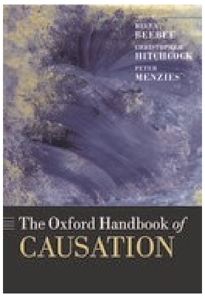


The Oxford Handbook of Causation

Helen Beebee (ed.) et al.

https://doi.org/10.1093/oxfordhb/9780199279739.001.0001

Published: 2009 **Online ISBN:** 9780191577246 **Print ISBN:** 9780199279739

CHAPTER

23 Causation and Statistical Inference

Clark Glymour

https://doi.org/10.1093/oxfordhb/9780199279739.003.0024 Pages 498–521

Published: 02 January 2010

Abstract

In the applied statistical literature, causal relations are often described equivocally or euphemistically as ‘risk factors’, or as part of ‘dimension reduction’. The statistical literature also tends to speak of ‘statistical models’ rather than of causal explanations, and to say that parameters of a model are ‘interpretable’, often means that the parameters make sense as measures of causal influence. These ellipses are due in part to the use of statistical formalisms for which a causal interpretation is wanted but unavailable or unfamiliar, and in part to a philosophical distrust of attributions of causation outside experimental contexts, misgivings traceable to the disciplinary institutionalization of claims of influential statisticians, notably Karl Pearson and Ronald Fisher. More candid treatments of causal relations have recently emerged in the theoretical statistical literature.

Keywords: causal relations, dimension reduction, risk factors, statistical models, causal influence, institutionalization

Subject: Philosophy of Science, Metaphysics, Philosophy

Series: Oxford Handbooks

1. Introduction

In the applied statistical literature, causal relations are often described equivocally or euphemistically as ‘risk factors’, or as part of ‘dimension reduction’. The statistical literature also tends to speak of ‘statistical models’ rather than of causal explanations, and to say that parameters of a model are ‘interpretable’, often means that the parameters make sense as measures of causal influence. These ellipses are due in part to the use of statistical formalisms for which a causal interpretation is wanted but unavailable or unfamiliar, and in part to a philosophical distrust of attributions of causation outside experimental contexts, misgivings traceable to the disciplinary institutionalization of claims of influential statisticians, notably Karl Pearson and Ronald Fisher. More candid treatments of causal relations have recently emerged in the theoretical statistical literature.

p. 499

2. Causal Interpretations of Statistical Models

In statistics, differences in the family of probability distributions considered are almost always accompanied by a new ‘model’ terminology, with the result that similarities and dissimilarities relevant to causal inference and prediction are sometimes obscured. Some of the most common frameworks include:

1. *ANOVA models*. Analysis of Variance is one of the most widely used methods for estimating the effect of one or more categorical variables on a continuous variable. A unit u_{ij} belongs to some group j having the same value for the potential cause, *X*, and the value of *Y* for u_{ij} is Y_{ij}. The model is

   $$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \tag{1}$$

   where μ is the average value across all values of *X*, α_j is the mean of the group with the jth value of *X*, and ►$_{ij}$ is the value of a Normally distributed random variable, with the same distribution for all units. The intuitive causal interpretation, where appropriate (which is not always), is that moving units from one group to group j, or intervening to give new units the jth value of *X*, would on average result in values of *Y* characterized by α_j.

2. *Recursive, linear structural equation models*. Variables are ordered and each is written as a linear function of a subset of its non-descendants in the ordering plus ‘noise’. Noises may be independently distributed or correlated. Variations include treating values of linear coefficients as random variables, and the use of binary variables as sources (exogenous variables). Some variables may be unrecorded or ‘latent’. The causal interpretation is that the equations (with fixed coefficients) specify the change that would occur in the dependent variable *Y* for a forced, exogenous unit change in any independent variable *X* occurring in the equation for *Y*, if other variables (including the noise variable for *Y*) were held constant by intervention. With random coefficients estimated by their means, the equations give the average change in *Y* for a unit forced, exogenous change in *X* (with similar constraints on other variables except the noise term). This class of statistical models includes as special cases factor analysis models, linear regression models, and principal components models.

3. *Non-recursive linear structural equation models* do not require an ordering of variables: *X* may occur in an equation for *Y*, and *Y* may reciprocally occur in an equation for *X*, and more generally a chain of equations may occur constituting (in graphical terms) a closed path from *X* to *X*. One causal interpretation is that each variable *X* corresponds to a time series X_t and if an equation such as *Y* = *a X* + ► occurs, then $Y_t = a X_t +$ ►$_t$ in the time series, with the ► variables all independently distributed. p. 500 An intervention on *X* then fixes (or randomizes) X_t at some arbitrary ↳ time t_0, and the effect on other variables is determined by the time series—results differ if *X* is held fixed throughout the resulting series, or merely ‘shocked’ at one time. Algorithms for computing the equilibrium effects using linear cyclic graphical models were developed in the engineering literature.

4. *Logistic regression models*. Suppose *Y* is a binary variable (say with values 0 and 1) and *X* a continuous variable. The *odds* that *Y* = 1 are Pr(*Y* = 1) / 1 – Pr(*Y* = 1) and the *log odds* or *logit* is the logarithm of that ratio. In logistic regression models, the logit of *Y* is set equal to a linear function of *X* plus noise, that is:

$$\log(\mathrm{pr}\,(\mathrm{Y}{=}1)\,/\mathrm{pr}\,(\mathrm{Y}{=}0)) = aX + \varepsilon \tag{2}$$

Versions of logistic regression are among the most widely used statistical models, especially in epidemiological contexts where causation is at issue; the model suggests that an intervention that changes *X* will change the *difference* (or the ratio) in the logs of the probabilities of the two values of *Y*. 5.*General Additive Models* allow for dependent variable *Y* to be any smooth additive function of other variables plus noise, e.g. $Y = a \ln(X^2) + b\, e^{cz} + ►$. Their causal interpretation is generally straightforward.

5. *Polynomial Regression Models* allow *Y* to depend on any polynomial function of other variables, plus noise. Again, the interpretation as causal models is straightforward.

6. *Time series models*. Consider an indexed set of vectors $\mathbf{V}_t$ with t ranging over the integers or the positive integers, and let there be a joint probability distribution that is stationary—the marginal distribution is the same for all times t. For such systems, and more specifically for linear systems in which the co-variances do not change with time, Granger (1969) proposed that the series X_t is a cause (now called a *Granger cause*) of the series Y_t with respect to the remaining variables provided the expected value of Y_t conditional on $\mathbf{V}_t \setminus \{Y_{t-}, X_{t-}\}$ is not equal to the expected value of Y_t conditional on $\mathbf{V}_t \setminus Y_{t-}$, where Y_{t-} and X_{t-} are all variables for *Y* and for *X*, respectively with indices smaller than t. The idea is a generalization of linear regression and related to Suppes (1970) more general proposal for understanding causal relations as conditional probability relations subject to a time constraint.
   Granger causation need not be causation if, for example, the difference in expected values is due to unrecorded common causes.

7. *Categorical variables* (Bishop, Fienberg, and Holland 1975) were the subject of various attempts over many years to provide a family of probability distributions and representations that could be estimated and could be naturally interpreted as specifying causal relations. One influential proposal is the log linear model, which specifies, not the dependence of variables on one another, but the joint probability of an assignment of values to categorical variables. With a system of variables of n × m values and a joint probability distribution one can associate an n × m table, with each cell of the table representing the probability that a case exhibits the combination of ↳ values in that cell. The log linear model treats the natural log of the cell probability as a linear function of an undetermined parameter raised to the power of each variable and each conjunction of variables. For example, for two binary variables, A, B, the natural log of the probability of a case lying in cell ij is $\ln(f_{ij}) = \lambda_i^{A} + \lambda_j^{B} + \lambda_{ij}^{AB}$. The parameter estimation problem is to determine the values of the λ variables. The log linear model has been given a causal interpretation for certain spatial statistics (Moore 2001) but has no evident general causal interpretation, although attempts have been made to give it one (Goodman 1978).

In its most general form, the causal Bayes net model for categorical variables assumes a multinomial distribution of the variables, and a directed acyclic graph of causal relations. The parameters of the distribution are simply the probabilities of each value of each variable conditional on each assignment of values to its parents—its direct causes—in the graph. Such models have the causal interpretations described in more detail below. More specialized parametric families of probability distributions for causal Bayes nets have also been used. For binary variables, a consistent parameterization (Pearl 1988; C. Glymour

2003) is obtained by treating each variable as a Boolean function of its parent variables, with each parent variable multiplied by a Boolean parameter, and taking probabilities of both sides.

3. Parameter Estimation

While in principle every method of searching for causal explanations based on sample data might be structured as a form of parameter estimation, the most common view of causal inference in statistics is that it involves estimating unknown values of numerical parameters that measure the strengths of potential causal relations that are themselves specified a-statistically—from prior theory or from experimental design. In the simple 'model'

$$crop\,yield/acre = \alpha\,tons\,of\,fertilizer/acre + \beta\,tons\,of\,water\,applied/acre + \chi + \varepsilon \tag{3}$$

the aim is to estimate the unspecified values of α, β, and χ under various assumptions, for example that εε is a normally distributed random variable and that values of α, β, and χ are the same for all cases ('fixed effects') or vary from case to case ('random effects'), etc. An estimator is simply a *statistic*: that is, a function from sample properties to some definite range of mathematical objects, in the present case to the real numbers that are the possible parameter values. In the simplest cases, estimates of the parameter values and assumptions made about the joint distributions of variables—in our example, εε, tons of fertilizer/acre and p. 502 tons of ↳ water applied/acre—determine a sampling distribution for a statistic—that is, for any given sample size, the probability distribution of values of the statistic among samples of that size. The oldest statistic of this kind is Legendre's least squares.

A large body of statistical results concerns which estimators meet various intuitively motivated criteria that can be assessed without knowledge of the true value of the parameter (Lehmann 1998). One criterion is 'consistency', which means, roughly, convergence of estimates to the true value of the quantity estimated. There are importantly different exact definitions. In all definitions below, α is a parameter or vector of parameters, *a* is its actual value, and 漾(α, N) denotes the value of the estimation function ω for α applied to a sample of size N and ϖ漾(α, *d*) the estimate of α from *d*, the true probability density, and Pr is a probability based on *d.*

(1) Fisher Consistency: Given the (true) density *d* for the observed variables, when the value of parameter α is *a*, 漾(α, *d*) = *a.*

(2) Pointwise Consistency:

$$\forall\delta > 0, \forall\varepsilon > 0, \forall\, a, \exists\, \mathrm{n} \,\forall\, \mathrm{N} > \mathrm{n}, [\mathrm{pr}\,(|\text{潛}(\alpha,\mathrm{N}) - a| > \delta) < \varepsilon]$$

(3) Uniform Consistency:

$$\forall\delta > 0,\ \forall\varepsilon > 0,\ \exists\ \mathrm{n} \,\forall\, \mathrm{N} > n, \forall\, a\,[\mathrm{pr}\,(|\text{潛}(\alpha,\mathrm{N}) - a \circ > \delta) < \varepsilon]\,.$$

All these consistency criteria have 'weak' versions that allow the estimator to be a partial function—that is, on some data the estimator can pass, and can continue to pass no matter how large the sample size, but must eventually provide an estimate as the sample size increases without bound. Analogous criteria apply as well to hypothesis testing and to procedures that search for graphical causal models, discussed below.

Uniform Consistency, but not Pointwise or Fisher Consistency, entails that confidence intervals for the estimates can be constructed that converge to zero width as the sample size increases without bound. For some statisticians, no estimation procedure is acceptable unless it satisfies Uniform Consistency, a requirement that sometimes excludes all possible estimators.

None of these consistency criteria suffice to distinguish among many possible estimators, and other desiderata are therefore imposed where they can be. For example, the mean squared error of estimates can be divided into a term representing the expected absolute value of the difference between the true value and the estimated values—the square of the bias—and a term representing the variance of the estimates (the variance of the estimates, that is, that would be obtained on samples of the given size obtained from the true p. 503 distribution.) A common ↳ requirement is that an estimator be unbiased and among unbiased estimators, have the minimum variance. A popular alternative, Fisher (1990), is maximum likelihood: that the estimated value be that for which, among the alternative possible values, the observed sample is the most probable. An enormous literature studies the applicability and interconnections of these and related criteria.

An increasingly popular alternative is Bayesian estimation of parameters, which, starting with a probability distribution over the values of the parameters, computes a probability distribution conditional on the observed sample (Lee 2004). Bayesian estimation, long merely a toy because of the difficulty of actually computing posterior distributions, has been made practical by simulation methods that allow such estimates for small samples (Casella and George 1992) and by an easily computed asymptotic formula, the Bayes Information Criterion (Schwarz 1978) that in many cases provides good approximations to the posterior probability for large samples. Disputes over the various consistency criteria above are of little relevance to Bayesians, who have weaker requirements, for example, that the set of values of the parameter for which the posterior probability converges, in the pointwise sense above, has probability 1, or that the expected error converges to 0.

Parameter estimation has an underdetermination problem. *Identifiability* fails when more than one assignment of parameter values determines the same marginal probability distribution over observed variables. Identifiability typically fails for parameters relating variables that are *confounded*, that is, jointly influenced by one or more common unobserved variables. When X is thought to be a confounded cause of Y, a standard solution is to find a *instrumental variable Z* that is thought to influence Y if at all, only through X; in some distributions this permits consistent estimation of the influence of X on Y, for example, of α in equation (3). The instrumental variables technique works only for special forms of dependency and probability distributions; while it holds for systems of binary variables parameterized as ‘noisy or gates’ (e.g. $\Pr(Y) = \Pr(\mathrm{a}\, X \oplus \mathrm{b}\, Z)$, where Y, X, Z, a, and b are Boolean and $\oplus$ is Boolean addition) (Glymour 2003), it fails, for example, for categorical variables distributed according to a multinomial distribution, although bounds on probabilities may be estimated (Galles and Pearl 1995).

Statistical literature and practice contain various other ad hoc or heuristic rules for avoiding confounding, in particular advice to condition on any variable found to be associated with both of two variables thought to be causally related, and to stop conditioning on new variables when the association under study does not change much. While widely used, this recommendation is not generally sound and can result in increased error compared to estimates with a smaller or larger set of co-variates.

Finally, some distributions and their parameters are regarded as ‘not causal’ for good reason. For example, in a linear system with normally distributed variables, the variables may be transformed, or ‘standardized’, by setting, for all variables X, $X_s = (X - \mu_X) / \sigma_X$, where μ_X is the mean of X in the sample and σ_X is the sample p. 504 ↳ standard deviation. The result is that linear coefficients become correlation coefficients, but do not predict the effect of an intervention that produces a unit change in a causal variable in any other sample governed by the same causal process but with different noise variances.

4. Hypothesis Testing

Parameter estimation is often an implicit step within testing a causal hypothesis expressed by an equation, such as (1), assuming a family of probability distributions characterized by values of the parameters in the equation. For example, from a maximum likelihood estimate of parameter values a sampling distribution of some statistic is obtained and the probability that the value of the statistic lies in some tail or tails of the distribution is computed, ideally in conjunction with the probability of the same tail membership of the statistic as a function of alternative values of the parameters—essentially, the *power function* of the test. Depending on the school of statistics, the use of the test is to reject the hypothesis, or not (Neyman–Pearson), or to regard the hypothesis as confirmed, or not (Fisher). There are as many tests of a model as there are relevant statistics, and a statistical standard is to search for ‘uniformly most powerful’ tests and recommend their use when they exist.

Philosophical commentators (Mayo1996) have emphasized that reliable inquiry requires some further criterion of severity of a test, or body of tests, in excluding alternatives, where the severity of a test of a hypothesis is, roughly, a function of the probability that the test would result in a worse fit to the sample data were the hypothesis false; Mayo proposes that experiment with hypothesis testing be extended to all of the assumptions of a ‘model’ such as (3), and to the sampling procedure, with the hope that unique explanatory features will eventually be identified. Assumptions about the distribution family may be tested, as may assumptions that the probability distribution for values of variables are the same for all sample units, and independent for each sample unit (i.i.d. sampling), and so on. A careful formulation of severity criteria and a more detailed discussion of foundational issues in hypothesis testing may be found in Mayo (1996). This perspective is illustrated vividly in work by Spanos (2007), who develops a heuristic search procedure based on a series of hypothesis tests for finding the appropriate family of probability distributions, and for detecting and correcting for statistical dependencies between units in a sample (usually called, rather misleadingly, autocorrelation).

p. 505

5. Estimating Interventions and Graphical Models

The graphical causal model framework exploits directed graph representations, representations that have a long history, reviewed in Pearl (2000). Variables are represented as nodes in a directed graph, and a directed edge, $X \rightarrow Y$, is the claim that X is a direct (relative to other variables represented in the graph) cause of Y. The diagrams are useful as a psychological aid, but so far as estimation is concerned, the key tools are consequences of ‘factorizations’ of the joint probability density on values of a set of variables. In graphs without cycles—paths of directed edges with the same orientation that begin and end with the same variable—the factorization is implied by the Markov assumption: each variable X, in a directed graph is independent in probability of variables in the graph that are not direct or indirect effects of X, conditional on the direct causes of X in the graph. Formulations of the Markov assumption (Kiiveri and Speed 1982) for causal systems emerged from studies in the late 1970s of factorizations of distributions. Let **V** be a set of variables V_i, the vertex set for a directed acyclic graph (DAG) G, and for each variable Vi in **V**, let **Par**(V_i) be the set of variables with edges in G directed into V_i. Let Pr be a probability distribution or density satisfying the Markov assumption for G. Then for all sets **v** consisting of one value, v_i, for each variable V_i in **V**,

$$\Pr(V = \nu)) = \Pi_i \Pr(Vi = \mathrm{v_i} \circ \mathrm{par}(V_i)) \tag{4}$$

The Markov factorization is a *necessary* consequence of either of the following:

a. the value of each variable is determined by the values of its parents and zero indegree variables are jointly independent;

b. the joint probability distribution is the marginal of a probability distribution satisfying the Markov assumption for a directed graph without cycles, zero indegree variables are jointly independent; and some (possibly empty) set of variables, each of which has a single direct effect in the graph and is the effect of no variable in the graph, is marginalized out.

The causal graphical framework aims (1) to enable the estimation of the probability of any represented variable conditional on values of any other variables represented; (2) to enable the estimation of the probability of any represented variable conditional on any hypothetical intervention that forces new probability distributions on other variables. Our concern here is with the second, causal, aim. The probabilities of variables in a graphical model can be interpreted either as actual or hypothetical population frequencies, or as propensities or chances attached to individuals, propensities that may differ from the value a variable actually has for the individual.

p. 506 Pearl, Geiger, and Verma (Pearl 1988) and Lauritzen and his collaborators (Lauritzen 1996), provided algorithms that decide whether the Markov assumption applied to a directed acyclic graph implies any particular conditional independence relation. Pearl's version has been more popular: V_1 and V_2 are *d-connected* conditional on set **Z** if and only if there is a sequence of edges between V_1 and V_2 such that every vertex touched by the sequence and having two of the sequence edges directed into it (every *collider* on the sequence) is in **Z** or is the source of a directed path leading to a member of **Z**, and no other vertex touched by the sequence is in **Z**. Vertices are *d-separated* with respect to Z if they have no d-connecting path with respect to **Z**. The Markov assumption applied to a directed acyclic graph implies that in *every* distribution satisfying the assumption for the graph, two vertices are independent conditional on a set of other variables if the vertices are d-separated conditional on the set. Many of the notions that are otherwise explained in terms of correlations of various kinds, or their absence, are explicable in terms of d-connection. For example *Z* is an *instrument* for the effect of *X* on *Y* provided (i) *Z*, *X*, and *Y* are not pairwise independent; (ii) there is no edge from *X* to *Z*; and (iii) every edge sequence that unconditionally d-connects *Z* and *Y* touches *X*. The d-separation relation also characterizes independence and conditional independence relations in systems in which the noise terms for some variables are specified to be correlated, without causal explanation, and it also necessarily characterizes those relations implied (for all non-zero values of linear coefficients) by linear systems represented as *cyclic* graphs under the two circumstances listed above as sufficient for the Markov condition.

Given an acyclic causal graph *G* = <**V**, **E**>, **V** a set of random variables and **E** a set of directed edges, and a probability distribution Pr, Markov for the graph, an intervention on *V* in **V** can be represented by an extension of *G* with a new variable I_V with at least two values and an edge directed into *V* and no edges into or out of I_V, and a probability distribution Pr*, Markov for the extended graph, such that Pr* conditional on one value i of I_V is equal to Pr, and conditional on at least one other value j of I_V is equal to Pr with a new factor Pr*(*V*) substituted for the original Pr(*V*) in the factorization. In general, if the factorization of **V** depends on **Par**(*V*), in the new distribution the factorization of **V** depends on **Par**(*V*) ∪ I_V The representation allows the computation, given the probability of *V* conditional on I_V = j, of the probability of any other variables produced by an intervention (Spirtes et al. 2001). For the special case of interventions that 'break' edges into *V*—i.e. that in the factorization of the original graph remove the dependence of *V* on other variables, Pearl (2000) provides an algorithm for the computation when the causal structure is known but may contain latent variables, extending an algorithm of Spirtes et al. (2001). Either procedure can be applied when causal and probabilistic input is incomplete, and Spirtes's form can do so even when the intervention alters, but does not remove, the conditional probability of *V* on **Par**(*V*). Spirtes's algorithm can give 'not computable' as an outcome, and it is known that the procedure is not p. 507 ↳ maximally informative, but Pearl's has been shown to be. Woodward (2003) has argued that the edge-breaking sense of intervention is fundamental to the notion of causation. The Markov condition and associated algorithms yield results about estimation of intervention effects that hold for *every* probability distribution Markov for *any* DAG. Further prediction results may hold for particular families of probability distributions; the consequences of restrictive distribution assumptions are not very well explored.

6. The Counterfactual Framework

A more influential framework in contemporary statistics (Rubin 1977; Robins 1986) analyses causal dependency as a counterfactual relation, roughly in the sense of Lewis (1973), although philosophical logicians seem not to have been read by the statistical community. The goal of inference is taken to be the effect of treatment assignment T = t on outcome O for unit u, defined to be the difference between the actual value of the outcome for u and the outcome u would have had under an alternative treatment assignment, T = t′. 'Treatment assignment' is sometimes a misnomer, since the counterfactual framework is meant to be applied to non-experimental data as well as to experimental, and merely means whichever variable of a pair is considered to be the potential cause. ('Treatment assignment' is distinguished from 'Treatment' in the epidemiological literature because patients do not always do as told.)

Since the alternative treatment is not given, the outcome that would have obtained had an individual been given an alternative treatment is not observed. The distinguishing idea of the counterfactual framework is to introduce for each outcome variable and each alternative treatment assignment a 'counterfactual variable' whose value is, for each individual in the sample, the value the outcome would have had for that individual if the alternative 'treatment' had been given. Contrasts between the outcome on one treatment assignment and on another treatment assignment then become a problem in estimating models with unobserved, counterfactual variables. This kind of estimation problem has no unique answer for the individual case. Estimation of average influences requires assumptions about the uniformity of dependencies across individuals (or the random distribution of dependencies independently of the variable values) and the absence of confounding variables influencing the putative cause and the putative effect. Statisticians using the framework tend to report 'propensity scores' showing how the contrast depends on the values of parameters in the model representing counterfactual dependencies and confounding influences.

p. 508 Models developed within the counterfactual framework implicitly use the Markov assumption, which suggests these models have graphical model equivalents that eschew counterfactual variables, as various authors have (e.g. Pearl 2000) have argued. There are differences. The counterfactual framework does not allow counterfactual variables that range over values the *treatment assignment* (or other upstream variable) would have had were the effects to have been different. Thus, unlike the graphical causal model framework, in order to specify the relevant model variables the counterfactual framework requires prior assumptions as to which variables are potential causes of which others. For this and other reasons, work in the counterfactual framework avoids automated search procedures.

7. Search for Causal Explanations

Experimental settings in which one or more variables are manipulated by the experimenter provide restrictions on plausible causal hypotheses. If the value of *X* in each case is determined by the experimenter, then values of *X* and of potential effects, *Y*, of *X* are not confounded by common causes, and *Y* is not an effect of *X*. The inquiry is reduced to estimating whether *X* has any effect on *Y*, and, if so, some measure of the strength of that effect. Early in the twentieth century, Fisher (1990) made popular designs that randomize the values assigned to variables under experimental control, both avoiding confounding and allowing statistical tests of the hypothesis of no effect, and estimates (as by ANOVA or regression) of the strengths of influences. Fisher also introduced strategies for making statistical inference more efficient—both in an informal and in a technical sense—by various sampling and control methods. For example, in estimating which of two kinds of shoes wear longest on boys, simple randomization would assign a pair of shoes of one or another kind at random to a representative sample of boys. But, on average, boys might wear out shoes on their right feet more quickly than on their left, and boys vary in how rapidly they wear out shoes. Estimates of the difference would be improved by randomly assigning, for each boy, one shoe of one type to one of the feet of a boy, and the other type to the other foot of the same boy and estimating the average across all boys of the individual differences in wear. Recent work in the graphical causal model tradition has shown that when the aim is to determine all causal relations among a set of variables, strategies that simultaneously and independently randomize multiple variables reduce exponentially the number of experiments required. Related results also suggest improved estimation of causal effects through such strategies (Eberhardt, Glymour, and Scheines, 2005).

Some Bayesian statisticians dispute the necessity of randomization, and the appropriateness of relying on randomization to remove confounding. A random sample may, by chance, be very unrepresentative of the population from which it is sampled.

Even with experimental manipulation, causal inference can be difficult. The treatment and its effects may, for example, cause some units in an experiment to drop out or not comply with the experimental design (e.g. cells die, mice die, people stop taking their drugs, people drop out of a long-term experiment), resulting in an unrepresentative ('biased') sample. Experiments may seek simultaneously to estimate the effects of a variable on multiple outcomes, but the several outcome variables may influence one another, or there may be unmeasured confounding variables that influence the outcomes and are not removed by the randomization—because the *outcome* variables are not randomized. In these respects, causal inference from experiments shares problems with causal inference without experimental controls.

Without experimental controls, search for causal relations from samples may be viewed essentially as a kind of estimation problem in which hypothesis testing may (or, as in Bayesian procedures, may not) be a tool or step. Causal estimation may be done in steps, first estimating the graphical causal structure and possibly the functional form of dependencies, then estimating parameters, or the functional form may be separately estimated (or assumed) and the estimation of parameters and graphical structure estimated simultaneously. The estimated causal structures are themselves hypotheses that can in most cases be subjected to statistical tests. The mathematical questions are of the same kind as in conventional statistical estimation: under what assumptions do which kinds of search procedures (i.e. estimators) have which kind of desirable statistical and computational properties connected with (probably) finding the truth?

Despite the parity of reasoning between estimation and search, a long and influential tradition in statistics has deprecated model search as 'fishing expeditions' or 'ransacking'. One intelligible thought behind the slogans was that using data to develop and test a statistical model would lead to 'overfitting', meaning that estimates of parameter values obtained from the data also used to obtain the model would in general not agree with estimates of the same parameters obtained from new data drawn from the same probability distribution (because the model obtained would sometimes be wrong, typically containing too many parameters). Statistical writers distinguished 'confirmatory' statistical analyses, usually meaning those that issued in a well-defined test of the hypothesis on data not 'used' in formulating the hypothesis, and those analyses resulting from data-driven search that had no such test. Bayesian statisticians have been especially concerned about 'double counting', that is, using the same datum in forming a prior probability distribution and in calculating a posterior distribution. These objections are now often addressed in practice p. 510 by holding out a sample of data for testing, or by ↳ repeating model search and parameter estimation with subsets of the original data and testing the model and estimates on the remainder of the original data (usually called *cross-validation*). Further, it was for a long time quite unclear what mathematical objects could represent causal relations, and without such objects model search could not be treated mathematically as estimation. The graphical representation of causal relations and formalization of the Markov condition have largely, but not entirely, resolved that problem. One could reasonably doubt that automated procedures could be devised that would substitute for knowledge of a domain and human consideration of the data. For example, the fact that test scores of students on an examination can be put in a series in which there are improbable sequences of similar scores would suggest nothing to a computer, but to a human, knowing that the series order is the seating order, it is evidence of cheating. The doubt is well-founded, but that does not mean that systematic, automated search procedures cannot help in discovery.

From early in the twentieth century, statisticians recognized that in many problems the number of potential alternative causal explanations for a body of data is infinite, or at least too large to survey explicitly, and that the fact that a particular model is not rejected by a test provides no guarantee that a particular alternative model is the true one, or at least relevantly closer to the truth. Further, the implicit statistical criterion of success involved succeeding—converging to the truth—regardless of what the truth might be, without sometimes having to say 'don't know'. This meant that causal parameters could never be pointwise estimated unless the graphical structure and parametric family were already known, and it was far from clear how such knowledge could be acquired systematically and reliably. There are a great many hazards to correct causal inference from non-experimental data, for example: (1) missing values of variables for cases; (2) unmeasured confounding variables; (3) measurement errors; (4) sample selection bias; (5) autocorrelation, in which values of variables for a sample unit influence values of variables in other sample units; (6) probability distributions and functional dependencies that are not among the familiar examples; (7) samples that are formed of sub-populations with distinct probability distributions and even distinct qualitative causal relations; (8) the data may be described best by a *cyclic* graph, and for reasons noted above, the causal content of such models is ambiguous and their discovery from data alone seemed implausible; (9) sometimes the causal relations of interest are among variables that are not measured, but whose effects or manifestations are measured, and it seemed implausible—some claimed impossible (Bartholomew and Knott 1999)—that data-driven methods could provide the information required.

The development of the formalism of graphical causal models prompted a burst of research beginning around 1990 on computerized search methods for which proofs of convergence to correct information could p. 511 be provided. The Markov assumption alone is insufficient for the existence of any consistent estimators of ↳ causal relations, and further assumptions are needed to form a subspace of possible models for which search is possible. One widely used assumption is *Faithfulness*: all conditional independence relations in the distribution are implied by the Markov assumption applied to a DAG. Faithfulness was later shown to hold probability 1 for DAGs with smooth measures on the parameters of linear models or the parameters of categorical variable models (Spirtes et al. 2001). Markov, Faithfulness, and samples that are independently and identically distributed have been proved to be sufficient for pointwise consistent inference to features of causal graphs in a variety of circumstances: (a) for acyclic causal structures, when there are no unrecorded confounding variables or 'correlated errors'; (b) for linear, cyclic causal structures when there are no unrecorded confounding variables or correlated errors; (c) for acyclic causal structures when there are unrecorded, and unknown (before data analysis), confounding variables or correlated errors; (d) for identifying sets of measured variables that share a single unmeasured common cause; and (e) for estimating features of the causal relations among the latent variables identified as in (d). The same algorithm that suffices for (c) is also pointwise consistent when there is sample selection bias. These procedures do not typically identify a unique directed graph of causal relations, but rather features (e.g. a directed edge, or a directed path, etc.) common to all members of a set of alternative graphs that might explain the data. More recent work has shown that unique DAGs for linearly related, non-normally distributed systems without latent variables can be consistently identified from i.i.d. data provided measured variables are not deterministic functions of one another. (Shimizu et al. 2006). Faithfulness is not required—effects due to different causal pathways can perfectly cancel and the causal structure can nonetheless be fully recovered. These methods have been combined with procedures for identifying latent variables to estimate a unique DAG among unrecorded common causes.

Linear autoregressive time series can be given the form $y_n = x_n + \Sigma_j a_j y_{n-j}$ where j ranges over some specified number of previous time steps, the a_j are real constants, and x_n is a 'random shock' at time step n. In the multivariate version each variable may depend on a linear combination of time delays of other variables. Moving average time series make y_n a linear function of past random shocks plus a current shock. When both sorts of dependencies are present, the system is called an autoregressive moving average, or ARMA model. Time series models have an autocorrelation between any two variables for any specified time difference, or lag—the correlation of y_n with x_{n-j} for some fixed j. Partial autocorrelation for a given time difference is autocorrelation conditional on values of the variables between the lags.

A standard search procedure for time series is owed to Box and Jenkins (1970). An empirical series is tested, and if necessary adjusted, for stationarity—the joint distribution of the variables at a time step must be independent of the time step, or, for Gaussian processes, the co-variance of variables must be independent of time. p. 512 ↳ Patterns of autocorrelation and partial autocorrelation are then used to determine which variables influence which others at which lags, and the values of parameters are estimated by standard statistical procedures. It can happen that two series are not stationary, but some linear combination of them is. If, of two such series, the two derivative series obtained by taking the difference of values between each time step are both stationary, then the two original series are said to be co-integrated. One analogy for co-integrated series is a man walking a dog on a leash. As the man pulls the dog and the dog pulls the man, each of their trajectories is irregular, but the average of their positions follows a much smoother trajectory.

A simplified procedure is sometimes used to find Granger causes: each pair of variables is regressed in each direction, controlling for other variables at a large number of lags; significant regressions are taken to indicate a causal relation. Many complexities arise in searching for multivariate causal time series: dependencies may be non-linear, distributions non-normal, stationarity may not hold, a series of unobserved common causes may exist, and, because many time series are in discrete steps, 'contemporaneous' causal processes may occur between the time steps. The last problem has been essentially solved for contemporaneous causes by first regressing each variable on all previous time steps of all variables, and then applying graphical search algorithms referenced above to the residuals (Demiralp and Hoover 2003; Moneta and Spirtes 2006). The procedure allows unobserved contemporaneous common causes. Time series problems remain very much open.

8. Causal Interpretation and Causal Puzzles

The statistical literature has developed various standard puzzles about probability and causality that sometimes take on a didactic role.

1. *Mistaken mechanisms*. The overall rate of acceptance of female applicants to graduate programmes at UC Berkeley was lower than the overall rate of acceptance of male applicants, prima facie evidence of

discrimination against women. But not really: within each department, women were accepted at the same rate as men, but women tended more often to apply to programmes for which admission was more competitive.

2. *Zero correlation*. The correlation between per capita foreign aid that nations receive and the proportion of a nation's population living on less than one dollar a day, is zero. It does not follow and is not true that increasing foreign aid to any particular nation would not decrease poverty in that nation. What does follow, or p. 513 ↳ at least is suggested, is that a marginal increase in the amount of foreign aid, if distributed among nations according to current mechanisms, would not decrease extreme poverty.

3. *Correlation and aggregation*. Suppose the RNA is extracted from each cluster of cells and the concentrations measured, for many cell clusters. Suppose the concentrations of two molecular species of RNA are strongly correlated, and remain correlated conditional on a third measured species correlated with the first two. Each RNA species can be mapped to a 'reading frame'—a gene fragment from which that RNA molecule is produced by transcription. Does the correlation mean that transcription of one of the genes influences transcription of the other, or that there is an unrecorded common cause? It does not, because the measurements are effectively of *sums* of concentrations over many cells. If the relations between transcription of one gene and transcripts of other genes is non-linear, as it is thought to be in some cases, then conditional independence relations among concentrations at the cellular level can become conditional *dependence* relations among aggregated concentrations.

4. *The Monty Hall Problem*. Monty Hall places a pile of money behind one of three doors. A contestant arrives and chooses a door. If the door the contestant chooses actually hides the money, Monty opens one of the other doors at random; otherwise Monty opens the door that the contestant did not choose and that does not hide the money. Seeing the open, empty door, the contestant now has the option of changing his selection or staying with his original choice. He wins the money if the door he finally chooses, whichever it is, hides the money. Should he switch doors or stand pat?
   In repeated plays, contestants will win twice as often if they switch doors. The argument is simple: there was a 2/3 chance the original choice was incorrect, and subsequently removing an alternative does not change that fact. A similar result would obtain with 100 doors: one would win 99 times out of a hundred by switching after 98 randomly chosen doors were opened. The connection with causality is as follows: Monty's choice of where to put the money and the contestant's original choice of doors are independent variables, each with three possible values. Each of these variables influences which door Monty opens, another variable with three values. Independent variables that mutually influence a third variable are (almost always, assuming faithfulness) dependent conditional on the value of the variable they both affect, and that holds in this case. Given the information about which door Monty opens, the contestant's original choice of doors provides information about where Monty put the money.

5. *Simpson's Paradox*. Simpson (1951) produced an imaginary case in which the story suggests that neither *X* nor *Y* influence *Z*, *X* and *Y* are independent, but *X* and *Y* are dependent conditional on *Z*. His imaginary example produced a considerable statistical (and, eventually, philosophical) literature on reversals of association by conditioning on other variables. From the point of view of graphical models, p. 514 ↳ Simpson's example is a contrived story with an unfaithful distribution, but the changes in association, including reversals of sign, resulting from conditioning is a fundamental problem that besets causal inference from regression.

6. *Lindley and Novick's Puzzl e* . Suppose we have the data for variable *Y*, with values y1 and y2, and variable *X* with values x1 and x2, for two groups, one with value z1 for variable Z and the other with value z2 for Z (Fig. 23. 1).

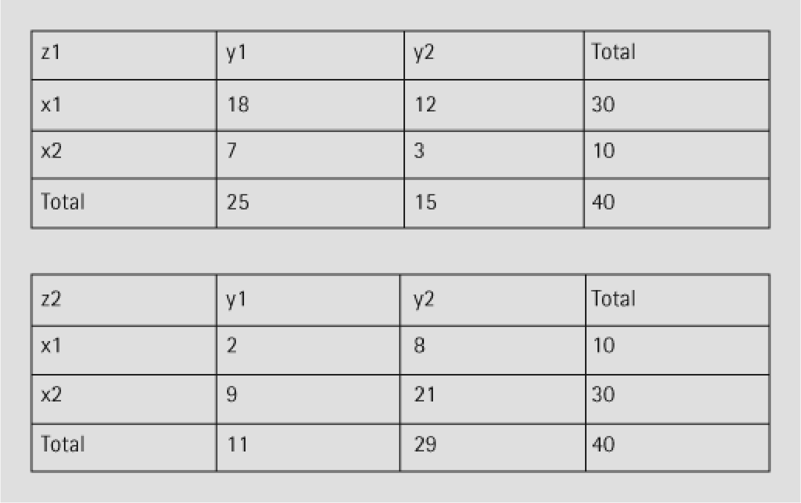
> Taking the sample as representative of a joint probability distribution, none of the variables are independent of any others. There is, however, something odd about the conditional probabilities: p(y1|x2, z1) > p(y1|x1, z1), and p(y1|x2, z2)> p(y1|x1, z2), but p(y1|x2) < p(y1|x1). If we condition on a value of *Z*, then no matter which value of *Z* we condition on, also conditioning on x2 gives y1 a higher probability than does also conditioning on x1. But if we do *not* condition on any value of *Z*, conditioning on x1 gives y1 a higher probability than does conditioning on x2. No pair of the variables is independent conditional on the third.

Lindley and Novick (1981) pose two different ways these data could have been generated:

1. A medical experiment: X = treatment; x1 = treated; x2 = not treated; Y = outcome; y1 = recovered; y2 = did not recover; Z = sex; z1 = male; z2 = female.

2. An agricultural experiment: X = variety of plant; x1 = white, x2 = black; Y = yield; y1 = high, y2 = low; Z = height of plant; z1 = tall; z2 = short. They raise these questions: if you want to produce the best medical effect, to whom if anyone should the treatment be given? If you want to produce the best yield, p. 515 ↳ what variety of plant should be planted? They answer: for the medical case, no treatment should be given, that is, x2 is the 'non-treatment' of choice, because it has better recovery probabilities for males and better recovery probabilities for females. In the agricultural experiment, white plants should be grown (i.e. x1) because, overall, it has a better probability of high yield.

Fig. 23.1



z1	y1	y2	Total
x1	18	12	30
x2	7	3	10
Total	25	15	40

z2	y1	y2	Total
x1	2	8	10
x2	9	21	30
Total	11	29	40

Their explanation is that under the second interpretation, but not the first, the cases in the data are 'exchangeable'. The idea of an exchangeable probability distribution is simply that for any finite ordered sample, the probability of obtaining that sample, given the sample size, is the same as the probability of obtaining a sample of the same size with any permutation of the given ordering. The exchangeability explanation is mysterious.

The two different stories, one medical and the other agricultural, naturally lead to different causal interpretations of the data, and the different causal interpretations suggest different effects of interventions (Meek and Glymour 1994; Pearl 2000). In the medical case the task is to estimate the relative effects of treatment versus no treatment in the experiment, and then use that information to recommend how the general population should be treated, or not treated. Assume that in the experiment someone's sex is not caused by the medical treatment. Sex and treatment, *Z* and *X*, are not independent in the data—more men received the treatment than did women. The dependency must then either be due to chance in selection of subjects, or due to the influence of sex on which patients were treated and which were untreated, or due to the influence of something unknown on both sex and treatment. The causal structure of the experiment in the two alternative later cases is one of those shown in Figs. 23.2 and 23.3.

The case of Fig. 23.3 is implausible if we think of sex as the *biological condition* of a person, but not if we think of sex as the biological condition of a *subject selected for the experiment*. In either case we know from the Markov Assumption how to compute the probability of recovery, y1, from an intervention—a forced value of x1 or of x2—in a system with this description. We eliminate the association between treatment choice and outcome due to the common cause (sex in Fig. 23.2; unknown in Fig. 23.3), and use the remaining association as our estimate of the effect of treatment choices on recovery. We can do that by conditioning on p. 516 Z, on the sex ↳ of the subjects. We compute the probability of y1 conditional on x1, for example, and conditional on Z. The probability of y1 conditional on x1, or respectively on x2, is of course different for different values of Z, for males or females, but x2 is better in both cases.

Fig. 23.2

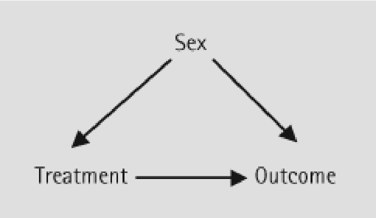


Fig. 23.3

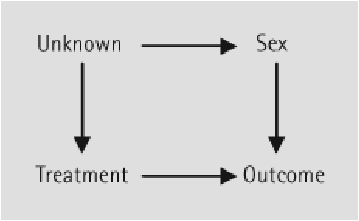


Fig. 23.4

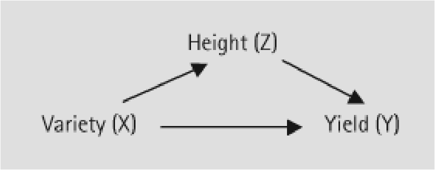
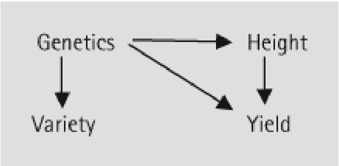


Fig. 23.5



In the agricultural interpretation of the experiment, heights of plants are correlated with variety of plant. That is most plausibly because the plant's variety influences its height, or something else—genetics, say—influences both the variety and the height of the plant, but the height of the plant doesn't influence the variety. (Both mechanisms are possible, of course.)

In recommending a planting policy, we know the variety of seed to be planted, black or white, but we do not know when we plant whether the plant will be short or tall. If in fact the plant variety influences height, as in Fig. 23.4, then the variety influences the yield through two mechanisms, one direct, and the other through height. In that case, to assess the influence of variety on yield, we *should not* condition on height. To do so would be to discount one of the paths by which variety influences yield. So we find Lindley and Novick's conclusion. Various complications arise if we think of genetics as a common cause of variety and height, as in Fig. 23.5.

A variety of other difficulties in the causal analysis of epidemiological data are clearly explained from the perspective of graphical causal models in M. M. Glymour (2006).

Further Reading

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer.
Google Scholar Google Preview WorldCat COPAC

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysi*s. Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

Fienberg, S. (1975). *The Analysis of Cross-Classified Categorical Data*. Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

Fisher R. A. (1990). *Annual Proceedings of the Conference on Uncertainty in Artificial Intelligence. The Journal of Machine Learning Research*.

Glymour, C. and Cooper G. (eds.), *Computation, Causation and Discovery*, Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

p. 517 Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

Pearl J. (2002). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge: Cambridge University Press.
Google Scholar Google Preview WorldCat COPAC

References

Bartholomew, D., and Knott, M. (1999). *Latent Variable Models and Factor Analysis*. 2nd edn. London: Edward Arnold
Google Scholar Google Preview WorldCat COPAC

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysi*s. Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

Blalock, H. (1961). *Causal Inferences in Nonexperimental Research*. New York: W. W. Norton.
Google Scholar Google Preview WorldCat COPAC

Box, G. E. P., and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
Google Scholar Google Preview WorldCat COPAC

Casella, G., and George, Edward I. (1992). 'Explaining the Gibbs Sampler'. *The American Statistician* 46: 167–74. 10.2307/2685208
Google Scholar WorldCat Crossref

Chickering, D., and Meek, C. (2002). 'Finding Optimal Bayesian Networks', *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002*). San Francisco: Morgan Kaufman, 94–102.

Demiralp, S., and Hoover, K. (2003). 'Searching for the Causal Structure of a Vector Autoregression', *Oxford Bulletin of Economics* 65: 745–67. 10.1046/j.0305-9049.2003.00087.x
Google Scholar WorldCat Crossref

Eberhardt, F., Glymour, C., and Scheines, R. (2005). 'On the Number of Experiments Sufficient and in the Worst Case Necessary to Identify All Causal Relations Among N Variables', in F. Bacchus and T. Jaakkola (eds.), *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*. Arlington, Va.: AUAI Press, 178–84.
Google Scholar Google Preview WorldCat COPAC

Fisher, R. A. (1990). *Statistical Methods, Experimental Design, and Scientific Inference*. New York: Oxford.
Google Scholar Google Preview WorldCat COPAC

Friedman, N. (1997). 'Learning Bayesian Networks in the Presence of Missing Values and Hidden Variables', in Fourteenth International Conference on Machine Learning.

Galles, D., and Pearl, J. (1995) 'Testing Identifiability of Causal Effects', in P. Besnard and S. Hands (eds.), *Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann, xi. 185–95.
Google Scholar Google Preview WorldCat COPAC

Geiger, D., and Meek, C. (1999). 'On Solving Statistical Problems with Quantifier Elimination', *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*. San Francisco: Morgan Kaufmann, 216–25.
Google Scholar Google Preview WorldCat COPAC

Glymour, C. (2003). *The Mind's Arrows*. Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

—— and Cooper, G. (eds.) (1999). *Computation, Causation and Discovery*. Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

Glymour, M. M. (2006) 'Using Causal Diagrams to Understand Common Problems in Social Epidemiology', in J. M. Oakes and J. S. Kaufman (eds.), *Methods in Social Epidemiology: Research Design and Methods*. San Francisco: Jossey-Bass.
Google Scholar Google Preview WorldCat COPAC

Goodman, L. (1978) *Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent Structure Analysis.* Cambridge, Mass.:

Abt Books.
Google Scholar Google Preview WorldCat COPAC

Granger, C. (1969). ‘Investigating Causal Relations by Econometric Models and Cross-spectral Methods’, *Econometrica* 37: 424–38. 10.2307/1912791
Google Scholar WorldCat Crossref

Greenland, S. (2003). ‘Quantifying Biases in Causal Models: Classical Confounding versus Collider-Stratification Bias’, *Epidemiology* 14: 300–6. 10.1097/00001648-200305000-00009
Google Scholar WorldCat Crossref

Heckerman, D. (1998). ‘A Tutorial on Learning with Bayesian Networks’, in M. Jordan (ed.), *Learning in Graphical Models*. Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

Hoover, K. (2001). *Causality in Macroeconomics*. New York: Cambridge University Press.
Google Scholar Google Preview WorldCat COPAC

—— (2005). ‘Automatic Inference of the Contemporaneous Causal Order of a System of Equations’, *Econometric Theory* 21: 69–77.
WorldCat

Huang Y., and Valtorta, M. (2006). ‘Pearl's Calculus of Intervention is Complete’, *Proceedings of the 2006 Conference on Uncertainty in Artificial Intelligence*.

Kiiveri, H., and Speed, T. (1982). ‘Structural Analysis of Multivariate Data: A Review’, in S. Leinhardt (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.
Google Scholar Google Preview WorldCat COPAC

—— —— and Carlin, J. (1984). ‘Recursive Causal Models’, *Journal of the Australian Mathematical Society* 36: 30–52. 10.1017/S1446788700027312
WorldCat Crossref

Koster, J., (1995). ‘Markov Properties of Non-Recursive Causal Models’, *Annals of Statistics* 24: 2148–77. 10.1214/aos/1069362315
Google Scholar WorldCat Crossref

Lauritzen, S. (1996). *Graphical Models*. Oxford: Oxford University Press.
Google Scholar Google Preview WorldCat COPAC

—— (2001). ‘Causal Inference from Graphical Models’, in O. Barnsdorff-Nielsen, D. Cox, and Kluppenlberg, C. (eds.), *Complex Stochastic Systems*. London: Chapman & Hall, 3–107.
Google Scholar Google Preview WorldCat COPAC

—— and Richardson, (2002). ‘Chain Graph Models and their Causal Interpretations (with Discussion)’, *Journal of the Royal Statistical Society, Series B* 64: 321–61. 10.1111/1467-9868.00340
Google Scholar WorldCat Crossref

Lee, P. (2004). *Bayesian Statistics: An Introduction*. New York: Wiley.
Google Scholar Google Preview WorldCat COPAC

Lehmann, E. (1998). *Theory of Point Estimation*. New York: Springer.
Google Scholar Google Preview WorldCat COPAC

Lewis, David (1973). *Counterfactuals*. Oxford: Blackwell.
Google Scholar Google Preview WorldCat COPAC

p. 518 —— (1973*b*). ‘Causality’, *The Journal of Philosophy* 70: 556–67.
WorldCat

Lindley, D., and Novick, M. (1981). ‘The Role of Exchangeability in Inference’, *Annals of Statistics* 9: 45–58. 10.1214/aos/1176345331
Google Scholar WorldCat Crossref

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
Google Scholar Google Preview WorldCat COPAC

Meek, C. (1995). ‘Causal Inference and Causal Explanation with Background Knowledge’, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. Philippe Besnard and Steve Hanks. San Mateo, Calif.: Morgan Kaufmann, 403–10.
Google Scholar Google Preview WorldCat COPAC

—— and Glymour, C. (1994). ‘Conditioning and Intervening’, *British Journal for the Philosophy of Science* 45: 1001–21. 10.1093/bjps/45.4.1001
Google Scholar WorldCat Crossref

Moneta, A., and Spirtes, P. (2006). ‘*Graphical Models for Identification of Causal Structures in Multivariate Time Series*’, in *Joint Conference on Information Sciences Proceedings*. Paris: Atlantis Press.
Google Scholar Google Preview WorldCat COPAC

Moore, M. (2001). *Spatial Statistics*. New York: Springer.
Google Scholar Google Preview WorldCat COPAC

Neal, R. (2000). ‘On Deducing Conditional Independence from *d*-Separation in Causal Graphs with Feedback’, *Journal of Artificial Intelligence Research* 12: 87–91.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, Calif.: Morgan Kaufmann.
Google Scholar Google Preview WorldCat COPAC

—— (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.
Google Scholar Google Preview WorldCat COPAC

—— and Dechter, R. (1996). ‘Identifying Independencies in Causal Graphs with Feedback’, *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, ed. Eric Horvitz and Finn Verner Jensen. San Mateo, Calif.: Morgan Kaufmann, 420–6.
Google Scholar Google Preview WorldCat COPAC

—— and Robins, J. (1995). ‘Probabilistic Evaluation of Sequential Plans from Causal Models with Hidden Variables’, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. Philippe Besnard and Steve Hanks. San Mateo, Calif.: Morgan Kaufmann, 444–53.
Google Scholar Google Preview WorldCat COPAC

Ramsey, J. Zhang, J., and Spirtes, P. (2006). ‘Adjacency Faithfulness and Conservative Causal Inference’, *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. Arlington, Va.: AUAI Press.
Google Scholar Google Preview WorldCat COPAC

Richardson, T. (1996). ‘A Polynomial-Time Algorithm for Deciding Equivalence of Directed Cyclic Graphical Models’, in *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*’, ed. Eric Horvitz and Finn Verner Jensen. San Mateo, Calif.: Morgan Kaufmann, 462–9.
Google Scholar Google Preview WorldCat COPAC

Robins, J. (1986). ‘A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect’, *Mathematical Modeling* 7: 1393–512. 10.1016/0270-0255(86)90088-6

Google Scholar WorldCat Crossref

—— Scheines, R., Spirtes, P., and Wasserman, L. (2003). 'Uniform Consistency in Causal Inference', *Biometrika* 90: 491–515. 10.1093/biomet/90.3.491
Google Scholar WorldCat Crossref

Rubin, D. (1977). 'Assignment to Treatment Group on the Basis of a Covariate', *Journal of Educational Statistics* 2: 1–26. 10.2307/1164933
Google Scholar WorldCat Crossref

Schwarz, G. (1978). 'Estimating the Dimension of a Model', *Annals of Statistics* 6: 461–4. 10.1214/aos/1176344136
Google Scholar WorldCat Crossref

SMIMIZU, S. HOYER, P. MYARINEN, and KERMINEN, A (2006). 'A Linear Non-Gaussian Acyclic Model for Causal Discovery', *Journal of Machine Learning Research* 7: 2003–30.
WorldCat

Shipley, W. (2000). *Cause and Correlation in Biology*. Cambridge: Cambridge University Press.
Google Scholar Google Preview WorldCat COPAC

Silva, R., Scheines R., Glymour, C., and Spirtes, P. (2006). 'Learning the Structure of Linear Latent Variable Models', *Journal of Machine Learning Research* 7: 191–246.
Google Scholar WorldCat

Simpson, E. (1951). 'The Interpretation of Interaction in Contingency Tables', *Journal of the Royal Statistical Society, Series B* 13: 248–51.
Google Scholar WorldCat

Spanos, A. (2007). 'Curve Fitting, the Reliability of Inductive Inference and the Error- Statistical Approach', *Philosophy of Science* 74: 1046–66. 10.1086/525643
Google Scholar WorldCat Crossref

p. 519 Spirtes, P. (1995). 'Directed Cyclic Graphical Representation of Feedback Models', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. Philippe Besnard and Steve Hanks. San Mateo, Calif.: Morgan Kaufmann, 491–8.
Google Scholar Google Preview WorldCat COPAC

—— and Richardson, T. (1997). 'A Polynomial Time Algorithm for Determining DAG Equivalence in the Presence of Latent Variables and Selection Bias', *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*.

—— Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. Cambridge, Mass.: MIT.
Google Scholar Google Preview WorldCat COPAC

Strotz, R., and Wold, H. (1960). 'Recursive versus Nonrecursive Systems: An Attempt at Synthesis', *Econometrica* 28: 417–27. 10.2307/1907731
Google Scholar WorldCat Crossref

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Acta Philosophica Fennica 24. Amsterdam: North-Holland.
Google Scholar Google Preview WorldCat COPAC

Swanson, N., and Granger, C. (1997). 'Impulse Response Function Based on a Causal Approach to Residual Orthogonalization in Vector Autoregression', *Journal of the American Statistical Association* 92/437: 357–67. 10.2307/2291481
Google Scholar WorldCat Crossref

Van Der Laan, M., and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
Google Scholar Google Preview WorldCat COPAC

Verma, T., and Pearl, J. (1990). ‘Equivalence and Synthesis of Causal Models’, in *Proceedings of the Sixth Conference on Uncertainty in AI*. Mountain View, Calif.: Association for Uncertainty in AI.
Google Scholar Google Preview WorldCat COPAC

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.
Google Scholar Google Preview WorldCat COPAC

Woodward, J. (2003). *Making Things Happen*. Oxford: Oxford University Press.
Google Scholar Google Preview WorldCat COPAC

Wright, S. (1934). The Method of Path Coefficients, *Annals of Mathematical Statistics* 5: 161–215. 10.1214/aoms/1177732676
Google Scholar WorldCat Crossref

Zhang, J., and Spirtes, P. (2003). ‘Strong Faithfulness and Uniform Consistency in Causal Inference’, *Proceedings of the 19th Conference on Uncertainty and Artificial Intelligence*, ed. C. Meek and U. Kjorulff, 632–9. ↳ p. 520

p. 521

↳