Algebraic discrete causal models

Eva Riccomagno · Jim Q. Smith · Peter Thwaites

Abstract The main feature of the paper is to show that Algebraic Statistics is a natural framework to address issues of causality and to help discern a total cause. Indeed identifiability of an effect of a cause in discrete models is almost algebraic rather than graphical in nature. It is useful to think of it as such and it leads to the definition of a large class of discrete models which comprises popular ones.

Keywords Causality · Algebraic Statistics · Identifiability

1 Introduction

Much recent work in the field of causality has focused on how cause relates to control, and the analysis of controlled models. We assume the existence of a background idle system which is subjected to some sort of intervention or manipulation. Indeed in a common scenario, in fields such as epidemiology and economics, an observer collects data from a system and wants to make inference about what would happen were the system been controlled, for example by imposing a new treatment regime. The data generating process and hypotheses on the causal mechanism, governing both the idle system and the manipulated system, are to be specified in order to make predictions. The framework of Markov or Semi-Markovian models has been extensively adopted.

The Bayesian network has been one of the most successful graphical tools for representing complex dependency relationships and the directionality of

E. Riccomagno
Department of Mathematics, Università di Genova
Via Dodecaneso 35
Tel.: +39-010-5646938
E-mail: riccomagno@dima.unige.it

J.Q. Smith and P. Thwaites
The University of Warwick

edges has been interpreted as causal in some way. This has led to the development of the Causal Bayesian Network, using a non-parametric representation based on structural equation models. These provide a framework for expressing assertions about what might happen when the system under study is externally manipulated and some of its variables are assigned certain values. The causal Bayesian network is often used to determine whether or not a causal effect can be deduced from non-experimental data obtained from an idle system.

Many authors have developed useful methods for defining causality and investigating its identifiability under various sampling schemes. Our main references are Pearl (2000), Spirtes et al. (1993) and Shafer (1996).

In this paper by collecting together and developing some results in the literature, we aim to illustrate how looking at the formal algebraic aspects of algebraic statistical models and of the notion of causality allows us to capture the mathematical essence of a causal statistical model and allows for a fully general modelling framework freed of the regularity constraints of Bayesian networks.

2 Algebraic set-up for causal Bayesian networks

This is a review section. The remainder of the paper generalises models and ideas presented here showing that the applicable mathematical technologies are the same as for Bayesian networks. For an ample discussion of formal mathematical structures for conditional independence see Studený (2004). Here we consider only discrete setting. Little is available in the algebraic statistics literature for the continuous case. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random vector where X_i takes values in the finite set $\mathbb{X}_i$, for $i = 1, \ldots, n$ and the sample space for $\mathbf{X}$ becomes $\mathbb{X} = \prod_{i=1}^{n} \mathbb{X}_i$.

Consider a directed acyclic graph whose nodes are in one-to-one correspondence with the components of $\mathbf{X}$. Label the nodes compatibly with the graph, namely $1 \leq j < i \leq n$, hence $X_j < X_i$, whenever there is a direct path from node j to node i. This can be interpreted causally by stating that an edge from j into i corresponds to a direct causal influence of X_j on X_i.

The graph can also be taken to represent conditional independence constraints on the components of $\mathbf{X}$, namely each variable is independent of all its non-descendants given its direct parents in the graph. This corresponds to the following factorization of a joint probability p for $\mathbf{X}$

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | pa_i) \quad \text{for all } (x_1, \ldots, x_n) \in \mathbb{X} \tag{1}$$

where $p(x_i|pa_i)$ is the probability of X_i taking the value $x_i \in \mathbb{X}_i$ given that its parents among $X_1, \ldots, X_v$, $v < i$, take values pa_i. The set of all distribution for $\mathbf{X}$ satisfying Equations (1) gives a statistical model based on a causal Bayesian network.

We consider definitions of interventions which are compatible with the graph structure; that is, which corresponds to modification of some factors in

the right-hand side of Equation (1). Furthermore, we focus on the problem of prediction of intervention effects. The simplest kind of intervention, which we call atomic intervention, is as follows. The jth component of X_j is forced to take a specific value, say $\hat{x_j} \in \mathbb{X}_j$, with probability one and a new density, or post-intervention density, is defined on $\{X_1, \ldots, X_n\} \setminus \{X_j\}$ by the recursive formula

$$p(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n || \hat{x_j}) = \prod_{i=1, i \neq j}^{n} p(x_i | pa_i) \tag{2}$$

with $p(x_i|pa_i)$ as above, but noting that if X_j is a parent of X_i then the component of pa_i corresponding to X_j takes the value $\hat{x}_j$. Equation (2) corresponds to a sub-graph obtained by removing the node j together with its edges. Successive atomic interventions of this type produce an atomic intervention on a subset of $\{X_1, \ldots, X_n\}$.

The pre-intervention and post-intervention densities can be combined in a single framework by adding an extra node $j^\star$ and an extra edge from $j^\star$ in j. The new node corresponds to a random variable F_j taking values in $\mathbb{X}_j \cup \{idle\}$. A joint probability on the augmented causal graph is defined as

$$p(x_1, \ldots, x_n, f_{j^\star}) = \begin{cases} p(x_1, \ldots, x_n) & \text{if } f_{j^\star} = idle \\ p(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n || \hat{x}_j) & \text{if } f_{j^\star} = \hat{x}_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Often it is of interest to estimate the post-intervention distribution, some of its marginal or some other functions, from non-observational data from the pre-intervention distribution. Each factor in the right hand side of Equation (1) can be interpreted as a data simulator, or data generating process. At a structural level, we could ask whether some function e of the post-intervention distribution can be explicitly written as a function of the variables appearing in Equation (1) either on the right hand side or left hand side. Clearly, this is possible for any e if all values of $\mathbf{X}$ are observables prior intervention. But, if there are unobservable or hidden variables in the model, the questions is more difficult and the answer depends on the modular structure imposed by the graph on the model, on the structure of the observed, or manifest, variables, and on the intervention and importantly on how these three are related. We shall go back to this in Section 4.

2.1 Algebraic representation of (causal) Bayesian networks

Bayesian networks can be parametrized in at least two ways: through an explicit mapping of a set of parameters to a set of distributions or via a set of independence constraints that the distributions must satisfy. The first parametrization is given by the transition parameters on the right hand side of Equation (1), and the second one by the $p(x_1, \ldots, x_n)$ for $(x_1, \ldots, x_n) \in \mathbb{X}$.

Consider a simple example. For X_1, X_2, X_3 binary with levels $1-2$, the graph $X_1 \to X_2 \to X_3$ and the atomic intervention $\hat{x}_2 = 1$, Equations (1) are

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2), \text{ for } x_1, x_2, x_3 \in \{1, 2\} \tag{4}$$

and Equations (2) become

$$p(x_1, x_3||\hat{x}_2) = p(x_1)p(x_3|\hat{x}_2), \text{ for } x_1, x_3 \in \{1, 2\}$$

It can be shown that the first set of equations is satisfied by any function on $\{1, 2\}^3$ taking non-negative values summing to one and satisfying the following two polynomial equations

$$-p(1, 2, 2)p(2, 2, 1) + p(1, 2, 1)p(2, 2, 2) = 0 \tag{5}$$
$$-p(1, 1, 2)p(2, 1, 1) + p(1, 1, 1)p(2, 1, 2) = 0 \tag{6}$$

The polynomials in the right hand side of (5) above can be taken to provide an implicit description of the graphical model $X_1 \to X_2 \to X_3$. These are obtained by eliminating the conditional probabilities from Equations (4) and are generators of the set of all polynomials which are invariant under the model. Analogously, the polynomial invariants of the manipulated graph, expressed in the joint probabilities, are of the form af with $f = p(1, 1, 2)p(2, 1, 1) - p(1, 1, 1)p(2, 1, 2)$ and a any polynomial in the $p(x_1, x_2, x_3)$, $(x_1, x_2, x_3) \in \{1, 2\}^3$.

Together with the polynomial constraints in (4) imposed by the conditional independence model, there are other obvious algebraic constraints like $\sum_{x_1,x_2,x_3} p(x_1, x_2, x_3) = 1$ in the joint probabilities and $\sum_{x_2} p(x_2|x_1) - 1$ in the conditional probabilities and there are some inequalities such as the non-negativity of probabilities which render the set of probability densities satisfying the model $X_1 \to X_2 \to X_3$ a semi-algebraic set, namely a space defined by polynomial identities and inequalities.

The theory behind the elimination process that leads to the implicit description of Bayesian networks is a branch of Algebraic Geometry called Elimination Theory (Cox et al. (2008)). It is fairly well understood and implemented in general computer algebra software such as Maple and Matlab. However, its use in applied contexts may be limited by the fact that actual computations are often unfeasible and more refined ways to eliminate are to be sought. For Bayesian networks with and without hidden variables, Garcia, Stillman, and Sturmfels (2005) study the algebraic varieties defined, in particular they show that the naive Bayes model corresponds to the higher secant varieties of Segre varieties.

To handle the semi-algebraic aspects, Drton and Sullivant (2007) use the Tarski-Seidenberg theorem on projections to show that the (well-defined) image of a semi-algebraic set through the rational map defined by Equations (1) on the simplex is still semi-algebraic.

Identifiability problems are typical problems of elimination theory and can be set up within this algebraic framework and used in the causal analysis

together with more standard topological arguments. Typical topological identifiability theorems are the back door and front door criteria for identifiability of a cause of an effect. The set of variables is now augmented to comprise also variables expressing the polynomial equalities and inequalities for the observed (or observable) functions. If in the simple example above we observe the joint probability of X_2 and X_3 then we would have four more variables $m(x_2, x_3) = p(1, x_2, x_3) + p(2, x_2, x_3)$, $x_2, x_3 \in \{1, 2\}^2$.

In computational biology, several authors including Allman, Rhodes and co-workers (2003, 2008) use the algebraic representations of Markov models to address identifiability issues within the field of molecular phylogenetics where it is of interest to infer evolutionary trees from DNA or protein sequences. Identification problems associated with the estimation of some probabilities after manipulation from passive observations (manifest variables measured in the idle system) have been formulated as an elimination problem in computational commutative algebra, for example in the case of Bayesian network the case study in Kuroki (2007). In general, a systematic implementation of these problems in computer algebra softwares will be slow to run. At times some pre-processing can be performed in order to exploit the symmetries and invariances to various group action for certain classes of statistical models (Mond et. al (2003)). Other times a re-parametrisation in terms of non-central moments loses an order of magnitude effect on the speed of computation and hence can be useful (e.g. Settimi and Smith (2000) and Zwiernik and Smith (2009)).

In Algebraic Statistics, the identifiability of causal effects for causal Bayesian networks with hidden variables is addressed by Kang and Tian (2007) via the notion of c-component (Tian and Pearl (2002)). They provide an algorithm to determine polynomial constraints that a causal Bayesian network must satisfy "to be compatible with given observational and experimental data". In the algebraic framework of this paper many non-graphically based symmetries which appear in common models are much easier to exploit than in a solely graphical setting. This suggests that the algebraic representation of causality is a promising way of computing the identifiability of a causal effect in much wider classes of models than Bayesian network.

3 Generalisations

Wider classes of discrete statistical models than causal Bayesian networks have an analogous algebraic formulation. These can overcome symmetry restrictions implied by the requirement of a product sample space in a Bayesian network and allow a larger class of manipulations. In this section, we present a point summary for causal Bayesian networks and define some of these wider classes by generalising one or more of those items.

1. a partial order on $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ and an associated multiplication rule as in Equation (1) define direct causal influences on the X_i's and a statistical model;

2. a discrete Bayesian network can be described through a set of linear equations together with inequalities to express non negativity of probabilities and linear equations for the sum-to-one constraints. Namely, for all $i = 1, \ldots, n$ set $p(x_i|x_1, \ldots, x_{i-1}) = p(x_i|x'_1, \ldots, x'_{i-1})$ whenever $(x_1, \ldots, x_{i-1})$ and $(x'_1, \ldots, x'_{i-1})$ coincide on the parent set of X_i (see Dawid and Studeny (1999)). We refer to this as the transition or primitive probabilities;
3. a Bayesian network is based on the assumption that the factorization in Equation (1) holds across all values of $\boldsymbol{x}$ in a cross product sample space. But in Settimi and Smith (2000) it is shown that identification depends on the sample space structure, in particular on the number of levels a variable takes;
4. within a graphical framework subsets of whole variables in $\boldsymbol{X}$ are considered manifest or hidden;
5. mainly the causal controls being studied in e.g. Pearl (2000); Spirtes et al. (1993) correspond to setting subsets of variables in $\boldsymbol{X}$ to take particular values and often the effect of a cause is expressed as a polynomial function of the primitive probabilities, for example marginals;
6. identification problems are basically elimination problems and can be addressed using elimination theory from computational commutative algebra coupled with a pre-processing based e.g. on topological arguments.

3.1 Bayesian linear models

A Bayesian network can defined through linear constraints on some conditional probabilities according to an ordering of the nodes compatible with the graph, see Item 2 above. A Bayesian linear constraint model (Riccomagno and Smith (2004)) is a straightforward generalisation given by 1. a total order on the components of $\mathbf{X}$ and an associated collection of factorisation formulae like Equations (1); 2. a set of linear equations in transition probabilities

$$L_i(p(x_i|x_1, \ldots, x_{i-1}) = \sum_{x_i \in \mathbb{X}_i} a_{x_i, x_1, \ldots, x_{i-1}} p(x_i|x_1, \ldots, x_{i-1}) = 0$$

with real coefficients $a_{x_i, x_1, \ldots, x_{i-1}}$; 3. a set of linear inequalities with real coefficients on the transition probabilities of the form

$$L_i(p(x_i|x_1, \ldots, x_{i-1})) \leq M_i(p(x_i|x_1, \ldots, x_{i-1})).$$

A Bayesian linear constraint model is called feasible if there is at least a probability distribution over the sample space $\mathbb{X}$ satisfying 1-3 above. Again the Tarski-Seidenberg can be invoked to guarantee that Bayesian linear constraint model are algebraic models.

Examples include censoring of sample information where some joint probabilities, and hence transition probabilities, are set to zero; context specific Bayesian networks where associated factorisations may differ for different instantiations of some X_i, $i = 1, \ldots, n$; parametric discrete Bayesian networks

especially for large sample size where some variables in the network have a distribution conditional on their parents which is parametric and the transition probabilities are a (rational) polynomial function of the distribution parameters, for example a Binomial distribution with a known number of independent trials and unknown success probability; and finally chain graph models.

Intervention is defined for a feasible Bayesian linear constraint model as in a Bayesian network by forcing some variables to assume given values and the post intervention distribution is expressed by equations resembling Equations (2). Again the post intervention distribution needs to continue to respect the equality and inequality constraints in 2-3 above.

3.2 Tree based generalization

For a probability tree $\mathcal{T}$ with vertex set V and edge set E and for $v, v' \in V$, $(v, v') \in E$, let $\pi(v'|v)$ be the possibly unknown transition probability from v to v', under the constraint $\sum_{v':(v,v')\in E} \pi(v'|v) = 1$. The values $\pi(v'|v)$, $(v, v') \in E$, parametrize the model and are the primitive or transition probabilities. The probability of the event corresponding to the root-to-leaf paths $\lambda = (v_0, \ldots, v_{n(\lambda)})$, where v_0 is the root vertex and $v_{n(\lambda)}$ a leaf vertex is

$$p(\lambda) = \prod_{i=0}^{n(\lambda)-1} \pi(v_{i+1}|v_i) \qquad \text{see Equation (1)} \tag{7}$$

The nodes of the tree and the root-to-leaf paths, which are analogue to the joint probabilities for Bayesian networks, are the topological keys to the two parametrisations.

There is a natural partial order associated with the tree which can be used as a framework to express causality. If in the tree the event expressed by node v occurs before the one represented by v' whenever $v \prec v'$, then the effects of a control on a regular tree T can now be defined in total analogy to Item 5 above by modifying the values of some primitive probabilities or more generally by defining constraints in the primitive probabilities that have a causal interpretation. Hence a manipulation of the tree is given by a subset $F \subset E$ and new transition probabilities $\hat{\pi}$ on the edges in F which are functions of the primitive probabilities and compatible with the existing model.

Issues of feasibility are similar to those for Bayesian linear models. Finite discrete Bayesian networks are a special case of this model class. Indeed once an order on $\boldsymbol{X}$ has been chosen, a Bayesian network corresponds to a tree whose root-to-leaf paths have all the same length and whose independence structure is translated into equalities of some primitive probabilities. The basic saturated model in Equations (7) augmented with a set of algebraic equations in the transition probabilities has been called algebraically constraint tree in Riccomagno and Smith (2009) where their superior modelling performance on straightforward Bayesian network is proven. Also Bayesian linear constraint model and chain event graphs below are models of this type. For a detailed example see Riccomagno and Smith (2009).

3.3 Extreme causality

Note that to discuss causal maps we need 1. a finite set of controllable "circumstances" and a finite set of outcomes of an experiment, and 2. a partial order defined on these circumstances, more abstractly 1. a finite set $V = \{v\}$ and 2. a partial order on V, $\prec$. A chain of the Hasse diagram of the ordering $\prec$ represents a possible unfolding of the experiment. The structure is that of a directed acyclic graph, like in a Bayesian network, but where each node stands for an event, like in a probability tree.

A saturated statistical model on V can be defined giving a set of transition probabilities: $\pi(v'|v) \in [0,1]$ where $v, v' \in V$ are in the same chain and there is no v^* in the chain such that $v \prec v^* \prec v'$.

The probability of a chain $\lambda = (v_0, \ldots, v_{n(\lambda)})$ is $p(\lambda) = \prod_{i=1}^{n(\lambda)} \pi(v_i|v_{i-1})$ (cf. Equation (1)). The sum to one condition holds and states that $\sum_{v'} \pi(v'|v) = 1$. An algebraic sub-model is defined by setting to zero suitable algebraic equations of the transition probabilities and by imposing some polynomial inequalities among them.

A manipulation or control can now be defined implicitly by considering a set F of edges of the Hasse diagram and assigning to $(v, v') \in F$ a new primitive probability $\widehat{\pi}(v'|v)$ which we take to be a polynomial function of the vector of primitive probabilities π's. This could be of the atomic type by setting to one some $\pi(\cdot|\cdot)$, or more generally, and often realistic, simply another probability densities on the Hasse diagram

3.4 An example

Consider a statistical model built to study whether watching a violent movie might induce a man into a fight, allowing for testosterone levels to, at least partially, explain a violent behaviour. This could be modelled with a Bayesian network. Let X_2 denote whether a man watches a violent movie early one evening $\{x_2 = 1\}$ or not $\{x_2 = 2\}$ and let X_4 be an indicator of whether he is arrested for fighting $\{x_4 = 1\}$ or not $\{x_4 = 2\}$ late that evening. If he watches the movie, let X_1 denote his testosterone level just before seeing it and X_3 his testosterone level late that evening. For a man who does not watch the movie let $X_1 = X_3$ denote his testosterone level that evening.

Assume X_1 and X_3 take three values: 1 for low levels of testosterone, 2 for medium levels and 3 for high levels, so that $(r_1, r_2, r_3, r_4) = (3, 2, 3, 2)$ and $r = 36$. Then this can be depicted as the following Bayesian example

$$\begin{array}{ccc} X_1 & \rightarrow & X_3 \\ & \nearrow & \downarrow \\ X_2 & \rightarrow & X_4 \end{array}$$

The graph of this Bayesian example embodies two substantive statements: $X_2 \perp\!\!\!\perp X_1$ and $X_4 \perp\!\!\!\perp X_1 | (X_2, X_3)$. The first one states that that fact the man watched the movie would not depend on his testosterone level and the second

one states that the testosterone level before watching the movie gives no additional relevant information about the man's inclination to violence provided that we happen to know both whether he watched the movie and his current testosterone levels.

There are 36 elements in the sample space $\mathbb{X}$ and a general joint mass function on (X_1, X_2, X_3, X_4) is given by the 36 quartic equations

$$p(\boldsymbol{x}) = \pi_1(x_1)\pi_2(x_2|x_1)\pi_3(x_3|x_1, x_2)\pi_4(x_4|x_1, x_2, x_3). \tag{8}$$

The conditional independence statements in Item 2 are given by

$$\begin{aligned} \pi_2(x_2|x_1) = \pi_2(x_2|x_1') &\triangleq \pi_2(x_2) \text{ (say)} \\ \pi_4(x_4|x_1, x_2, x_3) = \pi_4(x_4|x_1', x_2, x_3) &\triangleq \pi_4(x_4|x_2, x_3) \text{ (say)} \end{aligned} \tag{9}$$

for all $x_1, x_1' = 1, 2, 3$. The simple substitution of Equations (9) into (8) allows us to reduces the number of parameters and of constraints. Indeed the resulting vectors $(\pi_1(1), \pi_1(2), \pi_1(3))$ lie in three-dimensional simplex Δ_2 as do each of the vectors $(\pi_3(1|x_1, x_2), \pi_3(2|x_1, x_2), \pi_3(3|x_1, x_2))$ for $x_1 = 1, 2, 3$ and $x_2 = 1, 2$ whilst the vectors $(\pi_2(1), \pi_2(2))$ and each of the vectors $(\pi_4(1|x_2, x_3), \pi_4(2|x_2, x_3))$ for $x_2 = 1, 2$ and $x_3 = 1, 2, 3$ lies in Δ_1. Each of the 14 simplices also embodies a linear constraint through its sum-to-one condition making the interior of the domain a 21 dimensional linear manifold.

Now, other non-graphical hypotheses are added to the condition independence statements expressed by the Bayesian network. The new hypotheses can still be expressed as a set of algebraic equations or inequalities on the primitive probabilities. We list some for our example.

– If the movie is not watched then we would expect $X_3 = X_1|(X_2 = 2)$, equivalently

$$\pi_3(x_3|x_1, x_2 = 2) = \begin{cases} 1 & \text{if } x_3 = x_1 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

– If a unit did watch the movie, we would not expect this to reduce his testosterone level. This sets some of the primitive probabilities to zero, namely

$X_3\|X_1 = x_1, X_2 = 1$	$x_3 = 1$	$x_3 = 2$	$x_3 = 3$
$x_1 = 1$	$\pi_3(1\|1,1)$	$\pi_3(2\|1,1)$	$\pi_3(3\|1,1)$
$x_1 = 2$	0	$\pi_3(2\|2,1)$	$\pi_3(3\|2,1)$
$x_1 = 3$	0	0	1

(11)

– The assumption that the higher the prior testosterone levels the higher the posterior ones, is given by

$$\begin{aligned} \pi_3(1|2,1) &= r_{3,2}\pi_3(1|1,1) \\ \pi_3(3|1,1) &= r_{3,3}\pi_3(3|2,1) \end{aligned} \tag{12}$$

where $0 \leq r_{3,2}, r_{3,3} \leq 1$ are additional semi parametric parameters.

– Similarly it is reasonable to expect that higher levels of testosterone together with having seen the movie would make more probable that a man would be arrested for fighting. This can be expressed as

$$\begin{aligned} \pi_4(1|1,x_3) &= r_{4,x_3}\pi_4(1|2,x_3) \quad \text{for } x_3 = 1,2,3 \\ \pi_4(1|1,x_3+1) &= r'_{4,x_3}\pi_3(1|1,x_3) \quad \text{for } x_3 = 1,2 \\ \pi_4(1|2,x_3+1) &= r''_{4,x_3}\pi_3(1|2,x_3) \end{aligned}$$

where $0 \leq r_{4,x_3}, r'_{4,x_3}, r''_{4,x_3} \leq 1$, similarly to the previous bullet point.
– Finally a common simple log-linear response model might assume $r_{4,1} = r_{4,2} = r_{4,3}$.

The point here is not that these supplementary equations and inequalities provide the most compelling model, but rather that embellishments of this type, whilst not graphical, are common, are easily expressed in the primitive probability parametrization, and often have an almost identical type of algebraic description as the Bayesian example.

4 (Algebraic) identifiability

Identifiability problems are formulated in the classes of models in Section 3 in the same way. Some polynomial equalities of the transition probabilities, say $m = m(\boldsymbol{\pi})$, are observed or observable together with some inequalities. The interest is in checking whether a total cause (Pearl (2000)) expressed as a function of the post intervention transition probabilities, say $e = e(\widehat{\boldsymbol{\pi}})$, can be written as functions of the m's.

In principle this computation can be done by using elimination techniques from algebraic geometry but, usually, this is not feasible in practice, even less so than in Bayesian networks. If an effect e is identifiable then it is a ratio of polynomials in the observable variables by the Tarski-Seidenberg theorem and could be found by a powerful enough computer.

In Section 5 below, we define a class of models with a strong topological structure and generalise the notion of intervention to these models. Issues of structural identifiability can be discussed via topological arguments in analogy to the back door and front-door criteria for causal Bayesian networks, as well as via a straightforward elimination using techniques from Algebraic Geometry. For particular sub-models of the models in Section 3, the analogue of c-components could be sought but this is beyond the scope of this paper, which aims to illustrate how looking at the formal algebraic aspects of algebraic statistical models and of the notion of causality allows us to capture the mathematical essence of a causal statistical model and allows for a fully general modelling framework freed of the regularity constraints of Bayesian networks.

4.1 Non-regular observables and effects

With refer to the example in Section 3.4 we show that sometimes the manifest/hidden variables are not full marginal, but are rather some (polynomial) function of transition probabilities. Likewise the causal effect that it is necessary to identify might be not a full marginal. Hence an algebraic model of the type discussed in this paper might be appropriate.

Consider collecting data when X_4 is hidden and it is the variable of central interest with its associated probabilities $\pi_4(x_4|x_1, x_2, x_3)$. It might be possible to randomly sample men and measure their testosterone levels before and after watching a violent movie. Call this Experiment 1. However if it were believed that watching a violent movie might induce a fight, it would be unethical to release the subjects after watching the movie, while any therapy either in the form of drugs or counselling will corrupt the experiment. In any case recording the proportions of subjects who later fought would not give an appropriate estimate of probabilities associated with X_4 and conditional on its parents. So values like $\pi_4(2|x_1, 1, x_3)$ cannot be estimated from such samples. To identify the system we therefore need to supplement this type of experiment with another measuring "willingness to fight". Other experiments might be envisaged leading to analogues problems.

Partial information about the joint distribution of X_4 with other variables might be obtained from a random sample of men arrested for fighting $\{x_4 = 1\}$. Their current testosterone levels X_3 and whether they had recently watched a violent movie X_2 could be measured. But we could not measure (X_2, X_3) for men that are not caught fighting. Thus the finest partition of probabilities we could hope for in a population under this kind of survey is based on the sample space partition $\{\overline{A}, A(x_2, x_3\} : x_2 = 1, 2, x_3 = 1, 2, 3\}$ where $\overline{A} = \{\boldsymbol{x} : X_4 = 2\}$ and $A(x_2, x_3) = \{\boldsymbol{x} : X_2 = x_2, X_3 = x_3, X_4 = 1\}$ i.e. $q(\overline{A}) = \sum_{\boldsymbol{x}_i \in \overline{A}} p(\boldsymbol{x})$ and for $x_2 = 1, 2$ and $x_3 = 1, 2, 3$, $q(A(x_2, x_3)) = \sum_{\boldsymbol{x}_i \in A(x_2,x_3)} p(\boldsymbol{x})$. Call this Experiment 2.

In the example the partial order on the nodes of the Bayesian example is $X_1, X_2 \prec X_3$ and $X_1, X_2, X_3 \prec X_4$. But note that, in our statement of the problem, if the man does not watch the movie then by definition $X_1 = X_3$. Under this definition, manipulating X_3 and leaving X_1 unaffected, as would be required by the causal Bayesian network, is not possible. If we follow the two different types of unfoldings of history: {prior testosterone level $X_1 = 1, 2, 3$, watch movie, $X_2 = 1$ posterior testosterone level $X_3 = 1, 2, 3$, arrested $X_4 = 1, 2$} and {prior testosterone level $X_1 = 1, 2, 3$, don't watch movie, $X_2 = 2$, arrested $X_4 = 1, 2$} this sort of ambiguity disappears and we could reasonably conjecture that these unfoldings are consistent with their "causal order". This might be expressed by the two context specific graphs below

$$\begin{array}{ccc} X_1 & \rightarrow X_3 & \qquad\qquad X_1 \\ & \nearrow \ \downarrow & \qquad\qquad\qquad \searrow \\ X_2 = 1 & \rightarrow X_4 & \qquad X_2 = 2 \rightarrow X_4 \end{array}$$

The joint mass function is no longer defined on the product space $S_{\boldsymbol{X}}$ with $X = \{X_1, X_2, X_3, X_4\}$. However the joint mass function of each of these possible unfoldings is well defined and furthermore each unfolding is expressible as a monomial in the primitive probabilities.

5 Chain Event Graphs

Chain Event Graphs are statistical models based on a probability tree, which can have more topological structure than the model classes in Section 3 and still allow a similar algebraic set-up. They are defined through two equivalence relationships on the nodes of a probability tree, $\mathcal{T}$.

Two non-leaf vertices v, v^* of $\mathcal{T}$ are stage equivalent if there is a one-to-one map between the primitive probabilities on the two sets of edges out of the two vertices. They are position equivalent if 1. the two sub-trees rooted at v and v^*, $\mathcal{T}(v)$ and $\mathcal{T}(v^*)$, have the same support/tree structure (call μ this map) and 2. for every non-leaf vertex $w \in \mathcal{T}(v)$, w and $\mu(w)$ are in the same stage.

The chain event graph of a probability tree is a mixed graph whose vertex set is the set of positions union a sink node, whose undirected edges join positions in the same stage and whose directed edges are of two types. Namely, for each edge (v, v^*) in the tree, with v in position w, if v^* is in position w^* then add a directed edge from w to w^* and if v^* is a leaf node then add a directed edge from w to the sink node.

Figure 1 gives a probability tree with thirteen root-to-leaf path. The stage equivalence classes are $\{v_1, v_3, v_{13}, v_{17}\}$, $\{v_5, v_9\}$ and $\{v_2, v_7\}$ leading to the chain event graph with eight position nodes in Figure 2.

Anderson and Smith (2005) discuss conditional independence for Chain Event Graphs and an application to the study of biological regulation models. Propagation of probabilities over chain event graphss has been discussed in Thwaites et al. (2008) and conjugate estimation in Riccomagno and Smith (2005).

6 Manipulations

A manipulation of a probability tree is called positioned (staged) if the set of positions (stages) after manipulation is equal to or a coarsening of the set of positions (stages) before manipulation. Positioned manipulations are those that can be enacted directly on the chain event graphs.

Standard manipulation of a Bayesian network force some components of the network to take preassigned values. The analogue for a chain event graph is a manipulation which forces all paths to pass through a specified set of positions W. E.g. assign patients with particular values of a set of covariates (their current positions) to a particular treatment regime (a set of subsequent positions W).

For two detailed examples see Riccomagno and Smith (2005) and Riccomagno, Smith and Thwaites (2009). Below we present a result in this last

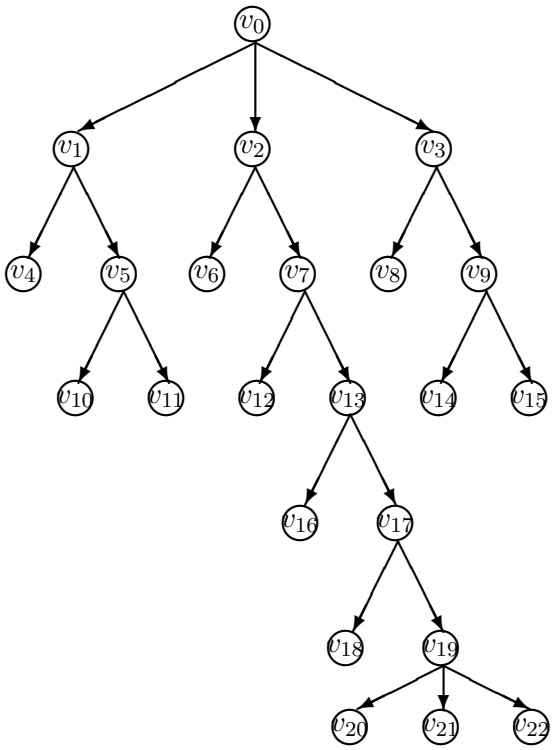

Fig. 1 A probability tree

paper to show how fairly intricated topological arguments lead to a simple elimination formula for the estimation of a causal effect. It works because of how manipulation, observables and effects are defined relative to each other.

7 A back door criterion for chain event graphs

The following is rather technical. This particular example is build in analogy to the back door theorem in Pearl (2000). For a set of positions W in the chain event graphs and $w \in W$ let $C^*(w)$ be the sub-graph from the root to w and $K(C^*(w))$ its positions; $K(C^*(W))$ is $\bigcup_{w \in W} K(C^*(w))$; $\Lambda(w)$ gives

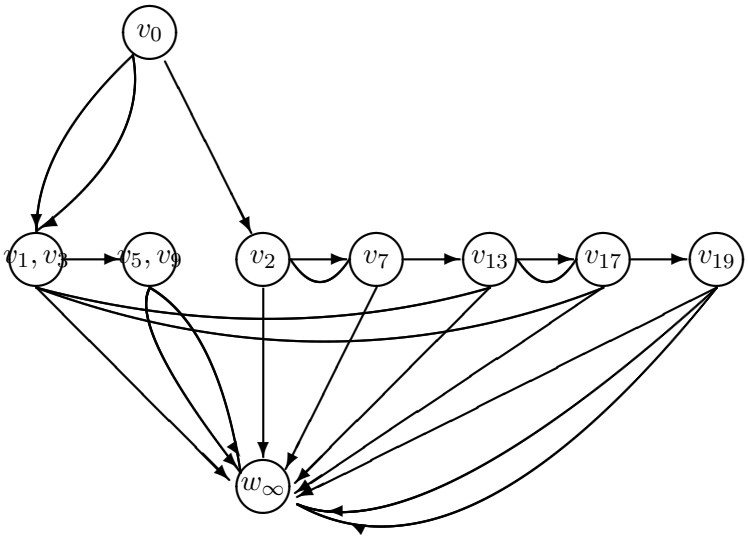

Fig. 2 A chain event graph

the directed paths in the chain event graphs through w, equivalently paths in the tree through nodes in the equivalence class w; finally let $\pi(\Lambda(w))$ be probability of $\Lambda(w)$.

The set W is called simple if 1. no two positions of W lie on the same directed path of the chain event graphs, 2. for $w \in W$ $\pi(\Lambda(w)) = A(w)B(w)$ where $A(w)$ is a function of a set $K^a \subset K(C^*(W))$ and $B(w)$ of $K^b = K(C^*(W)) \setminus K^a$ and 3. $A(w)$ is constant for $w \in W$ (a stands for active, b for background).

A manipulation is called forced to W if 1. it assigns probability one to passing through W and 2. it leaves unchanged the probabilities on edges downstream of w. A manipulation is called amenable forced to W if 1. W is simple in the idle chain event graphs, 2. W is simple in the manipulated chain event graphs and the manipulation assigns probability one to passing through W, and 3. the manipulation changes only the probability of edges in K^a.

Let Ω be the set of root-to-leaf paths in the tree, that is the set of directed root-to-sink paths in the chain event graphs. A random variable $Y : \Omega \to \mathbb{R}$, measurable w.r.t. the (directed) path σ-algebra, partitions Ω into $\{\Lambda_y : y \text{ values taken by } Y\}$.

Let Z (a back door variable) be a random variable on the chain event graphs. A set of positions W is called simple conditioned on Z if 1. no two positions of W lie on the same directed path of the chain event graphs, 2. $W = \bigcup_{z:Z=z} W_z$ and W_z is simple in $C(\Lambda_z)$[1], and 3. there is a (directed) path from each $w'_z \in \Lambda_z$ through a $w^2 \in W$ and W_z is the set of precisely those positions in W which lie on a root-to-sink path including w'_z.

Given a random variable $\hat{Y}$ (an effect) on a chain event graphs after a manipulation forced to w $[W]$, it is possible to construct a random variable Y on the idle chain event graphs that coincides with $\hat{Y}$ on the paths through w $[W]$.

7.1 Back door theorem

If a set W is simple conditioned on Z then the distribution of an effect random variable Y after an amenable manipulation to W is identified from the probabilities (in the idle system) of the event $\{Y = y, W, Z = z : y, z\}$, namely

$$\hat{\pi}(\hat{Y} = y) = \sum_{z:Z=z} \frac{\pi(Y = y, W \mid Z = z)}{\pi(W \mid Z = z)} \pi(Z = z)$$

equivalently

$$\hat{\pi}(\Lambda_y) = \sum_{z:Z=z} \pi(\Lambda_y \mid \Lambda(W), \Lambda_z)\, \pi(\Lambda_z).$$

7.2 Identifying effects algebraically

We end this paper by a short discussion of how the identifiablity issues associated with the non-graphical example of Section 3.4 can be addressed algebraically.

For the example in Section 3.4 assume conditions (10) and (11) hold. Hence, for $x_1 = 1, 2, 3$ the non-zero probabilities associated with not viewing the movie are $p(x_1, 2, x_1, 1) = \pi_1(x_1)\pi_2(2)\pi_4(1|2, x_1)$ and $p(x_1, 2, x_1, 2) = \pi_1(x_1)\pi_2(2)\pi_4(2|2, x_1)$ whilst the probabilities associated with viewing it are given in Table 1.

Consider the two controls described in Section 4.1, manipulating $X_2 = 2$ and $X_1 = X_3 = 1$ respectively. The first, banning the film, gives non-zero probabilities for $x_1 = 1, 2, 3$ satisfying the equations $\widehat{p}(x_1, 2, x_1, 1) = \pi_1(x_1)\pi_4(1|2, x_1)$ and $\widehat{p}(x_1, 2, x_1, 2) = \pi_1(x_1))\pi_4(2|2, x_1)$. The second, the fixing of testosterone levels to low for all time, gives manipulated probabilities

$$\widehat{\widehat{p}}(1, 2, 1, 1) = \pi_2(2)\pi_4(1|2, 1) \qquad \widehat{\widehat{p}}(1, 2, 1, 2) = \pi_2(2)\pi_4(2|2, 1)$$
$$\widehat{\widehat{p}}(1, 1, 1, 1) = \pi_2(1)\pi_4(1|1, 1) \qquad \widehat{\widehat{p}}(1, 1, 1, 2) = \pi_2(1)\pi_4(2|1, 1).$$

[1] This is the sub-chain event graphs including all paths that give $Z = z$.

$$
\begin{aligned}
p(1,1,1,1) &= \pi_1(1)\pi_2(1)\pi_3(1|1,1)\pi_4(1|1,1)\\
p(1,1,1,2) &= \pi_1(1)\pi_2(1)\pi_3(1|1,1)\pi_4(2|1,1)\\
p(1,1,2,1) &= \pi_1(1)\pi_2(1)\pi_3(2|1,1)\pi_4(1|1,2)\\
p(1,1,2,2) &= \pi_1(1)\pi_2(1)\pi_3(2|1,1)\pi_4(2|1,2)\\
p(1,1,3,1) &= \pi_1(1)\pi_2(1)\pi_3(3|1,1)\pi_4(1|1,3)\\
p(1,1,3,2) &= \pi_1(1)\pi_2(1)\pi_3(3|1,1)\pi_4(2|1,3)\\
p(2,1,2,1) &= \pi_1(2)\pi_2(1)\pi_3(2|2,1)\pi_4(1|1,2)\\
p(2,1,2,2) &= \pi_1(2)\pi_2(1)\pi_3(2|2,1)\pi_4(2|1,2)\\
p(2,1,3,1) &= \pi_1(2)\pi_2(1)\pi_3(3|2,1)\pi_4(1|1,3)\\
p(2,1,3,2) &= \pi_1(2)\pi_2(1)\pi_3(3|2,1)\pi_4(2|1,3)\\
p(3,1,3,1) &= \pi_1(3)\pi_2(1)\pi_4(1|1,3)\\
p(3,1,3,2) &= \pi_1(3)\pi_2(1)\pi_4(2|1,3)
\end{aligned}
$$

Table 1 Probabilities associated with viewing the movie

Now consider three experiments. Experiment 1 of Section 4.1 exposes men to the movie, measuring their testosterone levels before and after viewing the film. This obviously provides us with estimates of $\pi_1(x_1)$, for $x_1 = 1, 2, 3$ and $\pi_3(x_3|1, x_1)$ $1 \leq x_1 \leq x_3 \leq 3$. Under Experiment 2 of Section 4.1 a large random large sample is taken over the relevant population providing good estimates of the probability of the margin of each pair of X_2 and the level of testosterone X_3 on those who fought, $\{X_4 = 1\}$, but only the probability of not fighting otherwise. So you can estimate the values of and sample for $x_1 = 1, 2, 3$ $p(x_1, 2, x_1, 1) = \pi_1(x_1)\pi_2(2)\pi_4(1|2, x_1)$ and

$$
\begin{aligned}
p(1,1,1,1) &= \pi_1(1)\pi_2(1)\pi_3(1|1,1)\pi_4(1|1,1)\\
p(1,1,2,1) &= \pi_1(1)\pi_2(1)\pi_3(2|1,1)\pi_4(1|1,2)\\
p(1,1,3,1) &= \pi_1(1)\pi_2(1)\pi_3(3|1,1)\pi_4(1|1,3)\\
p(2,1,2,1) &= \pi_1(2)\pi_2(1)\pi_3(2|2,1)\pi_4(1|1,2)\\
p(2,1,3,1) &= \pi_1(2)\pi_2(1)\pi_3(3|2,1)\pi_4(1|1,3)\\
p(3,1,3,1) &= \pi_1(3)\pi_2(1)\pi_4(1|1,3).
\end{aligned}
$$

Note the last probability is redundant since it is one minus the sum of those given above. Finally Experiment 3 is a survey that informs us about the proportion of people watching the movie on any night, i.e tells us $(\pi_2(1), \pi_2(2))$.

Now suppose we are interested in the total cause

$$
e = \sum_{x_1,x_3} \widehat{p}(x_1, 2, x_3, 1) = \sum_{x_1} \pi_1(x_1)\pi_4(1|2, x_1)
$$

of fighting if forced not to watch. Clearly this is identified from an experiment that includes Experiments 2 and 3 by summing and division by $\pi_2(2)$, but by no other combination of experiments. Similarly $e' = \widehat{\widehat{p}}(1, 1, 1, 1) = \pi_2(1)\pi_4(1|1, 1)$, the probability a man with testosterone levels held low watches the movie and

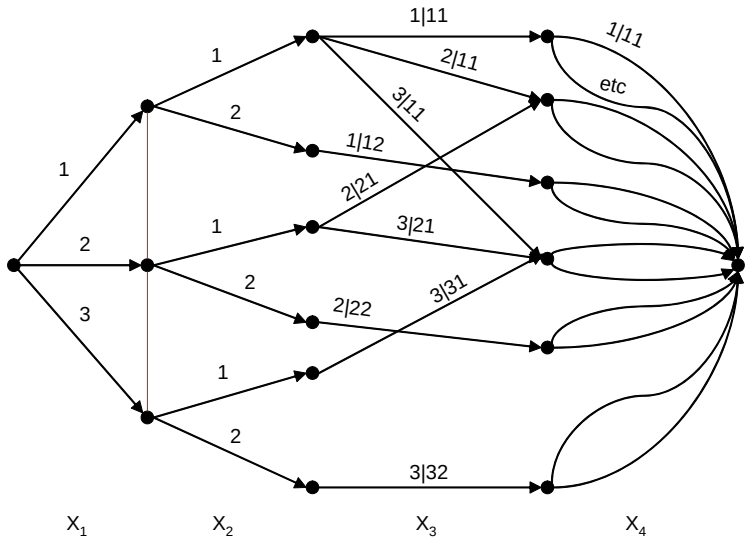

Fig. 3 Chain even graph for the movie example

fights, is identified from $p(1,1,1,1)$ obtained from Experiment 1 and 2 by division.

The chain event graph in Figure 3 illustrate the asymmetries in our movie example and can also portray the possible manipulations. For example the manipulation which sets $X_2 = 2$ simply prunes all the edges that appear only on $(X_2 = 1)$-consistent paths, and assigns a probability of one to $(X_2 = 2)$ edges. The resultant graph is of course consistent with the $\hat{p}$ expressions above. Similarly, the manipulation which sets $X_1 = X_3 = 1$ is easily represented by pruning appropriate edges, and gives $\hat{\hat{p}}$ expressions consistent with those above.

8 Comments

The movie example falls within the general scheme of Section 4.1. Of course a graphical representation of the movie example, e.g. over a tree or even a Bayesian network or even a chain event graph, is possible and useful. But one of the point of this paper is to show that when discussing causal modelling the first step does not need to be the elicitation of a graphical structure whose geometry can then be examined through its underlying algebra. Rather an algebraic formulation based on the identification of the circumstances of interest, e.g. the set V, and the elicitation of a causal order, e.g. the partial order on V, is a more naturally starting point. Clearly in such framework on one hand the graphical type of symmetries embedded and easily visualised on e.g. a Bayesian network are not immediately available but they can be retrieved

(Smith and Anderson (2008)). On the other hand algebraic type of symmetries might be easily spotted and be exploited in the relevant computations.

In this example computation was simple algebraic operation while in more complex case we might need to recur to a computer. Of course the usual difficulties of using current computer code for elimination problems of this kind remain, because inequality constraints are not currently integrated into software and because of the high number of primitive probabilities involved. The advantages of ad-hoc parametrizations apply to these structures based on trees and/or defined algebraically.

References

Allman, E. S. and Rhodes, J. A. (2003). Phylogenetic invariants for the general Markov model of sequence mutation. Math. Biosci. 186, no. 2, 113–144.

Allman, E. S., Ané, C., Rhodes, J. A. (2008). Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. Adv. in Appl. Probab. 40, no. 1, 229–249.

Anderson, P.E., Smith, J.Q. (2005). A graphical framework for representing the semantics of asymmetric models. Technical report 05-12, CRiSM.

Cox, D., Little, J. & OShea, D. (2008). *Ideals, Varieties, and Algorithms* 3edn. Springer-Verlag, New York.

awid, A.P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, 70, 161–189.

Dawid, A.P., Studený, M. (1999). Conditional products: an alternative approach to conditional independence. In D. Heckerman, J. Whittaker (Eds.), *Artificial Intelligence and Statistics 99* (pp. 32–40) Morgan Kaufmann Publishers, S. Francisco.

Drton, M., Sullivant, S. (2007). Algebraic statistical models. Statist. Sinica 17 (2007), no. 4, 1273–1297.

Garcia, L. D., Stillman, M., Sturmfels, B. (2005). Algebraic geometry of Bayesian networks. J. Symbolic Comput. 39, no. 3-4, 331–355.

Glymour, D., Cooper, G.F. (1999). *Computation, Causation, and Discovery*. MIT Press, Cambridge.

Hausman, D. (1998). *Causal asymmetries*. Cambridge University Press, Cambridge.

Kang, C., Tian, J. (2007). Polynomial Constraints in Causal Bayesian Networks. In *Proceedings of the Conference on UAI* (pp. 200–208), AUAI Press.

M. Kuroki, *Graphical identifiability criteria for causal effects in studies with an unobserved treatment/response variable*, Biometrika **94**(1) 37–47, 2007.

Lauritzen, S. (2000). Graphical models for causal inference. In O.E. Barndorff-Nielsen, D. Cox, and C. Kluppelberg (Eds.), *Complex Stochastic Systems* (pp. 67–112), Chapman and Hall/CRC Press, London.

McAllester, D., Collins, M., Periera, F. (2004). Case factor diagrams for structured probabilistic modeling. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 382–391), AUAI Press.

D.M.Q. Mond, J.Q. Smith and D. Van Straten, *Stochastic factorisations, sandwiched simplices and the topology of the space of explanations*, Proc. R. Soc. London. A **459**: 2821-2845, 2003.

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

Riccomagno, E., Smith, J.Q. (2009). The geometry of causal probability trees that are algebraically constrained. In L. Pronzato and A. A. Zhigljavsky (Eds.), *Search for Optimality in Design and Statistics* (pp. 133–154), Springer-Verlag, Berlin.

Riccomagno, E., Smith, J.Q. (2004). Identifying a cause in models which are not simple Bayesian networks. In *IPMU 2004* (pp. 1315–1322).

Riccomagno, E., Smith, J.Q. (2005). The causal manipulation and Bayesian estimation of chain event graphs, http://arxiv.org/abs/0709.3380.

Riccomagno, E., Smith, J.Q., Thwaites, P.A. (submitted). Causal analysis with chain event graphs.

Robins, J.M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality* (pp. 69–117), Springer, New York.

R. Settimi and J.Q. Smith, *Geometry, moments and conditional independence trees with hidden variables*, The Annals of Statistics, **28**(4):1179-1205, 2000.

Shafer, G. (1996). *The Art of Casual Conjecture* MIT Press.

Smith, J.Q., Anderson, P.E. (2008). Conditional independence and Chain Event Graphs. *Artificial Intelligence*, 172, 42–68.

Spirtes, P., Glymour, C., Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

Studený, M. (2004). *Probabilistic Conditional Independence Structures*. Springer-Verlag, London.

Thwaites, P.A., Smith, J.Q., Cowell, R.G. (2008). Propagation using Chain event graphs. In *Proceedings of the 24th Conference on UAI*, Helsinki.

Tian, J., Pearl, J. (2002). On the testable implications of causal models with hidden variables. In Proceedings of UAI 2002, pp.567–573.

Zwiernik, P. and Smith, J.Q. (2009). *The Geometry of Conditional Independence Tree Models with Hidden Variables* (submitted to Annals of Statistics).