Consequences of near-unfaithfulness in a finite sample – a simulation study

Björn Andersson

Student
VT 2010
One year master thesis, 15 hp
Department of Statistics
Supervisor: Ingeborg Waernbaum

Abstract

Directed acyclic graphs can be used to infer associations between variables and specify their joint distribution. To infer these associations, an assumption called faithfulness must be satisfied. Faithfulness entails that the independencies that exist in the distribution are equivalent to those in the graph. Violations of the faithfulness assumption can occur because of parameter cancellations or because of certain deterministic relationships between variables. In this study, simulations have been conducted comparing inference between settings when near-unfaithfulness due to parameter cancellations is disallowed versus when it is allowed. The results of the simulations suggest that inference under a setting when near-unfaithfulness is allowed is negatively affected only slightly compared to when near-unfaithfulness is not allowed. For the more sparse models with three and four variables, the differences tended to zero with increasing sample sizes. In simulations of the largest model used in this study, with seven variables, the differences did not tend to zero with increasing sample sizes, likely because of the increased complexity of the inference method when using larger models.

Sammanfattning

Riktade acykliska grafer kan användas för att fastställa associationer mellan variabler och specificera deras simultana fördelning. Vid inferens för sådana associationer behövs ett antagande kallat stabilitet, som innebär att de oberoenden som finns i fördelningen är identiska med de oberoenden som finns i grafen. Stabilitet kan vara ouppfyllt vid vissa parametervärden då parametrarna tar ut varandra eller när särskilda deterministiska förhållanden råder mellan variablerna. I denna uppsats har simuleringar genomförts under förutsättningar som tillåter brister i stabilitet på grund av att parametrar tar ut varandra respektive ej tillåter att stabilitet brister på grund av att parametrar tar ut varandra, för att se vilken skillnad, om någon, som finns för inferens under de olika förutsättningarna. Resultaten av simuleringarna indikerar att inferens under ett tillstånd där stabilitet kan vara ouppfyllt enbart påverkas negativt i mycket liten utsträckning. Vid simulering av de mindre modellerna med tre och fyra variabler gick skillnaderna mellan de olika förutsättningarna mot noll när stickprovsstorleken ökade. Detta var inte fallet för modellen med sju variabler, vilket kan förklaras med den ökade komplexiteten i inferensmetoden när antalet variabler stiger.

1

1. Introduction

A multivariate statistical model with a joint distribution can be represented in a graph by evaluating conditional independence statements regarding the variables in the model. Often, a research question involves a background variable and a response variable of interest. If there is strong background knowledge to suggest that these variables can be ordered in a sequence from the background variable to the response variable such as in time or in a causal ordering, a directed graphical model may be appropriate. Connecting the two variables of interest, there may be intermediate variables which are both potential responses to previous variables but also potentially explanatory for coming variables. In a case like this there are no variables that are explanatory in and of themselves. An independence graph that is directed and acyclic can be created to represent the relationships between the variables and draw conclusions about their possible connections.[1] Using graphical models to infer causal links in non-experimental data is of particular interest.

Directed acyclic graphs have been used in different fields to represent connections between variables graphically and determine how they are related. An example is Collinet et al, who constructed directed acyclic graphs to infer characteristics of endosomes in endocytosis, which is the process where molecules outside a cell are engulfed by the cell membrane.[2] Several other areas of application have been proposed, such as within microarray genetic experiments[3] and forensic DNA analysis[4].

In order to learn a directed acyclic graph from observations an assumption called faithfulness is needed. Faithfulness entails that the independencies that are embedded in the distribution are equivalent to the independencies in the graphical model specified, which implies that an altered parametrization of the distribution should not affect the independence structure of the graph. If

---

[1] Wermuth, N. (2003). Analysing social science data with graphical Markov models, In *Highly Structured Stochastic Systems,* 47-52, Oxford University Press

[2] Collinet et al, (2010). Systems survey of endocytosis by multiparametric image analysis, *Nature* **464** 243-249

[3] Spirtes, Glymour et al, (2001) Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data *Proceedings of the Atlantic Symposium on Computational Biology* Genome Information Systems & Technology, 2000

[4] Mortera, Dawid, Lauritzen, (2002). Probabilistic expert systems for DNA mixture profiling *Theoretical Population Biology* **63** *191–205*

faithfulness is not satisfied, then recovering the correct underlying relationships in the graph from the distribution is not possible.

The object of this paper is to evaluate the faithfulness assumption with a simulation study, to see how inference is affected when faithfulness is guaranteed versus when it is not guaranteed. The reasons for assuming faithfulness will be discussed and also whether or not faithfulness is tenable as an assumption when working with applications.

This paper is organized as follows. Section 2 introduces graphical models and involves the theoretical underpinnings of graphical models such as the concept of conditional independence and the Markov properties of directed graphs. Faithfulness is defined and the discussion regarding the assumption is also presented. Section 3 describes the simulation study and the results are presented. In section 4 the results are discussed.

2. Theoretical background

2.1 Conditional independence

The concept of conditional independence is the basis for most of the theory behind directed acyclic graphs. If X, Y, Z are random variables with joint probability distribution P, then X is said to be conditionally independent of Y given Z if P(X|Y,Z)=P(X|Z). That is, knowing Y does not affect what we can say about X given that we already know Z. Conditional independence between X and Y given Z is written X⊥Y|Z. Conditional independence satisfies a number of properties, some of which are written below together with an informal interpretation of what they mean:

$X {\perp\!\!\!\perp} Y|Z \rightarrow Y {\perp\!\!\!\perp} X|Z$, meaning that if X is irrelevant for knowing Y when you know Z, then Y is also irrelevant for knowing X given that you know Z.

$X {\perp\!\!\!\perp} Y|Z \wedge U = h(X) \rightarrow U {\perp\!\!\!\perp} Y|Z$, meaning that if you know Z and that X then is irrelevant for knowing Y, and U is a function of X, then, knowing Z, U is also irrelevant for knowing Y.

$X {\perp\!\!\!\perp} Y|Z \wedge U = h(X) \rightarrow X {\perp\!\!\!\perp} Y|(Z, U)$, meaning that if, knowing Z, X is irrelevant for knowing Y and U is a function of X, then X is irrelevant for knowing Y given that you know Z and U.

$X {\perp\!\!\!\perp} Y|Z \wedge X {\perp\!\!\!\perp} W|(Y, Z) \rightarrow X {\perp\!\!\!\perp} (W, Y)|Z$, meaning that if, knowing Z, X is irrelevant for knowing Y and, knowing Y and Z, X is irrelevant for knowing W, then X is irrelevant for knowing both W and Y given that you know Z.[5]

If f denotes the probability density of the random variables given, the following statements hold:

$X {\perp\!\!\!\perp} Y|Z \iff f(x,y,z) = f(x,z)f(y,z)/f(z)$

$X {\perp\!\!\!\perp} Y|Z \iff f(x|y,z) = f(x|z)$

$X {\perp\!\!\!\perp} Y|Z \iff f(x,y|z) = f(x|z)f(y|z)$

$X {\perp\!\!\!\perp} Y|Z \iff f(x,y,z) = h(x|z)k(y,z)$, *for some functions h and k,*

$X {\perp\!\!\!\perp} Y|Z \iff f(x,y,z) = f(x|z)f(y,z)$.[6,7]

If the joint density is strictly positive then $X {\perp\!\!\!\perp} Y|Z \wedge X {\perp\!\!\!\perp} Z|Y \rightarrow X {\perp\!\!\!\perp} (Y, Z)$ also holds.

[5] Dawid, (1979). Conditional Independence in Statistical Theory, *Journal of the Royal Statistical Society. Series B (Methodological)*, **41(1)**, 1-31

[6] Ibid.

[7] Lauritzen, S.L., (2000) Causal Inference from Graphical Models, In *Complex Stochastic Systems*, 63-107, Chapman and Hall/CRC Press

To exemplify the concept of conditional independence, consider the rolling of a fair six-sided dice. The sample space for the rolling of a dice once is S={1,2,3,4,5,6}. Let A be the event {1,2,3,4}, that is the dice shows either a 1, a 2, a 3 or a 4, and B be the event {2,4,6}. A and B are independent events since

P(A)*P(B)=(4/6)*(3/6)=2/6=P(A,B).

Now, let C be the event {3,6}. P(A|C), the probability of event A given that we have observed event C, can be written

$$\frac{P(A,C)}{P(C)} = \frac{P(\{3\})}{P(\{3,6\})} = \frac{1/6}{2/6} = \frac{1}{2} = \frac{P(\{6\})}{P(C)} = \frac{P(B,C)}{P(C)} = P(B|C).$$

From this it follows that

$$P(A,B|C) = \frac{P(\emptyset)}{P(C)} = 0 \neq P(A|C) * P(B|C),$$

that is while events A and B are independent, they become marginally dependent given event C and the reason is that their intersections with C are disjoint. These types of conditional relations between variables are of relevance when learning a graph from data.

2.2 Graphical models and directed acyclic graphs

A graph is herein defined as a set **G**={**V**, **E**}, where **V** is a set of vertices and **E** is set of edges, a subset of the ordered pairs of **V**, indicating the pairwise relationships between vertices. The existence of an edge between two vertices represents dependence between those vertices. A graph can thus be said to contain a number of dependence and independence statements about a set of vertices. Since **E** is a set it contains no multiple edges. Also, it is required to only contain pairs of distinct vertices, disallowing loops.[8]

In case the pairs (α, β) and (β, α) are both in **E**, then there exists an undirected edge between α and β, written α~β. If two vertices are connected by an undirected edge, those vertices are called neighbours and the set of all neighbours to a vertex β is written ne(β). If the set **E** contains the pair (α, β) but does not contain the pair (β, α), the edge between α and β is said to be directed and is written α → β. Furthermore, α is then called a parent of β and β is called a child of α. The set of parents to α is written pa(α) and the set of children to α is written ch(α). α and β are joined if (α,

---

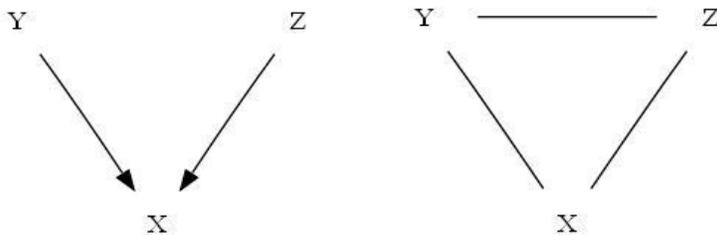[8] Lauritzen, S.L., (2000) Causal Inference from Graphical Models, In *Complex Stochastic Systems*, 63-107, Chapman and Hall/CRC Press

β) ∨ (β, α) ∈ **E**. The statement α↛β then says that α and β are not joined. If **A** ⊂ **V** then pa(**A**), ne(**A**) and ch(**A**) are the collections of parents, neighbours and children of the elements in **A**, excluding those elements that are in **A**.[9]

The primary interest in this paper is directed graphs, and they are defined as graphs in which all edges are directed. **G$_A$**= {**A**, **E$_A$**} is called a subgraph of **G**={**V**, **E**} if **A** ⊆**V** and **E$_A$** ⊆ **E** ∩(**A** × **A**), that is **E$_A$** is a subset of the intersection of **E** and the ordered pairs of **A**. A complete graph is a graph where all pairs of vertices are joined and a subset of vertices from **G** is said to be complete if they form a complete subgraph.[10]

A path from α to β with length n is defined as the sequence of consecutive vertices α=α$_0$, α$_1$ … α$_n$=β, where (α$_{i-1}$, α$_i$) ∈ **E**, ∀i=1,2, … , n. A cycle is a path where α=α$_0$, α$_1$, …, α$_n$=α. If a directed graph contains no cycles it is called a directed acyclic graph (DAG). A path from α to β is written α ↦ β. In a directed acyclic graph **G**={**V**,**E**}, the set of vertices fulfilling α ↦ β but not β ↦ α are called ancestors of β and the set of vertices fulfilling β ↦ α but not α ↦ β are called descendents of β. They are written an(β) and de(β), respectively. An ancestral set of β is then the set of ancestors to β. The moral graph **G^m** of a directed acyclic graph **G** is defined as the graph created when undirected edges are added between all pairs of vertices which have common children, given that they are not already joined, and then forming the undirected graph of the created graph.[11]

Figure 1

A directed acyclic graph (left) and the moral graph corresponding to it (right)

[9] Lauritzen, S.L., (2000) Causal Inference from Graphical Models, In *Complex Stochastic Systems*, 63-107, Chapman and Hall/CRC Press
[10] Ibid.
[11] Ibid.

2.3 The DAG as an independence model

In a directed acyclic graph the dependencies between variables may be viewed as having a causal element or they may be viewed as merely having an association. The latter view is said to be a probabilistic perspective which then means that the interest is to specify the joint distribution of the given variables and not attribute any causal element to the relationships between them. In this paper structural models that are causal are used. However, a causal interpretation of a directed acyclic graph is not a requirement. In a directed graph a missing edge between two vertices implies conditional independence given all the prior variables whereas in an undirected graph a missing edge between two vertices implies conditional independence given all the remaining variables. To exemplify this, consider the model involving the variables age, gender and wages shown in Figure 2. Using a directed graph as opposed to an undirected graph is motivated by the fact that it is not reasonable to specify that wages can precede or influence age or gender, such as is the implication when using an undirected graph.

Figure 2
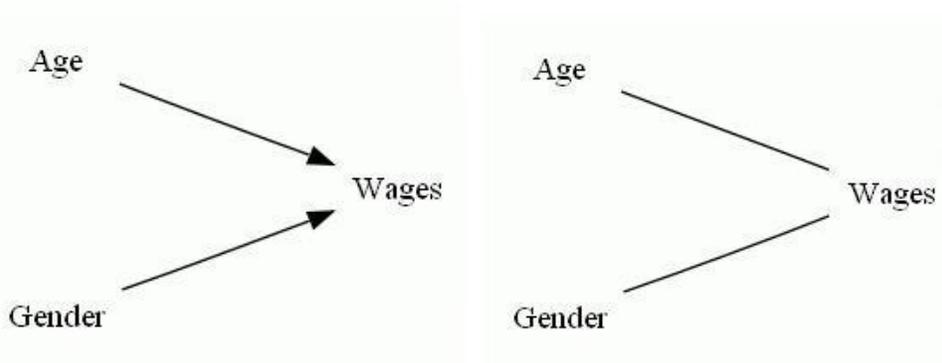
A directed acyclic graph and an undirected graph containing the same variables and adjacencies; the conditional independence statements are however not equivalent

As mentioned, a directed edge between two vertices corresponds to dependence between those two vertices. However, it is not entirely clear what can be deduced about independencies and dependencies in a more complex directed acyclic graph where there are paths stretching over many vertices.

Now, if a distribution P is compatible with G we have that

8

$A \perp\!\!\!\perp B | S_G \implies A \perp\!\!\!\perp B | S_P,$

that is if A and B are independent conditional on S in the graph G, then A and B are independent conditional on S in the distribution P. This entails that P can be factorized in accordance to the graph like

$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i | pa(x_i)),$

where each term on the right hand side is the density of one of the variables conditional on its parents in the graph. The global Markov property states that two subsets of variables are conditionally independent given a third subset, that is $\mathbf{X_A} \perp\!\!\!\perp \mathbf{X_B} | \mathbf{X_S}$ where A, B and S denote some subsets of variables. If P, the joint distribution of the variables, factorizes recursively according to a directed acyclic graph $\mathbf{G}$, then it factorizes according to the moral graph $\mathbf{G^m}$ and thus obeys the global Markov property relative to $\mathbf{G^m}$. Furthermore, if $\mathbf{A}$ is an ancestral set of $\mathbf{G}$ and P factorizes recursively according to $\mathbf{G}$, then the marginal distribution P_A factorizes recursively according to $\mathbf{G_A}$. A consequence of this result is that if P factorizes recursively according to $\mathbf{G}$ then $\mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{S}$, when $\mathbf{A}$ and $\mathbf{B}$ are separated by $\mathbf{S}$ in $\mathbf{G^m_{an(A \cup B \cup S)}}$, the moral graph of the smallest ancestral set that contains $\mathbf{A}$, $\mathbf{B}$ or $\mathbf{S}$. This property is called the directed global Markov property.[12] Identical to this property is the concept of d-separation, introduced by Pearl. By his terminology, a vertex is called a collider on a path whenever there are two directed edges pointing toward it and the vertex is not an endpoint in the graph. In Figure 3 the vertex C is a collider. If a vertex is not an endpoint and not a collider it is called a noncollider.

Figure 3

A DAG illustrating a collider, here C is a collider

A path from X to Y is now said to be d-separated by $\mathbf{Z}$, a set of vertices, if and only if

1) the path contains a sequence a → b → c or a sequence a ← b → c such that b is in $\mathbf{Z}$, or

2) the path contains a collider a → b ← c such that b is not in Z and no descendant of b is in $\mathbf{Z}$.

[12] Lauritzen, S.L., (2000) Causal Inference from Graphical Models, In *Complex Stochastic Systems*, 63-107, Chapman and Hall/CRC Press

To explain further, let a → b → c and a ← b → c be two sequences. a and c are then marginally dependent but become independent when conditioning on b. Figuratively speaking conditioning on b blocks the information along the path, since knowing b renders information about a of no use when the object is to know the probability of c. A collider works in the opposite way, a and c are marginally independent but they will become dependent once we know b or any descendent of b. A path can thus be either blocked or active, with active meaning that the endpoints in the path are dependent and blocked meaning that they are independent.[13] The relationship between d-separation and conditional independence is established in the following theorem.

Theorem 1:

If the sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every distribution compatible with G. Conversely, if X and Y are not d-separated by Z in a DAG G, then X and Y are marginally dependent conditional on Z in at least one distribution compatible with G.[14]

Identical conditional independence relationships between a set of variables can be represented by different directed acyclic graphs. An example is the statement A⊥⊥B|C which can be represented by the three different directed acyclic graphs shown in Figure 4.[15] A set of DAGs representing exactly the same conditional independence statements is said to constitute a Markov equivalence class.

Figure 4
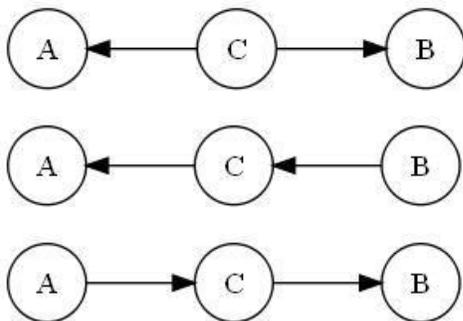
Three different DAGs illustrating the same conditional independence statement, A⊥⊥B|C

---

[13] Edwards, D. (2000) *Introduction to Graphical Modelling* Springer
[14] Pearl, J., (2000) *Causality*, Cambridge University Press
[15] Zhang, J., Spirtes, P. (2008). Detection of Unfaithfulness and Robust Causal Inference, *Minds & Machines* **18**, 239-271
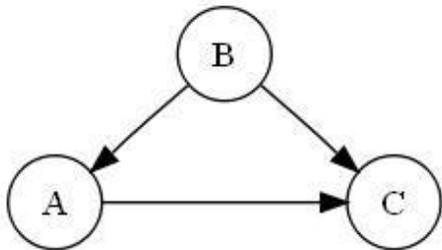
2.4 Faithfulness

The structure of a graphical model is impossible to learn from data without an assumption called faithfulness (also called DAG-isomorphism or stability). Faithfulness is defined as

$A \perp\!\!\!\perp B | S_G \iff A \perp\!\!\!\perp B | S_P,$

meaning that the conditional independence statements that are in a distribution P are equivalent to those in the graph G. A result of this assumption is that the independencies in a model should remain invariant as the parametrization of the distribution changes.[16]

An example where faithfulness is not necessarily fulfilled (taken from Pearl 2000) is as follows. Consider the binary variable A, which takes the value 1 whenever the coins B and C turn up the same and the value 0 in all other cases. In the distribution that is created from this scenario all pairs of variables are marginally independent yet are all also dependent conditional on the third variable.

Figure 5

A DAG of a case where faithfulness can be violated

In this case, only when the faithfulness assumption is fulfilled will the structure retain its independence pattern when the parameters change.[17]

Violations of faithfulness can also occur if the parameters in a DAG cancel each other out exactly. Some independencies are structural and remain in every possible parametrization of a particular DAG. In certain DAGs, this may not be the case. Take for example the DAG Z←X→

---

[16] Lauritzen, S.L., (2000) Causal Inference from Graphical Models, In *Complex Stochastic Systems*, 63-107, Chapman and Hall/CRC Press

[17] Pearl, J., (2000) *Causality*, Cambridge University Press

Y, where Z and Y are independent conditional on X in every possible parametrization. If an arrow is added between Z and Y and a linear relationship between the variables exist, then a certain very precise parametrization will cause X and Y to become independent. This independence is however not "stable", since it disappears when the very precise parametrization is slightly altered. Thus faithfulness is violated.[18] Such very fine-tuned parametrizations have been shown to have zero probability of occurring if parameters are allowed to vary independently over a joint continuous distribution.[19] Objections to this line of reasoning have been made, suggesting that these types of instances are not nearly as rare in the real world as to have zero probability of occurring, with arguments involving the existence of non-causal events in the macroscopic world.[20] Steel suggests that the arguments made against the assumption of faithfulness ultimately involve two issues: selection of parameters and homogeneity of parameters. Selection is defined as a process that concentrates the weight of the distribution of parametrizations to a subset in which faithfulness is violated, and homogeneity is defined as the absence of factors that alter the distribution of the parameters in uncontrolled ways. Strict violations of faithfulness indeed have zero probability of occurring, he says, but near-violations can realistically occur when homogeneity and selection are both satisfied. These near-violations can be just as damning as strict violations, since the probability distributions must be estimated by way of finite sample data and are never known exactly. Despite the possibility of faithfulness being violated in real-world situations, Steel maintains that in complex systems like social groups, biological organisms and ecosystems, the primary fields of applications for directed acyclic graphs according to Steel, such occurrences are very unlikely.[21]

In simulation studies, violations of faithfulness have nevertheless been demonstrated to occur frequently. The reason for this is that accidental correlations occur and dependencies are weakened when faced with limited sample sizes.[22]

---

[18] Pearl, J., (2000) *Causality*, Cambridge University Press
[19] Meek, C., (1995). Strong completeness and faithfulness in Bayesian networks, In P. Besnard & S. Hanks (Eds.), *Uncertainty in artificial intelligence: Proceedings of the eleventh conference (pp. 411–418).* San Francisco: Morgan Kaufmann.
[20] Cartwright, N., (1999). Causal diversity and the Markov condition *Synthese* **121** 3-27
[21] Steel, D., (2006). Homogeneity, selection and the faithfulness condition, *Minds & Machines* **16** 303–317
[22] Lemeire, J., Steenhaut, K., (2009). Constraint-based Causal Structure Learning when Faithfulness Fails, Annual machine learning conference of Belgium and The Netherlands (BeneLearn 2009), Tilburg, The Netherlands, 2009.

Objections have been raised against the faithfulness assumption in that it is not empirically testable. In itself, the faithfulness assumption is untestable because it states that the possible parametrizations of the variables are invariant to the true structure of the DAG. Since the true structure is unknown, faithfulness cannot be tested. Zhang and Spirtes have suggested a decomposition of the faithfulness assumption, and parts of this decomposition are testable. Since faithfulness entails that the probability distribution of the variables and the independence structure of those variables are intertwined, it entails that there exists an independence structure that the probability distribution is faithful to, and this is possible to test for. Assuming that the directed Markov property holds, it is then possible to determine if the population distribution is faithful to any DAG at all. They introduce the adjacency-faithfulness condition, in itself untestable, stated as

*Given a set of variables **V** whose true causal DAG is **G**, if two variables X, Y are adjacent in **G**, then they are dependent conditional on any subset of **V**\\{X,Y}.*

They also introduce the orientation-faithfulness condition

*Given a set of variables **V** whose true causal DAG is **G**, let <X, Y, Z> be a triple in **G** such that Z is adjacent to Y but Z and X are not adjacent to each other.*

1. *If X → Y ← Z, then X and Z are dependent given any subset of **V**\\{X,Z} that contains Y*
2. *Otherwise, X and Z are dependent conditional on any subset of **V**\\{X, Z} that does not contain Y.*

These two conditions are not equivalent to faithfulness, but they are consequences of faithfulness. Zhang and Spirtes show that if the adjacency-faithfulness condition is assumed to hold, then the orientation-faithfulness condition can be tested. [23]

To learn a DAG from data, tests of conditional independencies between variables is performed and utilized in an algorithm. A widely used such algorithm is the PC algorithm (PC stands for the original developers Peter Spirtes and Clark Glymour). It takes as input $\widehat{P}$, a faithful empirical distribution on a set of variables **V**, and outputs a DAG compatible with $\widehat{P}$ as follows:

---

[23] Zhang, J., Spirtes, P. (2008). Detection of Unfaithfulness and Robust Causal Inference, *Minds & Machines* **18**, 239-271

1.) For all vertices (a,b) ∈ **V**, search for a set $\mathbf{S_{ab}}$ such that a ⊥ b|$\mathbf{S_{ab}}$ holds in $\widehat{P}$. That is, a and b should be independent in $\widehat{P}$, given $\mathbf{S_{ab}}$. Then, construct an undirected graph **G** such that vertices a and b are connected with an edge if and only if no set $\mathbf{S_{ab}}$ can be found.

2.) For all (a,b) such that a and b are non-adjacent while having a common neighbour c, check if c belongs to $\mathbf{S_{ab}}$. If it does, continue to the next step. If it does not, then add arrowheads pointing at c.

3.) In the partially directed graph that results from the first two steps, orient as many of the undirected edges as possible subject to two conditions:

   i)   the orientation should not create a new v-structure, that is no new triples of variables x, y, z satisfying x→y and z→y, where x and z are not adjacent, should be created

   ii)   the orientation should not create a directed cycle

The PC algorithm is a derivative of the IC (Inductive Causation) algorithm, adding to it a specific way to search for the sets $\mathbf{S_{ab}}$ in step 1. This search starts with sets $\mathbf{S_{ab}}$ of cardinality 0, then of cardinality 1 and so on, and edges are removed recursively from a complete graph when a separating set is found.[24] Since multiple DAGs can represent the same conditional independence relations, the resulting DAG of the PC algorithm may not be unique. An implementation of the PC algorithm in the statistical software environment R[25], called pcalg, is used in this paper.

---

[24] Spirtes, P., Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs, *Social Science Computer Review*, **9(1)**, 62-72

[25] R Development Core Team (2009). R: A language and environment for statictical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-97-1, URL http://www.R-project.org.

3. Simulation

3.1 Design

Parameter cancellations may occur if faithfulness is not satisfied, causing incorrect inference. In this study, data has been simulated in accordance with a causal linear structural equation model with one set of parameters guaranteeing faithfulness and with another set where faithfulness is not guaranteed. Inference was then compared between the two settings.

Three models have been used in this study. The first model, which can be seen in Figure 6 panel a), contains three variables and is written

$$A = \varepsilon_A$$
$$B = \theta_{AB} * A + \varepsilon_B$$
$$C = \theta_{AC} * A + \theta_{BC} * B + \varepsilon_C \, ,$$

the second model, seen in Figure 6 panel b), contains four variables and is written

$$A = \varepsilon_A$$
$$B = \theta_{AB} * A + \varepsilon_B$$
$$C = \theta_{AC} * A + \varepsilon_C$$
$$D = \theta_{BD} * B + \theta_{CD} * C + \varepsilon_D$$

and the third model, seen in Figure 6 panel c), contains seven variables and is written

$$A = \varepsilon_A$$
$$B = \theta_{AB} * A + \varepsilon_B$$
$$C = \theta_{AC} * A + \varepsilon_C$$
$$D = \theta_{BD} * B + \varepsilon_D$$
$$E = \theta_{BE} * B + \varepsilon_E$$
$$F = \theta_{CF} * C + \theta_{EF} * E + \varepsilon_F$$
$$G = \theta_{DG} * D + \theta_{EG} * E + \varepsilon_G$$

In the first model parameter cancellations occur if

$$\theta_{AC} = -\theta_{AB} * \theta_{BC},$$

as shown in Appendix A. In the second model parameter cancellations occur if

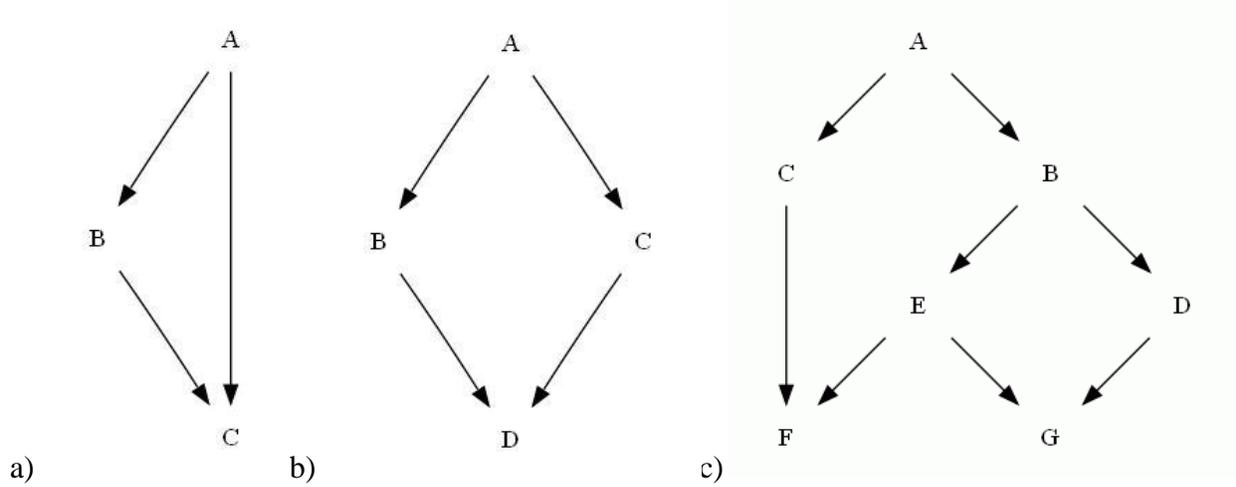$$\theta_{AC} * \theta_{CD} = -\theta_{AB} * \theta_{BD},$$

and in the third model parameter cancellations occur if

15

$$\theta_{AB} * \theta_{BD} * \theta_{CD} = -\theta_{AB} * \theta_{BE} * \theta_{EG}$$

or

$$\theta_{AB} * \theta_{BE} * \theta_{EF} = -\theta_{AC} * \theta_{CF}.$$

Figure 6

a) b) c)

The model with three variables (a), the model with four variables (b) and the model with seven variables (c)

In order to guarantee faithfulness in the simulation, parameter values corresponding to these cancellations must be disallowed. As stated in the previous section, it has been shown that exact cancellations of the sort described above have zero probability of occurring under assumptions that are fulfilled in this simulation study. The object of restricting the parameters thus becomes to impose a limitation such that near-unfaithfulness is disallowed. Near-unfaithfulness is defined as an interval around the parameter values generating unfaithfulness (e.g. an interval around the parameter combination $\theta_{AC} = -\theta_{AB} * \theta_{BC}$ for the model with three variables).

The simulation was performed using the standard normal distribution $\forall \varepsilon_i, i = A, B, C, D, E, F, G$ and with varying sample sizes (100, 500, 1000 and 2000). The parameters $\theta_{i,j}, i, j = A, B, C, D, E, F, G \ i \neq j$ were drawn independently from a uniform distribution with interval [-1.5, 1.5], in accordance with similar studies. In deciding which restrictions to impose, near-unfaithfulness was defined to be the interval corresponding to the 10% of the parameter combinations closest to unfaithfulness. The distributions for the four different parameter

16

combinations corresponding to unfaithfulness are not explicitly formulated, so simulations of the distributions were conducted when deciding the particular restrictions. The density plots of these simulations are found in Appendix B and the loops used in this procedure can be found in Appendix C.

The simulation was done as follows. First the parameters to be used in the structural equation model were drawn from a uniform distribution. These parameters were then checked to see if near-unfaithfulness existed given the model. If so, the parameters were drawn again and this persisted until faithfulness was guaranteed. Data was then generated in accordance with the specified model and used as input to the PC algorithm. The resulting estimated graph was compared with the true graph and three measurements of model fit were recorded. This procedure was repeated 1000 times. For the models without a restriction the same method was used, with the exception that the parameters were not checked for near-unfaithfulness. The functions pertaining to the simulations of the model with seven variables can be found in Appendix D and a description of the loops used is found in Appendix E.

The three measurements of model fit are true positive rate (TPR), defined as the number of correct edges found divided by the total number of true edges in the model, false positive rate (FPR), defined as the number of incorrect edges found divided by the number of unjoined pairs in the true model, and true discovery rate (TDR), defined as the number of correct edges found divided by the total number of edges found. If the true model does not contain any unjoined pairs, FPR is defined as 1. In the model with three variables FPR and TDR are always 1, since the true model does not contain any unjoined pairs and it is not possible to find incorrect edges (if an edge is found it is correct). Higher values of TPR and TDR are thus considered a better result, while lower values of FPR are considered a better result.

3.2 Results

The results are presented in tables below. As can be seen in Table 1, there are only minute differences between the two settings in the model with three variables. TPR is slightly higher under the guarantee of faithfulness compared to near-unfaithfulness but the differences get

17

progressively smaller as the sample size increases, with a average difference in TPR of only 0.0047 percentage points using a sample size of 2000.

Table 1: *Results for the model with 3 variables, average TPR, FPR and TDR from1000 replicates*

Faithfulness	Sample size	True positive rate (TPR)	False positive rate (FPR)	True discovery rate (TDR)
guaranteed	100	0.6860	1	1
not guaranteed	100	0.6720	1	1
guaranteed	500	0.8517	1	1
not guaranteed	500	0.8403	1	1
guaranteed	1000	0.8917	1	1
not guaranteed	1000	0.8840	1	1
guaranteed	2000	0.9197	1	1
not guaranteed	2000	0.9150	1	1

In the model with four variables the same tendency prevails; there are only very small differences which tend to zero as the sample size increases, seen in Table 2. The differences fall within one percentage point using all measurements, and they do not side with any one of the two settings.

Table 2: *Results for the model with 4 variables, average TPR, FPR and TDR from1000 replicates*

Faithfulness	Sample size	True positive rate (TPR)	False positive rate (FPR)	True discovery rate (TDR)
guaranteed	100	0.6955	0.0240	0.9840
not guaranteed	100	0.6903	0.0235	0.9848
guaranteed	500	0.8308	0.0075	0.9961
not guaranteed	500	0.8363	0.0040	0.9982
guaranteed	1000	0.8748	0.0060	0.9973
not guaranteed	1000	0.8775	0.0060	0.9970
guaranteed	2000	0.9095	0.0070	0.9968
not guaranteed	2000	0.9068	0.0075	0.9967

With seven variables similar results are achieved. The simulation guaranteeing faithfulness renders slightly better TPR results for all sample sizes, but the differences are very small (Table 3). The differences do not decrease with increasing sample sizes, however. Under settings that allow for near-unfaithfulness due to parameter cancellations, the PC algorithm is worse by about one edge in every 200 estimates compared to when near-unfaithfulness is disallowed.

Table 3: *Results for the model with 7 variables, average TPR, FPR and TDR from1000 replicates*

Faithfulness	Sample size	True positive rate (TPR)	False positive rate (FPR)	True discovery rate (TDR)
guaranteed	100	0.6779	0.0052	0.9882
not guaranteed	100	0.6724	0.0048	0.9892
guaranteed	500	0.8339	0.0019	0.9962
not guaranteed	500	0.8286	0.0018	0.9967
guaranteed	1000	0.8728	0.0022	0.9960
not guaranteed	1000	0.8683	0.0019	0.9966
guaranteed	2000	0.9048	0.0015	0.9975
not guaranteed	2000	0.8973	0.0020	0.9966

4. Discussion

Other simulation studies have shown that inference can fail when data is generated in accordance with a certain structural model and the graphical model then is estimated with the PC algorithm. This means that the conditional independence statements in the structural model were not identical to those in the estimated graph. In this study, three structural models have been fixed and data has been generated in accordance with the models in two ways such that faithfulness is guaranteed and not guaranteed. The results of this simulation study suggest that when parameters in a structural model are drawn independently from a continuous uniform distribution, limiting the possible parameter values to cases where faithfulness is guaranteed as opposed to not guaranteed has little effect on inference using the PC algorithm. In the more sparse models with three and four variables, the already small differences decreased or disappeared with increasing sample sizes. The structural model with seven variables did however not exhibit this behavior when the sample size increased. This may be caused by the fact that as the number of variables in the model grows, the larger the subsets that must be evaluated are and thus the more variables need be checked at every step in the PC algorithm. Guaranteeing faithfulness thus becomes more important in larger models. The difference is likely to disappear once the sample size becomes greater, this point was simply not reached in this simulation study. It must however be stressed that the differences between the settings are very small and that other error sources such as weak dependencies as a result of limited sample sizes are likely to be more important than near-unfaithfulness is, even with larger models.

Different authors have suggested conditions under which unfaithfulness may be a problem in inference. Among these, there is the condition Steel calls selection – the concentration of the weight of the parameter distribution to situations where near-unfaithfulness exists. Steel proposes that whenever the conditions selection and homogeneity of parameters exist, inference in graphical models may be compromised. The parameter realizations which have been excluded can be seen as excluding selection as defined by Steel, and thus this study indicates that selection alone is unlikely to have a sizable effect on inference should it occur since there were small or no differences in inference under the two different settings.

In this study, only normally distributed variables have been used and as such the results are not necessarily applicable for other distributions. Furthermore, the models used are chosen arbitrarily and may not be indicative of what might occur in more elaborate models. Instead of choosing models arbitrarily, model choice at random could have been utilized. However, doing so would not necessarily create a model where unfaithfulness could occur, making discretion necessary anyway. An automated procedure of this sort is feasible but was considered beyond the scope of this paper.

This study has used a causal structural equation model when generating data. Using graphical models to infer causal relations is a topic of much debate. The main point of contention is the attribution of causation in directed acyclic graphs. Some of the arguments put forward against inferring causal links in graphical models rely on a non-deterministic assumption akin to the quantum mechanical indeterminism which rules over the elementary particles. However, such non-deterministic views are generally never applicable to the macroscopic world, making objections against causal inference on the basis of such reasoning dubious.

Unfaithfulness can occur other than through parameter cancellations. When certain deterministic relationships between variables exist, the underlying relationships may be impossible to infer using graphical models. Graphical models using inference by way of the PC algorithm should not be used under such circumstances since the issue is not that certain improbable parameter instances cause incorrect inference, but the fact that the relationships themselves cause incorrect inference given the estimation procedure.

The faithfulness assumption can be decomposed into adjacency-faithfulness and orientation-faithfulness, which are consequences of but not identical to faithfulness. This decomposition has not been directly studied in this paper. A study which is similar to this paper but that makes sure that orientation-faithfulness is satisfied as opposed to making sure that faithfulness is satisfied could be conducted, to see whether or not a decomposition of the faithfulness assumption and the subsequent testing of orientation-faithfulness yields any beneficial results on inference. Since orientation-faithfulness is testable, less strict assumptions would be required when inferring

graphical models. However, if the satisfaction of orientation-faithfulness yields little benefit the decomposition is unnecessary.

References

Cartwright, N., (1999). Causal diversity and the Markov condition *Synthese* **121** *3-27*

Collinet et al, (2010). Systems survey of endocytosis by multiparametric image analysis, *Nature* **464** 243-249

Dawid, (1979). Conditional Independence in Statistical Theory *Journal of the Royal Statistical Society. Series B (Methodological)* **41(1)** 1-31

Edwards, D. (2000) *Introduction to Graphical Modelling* Springer

Lauritzen, S.L., (2000) Causal Inference from Graphical Models, In *Complex Stochastic Systems*, 63-107, Chapman and Hall/CRC Press

Lemeire, J., Steenhaut, K., (2009). Constraint-based Causal Structure Learning when Faithfulness Fails, Annual machine learning conference of Belgium and The Netherlands (BeneLearn 2009), Tilburg, The Netherlands, 2009.

Meek, C., (1995). Strong completeness and faithfulness in Bayesian networks, In P. Besnard & S. Hanks (Eds.), *Uncertainty in artificial intelligence: Proceedings of the eleventh conference (pp. 411–418).* San Francisco: Morgan Kaufmann.

Mortera, Dawid, Lauritzen, (2002). Probabilistic expert systems for DNA mixture profiling *Theoretical Population Biology* **63** *191–205*

Pearl, J., (2000) *Causality*, Cambridge University Press

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Spirtes, P., Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs, *Social Science Computer Review* **9(1)** 62-72

Spirtes, Glymour et al, (2001) Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data *Proceedings of the Atlantic Symposium on Computational Biology* Genome Information Systems & Technology, 2000

Steel, D., (2006). Homogeneity, selection and the faithfulness condition, *Minds & Machines* **16** *303–317*

Wermuth, N. (2003). Analysing social science data with graphical Markov models, In *Highly Structured Stochastic Systems,* 47-52, Oxford University Press

Zhang, J., Spirtes, P. (2008). Detection of Unfaithfulness and Robust Causal Inference *Minds & Machines* **18** 239-271

Appendices

Appendix A

Consider the structural equation model

$$A = \varepsilon_A$$

$$B = \theta_{AB} * A + \varepsilon_B$$

$$C = \theta_{AC} * A + \theta_{BC} * B + \varepsilon_C.$$

We want to show that $A \perp\!\!\!\perp C | (\theta_{AC} = -\theta_{AB} * \theta_{BC})$.

If $\theta_{AC} = -\theta_{AB} * \theta_{BC}$, we have

$$C = \theta_{AC} * A - (\theta_{AB} * A + \varepsilon_B) * \frac{\theta_{AC}}{\theta_{AB}} + \varepsilon_C$$

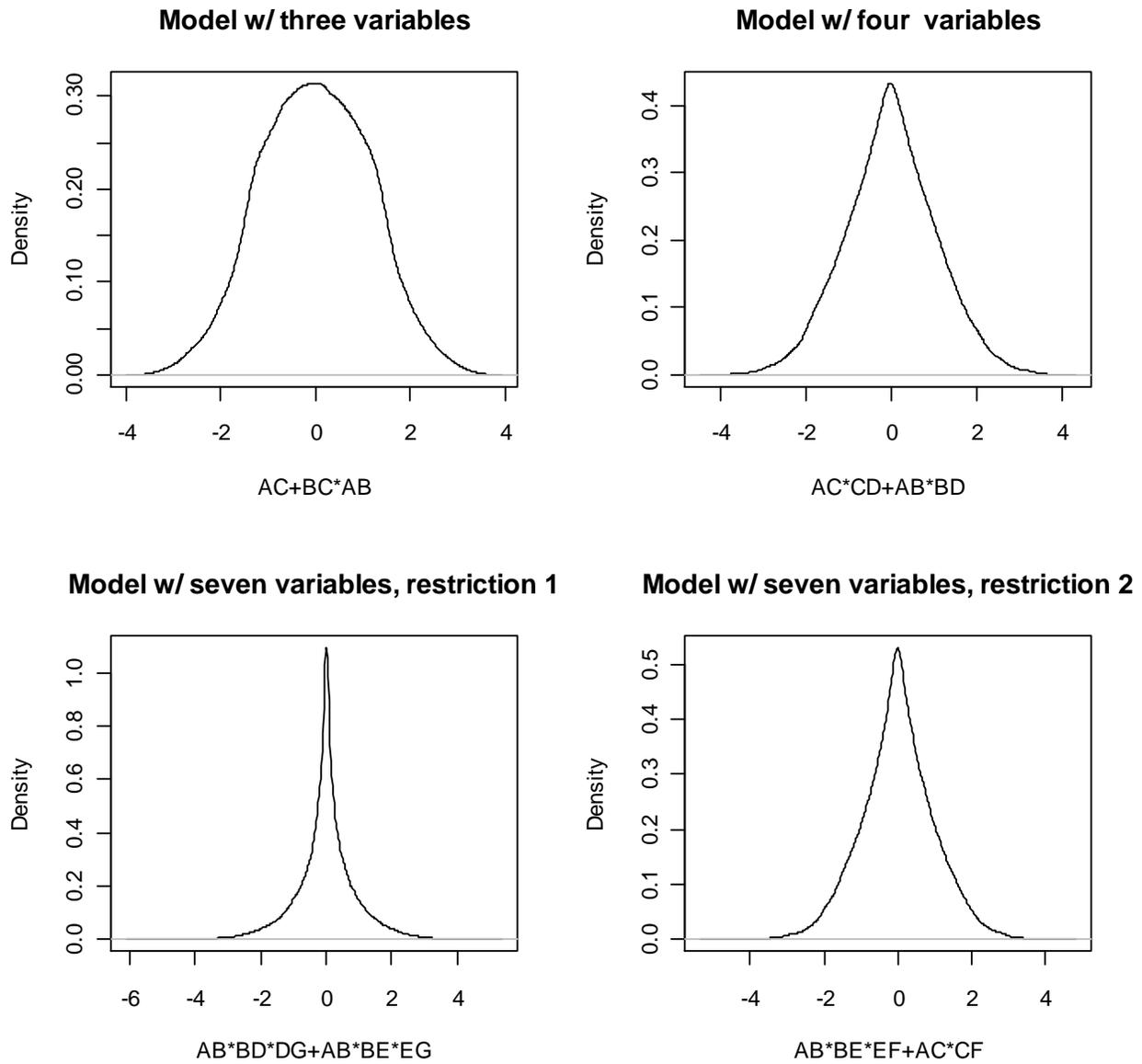$$= \theta_{AC} * A - \theta_{AC} * A - \frac{\theta_{AC}}{\theta_{AB}} * \varepsilon_B + \varepsilon_C$$

$$= -\frac{\theta_{AC}}{\theta_{AB}} * \varepsilon_B + \varepsilon_C.$$

Since $A = \varepsilon_A$ and $\varepsilon_A \perp\!\!\!\perp \varepsilon_B, \varepsilon_C$ we have that $A \perp\!\!\!\perp C | (\theta_{AC} = -\theta_{AB} * \theta_{BC})$.

Equivalent results can be obtained for the other two models used.

Appendix B

Figure 7

Model w/ three variables

Density

AC+BC*AB

Model w/ four variables

Density

AC*CD+AB*BD

Model w/ seven variables, restriction 1

Density

AB*BD*DG+AB*BE*EG

Model w/ seven variables, restriction 2

Density

AB*BE*EF+AC*CF

Approximate density functions for the different parameter combinations considered in the three models used

Appendix C

The simulations of the distributions of the parameter combinations were done in R using the following loops.

```
i=1
t=100000
uniform3=c(1:t)

while(i<=t){
par_AB=runif(1, min=-1.5, max=1.5)
par_AC=runif(1, min=-1.5, max=1.5)
par_BC=runif(1, min=-1.5, max=1.5)
uniform3[i]=(par_AC+par_BC*par_AB)
i=i+1
}
plot(density(uniform3))
sort(abs(uniform3))[10000]#select the value corresponding to the 10th percentile to use as the
restriction

i=1
t=100000
uniform4=c(1:t)

while(i<=t){
par_AB=runif(1, min=-1.5, max=1.5)
par_AC=runif(1, min=-1.5, max=1.5)
par_BD=runif(1, min=-1.5, max=1.5)
par_CD=runif(1, min=-1.5, max=1.5)
uniform4[i]=(par_AC*par_CD+par_AB*par_BD)
i=i+1
}
plot(density(uniform4))
sort(abs(uniform4))[10000]#select the value corresponding to the 10th percentile to use as the
restriction

i=1
t=100000#100000 replicates
uniform7_1=c(1:t)

while(i<=t){
AB=runif(1, min=-1.5, max=1.5)#draw parameters from a uniform distribution
AC=runif(1, min=-1.5, max=1.5)
BD=runif(1, min=-1.5, max=1.5)
BE=runif(1, min=-1.5, max=1.5)
CF=runif(1, min=-1.5, max=1.5)
```

26

```
DG=runif(1, min=-1.5, max=1.5)
EG=runif(1, min=-1.5, max=1.5)
EF=runif(1, min=-1.5, max=1.5)

uniform7_1[i]=(AB*BD*DG+AB*BE*EG)#store the parameter combination of interest
i=i+1
}
plot(density(uniform7_1))#plot the approximate density function
sort(abs(uniform7_1))[5000]#select the value corresponding to the 5th percentile to use as the
restriction

i=1
t=100000
uniform7_2=c(1:t)

while(i<=t){
AB=runif(1, min=-1.5, max=1.5) #draw parameters from a uniform distribution
AC=runif(1, min=-1.5, max=1.5)
BE=runif(1, min=-1.5, max=1.5)
CF=runif(1, min=-1.5, max=1.5)
EF=runif(1, min=-1.5, max=1.5)

uniform7_2[i]=(AB*BE*EF+AC*CF)# store the parameter combination of interest
i=i+1
}

plot(density(uniform7_2)) #plot the approximate density function
sort(abs(uniform7_2))[5000]#select the value corresponding to the 5h percentile to use as the
restriction
```

Appendix D

Six different functions were created in order to conduct the simulation. The functions used to

simulate the model with seven variables, newsim7.R and newsim7f.R, are provided below. The

functions for the other models can be provided upon request.


```
newsim7=function(n=n, alpha=alpha, p=p, true=true)

{
AB=runif(1, min=-1.5, max=1.5)
AC=runif(1, min=-1.5, max=1.5)
BD=runif(1, min=-1.5, max=1.5)
BE=runif(1, min=-1.5, max=1.5)
CF=runif(1, min=-1.5, max=1.5)
DG=runif(1, min=-1.5, max=1.5)
EG=runif(1, min=-1.5, max=1.5)
EF=runif(1, min=-1.5, max=1.5)

A=rnorm(n)
B=AB*A+rnorm(n)
C=AC*A+rnorm(n)
D=BD*B+rnorm(n)
E=BE*B+rnorm(n)
F=CF*C+EF*E+rnorm(n)
G=DG*D+EG*E+rnorm(n)

d.mat=matrix(c(A,B,C,D,E,F,G),n)
indepTest=gaussCItest
suffStat=list(C=cor(d.mat), n=n)
est=skeleton(suffStat, indepTest, p, alpha)
true=true
res=compareGraphs(est@graph, true)
res=as.data.frame(res)
res
}

newsim7f=function(n=n, alpha=alpha, p=p, true=true)
{
AB=runif(1, min=-1.5, max=1.5)
AC=runif(1, min=-1.5, max=1.5)
BD=runif(1, min=-1.5, max=1.5)
BE=runif(1, min=-1.5, max=1.5)
CF=runif(1, min=-1.5, max=1.5)
DG=runif(1, min=-1.5, max=1.5)
EG=runif(1, min=-1.5, max=1.5)
EF=runif(1, min=-1.5, max=1.5)
```

28

```
while((abs(AB*BD*DG+AB*BE*EG)<=0.01611097)|(abs(AB*BE*EF+AC*CF)<=0.04499905)
){
AB=runif(1, min=-1.5, max=1.5)
AC=runif(1, min=-1.5, max=1.5)
BD=runif(1, min=-1.5, max=1.5)
BE=runif(1, min=-1.5, max=1.5)
CF=runif(1, min=-1.5, max=1.5)
DG=runif(1, min=-1.5, max=1.5)
EG=runif(1, min=-1.5, max=1.5)
EF=runif(1, min=-1.5, max=1.5)
}

A=rnorm(n)
B=AB*A+rnorm(n)
C=AC*A+rnorm(n)
D=BD*B+rnorm(n)
E=BE*B+rnorm(n)
F=CF*C+EF*E+rnorm(n)
G=DG*D+EG*E+rnorm(n)

d.mat=matrix(c(A,B,C,D,E,F,G),n)
indepTest=gaussCItest
suffStat=list(C=cor(d.mat), n=n)
est=skeleton(suffStat, indepTest, p, alpha)
true=true
res=compareGraphs(est@graph, true)
res=as.data.frame(res)
res
}
```

Appendix E
The different simulations were performed using loops such as the one below, with specific inputs used, e.g. n=1000 and true=true7, for a sample size of 1000 and the model with seven variables.

```
library(pcalg)#load the required package pcalg
i=1
t=1000#set no of replicates, here 1000
newresults7f_100=matrix(1:(3*t), nrow=3, ncol=t)#create matrix to store results in
while(i<=t){xyz=newsim7f(100, 0.01, 7, true7)#estimate graph under conditions guaranteeing
faithfulness t times
newresults7f_100[,i]=xyz[,1]#store results from the i:th simulation
i=i+1
}
```