Journal of Machine Learning Research 17 (2016) 1-23 Submitted 11/14; Revised 7/15; Published 4/16

Estimating Causal Structure Using Conditional DAG Models

Chris. J. Oates CHRISTOPHER.OATES@UTS.EDU.AU
School of Mathematical and Physical Sciences
University of Technology Sydney
NSW 2007, Australia

Jim Q. Smith J.Q.SMITH@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Sach Mukherjee SACH.MUKHERJEE@DZNE.DE
German Center for Neurodegenerative Diseases (DZNE)
53175 Bonn, Germany

Editor: Samuel Kaski

Abstract

This paper considers inference of causal structure in a class of graphical models called conditional DAGs. These are directed acyclic graph (DAG) models with two kinds of variables, primary and secondary. The secondary variables are used to aid in the estimation of the structure of causal relationships between the primary variables. We prove that, under certain assumptions, such causal structure is identifiable from the joint observational distribution of the primary and secondary variables. We give causal semantics for the model class, put forward a score-based approach for estimation and establish consistency results. Empirical results demonstrate gains compared with formulations that treat all variables on an equal footing, or that ignore secondary variables. The methodology is motivated by applications in biology that involve multiple data types and is illustrated here using simulated data and in an analysis of molecular data from the Cancer Genome Atlas.

Keywords: Graphical Models, Causal Inference, Directed Acyclic Graphs, Instrumental Variables, Data Integration

1. Introduction

This paper considers learning causal structure among a set of *primary variables* $(Y_i)_{i \in V}$, using additional *secondary variables* $(X_i)_{i \in W}$ to aid in estimation. The primary variables are those of direct scientific interest while the secondary variables are variables that are known to influence the primary variables, but whose mutual relationships are not of immediate interest and perhaps not amenable to estimation using the available data. As we discuss further below, the primary/secondary distinction is common in biostatistical applications and is often dealt with in an *ad hoc* manner, for example by leaving some relationships or edges implicit in causal diagrams. Our aim is to define a class of graphical models for this setting and to clarify the conditions under which secondary variables can aid in causal estimation. We focus on causal estimation in the sense of estimation of the presence or absence of edges in the causal graph rather than estimation of quantitative causal effects.

The fact that primary variables of direct interest are often part of a wider context, including additional secondary variables, presents challenges for graphical modelling and causal inference, since in general the secondary variables will not be independent and simply marginalising may introduce spurious dependencies (Evans and Richardson, 2014). Motivated by this observation, we define conditional DAG (CDAG) models and discuss their semantics. Nodes in a CDAG are of two kinds, corresponding to primary and secondary variables, and as detailed below the semantics of CDAGs allow causal inferences to be made about the primary variables $(Y_i)_{i \in V}$ whilst accounting for the secondary variables $(X_i)_{i \in W}$. To limit scope, we focus on the setting where each primary variable has a known cause among the secondary variables. Specifically we suppose there is a bijection $\phi : V \to W$, between the primary and secondary index sets V and W, such that for each $i \in V$ a direct[1] causal dependency $X_{\phi(i)} \to Y_i$ exists. Under explicit assumptions we show that such secondary variables can aid in causal inference for the primary variables, because known relationships between secondary and primary variables render "primary-to-primary" causal links of the form $Y_i \to Y_j$ identifiable from joint data on primary and secondary variables. We put forward score-based estimators of CDAG structure that we show are asymptotically consistent under certain conditions; importantly, independence assumptions on the secondary variables are not needed.

This work is motivated by current efforts in biology aimed at exploiting high-throughput data to better understand causal molecular mechanisms, such as those involved in gene regulation or protein signaling. A notable feature of molecular biology is the fact that some causal links are relatively clearly defined by known sequence specificity. For example, the DNA sequence has a causal influence on the level of corresponding mRNA; mRNAs have a causal influence on corresponding total protein levels; and total protein levels have a causal influence on levels of post-translationally modified protein. This means that in a study involving a certain molecular variable (a protein, say), a subset of the causal influences upon it may be clear at the outset (e.g. the corresponding mRNA) and typically it is the unknown influences that are the subject of the study. Then, it is natural to ask whether accounting for the known influences can aid in identification of the unknown influences. For example, if interest focuses on causal relationships between proteins, known mRNA-protein links could be exploited to aid in causal identification at the protein-protein level. We show below an example using protein data.

Our development of the CDAG can be considered dual to the acyclic directed mixed graphs (ADMGs) developed by Evans and Richardson (2014), in the sense that we investigate conditioning as an alternative to marginalisation. In this respect our work mirrors recently developed undirected graphical models called conditional graphical models (CGMs; Li *et al*, 2012; Cai *et al.*, 2013) In CGMs, Gaussian random variables $(Y_k)_{k \in V}$ satisfy

$$Y_i \perp\!\!\!\perp Y_j | (Y_k)_{k \in V \setminus \{i,j\}}, (X_k)_{k \in W} \text{ if and only if } (i,j) \notin G \tag{1}$$

where G is an undirected acyclic graph and $(X_k)_{k \in W}$ are auxiliary random variables that are conditioned upon. CGMs have recently been applied to gene expression data $(Y_i)_{i \in V}$. In that setting Zhang and Kim (2014) used single nucleotide polymorphisms (SNPs) as the

[1] Throughout, we use the term "direct" in the sense of Pearl (2009) and note that the causal influence need not be *physically* direct.

$(X_i)_{i \in W}$, while Logsdon and Mezey (2010); Yin and Li (2011); Cai *et al.* (2013) used expression qualitative trait loci (e-QTL) as the $(X_i)_{i \in W}$. The latter work was recently extended to jointly estimate several such graphical models in Chun *et al.* (2013). Also in the context of undirected graphs, van Wieringen and van de Wiel (2014) recently considered encoding a bijection between DNA copy number and mRNA expression levels into inference. Our work complements these efforts by using directed models that are arguably more appropriate for causal inference (Pearl, 2009). Evans (2015) uses a similar directed graphical characterisation and nomenclature. However, secondary variables in Evans (2015) are unobserved and can have multiple children in the set of primary variables, while ours are observed and have exactly one child among the primary variables. CDAGs are also related to instrumental variables and Mendelian randomisation approaches (Didelez and Sheehan, 2007) that we discuss below (Section 2.2).

CDAGs share some similarity with the influence diagrams (IDs) introduced by Dawid (2002) as an extension of DAGs that distinguish between variable nodes and decision nodes. This generalised the augmented DAGs of Spirtes *et al.* (2000); Lauritzen (2000); Pearl (2009) in which each variable node is associated with a decision node that represents an intervention on the corresponding variable. However, the semantics of IDs are not well suited to the scientific contexts that we consider, where secondary nodes represent variables to be observed, not the outcomes of decisions. The notion of a non-atomic intervention (Pearl, 2003), where many variables are intervened upon simultaneously, shares similarity with CDAGs in the sense that the secondary variables are in general non-independent. However again the semantics differ, since our secondary nodes represent random variables rather than interventions. In a different direction, Neto *et al.* (2010) recently observed that the use of e-QTL data $(X_i)_{i \in W}$ can help to identify causal relationships among gene expression levels $(Y_i)_{i \in V}$. However, Neto *et al.* (2010) require independence of the $(X_i)_{i \in W}$; this is too restrictive for general settings, including in molecular biology, since the secondary variables will typically themselves be subject to regulation and far from independent. An important and novel aspect of our approach is that no independence or conditional independence assumptions need to be placed on the secondary variables in order to draw causal conclusions about the primary variables.

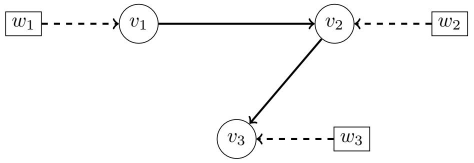
This paper begins in Sec. 2 by defining CDAGs and discussing identifiability of their structure from observational data on primary and secondary variables. Sufficient conditions are then given for consistent estimation of CDAG structure along with an algorithm based on integer linear programming. The methodology is illustrated in Section 3 on simulated data, including data sets that violate CDAG assumptions, and on protein data from cancer samples, the latter from the Cancer Genome Atlas (TCGA).

2. Methodology

Below we define CDAGS, study their theoretical properties and propose consistent estimators for CDAG structure.

2.1 A Statistical Framework for Conditional DAG Models

Consider index sets V, W and a bijection $\phi : V \to W$ between them. We will distinguish between the nodes in graphs and the random variables (RVs) that they represent. Specif-



(a)

type	index	node	variable
primary	$i \in V$	$v_i \in N(V)$	Y_i
secondary	$i \in W$	$w_i \in N(W)$	X_i

(b)

Figure 1: A conditional DAG model with primary nodes $N(V) = \{v_1, v_2, v_3\}$ and secondary nodes $N(W) = \{w_1, w_2, w_3\}$. Here primary nodes represent primary random variables $(Y_i)_{i \in V}$ and solid arrows correspond to a DAG G on these vertices. Square nodes are secondary variables $(X_i)_{i \in W}$ that, in the causal interpretation of CDAGs, represent known direct causes of the corresponding $(Y_i)_{i \in V}$ (dashed arrows represent known relationships; the random variables $(X_i)_{i \in W}$ need not be independent). The name conditional DAG refers to the fact that conditional upon $(X_i)_{i \in W}$, the solid arrows encode conditional independence relationships among the $(Y_i)_{i \in V}$.

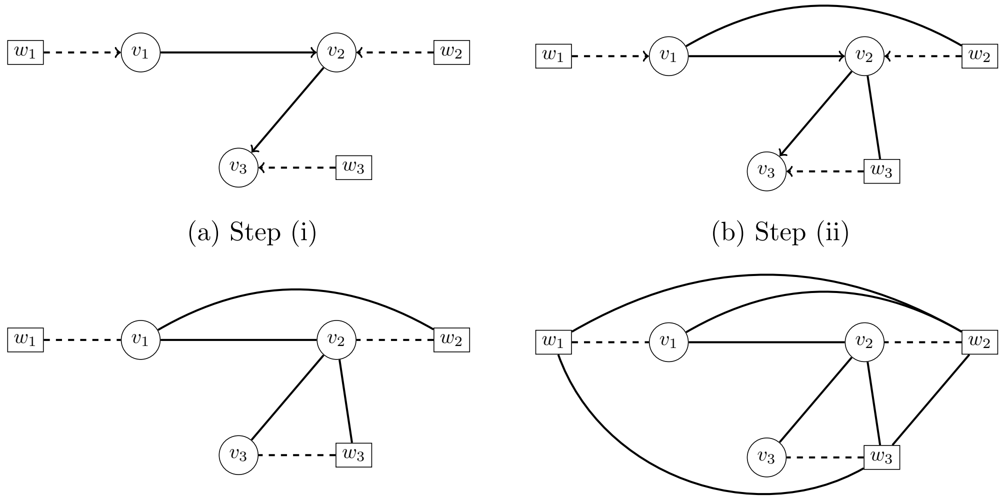
ically, indices correspond to nodes in graphical models; this is signified by the notation $N(V) = \{v_1, \ldots, v_p\}$ and $N(W) = \{w_1, \ldots, w_p\}$. Each node $v_i \in N(V)$ corresponds to a primary RV Y_i and similarly each node $w_i \in N(W)$ corresponds to a secondary RV X_i.

Definition 1 (CDAG) *A* conditional DAG *(CDAG) $\overline{G}$, with primary and secondary index sets V, W respectively and a bijection ϕ between them, is a DAG on the primary node set $N(V)$ with additional directed edges from each secondary node $w_{\phi(i)} \in N(W)$ to its corresponding primary node $v_i \in N(V)$.*

In other words, a CDAG $\overline{G}$ has node set $N(V) \cup N(W)$ and an edge set that can be generated by starting with a DAG on the primary nodes $N(V)$ and adding a directed edge from each secondary node in $N(W)$ to its corresponding primary node in $N(V)$, with the correspondence specified by the bijection ϕ. An example of a CDAG is shown in Fig. 1a. To further distinguish V and W in the graphical representation we employ circular and square vertices respectively. In addition we use dashed lines to represent edges that are required by definition and must therefore be present in any CDAG $\overline{G}$.

Since the DAG on the primary nodes $N(V)$ is of particular interest, throughout we use G to denote a DAG on $N(V)$. We use $\mathcal{G}$ to denote the set of all possible DAGs with $|V|$ vertices. For notational clarity, and without loss of generality, below we take the bijection to simply be the identity map $\phi(i) = i$. The parents of node v_i in a DAG G are indexed by $\text{pa}_G(i) \subseteq V \setminus \{i\}$. Write $\text{an}_{\overline{G}}(S)$ for the ancestors of nodes $S \subseteq N(V) \cup N(W)$ in the CDAG $\overline{G}$ (which by definition includes the nodes in S). For disjoint sets of nodes A, B, C in an undirected graph, we say that C *separates* A and B if every path between a node in A and a node in B in the graph contains a node in C.

Definition 2 (c-separation) *Consider disjoint $A, B, C \subseteq N(V) \cup N(W)$ and a CDAG $\overline{G}$. We say that A and B are* c-separated *by C in $\overline{G}$, written $A \perp\!\!\!\perp B | C \; [\overline{G}]$, when C separates*



(a) Step (i)

(b) Step (ii)

(c) Step (iii)

(d) Step (iv)

Figure 2: Illustrating c-separation. Here we ask whether v_3 and v_1 are c-separated by $\{v_2\}$ in $\overline{G}$, the CDAG shown in Fig. 1a. [Step (i): Take the subgraph induced by $\mathrm{an}_{\overline{G}}(\{v_1, v_2, v_3\})$. Step (ii): Moralise this subgraph (i.e. join with an undirected edge any parents of a common child that are not already connected by a directed edge). Step (iii): Take the skeleton of the moralised subgraph (i.e. remove the arrowheads). Step (iv): Add an undirected edge between every pair (w_i, w_j). In the final panel (d) we ask whether there exists a path from v_3 to v_1 that does not pass through v_2; the answer is positive (e.g. $v_3 - w_3 - w_1 - v_1$) and hence we conclude that $v_3 \not\perp\!\!\!\perp v_1 | v_2 [\overline{G}]$, i.e. v_3 and v_1 are not c-separated by $\{v_2\}$ in $\overline{G}$.]

A and B in the undirected graph U_4 that is formed as follows: (i) Take the subgraph U_1 induced by $an_{\overline{G}}(A \cup B \cup C)$. (ii) Moralise U_1 to obtain U_2 (i.e. join with an undirected edge any parents of a common child that are not already connected by a directed edge). (iii) Take the skeleton of the moralised subgraph U_2 to obtain U_3 (i.e. remove the arrowheads). (iv) Add an undirected edge between every pair of nodes in $N(W)$ to obtain U_4.

The c-separation procedure is illustrated in Fig. 2, where we show that v_3 is not c-separated from v_1 by the set $\{v_2\}$ in the CDAG from Fig. 1a.

Remark 3 *The classical notion of d-separation for DAGs is equivalent to omitting step (iv) in c-separation. Notice that v_3* is *d-separated from v_1 by the set $\{v_2\}$ in the DAG G, underlining the need for a custom notion of separation for CDAGs.*

The topology (i.e. the set of edges) of a CDAG carries formal (potentially causal) semantics on the primary RVs, conditional on the secondary RVs, as specified below. Write $T(S)$ for the collection of triples $\langle A, B | C \rangle$ where A, B, C are disjoint subsets of S.

Definition 4 (Independence model) *The CDAG $\overline{G}$, together with c-separation, implies a formal* independence model *(p.12 of Studený, 2005)*

$$\mathcal{M}_G = \{\langle A, B|C\rangle \in T(N(V) \cup N(W)) : A \perp\!\!\!\perp B|C \ [\overline{G}]\} \tag{2}$$

where $\langle A, B|C\rangle \in \mathcal{M}_G$ carries the interpretation that the RVs corresponding to A are conditionally independent of the RVs corresponding to B when given the RVs corresponding to C. We will write $A \perp\!\!\!\perp B|C \ [\mathcal{M}_G]$ as a shorthand for $\langle A, B|C\rangle \in \mathcal{M}_G$.

Remark 5 *An independence model $\mathcal{M}_G$ does not contain any information on the structure of the marginal distribution $\mathbb{P}^{(X_i)}$ of the secondary variables, due to the additional step (iv) in c-separation.*

Lemma 6 (Equivalence classes) *The map $G \mapsto \mathcal{M}_G$ is an injection.*

Proof Consider two distinct DAGs $G, H \in \mathcal{G}$ and suppose that, without loss of generality, the edge $v_i \to v_j$ belongs to G and not to H. It suffices to show that $\mathcal{M}_G \neq \mathcal{M}_H$. First notice that G represents the relations (i) $w_i \not\perp\!\!\!\perp v_j|w_j \ [\overline{G}]$, (ii) $w_i \not\perp\!\!\!\perp v_j|w_j, (v_k)_{k\in V\setminus\{i,j\}} \ [\overline{G}]$, and (iii) $w_j \perp\!\!\!\perp v_i|w_i \ [\overline{G}]$. (These can each be directly verified by c-separation.) We show below that H cannot also represent (i-iii) and hence, from Def. 4, it follows that $\mathcal{M}_G \neq \mathcal{M}_H$. We distinguish between two cases for H, namely (a) $v_i \leftarrow v_j \notin H$, and (b) $v_i \leftarrow v_j \in H$.

Case (a): Suppose (i) also holds for H; that is, $w_i \not\perp\!\!\!\perp v_j|w_j \ [\overline{H}]$. Then since $v_i \to v_j \notin H$, it follows from c-separation that the variable v_i must be connected to v_j by directed path(s) whose interior vertices must only belong to $N(V) \setminus \{v_i, v_j\}$. Thus H implies the relation $w_i \perp\!\!\!\perp v_j|w_j, (v_k)_{k\in V\setminus\{i,j\}} \ [\overline{H}]$, so that (ii) cannot also hold.

Case (b): Since $v_i \leftarrow v_j \in H$ it follows from c-separation that $w_j \not\perp\!\!\!\perp v_i|w_i \ [\overline{H}]$, so that (iii) does not hold for H. ■

Remark 7 *More generally the same argument shows that a DAG $G \in \mathcal{G}$ satisfies $v_i \to v_j \notin G$ if and only if $\exists S \subseteq pa_G(j) \setminus \{i\}$ such that $w_i \perp\!\!\!\perp v_j|w_j, (v_k)_{k\in S} \ [\overline{G}]$. As a consequence, we have the interpretation that conditional upon the (secondary variables) $(X_i)_{i\in W}$, the solid arrows in Fig. 1a encode conditional independence relationships among the (primary variables) $(Y_i)_{i\in V}$, motivating the CDAG nomenclature.*

It is well known that conditional independence (and causal) relationships can usefully be described through a qualitative, graphical representation. However to be able to use a graph for *reasoning* it is necessary for that graph to embody certain assertions that themselves obey a logical calculus. Pearl and Verma (1987) proposed such a set of rules (the semi-graphoid axioms) that any reasonable set of assertions about how one set of variables might be irrelevant to the prediction of a second, given the values of a third, might hold (see also Dawid, 2001; Studený, 2005). This can then be extended to causal assertions (Pearl, 2009), thereby permitting study of causal hypotheses and their consequences without the need to first construct elaborate probability spaces and their extensions under control. Below we establish that the independence models $\mathcal{M}_G$ induced by c-separation on CDAGs are semi-graphoids and thus enable reasoning in the present setting with two kinds of variables:

Lemma 8 (Semi-graphoid) *For any DAG $G \in \mathcal{G}$, the set $\mathcal{M}_G$ is semi-graphoid (Pearl and Paz, 1985). That is to say, for all disjoint $A, B, C, D \subseteq N(V) \cup N(W)$ we have*

(i) (triviality) $A \perp\!\!\!\perp \emptyset | C \; [\mathcal{M}_G]$

(ii) (symmetry) $A \perp\!\!\!\perp B | C \; [\mathcal{M}_G]$ *implies* $B \perp\!\!\!\perp A | C \; [\mathcal{M}_G]$

(iii) (decomposition) $A \perp\!\!\!\perp B, D | C \; [\mathcal{M}_G]$ *implies* $A \perp\!\!\!\perp D | C \; [\mathcal{M}_G]$

(iv) (weak union) $A \perp\!\!\!\perp B, D | C \; [\mathcal{M}_G]$ *implies* $A \perp\!\!\!\perp B | C, D \; [\mathcal{M}_G]$

(v) (contraction) $A \perp\!\!\!\perp B | C, D \; [\mathcal{M}_G]$ *and* $A \perp\!\!\!\perp D | C \; [\mathcal{M}_G]$ *implies* $A \perp\!\!\!\perp B, D | C \; [\mathcal{M}_G]$

Proof The simplest proof is to note that our notion of c-separation is equivalent to classical d-separation applied to an extension $\underline{G}$ of the CDAG $\overline{G}$. The semi-graphoid properties then follow immediately by the facts that (i) d-separation satisfies the semi-graphoid properties (p.48 of Studený, 2005), and (ii) the restriction of a semi-graphoid to a subset of vertices is itself a semi-graphoid (p.14 of Studený, 2005).

Construct an extended graph $\underline{G}$ from the CDAG $\overline{G}$ by the addition of a node z and directed edges from z to each of the secondary vertices $N(W)$. Then for disjoint $A, B, C \subseteq N(V) \cup N(W)$ we have that A and B are c-separated by C in $\overline{G}$ if and only if A and B are d-separated by C in $\underline{G}$. This is because every path in the undirected graph $U_4(\overline{G})$ (recall the definition of c-separation) that contains an edge $w_i \to w_j$ corresponds uniquely to a path in $U_3(\underline{G})$ that contains the sub-path $w_i \to z \to w_j$. ■

2.2 Causal CDAG Models

The previous section defined CDAG models using the framework of formal independence models. However, CDAGs can also be embellished with a causal interpretation, that we make explicit below. In this paper we make a causal sufficiency assumption that the $(X_i)_{i \in W}$ are the only source of confounding for the $(Y_i)_{i \in V}$ and below we talk about direct causes at the level of $(X_i)_{i \in W} \cup (Y_i)_{i \in V}$.

Definition 9 (Causal CDAG) *A CDAG is* causal *when an edge $v_i \to v_j$ exists if and only if Y_i is a direct cause of Y_j. It is further assumed that X_i is a direct cause of Y_i and not a direct cause of Y_j for $j \neq i$. Finally it is assumed that no Y_i is a direct cause of any X_j.*

Remark 10 *Here* direct cause *is understood to mean that the parent variable has a "controlled direct effect" on the child variable in the framework of Pearl (e.g. Def. 4.5.1 of Pearl, 2009) (it is not necessary that the effect is physically direct). No causal assumptions are placed on interaction between the secondary variables $(X_i)_{i \in W}$.*

Remark 11 *In a causal CDAG the secondary variables $(X_i)_{i \in W}$ share some of the properties of instrumental variables (Didelez and Sheehan, 2007). Consider estimating the average causal effect of Y_i on Y_j. Then, conditioning on X_j in the following, X_i can be used as a natural experiment (Greenland, 2000) to determine the size and sign of this causal effect.*

When we are interested in the controlled direct effect, we can repeat this argument with additional conditioning on the $(Y_k)_{k \in V \setminus \{i,j\}}$ *(or a smaller subset of conditioning variables if the structure of* G *is known).*

2.3 Identifiability of CDAGs

There exist well-known identifiability results for independence models $\mathcal{M}$ that are induced by Bayesian networks; see for example Spirtes *et al.* (2000); Pearl (2009). These relate the observational distribution $\mathbb{P}^{(Y_i)}$ of the random variables $(Y_i)_{i \in V}$ to an appropriate DAG representation by means of d-separation, Markov and faithfulness assumptions (discussed below). The problem of identification for CDAGs is complicated by the fact that (i) the primary variables $(Y_i)_{i \in V}$ are insufficient for identification, (ii) the joint distribution $\mathbb{P}^{(X_i) \cup (Y_i)}$ of the primary variables $(Y_i)_{i \in V}$ and the secondary variables $(X_i)_{i \in W}$ need not be Markov with respect to the CDAG $\overline{G}$, and (iii) we must work with the alternative notion of c-separation. Below we propose novel "partial" Markov and faithfulness conditions that will permit, in the next section, an identifiability theorem for CDAGs. We make the standard assumption that there is no selection bias (for example by conditioning on common effects). We also assume throughout that there exists a true CDAG $\overline{G}$. In other words, the observational distribution $\mathbb{P}^{(X_i) \cup (Y_i)}$ induces an independence model that can be expressed as $\mathcal{M}_G$ for some DAG $G \in \mathcal{G}$.

Definition 12 (Partial Markov) *Let* G *denote the true DAG. We say that the observational distribution* $\mathbb{P}^{(X_i) \cup (Y_i)}$ *is* partially Markov *with respect to* G *when the following holds: For all disjoint subsets* $\{i\}, \{j\}, C \subseteq \{1, \ldots, p\}$ *we have* $w_i \perp\!\!\!\perp v_j | w_j, (v_k)_{k \in C} \; [\mathcal{M}_G] \;\; \Rightarrow \;\; X_i \perp\!\!\!\perp Y_j | X_j, (Y_k)_{k \in C}$.

Definition 13 (Partial faithfulness) *Let* G *denote the true DAG. We say that the observational distribution* $\mathbb{P}^{(X_i) \cup (Y_i)}$ *is* partially faithful *with respect to* G *when the following holds: For all disjoint subsets* $\{i\}, \{j\}, C \subseteq \{1, \ldots, p\}$ *we have* $w_i \perp\!\!\!\perp v_j | w_j, (v_k)_{k \in C} \; [\mathcal{M}_G] \Leftarrow X_i \perp\!\!\!\perp Y_j | X_j, (Y_k)_{k \in C}$

Remark 14 *The partial Markov property implies that* $\mathbb{P}^{(X_i) \cup (Y_i)}$ *admits the factorisation*

$$p((x_i)_{i=1,\ldots,p}, (y_i)_{i=1,\ldots,p}) = p((x_i)_{i=1,\ldots,p}) \prod_{i=1}^{p} p(y_i | (y_k)_{k \in pa_G(i)}, x_i), \tag{3}$$

while partial faithfulness ensures that none of the terms inside the product in Eqn. 3 can be further decomposed. In this work we will prove only estimation consistency; for more refined convergence analysis the partial faithfulness property must be strengthened. See e.g. Uhler et al. (2013) for a recent discussion of faithfulness in the DAG setting.

Remark 15 *The partial Markov and partial faithfulness properties do not place any constraint on the marginal distribution* $\mathbb{P}^{(X_i)}$ *of the secondary variables.*

The following is an immediate corollary of Lem. 6:

Theorem 16 (Identifiability) *Suppose that the observational distribution* $\mathbb{P}^{(X_i) \cup (Y_i)}$ *is partially Markov and partially faithful with respect to the true DAG* G*. Then* *(i) It is not possible to identify the true DAG G based on the observational distribution $\mathbb{P}^{(Y_i)}$ of the primary variables alone.*

(ii) It is possible to identify the true DAG G based on the observational distribution $\mathbb{P}^{(X_i) \cup (Y_i)}$.

Proof (i) We have already seen that $\mathbb{P}^{(Y_i)}$ is not Markov with respect to the DAG G: Indeed a statistical association $Y_i \not\perp\!\!\!\perp Y_j | (Y_k)_{k \in V \setminus \{i,j\}}$ observed in the distribution $\mathbb{P}^{(Y_i)}$ could either be due to a direct interaction $Y_i \to Y_j$ (or $Y_j \to Y_i$), or could be mediated entirely through variation in the secondary variables $(X_k)_{k \in W}$. (ii) It follows immediately from Lemma 6 that observation of both the primary and secondary variables $(Y_i)_{i \in V} \cup (X_i)_{i \in W}$ is sufficient for identification of G. ■

2.4 Estimating CDAGs From Data

In this section we assume that the partial Markov and partial faithfulness properties hold, so that the true DAG G is identifiable from the joint observational distribution of the primary and secondary variables. Below we consider score-based estimation for CDAGs and prove consistency of certain score-based CDAG estimators.

Definition 17 (Score function; Chickering (2003)) *A* score function *is a map $S : \mathcal{G} \to [0, \infty)$ with the interpretation that if two DAGs $G, H \in \mathcal{G}$ satisfy $S(G) < S(H)$ then H is preferred to G.*

We will study the asymptotic behaviour of $\hat{G}_S$, the estimate of graph structure obtained by maximising $S(G)$ over all $G \in \mathcal{G}$ based on observations $(X_i^j, Y_i^j)_{i=1,\dots,p}^{j=1,\dots,n}$. Let $\mathbb{P}_n = \mathbb{P}^{(X_i^j, Y_i^j)}$ denote the finite-dimensional distribution of the n observations.

Definition 18 (Partial local consistency) *We say the score function S is* partially locally consistent *if, whenever H is constructed from G by the addition of one edge $Y_i \to Y_j$, we have*

1. $X_i \not\perp\!\!\!\perp Y_j | X_j, (Y_k)_{k \in pa_G(j)} \Rightarrow \lim_{n \to \infty} \mathbb{P}_n[S(H) > S(G)] = 1$
2. $X_i \perp\!\!\!\perp Y_j | X_j, (Y_k)_{k \in pa_G(j)} \Rightarrow \lim_{n \to \infty} \mathbb{P}_n[S(H) < S(G)] = 1$.

Theorem 19 (Consistency) *If S is partially locally consistent then $\lim_{n \to \infty} \mathbb{P}_n[\hat{G}_S = G] = 1$, so that $\hat{G}_S$ is a consistent estimator of the true DAG G.*

Proof It suffices to show that $\lim_{n \to \infty} \mathbb{P}_n[\hat{G}_S = H] = 0$ whenever $H \neq G$. There are two cases to consider:

Case (a): Suppose $v_i \to v_j \in H$ but $v_i \to v_j \notin G$. Let H' be obtained from H by the removal of $v_i \to v_j$. From c-separation we have $w_i \perp\!\!\!\perp v_j | w_j, (v_k)_{k \in \mathrm{pa}_G(j)}$ $[\overline{G}]$ and hence from the partial Markov property we have $X_i \perp\!\!\!\perp Y_j | X_j, (Y_k)_{k \in \mathrm{pa}_G(j)}$. Therefore if S is partially locally consistent then $\lim_{n \to \infty} \mathbb{P}_n[S(H) < S(H')] = 1$, so that $\lim_{n \to \infty} \mathbb{P}_n[\hat{G}_S = H] = 0$.

Case (b): Suppose $v_i \to v_j \notin H$ but $v_i \to v_j \in G$. Let H' be obtained from H by the addition of $v_i \to v_j$. From c-separation we have $w_i \not\perp\!\!\!\perp v_j | w_j, (v_k)_{k \in \mathrm{pa}_G(j)}$ $[\overline{G}]$ and hence from the partial faithfulness property we have $X_i \not\perp\!\!\!\perp Y_j | X_j, (Y_k)_{k \in \mathrm{pa}_G(j)}$. Therefore if S is partially locally consistent then $\lim_{n\to\infty} \mathbb{P}_n[S(H) < S(H')] = 1$, so that $\lim_{n\to\infty} \mathbb{P}_n[\hat{G}_S = H] = 0$. ■

Remark 20 *In this paper we adopt a maximum a posteriori (MAP) -Bayesian approach and consider score functions given by a posterior probability $p(G|(x_i^l, y_i^l)_{i=1,\dots,p}^{l=1,\dots,n})$ of the DAG G given the data $(x_i^l, y_i^l)_{i=1,\dots,p}^{l=1,\dots,n}$. This requires that a prior $p(G)$ is specified over the space $\mathcal{G}$ of DAGs. From the partial Markov property we have that, for n independent observations, such score functions factorise as*

$$S(G) = p(G) \prod_{l=1}^{n} p((x_i^l)_{i=1,\dots,p}) \prod_{i=1}^{p} p(y_i^l | (y_k^l)_{k \in pa_G(i)}, x_i^l). \tag{4}$$

We further assume that the DAG prior $p(G)$ factorises over parent sets $pa_G(i) \subseteq V \setminus \{i\}$ as

$$p(G) = \prod_{i=1}^{p} p(pa_G(i)). \tag{5}$$

This implies that the score function in Eqn. 4 is decomposable *and the maximiser $\hat{G}_S$, i.e. the MAP estimate, can be obtained via integer linear programming (ILP). In Sec. 2.6 we derive an ILP that targets the CDAG $\hat{G}_S$ and thereby allows exact (i.e. deterministic) estimation in this class of models.*

Lemma 21 *A score function of the form Eqn. 4 is partially locally consistent if and only if, whenever H is constructed from G by the addition of one edge $v_i \to v_j$, we have*

1. $X_i \not\perp\!\!\!\perp Y_j | X_j, (Y_k)_{k \in pa_G(j)} \Rightarrow \lim_{n\to\infty} \mathbb{P}_n[B_{H,G} > 1] = 1$
2. $X_i \perp\!\!\!\perp Y_j | X_j, (Y_k)_{k \in pa_G(j)} \Rightarrow \lim_{n\to\infty} \mathbb{P}_n[B_{H,G} < 1] = 1$

where

$$B_{H,G} = \frac{p((Y_j^l)^{l=1,\dots,n} | (Y_k^l)_{k \in pa_H(j)}^{l=1,\dots,n}, (X_j^l)^{l=1,\dots,n})}{p((Y_j^l)^{l=1,\dots,n} | (Y_k^l)_{k \in pa_G(j)}^{l=1,\dots,n}, (X_j^l)^{l=1,\dots,n})} \tag{6}$$

is the Bayes factor between two competing local models $pa_G(j)$ and $pa_H(j)$.

2.5 Bayes Factors and Common Variables

The characterisation in Lemma 21 justifies the use of any consistent Bayesian variable selection procedure to obtain a score function. To be clear, although our local scores derive from the variable selection literature, we do *not* perform node-wise variable selection, which would not produce a DAG in general. Instead, the local scores form the basis for a global search over DAGs via ILPs; see Sec. 2.6.

The secondary variables $(X_i^l)^{l=1,\dots,n}$ are included in all models, and parameters relating to these variables should therefore share a common prior. Below we discuss a formulation of the Bayesian linear model that is suitable for CDAGs. Consider variable selection for node j and candidate parent (index) set $\mathrm{pa}_G(j) = \pi \subseteq V \setminus \{j\}$. We construct a linear model for the observations

$$Y_j^l = [1 \; X_j^l]\boldsymbol{\beta}_0 + \boldsymbol{Y}_\pi^l \boldsymbol{\beta}_\pi + \epsilon_j^l, \quad \epsilon_j^l \sim N(0, \sigma^2) \tag{7}$$

where $\boldsymbol{Y}_\pi^l = (Y_k^l)_{k \in \pi}$ is used to denote a row vector and the noise ϵ_j^l is assumed independent for $j = 1, \dots, p$ and $l = 1, \dots, n$. Although suppressed in the notation, the parameters $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_\pi$ and σ are specific to node j. This regression model can be written in vectorised form as

$$\boldsymbol{Y}_j = \boldsymbol{M}_0 \boldsymbol{\beta}_0 + \boldsymbol{Y}_\pi \boldsymbol{\beta}_\pi + \boldsymbol{\epsilon} \tag{8}$$

where $\boldsymbol{M}_0$ is the $n \times 2$ matrix whose rows are the $[1 \; X_j^l]$ for $l = 1, \dots, n$ and $\boldsymbol{Y}_\pi$ is the matrix whose rows are $\boldsymbol{Y}_\pi^l$ for $l = 1, \dots, n$.

We orthogonalize the regression problem by defining $\boldsymbol{M}_\pi = (\boldsymbol{I} - \boldsymbol{M}_0(\boldsymbol{M}_0^T \boldsymbol{M}_0)^{-1}\boldsymbol{M}_0^T)\boldsymbol{Y}_\pi$ so that the model can be written as

$$\boldsymbol{Y}_j = \boldsymbol{M}_0 \tilde{\boldsymbol{\beta}}_0 + \boldsymbol{M}_\pi \tilde{\boldsymbol{\beta}}_\pi + \boldsymbol{\epsilon} \tag{9}$$

where $\tilde{\boldsymbol{\beta}}_0$ and $\tilde{\boldsymbol{\beta}}_\pi$ are Fisher orthogonal parameters (see Deltell, 2011, for details).

In the conventional approach of Jeffreys (1961), the prior distribution is taken as

$$\begin{aligned} p_{j,\pi}(\tilde{\boldsymbol{\beta}}_\pi, \tilde{\boldsymbol{\beta}}_0, \sigma) &= p_{j,\pi}(\tilde{\boldsymbol{\beta}}_\pi | \tilde{\boldsymbol{\beta}}_0, \sigma) p_j(\tilde{\boldsymbol{\beta}}_0, \sigma) & (10)\\ p_j(\tilde{\boldsymbol{\beta}}_0, \sigma) &\propto \sigma^{-1} & (11) \end{aligned}$$

where Eqn. 11 is the reference or independent Jeffreys prior. (For simplicity of exposition we leave conditioning upon $\boldsymbol{M}_0$, and $\boldsymbol{M}_\pi$ implicit.) The use of the reference prior here is motivated by the observation that the common parameters $\boldsymbol{\beta}_0, \sigma$ have the same meaning in each model π for variable Y_j and should therefore share a common prior distribution (Jeffreys, 1961). Alternatively, the prior can be motivated by invariance arguments that derive $p(\boldsymbol{\beta}_0, \sigma)$ as a right Haar measure (Bayarri *et al.*, 2012). Note however that σ does not carry the same meaning across $j \in V$ in the application that we consider below, so that the prior is specific to fixed j. For the parameter prior $p_{j,\pi}(\tilde{\boldsymbol{\beta}}_\pi | \tilde{\boldsymbol{\beta}}_0, \sigma)$ we use the g-prior (Zellner, 1986)

$$\tilde{\boldsymbol{\beta}}_\pi | \tilde{\boldsymbol{\beta}}_0, j, \pi \sim N(\boldsymbol{0}, g\sigma^2(\boldsymbol{M}_\pi^T \boldsymbol{M}_\pi)^{-1}) \tag{12}$$

where g is a positive constant to be specified. Due to orthogonalisation, $\mathrm{cov}(\hat{\boldsymbol{\beta}}_\pi) = \sigma^2(\boldsymbol{M}_\pi^T \boldsymbol{M}_\pi)^{-1}$ where $\hat{\boldsymbol{\beta}}_\pi$ is the maximum likelihood estimator for $\tilde{\boldsymbol{\beta}}_\pi$, so that the prior is specified on the correct length scale (Deltell, 2011). We note that many alternatives to Eqn. 12 are available in the literature (including Johnson and Rossell, 2010; Bayarri *et al.*, 2012). For discrete data we mention recent work by Massam and Wesołowski (2014).

Under the prior specification above, the marginal likelihood for a candidate model π has the following closed-form expression:

$$\begin{aligned} p_j(\boldsymbol{y}_j|\pi) &= \frac{1}{2}\Gamma\left(\frac{n-2}{2}\right)\frac{1}{\pi^{(n-2)/2}}\frac{1}{|\boldsymbol{M}_0^T\boldsymbol{M}_0|^{1/2}}\left(\frac{1}{g+1}\right)^{|\pi|/2} b^{-(n-2)/2} & (13)\\ b &= \boldsymbol{y}_j^T\left(\boldsymbol{I} - \boldsymbol{M}_0(\boldsymbol{M}_0^T\boldsymbol{M}_0)^{-1}\boldsymbol{M}_0^T - \frac{g}{g+1}\boldsymbol{M}_\pi(\boldsymbol{M}_\pi^T\boldsymbol{M}_\pi)^{-1}\boldsymbol{M}_\pi^T\right)\boldsymbol{y}_j & (14) \end{aligned}$$

Following Scott and Berger (2010), we control multiplicity via the prior

$$p(\pi) \propto \binom{p}{|\pi|}^{-1}. \tag{15}$$

Proposition 22 (Consistency) *Let $g = n$. Then the Bayesian score function $S(G)$ defined above is partially locally consistent, and hence the corresponding estimator $\hat{G}_S$ is consistent.*

Proof This result is an immediate consequence of Lemma 21 and the well-known variable selection consistency property for the unit-information g-prior (see e.g. Fernández *et al.*, 2001). ■

2.6 Computation via Integer Linear Programming

Structure learning for DAGs is a well-studied problem, with contributions including Friedman and Koller (2003); Silander and Myllymäkki (2006); Tsamardinos *et al.* (2006); Cowell (2009); Cussens (2011); Yuan and Malone (2013). Discrete optimisation via ILP can be used to allow efficient estimation for graphical models, exploiting the availability of powerful (and exact) ILP solvers (Nemhauser and Wolsey, 1988; Wolsey, 1998; Achterberg, 2009), as discussed in Bartlett and Cussens (2013). Below we extend the ILP approach to CDAG models.

We begin by computing and caching the quantities

$$p((y_i^l)^{l=1,\dots,n} | (y_k^l)_{k\in\pi}^{l=1,\dots,n}, (x_i^l)^{l=1,\dots,n}) \tag{16}$$

that summarise evidence in the data for the local model $\mathrm{pa}_G(i) = \pi \subseteq V \setminus \{i\}$ for variable i. These cached quantities are transformed to obtain "local scores"

$$s(i,\pi) := \log(p((y_i^l)^{l=1,\dots,n} | (y_k^l)_{k\in\pi}^{l=1,\dots,n}, (x_i^l)^{l=1,\dots,n})) + \log(p(\pi)). \tag{17}$$

These are the (log-) evidence from Eqn. 16 with an additional penalty term that provides multiplicity correction over varying $\pi \subseteq V \setminus \{i\}$, arising from Eqn. 5. Then we define binary indicator variables that form the basis of our ILP as follows:

$$\Pi(i,\pi) := [\mathrm{pa}_G(i) = \pi] \quad \forall i = 1,\dots,p,\ \pi \subseteq V \setminus \{i\}. \tag{18}$$

The information in the variables Π contains all information on the DAG G. However it will be necessary to impose constraints that ensure the Π correspond to a well-defined DAG:

$$\sum_{\pi \subseteq V\setminus\{i\}} \Pi(i,\pi) = 1 \quad \forall i = 1,\dots,p \tag{C1; convexity}$$

Constraint (C1) requires that every node i has exactly one parent set (i.e. there is a well-defined graph G). To ensure G is acyclic we require further constraints:

$$\sum_{i\in C} \sum_{\substack{\pi\subseteq V\setminus\{i\} \\ \pi\cap C=\emptyset}} \Pi(i,\pi) \geq 1 \quad \forall\, C \subseteq V, C \neq \emptyset. \tag{C2; acyclicity}$$

(C2) states that for every non-empty set C there must be at least one node in C that does not have a parent in C.

Proposition 23 *The MAP estimate $\hat{G}_S$ is characterised as the solution of the ILP*

$$\hat{G}_S = \arg\max_{G \in \mathcal{G}} \sum_{i=1}^{p} \sum_{\pi \subseteq V \setminus \{i\}} s(i,\pi)\Pi(i,\pi) \tag{19}$$

subject to constraints (C1) and (C2).

Proof It was proven in Jaakola *et al.* (2010) that (C1) and (C2) together exactly characterise the space $\mathcal{G}$ of DAGs. ■

For the applications in this paper, all ILP instances were solved using the GOBNILP software that is freely available to download from `http://www.cs.york.ac.uk/aig/sw/gobnilp/`.

3. Results

Below we present results based on simulated data and data from molecular biology.

3.1 Simulated Data

We simulated data from linear-Gaussian structural equation models (SEMs). Here we summarise the simulation procedure, with full details provided in the supplement. We first sampled a DAG G for the primary variables and a second DAG G' for the secondary variables (independently of G), following a sampling procedure described in the supplement. That is, G is the causal structure of interest, while G' governs dependence between the secondary variables. Data for the secondary variables $(X_i)_{i \in W}$ were generated from an SEM with structure G'. The strength of dependence between secondary variables was controlled by a parameter $\theta \in [0,1]$. Here $\theta = 0$ renders the secondary variables independent and $\theta = 1$ corresponds to a deterministic relationship between secondary variables, with intermediate values of θ giving different degrees of covariation among the secondary variables. Finally, conditional on the $(X_i)_{i \in W}$, we simulated data for the primary variables $(Y_i)_{i \in V}$ from an SEM with structure G. To manage computational burden, for all estimators we considered only models of size $|\pi| \leq 5$. Performance was quantified by the structural Hamming distance (SHD) between the estimated DAG $\hat{G}_S$ and true, data-generating DAG G, taking into account directionality; we report the mean SHD as computed over 10 independent realisations of the data. We emphasise that the secondary variables need not be generated from a DAG model, however this is a convenient and familiar approach. We note that in the biological application below a DAG for the secondary variables might not be appropriate.

In Table 1 we compare the proposed score-based CDAG estimator with the corresponding score-based DAG estimator that uses only the primary variable data $(Y_i^l)_{i=1,\dots,p}^{l=1,\dots,n}$. We also considered an alternative (DAG2) where a standard DAG estimator is applied to *all* of the variables $(X_i)_{i \in W}$, $(Y_i)_{i \in V}$, with the subgraph induced on the primary variables giving the estimate for G. We considered values $\theta = 0$, 0.5, 0.99 corresponding to zero, mild

$\theta = 0$	$n = 10$			$n = 100$			$n = 1000$		
	DAG	DAG2	CDAG	DAG	DAG2	CDAG	DAG	DAG2	CDAG
$p = 5$	2.9 ± 0.48	2.6 ± 0.48	3.4 ± 0.56	3.2 ± 0.81	1.9 ± 0.43	0.8 ± 0.25	3 ± 0.76	1.6 ± 0.48	0.3 ± 0.21
$p = 10$	9.8 ± 0.55	9.2 ± 0.66	8.8 ± 0.81	8.2 ± 0.99	5 ± 0.84	2.8 ± 0.51	5.5 ± 1.2	5.4 ± 1.1	0.3 ± 0.15
$p = 15$	15 ± 1.3	14 ± 1.1	15 ± 1.2	11 ± 1	6.8 ± 0.83	4.4 ± 0.58	6.3 ± 1.5	8.2 ± 0.92	0.8 ± 0.25
$\theta = 0.5$	$n = 10$			$n = 100$			$n = 1000$		
	DAG	DAG2	CDAG	DAG	DAG2	CDAG	DAG	DAG2	CDAG
$p = 5$	5.7 ± 0.56	4.5 ± 0.48	4 ± 0.49	3.9 ± 0.75	2.1 ± 0.62	0.6 ± 0.31	3.8 ± 0.74	1.8 ± 0.81	0 ± 0
$p = 10$	9.8 ± 0.96	8.1 ± 0.95	7.8 ± 1.1	7.2 ± 1.7	4.6 ± 1.2	1.7 ± 0.84	7.9 ± 1.3	3 ± 0.52	0.6 ± 0.31
$p = 15$	14 ± 0.85	13 ± 0.86	13 ± 0.7	14 ± 1.1	6.2 ± 0.55	3.6 ± 0.76	11 ± 0.93	7.5 ± 1.7	1 ± 0.45
$\theta = 0.99$	$n = 10$			$n = 100$			$n = 1000$		
	DAG	DAG2	CDAG	DAG	DAG2	CDAG	DAG	DAG2	CDAG
$p = 5$	5.1 ± 0.72	4 ± 0.56	4.2 ± 0.49	4.7 ± 0.79	2.1 ± 0.5	0.5 ± 0.31	2.7 ± 0.42	1.7 ± 0.56	0.9 ± 0.48
$p = 10$	9.1 ± 0.84	8.1 ± 0.72	7.8 ± 1.3	9.7 ± 1.1	3.7 ± 0.52	2.5 ± 0.45	9.4 ± 1	4.4 ± 0.64	0.3 ± 0.3
$p = 15$	15 ± 1.1	13 ± 0.94	13 ± 1.2	13 ± 1.4	6.6 ± 0.91	4.7 ± 0.86	17 ± 1.8	9.1 ± 0.69	0.7 ± 0.4

Table 1: Simulated data results. Here we display the mean structural Hamming distance from the estimated to the true graph structure, computed over 10 independent realisations, along with corresponding standard errors. [Data were generated using linear-Gaussian structural equations. $\theta \in [0, 1]$ quantifies dependence between the secondary variables $(X_i)_{i \in W}$, n is the number of data points and p is the number of primary variables $(Y_i)_{i \in V}$. "DAG" = estimation based only on primary variables $(Y_i)_{i \in V}$, "DAG2" = estimation based on the full data $(X_i)_{i \in W} \cup (Y_i)_{i \in V}$, "CDAG" = estimation based on the full data and enforcing CDAG structure.]

and strong covariation among the secondary variables. In each data-generating regime we found that CDAGs were either competitive with, or (typically) more effective than, the DAG and DAG2 estimators. In the $p = 15$, $n = 1000$ regime (that is closest to the biological application that we present below) the CDAG estimator dramatically outperforms these two alternatives. Inference using DAG2 and CDAG (that both use primary as well as secondary variables) is robust to large values of θ, whereas in the $p = 15$, $n = 1000$ regime, the performance of the DAG estimator based on only the primary variables deteriorates for large θ. This agrees with intuition since association between the secondary X_i's may induce correlations between the primary Y_i's. CDAGs outperformed DAG2 at large n (see also Table 1).

To better understand the limitations of CDAGs we considered three data-generating regimes that violate the CDAG assumptions. Firstly we focused on the $\theta = 0$, $p = 15$, $n = 1000$ regime where the CDAG estimator performs well when data are generated "from the model". We then introduced a number E of edges of the form $X_i \to Y_j$ where $Y_i \to Y_j \in G$. These edges (strongly) violate the structural assumptions implied by the CDAG model because their presence means that X_i is no longer a suitable instrument for $Y_i \to Y_j$ as it is no longer conditionally independent of the variable Y_j given Y_i. We assessed performance of the CDAG, DAG and DAG2 estimators as the number E of such misspecified edges is increased (Fig. 3). We find that whilst CDAG continues to perform well up to a moderate fraction of misspecified edges, for larger fractions performance degrades and eventually coincides with DAG and DAG2. Secondly, we fixed $\theta = 0$, $p = 15$, $n = 1000$ and reduced the dependence of the Y_i on the X_i from 100% (relative to values used in the above simulations)

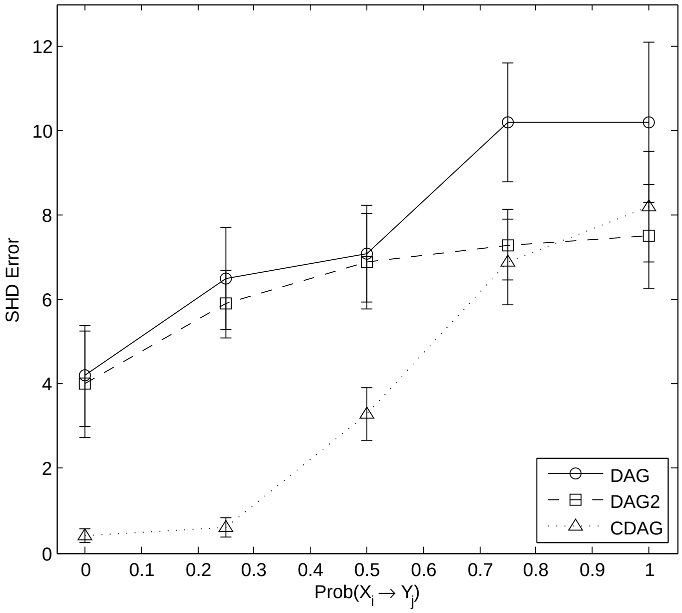


Figure 3: Simulated data results; model misspecification. [Data were generated using linear-Gaussian structural equations. Here we fixed $\theta = 0$, $p = 15$, $n = 1000$ and considered varying the number E of misspecified edges as described in the main text. On the x-axis we display the marginal probability that any given edge $Y_i \to Y_j$ has an associated misspecified edge $X_i \to Y_j$, so that when $\text{Prob}(X_i \to Y_j) = 1$ the number E of misspecified edges is equal to the number of edges in G. "DAG" = estimation based only on primary variables $(Y_i)_{i \in V}$, "DAG2" = estimation based on the full data $(X_i)_{i \in W} \cup (Y_i)_{i \in V}$, "CDAG" = CDAG estimation based on the full data $(X_i)_{i \in W} \cup (Y_i)_{i \in V}$.]

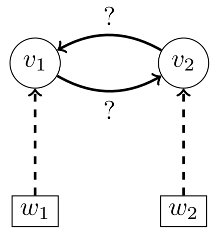
down to 0%. At 0% the partial faithfulness assumption is violated and estimators are no longer consistent. Results (SFig. 1) show that CDAG remains effective over a wide range of data-generating regimes, but eventually degrades when faithfulness is strictly violated. Thirdly, we considered removing the $X_i \to Y_i$ dependence from a random subset of the indices $i \in \{1, \ldots, p\}$, so that faithfulness is violated in a more heterogeneous way across the CDAG, similar to the situation considered by Neto *et al.* (2010). Results (SFig. 2) showed CDAG remained effective when only a handful of deletions occur, but eventually degraded as all the dependencies on secondary variables were removed.

3.2 Molecular Biological Data

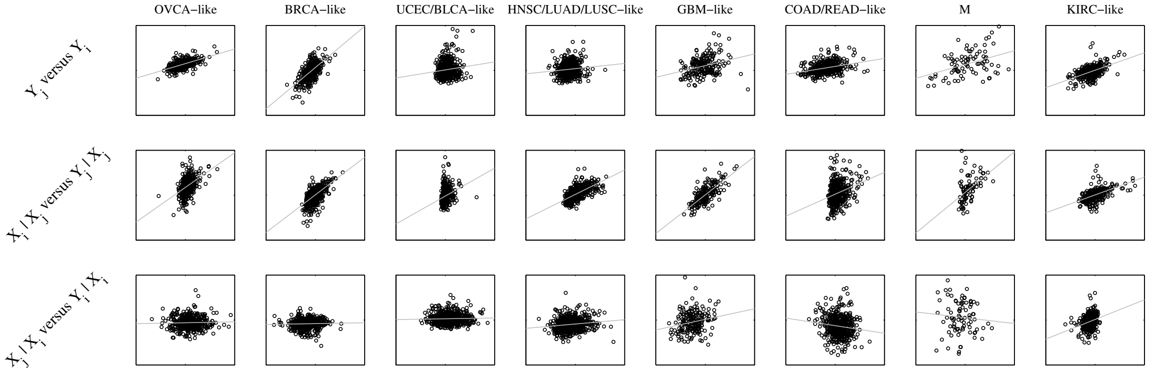
In this section we illustrate the use of CDAGs in an analysis of molecular data from cancer samples. We focus on causal links between post-translationally modified proteins involved in a process called cell signalling. The estimation of signalling networks has been a prominent topic in computational biology for some years (see, among others, Sachs *et al.*, 2005;

protein	node	variable
CHK1	w_i	X_i
p-CHK1	v_i	Y_i
CHK2	w_j	X_j
p-CHK2	v_j	Y_j

(a)



(b)



(c)

Figure 4: CHK1 total protein (t-CHK1) as a natural experiment for phosphorylation of CHK2 (p-CHK2). (a) Description of the variables. (b) A portion of the CDAG relating to these variables. It is desired to estimate whether there is a causal relationship $Y_i \to Y_j$ (possibly mediated by other protein species) or *vice versa*. (c) Top row: Plotting phosphorylated CHK1 (p-CHK1; Y_i) against p-CHK2 (Y_j) we observe weak correlation in some of the cancer subtypes. Middle row: We plot the residuals when t-CHK1 is regressed on total CHK2 (t-CHK2; x-axis) against the residuals when p-CHK2 is regressed on t-CHK2 (y-axis). The plots show a strong (partial) correlation in each subtype that suggests a causal effect in the direction p-CHK1 → p-CHK2. Bottom row: Reproducing the above but with the roles of CHK1 and CHK2 reversed, we see much reduced and in many cases negligible partial correlation, suggesting lack of a causal effect in the reverse direction, i.e. p-CHK1 $\not\leftarrow$ p-CHK2. [The grey line in each panel is a least-squares linear regression.]

Nelander *et al.*, 2008; Hill *et al.*, 2012; Oates and Mukherjee, 2012). Aberrations to causal signalling networks are central to cancer biology (Weinberg, 2013).

In this application, the primary variables $(Y_i)_{i \in V}$ represent abundance of phosphorylated protein (p-protein) while the secondary variables $(X_i)_{i \in W}$ represent abundance of corresponding total proteins (t-protein). A t-protein can be modified by a process called phosphorylation to form the corresponding p-protein and the p-proteins play a key role in signalling. An edge $v_i \to v_j$ has the biochemical interpretation that the phosphorylated form of protein i acts as a causal influence on phosphorylation of protein j. The DAG2

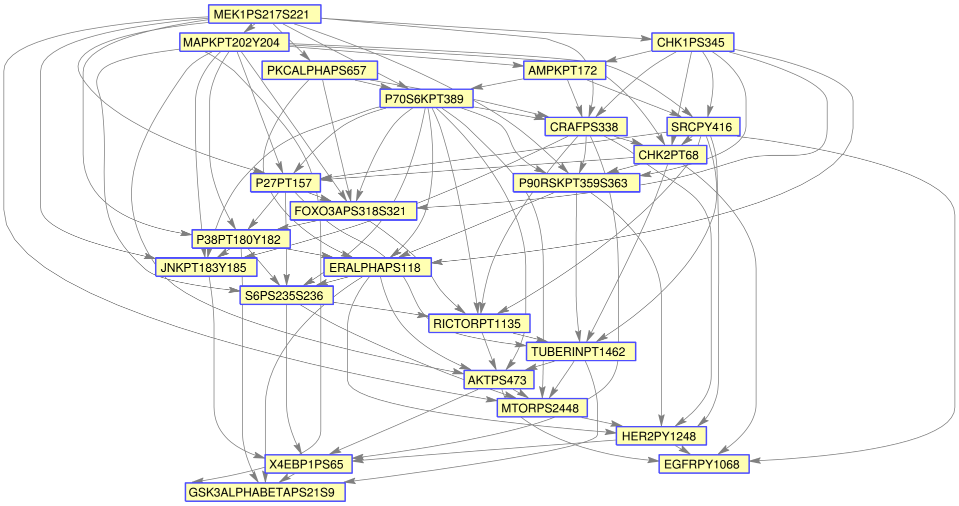


Figure 5: *Maximum a posteriori* conditional DAG, estimated from protein data from cancer samples (from the Cancer Genome Atlas, samples belonging to the BRCA-like group, see text). Here vertices represent phosphorylated proteins (primary variables) and edges have the biochemical interpretation that the parent protein plays a causal role in phosphorylation of the child protein.

method described in Sec. 3.1 may not be suitable for use here since the t-proteins are likely confounded by unobserved variables.

The data we analyse are from the TCGA "pan-cancer" project (Akbani *et al.*, 2014) and comprise measurements of protein levels (including both t- and p-proteins) using a technology called reverse phase protein arrays (RPPAs). We focus on $p = 24$ proteins for which (total, phosphorylated) pairs are available; the data span eight different cancer types (as defined by a clustering analysis due to Städler *et al.*, 2015) with a total sample size of $n = 3,467$ patients. We first illustrate the key idea of using secondary variables to inform causal inference regarding primary variables with an example from these data:

Example 1 (CHK1 t-protein as a natural experiment for CHK2 phosphorylation) *Consider RVs (Y_i, Y_j) corresponding respectively to p-CHK1 and p-CHK2, the phosphorylated forms of CHK1 and CHK2 proteins. Fig. 4c (top row) shows that these variables are weakly correlated in most of the 8 cancer subtypes. There is a strong partial correlation between t-CHK1 (X_i) and p-CHK2 (Y_j) in each of the subtypes when conditioning on t-CHK2 (X_j) (middle row), but there is essentially no partial correlation between t-CHK2 (X_j) and p-CHK1 (Y_i) in the same subtypes when conditioning on t-CHK1 (bottom row). Thus, under the CDAG assumptions, this suggests that there exists a directed causal path from p-CHK1 to p-CHK2, but not vice versa.*

Example 1 provides an example from the TCGA data where controlling for a secondary variable (here, t-protein abundance) may be important for causal inference concerning primary variables (p-protein abundance).

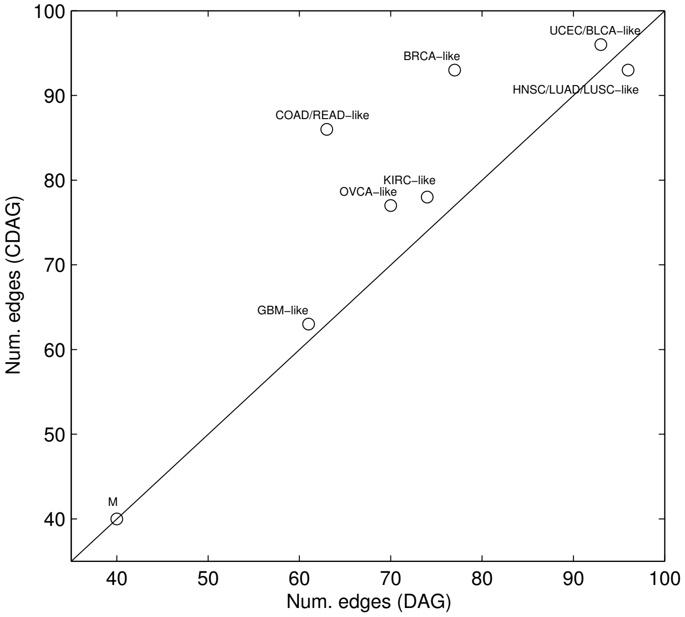


Figure 6: Cancer patient data; relative density of estimated protein networks. Here we plot the number of edges in the networks inferred by estimators based on DAGs (x-axis) and based on CDAGs (y-axis). [Each point corresponds to a cancer subtype (see text). "DAG" = estimation based only on primary variables $(Y_i)_{i\in V}$, "CDAG" = estimation based on the full data and enforcing CDAG structure.]

We now apply the CDAG methodology to all $p = 24$ primary variables that we consider, using data from the largest subtype in the study (namely BRCA-like). The estimated graph is shown in Fig. 5. We note that assessment of the biological validity of this causal graph is a nontrivial matter, and outside the scope of the present paper. However, we observe that several well known edges, such as from p-MEK to p-MAPK, appear in the estimated graph and are oriented in the expected direction. Interestingly, in several of these cases, the edge orientation is different when a standard DAG estimator is applied to the same data, showing that the CDAG formulation can reverse edge orientation with respect to a classical DAG (see supplement). We note also that the CDAG is denser, with more edges, than the DAG (Fig. 6), demonstrating that in many cases, accounting for secondary variables can render candidate edges more salient. These differences support our theoretical results insofar as they demonstrate that in practice CDAG estimation can give quite different results from a DAG analysis of the same primary variables but we note that proper assessment of estimated causal structure in this setting is beyond the scope of this paper.

4. Conclusions

Practitioners of causal inference understand that it is important to distinguish between variables of direct interest and others that can play a supporting role in analysis. In this work we put forward CDAGs as a simple class of graphical models that make this distinction explicit. Motivated by molecular biological applications, we developed CDAGs that use bijections between primary and secondary index sets. The general approach presented here could be extended to other multivariate settings where variables are in some sense non-exchangeable. Naturally many of the philosophical considerations and practical limitations and caveats of classical DAGs remain relevant for CDAGs and we refer the reader to Dawid (2010) for an illuminating discussion of these issues.

The application to biological data presented above represents a principled approach to integrate different molecular data types (here, total and phosphorylated protein, but the ideas are general) for causal inference. Our results suggest that integration in a causal framework may be useful in some settings. Theoretical and empirical results showed that CDAGs can improve estimation of causal structure relative to classical DAGs when the CDAG assumptions are even approximately satisfied.

We briefly mention three natural extensions of the present work: (i) The CDAGs put forward here allow exactly one secondary variable X_i for each primary variable Y_i. In many settings this may be overly restrictive. In biology there are many examples of known causal relationships that are one-to-many or many-to-one. It would be natural to extend CDAGs in this direction. Examples of a more general formulation along these lines were recently discussed by Neto *et al.* (2010) in the context of eQTL data. Conversely we could extend the recent ideas of Kang *et al.* (2014) by allowing for multiple secondary variables for each primary variable, not all of which may be valid as instruments. (ii) In many applications data may be available from multiple related but causally non-identical groups, for example disease types. It could then be useful to consider *joint* estimation of multiple CDAGs, following recent work on estimation for multiple DAGs (Oates *et al.*, 2015, 2014). (iii) Biotechnological advances now mean that the number p of variables is frequently very large. Estimation for high-dimensional CDAGs may be possible using recent results for high-dimensional DAGs (e.g. Kalisch and Bühlmann, 2007; Loh and Bühlmann, 2014, and others).

Acknowledgments

This work was inspired by discussions with Jonas Peters at the Workshop *Statistics for Complex Networks*, Eindhoven 2013, and Vanessa Didelez at the *UK Causal Inference Meeting*, Cambridge 2014. CJO was supported by the Centre for Research in Statistical Methodology (CRiSM) EPSRC EP/D002060/1 and the ARC Centre of Excellence for Mathematical and Statistical Frontiers. Part of this work was done while SM was at the Medical Research Council Biostatistics Unit and University of Cambridge, Cambridge, UK. SM acknowledges the support of the Medical Research Council and a Wolfson Research Merit Award from the UK Royal Society.

References

Achterberg, T. (2009). SCIP: solving constraint integer programs. *Mathematical Programming Computation* **1**(1), 1-41.

Akbani, R. *et al.* (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature Communications* 5:3887.

Bartlett, M., and Cussens, J. (2013). Advances in Bayesian network learning using integer programming. *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 182-191.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics* **40**(3), 1550-1577.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100**(1), 139-156.

Cai, X., Bazerque, J. A., and Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology* **9**(5), e1003068.

Chickering, D.M. (2003). Optimal structure identification with greedy search. *The Journal of Machine Learning Research* **3**, 507-554.

Chun, H., Chen, M., Li, B., and Zhao, H. (2013). Joint conditional Gaussian graphical models with multiple sources of genomic data. *Frontiers in Genetics* **4**, 294.

Consonni, G., and La Rocca, L. (2010). Moment priors for Bayesian model choice with applications to directed acyclic graphs. *Bayesian Statistics* **9**(9), 119-144.

Cowell, R.G. (2009). Efficient maximum likelihood pedigree reconstruction. *Theoretical Population Biology* **76**, 285-291.

Cussens, J. (2011). Bayesian network learning with cutting planes. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 153-160.

Dawid, A.P. (2001). Separoids: A mathematical framework for conditional independence and irrelevance. *The Annals of Mathematics and Artificial Intelligence* **32**(1-4), 335-372.

Dawid, A.P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review* **70**(2), 161-189.

Dawid, A.P. (2010). Beware of the DAG! *Journal of Machine Learning Research-Proceedings Track* **6**, 59-86.

Deltell, A.F. (2011). Objective Bayes criteria for variable selection. *Doctoral dissertation, Universitat de València.*

Didelez, V., and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* **16**, 309-330.

Evans, R. J., and Richardson, T. S. (2014). Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics* **42**(4), 1452-1482.

Evans, R. J. (2015). Margins of discrete Bayesian networks. arXiv:1501.02103.

Fernández, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**(2), 381-427.

Friedman, N., and Koller, D. (2003). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* **50**(1-2):95-126.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* **29**(4), 722-729.

Hill, S. M., Lu, Y., Molina, J., Heiser, L. M., Spellman, P. T., Speed, T. P., Gray, J. W., Mills, G. B., and Mukherjee, S. (2012). Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics* **28**(21), 2804-2810.

Jaakkola, T., Sontag, D., Globerson, A., and Meila, M. (2010). Learning Bayesian network structure using LP relaxations. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 358-365.

Jeffreys, H. (1961). *Theory of probability.* Oxford University Press, 3rd edition.

Johnson V.E., and Rossell D. (2010). Non-local prior densities for default Bayesian hypothesis tests. *Journal of the Royal Statistical Society B* **72**, 143-170.

Kalisch, M., and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* **8**, 613-636.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2014). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. arXiv:1401.5755.

Lauritzen, S.L. (2000). Causal inference from graphical models. In: *Complex Stochastic Systems, Eds. O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg.*CRC Press, London.

Lauritzen, S.L., and Richardson, T.S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B* **64**(3):321-348.

Li, B., Chun, H., and Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association* **107**(497), 152-167.

Logsdon, B. A., and Mezey, J. (2010). Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology* **6**(12), e1001014.

Loh, P. L., and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research* **15**, 3065-3105.

Massam, H., and Wesołowski, J. (2014). A new prior for the discrete DAG models with a restricted set of directions. arXiv:1412.0972.

Nelander, S., Wang, W., Nilsson, B., She, Q. B., Pratilas, C., Rosen, N., Gennemark, P., and Sander, C. (2008). Models from experiments: combinatorial drug perturbations of cancer cells. *Molecular Systems Biology* **4**, 216.

Nemhauser, G.L., and Wolsey, L.A. (1988). *Integer and combinatorial optimization.* Wiley, New York.

Neto, E. C., Keller, M. P., Attie, A. D., and Yandell, B. S. (2010). Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics* **4**(1), 320-339.

Oates, C. J., and Mukherjee, S. (2012). Network inference and biological dynamics. *The Annals of Applied Statistics* **6**(3), 1209-1235.

Oates, C. J., Costa, L., and Nichols, T.E. (2014). Towards a multi-subject analysis of neural connectivity. *Neural Computation* **27**, 1-20.

Oates, C. J., Smith, J. Q., Mukherjee, S., and Cussens, J. (2015). Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, to appear.

Pearl, J., and Paz, A. (1985). Graphoids: A graph-based logic for reasoning about relevance relations. Computer Science Department, University of California.

Pearl, J., and Verma, T. (1987). The logic of representing dependencies by directed graphs. In: *Proceedings of the AAAI*, Seattle WA, 374-379.

Pearl, J. (2003). Reply to Woodward. *Economics and Philosophy* **19**(2), 341-344.

Pearl, J. (2009). *Causality: models, reasoning and inference (2nd ed.).* Cambridge University Press.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**(5721), 523-529.

Scott, J. G., and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38**(5), 2587-2619.

Silander, T., and Myllymäkki, P. (2006). A simple approach to finding the globally optimal Bayesian network structure. *Proceedings of the 22nd Conference on Artificial Intelligence*, 445-452.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search (Second ed.).* MIT Press.

Städler, N., Dondelinger, F., Hill, S. M., Kwok, P., Ng, S., Akbani, R., Werner, H. M. J., Shahmoradgoli, M., Lu, Y., Mills, G. B., and Mukherjee, S. (2015). In preparation.

Studený, M. (2005). Probabilistic conditional independence structures. London: Springer.

Tsamardinos, I., Brown, L.E., and Aliferis, C.F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* **65**(1), 31-78.

Uhler, C., Raskutti, G., Bhlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics* **41**(2), 436-463.

van Wieringen, W. N., and van de Wiel, M. A. (2014). Penalized differential pathway analysis of integrative oncogenomics studies. *Statistical Applications in Genetics and Molecular Biology* **13**(2), 141-158.

Weinberg, R. (2013). *The biology of cancer*. Garland Science.

Wolsey, L.A. (1998). *Integer programming*. Wiley, New York.

Yin, J., and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics* **5**(4), 2630-2650.

Yuan, C., and Malone, B. (2013). Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research* **48**, 23-65.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *In A. Zellner, ed., Bayesian Inference and Decision techniques: Essays in Honour of Bruno de Finetti, Edward Elgar Publishing Limited*, 389-399.

Zhang, L., and Kim, S. (2014). Learning gene networks under SNP perturbations using eQTL datasets. *PLoS Computational Biology* **10**(2), e1003420.