Biometrika (2003), **90,** 3, *pp.* 491–515
© 2003 Biometrika Trust
Printed in Great Britain

Uniform consistency in causal inference

By JAMES M. ROBINS

Department of Epidemiology and Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.

robins@hsph.harvard.edu

RICHARD SCHEINES, PETER SPIRTES

Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

scheines@andrew.cmu.edu ps7z@andrew.cmu.edu

and LARRY WASSERMAN

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

larry@stat.cmu.edu

Summary

There is a long tradition of representing causal relationships by directed acyclic graphs (Wright, 1934). Spirtes (1994), Spirtes et al. (1993) and Pearl & Verma (1991) describe procedures for inferring the presence or absence of causal arrows in the graph even if there might be unobserved confounding variables, and/or an unknown time order, and that under weak conditions, for certain combinations of directed acyclic graphs and probability distributions, are asymptotically, in sample size, consistent. These results are surprising since they seem to contradict the standard statistical wisdom that consistent estimators of causal effects do not exist for nonrandomised studies if there are potentially unobserved confounding variables. We resolve the apparent incompatibility of these views by closely examining the asymptotic properties of these causal inference procedures. We show that the asymptotically consistent procedures are 'pointwise consistent', but 'uniformly consistent' tests do not exist. Thus, no finite sample size can ever be guaranteed to approximate the asymptotic results. We also show the nonexistence of valid, consistent confidence intervals for causal effects and the nonexistence of uniformly consistent point estimators. Our results make no assumption about the form of the tests or estimators. In particular, the tests could be classical independence tests, they could be Bayes tests or they could be tests based on scoring methods such as BIC or AIC. The implications of our results for observational studies are controversial and are discussed briefly in the last section of the paper. The results hinge on the following fact: it is possible to find, for each sample size n, distributions P and Q such that P and Q are empirically indistinguishable and yet P and Q correspond to different causal effects.

Some key words: Causation; Confounding; Directed acyclic graph.

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

1. Introduction

The problem of inferring causal relationships between variables is a topic of continuing interest. Common statistical wisdom dictates that causal effects cannot be consistently estimated from observational studies alone unless one observes and adjusts for all possible confounding variables, and knows the time order in which events occurred. However, Spirtes (1994), Spirtes et al. (1993) and Pearl & Verma (1991) developed a framework in which causal relationships are represented by arrows in a directed acyclic graph G. They also described asymptotically consistent procedures for determining features of causal structure from data even if we allow for the possibility of unobserved confounding variables and/or an unknown time order. For certain combinations of directed acyclic graphs and probability distributions, these procedures can infer the existence or absence of causal relationships. In particular, Spirtes et al. (1993, Ch. 5, 6) proved the Fisher consistency of these procedures. Pointwise consistency follows from the Fisher consistency and the uniform consistency of the test procedures for conditional independence relationships that the procedures use.

We will call the framework assumed in Spirtes et al. (1993) the Spirtes–Glymour–Scheines model. The model is closely related to Wright's path diagrams (Wright, 1934) and structural equation models with acyclic path diagrams and uncorrelated errors (Bollen, 1989, pp. 80–2; Spirtes et al., 2000, pp. 27–8; Pearl, 1995). The Spirtes–Glymour–Scheines model has also been extended to allow cyclic directed graphs and correlated errors, although this is not described here; see Spirtes et al. (1998) and Richardson (1996). The reader is referred to Lauritzen (1996) and Wermuth (1980) for further details on statistical graphical models. A related approach to causation is based on the theory of counterfactuals; for a recent discussion of this approach, see Dawid (2000). Philosophical comments on causality can be found in many places; see Humphreys & Freedman (1996) and Spirtes et al. (1997) for example.

This paper explores the apparent discrepancy between the common statistical wisdom and the aforementioned results. We show that under the Spirtes–Glymour–Scheines model there do, in certain canonical examples, exist 'pointwise consistent' tests but there do not exist 'uniformly consistent' tests for causal effects. This implies that these tests are guaranteed to yield correct answers with an infinite sample size, but that no test can make such guarantees in finite samples, even approximately, no matter how large the sample. Furthermore, we show that valid, consistent confidence intervals do not exist for causal effect parameters in the presence of potential confounding variables. The methodological implications of these results for observational studies are controversial, and are briefly discussed in § 10.

In § 2 we review the Spirtes–Glymour–Scheines model. In § 3 we introduce several canonical examples. In § 4 we discuss consistent tests. In § 5 we present the main result of the paper, the nonexistence of uniformly consistent causal inference procedures. Section 6 examines the implications for confidence intervals and point estimates. Section 7 discusses Bayes tests. In § 8 we discuss the problem of inferring time order. A generalisation of the results is discussed in § 9. Some further remarks are in § 10. Proofs of results are in Appendix 1. In the interest of brevity, some proofs are omitted. To make the results precise, it is necessary to introduce a certain amount of notation. For the reader's convenience, we have included a glossary of notation in Appendix 2, which also defines the notion of 'd-separation'. Throughout the paper, we write $X \perp\!\!\!\perp Y$ to mean that the random variables X and Y are independent. We write $X \perp\!\!\!\perp Y | Z$ to mean that the random variables X and

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

Y are independent given Z, where X, Y and Z may represent individual random variables or sets of random variables.

2. The Spirtes–Glymour–Scheines model for causal inference

A directed acyclic graph G is a set of vertices with arrows between some pairs of vertices such that there is no directed cycle, in that one cannot start at a vertex and follow a directed path back to that vertex; a directed path between vertices A and B is a sequence of vertices starting at A and ending at B of the form $A \to X_1 \to \ldots \to X_k \to B$ or of the form $A \leftarrow X_1 \leftarrow \ldots \leftarrow X_k \leftarrow B$. An example of a directed acyclic graph is given in Fig. 1.

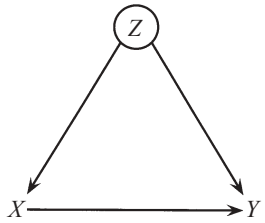

Fig. 1. Directed acyclic graph for Example 1.

The Spirtes–Glymour–Scheines model begins with a triple (G, V, P), where G is a directed acyclic graph with a set of vertices V, $V = (X_1, \ldots, X_k)$ is a vector of random variables and P is a joint probability distribution for V. The random variable X_j takes values in a set $\mathcal{X}_j$. Arrows represent causal relationships: if there is an arrow pointing from X_i to X_j, it means that X_i has a direct causal effect on X_j, relative to V. In what follows we will always assume that the distribution P is 'Markov' to G and we will sometimes assume that P is 'faithful' to G. We now proceed to define these two concepts.

The variable X_i is in the set of 'parents' of a variable X_j in a directed acyclic graph G, denoted by $\text{PA}_G(X_j)$, if there is an edge $X_i \to X_j$ in G; X_j is a 'descendant' of X_i in G if there is a directed path from X_i to X_j or $X_i = X_j$. A distribution P with density function p is 'Markov' to G if

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{n} p\{x_i | \text{pa}_G(x_i)\},$$

where $p\{x_i | \text{pa}_G(x_i)\} = p(x_i)$ when $\text{pa}_G(x_i) = \varnothing$. The formula above is called the 'Markov factorisation' of P according to G. Equivalently, P is Markov to G if for every subset W of V we have that $X_i \perp\!\!\!\perp \hat{X}_i | \text{PA}_G(X_i)$, where $\hat{X}_i$ are all variables in V that are not descendants of X_i in G. If each of the conditional probabilities $p\{X_i | \text{pa}_G(x_i)\}$ in the Markov factorisation is well defined, then P has 'full Markov support' relative to G. Let $\mathcal{P}(G)$ denote all distributions that are Markov to G. Note that if P is in $\mathcal{P}(G)$ then P is 'compatible' with G. The Markov assumption is that if G represents the data generating mechanism for P, then $P \in \mathcal{P}(G)$.

Given $P \in \mathcal{P}(G)$, let $\mathcal{I}(P)$ represent all independence and conditional independence relationships that hold for the variables in V under P. Let $\mathcal{I}_G = \bigcap_{Q \in \mathcal{P}(G)} \mathcal{I}(Q)$ be all independence relationships that are common to all the distributions in $\mathcal{P}(G)$. We say that P is 'faithful' to G if $\mathcal{I}(P) = \mathcal{I}_G$. Otherwise, P is called 'unfaithful'. In other words, P is faithful if it does not possess extra independence relationships not shared by all the other distributions in $\mathcal{P}(G)$.

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

What reasons are there for considering the assumption of faithfulness to directed acyclic graphs to be interesting? After all, as Matus & Studeny have pointed out, there are 18 300 distinct sets of conditional independence relationships that hold in some probability distribution over four variables, but directed acyclic graphs over four variables can faithfully represent fewer than 1000 patterns of conditional independence relationships (Matus & Studeny, 1995a, b).

Why expect the patterns of conditional independence relationships that hold among a set of variables to be among the small percentage that can be faithfully represented by directed acyclic graphs? One reason that we have chosen to concentrate on inferring directed acyclic graph models and their generalisations is that they provide a clear causal semantics for explaining mechanisms by which a probability distribution can be generated (Spirtes et al., 2000, Ch. 3; Pearl, 2000, pp. 27–40). In addition, such models have long played an important role in various sciences, such as econometrics.

One reason that the faithfulness assumption should be of interest to statisticians is that it is implicit in a variety of statistical practices. For example, it is common practice to select variables that are of interest in an observational causal study by regressing a variable Y on a set of regressors X and then removing from the study those variables with small regression coefficients. The regression coefficient of a variable Z is zero if and only if Z is independent of Y conditional on X. Hence this practice implicitly assumes that a conditional independence relationship indicates a zero causal effect, which is the faithfulness assumption. In addition, in cases where $\mathscr{P}(G)$ can be parameterised by a family of distributions with a parameter of finite dimension, the set of unfaithful distributions typically has Lebesgue measure zero (Spirtes et al., 2000, pp. 41–2; Meek, 1995).

Remark 1. It is important to note that the Spirtes–Glymour–Scheines procedures, in their most general form, allow for the possibility that the observed variables were measured with selection bias, leave out latent variables, and include correlated errors and cycles. All of these possibilities imply that the set of conditional independence relationships that hold among the observed variables may not be faithful to a directed acyclic graph over the observed variables. Spirtes et al. (2000) introduce generalisations of directed acyclic graphs, closely related to the extensions of directed acyclic graphs found in the structural equation modelling literature, to handle these various possibilities. They still assume, however, that the set of conditional independence relationships among the observed variables are faithful to some generalised graph, and this set of conditional independence relationships is still a small fraction of the total number of possible sets of conditional independence relationships. The extension of the Spirtes–Glymour–Scheines methods to generalised graphs does not materially affect any of the arguments or conclusions of this paper.

Let $\Omega(G) \subset \mathscr{P}(G)$ denote the set of all distributions that are faithful to G. The faithfulness assumption is that, if G represents the data generating mechanism for P, then $P \in \Omega(G)$. As Spirtes et al. (2000) say, '... the Faithfulness Condition can be thought of as the assumption that conditional independence relations are due to causal structure rather [than] to accidents of parameter values'.

Let $\mathscr{G}$ be some set of directed acyclic graphs of interest. Define $\mathscr{P}(\mathscr{G}) = \cup_{G \in \mathscr{G}} \mathscr{P}(G)$ and $\Omega(\mathscr{G}) = \cup_{G \in \mathscr{G}} \Omega(G)$. In many of our examples, $\mathscr{G}$ has the following special structure. We start with time-ordered random variables and we let G be the unique, complete graph consistent with this time ordering; complete means no missing arrow. Then $\mathscr{G}$ is G together with all subgraphs of G where a subgraph is obtained by deleting some of the arrows in G. This is the structure of $\mathscr{G}$ throughout the paper except in § 8.

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

Typically, we do not observe all the random variables V. In this case we write $V=(O, U)$, where O are the observed random variables and U are the unobserved, or latent, random variables. Let $\tilde{P}$ denote the marginal distribution of P for the observed O. We suppose that we have independent and identically distributed observations $O^n=(O_1, \ldots, O_n)$ from $\tilde{P}$.

We can now summarise the Spirtes–Glymour–Scheines model. We start with a set of graphs $\mathcal{G}$ and we assume there exists a triple (G, V, P), where $G \in \mathcal{G}$, $V=(O, U)$, and that P is Markov and faithful to G. We observe n independent and identically distributed observations $O^n=(O_1, \ldots, O_n)$ from $\tilde{P}$, the O-marginal of P. We shall address the following question: given two variables of interest X_i and X_j in O, and given a sample O^n from $\tilde{P}$, can we test whether or not there is an arrow from X_i to X_j in G, or, more precisely, is there a directed path from X_i to X_j in G which except for the endpoints contains only unobserved variables? We also address a related question: can we estimate the size of the causal effect, as defined in § 4·2, of X_i on X_j? It is critical to recognise that in general the analyst does not know the graph G that generated the data and the only available information is the data O^n and possible background knowledge, such as time order.

Even given the Markov and faithfulness assumptions, many distributions over the random variables in a directed acyclic graph are compatible with more than one directed acyclic graph which yield different predictions about the causal effect of a variable. For example, if there are just two measured variables X and Y which are dependent, the joint distribution over X and Y is compatible with X causing Y or Y causing X. In the former case, the causal effect of X on Y is not zero, while in the latter case it is zero. When faced with this situation, the Spirtes–Glymour–Scheines procedures generate the output 'no conclusion'. However, there are some distributions that are compatible only with a set of directed acyclic graphs all of which yield the same prediction about the causal effect of a variable. The Spirtes–Glymour–Scheines approach is to characterise which distributions have this property, under a variety of background assumptions.

3. Some key examples

Before proceeding with any technical details, we introduce some examples. In this section, we deal with the examples theoretically: we assume that, rather than given a sample O^n of size n, we are given the marginal distribution of O. In §§ 4 to 8, we examine the realistic setting in which the marginal distribution of O is unknown but we obtain a sample O^n.

Example 1. Let $V=(Z, X, Y)$, where X is the number of cigarettes smoked in one year, Y is a measure of disease at a later time, and Z represents all potential common causes of X and Y. Suppose, for the purposes of this example, that the following structural equation model holds, where X, Y and Z are standardised Normal variables:

$$Z=\varepsilon_Z, \quad X=\alpha Z+\varepsilon_X, \quad Y=\beta X+\gamma Z+\varepsilon_Y, \tag{1}$$

where ε_X, ε_Y and ε_Z are Normally distributed with mean 0. The directed acyclic graph G is given in Fig. 1. In this model, X is a direct cause of Y if and only if $\beta \neq 0$. In other words, there is an arrow from X to Y if and only if $\beta \neq 0$. Also, $\mathrm{cov}(X, Y)=\beta+\alpha\gamma$. Suppose we do not observe Z. Thus, $O=(X, Y)$ and $U=Z$. Suppose we are given that X and Y are uncorrelated. We will show that this information is sufficient to deduce that X is not a cause of Y under the Spirtes–Glymour–Scheines model. First note that there are two explanations for the zero correlation. One possibility is that β and at least one of α and

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

γ are 0. In this case, inferring that X does not cause Y from the data is correct. On the other hand, even if β is large we can still obtain a zero correlation between X and Y. This will happen if $\beta = -\alpha\gamma$. For example, we might have $\beta = 0{\cdot}72$, $\alpha = 0{\cdot}9$ and $\gamma = -0{\cdot}8$. In this case, X does cause Y, indeed the causal effect β is large compared to the variance, and inferring that X does not cause Y would be an error. Intuitively, it seems 'unlikely' that $\beta = -\alpha\gamma$. To be specific, the set of (α, β, γ) values that satisfy $\beta = -\alpha\gamma$ has Lebesgue measure 0 in R^3. A distribution for which $\beta = -\alpha\gamma$ is unfaithful to G since it has an independence between some variables not because of missing arrows in the directed acyclic graph but because of parameters cancelling each other out as in the equation $\beta = -\alpha\gamma$. If a priori we rule out these unfaithful distributions, which have Lebesgue measure 0, then the conclusion must be that X does not cause Y.

The theorems that we will present about what causal conclusions can be reliably inferred from marginal probability distributions are consequences of the following relationships between the set $\mathscr{P}$ of all joint probability distributions over X, Y and Z, the set B of all values of θ and the set $\tilde{\mathscr{P}}$ of all marginal probability distributions over X and Y. Here, θ denotes the causal effect or treatment effect, defined formally in § 4·2. These relationships are illustrated in Fig. 2. Each joint distribution P uniquely determines both a marginal $g(P)$ and a value of $\theta = f(P)$. However, both the function g and the function f are many-to-one functions: for each marginal distribution $\tilde{P}$ there are many distributions P such that $\tilde{P} = g(P)$, and for each value θ there are many distributions P such that $\theta = f(P)$.

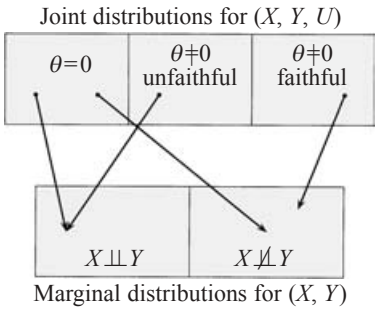

Fig. 2. The mapping from joint distributions for (X, Y, U) to marginal distributions for (X, Y). Unfaithful joint distributions with a nonzero causal effect can make X and Y independent. If the unfaithful distributions, which have Lebesgue measure 0, are ruled out, then $X \perp\!\!\!\perp Y$ if and only if $\theta = 0$.

For any subset $\mathscr{Q}$ of $\mathscr{P}$, let

$$f(\mathscr{Q}) = \{\theta \in B : \text{there exists } P \in \mathscr{Q} \text{ with } f(P) = \theta\}.$$

Assume that the marginal distribution $\tilde{P}$ is observed. Then, if faithfulness is not assumed, for all P, $f(g^{-1}\tilde{P}) = B$. If faithfulness is assumed, it is still the case that for all $\tilde{P}$, in which X and Y are dependent, $f(g^{-1}\tilde{P}) = B$; however, unlike the unfaithful case, for all $\tilde{P}$ in which X and Y are independent, $f(g^{-1}\tilde{P}) = 0$. Thus, if we assume faithfulness, and if X is observed to be independent of Y, then the conclusion $\theta = 0$ can reliably be returned, but if X is observed to be dependent on Y then the conclusion 'don't know' must be returned. Given just samples from the observed marginal distribution, this is what entails the existence of pointwise consistent tests of $\theta = 0$, but not of other values of θ. However, it is also

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

the case that, even if we assume faithfulness, there are distributions $P \in \mathscr{P}$ such that $f(P)$ is arbitrarily large, but the correlation between X and Y is arbitrarily small; as we will show, this entails that, given samples from the marginal distribution, there is no uniformly consistent test of any value of θ.

Example 2: *Deducing noncausation.* This example is the same as Example 1 but we drop the Normality assumption and state the set-up in more generality. Let $V = (Z, X, Y)$, where Z represents unobserved confounders, X is an exposure variable and Y is an outcome. We observe X and Y but not Z. Thus, $U = Z$ and $O = (X, Y)$. We further assume that the variables are in known time order with Z preceding X and X preceding Y. There are exactly eight possible directed acyclic graphs for these variables as shown in Fig. 3, all of which are subgraphs of G. Then $\mathscr{G}$ is this set of eight directed acyclic graphs. We have partitioned the directed acyclic graphs into a set A with five graphs and a set B of three. If we assume faithfulness, $X \perp\!\!\!\perp Y$ for the directed acyclic graphs in the set B, and X and Y are dependent for the graphs in A.

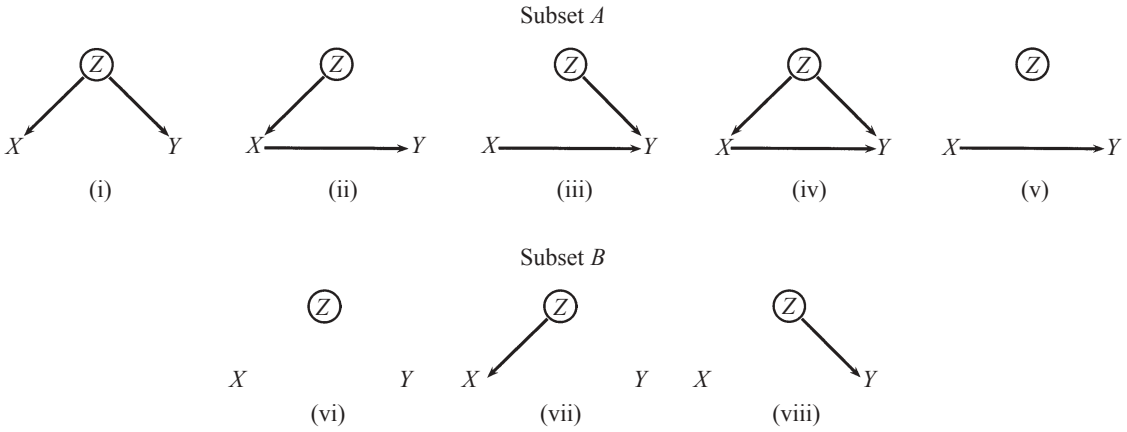

Fig. 3. Subgraphs for Example 2 partitioned into subsets A and B.

Suppose $X \perp\!\!\!\perp Y$ so that the true directed acyclic graph G generating the data is in set B. Since there is no arrow from X to Y in any graph in B, we then conclude that X does not cause Y. Suppose now that X is not independent of Y so the true graph G is in set A. Since, in set A, some graphs have an arrow from X to Y while others do not, we report 'no conclusion'. Unlike most familiar statistical procedures, a key component of these causal procedures is that they can produce the output 'no conclusion'. This is formalised in § 4.

In this example, if the joint distribution of X, Y and U is known then the hypothesis that X does not cause Y is identifiable because the hypothesis is true if and only if X and Y are independent given U. The hypothesis is not a function of the marginal distribution of the observables X and Y if the assumption of faithfulness is not imposed. If however the assumption of faithfulness is imposed then the hypothesis is a function of the marginal distribution of X and Y if and only if X and Y are independent, in which case the hypothesis is true.

In the realistic setting where the population distribution of O is unknown, Spirtes et al. (2000, Ch. 6) suggest that we perform a test of the null hypothesis $X \perp\!\!\!\perp Y$ based on data O^n where the level of the test may depend on n. If the test does not reject, we will accept $X \perp\!\!\!\perp Y$ and conclude that X did not cause Y.

We shall see that this procedure can result in pointwise consistent but not uniformly consistent tests. Pointwise consistency in this setting means the following. For any fixed

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

faithful distribution for X, Y, and U under which X and Y are marginally independent, there exists a sample size n, depending on the alternative, such that we reject the alternative and accept the hypothesis of no causation with probability arbitrarily close to one. If X and Y are not independent, there exists a sample size n such that the test produces the output 'no decision' with probability arbitrarily close to one. On the other hand, there is no uniformly consistent test; that is, there does not exist a sample size n, independent of the alternative, such that we reject all alternatives outside a neighbourhood of the null which we will define precisely in § 4, with probability arbitrarily close to one. In the absence of uniform consistency nontrivial confidence intervals do not exist. Sections 4 to 8 make these ideas precise.

Example 3: *Deducing causation.* Now suppose we have three time ordered variables (X, Y, Z). We want to know if Y causes Z. We allow for a confounding variable R between Y and Z and another possible confounding variable S between X and Z; see Fig. 4. Hence, $V = (X, Y, Z, R, S)$. The decomposition into observables and unobservables is $O = (X, Y, Z)$ and $U = (R, S)$. We could also allow for a third confounding variable between X and Y but this is not crucial to the example.

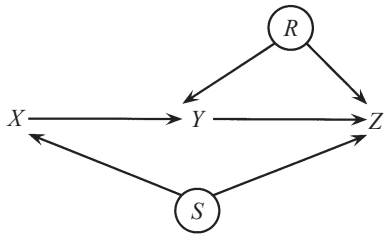

Fig. 4. Directed acyclic graph for Example 3.

Suppose we are given that

(i) X and Y are not independent, and not independent given Z,

(ii) Y and Z are not independent and not independent given X,

(iii) X and Z are not independent but are independent given Y.

If we assume faithfulness and use the d-separation rules of Pearl (1988, pp. 116–22), it follows that at least one arrow from R and one arrow from S must be removed. Moreover, the arrow from Y to Z must be present. We have thus inferred that Y causes Z. The reader who is unfamiliar with the d-separation rules will not find this immediately obvious from looking at the directed acyclic graph; d-separation is defined in Appendix 2.

Spirtes et al. (2000) suggest testing the joint null hypothesis (i)–(iii) with an α-level test, and, if it does not reject, concluding that (i)–(iii) are true and thus that Y causes Z. Again, this test can be made to be pointwise but not uniformly consistent.

These examples give a sense of how causal inferences are made in this framework; see Spirtes et al. (2000) for the details of the general theory, which does not require either time order or the assumption of no latent variables. For clarity, the results in this paper mostly focus on these last two key examples.

4. Consistent tests

4·1. *Introduction*

To answer questions about the presence or absence of causal arrows from data in the Spirtes–Glymour–Scheines model, we need to use some test or model-search technique

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

for choosing between alternative causal models. In this section we review some basic facts about statistical tests.

Recall that G is a directed acyclic graph for variables $V=(X_1,\ldots,X_k)$, $\mathscr{P}(G)$ are all probability distributions Markov to G and $\Omega(G)\subset\mathscr{P}(G)$ are the distributions that are faithful to G. Define

$$\mathscr{P}\mathscr{G}=\bigcup_{G\in\mathscr{G}}\{(P,G);\,P\in\mathscr{P}(G)\},\quad \Omega\mathscr{G}=\bigcup_{G\in\mathscr{G}}\{(P,G);\,P\in\Omega(G)\}.$$

Thus, (P,G) is in $\Omega\mathscr{G}$ if and only if P is faithful to $G\in\mathscr{G}$. Note that $\Omega\mathscr{G}$ is not the Cartesian product of $\Omega(\mathscr{G})$ and $\mathscr{G}$. Recall also that $V=(O,U)$, where O is the subset of the variables of V that are observable and U are the remaining variables. For any $P\in\mathscr{P}(\mathscr{G})$, $\tilde{P}$ denotes the marginal distribution of O under P. We assume that we have a random sample $O^n=(O_1,\ldots,O_n)$ from $\tilde{P}$.

4·2. *Consistent tests*

Let $\mathscr{O}^n=\mathscr{O}\times\ldots\times\mathscr{O}$, where $\mathscr{O}$ is the range of the set of random variables O, and let P^n denote the n-fold product measure corresponding to P. Let $\theta=T(P,G)$, where T maps $\Omega\mathscr{G}$ into $\mathscr{R}$. We call θ a 'parameter.' In § 4·2 we will formally define the 'causal effect' or 'treatment effect' which will be the parameter we are interested in. For now we let θ be an arbitrary parameter. We will use the terms 'causal effect' and 'treatment effect' interchangeably.

Let θ_0 be a fixed constant and consider testing $H_0:\theta=\theta_0$ versus $H_1:\theta\neq\theta_0$. A test is a mapping $\phi:\mathscr{O}^n\to\{0,1,2\}$. The interpretation is that $\phi(O^n)=0$ means 'choose H_0', $\phi(O^n)=1$ means 'choose H_1', and $\phi(O^n)=2$ means 'make no decision'; we ignore randomised tests which make a random rather than a deterministic decision. We shall be studying the asymptotic, large-sample, properties of tests. Thus, suppose we specify a test ϕ_n for each sample size n and let $\phi=(\phi_1,\phi_2,\ldots)$ represent this sequence of tests. In what follows, all limits refer to the sample size n tending to ∞. When we refer to a test, we mean a sequence of tests ϕ. A test which always reports 'no decision' is not useful. To rule out such trivial cases, we make the following definition.

DEFINITION 1. *A test ϕ is nontrivial if there exists some $P\in\Omega(\mathscr{G})$ and some $j\in\{0,1\}$ such that*

$$\lim_{n\to\infty}P^n\{\phi_n(O^n)=j\}=1. \tag{2}$$

In other words, a test is nontrivial if there is at least one distribution P in the model such that, with probability tending to one, it eventually settles on a definite decision. Throughout the rest of this paper, we restrict attention to nontrivial tests. A test is consistent if, at least for large sample sizes, it does not report an incorrect decision. This is formalised as follows. Let

$$\mathscr{P}_{\mathscr{G}0}=\bigcup_{G\in\mathscr{G}}\{P\in\mathscr{P}(G);\,T(P,G)=\theta_0\},\quad \mathscr{P}_{\mathscr{G}1}=\bigcup_{G\in\mathscr{G}}\{P\in\mathscr{P}(G);\,T(P,G)\neq\theta_0\},$$
$$\Omega_{\mathscr{G}0}=\bigcup_{G\in\mathscr{G}}\{P\in\Omega(G);\,T(P,G)=\theta_0\},\quad \Omega_{\mathscr{G}1}=\bigcup_{G\in\mathscr{G}}\{P\in\Omega(G);\,T(P,G)\neq\theta_0\}.$$

DEFINITION 2. *A test ϕ is pointwise consistent if*

(i) *for every $P\in\Omega_{\mathscr{G}0}$, $\lim_{n\to\infty}P^n\{\phi_n(O^n)=1\}=0$ and*

(ii) *for every $P\in\Omega_{\mathscr{G}1}$, $\lim_{n\to\infty}P^n\{\phi_n(O^n)=0\}=0$.*

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

Let $\mathscr{P}_{\mathscr{G}\delta} = \cup_{G \in \mathscr{G}}\{P \in \Omega(G); |T(P, G) - \theta_0| \geqslant \delta\}$; that is $\mathscr{P}_{\mathscr{G}\delta}$ contains distributions P compatible with a directed acyclic graph G such that the treatment effect is at least δ away from θ_0. The notation for the subset of faithful distributions of these sets is similar but with Ω in place of $\mathscr{P}$.

Definition 3. *A test ϕ is uniformly consistent if*
(i) $\lim_{n\to\infty} \sup_{P \in \Omega_{\mathscr{G}0}} P^n\{\phi_n(O^n) = 1\} = 0$ *and*
(ii) *for every* $\delta > 0$, $\lim_{n\to\infty} \sup_{P \in \Omega_{\mathscr{G}\delta}} P^n\{\phi_n(O^n) = 0\} = 0$.

The difference between a pointwise and a uniformly consistent test is the presence of the suprema in the definitions. Uniform consistency is what links asymptotic, i.e. large-sample, procedures to finite samples. To see this, suppose that, given $\varepsilon > 0$, we wish to find a sample size $n_0(\varepsilon)$ such that the probability of falsely rejecting the null hypothesis and accepting the alternative is bounded above by ε if $n \geqslant n_0(\varepsilon)$. To achieve this goal with a test that is pointwise but not uniformly consistent requires one to know the true distribution that generates the data; that is, $n_0(\varepsilon)$ is a function of the unknown P. Hence, there is no sample size n which guarantees that the probability of choosing the wrong hypothesis is less than ε, if P is unknown.

If a test is uniformly consistent then we can find $n_0(\varepsilon)$ which does not depend on P such that $\sup_{P \in \Omega_{\mathscr{G}0}} P^n\{\phi_n(O^n) = 1\} \leqslant \varepsilon$ for all $n \geqslant n_0(\varepsilon)$. Furthermore, if the test is uniformly consistent, then, given an error rate $\varepsilon > 0$ and a $\delta > 0$, we can find an n_0 such that, for all $n \geqslant n_0$, the probability of falsely rejecting the null is bounded above by ε and the probability of falsely accepting the null and rejecting the alternative when $|T(P, G)| > \delta$ is also bounded above by ε for all $n \geqslant n_0$, without knowledge of P or G.

There is a relationship between tests and confidence intervals. Loosely speaking, tests cannot be inverted to form uniformly consistent confidence intervals unless they are uniformly consistent. We discuss this further in § 5.

4·3. *Causal effects and hypotheses about causal graphs*

In our earlier informal description of Example 2, we constructed tests as follows. We began with a directed acyclic graph G, and let $\mathscr{G}$ be the set of all subgraphs of G. We took A and B to be two sets of subgraphs. Define $\Omega(A)$ to be all distributions faithful to some directed acyclic graph in A and $\Omega(B)$ to be all distributions faithful to some directed acyclic graph in B. We constructed A and B such that $\Omega(\mathscr{G}) = \Omega(A) \cup \Omega(B)$ and $\Omega(A) \cap \Omega(B) = \varnothing$. Finally, we tested A versus B by testing $P \in \Omega(A)$ versus $P \in \Omega(B)$. To put this in the terminology of § 4·2, deciding that $P \in \Omega(B)$ is like reporting $\phi_n(O^n) = 0$, while deciding $P \in \Omega(A)$ is like reporting $\phi_n(O^n) = 2$. At this point, it will clarify the discussion if we introduce the notion of a causal effect and then express our tests in terms of the causal effect parameter.

The causal effect of a variable X_i on another variable X_j can be thought of as the mean, or some other summary, of the distribution of X_j when X_i is set to a particular value x. To make this precise, let

$$p(v) = \prod_{r=1}^{k} p\{x_r | \mathrm{pa}_G(x_r)\}$$

be the joint density of V, where $\mathrm{PA}_G(X_r)$ is the set of parents of X_r in G, and $\mathrm{pa}_G(x_r)$ is some particular configuration of values of $\mathrm{PA}_G(X_r)$. Define a new distribution $p_{G,X_i=x}(v)$ obtained by starting with the joint density $p(v)$ and replacing the factor $p\{x_i | \mathrm{pa}_G(x_i)\}$ with

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

a distribution which is a point mass at x; that is

$$p_{G,X_i=x}(v) = \delta(x_i) \prod_{r \neq i} p\{x_r | \mathrm{pa}_G(x_r)\},$$

where $\delta(x_i) = 1$ if $x_i = x$, and $\delta(x_i) = 0$ otherwise.

Graphically, $p_{G,X_i=x}(x_j)$ has the following interpretation. Start with the graph G and create a new graph by breaking all arrows into X_i. Then $p_{G,X_i=x}$ is the distribution in the new graph. We regard $p_{G,X_i=x}$ as the density obtained by setting $X_i = x$. It should not be confused with $p(.|X_i = x)$, which is the distribution when X_i is observed to be x. We call the distribution with density $p_{G,X_i=x}$ the causal distribution. Note that, unlike $p(.|X_i = x)$, $p_{G,X_i=x}$ depends upon both the joint distribution p, and on the graph G; two different data generating mechanisms represented by G_1 and G_2 respectively could generate the same distribution, but disagree on their respective causal distributions, $p_{G_1,X_i=x}$ and $p_{G_2,X_i=x}$.

Let

$$\tilde{p}_{G,X_i=x}(x_j) = \sum_{x_i; i=j} p_{G,X_i=x}(v)$$

be the marginal density for X_j obtained from $p_{G,X_i=x}(v)$.

The causal effect is usually defined to be some contrast functional of $\tilde{p}_{G,X_i=x}(x_j)$. In particular, assume that X_i is binary and define the causal effect θ as the mean of X_j when X_i is set to 1 minus the mean of X_j when X_i is set to 0. To be specific, if $\mathscr{X}_j$ denotes the possible values that X_j can take, then the 'causal effect' is

$$\theta = \sum_{x_j \in \mathscr{X}_j} x_j \tilde{p}_{G,X_i=1}(x_j) - \sum_{x_j \in \mathscr{X}_j} x_j \tilde{p}_{G,X_i=0}(x_j) = E_{\tilde{P}_{G,X_i=1}}(X_j) - E_{\tilde{P}_{G,X_i=0}}(X_j).$$

The parameter θ is sometimes called the 'treatment effect'. Since θ is a function of the joint distribution P and G we write $\theta = T(P, G)$. An alternative way to define the causal effect is by way of the theory of counterfactuals used by Rubin (1974) and Robins (1986, 1987, 1995, 1997). This approach leads to exactly the same formulae for the causal effects.

If G is the directed acyclic graph in Example 2, the formula for the causal effect of X on Y turns out to be

$$\theta \equiv T(P, G) \equiv \int \{E(Y|X = 1, Z = z) - E(Y|X = 0, Z = z)\}\, dP(x). \tag{3}$$

If G is the directed acyclic graph in Example 3, the formula for the causal effect of Y on Z is

$$\begin{aligned} \theta &\equiv T(P, G) \\ &\equiv \int \{E(Z|Y = 1, R = r, S = s) - E(Z|Y = 0, R = r, S = s)\}\, dP(r, s). \end{aligned}$$

5. Nonexistence of uniformly consistent tests in the key examples

5·1. *Overview*

In this section we discuss the nonexistence of uniformly consistent tests for causal hypotheses in the two key examples from § 3. Recall that the first example involves a potential cause X, an outcome Y and a potential confounder Z. The second example

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

involves a covariate X, a potential cause Y, an outcome Z and two potential confounders R and S. In each example we will take X and Y to be binary. The variables Z, R and S are discrete and take only finitely many values.

5·2. *Example* 2

The variables $V=(Z, X, Y)$ are assumed to be time ordered as Z then X then Y. The variable Z represents possible confounding variables. The question of interest is whether or not X causes Y; that is, is there an arrow from X to Y?

Recall that, if we observe X and Y to be nearly independent in a large sample, the conclusion from the informal reasoning in § 3 is that X does not cause Y, that is $\phi_n(O^n)=0$. However, when X and Y were observed to be dependent, the outcome of the test was 'no decision about causation', that is $\phi_n(O^n)=2$. Let G be the complete directed acyclic graph for V, with an arrow from X to Y, as in Fig. 1. Let $\mathscr{G}$ be G and all subgraphs of G. Following the discussion in § 4, we test $H_0: \theta=0$ versus $H_1: \theta \neq 0$.

In this example, where the time order is given, the results about the existence of pointwise and uniformly consistent tests depend on whether or not Z is observed. In practice, we are mainly interested in the case where Z is unobserved. We include the case where Z is observed to make it clear exactly what is lost by the presence of the possibility of unobserved confounding. Note that when Z is observed we have $O=(Z, X, Y)$ and $U=\varnothing$ while if Z is unobserved then we have $O=(X, Y)$ and $U=Z$.

Theorem 1. *Consider testing $H_0: \theta=0$ versus $H_1: \theta \neq 0$. Given the Markov assumption and full Markov support, if Z is observed then there exist pointwise and uniformly consistent tests of this hypothesis. Given the Markov assumption, the faithfulness assumption and full Markov support, if Z is not observed then there exist pointwise consistent tests but there is no uniformly consistent test.*

Remark 2. Theorem 1 may be extended to testing $H_0: \theta=\theta_0$ versus $H_1: \theta \neq \theta_0$ with $\theta_0 \neq 0$. In this case, when Z is unobserved, there is not even a nontrivial, pointwise consistent test.

The proof of the theorem is in Appendix 1; related results may be found in Spirtes et al. (2000). Note that, in the absence of the assumption that there is no latent variable, if faithfulness is not assumed then there are not even pointwise consistent tests of the hypothesis when $\theta_0=0$. This is because the independence of X and Y can always be explained either faithfully by no edge from X to Y and either no edge from Z to X or no edge from Z to Y in which case the treatment effect is zero, or unfaithfully, in which case the treatment effect is not zero.

An intuitive explanation of the proof is as follows. For each sample size n we can always choose a P_n (i) which is faithful but arbitrarily close to unfaithful, (ii) which has a large causal effect, and (iii) whose marginal makes X and Y nearly independent. The existence of a single such P_n precludes a uniformly consistent test. As n grows, the offending P_n is closer and closer to being unfaithful.

It is worth recalling that there do exist uniformly consistent tests for testing associations. For example, a parameter that measures the association between X and Y is the risk difference $\alpha=\mathrm{pr}(Y=1|X=1)-\mathrm{pr}(Y=1|X=0)$. Another is the odds ratio

$$\psi=\frac{\mathrm{pr}(Y=1|X=1)\,\mathrm{pr}(Y=0|X=0)}{\mathrm{pr}(Y=0|X=1)\,\mathrm{pr}(Y=1|X=0)}.$$

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

PROPOSITION 1. *Let α and ψ be the parameters defined above, and let $O=(X,Y)$. For any α_0, there exists a uniformly consistent test for $H_0:\alpha=\alpha_0$ versus $H_1:\alpha\neq\alpha_0$, and similarly for ψ.*

This result follows since, for associations, we have that α and ψ are functions of P only through $\tilde{P}$.

Unless one imposes additional a priori smoothness or dimension reducing assumptions on either $\mathrm{pr}(Y=1|X,Z)$ or $\mathrm{pr}(X=1|Z)$, there is no uniformly consistent test of the hypothesis $Y \perp\!\!\!\perp X|Z$ even when X and Y are Bernoulli when Z is discrete with many levels, more precisely, with the number of levels increasing with sample size, or is absoutely continuous with respect to Lebesgue measure (Ritov & Bickel, 1990; Robins & Ritov, 1997). Thus in the context of Theorem 1 there is no uniformly consistent test of $H_0:\theta=0$ versus $H_1:\theta\neq 0$ even when Z is observed but no smoothness assumption is imposed. The philosophical implication is that even if somehow one were able to measure all variables that precede X and Y and thereby have measured all potential confounding factors, nonetheless, in the absence of substantive prior information, the problem of causal inference would still not be solved in the sense that no uniformly consistent test of causal hypotheses would exist.

Remark 3. In a randomised experiment, it is possible to construct uniformly consistent tests of causal hypotheses. To see this, consider our example. Randomisation breaks the arrow from Z to X in Fig. 1. It then can be shown that

$$\theta=\alpha=\mathrm{pr}(Y=1|X=1)-\mathrm{pr}(Y=1|X=0)$$

and, as noted above, there exist uniformly consistent tests for this risk difference α.

5·3. *Example* 3

Now consider time ordered variables X, Y and Z and potential confounding variables R and S. We are interested in the causal effect of Y on Z. The directed acyclic graph that describes all of the causal relationships in the model is given in Fig. 4. Again, we take all the variables to be discrete. Recall that θ is now the causal effect of Y on Z as defined in § 4.

THEOREM 2. *Consider testing $H_0:\theta=\theta_0$ versus $H_1:\theta\neq\theta_0$. Given the Markov assumption and full Markov support, if R and S are observed, there exist pointwise and uniformly consistent tests for this hypothesis. Given the Markov assumption, the faithfulness assumption and full Markov support, if R and S are unobserved, there exist pointwise consistent tests for this hypothesis but there do not exist uniformly consistent tests.*

Remark 4. In contrast to the situation in Theorem 1, there do exist nontrivial, pointwise consistent tests for the case where $\theta_0\neq 0$.

6. CONFIDENCE INTERVALS AND POINT ESTIMATES

6·1. *Estimability of causal effects*

An 'estimator' is a sequence of functions $\hat{\theta}_n$ of $O^n=(O_1,\ldots,O_n)$.

An estimator is consistent if, for every $(P,G)\in\Omega\mathcal{G}$, and every $\varepsilon>0$,

$$P^n\{|T(P,G)-\hat{\theta}_n|>\varepsilon\}\to 0.$$

An estimator $\hat{\theta}_n$ is 'uniformly consistent' if, for every $\varepsilon>0$,

$$\sup_{(P,G)\in\Omega\mathcal{G}} P^n\{|\hat{\theta}_n-T(P,G)|>\varepsilon\}\to 0.$$

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

To discuss estimation in the Spirtes–Glymour–Scheines framework, we need to allow for a more general notion of an estimator. Recall that their tests sometimes return 'no conclusion', which is what allows the tests to be pointwise consistent. Similarly, for estimation, we should allow the possibility that the answer is sometimes 'no estimate is provided'. Formally, we define a 'generalised estimator' $\hat{\theta}_n$ to be any measurable mapping from $\mathcal{O}^n$ into the Borel sets of the real line. In particular, when $\hat{\theta}_n(O^n) = \mathcal{R}$, as opposed to a point, this corresponds to 'no estimate is provided'. To define consistency for generalised estimators we need to define the distance between the estimator and the true value. Given a set A and a point b define $\tau(A, b) = \inf_{a \in A} d(a, b)$. If $A = \{a\}$ consists of a single point, then $\tau(A, b) = |a - b|$, the usual Euclidean distance. For our purposes, we say that a generalised estimator is 'consistent' at (P, G) if, for every $\varepsilon > 0$, $P^n\{\tau(\hat{\theta}_n, T(P, G)) > \varepsilon\} \to 0$. A generalised estimator $\hat{\theta}_n$ is 'uniformly consistent' if, for every $\varepsilon > 0$,

$$\sup_{(P,G) \in \Omega\mathcal{G}} P^n\{\tau(\hat{\theta}_n, T(P, G)) > \varepsilon\} \to 0.$$

A generalised estimator $\hat{\theta}_n$ is 'nontrivial' if there exists $P \in \Omega(\mathcal{G})$ such that

$$P^\infty(A_n^c \text{ finitely often}) = 1,$$

where A_n is the event that $\hat{\theta}_n$ consists of a single point. In other words, $\hat{\theta}_n$ is nontrivial if there is at least one P for which $\hat{\theta}_n$ eventually becomes a point.

Theorem 3. *Given the Markov assumption and full Markov support, in Example* 2, *if* Z *is observed, and in Example* 3, *if* R *and* S *are observed, and there is no unobserved variable, there is a uniformly consistent estimator of* θ. *Given the Markov assumption, the faithfulness assumption and full Markov support, in Example* 2, *if* Z *is not observed, and in Example* 2, *if* R *and* S *are not observed, there is no uniformly consistent estimator of* θ *but there exist pointwise consistent estimators.*

Some insight can be gained into the lack of consistency in the following way. Let H be the map that takes P into the marginal $\tilde{P}$. The parameter θ is 'identified at P' if $\theta(P)$ is the same for all $H^{-1}(\tilde{P})$. It is well known that if there is any P such that a parameter is not identified then it cannot be consistently estimated. In our case, the situation is a bit subtle. If the parameter were defined on $\mathcal{P} \equiv \mathcal{P}(\mathcal{G})$ instead of $\Omega \equiv \Omega(\mathcal{G})$, then it would follow immediately that θ is not identifiable for any P and hence not consistently estimable. One might hope that restricting ourselves to Ω instead of $\mathcal{P}$ might help since, for example, in Example 2, if $\tilde{P}$ is a distribution with X independent of Y then θ is identified at P and is equal to 0. However it can be shown that Ω is dense in $\mathcal{P}$, so uniform consistent estimation is still precluded. As with testing, there are uniformly consistent estimators of the association parameters α and ψ defined in § 5.

6·2. *Confidence intervals*

Consider a collection of maps

$$\mathcal{C} = \{I_{\alpha,n};\ n = 1, 2, \ldots, \alpha \in [0, 1]\},$$

where each $I_{\alpha,n}$ takes $\mathcal{O}^n$ into the Borel sets of $\mathcal{R}$. If

$$\liminf_n \inf_{(P,G) \in \Omega\mathcal{G}} P^n\{T(P, G) \in I_{\alpha,n}(O^n)\} \geqslant 1 - \alpha$$

then we call $\mathcal{C}$ a 'confidence map' and we call $I_{\alpha,n}$ a $1 - \alpha$ asymptotic confidence region.

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

Note that confidence sets are, by definition, uniform. There does not appear to be a useful 'pointwise' notion of a confidence map. A confidence map $\mathscr{C}$ is 'consistent' if it eventually omits all false values; that is, if for some (P, G) we have that, for every $\delta > 0$ and for every $\alpha \in [0, 1]$,

$$\sup_{(Q,H) \in \Omega\mathscr{G}; |T(Q,H) - (P,G)| > \delta} P^n\{T(Q, H) \in I_{\alpha,n}(O^n)\} \to 0.$$

THEOREM 4. *Given the Markov assumption, the faithfulness assumption and full Markov support, if in Example* 2 *Z is unobserved then there does not exist a consistent, confidence map for* θ. *The same holds in Example* 3, *if R and S are unobserved.*

7. BAYESIAN INFERENCE

So far, we have concentrated on frequentist inference. Robins & Wasserman (1999) discuss related Bayesian results; see also Heckerman et al. (1994). This section gives a brief summary of the Bayesian results and how they relate to the frequentist results; see Robins & Wasserman (1999) for a full account and Glymour et al. (1999) for a discussion. The goal of the section is to elucidate the behaviour of the causal test as a function of the prior distribution. We focus on Example 2.

There are weaker and stronger senses of Bayesian convergence, which are analogous to pointwise and uniform consistency. In the weaker sense, which uses a fixed prior, of Bayesian convergence, the procedures described in Spirtes et al. (2000) asymptotically converge to the truth for any prior over the parameters which is absolutely continuous with Lebesgue measure and in which the true causal directed acyclic graph has nonzero probability; in the stronger sense of Bayesian convergence, which uses a prior that changes with sample size, whether or not these procedures asymptotically converge to the truth is more sensitive to the prior and in particular depends upon how probable confounding is relative to the sample size.

There are eight subgraphs in Example 2; see Fig. 2. Denote these subgraphs by $G_1, \ldots, G_8$ and let γ_j represent the parameters of the joint distribution for G_j. For example, in the complete graph G_4, we could define the parameters γ_4 by

$$\gamma_4 = \{\gamma_{4z}, \gamma_{4xz}, \gamma_{4yxz}, z \in \mathscr{Z}, x, y \in \{0, 1\}\},$$

where

$$\gamma_{4z} = \mathrm{pr}(Z = z), \quad \gamma_{4xz} = \mathrm{pr}(X = x | Z = z), \quad \gamma_{4yxz} = \mathrm{pr}(Y = y | X = x, Z = z).$$

For a Bayesian analysis we put a prior on the subgraphs and on the parameters of each subgraph. We can write the prior as $\pi(G_j)\pi(\gamma_j | G_j)$. We assume that the prior on γ_j is smooth, i.e. absolutely continuous with respect to Lebesgue measure with bounded density. After we have observed the data O^n, Bayes' theorem gives us a posterior probability $\pi(G_j | O^n)$ for each possible subgraph. We can then test for the presence of a causal effect by finding the posterior odds B_n of 'X causes Y' versus 'X does not cause Y':

$$B_n = \frac{\mathrm{pr}(M | O^n)}{\mathrm{pr}(M^c | O^n)} = \frac{\mathrm{pr}(H_1 | O^n)}{\mathrm{pr}(H_0 | O^n)},$$

where M is the set of all graphs G in which there is an arrow from X to Y, that is M corresponds to H_1. The number B_n tells us the strength of evidence in favour of the hypothesis of causation. In testing consistently estimable parameters such as measures of

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

association, it is typically the case that B_n will tend to 0 or infinity in probability as the sample size increases and thus B_n will be decisive.

Let us compare the Bayesian and frequentist analyses. In contrast to the Bayesian analysis, the frequentist analysis is a worst case analysis: for each sample size n we can always choose a P_n which is faithful but arbitrarily close to unfaithful, which has a large causal effect, and whose marginal makes X and Y nearly independent. The existence of a single such P_n precludes a uniformly consistent test. However, as n grows, the offending P_n must be closer and closer to being unfaithful. In other words, the set $\mathscr{P}_n$ of offending P_n gets smaller as n increases. With fixed prior mass on each subgraph and a smooth, fixed prior density on the parameters of each subgraph, the weaker sense of Bayesian convergence, we see that $\pi(\mathscr{P}_n) \to 0$. Thus, these distributions play a lesser and lesser role as n increases, which is why $B_n \to 0$ in probability, in the cases mentioned above.

There is a stronger sense of Bayesian convergence, which more closely links the asymptotic results with finite sample sizes. In this stronger sense of Bayesian convergence, instead of using a fixed prior the prior mass on the graphs changes with sample size n. This is only meant to reflect the fact that the prior on $\mathscr{H}$, which denotes all subgraphs where Z does not have arrows into both X and Y, can be small relative to sample size. We do not literally envisage the prior changing with n. In this stronger sense of Bayesian convergence, some prior distributions will lead to tests that are pointwise consistent while others will lead to trivial tests. Here $\mathscr{H}$ can be thought of as the hypothesis that there is no unobserved confounding. Robins & Wasserman (1999) showed that, under weak conditions, if $\pi(\mathscr{H}) = o(n^{-\frac{1}{2}})$, then B_n will stay bounded, in probability, away from 0 and infinity and hence will not be decisive. Thus, if one's prior belief that there is no confounding, relative to the sample size, is small, the Bayes test provides no decision. On the other hand, if there were reason to believe that there might not be unobserved confounding, then we would set $\pi(\mathscr{H})$ to be not small relative to sample size; for example, we might put fixed positive mass on each subgraph. In this case, when $X \perp\!\!\!\perp Y$, $B_n \to 0$ in probability.

As pointed out previously, we can always choose a P_n which is faithful but arbitrarily close to unfaithful, which has a large causal effect and whose marginal makes X and Y nearly independent. With a fixed prior, as sample size grows these distributions play a lesser and lesser role. However, suppose that, as the sample size increases, we put more probability near the unfaithful distributions or we put less and less mass on the subgraphs with no confounder. We capture this by using a prior π_n that depends on n. Then $\pi_n(\mathscr{P}_n)$ does not tend to 0 and the Bayes test makes no decision. Which priors are more reasonable depends on subject matter knowledge; see Robins & Wasserman (1999) and Spirtes et al. (1998, 2000) for a discussion of the role of the prior. There are parallel results for point and interval estimation but we shall not pursue the details here.

8. Inferring time order

Throughout this paper we have assumed that the variables have known time order. The lack of uniform consistency arose because of the presence of potential unobserved confounding variables. The Spirtes–Glymour–Scheines procedures that we have discussed also permit one to infer the time order of the variables in some cases. The purpose of this section is to show that, for the problem of inferring time order, there again do not exist uniformly consistent procedures, but in this case the lack of uniformity does not even require unobserved confounding variables.

Suppose we have four binary random variables $V = (X_1, \ldots, X_4)$ and consider the two

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

graphs in Fig. 5. Suppose we are interested in the causal effect of X_3 on X_4. In Fig. 5(a), it can be shown that the causal effect is

$$\theta = \mathrm{pr}(X_4 = 1 \,|\, X_3 = 1) - \mathrm{pr}(X_4 = 1 \,|\, X_3 = 0).$$

In Fig. 5(b), θ is identically 0. This follows because the arrow points from X_4 to X_3, so manipulating X_3 will have no effect on X_4. Note that getting the direction of the arrow wrong thus implies getting the value of θ wrong, so we will cast questions about getting the directions of the arrows correctly as questions about the causal effect θ.

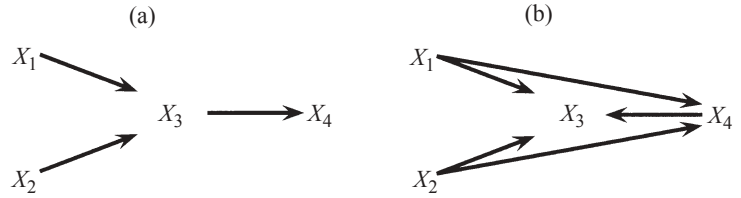

Fig. 5. Directed acyclic graph for § 8.

Is it possible to distinguish these two directed acyclic graphs? Suppose it were known that $X_1 \perp\!\!\!\perp X_2$ and $\{X_1, X_2\} \perp\!\!\!\perp X_4 | X_3$, and that no other independence relationship not entailed by these holds. Assume there is no unobserved confounding variable. Then only the first directed acyclic graph is consistent with these relationships assuming faithfulness, and there are pointwise consistent tests of the effect of X_3 on X_4. However, there is no uniformly consistent test of whether the effect of X_3 on X_4 is nonzero.

For inferring time order, G is also unknown. Define $\mathscr{G}$ to be all directed acyclic graphs for $X_1, \ldots, X_4$.

THEOREM 5. *Suppose we have n independent and identically distributed observations* $V^n = (V_1, \ldots, V_n)$ *and that there is no unobserved confounding variable. Given the Markov assumption, the faithfulness assumption and full Markov support, in the above example there are pointwise consistent tests but no uniformly consistent test of* $H_0 : \theta = \delta$ *versus* $H_1 : \theta \neq \delta$, *where* $\delta \neq 0$. *Similarly, there is no uniformly consistent point estimator and there is no consistent* $1 - \alpha$ *confidence interval for* θ.

Remark 5. Note that the case $H_0 : \theta = 0$ is not included in Theorem 5. When we are trying to find a causal time order with no confounder, the value 0 plays a special role. To see this, consider the simple case of two variables X and Y. Suppose we know that there is no confounding variable but we do not know the time order, so we are trying to decide if $X \to Y$ or $Y \to X$. In particular, let θ be the causal effect of X on Y. Clearly, it is not possible to test consistently whether or not $\theta = \delta$, for $\delta \neq 0$. Any correlation between X and Y can be explained equally well by the graph $X \to Y$ or the graph $Y \to X$. However, we can test whether or not $\theta = 0$ merely by testing for an association between X and Y. Even without time order, the only explanation for no association between X and Y is that $\theta = 0$.

9. A GENERALISATION

Theorems 1 and 2 referred to specific examples but the results can be made more general. In this section we give a generalisation to linear, Normal structural equation

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

models (Bollen, 1989, Ch. 6). Although we believe that the results can be generalised even further, we do not pursue that here.

Consider the model in § 8. Suppose that the true directed acyclic graph is the one in Fig. 5(a), that faithfulness is assumed, and the set of alternative causal structures are those represented by any other directed acyclic graph containing (X_1, X_2, X_3, X_4) and any number of other variables, even latent variables. Then there are pointwise consistent tests of the treatment effect of X_3 on X_4 that are functions of (X_1, X_2, X_3, X_4), but there is no such pointwise consistent test that is a function of (X_3, X_4) alone.

The question arises whether or not there is an analogous phenomenon for uniform consistency. Our examples have shown that, even assuming faithfulness, if the possibility of latent variables is allowed, then there is no nontrivial uniformly consistent test of the treatment effect of X_1 on X_2 that is a function only of the joint distribution of (X_1, X_2). Are there nonetheless causal structures for which, even allowing for the possibility of hidden variables, there are nontrivial uniformly consistent tests of the treatment effect of X_1 on X_2 which are functions of some larger set of variables containing X_1 and X_2? The theorem stated in this section shows that this is not the case: no matter what other variables are measured, there is no uniformly consistent test of the treatment effect of X_1 on X_2.

Theorem 6. *If $\mathscr{G}$ is the set of all directed acyclic graphs containing B and C, Z is the causal effect of B on C, and $\mathscr{O}$ is any set of variables containing B and C, there is no nontrivial uniformly consistent test of $\theta = z$ against $\theta \neq z$ with respect to $\Omega_{\mathscr{G}z}$ and $\Omega_{\mathscr{G}\delta z}$. Here,*

$$\Omega_{\mathscr{G}z} = \bigcup_{G \in \mathscr{G}} \{P \in \Omega(G);\ T(P, G) = z\}, \quad \Omega_{\mathscr{G}\delta z} = \bigcup_{G \in \mathscr{G}} \{P \in \Omega(G);\ |T(P, G) - z| > \delta\}.$$

10. Remarks

Many results in statistical inference are asymptotic; that is, they hold true for large sample sizes. If asymptotic results are to be used, it is important that the large-sample results can be approximated with finite samples. One way of ensuring this is to insist that our tests and estimates are uniformly consistent. We have shown that causal procedures based on the Markov and faithfulness assumptions are not uniformly consistent. The same problem affects model-search or model-averaging methods, such as stepwise regression. In these methods, each subgraph obtained by deleting some set of arrows is regarded as a separate model. Then a model-search technique, such as Bayesian model selection, BIC, AIC and so on, is used to find a best graph or to average over graphs. These methods will also fail to have uniform consistency properties, as will any variable selection technique that attempts to eliminate variables with small treatment effects on the basis of the observed data.

This leaves open the question of what to do when analysing observational studies. Each of the suggested strategies is controversial. The problems stem from unobserved confounding or unknown time order. Given a time order, one suggestion is to follow standard epidemiological advice and measure as many confounders as possible, but even then we cannot be sure that there are not further unobserved confounding variables. Furthermore, conditioning on inappropriate variables can introduce bias into the procedures; see Spirtes et al. (1998) and Greenland et al. (1999). A second possible approach is to perform a sensitivity analysis in which we quantify how the estimates and tests change as a function of the amount of unobserved confounding. Examples of sensitivity analysis can be found in Rosenbaum (1993), Robins et al. (1999), Manski (1990, 1995, Ch. 3–4) and Balke &

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

Pearl (1997). A third strategy is to continue to pursue search strategies which are not uniformly consistent but which do satisfy weaker consistency conditions, such as Bayesian consistency conditions.

Finally, as we pointed out after Proposition 1, even if we knew the time order and that there was no unmeasured confounder, uniformly consistent tests of causal effects do not exist when the measured potential confounders are either continuous or take many levels; see Robins & Ritov (1997) for further discussion.

Acknowledgement

Spirtes' research was supported by the National Science Foundation. Wasserman's research was supported by the National Institutes of Health and the National Science Foundation. We thank the reviewers for helpful comments.

Appendix

Proofs

In the proofs, we make use of the total variation distance. If P and Q are two probability measures, then the total variation distance between P and Q is defined to be $d(P, Q) = \sup_A |P(A) - Q(A)|$, the supremum being over all measurable sets. If P and Q live on a finite sample space $S = \{s_1, \ldots, s_m\}$ then it can be shown that $d(P, Q) = (\frac{1}{2})\sum_{j=1}^{m} |p_j - q_j|$, where $p_j = P(\{s_j\})$ and $q_j = Q(\{s_j\})$. Furthermore, $d(P, Q)$ is a continuous function of the p_j's and the q_j's. For any vector $(a_1, \ldots, a_K)$ we define $\|a\|_\infty = \max_j |a_j|$.

Proof of Theorem 1. We only prove the case $\theta_0 = 0$, which is the more difficult case. First suppose that Z is observed. We construct a test that is both uniformly and pointwise consistent. Let $\Delta = \Delta(P) = (\beta_{000}, \ldots, \beta_{11B})$, where $\beta_{rst} = \mathrm{pr}(X = r, Y = s, Z = t)$ for $r, s \in \{0, 1\}$ and $t \in \{0, \ldots, B\}$. In this example, because the time order is given and there is no latent variable, the treatment effect is completely determined by Δ. Let $\hat{\Delta} = (\hat{\beta}_{000}, \ldots, \hat{\beta}_{11B})$, where $\hat{\beta}_{rst}$ is the observed sample proportion corresponding to β_{rst}, that is

$$\hat{\beta}_{rst} = \mathrm{card}\{(X_i, Y_i, Z_i);\ X_i = r, Y_i = s, Z_i = t\}/n.$$

Let $\varepsilon_n = \sqrt{(\log n/n)}$. It follows from Hoeffding's inequality that

$$\sup_{P \in \Omega(\mathscr{G})} P^n\{\|\hat{\Delta} - \Delta(P)\|_\infty > \varepsilon_n\} \to 0.$$

Define $\hat{\theta} = T(\hat{\Delta}) := T(\hat{\Delta}, G)$. Since θ is a continuous function of Δ,

$$\sup_{P \in \Omega(\mathscr{G})} P^n\{|\hat{\theta} - \theta(P)| > \varepsilon_n\} \to 0. \tag{A1}$$

Define $\phi_n(O^n) = 1$ if $|\hat{\theta}| > \varepsilon_n$ and $\phi_n(O^n) = 0$ otherwise. It follows from (A1) that this test is uniformly and pointwise consistent.

Now assume that Z is not observed. First we construct a pointwise consistent test. Define $\alpha = \mathrm{pr}(Y = 1 | X = 1) - \mathrm{pr}(Y = 1 | X = 0)$ and let

$$\hat{\alpha} = \left\{ \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i} - \frac{\sum_{i=1}^n Y_i (1 - X_i)}{\sum_{i=1}^n (1 - X_i)} \right\}.$$

Again, we have that $\sup_{P \in \Omega(\mathscr{G})} P^n\{|\hat{\alpha} - \alpha(P)| > \varepsilon_n\} \to 0$. Define $\phi_n(O^n) = 0$ if $|\hat{\alpha}| < \varepsilon_n$ and $\phi_n(O^n) = 2$ otherwise.

If $\alpha(P) = 0$ then $P^n\{\phi_n(O^n) = 0\} \to 1$ so the test is nontrivial. Now we show that the test is pointwise consistent. Suppose that $P \in \Omega_0$. Then $P^n\{\phi_n(O^n) = 1\} \to 0$ by definition of ϕ. If $P \in \Omega_{\mathscr{G}1}$ and P is

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

faithful, then $\alpha(P) \neq 0$. Hence, from the uniform consistency of $\hat{\alpha}$ we see that $P^n\{\phi_n(O^n) = 2\} \to 1$. This establishes the pointwise consistency of the test.

Now we show that no test can be uniformly consistent. Basically we will show that if $\theta \neq 0$ we will never be able to establish this by observation of O^n alone. This is intuitively clear from Fig. 2. Suppose there is a nontrivial, uniformly consistent test. We will show that this leads to a contradiction. Since ϕ is nontrivial, (2) holds for some P. Suppose $T(P, G^*) = \delta$ for some $G^* \in \mathcal{G}$ and $\delta \neq 0$, so that $P \in \Omega_{\mathcal{G}\delta}$ and $P^n\{\phi_n(O^n) = 1\} \to 1$ as $n \to \infty$. Let $p(x, y, z)$ denote the density of P. Let G be directed acyclic graph (i) in Fig. 2. Then $\mathcal{P}(G)$ is all Q whose densities satisfy $q(x, y, z) = q(z)q(x|z)q(y|z)$. Note that $Q \in \mathcal{P}(G)$ implies that $T(Q, G) = 0$ if Q has full Markov support. If $\tilde{P}$ has full support, let $P^*(X, Y) = \tilde{P}$, and otherwise let $P^*(X, Y)$ be a distribution with full support such that $d(\tilde{P}, P^*) < \varepsilon$.

There exists a $Q \in \mathcal{P}(G)$ such that Q has full Markov support, is faithful to G, $\tilde{Q} = P^*$ and $T(Q, G) = 0$. To see this, define $\alpha_{rs} = P^*(X = r, Y = s)$ for $r, s \in \{0, 1\}$. Define a distribution Q as follows. Let $Q(Z = 0) = \alpha_{00}$, $Q(Z = 1) = \alpha_{01}$, $Q(Z = 2) = \alpha_{10}$ and $Q(Z = 3) = \alpha_{11}$. Also, let

$$Q(X = 0, Y = 0 | Z = 0) = 1, \quad Q(X = 0, Y = 1 | Z = 1) = 1,$$

$$Q(X = 1, Y = 0 | Z = 2) = 1, \quad Q(X = 1, Y = 1 | Z = 3) = 1,$$

Let q be the corresponding density. It can be verified directly that $q(x, y, z) = q(z)q(x|z)q(y|z)$, so that $Q \in \mathcal{Q}$, has full Markov support, is faithful to G, $\tilde{Q} = P^*$, and $T(Q, G) = 0$. Hence $Q \in \Omega_{\mathcal{G}0}$. Since the test depends only on O^n and since $\tilde{Q} = P^*$, we have that

$$\sup_{R \in \Omega_{\mathcal{G}0}} R^n\{\phi_n(O^n) = 1\} \geq Q^n\{\phi_n(O^n) = 1\} = \tilde{Q}^n\{\phi_n(O^n) = 1\} = P^{*n}\{\phi_n(O^n) = 1\}$$

$$> \tilde{P}^n\{\phi_n(O^n) = 1\} - \varepsilon = P^n\{\phi_n(O^n) = 1\} - \varepsilon \to 1 - \varepsilon.$$

Hence, $\sup_{R \in \Omega_{\mathcal{G}0}} R^n\{\phi_n(O^n) = 0\}$ does not tend to 0 as required for a uniformly consistent test.

Now suppose instead that $P \in \Omega_{\mathcal{G}0}$. Thus $P^n\{\phi_n(O^n) = 0\} \to 1$. If $X \perp\!\!\!\perp Y$ in $\tilde{P}$, let

$$\delta = P(Y = 0)P(Y = 1)/4,$$

and otherwise let

$$\delta = \{P(Y = 1 | X = 1) - P(Y = 1 | X = 0)\}/4.$$

For every ε, there exists a $Q \in \mathcal{P}_{\mathcal{G}\delta}$ such that $d(\tilde{P}^n, \tilde{Q}^n) > \varepsilon/2$. This fact is proved in Proposition A1 below. However, Q might not be faithful. Since T and $d(.,.)$ are both continuous functions of P, there exists a faithful $Q_n \in \Omega_{\mathcal{G}\delta}$ with full Markov support such that $|T(Q_n)| > \delta$ and $d(Q_n^n, Q^n) < \varepsilon/2$. It follows that $d(\tilde{Q}_n^n, \tilde{Q}^n) < \varepsilon/2$. Hence,

$$\sup_{R \in \Omega_{\mathcal{G}\delta/2}} R^n\{\phi_n(O^n) = 0\} \geq Q_n^n\{\phi_n(O^n) = 0\} > Q^n\{\phi_n(O^n) = 0\} - \varepsilon/2 = \tilde{Q}^n\{\phi_n(O^n) = 0\} - \varepsilon/2$$

$$> \tilde{P}^n\{\phi_n(O^n) = 0\} - \varepsilon/2 - \varepsilon/2 = P^n\{\phi_n(O^n) = 0\} - \varepsilon \to 1 - \varepsilon.$$

Since this is true for any $\varepsilon > 0$ we conclude that

$$\sup_{R \in \Omega_{\mathcal{G}\delta}} R^n\{\phi_n(O^n) = 0\} \to 1,$$

and hence the test is not uniformly consistent. □

PROPOSITION A1. *In Example 2, let $O = (X, Y)$. Consider $P \in \Omega_{\mathcal{G}0}$. If $X \perp\!\!\!\perp Y$ under P, let $\delta = P(Y = 1)P(Y = 0)/4$; otherwise let*

$$\delta = |P(Y = 1 | X = 1) - P(Y = 1 | X = 0)|/4.$$

For every $\varepsilon > 0$ there exists a $Q \in \mathcal{P}_{\mathcal{G}\delta}$ such that $d(\tilde{P}, \tilde{Q}) < \varepsilon/2$.

Proof. First suppose that $p(x, y, z) = p(z)p(x|z)p(y|z)$. Thus, P is faithful to subgraph (i) in A in Fig. 3. It follows that $\alpha(P) \neq 0$; that is X and Y are not marginally independent. If $\tilde{P}$ has full

support, let $P^*(X, Y) = \tilde{P}$, and otherwise let $P^*(X, Y)$ be a distribution with full support such that $d(\tilde{P}, P^*) < \varepsilon/2$, and

$$P^*(Y = 1 | X = 1) - P^*(Y = 1 | X = 0) > \{P(Y = 1 | X = 1) - P(Y = 1 | X = 0)\}/2.$$

Define Q as follows. Let $Q(Z = 0) = \frac{1}{2}$ and let

$$Q(X = x, Y = y | Z = z) = \tilde{P}^*(X = x, Y = y) \quad (z = 0, 1).$$

By construction, $\tilde{Q} = P^*$. Let G be directed acyclic graph (iv) in Fig. 2. By equation (3),

$$T(Q, G) = P^*(Y = 1 | X = 1) - P^*(Y = 1 X = 0) > \{P(Y = 1 | X = 1) - P(Y = 1 | X = 0)\}/2 > \delta.$$

Since $d(\tilde{P}, P^*) < \varepsilon/2$ and $\tilde{Q} = P^*$, $d(\tilde{P}, \tilde{Q}) < \varepsilon/2$.

For all other $P \in \Omega_{\mathscr{G}0}$ we have that $\alpha(P) = 0$, that is X and Y are marginally independent, so the above strategy does not work. Instead we proceed as follows. Let $P^*(X, Y)$ be a distribution with full support such that $d(\tilde{P}, P^*) < \varepsilon/2$, $X \perp\!\!\!\perp Y$ under P^*, and

$$P^*(Y = 0)P^*(Y = 1) > P(Y = 0)P(Y = 1)/2.$$

Let $a = P^*(X = 0)$ and $b = P^*(Y = 0)$, and let c and d be two reals between 0 and 1. Since X and Y are marginally independent under P^*, a and b completely determine P^*. Define Q as follows:

$$Q(X = 0) = a, \quad Q(Y = 0) = b, \quad Q(Z = 0 | X = 0, Y = 0) = 1 - c,$$

$$Q(Z = 0 | X = 0, Y = 1) = 0, \quad Q(Z = 0 | X = 1, Y = 0) = 1 - d, \quad Q(Z = 0 | X = 1, Y = 1) = 0,$$

and $q(x, y, z) = q(x)q(y)q(z | x, y)$ where q is the density of Q.

By inspection, $\tilde{Q} = \tilde{P}^*$. Also Q has the following properties:

$$Q(Z = 1) \geqslant 1 - b, \quad Q(Y = 1 | X = 0, Z = 0) = 0, \quad Q(Y = 1 | X = 1, Z = 0) = 0,$$

$$Q(Y = 1 | X = 0, Z = 1) = \frac{1 - b}{1 - b + cb}, \quad Q(Y = 1 | X = 1, Z = 1) = \frac{1 - b}{1 - b + db}.$$

As $d \to 0$, $Q(Y = 1 | X = 1, Z = 1) \to 1$ and, as $c \to 1$, $Q(Y = 1 | X = 0, Z = 1) \to 1 - b$. Let G be subgraph (iv) in Fig. 2. Then, by choosing the proper values for c and d, and using equation (3), we have that

$$T(Q, G) > b(1 - b)/2 = P^*(Y = 0)P^*(Y = 1)/2 > P(Y = 0)P(Y = 1)/4 = \delta.$$

Since $d(\tilde{P}, P^*) < \varepsilon/2$ and $\tilde{Q} = P^*$, $d(\tilde{P}, \tilde{Q}) < \varepsilon/2$. □

Proof of Theorem 4. Suppose a consistent confidence interval exists. Define a test as follows. Let α_n be a decreasing sequence such that $\alpha_n \to 0$ as $n \to \infty$. Let $\phi_n(O^n) = 1$ if $0 \notin I_{\alpha_n, n}$ and $\phi_n(O^n) = 0$ if $0 \in I_{\alpha_n, n}$. It is easy to see that this defines a uniformly consistent test, which contradicts Theorem 1. □

Proof of Theorem 5. Let P have density p such that

$$p(v) = p(x_1)p(x_2)p(x_3 | x_1, x_2)p(x_4 | x_3).$$

It follows that $\theta = T(P, G_1) = P(X_4 = 1 | X_3 = 1) - P(X_4 = 1 | X_3 = 0)$. Moreover, we can choose such a p so that $|\theta| \geqslant \delta$ for some $\delta > 0$. Fix $\varepsilon > 0$. It is possible to choose Q_n with density q_n such that Q_n is Markov and faithful to the graph G_2 in Fig. 5(b) and such that $d(P^n, Q_n^n) < \varepsilon$. Note that $T(Q_n, G_2) = 0$. The remainder of the proof is similar to the previous proofs. □

To prove Theorem 6, we first need two preliminary results. Also, we use the following facts concerning directed acyclic graphs. Consider a directed acyclic graph G and an associated set of linear coefficients D; for example, α might be the linear coefficient associated with the $X \to Z$ edge in Example 1. We will transform the variables so that each variable has variance 1. Each correlation matrix is then a polynomial in the linear coefficients, and hence a continuous function of the

coefficients; let the correlation matrix associated with D be $R(D)$. Then $R(D)$ is not faithful to G when some partial correlation is zero, but need not be zero; that is the polynomial in the coefficients is equal to zero but is not identically equal to zero for all values of the coefficients. We will refer to a polynomial that is set equal to zero as a polynomial equation. Each zero partial correlation is a polynomial equation in some subset of the correlations in $R(D)$, and hence a polynomial equation in some subset of the linear coefficients. Since the set of solutions of a polynomial equation that is not an identity is of Lebesgue measure zero, it follows that the Lebesgue measure of the coefficients that are unfaithful to a given graph G is zero.

THEOREM A1. *Let x be the causal effect of A on B, y the causal effect of A on C and z the causal effect of B on C. If $z \neq 0$, let M be a directed acyclic graph with edges from a latent variable A to B and C, and an edge from B to C; otherwise, let M be a directed acyclic graph with edges from a latent variable A to B and C, and no edge from B to C. Let P be a Normal distribution with the correlation matrix of a standard normal bivariate probability distribution with means 0 over B and C, let r be the correlation between B and C in P, and let $r\{R(D)\}$ be the correlation between B and C in the distribution with correlation matrix $R(D)$. For every r, z and $\delta > 0$, there is a set of coefficients D' such that the causal effect of B on C in D' is z, the Normal distribution with correlation matrix $R(D')$ is faithful to graph M, and $|r - r\{R(D')\}| < \delta$.*

Proof. First, consider the case where $z \neq 0$. We will set the values of x and y in a set of linear coefficients D in such a way that $R(D)$ satisfies the given constraints upon $r\{R(D)\}$. The constraints upon x, y and z that guarantee that the covariances are correlations and that the correlation $r\{R(D)\}$ between B and C is within δ of r are the following:

(i) $-1 \leqslant x \leqslant 1$,
(ii) $y^2 + x^2 + 2xyz \leqslant 1$,
(iii) $|z + xy - r| < \delta$.

These constraints can be derived either from noting that the contribution to the variance of B from A and the contribution to the variance of C from A and B have to be less than 1, or by calculating the constraints needed to make the entailed matrix with 1's along the diagonal positive definite. In order to satisfy these constraints, we will define values for x_1 and y, and then form the values for x. Set $\varepsilon = r - z + \sqrt{(z^2 - 2rz + 1)}$ if $z - r > 0$, and $\varepsilon = r - z - \sqrt{(z^2 - 2rz + 1)}$ if $z - r < 0$. Set $y = z - r + \varepsilon$ and $x_1 = (r - z)/y$. Note that y and $z - r$ have the same sign, and hence $x_1 \leqslant 0$. Also, $y \neq 0$ because, for positive z, $\sqrt{(z^2 - 2rz + 1)}$ is strictly bounded below at 0 when $|r| = 1$. We will now show how each of the conditions (i)–(iii) listed above can be satisfied.

(i) First we show how to set $-1 < x \leqslant 1$. Note that $x_1 = (r - z)/(z - r + \varepsilon)$. Since $-1 < r < 1$, $|r - z|$ is strictly bounded above by $|1 - z|$ or $|1 + z|$, depending on the sign of z. Now $z - r + \varepsilon = \pm\sqrt{(z^2 - 2rz + 1)}$ depending on the sign of $z - r$. Also, $\sqrt{(z^2 - 2rz + 1)}$ is strictly bounded below when $r = 1$ or $r = -1$, depending on the sign of z. Hence, $\sqrt{(z^2 - 2rz + 1)}$ is strictly bounded below at either $|1 - z|$ or $|1 + z|$. It follows that $|x_1| < 1$. Hence $-1 < x_1$. It follows that $-1 < x_1 \leqslant 0$. If y and z are of the same sign, set x to a value between x_1 and -1 and within $|\delta/(2y)|$ of x_1; if y and z are of opposite signs, set x to a value between x_1 and 0 and within $|\delta/(2y)|$ of x_1.

(ii) Next we show that $y^2 + z^2 + 2xyz \leqslant 1$. We have $y^2 + z^2 + 2x_1 yz = 1$ and $2xyz < 2x_1 yz$, and hence $y^2 + z^2 + 2xyz < y^2 + z^2 + 2x_1 yz = 1$.

(iii) Finally we show that $|r - z + xy| < \delta$. Note that $|r - z + xy| < \delta/2$ follows from the facts that $x_1 = (z - r)/y$ and x is within $|\delta/(2y)|$ of x_1.

It is possible that $R(D)$ is not faithful to M. We will now show that then there is a set of coefficients D' such that the value of z is the same in D and D', $R(D')$ is faithful to M and $|r\{R(D)\} - r\{R(D')\}| < \delta/2$. The set of coefficients D is unfaithful to M when it fails to satisfy at least one of the following constraints, in which $\rho(A, B).C$ is the partial correlation of A and B conditional on C:

(iv) $x - (y + xz)(z + xy) = x(1 - y^2 - z^2 - xyz) - (yz) \neq 0 \quad (\rho(A, B).C \neq 0)$,
(v) $y + xz - \{x(z + xy)\} = y(1 - x^2) \neq 0 \quad (\rho(A, C).B \neq 0)$,
(vi) $z + xy - \{x(y + xz)\} = z(1 - x^2) \neq 0 \quad (\rho(B, C).A \neq 0)$,
(vii) $x \neq 0 \quad (\rho(A, B) \neq 0)$,

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

(viii) $y + xz \neq 0 \quad (\rho(A, C) \neq 0)$,
(ix) $z + xy \neq 0 \quad (\rho(B, C) \neq 0)$.

Fixing the value of z at any value other than 0 in the polynomials (iv)–(ix) does not make any of them identically zero for all values of x and y. Since the set of solutions in x and y is of Lebesgue measure 0, for each polynomial equation there is a non-solution $D' = \{x', y', z\}$ that has the same z value as D', $|x'y' - xy| < \delta/2$, and that is faithful to M. It follows that $|r\{R(D')\} - r| < \delta$.

For the case where $z = 0$, choose a value of ε such that $0 < \varepsilon < \delta$, $-1 < r + \varepsilon < 1$ and $\varepsilon \neq 0$. In D', set $x = y = \sqrt{(r + \varepsilon)}$. It follows that $x \neq 0$ and $y \neq 0$, and the correlation between B and $C = xy$ is within δ of r. The probability distribution with the correlation matrix $R(D')$ is faithful to M. □

Let $\text{KL}(P, P')$ denote the Kullback–Leibler distance between two distributions.

THEOREM A2. *Let V be a set of variables containing B and C, let $P(V)$ be a Normal probability distribution faithful to some directed acyclic graph H, and let the correlation between B and C in $P(V)$ be r. For every r, $z \neq 0$ and $\delta > 0$, there is a directed acyclic graph G over the set of variables V', and an associated set of linear coefficients D, where $V \subset V'$, the causal effect of B on C in D is z, the Normal distribution with correlation matrix $R(D)$ is faithful to G, the marginal of the Normal distribution with correlation matrix $R(D)$ over V is $P'(V)$, and $\text{KL}\{P(V), P'(V)\} < \delta$.*

Proof. We have that $\text{KL}\{P(V), P'(V)\}$ is a continuous function of the correlation matrices of $P(V)$ and $P'(V')$. Hence, there is a δ' such that, if the sum of the absolute values of the differences of the correlations is less than δ', $\text{KL}\{P(V), P'(V)\} < \delta/2$. Choose such a δ'.

Form G' by making B the first variable and C the second variable and making each pair of vertices adjacent. Since G' is a complete graph it can represent any Normal distribution over the measured variables with a set of coefficients D'. At this stage of the construction, however, there is no guarantee that the linear coefficient z' from B to C in D' is equal to the given z, nor that $R(D')$ is faithful to G'. Form G by adding a latent variable A, and adding edges from A to B and C. From Theorem A1 there is a set of coefficients $\{x, y, z\}$ for the linear coefficient from A to B, the linear coefficient from A to C and the linear coefficient from B to C respectively, such that the resulting distribution is faithful to the subgraph over A, B and C, and the correlation between B and C is arbitrarily close to r. Let the coefficients D'' for G be $\{x, y, z\} \cup (D' - \{z'\})$; that is, keep all of the coefficients from D', except for replacing the coefficient for the edge from B to C by z, and add the coefficients x and y. This makes the causal effect of B on C equal to the given z, but it still may be the case that $P(D'')$ is not faithful to G.

No partial correlation is required to be zero in G', and hence no polynomial equation for a partial correlation in terms of the coefficients in D' is identically equal to zero. Given G, the equation for any correlation between variables in V in terms of the coefficients of D'' is the same as the corresponding equation in D', except that, everywhere the equation in D' contains z', the corresponding equation in D'' contains $(z + yx)$. Hence, the equation for any partial correlation between variables in V in terms of the coefficients in D'' is the same as the corresponding equation in terms of D', except that, everywhere the equation in D' contains z', the equation in D'' contains $(z + yx)$. None of the partial correlation equation in D' is an identity; hence none of the partial correlation equations among variables in V in terms of the variables in D'' is an identity, even if z is held fixed, because xy can be varied. It follows that there is a set of coefficients D, which contains the same value of z as does D'', such that the sum of the absolute values of the differences between the correlations of $P(V)$ and $P'(V')$ is less than δ'. Hence $\text{KL}\{P(V), P'(V)\} < \delta$. □

Proof of Theorem 6. Suppose that on the contrary there is a uniformly consistent test ϕ of $\theta = z$ against $\theta \neq z$. Since ϕ is nontrivial, either

(i) for some $P \in \Omega_{\mathscr{G}}$, $\lim_{n\to\infty} P^n\{\phi^n(\mathscr{O}^n) = 0\} = 1$, or
(ii) for some $P \in \Omega_{\mathscr{G}}$, $\lim_{n\to\infty} P^n\{\phi^n(\mathscr{O}^n) = 1\} = 1$.

Suppose that (ii) is the case. If $P \in \Omega_{\mathscr{G}_z}$ then ϕ is not uniformly consistent. Suppose then that $P \in \Omega_{\mathscr{G}_{\delta z}}$. By Theorem A2, for every distribution $P \in \Omega_{\mathscr{G}_{\delta z}}$ there is a distribution $D_n \in \Omega_{\mathscr{G}_z}$ with a

marginal for $\mathcal{O}$ such that the Kullback–Leibler distance $\mathrm{KL}(\tilde{P}_n; \tilde{D}_n) < \varepsilon^2$, where $\tilde{P}_n$ is the marginal of P over $\mathcal{O}$ and $\tilde{D}_n$ is the marginal of D_n over $\mathcal{O}$. Now,

$$\sup_A |\tilde{P}_n(A) - \tilde{D}_n^n(A)| \leqslant \frac{1}{2}\sqrt{\mathrm{KL}(\tilde{P}_n; \tilde{D}_n^n)} = \frac{\varepsilon}{2}.$$

Since ϕ is $\mathcal{O}$-measurable,

$$P^n\{\phi_n(\mathcal{O}^n) = 1\} = \tilde{P}_n\{\phi_n(\mathcal{O}^n) = 1\} \leqslant \tilde{D}_n^n\{\phi_n(\mathcal{O}^n) = 1\} + \varepsilon/2 = D_n^n\{\phi_n(\mathcal{O}^n) = 1\} + \varepsilon/2.$$

Since $P^n\{\phi_n(\mathcal{O}^n) = 1\} \to 1$, for all $\varepsilon/2 > 0$ there exists an N such that, for all $n > N$, $P^n\{\phi_n(\mathcal{O}^n) = 1\} > 1 - (\varepsilon/2)$. Hence, for all $\varepsilon > 0$ there exists an N such that, for all $n > N$,

$$D_n^n\{\phi_n(\mathcal{O}^n) = 1\} > 1 - (\varepsilon/2) - (\varepsilon/2) = 1 - \varepsilon.$$

Since each $D_n \in \Omega_{\mathcal{G}_Z}$, it follows that

$$\lim_{n\to\infty} \sup_{P \in \Omega_{\mathcal{G}\delta_Z}} P^n\{\phi_n(\mathcal{O}^n) = 1\} = 1$$

and hence ϕ is not uniformly consistent. The proof for case (i) is similar. □

Appendix 2

Definitions and glossary

Definition of d-separation. Let a 'directed path' d in a directed acyclic graph G be a sequence of distinct vertices $V_1 \ldots V_n$ such that, for $1 \leqslant i < n$, there is an edge $V_i \to V_{i+1}$ in G. A vertex X is an 'ancestor' of Y if $X = Y$ or there is a directed path from X to Y. Let an 'undirected path' U in a directed acyclic graph be a sequence of distinct vertices $V_1 \ldots V_n$ such that, for $1 \leqslant i \leqslant n$, there is either an edge $V_1 \to V_{i+1}$ or $V_i \leftarrow V_{i+1}$ in G. Here V_i is a 'collider' on U if there are edges $V_{i-1} \to V_i \leftarrow V_{i+1}$ on U; otherwise V_i is a 'non-collider' on U. For three disjoint sets of vertices X, Y and Z, X is 'd-separated' from Y given Z if there is no undirected path U from a vertex in X to a vertex in Y such that every collider on U is an ancestor of a member of Z, and no non-collider on U is in Z. Pearl (1988) states that if $P \in \mathcal{P}(G)$ and X is d-separated from Y given Z then $X \perp\!\!\!\perp Y | Z$.

Glossary of notation. Let $\mathcal{G}$ be a set of directed acyclic graphs and let G denote a member of $\mathcal{G}$. Then $\mathcal{P}(G)$ means all distributions Markov to G, $\Omega(G)$ means all distributions faithful to G and

$$\mathcal{P}(\mathcal{G}) = \bigcup_{G \in \mathcal{G}} \mathcal{G}(G), \quad \Omega(\mathcal{G}) = \bigcup_{G \in \mathcal{G}} \Omega(G),$$

$$\mathcal{P}_{\mathcal{G}0} = \bigcup_{G \in \mathcal{G}} \{P \in \mathcal{P}(G);\ T(P, G) = \theta_0\}, \quad \mathcal{P}_{\mathcal{G}1} = \bigcup_{G \in \mathcal{G}} \{P \in \mathcal{P}(G);\ T(P, G) \neq \theta_0\},$$

$$\Omega_{\mathcal{G}0} = \bigcup_{G \in \mathcal{G}} \{P \in \Omega(G);\ T(P, G) = \theta_0\}, \quad \Omega_{\mathcal{G}1} = \bigcup_{G \in \mathcal{G}} \{P \in \Omega(G);\ T(P, G) \neq \theta_0\},$$

$$\mathcal{P}_{\mathcal{G}\delta} = \bigcup_{G \in \mathcal{G}} \{P \in \mathcal{P}(G);\ |T(P, G) - \theta_0| > \delta\}, \quad \Omega_{\mathcal{G}\delta} = \bigcup_{G \in \mathcal{G}} \{P \in \Omega(G);\ |T(P, G) - \theta_0| > \delta\},$$

$$\mathcal{P}\mathcal{G} = \bigcup_{G \in \mathcal{G}} \{(P, G);\ P \in \mathcal{P}(G)\}, \quad \Omega\mathcal{G} = \bigcup_{G \in \mathcal{G}} \{(P, G);\ P \in \Omega(G)\}.$$

References

Balke, A. & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Am. Statist. Assoc.* **92**, 1171–6.

Bollen, K. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *J. Am. Statist. Assoc.* **95**, 407–48.

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077

Donoho, D. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16**, 1390–420.
Glymour, C., Spirtes, P. & Richardson, T. (1999). On the possibility of inferring causation from association without background knowledge. In *Computation, Causation, and Discovery*, Ed. C. Glymour and G. Cooper, pp. 323–32. Cambridge, MA: MIT Press.
Greenland, S., Pearl, J. & Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48.
Heckerman, D., Geiger, D. & Chickering, D. (1994). Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, Ed. R. Lopez de Mantaras and D. Poole, pp. 293–302. San Francisco, CA: Morgan Kaufmann.
Humphreys, P. & Freedman, D. (1996). The grand leap. *Br. J. Phil. Sci.* **47**, 113–8.
Lauritzen, S. (1996). *Graphical Models.* Oxford: Clarendon Press.
LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1**, 38–53.
Manski, C. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev. Papers Proc.* **80**, 319–23.
Manski, C. (1995). *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.
Matus, F. & Studeny, M. (1995a). Conditional independence properties among four random variables, I. *Combinat. Prob. Comp.* **4**, 269–78.
Matus, F. & Studeny, M. (1995b). Conditional independence properties among four random variables, II. *Combinat. Prob. Comp.* **4**, 407–17.
Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in AI*, Ed. P. Besnard and S. Hanks, pp. 411–8. San Francisco, CA: Morgan Kaufmann.
Pearl, J. (1995). Causal diagrams for empirical research (with Discussion). *Biometrika* **82**, 669–709.
Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann.
Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge.
Pearl, J. & Verma, T. (1991). A theory of inferred causation. In *Principles of Knowledge, Representation and Reasoning: Proceedings of the Second International Conference*, Ed. J. A. Allen, R. Filkes and E. Sandewall, pp. 441–52. San Francisco, CA: Morgan Kaufmann.
Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the 12th Conference on Uncertainty in AI*, Ed. E. Horvitz and F. Jensen, pp. 454–61. San Francisco, CA: Morgan Kaufmann.
Ritov, Y. & Bickel, P. (1990). Achieving information bounds in semi-parametric models. *Ann. Statist.* **18**, 925–938.
Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math. Mod.* **7**, 1393–512.
Robins, J. M. (1987). Addendum to 'A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect'. *Comp. Math. Appl.* **14**, 923–45.
Robins, J. M. (1995). Discussion of 'Causal diagrams for empirical research' by J. Pearl. *Biometrika* **82**, 695–8.
Robins, J. M. (1997). Causal inference from complex longitudinal data In *Latent Variable Modeling and Applications to Causality*, Lecture Notes in Statistics **120**, Ed. M. Berkane, pp. 69–117. New York: Springer Verlag.
Robins, J. M. & Ritov, Y. (1997). A curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statist. Med.* **16**, 285–319.
Robins, J. M. & Wasserman, L. (1999). On the impossibility of inferring causation from association without background knowledge. In *Computation, Causation, and Discovery*, Ed. C. Glymour and G. Cooper, pp. 323–31. Cambridge, MA: MIT Press.
Rosenbaum, P. (1993). *Observational Studies.* New York: Springer.
Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.
Spirtes, P. (1994). Building causal graphs from statistical data in the presence of latent variables. In *Proceedings of the IX International Congress on Logic, Methodology, and Philosophy of Science, Uppsala, 1991*, Ed. D. Prawitz, B. Skyrms and D. Westerstahl, pp. 813–29. Amsterdam: Elsevier.
Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction and Search.* New York: Springer-Verlag.
Spirtes, P., Glymour, C & Scheines, R. (1997). Reply to Humphreys' and Freedman's review of 'Causation, Prediction, and Search'. *Br. J. Phil. Sci.* **48**, 555–68.
Spirtes, P., Richardson, T., Meek, C., Scheines, R. & Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociol. Meth. Res.* **27**, 148–81.
Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction and Search*, 2nd ed. Cambridge, MA: MIT Press.
Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *J. Am. Statist. Assoc.* **75**, 963–72.
Wright, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5**, 161–215.

[*Received September* 2000. *Revised March* 2003]

bio9030311 28-08-03 10:09:18 Rev 14.05
The Charlesworth Group, Huddersfield 01484 517077