the LEO.NB.ALL score down to 5%, we then have 53-60% (± 2.3% SE) power. Using a threshold of 4.21, the MAX.MAX method has 28-33% (± 2.1% SE) power at its 5% false positive rate threshold. These distributions are depicted in boxplots at the first SNP in the titration Figures 10.3, 10.4, 10.5, and 10.6 respectively.

An important comparison between LEO.NB.OCA and LEO.NB.CPA is drawn using Tables 10.1 and 10.2. At the recommended 0.3 significance threshold (RT or Recommended Threshold column), we obtain a LEO.NB.CPA false positive rate of 0.027, somewhat higher (six-fold higher) than for the LEO.NB.OCA method under the same circumstances. The power under the 50% confounded model is actually somewhat higher for the LEO.NB.CPA than the LEO.NB.OCA. However the power of the CPA under the fully causal model, while at an acceptable 83-86% is slightly lower than the LEO.NB.OCA power of 90-94% for the same data.
Continuing the analysis of Tables 10.1 and 10.2, we further note that if we compare the LEO.NB.CPA and the LEO.NB.OCA at the same false positive rate, a LEO.NB.CPA threshold of 1.552 would be needed to obtain the false positive rate of 0.0041 (± 0.00078 SE) shown by the LEO.NB.OCA. At this threshold LEO.NB.CPA would yield only 48% (± 2.3% SE) power on the 100% causal simulations. For comparable power, a LEO.NB.CPA threshold of -0.2 would obtain equivalent power to the LEO.NB.OCA at its 0.3 threshold, but would in turn bring the false positive rate of the LEO.NB.CPA to 0.05 (± 0.0066 SE), twelve-fold more false positives than obtained by LEO.NB.OCA.

While the conclusions of this analysis do assume normal score distributions, the distributions appeared normal for each score by visual inspection of quantile-quantile plots and showed non-significant p-value under the Shapiro-Wilk test for normality. In the case of Table 10.1, three extremely low LEO.NB.OCA scores under the null distribution were conservatively discarded to obtain normality of the null. Individual p-values for the Shapiro-Wilk test are given in their respective figure legends.

From these data we conclude that the LEO.NB.OCA offers simultaneously better false positive rate and better power than the alternatives studied, when orthogonal causal anchors are available.
Figure 10.3: **SNP selection robustness titration for the LEO.NB.OCA scores under a 100% confounded Null model.** This study highlights a strength of the LEO.NB.OCA scores: robustness in the face of confounded correlation. When association is due to a shared yet unobserved or hidden causal parent ($M_A \rightarrow A \leftarrow H \rightarrow B \leftarrow M_B$) rather than true $A \rightarrow B$ or $B \rightarrow A$ causal flow, the LEO.NB.OCA score—which compares against a local SEM model for confounding—remains non-significant and strongly negative. The titration study here shows that even when significant QTL are present for $A$ and $B$ individually, if the $A - B$ association is due to confounding then the LEO.NB.OCA scores stay substantially below their significance threshold at the dashed line of 0.3. Here $h_A' = h_B' = 0.4; \gamma^2 = 0.3; \omega = 0$. One-hundred simulations were performed at $N=200$ samples; forward SNP selection plotted (the other selections methods were similar). Default parameter R (version 2.4.1) boxplots are shown in each of this series of Null distribution boxplots: boxes show the inter-quartile range (IQR) with the horizontal line at the median. (Notches show +/-1.58 IQR/$\sqrt{N}$, approximating a 95% confidence interval for the difference in two medians). Whiskers extend to the most extreme data point which is no more than 1.5 times the length of the box away from the box. Outliers beyond the whiskers are plotted individually as circles.
Figure 10.4: **SNP selection robustness titration for the LEO.NB.CPA scores under a 100% confounded Null model.** Under the same simulation parameters as Figure 10.3, the null distribution of the LEO.NB.CPA score is shown. As in Table 10.2, the recommended threshold of 0.3 is shown as the dashed significance line. Under this threshold, the false positive rate for one noise SNP is 0.027, rising slightly and levelling off at 0.04 at eight noise SNPs.
Figure 10.5: **SNP selection robustness titration for the LEO.NB.ALL scores under a 100% confounded Null model.** Under the same simulation parameters as Figure 10.3, the null distribution of the LEO.NB.ALL score is shown. The variation at the single SNP point around the near-zero mean pushes the dashed 5% false-positive line to 4.53. Although a lower false positive threshold could be utilized at 5 or so SNPs, Figures 10.9 and 10.13 show that statistical detection power would diminish dramatically in turn.
Figure 10.6: **SNP selection robustness titration for the MAX.MAX scores under a 100% confounded Null model.** Under the same simulation parameters as Figure 10.3, the null distribution of the MAX.MAX score is shown. The variation of the score around zero requires a 4.21 significance threshold to obtain a 5% false positive rate.

For fairness of comparison, we also studied the situation in which no orthogonal causal anchors were available into the most downstream node. This is the scenario in which the CPA model is a perfect fit and the OCA model is forced to search for the best noise SNP it can to fill the OCA role. We use the recommended threshold of 0.3 and simulate under the same heritability and parameter
Table 10.1: Orthomarker model (LEO.NB.OCA) recommended score thresholds: false positive and power evaluation. Here we summarize single SNP/node statistics for single SNP/node genetic models. In silico results from 300 Monte Carlo simulations are shown, with sample size \( N = 200 \) in each simulation. Restricted heritabilities were \( h_A' = h_B' = 0.4 \) and observed \( A - B \) correlation = 0.3. The 5\% false positive threshold reported is the mean score (under the null hypothesis of completely confounded association; \( \gamma^2 = 0.3 \)) plus 1.644 standard deviations. The completely confounded null model (top line, A) corresponds to having independent SNPs associated with both \( A \) and \( B \) nodes. The recommended threshold (RT) of 0.3 for LEO.NB.OCA is substantially more conservative than the 5\% false positive rate, and secures a two-fold or greater probability ratio (\( \log_{10} 2 = 0.3 \)) for the oriented causal edge when compared against several realistic alternative models. In each simulation 100 background noise SNPs were included. While the false positive rate and power computations assume a normal distribution, LEO.NB.OCA score distributions appeared normal by quantile-quantile plot and by Shapiro-Wilk normality test (p-value 0.28) after removing the bottom 3 observations (1\% of the 300 scores). [Abbreviations: standard deviation (sd), Recommended Threshold (RT), Standard Error (SE), false positive threshold (fpt).]
<table>
<thead>
<tr>
<th>Key</th>
<th>% causal signal</th>
<th>LEO.NB.CPA Mean (SE)</th>
<th>sd</th>
<th>RT</th>
<th>5% fpt</th>
<th>Power at RT (SE)</th>
<th>False pos rate at RT (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0%</td>
<td>-3.09 (0.11)</td>
<td>1.76</td>
<td>0.3</td>
<td>-0.2</td>
<td>0.38 (0.023)</td>
<td>0.027 (0.004)</td>
</tr>
<tr>
<td>B</td>
<td>50%</td>
<td>-0.19 (0.1)</td>
<td>1.65</td>
<td>0.3</td>
<td>0.83</td>
<td>0.86 (0.013)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>100%</td>
<td>1.48 (0.07)</td>
<td>1.22</td>
<td>0.3</td>
<td>0.83</td>
<td>0.86 (0.013)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>100%</td>
<td>1.50 (0.06)</td>
<td>1.12</td>
<td>0.3</td>
<td>0.83</td>
<td>0.86 (0.013)</td>
<td></td>
</tr>
</tbody>
</table>

Table 10.2: **Common Pleiotropic Anchor (LEO.NB.CPA) recommended score thresholds: false positive and power evaluation.** Using the same single SNP/node simulations as in Table 10.1, we study the power and false positive rates in the CPA method using forward stepwise SNP selection (output also known as the LEO.MULTIONE.NB score in our code). Here we also suggest a more conservative threshold of 0.3 than is strictly necessary to obtain 5% false positives (the 5% false positive threshold or fpt column). At the recommended 0.3 significance threshold (RT or Recommended Threshold column), we obtain a false positive rate of 0.027, somewhat higher (six-fold higher) than the LEO.NB.OCA method. The power under the 50% confounded model is actually somewhat higher than the LEO.NB.OCA. However the power of the CPA under the fully causal model, while at an acceptable 83-86% is slightly lower than the LEO.NB.OCA power of 90-94% for the same data. To compare the LEO.NB.CPA and the LEO.NB.OCA at the same false positive rate, a LEO.NB.CPA threshold of 1.552 would be needed to obtain the false positive rate of 0.0041 (± 0.00078 SE). At this threshold LEO.NB.CPA would yield only 48% (± 2.3% SE) power on the 100% causal simulations. For comparable power, a LEO.NB.CPA threshold of -0.2 would obtain equivalent power to the LEO.NB.OCA at its 0.3 threshold, but would in turn bring the false positive rate of the LEO.NB.CPA to 0.05 (± 0.0066 SE), twelve-fold more false positives than obtained by LEO.NB.OCA. While the false positive rate and power computations assume a normal distribution, LEO.NB.CPA score distributions appeared normal by quantile-quantile plot and by Shapiro-Wilk normality test (p-value 0.78). [Abbreviations: standard deviation (sd), Recommended Threshold (RT), Standard Error (SE), false positive threshold (fpt).]
### Table 10.3: Averaging single anchor models LEO.NB.ALL recommended score thresholds: false positive and power evaluation.

Here we summarize single SNP/node LEO.NB.ALL statistics for single SNP/node genetic models. Restricted heritabilities were $h_A' = h_B' = 0.4$ and observed $A - B$ correlation = 0.3. The same simulations as in Table 10.1–in silico results from 300 Monte Carlo simulations, with sample size $N = 200$ in each simulation–are evaluated and the ALL method results shown. The 5% false positive threshold reported is the mean score (under the null hypothesis of completely confounded association; $\gamma^2 = 0.3$) plus 1.644 standard deviations. While the false positive rate and power computations assume normal a distribution, LEO.NB.ALL score distributions appeared normal by quantile-quantile plot and by Shapiro-Wilk normality test (p-value 0.45). [Abbreviations: standard deviation (sd), Recommended Threshold (RT), Standard Error (SE), false positive threshold (fpt).]

<table>
<thead>
<tr>
<th>Key</th>
<th>% causal signal</th>
<th>LEO.NB.ALL Mean (SE)</th>
<th>sd</th>
<th>RT</th>
<th>5% fpt</th>
<th>Power at RT (SE)</th>
<th>False pos rate at RT (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0%</td>
<td>0.03 (0.18)</td>
<td>2.74</td>
<td>4.53</td>
<td>4.53</td>
<td>0.28 (0.020)</td>
<td>0.05 (0.0066)</td>
</tr>
<tr>
<td>B</td>
<td>50%</td>
<td>3.02 (0.15)</td>
<td>2.55</td>
<td>4.53</td>
<td></td>
<td>0.53 (0.023)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>100%</td>
<td>4.71 (0.12)</td>
<td>2.06</td>
<td>4.53</td>
<td></td>
<td>0.60 (0.022)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>100%</td>
<td>5.01 (0.11)</td>
<td>1.88</td>
<td>4.53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 10.4: Best single anchor model 1 versus best single anchor model 2 MAX.MAX recommended score thresholds: false positive and power evaluation.

Here we summarize the MAX.MAX statistics for single SNP/node genetic models. The restricted heritabilities were $h_A' = h_B' = 0.4$ and observed $A - B$ correlation = 0.3. In silico results from 300 Monte Carlo simulations, with sample size $N = 200$ in each simulation are shown. The 5% false positive threshold reported is the mean score (under the null hypothesis of completely confounded association; $\gamma^2 = 0.3$) plus 1.644 standard deviations. While the false positive rate and power computations assume a normal distribution, the MAX.MAX score distributions appeared normal by quantile-quantile plot and by Shapiro-Wilk test for normality (p-value of 0.79). [Abbreviations: standard deviation (sd), Recommended Threshold (RT), Standard Error (SE), false positive threshold (fpt).]

<table>
<thead>
<tr>
<th>Key</th>
<th>% causal signal</th>
<th>MAX.MAX Mean (SE)</th>
<th>sd</th>
<th>RT</th>
<th>5% fpt</th>
<th>Power at RT (SE)</th>
<th>False pos rate at RT (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0%</td>
<td>0.03 (0.16)</td>
<td>2.54</td>
<td>4.21</td>
<td>4.21</td>
<td>0.19 (0.016)</td>
<td>0.05 (0.007)</td>
</tr>
<tr>
<td>B</td>
<td>50%</td>
<td>2.41 (0.12)</td>
<td>2.04</td>
<td>4.21</td>
<td></td>
<td>0.28 (0.020)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>100%</td>
<td>3.24 (0.10)</td>
<td>1.69</td>
<td>4.21</td>
<td></td>
<td>0.33 (0.021)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>100%</td>
<td>3.50 (0.09)</td>
<td>1.63</td>
<td>4.21</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

90
<table>
<thead>
<tr>
<th>Key</th>
<th>shorthand</th>
<th>Description of Monte Carlo study for directional $B \rightarrow A$ edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1a.1b.0causal</td>
<td>One signal SNP into A, one signal SNP into B; A-B association completely confounded</td>
</tr>
<tr>
<td>B</td>
<td>1a.1b.50</td>
<td>One signal SNP into A, one signal SNP into B; A-B association 50% confounded</td>
</tr>
<tr>
<td>C</td>
<td>1a.1b.100</td>
<td>One signal SNP into A, one signal SNP into B; A-B association 100% causal</td>
</tr>
<tr>
<td>D</td>
<td>1a.1b.100.4an.4bn</td>
<td>One signal SNP into A, one signal SNP into B; A-B association 100% causal; four neighboring SNPs into A and four into B. Neighbors had correlation with the primary signal SNPs sampled from $[0.4, 0.7]$.</td>
</tr>
</tbody>
</table>

Table 10.5: Descriptions of the four simulation studies reported in Tables 10.1, 10.2, 10.3 and 10.4. Each of these simulations had a single SNP for the A and a single SNP into the B node.

conditions above. Under these conditions we find that the LEO.NB.OCA has a false positive rate of $0.31\%$ ($\pm 0.055\%$ SE), while the improved LEO.NB.CPA has a false positive rate of $0.18\%$ ($\pm 0.034\%$ SE). The power of the improved LEO.NB.CPA under the fully causal model is $95\%$ ($\pm 0.59\%$ SE), and the power of the LEO.NB.OCA is $93\%$ ($\pm 0.77\%$ SE). Our NEO implementation reports both scores on the NEO results spreadsheet, so that the analyst may inspect both scores during a single run.

10.1.3 Robustness to SNP count parameter and handling of multiple SNP traits

In Figures 10.7, 10.8, 10.9, and 10.10 we compare the OCA to the CPA, ALL, and MAX.MAX algorithms by testing them on a data set that simulates a four SNP trait with restricted heritabilities $h_A^4 = h_B^4 = 0.4$ and observed correlation 0.3. The LEO.NB.OCA score, using the forward stepwise regression SNP aggregation with AIC halting criteria shows several desirable qualities. We see (1) a significant
score at the single SNP level, (2) a score that increases as we add signal SNPs to the model, and (3) a resistance to over-fitting as extraneous noise SNPs are made available. The LEO.NB.ALL and the MAX.MAX methods in contrast do not attain significant scores for this four SNP trait, and do not show corresponding improvement or robustness as the SNP selection count parameter is increased. While other authors use manual SNP selection with their CPA strategy (an option with our NEO package), we use the same forward stepwise selection method in both the LEO.NB.CPA and LEO.NB.OCA to study in isolation the effects of using a model which includes orthogonal causal anchors.
Figure 10.7: The LEO.NB.OCA score is robust against the addition of extraneous noise SNPs. This robustness study of the orthomarker LEO.NB scores used the OCA aggregation strategy on a four SNP trait generated under Monte Carlo simulation with $h'_A = h'_B = 0.4$. The observed correlation was 0.3 ($\omega = 0.3$) at 100% causal signal (no confounding; $\gamma = 0$). At $N=200$ we simulated 100 replicates simulated for each of the three SNP selection methods. The plot shows both the peak in signal as the 3rd and 4th contributing signal SNPs are added to the model, and addition shows that the LEO.NB.OCA algorithm is robust to the addition of extraneous noise SNPs, as the scores stay above the dashed significance threshold even beyond the addition of the fourth causal signal SNP. Although addition noise SNPs are added and their addition does depress the score somewhat, the LEO.NB.OCA score remains significant. One hundred simulations for each method with $N=200$ were run with independent simulations for each of the SNP selection methods. Error bars show the standard error of the mean. The SNP selection strategies were tested on independent simulation runs. The SNP three selection methods compared do not show statistically significant differences from each other over the SNP titration shown here.
Figure 10.8: **SNP selection robustness titration for the LEO.NB.CPA scores for a 4 SNP trait.** For the same simulated data as Figure 10.7, the CPA method using forward-stepwise SNP model building shows good sensitivity and robustness in the face of additional noise SNPs.
Figure 10.9: SNP selection robustness titration for the LEO.NB.ALL scores for a 4 SNP trait. The ALL aggregation method does not reach the 5% false positive (dashed line) threshold when the trait is controlled by four SNPs each with equal contribution. As before, the restrictedheritabilities were $h'_A = h'_B = 0.4$; the observed correlation was 0.3 ($\omega = 0.3$) at 100% causal signal (no confounding or $\gamma = 0$). At $N=200$ we simulated 100 replicates simulated for each of the three SNP selection methods.
Figure 10.10: **SNP selection robustness titration for the MAX.MAX scores for a 4 SNP trait.** The MAX.MAX method is not effective when a trait controlled by four SNPs is the subject of the study. The dashed line shows the significance threshold required to keep false positives at 5%. As before, the restricted heritabilities were \( h_A' = h_B' = 0.4 \); the observed correlation was 0.3 (\( \omega = 0.3 \)) at 100% causal signal (no confounding or \( \gamma = 0 \)). At \( N=200 \) we simulated 100 replicates simulated for each of the three SNP selection methods.
We implement two improvements to the basic CPA approach to make it more competitive before comparing it to the LEO.NB.OCA models which use OCA. Although manual SNP selection is also used, here we use forward stepwise selection (also used by Schadt et al. (SLY05)) to select the markers put into CPA. This ensures that modifier alleles are found and incorporated when available and lessens model-fitting problems due to multicollinearity. It also isolates the comparison of the models with and without orthogonal causal anchors, so as to specifically evaluate the impact of using OCA. Second, we apply the marker assignment consistency principle to the markers chosen for $M_A$ and $M_B$, improving the chances up-front of the CPA approach locating markers that will show good model fit. Users of our R-based NEO implementation have this improved CPA approach available to them under the LEO.NB.CPA column in the NEO spreadsheet output.

In Figures 10.11, 10.12, 10.13, and 10.14 we compare the same four strategies when the traits have the same heritabilities, but are effected each by a single locus. We include in this simulation neighboring SNPs that recreate the phenomenon of LOD peak shoulders in that the neighbor are correlated with the main signal SNPs, albeit less strongly correlated with the trait itself. The LEO.NB.OCA method is again significant and robust against extraneous SNP addition. The LEO.NB.ALL method is significant for the first and only signal SNP, but then fails to be robust against the addition of the neighboring SNPs. The MAX.MAX method does not reach its significance threshold even for a single SNP.
Figure 10.11: **SNP selection robustness titration for the LEO.NB.OCA scores for a single SNP trait.** The LEO.NB.OCA score remains significant at the recommended 0.3 threshold level even when neighboring LOD peak SNPs and other noise SNPs are added to the model. As before, the restricted heritabilities were $h_A' = h_B' = 0.4$; the observed correlation was 0.3 ($\omega = 0.3$) at 100% causal signal (no confounding or $\gamma = 0$). At N=200 we simulated 100 replicates simulated for each of the three SNP selection methods. To simulate the effect of LOD peak shoulders, four correlated neighbor SNPs surrounding the signal SNPs $M_A$ and $M_B$ respectively were added to the pool of measured markers. Each of these neighboring SNPs had a correlation sampled uniformly from $[0.4, 0.7]$ with their respective primary causal SNPs.
Figure 10.12: **SNP selection robustness titration for the LEO.NB.CPA scores for a single SNP trait.** The common pleiotropic anchor (CPA) approach using the automated SNP selection (forward-stepwise regression with AIC halting criteria, and application of the marker assignment consistency principle—see Definition 1) approach also shows robust signal detection. Simulation parameters were the same as in Figure 10.11.
Figure 10.13: SNP Selection robustness titration for the LEO.NB.ALL scores for a single SNP trait. The ALL aggregation method is not robust against the addition of neighboring LOD peak SNPs and noise SNPs; it is significant for a single SNP, but falls below the significance line upon signal dilution by weaker LOD peak SNPs. Simulation parameters were the same as in Figure 10.11.
Figure 10.14: SNP selection robustness titration of the MAX.MAX scores for a single SNP trait. The MAX.MAX edge orienting method fails to attain a significant score above its 5% false positive threshold at any point under a strong single SNP trait; in particular significance is not obtained even at the single SNP selection point. Simulation parameters were the same as in Figure 10.11.
We now return to the earlier depictions of the null distribution simulations to flesh out the additional detailed results provided by Figures 10.3, 10.4, 10.5, and 10.6. These figures show the null distribution of the scores under the scenario of 0% causal signal. These findings extend the results of Tables 10.1 - 10.4 by varying the SNP selection count parameter to NEO. The observed $A - B$ correlation in this model is due to confounding by an unobserved common causal parent. The LEO.NB.OCA recommended threshold of 0.3 requires (Figure 10.3) strong evidence at the single SNP level, and remains a conservative significance threshold even as the median null score rises from -4.75 to an asymptotic value of -1.75. The LEO.NB.ALL (Figure 10.5) shows a median score under the null that hovers around zero. The standard deviation of the ALL method does shrink as SNPs are added, however power suffers correspondingly, and given the power advantage of the OCA method, we conclude that the ALL method should not be employed for routine use. Similar to the ALL method, the MAX.MAX scores show very broad standard deviations for parts of the SNP count parameter titration, again requiring a high significance threshold under which power suffers greatly.

10.1.4 Conclusions about algorithm performance

In conclusion, the single signal marker results in Tables 10.1 - 10.4 show that when the false positive rate is kept at or below 5%, the LEO.NB.OCA score shows superior power to the other strategies. Moreover at the recommended 0.3 significance threshold–where the probability of model 1 is two-fold ($10^{0.3} = 2$) higher than any competing model–the LEO.NB.OCA score still obtains a false positive rate below 1% while simultaneously obtaining statistical power at or above 90%. These Monte Carlo simulation results used an aggressive null hypothesis in that there was observed the same degree of correlation between nodes $A$ and $B$, and between the marker into $A$ and the marker into $B$, however the
A – B correlation was entirely due to an unobserved confounder.

10.2 Automated SNP selection methods: comparing Greedy, Forward, and Both-Forward-and-Greedy

In Figures 10.15(a)-(f) illustrate the impact of LOD peak shoulders on the automated SNP selection algorithms. These plots show how the three selection methods compare as the number SNPs selected grows. In this study we simulated a multiple-marker complex disease model with |$M_A|$ = |$M_B$| = 5 and equal and additive influence of each of the five SNPs on their respective traits. Restricted heritabilities were set to $h'_A = h'_B = 0.5$; no confounding and an observed correlation of 0.4 was simulated. Figure 10.15(a) shows the situation with noise SNPs but no neighboring SNPs correlated with the main five signal SNPs. Figure 10.15(b) illustrates performance when each of the five signal SNPs has neighbors that would provide the shoulders of any QTL LOD peak. That is, for each signal SNP, six surrounding SNPs, each with correlation sampled from (0.75, 0.95), were simulated and added to the SNP data set supplied to the NEO scoring algorithms. Figures 10.15(c) and 10.15(d) show the same no neighbors/neighbors comparison with increasing numbers of selected SNPs but for Greedy-only selection. And the last pair of figures, Figures 10.15(e) and 10.15(f), show the impact of gradually increasing the greedy and the forward-step selected SNPs. Neighboring SNPs dilute the signal as shown by the top four panels in which each right hand score is diluted relative to its left hand counterpart. The exception to this signal dilution is when the forward and greedy SNP selection methods are combined and LEO.NB.OCA is employed. This is illustrated in the lower two panels. Here the scores are robust to neighborhood noise. The persistence of signal beyond five SNPs on the horizontal axis also
demonstrates that the combined selection methodology with the LEO.NB.OCA score is relatively robust to the addition of unimportant noise SNPs, 100 of which were supplied (simulating the output of a genome scan) here for each simulation.

Several conclusions can be drawn from these results. First, in general the LEO.NB.OCA score tends to increase with increasing SNPs supplied, as illustrated by rising scores with additional SNPs. This is consistent with the OCA aggregation strategy of utilizing the orthogonal information supplied by multiple SNPs about the complex phenotype. The LEO.NB.ALL score tends to fall with increasing numbers of SNPs, indicating that the additional causal signal is going unutilized.

However even the LEO.NB.OCA score is mildly susceptible to the decoy effect of the neighboring SNPs that lowers the score if only Greedy SNPs are chosen, as illustrated in Figures 10.15(d). Fortunately this effect is ameliorated by the use of both greedy and forward-step chosen SNPs, as illustrated by the stabilized and high LEO.NB.OCA scores in Figure 10.15(f). Since neighboring SNPs are increasingly common as whole genome SNP data comes online, our study shows that a strictly greedy approach should be supplemented by a forward-step approach to maximize one’s causal signal scores and detection power.

In summary we note that the studies in Figures 10.7 - 10.14 show rare disagreement between the three approaches in the signal SNP portion of the titration. We feel comfortable recommending the combined forward and greedy approach for application data analysis.
10.3 Speciality scores for diagnosis of confounding and acceleration of network edge orienting

10.3.1 BLV: Bilayer Verification Scores for detecting confounding

The Bilayer Verification theorem (Ate06; Ate08) provides theoretical foundation for the detection of confounded association. Here we extend the theoretical work with a practical score.

Specifically, the BLV score for and $A \rightarrow B$ causal direction is based on the relative strength of the correlation and partial correlation coefficients, as transformed by Fisher’s transform $F(\cdot)$ to Z-scores. Here $N$ is the number of observations and $\rho(a, b)$ is the correlation between $a$ and $b$, and $\rho(a, b|c)$ is the partial correlation between $a$ and $b$ given $c$.

$$BLV_{A\rightarrow B} = \left[ \sqrt{N - 3} \cdot F(|\rho(M_A, B)|) \right]$$

$$- (\sqrt{N - 4} \cdot F(|\rho(M_A, B|A)|))$$

(10.1)

$$BLV_{B\rightarrow A} = \left[ \sqrt{N - 3} \cdot F(|\rho(M_B, A)|) \right]$$

$$- (\sqrt{N - 4} \cdot F(|\rho(M_B, A|B)|))$$

(10.2)

Figure 10.16 demonstrates how the BLV score goes strongly negative as the correlation grows when that correlation is entirely due to confounding ($\gamma^2 = \rho_{AB}$).
10.3.2 ZEO: Z-score Edge Orienting for computational efficiency

When large numbers of nodes (more that 20 traits for instance) require evaluation, the computational time can be a concern. To address this, we developed a fast proxy score that can quickly discard edges not worth evaluating.

Since LEO score computation can be intensive, we further derive an efficiently computable proxy score that captures much of the LEO signal. It is derived from the model $M_A \rightarrow A \rightarrow B \leftarrow M_B$, where $M_A$ and $M_B$ are restricted here to single anchors. Specifically, the ZEO score is based on the relative strengths of the opposing BLV scores. The relevant component parts of the ZEO computation, as well as the BLV scores that are components of the ZEO, are defined as follows.

$$
ZEO_{A \rightarrow B} = \frac{1}{2}(BLV_{A \rightarrow B} - BLV_{B \rightarrow A})
$$

(10.3)

As illustrated by Figure 10.17, the ZEO score can be used to predict the LEO scores. The correlation between the two is in excess of 0.9 over 3 distinct causal signal-to-noise regimes. ZEO scores on our hardware and software are in excess of 100-fold faster to compute. Hence the ZEO score is used to quickly rule out edges that need not be evaluated by LEO.
Figure 10.15: The impact of LOD peak shoulders on the choice of SNP aggregation methods. Due to linkage, QTL peaks typically occur in a neighborhood of correlated markers. Each of these neighboring loci shows weaker correlation with the trait which might confuse the automated SNP selection process. Here we present Monte Carlo study results showing that the LEO.NB.OCA score in conjunction with forward and greedy combined SNP selection is immune to this concern. Simulation parameters were $h'_A = h'_B = 0.5$; $\omega = 0.4$; and $\gamma^2 = 0$.  

107
Negative BLV score implies confounding.
Monte Carlo study of Null: MA=>A; B independent.

Grid sampling as h'A and h'B step through [0.2,0.33,0.47,0.6].

BLV score = \[ Z(|\text{cor}(\text{MA},B)|) - Z(|\text{cor}(\text{MA},B|A)|) \]

Gamma-squared: correlation of A–B entirely due to confounding.

Figure 10.16: **Monte Carlo study of zero true signal and use of the BLV score for detecting false positives due to confounding.** When a quick check for confounding is desired, the Bilayer verification (BLV) scores can be consulted. The boxplot here shows the observed correlation when an A–B association is due entirely to confounding. The stronger the confounded correlation, the more strongly negative is the distribution of the BLV scores, and this conclusion holds as we grid-sample the heritability parameters from [0.2, 0.33, 0.47, 0.6]. We conclude that a negative BLV score serves as a strong flag to indicate that confounding may be a concern and the A–B association deserves further cautious investigation.
Figure 10.17: **Correspondence between LEO.NB and ZEO scores.** As shown here by Monte Carlo simulation, there is a high correlation (Pearson correlation 0.91, p-value = 2e-16) between the LEO.NB and the ZEO scores. The likelihood based LEO score is preferred, however the partial correlation based ZEO is computationally more efficient, and can be used to quickly screen out poor edges. Two marker MA, MB model was simulated using ha=hb=0.4; 100 simulations at each of three causal-signal-to-noise ratios; each model with N=200 individuals. Dashed lines give recommended LEO.NB = 0.3 and ZEO = 2 thresholds. As can be observed in the figure, this thresholds offer conservative delineation between the fully causal and the fully confounded edges.
10.4 Results: Application to data

10.4.1 BxH data set description

The data for these analysis are from a previously published BxH mouse F2 cross (CEZ05; WYS06; GDZ 5). The F2 intercross of C57BL/6J.ApoE--/- inbred strain with a C3H/HeJ.ApoE--/- inbred strain was examined. C57BL/6J is susceptible to a variety of atherosclerosis, diabetes, obesity, and heart disease related traits to which C3H is resistant. Particularly on the ApoE--/- background, the F2 mice are expected to show significant spectrum of atherosclerosis and metabolic syndrome responses to a high-fat diet. A total of 136 females and 129 males were studied. These were genotyped at 1278 loci across the mouse genome. A variety of clinical covariates were measured, including fat mass determination via NMR, aorta sectioning and measurement of aortic lesions, insulin, glucose, free fatty-acid levels in the blood, and cholesterol fractions (unesterified, HDL, and LDL+VLDL)(WYS06). Significantly, the liver, brain, adipose tissue and skeletal muscle tissue from each F2 was assayed via Agilent (HDS03; SMD03) spotted-oligo microarray, returning some 23K mRNA levels for each of the 4 tissues. Significant differences in the gene expression profiles between the male and female F2s were observed(WYS06), leading us to focus on the females. Of the available tissues, liver is a well understood tissue in that the literature documents the response of specific liver mRNA levels to insulin (MP06) and the known cholesterol biosynthesis pathways (Lus06). While these are just two of many known pathways in liver, the sterol regulatory pathway is of primary interest to our collaborators. Hence sterol regulation was chosen as the focus in our work to validate the NEO methodology by recovering known biology.
10.4.2 Cholesterol biosynthesis pathway recovery and implication of novel genes as downstream of \textit{Insig1}

We demonstrate our method by recovering known gene regulation relationships in the cholesterol biosynthesis pathway and implicating novel genes as being co-impacted by this pathway. Results are shown in Table 10.7. In particular we can see the relationship between \textit{Insig1} mRNA levels and the cholesterol synthetic pathway enzymes \textit{Fdlf1} and \textit{Dhcr7}, whose mRNA levels appear to be jointly downstream of \textit{Insig1}. \textit{Insig1} itself appears to be regulated by loci on chromosomes 8 and 16. The two upstream causal driver SNPs were modeled as \textit{Insig1} \( \sim \text{rs3677807.chr8.bp108220543 + rs3695481.chr16.bp68778145} \), and together had \( R^2 = 0.124 \) for \textit{Insig1}. This study suggests that \textit{Insig1} also regulates \textit{Tlcld1}, \textit{Eaf2}, \textit{Rdh11}, \textit{Slc25a44}, \textit{Slc23a1}, \textit{Slc25a1}, \textit{Galce}, and \textit{Qdpr}, and \textit{Frmd4b} in the mouse female liver. Literature searches for each of these genes revealed no previous knowledge of these particular genes in the liver, no placement downstream in the cholesterol biosynthesis pathway, nor co-regulation beside this pathway. They were therefore considered novel findings from this data set. Fourteen positive control genes, already known to be in the pathway and listed in Table 10.7, were also recovered.

To validate these results we then examined an independent test data set consisting of 129 male F2 mice from the same cross. Since the QTL appeared to differ by sex, we employed the automated modeling procedure available in NEO to give a blind evaluation of the causal paths between \textit{Insig1} and the downstream genes. The last two columns of Table 10.7 give the LEO.LB.OCA score from this blind automated model building and a validation (val†) column which contains a star (⋆) if the male LEO.NB.OCA is also positive. Of the novel genes suggested by the female analysis, four are replicated in the male liver: \textit{Tlcld1}, \textit{Slc25a44}, \textit{Slc23a1}, and \textit{Qdpr}. 

111
<table>
<thead>
<tr>
<th>no</th>
<th>edge</th>
<th>LEO. NB. OCA</th>
<th>( \rho )</th>
<th>path coef</th>
<th>path ( Z )</th>
<th>model prob</th>
<th>df</th>
<th>( \chi^2 )</th>
<th>RMSEA</th>
<th>BLV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs3705921 ( \rightarrow ) Insig1</td>
<td></td>
<td>0.22</td>
<td>0.18</td>
<td>2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>rs3670293 ( \rightarrow ) Insig1</td>
<td></td>
<td>-0.33</td>
<td>-0.31</td>
<td>-3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>rs3675054 ( \rightarrow ) Dhc7</td>
<td></td>
<td>-0.26</td>
<td>-0.15</td>
<td>-3.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Insig1 ( \rightarrow ) Dhc7</td>
<td>1.2</td>
<td>0.81</td>
<td>0.79</td>
<td>16.1</td>
<td>0.24</td>
<td>5</td>
<td>6.8</td>
<td>0.051</td>
<td>2.0</td>
</tr>
<tr>
<td>5</td>
<td>Insig1 ( \rightarrow ) Fdft1</td>
<td>1.4</td>
<td>0.67</td>
<td>0.64</td>
<td>10.7</td>
<td>0.75</td>
<td>5</td>
<td>2.7</td>
<td>0</td>
<td>2.0</td>
</tr>
<tr>
<td>6</td>
<td>rs3664397 ( \rightarrow ) Fdft1</td>
<td></td>
<td>0.34</td>
<td>0.27</td>
<td>4.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10.6: **Edge orienting report for the cholesterol biosynthesis pathway positive controls—models depicted in Figure 10.18.** A condensed summary of the NEO spreadsheet is shown here. The LEO.NB.OCA scores for the well established biology of \( \text{Insig1} \rightarrow \{ \text{Dhc7, Fdft1} \} \) gene regulation show a large margin of significance beyond their 0.3 threshold, with LEO.NB.OCA of 1.2 indicating that the \( \text{Insig1} \rightarrow \text{Fdft1} \) edge is \( 10^{1.2} \) = 15 times more likely than the next best edge orientation model, and the LEO.NB.OCA of 1.4 indicating that the \( \text{Insig1} \rightarrow \text{Fdft1} \) edge is \( 10^{1.4} \) = 25 times more likely than the next best edge orientation. The \( \rho \) column gives the marginal Pearson correlation coefficient, while the three path columns (standardized path coefficient, asymptotic Standard Error, and Z-score for the edge) give details of each individual edge in the SEM. The last five columns summarize the fits of the two best fitting SEM shown in Figures 10.18(a) and 10.18(b). The model probability column is defined in equation 8.4 and is based on the model degrees-of-freedom and the \( \chi^2 \) statistic. The Root Mean Square Error of Approximation (RMSEA) is a standard SEM fit evaluation index that, similar to the \( \chi^2 \) statistic, evaluates the overall fit of the model; a value of between 0 and 0.1 is desirable. These and related model fitting indices are reviewed in Appendix C. The Bilayer Verification (BLV) score checks for confounded paths and a positive score is desirable.
<table>
<thead>
<tr>
<th>Edge: Insig1</th>
<th>LEO, N.B. female</th>
<th>model prob</th>
<th>path coef</th>
<th>path pval</th>
<th>df</th>
<th>( \chi^2 )</th>
<th>BLV</th>
<th>+ control/ novel (a.k.a.)</th>
<th>LEO, N.B. OCA male</th>
<th>val†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fdrl1</td>
<td>1.4</td>
<td>.75</td>
<td>.64</td>
<td>△</td>
<td>5</td>
<td>2.7</td>
<td>2.0</td>
<td>+</td>
<td>0.2</td>
<td>*</td>
</tr>
<tr>
<td>Dicer7</td>
<td>1.2</td>
<td>.24</td>
<td>.79</td>
<td>△</td>
<td>5</td>
<td>6.8</td>
<td>1.4</td>
<td>+</td>
<td>1.9</td>
<td>*</td>
</tr>
<tr>
<td>Scol</td>
<td>1.2</td>
<td>.58</td>
<td>.63</td>
<td>△</td>
<td>5</td>
<td>3.8</td>
<td>1.7</td>
<td>+</td>
<td>0.4</td>
<td>*</td>
</tr>
<tr>
<td>Sc4mol</td>
<td>1.1</td>
<td>.35</td>
<td>.68</td>
<td>△</td>
<td>5</td>
<td>5.6</td>
<td>1.9</td>
<td>+</td>
<td>0.5</td>
<td>*</td>
</tr>
<tr>
<td>0010050501Rk</td>
<td>.901</td>
<td>.82</td>
<td>.67</td>
<td>△</td>
<td>5</td>
<td>2.2</td>
<td>1.8</td>
<td>novel (Tlc1)</td>
<td>2.4</td>
<td>*</td>
</tr>
<tr>
<td>Fads2</td>
<td>1.0</td>
<td>.64</td>
<td>.61</td>
<td>△</td>
<td>5</td>
<td>3.4</td>
<td>1.7</td>
<td>+</td>
<td>.02</td>
<td>*</td>
</tr>
<tr>
<td>Adipor2</td>
<td>.97</td>
<td>.98</td>
<td>.61</td>
<td>△</td>
<td>5</td>
<td>.73</td>
<td>1.8</td>
<td>+</td>
<td>-2.0</td>
<td></td>
</tr>
<tr>
<td>Fasn</td>
<td>.96</td>
<td>.72</td>
<td>.77</td>
<td>△</td>
<td>5</td>
<td>2.9</td>
<td>1.4</td>
<td>+</td>
<td>1.0</td>
<td>*</td>
</tr>
<tr>
<td>Eaf2</td>
<td>.89</td>
<td>.16</td>
<td>.54</td>
<td>6e-16</td>
<td>5</td>
<td>8</td>
<td>0.9</td>
<td>novel, Eaf2</td>
<td>-0.5</td>
<td></td>
</tr>
<tr>
<td>Stared4</td>
<td>.87</td>
<td>.46</td>
<td>.59</td>
<td>△</td>
<td>5</td>
<td>4.6</td>
<td>1.8</td>
<td>+</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Fads1</td>
<td>.86</td>
<td>.82</td>
<td>.73</td>
<td>△</td>
<td>5</td>
<td>2.2</td>
<td>1.3</td>
<td>+</td>
<td>-0.5</td>
<td></td>
</tr>
<tr>
<td>Dlat</td>
<td>.84</td>
<td>.81</td>
<td>.58</td>
<td>△</td>
<td>5</td>
<td>2.3</td>
<td>1.8</td>
<td>+</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Rhdl1</td>
<td>.82</td>
<td>.80</td>
<td>.73</td>
<td>△</td>
<td>5</td>
<td>2.3</td>
<td>1.0</td>
<td>novel, Rhdl1</td>
<td>-0.7</td>
<td></td>
</tr>
<tr>
<td>b170110c055Rk</td>
<td>.813</td>
<td>.87</td>
<td>.52</td>
<td>2e-13</td>
<td>5</td>
<td>1.8</td>
<td>1.6</td>
<td>novel (Scl23a44)</td>
<td>0.7</td>
<td>*</td>
</tr>
<tr>
<td>Apoc8</td>
<td>.72</td>
<td>.61</td>
<td>.59</td>
<td>△</td>
<td>5</td>
<td>3.6</td>
<td>1.4</td>
<td>+</td>
<td>-1.6</td>
<td></td>
</tr>
<tr>
<td>Scl23a1</td>
<td>.63</td>
<td>.58</td>
<td>.49</td>
<td>1e-11</td>
<td>5</td>
<td>3.8</td>
<td>1.0</td>
<td>novel, Scl23a1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Scl25a1</td>
<td>.63</td>
<td>.37</td>
<td>.65</td>
<td>△</td>
<td>5</td>
<td>5.4</td>
<td>1.8</td>
<td>novel, Scl25a1</td>
<td>-0.3</td>
<td></td>
</tr>
<tr>
<td>Acac</td>
<td>.59</td>
<td>.64</td>
<td>.73</td>
<td>△</td>
<td>5</td>
<td>3.4</td>
<td>0.9</td>
<td>+</td>
<td>-0.4</td>
<td></td>
</tr>
<tr>
<td>Acas2</td>
<td>.58</td>
<td>.19</td>
<td>.64</td>
<td>△</td>
<td>5</td>
<td>7.4</td>
<td>1.8</td>
<td>+</td>
<td>-2.5</td>
<td></td>
</tr>
<tr>
<td>Gale</td>
<td>.53</td>
<td>.65</td>
<td>.58</td>
<td>△</td>
<td>5</td>
<td>3.3</td>
<td>1.5</td>
<td>novel, Gale</td>
<td>-0.3</td>
<td></td>
</tr>
<tr>
<td>Med1</td>
<td>.38</td>
<td>.60</td>
<td>.59</td>
<td>△</td>
<td>5</td>
<td>3.7</td>
<td>0.9</td>
<td>+</td>
<td>0.6</td>
<td>*</td>
</tr>
<tr>
<td>Qdpr</td>
<td>.37</td>
<td>.74</td>
<td>.59</td>
<td>△</td>
<td>5</td>
<td>2.7</td>
<td>0.5</td>
<td>novel, Qdpr</td>
<td>0.4</td>
<td>*</td>
</tr>
<tr>
<td>000010c055Rk</td>
<td>.35</td>
<td>.70</td>
<td>.54</td>
<td>8e-15</td>
<td>5</td>
<td>3</td>
<td>1.5</td>
<td>novel (Frad4b)</td>
<td>-0.7</td>
<td></td>
</tr>
</tbody>
</table>

**Table 10.7:** **Fourteen positive control genes, and nine novel genes that appear downstream of Insig1 using NEO analysis.** This study shows that NEO analysis can reproduce known (positive control) genes that are known to be downstream of Insig1. Moreover, our analysis also implicates novel genes as being impacted by the same pathway. PubMed searches on these genes showed a lack of liver or sterol impacted gene expression reports available in the literature. As a caveat, these genes were implicated in Apoc-/- Female BxH mice and so may not replicate without the knockout of the Apoc gene. Nonetheless, since so many known cholesterol regulated positive controls are recovered simultaneously, these findings are expected to be robust. Using the 23388 genes on the array, and assuming that there are 200 known genes downstream of Insig1 (a conservative overestimate), we compute the Fisher exact test for the set of 9 predicted versus 14 known downstream genes giving a p-value of 1e-13 for the predicted novel gene set. The validation (val†) column shows a star (*) if the initial finding in the female liver was replicated in the independent male test set of 129 F2, in that a positive LEO.NB.OCA score was obtained using the automated model building and SNP selection. Of the nine novel genes suggested from the female analysis, the male liver analysis confirms four of these: Tlc1, Scl25a44, Scl23a1, and Qdpr. [P-value code: △ means the path p-value was < 10\(^{-20}\).]
10.5 Discussion

Figure 6.1 provides an overview of the edge orienting algorithm described here and also draws the distinction between earlier single anchor and common pleiotropic anchor (CPA) approaches and the orthogonal causal anchor (OCA) approach whose value is highlighted here.

Structural Equation Modeling (SEM) is the descendant of pioneering geneticist Sewall Wright’s path analysis, and since the addition of latent factors and maximum likelihood testing procedures introduced by Swedish psychometrician Karl Jöreskog in the early 1970s (J70), has become a widely used tool in the social sciences.

We utilize locally acyclic SEM in our Local-structure Edge Orienting (LEO) scores to assess the combined evidence for each possible directed network edge, and to assemble a medium scale pathway network that itself may not be acyclic. We have further proposed a computationally fast approximation of the LEO score, the Z-score edge orienting (ZEO) score to allow genome wide application of this methodology, and a Bilayer verification (BLV) score which detects spurious correlations due to confounding when an edge of particular interest is inspected.

We note that the influence of the causal anchor or SNP may be indirect. The variables through which the SNP are acting may or may not be included in the data set at hand, and our methods are robust with respect to variables on the causal path between SNP and the A and B nodes that are either not measured or not measurable. Furthermore, while SNP changes must be upstream of gene expression and phenotype manifestation, this is not to say that some SNPs may indeed modify the action of other SNPs, and the effect of such modifiers may become apparent only in particular contexts.

As noted in the introductory remarks, SEM based models are, by design,
linear models of interacting variables. The assumption of linearity is not always warranted, and non-parametric versions of these models can also be explored by replacing variables’ measurements with their ranks. Our NEO software currently implements this transformation as an option.

The definition of model probability herein is an operational definition, utilized as a practical device for making useful model comparisons during the process of network learning and edge orienting. We use the term model probability as a means to describe the components of LEO scores, and provide the user a reasonable intuition regarding the computations involved. It is also possible to utilize a Bayesian approach to the definition of model probability that incorporates a prior probability term for each model considered. Having no principled way to generate such priors, we excluded such an approach from our current work. The result is that each model considered here is currently given equal footing, in effect imposing an uninformative prior. Future work may explore imposing priors on the models, should methods, data, or persuasive approaches to prior distribution assignment become available. In fact we suggest that the LEO scores, which summarize the evidence for a particular edge, would make a rational basis upon which to articulate priors on a network, should additional data sets on the same network becomes available.

The frequently employed constraint of acyclicity in learning Bayesian networks, while highly desirable for interpretability and simplicity of causal interpretation, can be a double edged sword if local errors in edge orientation propagate through the rest of the network. Our approach relaxes the constraints imposed by global network acyclicity by deliberately discarding (during LEO scoring) the evidence from correlated trait neighbors in the undirected graph. Instead we weigh more significantly the causal anchors and the orthogonal causal anchors provided by the genetic variation seen in SNPs. We choose to fit only local edge acyclic
models (using possibly remote causal anchors) and to do so as accurately as possible. Only once all local models are fit would we consider whether a medium scale acyclic model is the appropriate next level of inference.

We find this approach is a more robust measure of directional edge strength, and that it resists the propagation of errors if the direction of a neighboring edge is miscalled. This containment of error comes at the price of not being able to guarantee an acyclic graph. Biological networks frequently employ feedback to maintain homeostasis. Rather than to impose the mathematical nicety of acyclicity on the observed graph, we prefer to allow the data to speak for the strength of each directed edge independently. Combined with the innovation of using orthogonal causal anchors, our edge orientation conclusions can provide a strong basis for further experimental validation of causal pathways.
Figure 10.18: **Orthomarker models reproduce known biology in the cholesterol biosynthesis pathway: Insig1 → Dhrc7 and Insig1 → Fdft1.** A total of 136 female BxH.Apoec−/− mice had liver gene expression profiled. Here we recover known cholesterol biosynthesis connections using previously unreported QTL for Insig1 on chromosomes 8 (rs3670293 with a LOD score of 3.5 by single marker QTL analysis) and chromosome 16 (LOD score 1.6 by single marker QTL analysis). A detailed report on the fit of the overall edge orientation models for the Insig1 → Dhrc7 (no. 4) edge and the Insig1 → Fdft1 (no. 5) edge is supplied in the NEO spreadsheet output, and is condensed and summarized in Table 10.6. The upper two panels show the full models used to obtain the LEO.NB.OCA score for the two edges. By single marker QTL analysis (c) neither of the two QTL peaks would have reached the 3.8 LOD score threshold (the 5% significance threshold). However these results suggest that our modeling approach may be used to rescue insignificant single marker QTL peaks by combining multiple smaller peaks in a model with previously validated biological relationships (edges 4 and 5 are well established) to discover novel pathway member genes (Table 10.7).
CHAPTER 11

Conclusion

This work has explored the modeling of genome scale data, especially mRNA expression data, using a number of organizing features. We have looked at algorithms with utilize no causal anchors at all and featured the RVL and RVV approaches in these cases. We have explored using single and multiple causal anchors in the form of genetic markers, and note the advantages of using multiple and orthogonal causal anchors in the LEO scoring algorithms and NEO methods. We have highlighted the power and specificity advantages obtained from these approaches, particularly when recovering networks in the biological setting where relaxing the constraint of network acyclicity may be most appropriate. The multivariate aggregation model building strategies shown in the Orthogonal Causal Anchor approaches have been applied to suggest novel genes in the cholesterol biosynthesis pathway in liver. We hope to have formulated our methods such that extensions to other sources of causal anchoring such as controlled and randomized experiments may be straightforward, but further domain specific modeling is usually appropriate and may be expected to improve one’s causal network. We hope these causal gene network hypotheses will contribute to the experimental testing of specific gene-trait relationships in follow-on laboratory validation work.
APPENDIX A

LEO Algorithm Additional Details

Here we give technical details for several of the LEO and ZEO algorithms such that the methods implemented in our software may be clearly discerned. While we do not document all of possible or available anchor aggregation methods, the details provided here should be sufficient that extrapolations and extensions are readily at hand.

In detail, NEO proceeds in several steps. First we align the data and designate specific variables as SNPs and specific variables as Traits. Second, we do preliminary SNP selection. We use the marginal association levels to select candidate SNPs from the background pool of SNPs without yet finalizing the choices of which SNPs will go into the models. Three methods: Greedy, Gram-Schmidt, and Agglomerative cutree centers are used for primary SNP selection. Third, SNPs and non-SNP variables with high correlation are merged into supernodes. The creation of supernodes prevents highly collinear variables from cancelling each others’ connections out during the subsequent filtering step. Fourth, we prune the candidate edges based on various filters, and do final SNP selection of those SNPs to be utilized in the final edge orienting models used for scoring and evaluation. In the fifth and final step, we carry out a consistency check and compute edge orientation scores on the network by multiple methods (RVV, ZEO, LEO). The choice here of which scores are computed depends to an extent on the number of SNPs chosen as connecting to each of the variables on either
end of an edge. Multiple methods can handle zero, one, two, and many incoming causally anchoring edges from SNP nodes, and the NEO software computes all available indices for each edge at hand. Lastly, the program prepares summaries and graphical output of the edge orienting scores, allowing the user to evaluate and iterate if desired.

A.0.1 Step One: Input

NEO takes as input a data frame of SNP and trait columns, over some sampled population. The NEO method models the relationships between any set of quantitative variables. The quantitative traits may correspond to individual genes or to module eigengenes, or to physiological quantitative trait measurements. Without loss of generality, we will use the language of gene expression networks and genetic markers as causal anchors.

To begin, we partition the observed variables $V$ in our data into two classes: SNPs or genetic markers in $M$, and non-SNP variables in $T$. Since $T$ can include microarray derived variables as well as clinical traits, we also designate by subset $T^* \subset T$ the specific trait or traits (typically a clinical phenotype such as aortic lesion scores) that forms the focus of the analysis, if such are present$^1$.

A.0.2 Step Two: Genetic Marker/SNP selection.

Given the millions of SNPs provided by SNP chips and high throughput genotyping platforms, we are faced with the challenge of selecting which SNPs from those measured to use as causal anchors in edge orienting. The task of SNP

$^1$Traits in $T^*$ are given special status in the algorithm in that, while usually we ignore edge orientation between members of a supernode, this rule is suspended for $T^*$ traits. Edges connecting to $T^*$, the primary trait of interest, are frequently those of most interest to the biologist in that they reflect the causal or reactive nature of the variables nearby. Hence we always score orientations for the edges incident upon $T^*$.
selection is a key component. During SNP selection the goal is to choose SNPs with strong statistical association with the traits or genes in the network node while at the same time avoiding false positives. As our studies show, we also wish to locate orthogonal interventions to score edge orientation and rule out possible confounding. The choice of SNPs may have a critical impact on determining the causal or reactive conclusion about a gene-gene or gene-trait interaction.

The marginal statistical tests of association between a SNP and trait or between a SNP and gene expression value is a reasonable place to start choosing SNPs. In addition to the greedy strategy of simply choosing the K most strongly correlated SNPs, two other strategies are employed. For complex diseases with epistatic interactions between loci, our stepwise forward regression method is used to capture those effects from loci that might not be the dominant signal on a panel, but may play important roles in terms of modulating or cooperating with a dominant effect. Finally we provide options to select representative SNPs from haplotype or linked blocks of alleles so that medium strength signals that may be interacting in a complex fashion may also be detected and incorporated into our analysis of complex traits.

While some authors prefer only cis-SNPs, we maximize the utility of our data and incorporate trans-SNPs as well.

SNP selection proceeds through a preliminary quality control and imputation phases followed by multiple selection phases that are summarized by equations (A.1) - (A.5). Prior to selection, missing SNP values for a particular individual are imputed from the nearest neighboring SNP (most similarly correlated) SNP in the data that has a value for a given NA (Not Available) entry in the SNP matrix. This imputation takes advantage of linkage in our F2 cross, and although it facilitates SNP selection and model building during the Stepwise regression phase, the imputed values do not carry forward into the correlations with
traits.

During quality control, SNPs that do not vary in our sample, and those that have a minor allele frequency of less than .01 are discarded from consideration.

For each of the traits in the network, NEO selects Greedy and Gram-Schmidt SNPs. Then agglomerative cutree center SNPs are chosen using only the SNPs data itself, independently of any trait. Newly selected SNPs that are substantially correlated with any member of the already chosen SNP set are discarded as redundant. In the following definition, Cor defaults to Pearson correlation, but could just as readily use common rank-based alternatives.

\[
\text{Greedy SNPs} = \bigcup_{\text{trait} \in T} \text{topK}\left(\{|\text{Cor}(\text{SNP}_i, \text{trait})|\}_{i=1}^{M}, K_{\text{greedy}} \right) \quad (A.1)
\]

\[
\text{Forward (Gram-Schmidt/}
\text{Forward) Stepwise SNPs} = \bigcup_{\text{trait} \in T} \text{Forward.regress(trait} \sim \text{SNP}_1 \ldots \text{SNP}_{M}) \quad (A.2)
\]

\[
\text{Agglomerative (cutree-}
\text{center) SNPs} = \text{Find.cluster.centers(cutree(hclust(SNP}_1 \ldots \text{SNP}_{M}))}(A.3)
\]

\[
\text{chosen SNPs} = \text{Greedy} \cup \text{Stepwise} \cup \text{Agglomerative SNPs} \quad (A.4)
\]

\[
\text{final SNPs} = \text{Non.redundant(\text{chosen SNPs}, \mathcal{Y}_{\text{redundant}})} \quad (A.5)
\]

In the first selection phase, as suggested by equation (A.1), SNPs are chosen greedily. We correlate each SNP in \( M \) with each trait in \( T \). At present this uses Pearson (optionally Spearman) correlation coefficients on the pairwise complete subset of observations. For each trait in \( T \), the top \( K_{\text{greedy}} \) SNPs that show maximum correlation with the trait are added to the list of chosen SNPs. The actual number used is a user adjustable parameter; currently we use (on the order of) \( K_{\text{greedy}} = 12 \).

In the second selection phase (equation (A.2)), a forward stepwise regression to predict each trait from the supplied SNPs is then employed, halting the
selection when either four (a parameter) more SNPs are chosen, or the Akaike Information Criterion (AIC) for the regression model fails to decrease for any additional SNPs (VR02). These Gram-Schmidt SNPs are added to the set of chosen SNPs.

Finally, as suggested by equation (A.3), we locate SNPs from larger clusters of co-varying SNPs in that they may provide potentially more orthogonal components in the marker data. To do so we perform a hierarchical clustering of the SNP data, using a distance metric of 1 minus the absolute value of the Pearson correlation between pairs of SNPs across the F2 sample. We then cut the resulting dendrogram horizontally (using the R function cutree) at a given correlation strength (0.7 was used in trials), to generate a finite and yet small number of SNP clusters. For each such cluster, a representative SNP is chosen by the following method.

All the SNPs (M nodes) in a cluster are correlated with all of the traits (T nodes), and the single SNP with largest absolute Pearson correlation is elected as the representative of that particular cluster. We typically use on the order of 20 clusters using this procedure, and hence add to our SNP set 20 more chosen candidates. Some will certainly be redundant given the forward stepwise regression choices, and hence following the filtering for tightly correlated SNPs (A.5) (at or above the threshold, \( T_{\text{redundant}} \), frequently set to 0.9), the total chosen set will increase in size by somewhat less than the total number of clusters. Finally, SNPs are sorted and the unique set of SNPs from all of the selection phases is supplied to the next part of the algorithm.

A.0.3 Step Three: Creating Supernodes.

**Definition 6 Supernode.** A supernode is a set of highly correlated node variables. Each variable in the supernode is correlated to every other variable in the
supernode above an absolute correlation threshold.

Many tightly correlated traits and SNPs can be problematic in that these correlations will tend to 1) mutually eliminate each other during the partial correlation filtering, and 2) introduce multicollinearity into the correlation matrices, making inversion more difficult. Therefore we take steps next to ensure that the variables entering into our graph are reasonably distinct for our uses in edge orienting. To accomplish this, we create a supernode out of any multiple traits that are correlated at or above a specific threshold (above 0.7 absolute correlation, for example, was frequently used in our experiments).

Variables within a supernode are considered completely connected, and partial correlation filters are applied and no edge orienting is done between them. In the final drawn graph, the supernodes show up as cliques-completely connected subgraphs.

A.0.4 Step Four: Filtering options to eliminate spurious edges.

As in the FTC algorithm of Cohen et al (CGB95) and in ways similar to the PC (SGS00a) and IC (Pea00) algorithms, we next employ a series of filter conditions to eliminate particularly unlikely edges from consideration. These filters are summarized in equations (A.6)-(A.10). We start with the fully connected directed graph over all \( M \) and \( T \) nodes.

Notationally, given the vertex set \( V \) of size \( |V| = |M \cup T| \), we represent the edges in the directed graph with a \( |V| \times |V| \) adjacency matrix \( E : E_{ij} \in \{0, 1\} \). We take \( E_{ij} = 1 \) iff there exists an edge from node \( i \) to node \( j \).

Edge filtering then applies each of the following rules to the adjacency matrix \( E \), after initializing \( E_{ij} = 1 \) for all \( i \neq j \); and \( E_{ij} = 0 \) for all \( i = j \). The user may fix the stringency of the \( \Upsilon \) parameters for each NEO edge orienting run,
although the program assumes some reasonable, albeit statistically conservative, default values.

\[ E_{ij} = 0 \text{ for all } j \in M \]  \hspace{1cm} (A.6)

\[ E_{ij} = 0 \text{ for all } i, j : |\rho(i, j)| < \Upsilon_{\text{cor}} \]  \hspace{1cm} (A.7)

\[ E_{ij} = 0 \text{ for all } i, j : |\rho(i, j|V\{i, j\})| < \Upsilon_{\text{full,partial}} \]  \hspace{1cm} (A.8)

\[ E_{ij} = 0 \text{ if } \left| 1 - \frac{|\rho(i, j|V\{i, j\})|}{|\rho(i, j)|} \right| > \Upsilon_{\text{Cohen}} \]  \hspace{1cm} (A.9)

\[ E_{ij} = 0 \text{ if } \left| 1 - \frac{|\rho(i, j|k^*)|}{|\rho(i, j)|} \right| > \Upsilon_{k_{\text{max}}}, \text{ with } k^* = \arg \min_{k} |\rho(i, j|k)| \] (A.10)

In (A.6), we eliminate all edges coming into the SNPs to reflect the fact that the SNPs are fixed from the zygote onward and be readily considered nodes without parents in the developed tissues. Then in (A.7) any edges below a given threshold for the absolute value of their correlation are eliminated. In (A.8), if the partial correlation between the two variables conditioned on all the remaining variables is small (below a parameterized threshold) we eliminate that edge.

Our next filter (A.9) is based on Cohen’s omega filter (CGB95). We first compute the ratio of the absolute value of the partial correlation for an edge (conditioning upon all the remaining variables using a shrinkage algorithm for speed an numerical stability when generating the pseudo-inverse (SS04)) to the absolute value of the marginal correlation for that edge. The absolute value of one minus this ratio must be below a given threshold in order to have the edge retained. This filter attempts to weed out edges that have large indirect influences, and to keep those edges that are capturing large direct effects, as suggested by a high partial to marginal correlation ratio.

Our last filter (A.10) is a variation on the previous filter that changes the conditioning set. Instead of using the partial correlation obtained by conditioning
an association on all other variables in the graph, we condition each correlation upon only the single variable that maximally drops the partial correlation relative to the marginal correlation.

Clinical trait $T^*$ nodes are handled specially during filtering, because we focus our primary interest on the edges that lead into (causal) of the trait, and lead away from (reactive to) the trait. Hence, we ignore the partial correlation and omega filters when we consider edges between a clinical trait and any other type of node.

A.0.5 Step Five: Scoring the structural confidence of each directed edge

Next the computation of edge orienting scores takes place. The Local-structure edge orienting scores (LEO scores) are derived from local SEM models fit to compute the evidence for a particular edge orientation. Single marker LEO scores use one marker at a time. Multiple marker LEO scores compare models utilizing two or more markers at a time. The LEO scores measure the transitive agreement of correlation values since these are valued significantly by the local structural equation model fit functions.

The Z-score based edge orienting (ZEO) scores utilize Fisher’s transform (Fis15; Fis21) of the marginal correlation and partial correlation to normality. The ZEO computes the correlation between nodes with and without conditioning on an intermediate node, and as such emphasizes conditional independence features. One component of the ZEO score is known as the Bilayer verification (BLV) score, and is particularly useful for detecting confounded correlation. The ZEO scores are, in additional, very useful as fast computational proxies for the LEO scores.

Three strategies for weighting, combining, and integrating evidence from
multiple markers are explored for the LEO and the ZEO scores.

### A.0.5.1 LEO score definitions

We define the LEO scores based on the local fits of the structure equation models shown in Figures 8.1 and 8.2. Referring to model numbers in Figure 8.1, the single marker version of these scores are stated as follows.

\[
LEO.NB = \log_{10} \left( \frac{P(\text{model 1: } M \to A \to B)}{\text{max} \left( P(\text{model 2: } M \to B \to A), P(\text{model 3: } A \leftarrow M \to B), P(\text{model 4: } M \to A \leftarrow B), P(\text{model 5: } M \to B \leftarrow A) \right)} \right) \tag{A.11}
\]

\[
LEO.I = \log_{10} \left( \frac{P(\text{model 1: } M \to A \to B)}{P(\text{model 3: } A \leftarrow M \to B)} \right) \tag{A.12}
\]

\[
LEO.O = \log_{10} \left( \frac{P(\text{model 1: } M \to A \to B)}{P(\text{model 4: } M \to A \leftarrow B)} \right) \tag{A.13}
\]

\[
LEO.R = \log_{10} \left( \frac{P(\text{model 1: } M \to A \to B)}{P(\text{model 2: } M \to B \to A)} \right) \tag{A.14}
\]

We prefer the use of the \(LEO.NB\) score as a primary analysis metric. A more subtle and detailed analysis benefits from inspection of the \(LEO.I\), \(LEO.O\), and \(LEO.R\) scores. The \(LEO.I\) compares the forward orientation to the confounded model. The \(LEO.O\) compares directly the forward \(A \to B\) orientation to the reverse. The \(LEO.R\) scores assesses the confidence in the marker assignment by comparing the causal to the reactive model.

Referring to model numbers in Figure 8.2, the multiple marker versions the
LEO scores are defined, with $H$ being a hidden latent variable, as follows.

\[
LEO_{NB} = \log_{10} \frac{P(\text{model 1}: M_A \rightarrow A \rightarrow B \leftarrow M_B)}{\max \left( P(\text{model 2}: M_A \rightarrow A \leftarrow B \leftarrow M_B), \right.}
\left. P(\text{model 3}: B \leftarrow M_A \rightarrow A; B \leftarrow M_B \rightarrow B), \right.}
\left. P(\text{model 4}: M_A \rightarrow A \leftarrow H \rightarrow B \leftarrow M_B) \right)}
\tag{A.15}
\]

\[
LEO_{I} = \log_{10} \frac{P(\text{model (a)}: M_A \rightarrow A \rightarrow B \leftarrow M_B)}{P(\text{model 4}: M_A \rightarrow A \leftarrow H \rightarrow B \leftarrow M_B)}
\tag{A.16}
\]

\[
LEO_{O} = \log_{10} \frac{P(\text{model (a)}: M_A \rightarrow A \rightarrow B \leftarrow M_B)}{P(\text{model 2}: M_A \rightarrow A \leftarrow B \leftarrow M_B)}
\tag{A.17}
\]

\[
LEO_{U} = \log_{10} \frac{P(\text{model (a)}: M_A \rightarrow A \rightarrow B \leftarrow M_B)}{P(\text{model 3}: B \leftarrow M_A \rightarrow A; B \leftarrow M_B \rightarrow A)}
\tag{A.18}
\]

The model probabilities are the computed from the right tail of the null distribution of the $\chi^2$ fit statistic for each model, given the degrees of freedom specific to that model. The SEM fit statistics are computed by currently using the sem R package (Fox06) by John Fox.

A.0.6 Derivation of the ZEO score and method

A.0.6.1 ZEO: Z-score Edge Orienting

Since the LEO score computation can be intensive, we here derive an efficiently computable proxy score that captures much of the LEO signal and can be used to accelerate LEO score computations by pre-processing to eliminate non-significant edges. The ZEO score is the output of a novel procedure for edge orienting using the combined Z-scores returned by the difference of the Fisher transform of the absolute value of two correlation coefficients to select between models 1 and 2 depicted in Figure 8.1(a).

To derive the ZEO score, consider the orthomarker in the lower half of
Figure 6.1(c). The following relationships must hold in either edge orientation
(A → B or B → A), and as such constitute a minimum entry conditions that
a set of two markers and two traits must satisfy in order to be considered for
the subsequent step of edge ordering in multimarker analysis. Here ⊥ denotes
independence, while \(\not\perp\) denotes dependence.

\[
\begin{align*}
M_A & \perp M_B & \quad (A.19) \\
M_A & \not\perp A & \quad (A.20) \\
M_B & \not\perp B & \quad (A.21) \\
A & \not\perp B & \quad (A.22)
\end{align*}
\]

In contrast, there are several distinguishing characteristics that allow us to discern
between the two edge orientations. Specifically, consider for contrast Figure 8.2
model 1, and Figure 8.2 model 2. In both cases let the \(M_A\) and \(M_B\) sets consist
of a single marker each. Now examine the differences in the independence and
conditional independence relationships each model predicts.

Model 1: \((M_A \rightarrow A \rightarrow B \leftarrow M_B)\) Model 2: \((M_A \rightarrow A \leftarrow B \leftarrow M_B)\)

\[
\begin{align*}
M_A & \not\perp B \quad \text{versus} \quad M_A \perp B & \quad (A.23) \\
M_B \perp A & \quad \text{versus} \quad M_B \not\perp A & \quad (A.24) \\
M_A \perp B|A & \quad \text{versus} \quad M_A \not\perp B|A & \quad (A.25) \\
M_B \not\perp A|B & \quad \text{versus} \quad M_B \perp A|B & \quad (A.26)
\end{align*}
\]

The entry point of subsequent analysis is that properties A.19-A.22 are
satisfied. Next, as a means of deciding between the models 1 and 2 of Figure 8.2,
and keeping the properties of A.23 - A.26 in mind, consider the following four
step decision analysis.

1. Is $M_B \perp A$ and simultaneously $M_A \perp B$? If so, then we don’t have a
good causal anchor SNP, and perhaps most likely the situation is that of Figure
8.2, model 4, hidden variable confounding. In this case we halt any further testing
and avoid any $A \to B$ and $B \to A$ conclusions whatsoever.

2. If both $M_B \not\perp A$ and simultaneously $M_A \not\perp B$ then we have some form
of epistasis (Figure 8.2 model 3) happening, and we refuse to draw directional
conclusions.

3. Next, if we do find the asymmetry we need, in that either $M_A \perp B$
while $M_B \not\perp A$, or instead $M_A \not\perp B$ while $M_B \perp A$, then we can proceed to the

4. Check $M_A \perp B|A$ (A.25) and $M_B \perp A|B$ (A.26)

If all these tests vote in favor of model 1 or model 2, then conclude that is
the correct model. If one test favors one model while second test favors the other
model, then we stop and avoid concluding any directionally verified edges at all
here.

The above analysis gives a discrete decision based on specified thresholds for
distinguishing between edge orientations. Similarly, we now derive a continuous
measure that expresses how confident we are that the directed arrow exists in
the graph. The Z-score Edge Orienting (ZEO) approach is based on the degree
of strength of the correlations, as transformed by Fisher’s transform $F(\cdot)$ of the
correlation coefficient to normality, with variance of $\frac{1}{N-|Z|-3}$ where $|Z|$ is the
size of the conditioning set for partial correlations. Here $N$ is the number of
observations and $\rho(a, b)$ is the correlation between $a$ and $b$, and $\rho(a, b|c)$ is the
partial correlation between $a$ and $b$ given $c$. The relevant component parts of the
ZEO computation, as well as the BLV scores that are a components of the ZEO,
are defined as follows.

\[
F(r) = \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right) \quad (A.27)
\]

\[
Z_{M_{A,B}} = \sqrt{N - 3} \cdot F(|\rho(M_A, B)|) \quad (A.28)
\]

\[
Z_{M_{B,A}} = \sqrt{N - 3} \cdot F(|\rho(M_B, A)|) \quad (A.29)
\]

\[
Z_{M_{A,B}|A} = \sqrt{N - 4} \cdot F(|\rho(M_A, B|A)|) \quad (A.30)
\]

\[
Z_{M_{B,A}|B} = \sqrt{N - 4} \cdot F(|\rho(M_B, A|B)|) \quad (A.31)
\]

\[
BLV_{A\rightarrow B} = Z_{M_{A,B}} - Z_{M_{A,B}|A} \quad (A.32)
\]

\[
BLV_{B\rightarrow A} = Z_{M_{B,A}} - Z_{M_{B,A}|B} \quad (A.33)
\]

\[
ZEO_{A\rightarrow B} = \frac{1}{2} (BLV_{A\rightarrow B} - BLV_{B\rightarrow A}) \quad (A.34)
\]

**A.0.6.2 Illustration and demonstration of the ZEO**

The construction and function of the ZEO score is depicted in Figure A.1. The three groups shown at 0, 50%, and 100% causal signal to noise are for a fixed observed correlation and fixed heritabilities. As the causal-directed signal from B to A increases, the \( BLV_{A\rightarrow B} \) rises in proportion, in turn pushing the total ZEO score above the given 2.0 significance threshold.
Figure A.1: **Box plots illustrating the decomposition of the ZEO score** (gray boxes) into its constituent BLV scores ($BLV_{A\rightarrow B}$ in violet boxes and $BLV_{B\rightarrow A}$ in yellow boxes) as the level of confounding goes from fully confounded (0% proportion $B \rightarrow A$ causal), to 50%, to not-confounded (100% causal). The negative $BLV_{B\rightarrow A}$ (yellow) score matches the degree to which the $B \rightarrow A$ signal is confounded. The Monte Carlo simulation here titrated between multimarker model 2 and multimarker model 4 in Figure 8.2. Two marker MA, MB model. ha=hb=0.4; 100 simulations at each of three causal-signal-to-noise ratios. N=200.
APPENDIX B

Aggregation methods: detailed methods for select algorithms

Here we provide mathematically precise definitions and details of the score computation for five of the studied edge orienting scores. The LEO.NB.OCA, the LEO.NB.CPA, the LEO.MAX, the LEO.NB.ALL, and the ZEO.FOR are described. The scores differ mainly in terms of whether they employ the single anchor or the multiple anchor models, and in terms of what type of SNP aggregation is employed.

B.0.7 Detailed aggregation algorithms

B.0.7.1 LEO.NB.OCA: Local-structure Edge Orienting, using Forward stepwise SNP aggregation and Orthogonal Causal Anchors

Under the Orthogonal Causal Anchor (OCA) strategy, also referred to as the Forward strategy, we build a forward stepwise regression model to predict A from the M.A SNPs. The M.A SNPs are, by definition and application of the marker assignment consistency principle, those SNPs that more strongly associated with A than with B. We then repeat this model building to predict B from the M.B SNPs—those more strongly associated with B.

1. Begin forward stepwise building a linear regression model for A; the start-
ing model contains the strongest associated SNP from $M_A$ and using it to generate the starting model.

2. Step forward, adding the next optimal SNP from $M_A$ as explanatory for the residuals of the model. Repeat until the Akaike Information Criterion (AIC is $-2(\log \text{ likelihood} - \# \text{ free parameters})$) fails to decrease for subsequent additional terms. Presently the R function MASS:stepAIC is used for this selection, with a customize pre-selection wrapper to enable it to handle large numbers of candidate SNPs.

3. If $M_B$ is not an empty set, repeat above the forward step model building to predict response $B$ from candidate predictors in $M_B$. Designate by $Ka$ and $Kb$ the final number of SNPs chosen as predictors for $A$ and $B$ respectively. Let $M_A^{(1,Ka)}$ and $M_B^{(1,Kb)}$ designate the set of chosen SNPs.

4. Fit by two-stage least squares the local structural equation models shown in Figure 8.2, evaluating and retaining the right tail distribution probabilities for the first two model’s (model 1 and model 2) $\chi^2$ fit statistic, $P_{\chi^2}(A \rightarrow B)$ and $P_{\chi^2}(B \rightarrow A)$. (Thence smaller $\chi^2$ means bigger probability and indicates a better fit.) The other two models (model 3 and model 4) are fit for reference and confirmation if the edge score is deemed significant.

(a) Model 1: Orthomarker $A \rightarrow B$: $M_A^{(1,Ka)} \rightarrow A \rightarrow B \leftarrow M_B^{(1,Kb)}$

(b) Model 2: Orthomarker $B \rightarrow A$: $M_A^{(1,Ka)} \rightarrow A \leftarrow B \leftarrow M_B^{(1,Kb)}$

(c) Model 3: Multimarker unresolvable: $A \leftarrow M_A^{(1,Ka)} \rightarrow B; A \leftarrow M_B^{(1,Kb)} \rightarrow B$

(d) Model 4: Multimarker confounded: $M_A^{(1,Ka)} \rightarrow A \leftarrow \{\text{Hidden confounder}\} \rightarrow B \leftarrow M_B^{(1,Kb)}$

5. Output $LEONB, OCA_{A \rightarrow B} = \log_{10}(\frac{P_{\chi^2(\text{model 1})}}{\max_{i \geq 2}(P_{\chi^2(\text{model } i)})})$.
B.0.7.2 LEO.NB.CPA: Local-structure Edge Orienting, using Forward stepwise SNP aggregation and Common Pleiotropic Anchors

The LEO.NB.CPA score for the edge $A \rightarrow B$ is distinguished from the LEO.NB.CPA score for the reversed or oppositely oriented edge $B \rightarrow A$ in that different sets of markers are used, and a total of ten models must be computed, rather than five, to obtain scores for both edge orientations. To compute the LEO.NB.CPA($A \rightarrow B$), only the $M_A$ set of markers is utilized. To compute the LEO.NB.CPA($B \rightarrow A$), only the $M_B$ set of markers is utilized. We describe below the details for only one orientation evaluation, that for $A \rightarrow B$; the other should be clear by symmetry. We also note for reference that the LEO.NB.CPA score is also known in the codebase by its earlier name, the MULTIONE score.

1. Begin forward stepwise building a linear regression model for $A$; the starting model contains the strongest associated SNP from $M_A$ and using it to generate the starting model.

2. Step forward, adding the next optimal SNP from $M_A$ as explanatory for the residuals of the model. Repeat until the Akaike Information Criterion (AIC is $-2(\log \text{likelihood} - \# \text{free parameters})$) fails to decrease for subsequent additional terms. Presently the R function MASS:stepAIC is used for this selection, with a customize pre-selection wrapper to enable it to handle large numbers of candidate SNPs.

3. Fit by two-stage least squares the local structural equation models shown in Figure 8.1 but using the full multiple marker set $M_A$ in place of $M$. Evaluate and retain the right tail distribution probabilities for the models’ $\chi^2$ fit statistic. Note that the degrees of freedom for the $\chi^2$ distribution will vary depending on the number of nodes in $M_A$. 

135
4. Output $\text{LEO.NB.CPA}_{A\to B} = \log_{10}(\frac{P_{\chi^2}(\text{model 1})}{\max_{i\geq 2}(P_{\chi^2} \text{model } i)})$.

**B.0.7.3 LEO.MAX: Single-best-A-versus-single-best-B SNP treatment**

To find the LEO.MAX score, a procedure very similar to the LEO.NB.OCA score is employed. However, instead of finding a set of markers to predict $A$ and $B$ respectively, the single best marker for $A$ and the single best marker for $B$ are sought. This combination of the “best” markers is tempered, however, by a search for orthogonal anchors.

1. Let $W_{A\to B}^{(i)} = \max_i \left( \log_{10}(P(X \geq \chi^2 \text{for model 1: } M_A^{(i)} \to A \to B)) - \log_{10}(P(X \geq \chi^2 \text{for model 4: } M_A^{(i)} \to A \leftarrow B)) \right)$. Let $j_A$ be the choice of marker index $i$ at which the maximum value $W_{A\to B}^{(i)}$ is obtained.

2. Let $W_{B\to A}^{(i)} = \max_i \left( \log_{10}(P(X \geq \chi^2 \text{for model 1: } M_B^{(i)} \to B \to A)) - \log_{10}(P(X \geq \chi^2 \text{for model 4: } M_B^{(i)} \to B \leftarrow A)) \right)$. Let $j_B$ be the choice of marker index $i$ at which the maximum value $W_{B\to A}^{(i)}$ is obtained.

3. If marker $M_A^{(j_A)}$ is independent of marker $M_B^{(j_B)}$ below a user defined correlation threshold parameter for dependence, then output $\text{LEO.OA}_{A\to B}^{\text{max}} = W_{A\to B}^{(j_A)} - W_{B\to A}^{(j_B)}$ and halt, we are done.
4. Otherwise conduct a search to through $M_A$ and $M_B$ to find independent markers that maximize their respect orienting scores under the constraint of marker-marker independence, as follows:

(a) First allow priority to the $M_A$. Choose the strongest score for the forward direction $W^{(j_A)}_{A\rightarrow B}$ and scan down the list of ranked $W^{(i_B)}_{B\rightarrow A}$ scores until we find an $i_B$ such that $M^{(i_B)}_B \in M_B$ that is independent of $M^{(j_A)}_A \in M_A$. Compute $LEO.O^{(i_B)} = W^{(j_A)}_{A\rightarrow B} - W^{(i_B)}_{B\rightarrow A}$.

(b) Then allow priority to the $M_B$ set. Choose the strongest score for the reverse direction $W^{(j_B)}_{B\rightarrow A}$ and scan down the list of ranked $W^{(i_A)}_{A\rightarrow B}$ scores until we find an $i_A$ such that $M^{(i_A)}_A \in M_A$ that is independent of $M^{(j_B)}_B \in M_B$. Compute $LEO.O^{(i_A)} = W^{(j_A)}_{A\rightarrow B} - W^{(i_B)}_{B\rightarrow A}$.

(c) If the resulting $LEO.O^{(i_A)}$ and $LEO.O^{(i_B)}$ scores differ in sign, then we mark that edge as inconclusive.

(d) Otherwise, reassign $j_A = i_A$ and $j_B = i_B$.

5. Fit by two-stage least squares the local structural equation models shown in Figure 8.1, evaluating and retaining the right tail distribution probabilities for the first two model’s (model 1 and model 2) $\chi^2$ fit statistic, $P_{\chi^2}(A \rightarrow B)$ and $P_{\chi^2}(B \rightarrow A)$.

(a) Multimarker $A \rightarrow B$: $M^{(j_A)}_A \rightarrow A \rightarrow B \leftarrow M^{(j_B)}_B$

(b) Multimarker $B \rightarrow A$: $M^{(j_A)}_A \rightarrow A \leftarrow B \leftarrow M^{(j_B)}_B$

6. Output $LEO.MAX_{A\rightarrow B} = \log_{10}(\frac{P_{\chi^2}(A\rightarrow B)}{P_{\chi^2}(B\rightarrow A)})$.

B.0.7.4 LEO.NB.ALL: all vote SNP aggregation

The ALL aggregation methods utilizes only single marker scores. They computes the average of the scores for the $A \rightarrow B$ direction, and contrasts them with the
average of the scores for the opposing $B \rightarrow A$ direction.

For ease of illustration we begin with the details for the LEO.O score under ALL SNP aggregation.

1. For the sets $M_A$ and $M_B$, compute the unweighted averages of the votes of the local SEM fits as follows. Separately fit $M_A$ and $M_B$ both for the single marker models of Figure 8.1. Use the marker sets in either direction to compute mean evidence for each direction in turn.

$$W_{A\rightarrow B}^\text{mean} = \frac{1}{|M_A|} \sum_{i=1}^{|M_A|} \left( \log_{10}(P(X \geq \chi^2 \text{ for model 1: } M_A^{(i)} \rightarrow A \rightarrow B)) - \log_{10}(P(X \geq \chi^2 \text{ for model 4: } M_A^{(i)} \rightarrow A \leftarrow B)) \right)$$

$$W_{B\rightarrow A}^\text{mean} = \frac{1}{|M_B|} \sum_{i=1}^{|M_B|} \left( \log_{10}(P(X \geq \chi^2 \text{ for model 1: } M_B^{(i)} \rightarrow B \rightarrow A)) - \log_{10}(P(X \geq \chi^2 \text{ for model 4: } M_B^{(i)} \rightarrow B \leftarrow A)) \right)$$

2. Output $LEO.O.ALL_{A\rightarrow B} = W_{A\rightarrow B}^\text{mean} - W_{B\rightarrow A}^\text{mean}$.

The LEO.O.NB.ALL score is then defined analogously to the LEO.O.ALL score, by taking the best of models 2, 3, 4, and 5 in place of model 4 above.

**B.0.7.5 ZEO.FOR: Z-scoring with Forward stepwise SNP aggregation**

The ZEO.FOR score is assembled from weighted BLV components. After building the linear regression model using forward stepwise regression, we compute a weight for each SNP put into the model. For each SNP, the weight assigned is proportion of additional variance explained as that variable is added to the model. From the ANOVA table for the model, the sum of squares for all the entered SNPs is summed to get a total, and then each SNP is assigned weight relative to this explained variance total.
In effect, the weighted mean Z-score for the $A \rightarrow B$ edge from the set of markers $M_A$ for evaluating $M_A \rightarrow A \rightarrow B$ is computed as the sum of weighted individual two-step $BLV_{M_A^{(i)} \rightarrow B}$ scores, where weighting is by the strength of association between the individual markers and $A$.

1. Begin forward stepwise building a linear regression model for $A$; the starting model contains the strongest associated SNP from $M_A$ and using it to generate the starting model.

2. Step forward, adding the next optimal SNP from $M_A$ as explanatory for the residuals of the model. Repeat until the Akaike Information Criterion (AIC is $-2(\log$ likelihood $- \# \text{ free parameters})$) fails to decrease for subsequent additional terms. Presently the R function MASS::stepAIC is used for this selection, with a customize pre-selection wrapper to enable it to handle large numbers of candidate SNPs.

3. If $M_B$ is not an empty set, repeat above the forward step model building to predict response $B$ from candidate predictors in $M_B$. Designate by $K_a$ and $K_b$ the final number of SNPs chosen as predictors for $A$ and $B$ respectively. Let $M_A^{(1..K_a)}$ and $M_B^{(1..K_b)}$ designate the set of chosen SNPs.

4. From the ANOVA table for the multiple linear regression model just obtained, let $T_A$ be the total the sum-of-squares for all the entered SNPs $M_A^{(1..K_a)}$. Each SNP then is assigned weight relative to this explained variance total. Let $v_i$ be the percentage of variance $T_A$ explained by SNP $M_{A}^{(i)}$. 

139
5. For the set $M_A^{(i)}$, compute the sum

$$BLV_{M_A \rightarrow B} = \sum_{i=1}^{\vert M_A \vert} v_i (Z_{M_A^{(i)},B} - Z_{M_A^{(i)},B|A})$$

(B.3)

$$= \sum_{i=1}^{\vert M_A \vert} v_i BLV_{M_A^{(i)} \rightarrow B}$$

(B.4)

6. Repeat the computation in the last two steps for $M_B^{(1..Kb)}$ to obtain $BLV_{M_B \rightarrow A}$.

7. Output $ZEO_{A \rightarrow B}^\text{for} = \frac{1}{2} (BLV_{M_A \rightarrow B} - BLV_{M_B \rightarrow A})$
APPENDIX C

A Brief Review of Structural Equation Model
Fitting and Goodness of Fit Indices

Discrepancy or fit functions for Structural Equations Model fitting have been derived in various forms, and here we review several of these measures of how well a given model fits the data. The various fit indices all take as their starting point the observed covariance matrix $S_{m \times m}$, and the predicted (implied by the final model) covariance matrix $\Sigma(\hat{\theta})_{m \times m}$, also written for convenience and clarity as $\hat{\Sigma}$. The $\hat{\theta}$ notation reminds us that $\hat{\Sigma}$ is the covariance matrix implied by the model after the best fitting estimates $\hat{\theta}$ of the parameter vector $\theta$ have been obtained by optimization (see Figure C.1). Our discussion follows closely the exposition of Loehlin (Loe04) and the terminology and observations of Bollen (Bol89).

There are four “fairly standard” fitting methods discussed in Loehlin’s introduction to this topic, and we recapitulate his choice of methods here. The fit functions differ slightly given the method chosen. The four methods discussed here are Ordinary Least Squares (OLS), Generalized Least Squares (GLS), Maximum Likelihood, and one version of Browne’s asymptotically distribution-free criterion (ADF) which is also known in the LISREL software package terminology as Generally weighted least squares or “arbitrary distribution generalized least squares” in the EQS package.

To give the context for these global fit functions, we refer to the covariance matrix notation illustrated in figure C.1. The covariance structure hypothesis
is also referred to as the null hypothesis. Contrary to classical statistics, now the null hypothesis is the exciting outcome that any deviation from our model is simply due to random variation or sampling noise. In SEM therefore, \( H_0 \) is that the population covariance \( \Sigma \) is the same as the asymptotic expected covariance predicted from the model, denoted \( \Sigma(\theta) \). Hence in shorthand, \( H_0 \) is \( \Sigma = \Sigma(\theta) \). Now \( \Sigma \) and \( \Sigma(\theta) \) are attributes of the population, and hence are typically unavailable for direct inspection, and must be estimated from a smaller sample of the population, yielding the sample counterparts of \( S \) and \( \hat{\Sigma} \).

As Bollen (Bol89)[p256-257] discusses, overall model fit indices are distinguished from component fit indices in that they evaluate the entire model at once, and hence may reveal global inadequacies of the model that are not found by the more myopic measures of individual component (equation, parameter) fitness measures.

A limitation, however, of overall fit indices is that they can’t be used by exactly identified models (with no degrees of freedom remaining). In this case of
exact identification, $S = \hat{\Sigma}$ always exactly, so overall fit is a moot point and not a useful score for indicating global model robustness.

Another shortcoming of general fit or overall fit indices is that from their global perspective they may hide errors or misfits on the local component level. The overall fit might be quite satisfactory while the parameter signs and estimates may be reversed in sign or insignificant in terms of their individual statistical significance scores.

Lastly, overall fit measures do not indicate how well each individual variable is predicted by its parent variables. Global indices don’t reveal an $R^2$ or percentage of variance explained by the local parent-child graphical/regression models. Nor do global indices summarize the $R^2$ for all such relationships in the model. In line with Bollen’s conclusion here, it is wise to use global measures to supplement and to view in conjunction with local measures of model fitness.

To establish notation, assuming we have $m$ variables in the model, let $\vec{s}$ be the linearized $\frac{m(m+1)}{2} \times 1$ vector of variances and covariances from the lower triangle entries of the observed sample covariance $S$, and let $\vec{c}$ be the correspondingly linearized vector from of the lower triangle of the predicted covariance $\hat{\Sigma}$.

Certain fit functions which we seek to minimize are characterized by a variation in the weight matrix $W$ adopted. Here $*$ denotes transpose.

$$F = (\vec{s} - \vec{c})^*W(\vec{s} - \vec{c}) = \sum_i \sum_j w_{ij} s_i c_j$$

In ADF, $W$ is defined by the inverse of the matrix whose $i,j$-th entry is obtained from $m_{ijkl} - s_{ij} s_{kl}$ where $m$ is the fourth order moment and $s_{ij}$ and $s_{kl}$ are the two covariances of interest. ADF is useful for large samples and can then address some concerns about non-normal data.
If a multivariate normal distribution can be assumed for the observed variables, then the above quadratic form expression for $F$ can be written using the trace function in a manner to unify the consideration of the other three fitting methods.

$$F = \frac{1}{2} tr((S - \hat{\Sigma})V)$$

The choice of the weight matrix $V$ then distinguishes the choice of the other methods.

$$V = \mathbf{I} \quad \text{for ordinary least squares (OLS)}$$
$$V = S^{-1} \quad \text{for generalized least squares (GLS)}$$
$$V = \hat{\Sigma}^{-1} \quad \text{approximately maximum likelihood (ML)}$$

Maximum Likelihood fitting for the Multivariate Normal (MVN) density assumption uses the following function which is computationally equivalent to the MVN density function.

$$F_{ML} = \ln |\hat{\Sigma}| - \ln |S| + tr\hat{\Sigma}^{-1} - m$$

Although the choice of $V = \hat{\Sigma}^{-1}$ does not yield exactly the maximum likelihood fit, it approximates it (Loe04), and further serves to emphasize the similarities and illuminate the major differences between the Ordinary Least Squares (OLS), Generalized Least Squares (GLS), and Maximum Likelihood (ML) fit functions. Notice that ML technique is more computationally intensive, because the $\hat{\Sigma}^{-1}$ must be re-calculated during the iterative ML solution, whereas $S^{-1}$ is fixed and need only be computed once.

Supposing that residual sampling variation is multivariate normal and that the model is correctly capturing the underlying process, then both GLS and ML
methods can be evaluated by the chi-squared distribution, since the minimum
found fit $F_{\text{min}}$ should then be the sum of squared normal errors, so that:

$$X^2 \text{ defined as } = F_{\text{min}}(N - 1)$$

and $X^2 \sim \chi^2$ under the ‘null’ that our model is correct up to normal errors

That is, multiplying the minimized fit function by one minus the sample
size yields a chi-squared variable under the null hypothesis of model correctness
and multivariate normal errors. Notice that, as in all SEM, the usual preferences
in terms of null and alternative hypotheses are reversed.

In rough terms—because we do not address the cross terms, we can think
of the difference between three of the methods in terms of $E_i$ expected value
from the model versus $O_i$ observed value in the data of the $i$-th variable, and
characterize as follows.

$$F \approx \sum_{i} \frac{(O_i - E_i)^2}{E_i} \text{ for ML}$$

$$F \approx \sum_{i} \frac{(O_i - E_i)^2}{O_i} \text{ for GLS}$$

$$F \approx \sum_{i} (O_i - E_i)^2 \text{ for OLS}$$

Goodness of fit indices are generated for each of the methods to provide the
user an overall numeric score for comparing models fit by the same method. Fit
indices, in addition to describing the goodness of the fit of the data and model,
may also involve penalty terms which favor simpler models with fewer unknowns.
Some are normalized to fall in $[0, 1]$, while others are not. Some are specifically
comparing the model to a baseline simple model where, for instance, all variables
are uncorrelated. Certain fit indices try to combine an estimate of the error of
the approximation of the population by the model (which should be fixed by the complexity of the model and independent of sample size N) with an estimate of the error of estimation due to sampling (which should decrease as sample size grows). Such “population based” indices utilize the non-central chi-squared distribution—this is the distribution of the minimized fitting function when the model fits only approximately to the population. Specifically the non-centrality parameter is estimated by the best fit $X^2 - df$. Using the Bentler-Weeks SEM conventions, the degrees of freedom, $df$, is $\frac{m(m+1)}{2}$ minus {the count of variances and covariances of the independent (including endogenous error) variables plus the number of regression coefficients from the endogenous variables’ equations} (Ben06).

Here we will describe several fit indices available to judge the final fitted Structural Equation Model.

**RMSEA** Root Mean Square Error of Approximation is a population based index and hence relatively insensitive to sample size.

$$RMSEA = \sqrt{\frac{X^2 - df}{(N - 1)df}}$$

RMSEA is defined as 0 when $X^2$ is less than the degrees of freedom. Note that RMSEA is primarily capturing the ratio between $F_{min}$ and the degrees of freedom, which becomes apparent when we observe that $\frac{1}{N-1}$ should become small quickly here.

$$RMSEA^2 = \frac{F_{min}(N - 1) - df}{(N - 1)df} = \frac{F_{min}}{df} - \frac{1}{N - 1}$$

**NFI** Bentler and Bonnett’s normed fit index, NFI, uses the ratio the fit of a baseline model to the model being evaluated. This can be equivalently defined in terms of the $X^2$ statistic alone, because the (N-1) terms for the
two models cancel.

\[ \text{NFI} = 1 - \frac{F_{\text{eval}}}{F_{\text{baseline}}} = 1 - \frac{X^2_{\text{eval}}}{X^2_{\text{baseline}}} \]

If \( F_{\text{eval}} \) is very good (close to zero), then \( \text{NFI} \rightarrow 1 \), so larger \( \text{NFI} \) is better.

If \( F_{\text{eval}} \) is very close to \( F_{\text{baseline}} \), then \( \text{NFI} \rightarrow 0 \), so an \( \text{NFI} \) close to zero indicates not much better than baseline. If \( \text{NFI} < 0 \), we know our model is doing worse than the baseline performance.

**GFI** LISREL’s Goodness-of-fit index, GFI, relates the explained covariance to the total covariance by comparing the fit of the model being evaluated to the fit of the same model with the observed covariance matrix \( \mathbf{S} \) substituted for \( \mathbf{S} - \hat{\Sigma} \). Specifically for maximum likelihood estimates, if we define \( F_s \) as the minimum fit score obtained from minimizing

\[ F = \frac{1}{2} \text{tr}[\mathbf{S}\hat{\Sigma}^{-1}]^2 \]

which should be approximately \( \frac{1}{2}m \) asymptotically under the null when there are \( m \) variables in the covariance matrix. Using these terms, we have the definition.

\[ \text{GFI} = 1 - \frac{F_{\text{eval}}}{F_s} \approx 1 - \frac{F_{\text{eval}}}{\frac{1}{2}m} \]

Where the last expression, it may be noted, is somewhat similar the main term \( \frac{F_{\text{mis}}}{df} \) the RMSEA, only differing by whether the model size or the degrees of freedom is used in the denominator. As with the NFI, zero is poor, one is good, and bigger is better.
Of these scores, we tend to refer to the RMSEA most often in practice, as while it is not a p-value, it approximates the common significance thresholds of the familiar 0.05 for a p-value. To interpret the RMSEA, we quote Loehlin (Loe04)[p69] for some suggested guidelines.

Browne and Cudeck have suggested the following guidelines for interpreting RMSEAs: “Practical experience has made us feel that a value of the RMSEA of about .05 or less would indicate a close fit of the model in relation to the degrees of freedom...We are also of the opinion that a value of .08 or less for the RMSEA would indicate a reasonable error of approximation and would not want to employ a model with a RMSEA greater than .10” (BC93)[p144]. Its originator, Steiger, considers values below .10 “good” and below .05 “very good” (Ste89)[p81].
APPENDIX D

Table of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC and CCU</td>
<td>Two causal discovery rules articulated by Silverstein et. al</td>
</tr>
<tr>
<td>CPA</td>
<td>Common Pleiotropic Anchor</td>
</tr>
<tr>
<td>eQTL</td>
<td>Expression Quantitative Trait Loci, a genetic locus correlated with an mRNA level</td>
</tr>
<tr>
<td>FBD and FTC</td>
<td>Two SEM learning algorithms</td>
</tr>
<tr>
<td>IC Algorithm</td>
<td>The Inductive Causation algorithm</td>
</tr>
<tr>
<td>LCD Algorithm</td>
<td>The Local Causal Discovery algorithm</td>
</tr>
<tr>
<td>LEO scores</td>
<td>Local-structure Edge Orienting scores</td>
</tr>
<tr>
<td>NEO</td>
<td>Network Edge Orienting software</td>
</tr>
<tr>
<td>OCA</td>
<td>Orthogonal Causal Anchor</td>
</tr>
<tr>
<td>PC Algorithm</td>
<td>A causal graph discovery algorithm</td>
</tr>
<tr>
<td>QTL</td>
<td>Quantitative Trait Loci</td>
</tr>
<tr>
<td>RVL and RVV</td>
<td>Two algorithms described in this dissertation for recovering causal DAGs</td>
</tr>
<tr>
<td>SEM</td>
<td>Structural Equation Model</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


