Brit. J. Phil. Sci. **54** (2003), 527–551, axg401

Nonstationary Time Series, Cointegration, and the Principle of the Common Cause

Kevin D. Hoover

ABSTRACT

Elliot Sober ([2001]) forcefully restates his well-known counterexample to Reichenbach's principle of the common cause: bread prices in Britain and sea levels in Venice both rise over time and are, therefore, correlated; yet they are *ex hypothesi* not causally connected, which violates the principle of the common cause. The counterexample employs nonstationary data—i.e., data with time-dependent population moments. Common measures of statistical association do not generally reflect probabilistic dependence among nonstationary data. I demonstrate the inadequacy of the counterexample and of some previous responses to it, as well as illustrating more appropriate measures of probabilistic dependence in the nonstationary case.

1 *A challenge to the principle of the common cause*
2 *Sober's argument and the attempts to rescue the principle*
3 *Probabilistic dependence*
4 *Nonstationary time series*
5 *Probabilistic dependence in nonstationary time series*
6 *Do Venetian sea levels and British bread prices violate the principle of the common cause?*

1 A challenge to the principle of the common cause

Hans Reichenbach's ([1956]) principle of the common cause and the causal Markov condition stand at the core of several modern accounts of causality (e.g., Spirtes, Glymour and Scheines [1993]; Hausman and Woodward [1999]). Reichenbach ([1956], p. 156) states the principle of the common cause: '*If an improbable coincidence has occurred, there must exist a common cause*' (emphasis in the original). He then goes on to elaborate the conditions for a common cause in terms of probabilities (pp. 156–67).[1] Elliot Sober

[1] The causal Markov condition states that any variable, V, in a causal graph, conditional on its parents, is independent of all other variables that are neither its parents nor its descendants (Spirtes, Glymour and Scheines [1993], p. 54; Hoover [2001], p. 157).

(2001, p. 331) restates the principle as:

> **(P)** If events X and Y are correlated, then either X caused Y, Y caused X, or X and Y are joint effects of a common cause (one that renders X and Y conditionally probabilistically independent).

Sober ([1994], pp. 161–2) had earlier challenged the principle of the common cause with a counterexample. In Sober's scenario, bread prices rise monotonically through time in Great Britain, and sea levels rise monotonically in Venice. Each process is causally independent of the other by assumption, yet, he asserts, the series are highly correlated. Sober regarded this as a violation of the principle of the common cause, and a demonstration that the principle fails in its key role in accounts of causality. Sober ([2001]) reiterates the point and offers an analysis aimed at showing that various attempts to rescue the principle of the common cause from his counterexample fail.

The kind of time series that Sober employs in his counterexample is commonplace in macroeconomics. Prices, gross domestic product, consumption, investment, employment, and wages—to name just a few prominent economic time series—trend up in the manner of Sober's bread prices and sea levels. In the past quarter century, statisticians (most notably time-series econometricians) have developed special tools for the analysis of these 'nonstationary' time series. One well-known exposition of some of the issues in time-series econometrics uses an example analogous to Sober's example, in which the level of the consumer price index plays the role of bread prices, and cumulative rainfall in the United Kingdom plays that of sea levels (Hendry [1980]). There are several reasons to believe that the principle of the common cause is not a successful foundation for causal analysis (see, e.g., Cartwright [1999], Ch. 5, or Hoover [2001], Ch. 4, §3). Nevertheless, the recent work in time-series statistics suggests that Sober's counterexample to the principle of the common cause is defective. If the principle is to be rejected, it must be on some other grounds.

2 Sober's argument and the attempts to rescue the principle

The principle of the common cause states that any correlation demonstrates a causal connection between the variables displaying the correlation—either direct or through a third cause. If the connection is through the third cause, then the third cause will screen off the correlation in the sense that the correlation of X and Y *conditional* on Z (the third or common cause) will be zero. Sober's counterexample is, then, simple: Venetian sea levels and British bread prices are truly correlated and are not causally connected by construction; therefore, neither causes the other and there can be no common cause. The assumption that sea levels and bread prices are truly correlated is central to

Sober's argument, and I will challenge it presently. For the moment, however, notice that the evidence he offers is crude but commonsensical: 'higher than average sea levels tend to be associated with higher than average bread prices' (Sober [2001], p. 332). This evidence is born out in a set of artificial data that Sober created to illustrate his counterexample.[2] The familiar correlation statistic (i.e., Pearson's correlation coefficient) for these data is 0.99, a number typically taken to indicate that the two series are, as Sober concluded on other grounds, 'very strongly correlated' (Sober [2001], p. 332).

Sober considers various attempts to defuse his counterexample, rightly rejecting some as beside the point. There are two that require further scrutiny. First, Sober ([2001], pp. 332–3) cites an argument that predates his article, traceable to Yule ([1926]) and revived in Meek and Glymour ([1994]), to the effect that the counterexample mixes different causal structures and different probability distributions. On this argument, his counterexample would be closely related to Simpson's paradox. Sober rejects this on the grounds that it is question-begging: 'If we only knew the true causal relationships, we would not need principles like (P) to tell us how to infer causal relationships from probabilities' (Sober [2001], p. 333). Sober would be completely correct if the argument required prior causal knowledge in order to dismiss the counterexample. Yet, as we shall see in Section 4 below, there is an important sense in which the probabilities in the counterexample are in fact not homogeneous through time, and there is nothing question-begging about inferring the absence of homogeneity from the data. In particular, there is no appeal to prior knowledge of causal structure.

The second attempt to defuse Sober's counterexample notices that, while the *levels* of bread prices and the sea may be highly correlated, *changes* in the levels need not be correlated (Forster [1988], Papineau [1992], and Hausman and Woodward [1999]). The situation in which the levels are correlated but the changes are not is taken to be a hallmark of the lack of causal connection. Sober argues that this observation does not save the principle of the common cause as stated in (P); rather it shows that it is false. The bread prices/sea levels example, he believes, would not defeat a weaker version of the principle:

> **(P*)** If events X and Y are correlated *and so are changes in X and changes in Y*, then either X caused Y, Y caused X, or X and Y are joint effects of a common cause (one that renders X and Y conditionally probabilistically independent). (Sober [2001], p. 335; emphasis added)

[2] The data given in Sober ([2001], p. 334) are:

Period	1	2	3	4	5	6	7	8
Bread Prices	4	5	6	10	14	15	19	20
Sea Levels	22	23	24	25	28	29	30	31

He then constructs further counterexamples, in which two causally unrelated processes are nevertheless correlated in both levels and changes and so violate (P*).

Hoover ([2001], pp. 164–6) also argues against shifting the focus to changes on different grounds. The critics of Sober's counterexample worry that the correlation between the levels is causally spurious, and suggest that a true causal relationship must also show up in changes. I argue that this may throw the baby out with the bath water: the application of (P*) may overlook a true causal connection between the levels that does not show up as a correlation between the changes.

Sober is correct that, take as many differences as one likes, there is some way to construct a counterexample analogous to the case of bread prices and sea levels, so that there is no hope in trying to save the principle by weakening it further. Instead, Sober's counterexample is defective at the top level: first, because probabilities are nonhomogeneous in a relevant sense; and, second, because the correlation between levels is not a *true* correlation in the relevant sense. The next three sections try to make sense of these claims.

3 Probabilistic dependence

As Sober ([2001], p. 343) notes, there are two steps in applying the principle of the common cause. First, correlations are inferred from associations—that is, from actual frequencies in the data. Second, causes are inferred from correlations—that is, from features of the underlying probabilities. The principle of the common cause relates only to the second step. Sober writes as if the first step were unproblematic for purposes of his argument. My central claim is that this assumption is wrong: associations of the bread prices/sea levels type do not, in general, reflect genuine probabilistic dependence and do not imply true correlations in the sense implicit in the principle of the common cause.

Association is not, in general, correlation (as is clearly understood by both Reichenbach [1956], p. 157, and Sober [2001], pp. 333, 343). That finite draws from probabilistically unconnected distributions might nevertheless be highly correlated is one of the standard difficulties of practical statistics. Calculating how the expected frequency of such spurious correlations changes as sample sizes increase is a key aim of statistical theory. Such small-sample problems to one side, correlations are inferred, not read directly from data (facts of association). What is inferred is a probability model.[3] The pattern of inference that forms the basis for statistics is: the observed facts of association would be

[3] Sober ([2001], p. 343) refers to 'causal models', but does not explicitly refer to probability models, despite his recognition that correlations are inferred from observed associations.

highly probable if the data were realizations from a particular model of the probability-generating process; therefore, that model is supported by the data.

Although Sober, like other philosophers and applied researchers uses the term 'correlation' as if its meaning were obvious, it is rarely clearly defined. One standard dictionary of statistics defines *correlation* as '[a] general term for interdependence between pairs of variables. See also **association**' (Everitt [1998], p. 80; bold in original). The same dictionary defines *association* as '[a] general term used to describe the relationship between two variables. Essentially synonymous with correlation' (Everitt [1998], p. 17). Standard textbooks generally do not define correlation or association at all, but leave the notions *implicit* in defining *measures* of correlation or association. For example, the *correlation coefficient* is defined as:

$$\rho = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \frac{E[(x - \mu_x)(y - \mu_y)]}{[E(x - \mu_x)^2(y - \mu_y)^2]^{1/2}}$$

where cov(..., ...) is the *covariance*, defined by the numerator of the right-hand term; var(..., ...) is the *variance*, defined by the denominator; $E(\ldots)$ is the *mathematical expectations operator*; and μ_j is the *mean* of j ($j = x$ or y) (Lindgren [1976], p. 135; Mood, Graybill and Boes [1974], p. 155).

Statisticians, however, clearly distinguish properties of the sample (the observed data) and properties of the population (the unobserved process that generates the data). It is useful to distinguish between *probability* (a population property) and *frequency* (a sample property). Reichenbach's ([1956]) statement of the principle of the common cause refers to facts of probability. The reference to correlation in Sober's principle (P) also refers to probability. I shall, therefore, draw a parallel distinction to that between probability and frequency and not treat 'correlation' and 'association' as synonyms. Instead, I shall use *correlation* to refer to a population property and *association* to a sample property. Probabilities are inferred from frequencies, and correlations are inferred from associations. A central message of this essay is that these inferences are not always obvious or straightforward. A co-occurrence may be a mere coincidence and, therefore, not a proper antecedent to the principle of the common cause.[4]

Hacking ([1965]), Mellor ([1971]) and Cartwright ([1999], Ch. 7) argue that probabilities are not robust facts about the world, always there for the asking. Rather they arise only in special arrangements of parts of the world, in set-ups that have a propensity to display probabilistic behavior. Cartwright goes too far, perhaps, in arguing that useful probabilities are almost always generated in artificial set-ups (her term is 'nomological machines') and only rarely in

[4] Reichenbach ([1956], p. 157) writes: 'Chance coincidences, of course, are not impossible [...] the existence of a common cause is therefore in such cases not absolutely certain, but only probable. This probability is greatly increased if coincidences occur repeatedly.'

nature. But the key point remains: probabilities are not observable, unmediated facts. We can reasonably speak of probabilities only in the context of particular models of the probability-generating process. Observable frequencies and facts of association are, at best, evidence that support inference to particular probability models.

Again, the typical pattern of inference runs: the observed facts of association would be highly probable *if* a certain probability model governed the generation of those facts. Statistics is largely about getting the 'right' model for the facts of association. The possible models vary *inter alia* in their degree of specificity (parametric, semi-parametric, nonparametric), in the degree of accuracy or approximation claimed for them, and in whether they deal with discrete or continuous variables (or even some complex combinations).

Although these details are beyond our present purpose, it is useful to note that most statistical inference and most of our own probabilistic intuitions are based on stationary probability distributions. There are several concepts of stationarity, all of which have in common the idea that the probability distribution is unaffected by the passage of time.

> **Definition:** A *time series* (or *time-series process* or *stochastic process*) is a random variable (or vector of variables) whose realizations are time ordered. (Hamilton [1994], p. 43)

For example, $\{X_t\}_{t=-j}^{k} = \{X_{-j}, X_{-j+1}, \ldots X_{-1}, X_0, X_1, \ldots X_{k-1}, X_k\}$. Gross domestic product (*GDP*) provides a concrete economic illustration of a time series: for example, $GDP_{1999} = \$9{,}301\text{b}$, $GDP_{2000} = \$9{,}704\text{b}$, $GDP_{2001} = \$10{,}239\text{b}$.

> **Definition:** A time series is *weakly* (or *covariance*) *stationary* if, and only if, its mean and variance are both finite and independent of time, and the covariance between the values of the series at different times depends only on the temporal distance between them. (Hendry [1995], p. 42).

The last condition means that if, for instance, rainfall is weakly stationary, the covariance between rainfall in year t and rainfall in year $t-2$ depends only on the fact that the observations are two years apart and not on the fact that $t = 1983$ or $t = 2001$.[5] The *normal* or *Gaussian* distribution is a paradigm of continuous stationary distributions. Many powerful results in statistics (e.g., laws of large numbers or central-limit theorems) require assumptions that means, variances, or higher moments are constant or, at least, finite.

To be absolutely clear—though at some risk of being tedious—let us examine such inference in a textbook case. I take the continuous normal distribution as the paradigm, but the arguments generalize readily. We can

[5] Alternative concepts exist. For example: **Definition:** A series is *strictly stationary* if, and only if, the entire probability distribution function is invariant to time (Hendry [1995], p. 42).

illustrate the key points for our purposes with a simple example. Consider a classroom of six-year-olds. Measure their heights (H) and weights (W). These variables are associated in the data. As Sober puts it, above-average heights are typically found with above-average weights. More precisely, we can calculate the sample means, $m_X = (1/J)\sum_j X_j$, where $X = H$ or W, and the index of the pupils is $j = 1, 2, 3, \ldots$, and J is the number of pupils. Similarly, we can calculate the sample variances, $s_{XX} = (1/J-1)\sum_j (X_j - m_X)^2$, and the sample covariance $s_{HW} = (1/J-1)\sum_j (H_j - m_H)(W_j - m_w)$. As they stand, these statistics are simply facts of association. To draw probabilistic inferences from them, we must identify them with the population mean, $\mu_x = E(X)$, the population variances, $\sigma_{XX} = E[(X-\mu_X)^2]$, and the population covariance, $\sigma_{HW} = E[(H-\mu_H)\,(W-\mu_W)]$. (For clarity, sample moments are indicated by Latin letters and population moments by the corresponding Greek letters.) The bivariate normal distribution is completely defined by the two means, the variances, and the covariance.[6]

The key step in moving from facts of association to probabilities is the assertion that $\mu_X = m_X$, $\sigma_{XX} = s_{XX}$ and $\sigma_{HW} = s_{HW}$. The validity of this assertion is not directly testable. It is what econometricians refer to as an 'identifying assumption'. It can be supported indirectly by comparing higher sample moments to the analogous population moments. For example, a standard test for the normality of a set of data asks whether the third moment (skewness) and the fourth moment (kurtosis) calculated from the data match the third and fourth population moments calculated from the normal distribution parameterized using the sample mean and variance. The usual metric of closeness is to ask: *if* the sample had been drawn from a normal distribution, how probable is it that we would observe data that generate the test statistics (in the example, the sample third and fourth moments). In the Neyman-Pearson framework that dominates practical statistical inference, a decision rule is employed which sets a rejection probability (the *size* of the test, defined as the probability of observing a test statistic larger than the one calculated on the assumption that the sample is correctly described by the probability model). The critical rejection probability is ideally chosen to balance the risks of type-I error (rejecting the probability model when it is true) against type-II error (accepting the probability model when it is false).

[6] The probability density function for the bivariate normal is:

$$f(x, y) = (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1} \times \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

where σ_X and σ_Y are the standard deviations of X and Y (i.e., $\sqrt{\sigma_X}$ and $\sqrt{\sigma_Y}$) and the population correlation coefficient is defined as $\rho = \sigma_{XY}/\sigma_X\sigma_Y$, which is just another way of stating the definition of ρ previously given (see Lindgren ([1976], Ch. 10, §2.1).

While sample moments can always be calculated, no inference beyond the current sample is warranted without the mediation of the probability model. Suppose that we wish to use our knowledge of one classroom to extrapolate the probable distribution of heights and weights in another classroom of six-year-olds. This can be done only by asserting that the pupils in the second class are like those in the first class and appealing to the probability model. If we can actually measure the heights and weights in the second class, the sample means, variances, and covariances typically will not be identical to those of the first class. Can they be characterized by the same probability model? We can again follow a testing strategy analogous to the test of normality: on the assumption that the true probability model is the one parameterized by equating the sample moments of the first class to the population moments of the model, how probable are the sample moments calculated in the second class? If the probability falls below the threshold, we reject the hypothesis that the same probability model describes both classes.

Originally Simpson's paradox referred to the case in which a positive association in subpopulations disappeared when the whole population was considered. I will use the term, however, to refer to any case in which the statistical associations differ systematically between subpopulations and the whole population (see Hoover [2001], p. 19, esp. note 25 for historical references on Simpson's paradox). One strategy for dealing with Simpson's paradox is to insist on the rule: avoid 'mixing populations with different causal structures and different probability distributions' (Sober ([2001], p. 333). Sober objects to this strategy on the grounds that it saves the principle of the common cause only by depriving it of its epistemic force (Sober ([2001], p. 333, note 4). Again, the argument seems to be that we would already have to know that the populations were mixed in order to apply the rule, which begs the question.

Surely, this is wrong. The standard statistical strategy just described for testing the equality of the population moments between the two classrooms provides evidence for the homogeneity or lack thereof of the probability distributions. It is not, of course, completely decisive. For example, the sample moments of the two classes could be wildly different even though the population moments were identical because of an atypical draw from the distribution. Nonetheless, the probability distribution itself provides a measure of the improbability of such a draw. Such inferences are not free from *a priori* assumptions—particularly assumptions about the functional form of the probability distribution—yet they are not arbitrary. (The normal distribution is justified through central-limit theorems.) And they are not special to the question of homogeneity, Simpson's paradox, or the principle of the common cause, but are the assumptions that lie behind most statistical inference. What is more, they are assumptions about the probability distributions and not about causal structures. The principle of the common cause infers casual

structure from probability distributions. Any statistical test is fallible. It may lead to incorrect inference when applied to the principle of the common cause, just as in every other application. As a result, causal connections may be incorrectly inferred using principle (P). But the inferential chain is from facts of association to conclusions about probability models to conclusions about causal structure. The inferences are one-way. In particular, no causal assumptions are used in the inference of probabilities from facts of association. No questions are begged.

The identification of the population moments with the sample moments and the use of the sample moments to parameterize the probability model are justified under stationarity by the fact that the expected value of the sample moments is, in fact, the corresponding population moment. So, for example,

$$E(m_X) = E\left(\frac{1}{J}\sum_j X_j\right) = \left(\frac{1}{J}\right)\sum_j E(X_j) = \left(\frac{1}{J}\right)J\mu_X = \mu_X$$

Analogous proofs exist for the expectations of sample variances and covariances.

Pearson's correlation coefficient measures the strength of association between two series. It is a normalization of the sample covariance bounded between -1 and $+1$.[7] We can reasonably assume that the sample correlation coefficient between height and weight for the pupils (r_{HW}) is positive. Despite the fact that r is called a *correlation* coefficient, it is, according to the terminology that we have adopted, a sample statistic expressing a fact of association and not a fact of probability. Still, in a stationary world, we can reasonably take the value of r as an estimate of the corresponding population correlation—ρ. Given the maintained probability model, the likelihood of a true correlation of $\rho = 0$ generating the observed $r \neq 0$ can be calculated. This is a measure of statistical significance. It is reasonable to assume that in a class of six-year-olds, r_{HW} would be positive and significantly different from zero and that such a fact of association would provide evidence for the true correlation of height and weight (ρ_{HW}).

As with the sample mean, the expected value of the sample correlation coefficient is the true population correlation coefficient ($E(r) = \rho$) when the data are generated by a stationary probability distribution. What is more, as the sample size increases, the value of the sample mean, the sample variances, and the sample covariances (and, therefore, the sample correlation coefficient) all converge to their true population values in the sense that deviations from the population value are less and less likely.

Reichenbach's principle of the common cause is defined with respect to the improbability of a coincidental relationship between two events. Sober's

[7] It may defined as $r_{XY} = s_{XY}/\sqrt{s_{XX}s_{YY}}$.

principle (P) refers to correlation between two variables. Both require a notion of probabilistic dependence. *Probabilistic dependence* can be defined as the absence of probabilistic independence.

> **Definition:** Two variables X and Y are *probabilistically independent* if $P(XY) = P(X)P(Y)$, where $P(XY)$ is the joint probability of X and Y, $P(X)$ is the (marginal) probability of X, and $P(Y)$ is the (marginal) probability of Y. (See Everitt [1998], p. 163)

Probabilistic independence implies that the correlation coefficient $\rho = 0$, but $\rho = 0$ does not imply independence.[8]

Probabilistic dependence means roughly that the probability distribution (or the likelihood of various draws) of one series is different for different realizations (or draws) of another series.[9] A probabilistically dependent relationship between two series (X and Y) can be written:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

where the residuals, ε, capture the variability of Y not tracked by variations in X.

Equation (1) is a population relationship. Regressions assign values to the βs using sample data. The regression equation can be written as

$$Y = b_0 + b_1 X + e \tag{2}$$

where the bs are chosen in such a way that the mean of the implied residuals, e, is zero, and the sample covariance between the X and e, which of course depends on the choice of the bs, is zero.[10] Regressions are, therefore, directional: reversing the roles of X and Y, the bs of the new equation could not be calculated algebraically from (2), but must take account of the sample variances and covariances of the variables. Equations (1) and (2) nevertheless do not introduce any causal presuppositions into the notion of probabilistic dependence. If X and Y are probabilistically dependent, $X = \beta_{X0} + \beta_{X1} Y + \varepsilon_X$ and equally $Y = \beta_{Y0} + \beta_{Y1} X + \varepsilon_Y$. The population analogue of regression calculates the value of and $\beta_{X1} = \rho\sqrt{\sigma_{XX}/\sigma_{YY}}$ and $\beta_{Y1} = \rho\sqrt{\sigma_{YY}/\sigma_{XX}}$. Each is just a normalization of the correlation coefficient, and correlation, unlike causation, is not directional.

[8] Assuming, without loss of generality, that the means of X and Y are both zero, $\mathrm{cov}(X, Y) = E(X, Y) - E(X)E(Y)$, so that X and Y independent implies $\mathrm{cov}(X, Y) = 0$ and, therefore, that $\rho_{XY} = 0$. A probability distribution may imply expectations that cancel so that $\mathrm{cov}(X, Y) = 0$ and $\rho_{XY} = 0$ even though $P(XY) \neq P(X)P(Y)$. (Lindgren [1976], p. 136, gives the example of a symmetric, discrete bivariate distribution in which such cancellation occurs.)

[9] Before formally defining 'probabilistic independence' as we have just done, Everitt ([1998], p. 163) says, 'Essentially, two events are said to be independent if knowing the outcome of one tells us nothing about the other.' This captures the same 'rough' idea, although it gives the definition an unnecessarily epistemic cast.

[10] It is straightforward to extend regression analysis to three or more variables.

Statistically significant estimates of b_1 provide evidence for probabilistic dependence. As always, that evidence is evaluated on the maintained hypothesis of a probability model. Typically, that probability model is not maintained as a pure act of faith, but because facts of association not used in assessing probabilistic dependence are, conditional on the model, rendered likely. Estimates of b_1 are facts of association and evidence of probabilistic dependence, which is itself a fact of probability—a value for β_1 or ρ.

4 Nonstationary time series

The importance of the stationarity assumption can be highlighted through an extension of the example of the six-year-olds. In that case, when we sought to apply the probability model parameterized using one classroom's statistics to another, the comparison was *cross-sectional*: time was not an issue. We could also apply the probability model over time.

For example, suppose that we observe a sequence of classes of six-year-olds, each resident in the same classroom over a number of years. Each year we measure the heights and weights and calculate the sample statistics. It would not be surprising if the same probability model with the same parameters applied to each class in the sequence; that is, it would not be surprising if there were no statistically significant differences between their sample statistics. In this case, the time series would be stationary.

We could instead follow the same class as it grew up and advanced to higher grades. It would be startling if heights and weights did not advance with the age of the pupils. The same stationary probability model could not describe a class of six-year-olds and a class of twelve-year-olds, six years later. At a minimum, the heights and weights would have risen sufficiently, so that the difference between the sample means of the class at twelve were statistically significantly different from the class at six. The data would be nonstationary. In this case, we nonetheless believe that the heights and weights are probabilistically dependent. We shall return to this case in the next section. In the meantime, what about nonstationary series that are *not* probabilistically dependent?

Consider the class of six-year-olds and now measure height and knowledge of mathematics. It would not be surprising to find that there were no statistically significant correlation between the two. Yet, as years pass, both the heights and the knowledge of the class should increase. Assume that the sample correlation coefficient between height and knowledge among the twelve-year-olds is also not significantly different from zero. Nevertheless, if we were to pool the data from the class at six and the class at twelve, we would surely find that the sample correlation coefficient (r) was positive.

Similarly, if for each year from six to twelve, we were to calculate the mean height and mean knowledge for the class to form two time series, the sample correlation between the time series would be positive.

The pupils present a version of Simpson's paradox: height and knowledge are not associated among the six-year-olds nor later among the twelve-year-olds. But they are highly associated when the data are pooled. And their means are highly associated over time. But are they correlated in the relevant sense of probabilistically dependent? That is a question of what probabilistic process actually generates the data. It can be answered only by determining what probability model is justified by the facts of association. The data, for example, *could* have been generated by a stationary, normal distribution. The large shifts in the sample means, however, suggest that in that case the data would have to have been a highly atypical draw with a near zero probability. They are more likely to have been generated by a nonstationary distribution—that is, one in which the population moments depend on time.

To understand what difference it makes to work with nonstationary data, let us reconsider the case of the pupils. Suppose that we wanted to use the probability model with its parameters estimated on a particular sample to extrapolate – that is, to guess what value an as-yet unobserved additional datum would take. In the cross-sectional example of six-year old pupils, the best guess is that the new observation will take the value of the observed sample mean. This guess is justified by the fact noted in Section 3 that the expected value of the sample mean is, in fact, the population mean. Conditional on the probability model, we could use the sample variance to assign probabilities to the new observation taking values other than the mean.

Time series carry additional structure: the fact that the data are arranged in a specific temporal order. A new observation is generally a later observation. Now suppose that the data are nonstationary, as for example they could be if we calculated the means of height, weight, or knowledge from observing the same group of pupils as they progress through the grades. Again suppose, for example, that we have observed the class from ages six to twelve and want to guess what value the mean height will take at age thirteen. The sample average would be a poor basis for that guess. It is almost certainly too low. The problem is that the true population mean—that is, the expected value of a realization of the variable—now depends on the time. It is no longer related to the sample mean in the simple way that it was with stationary data.

Two important types of nonstationary time series are the 'trend stationary' and the 'integrated' series. To get our ideas clear we begin with the simpler case.

> **Definition**. A time series $\{X_t\}$ is *trend stationary* if, and only if, $X_t - g(t) = \varepsilon_t$, where $g(t)$ is a deterministic function of t and $\{\varepsilon_t\}$ is a weakly stationary time series with mean zero. (Hamilton [1994], p. 435)

Whereas a stationary series with a constant variance is concentrated about its mean, a trend stationary series is concentrated about $g(t)$, where, in general, $g(t)$ takes different values at different times. In principle, $g(t)$ might be a complicated function, but typically it is modeled as a simple function.

To illustrate, consider first a simple nonstationary process—a deterministic trend without any random element—for example the series $X_t = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$ for $t = 1, 2, 3, \ldots, 10$. The sample mean of the series is $m_X = 11$, but the next value in the sequence is not 11, but 22.

The point generalizes to a trend-stationary process. Let a series be described by

$$X_t = 2t + \varepsilon_t \tag{3}$$

where ε_t is a mean-zero, stationary random variable. (Here $g(t) = 2t$.)

Suppose that at time $t = \tau$ we know the value of ε_t. What is our best guess for the next observed datum (i.e., $X_{\tau+1}$)? Answer: $E_\tau(X_{\tau+1}) = 2(\tau + 1)$, where the subscript on the expectations operator indicates that it is based on information up to and including $t = \tau$. The population mean increases with time and is always greater than the sample mean. This is the sense in which the data are nonhomogeneous: they are drawn from a distribution whose true population mean is different in each period.

The sample correlation coefficient is as misleading as the sample mean as an indicator of the underlying population correlation. It will not serve as a reliable guide to the probabilistic dependence needed to invoke the principle of the common cause. Consider two time series governed by deterministic trends. One develops as $U_t = \{1, 2, 3, 4, \ldots\}$ plus a stationary random term, and the other as $V_t = \{3, 6, 9, 12, \ldots\}$ plus a stationary random term. The random terms are assumed to be mean zero and independent of each other; for example, they might be drawn from a normal or a uniform distribution centered on zero.[11]

Without the random terms, the sample correlation coefficient for U and V is $r = +1$; that is, they would be perfectly positively correlated. With the random terms and a short sample of, say, ten periods, r could take virtually any value—positive, negative, high, low. As the sample grows, the nonrandom components of U and V will come to dominate, and r will approach $+1$. But these series are *not* probabilistically dependent. The probability *distribution* of one series is not conditional on the realization of the other. A particular realization of the random term of U conveys no information about the probability distribution of V, or vice versa.

It is true, as Sober ([2001]) observes, that knowing the value of U (the deterministic trend + the random component) does convey information about

[11] It is trivial to create such series on Microsoft Excel or some other spreadsheet and to use them to verify the claims made here for their properties.

the value of V. But none of this information is probabilistic. It arises only because the deterministic components of both U and V are functions of time.[12] Any well-functioning time piece conveys information about any other. But an unpredictable stoppage of, say, your mantel clock does not give you reason to think that your wristwatch has stopped as well. Another way to see this is to observe that, were one to predict the value of U_6 on the basis of information up to $t = 5$, the probabilistic accuracy of that prediction would not improve from knowing V_6. It is the hallmark of probabilistic dependence that knowing the realizations of one series would improve the predictions of the dependent series. Here it does not.

Notice that nothing whatsoever has been said about causality. The issue is not whether observed associations sometimes do not warrant causal inferences according to the principle of the common cause. Rather it is whether observed associations do not warrant the inference that two series are correlated in the sense of being probabilistically dependent. When data are nonstationary, association generally implies neither true correlation nor probabilistic dependence.

Once again, it is reasonable to view this as a case of Simpson's paradox. The true population mean at each time is different. Each observation can be seen as drawn from a different subsample. For example, imagine that time could be rerun so that we had a million realizations of the processes that generate U and V—each differing according to the different realizations of the stationary error term. At each time, say $t = \tau_1$, there would be a million observations of $U_{\tau 1}$ and $V_{\tau 1}$. Given the way in which these data are generated, there is a very low probability of a statistically significant sample correlation between them. Similarly, for the data at a different time, say $t = \tau_2 \neq \tau_1$, there would be another population described by different stationary probability distribution with a *different* mean. Again, there is a low probability that $U_{\tau 2}$ and $V_{\tau 2}$ have a statistically significant sample correlation. Because the means are clearly different, the subpopulations are not homogeneous. And yet, if we pool the subpopulations and calculate the sample correlation between $U_{\tau 1} \cup U_{\tau 2}$ and $V_{\tau 1} \cup V_{\tau 2}$, then the sample correlation is likely to be significant—the more likely, the longer the gap between τ_1 and τ_2 (that is, the more likely, the less homogenous the subsamples). This is precisely the form of Simpson's paradox. And yet, it would not be a good idea to stop with showing that Simpson's paradox arises in nonstationary data, for the differences between the subsamples defined by the nonstationary data are not arbitrary, but are structured in a way that may permit avenues of probabilistic dependence

[12] One might be tempted to say that time is a common cause of both series. This is a deeply unattractive suggestion for a variety of reasons that would take us too far afield for present purposes.

opaque to statistical methods that rely solely on stationary probability models. I return to this point in the next section.

In the meantime, let us consider the U and V processes somewhat further. The sample correlation of the first differences of the series (ΔU_t and ΔV_t) is statistically insignificantly different from $r = 0$ in a long sample. Commenting on the analogous situation with respect to bread prices and sea levels, Sober ([2001], p. 334) argues that such a fact about changes gives no ground to dismiss the observed correlation in levels because 'higher than average bread prices *are* correlated with higher than average sea levels' (emphasis in the original).

Sober's statement equivocates on the word 'average'. It is true that higher than *sample-mean* bread prices are associated with higher than sample mean sea levels. But the necessary sense of 'average' is population mean, *not* sample mean. And, as we have already seen, the two do not coincide for trend-stationary time series with deterministic trends. In fact, the relationship of any observed value to the sample mean is not a stable fact in such a nonstationary case. For example, ignoring the random terms, the sample mean of the first four observations of U_t is 2.5. One half of the sample is above the sample mean and one half below. But, if we could observe four more terms, the sample mean increases to 4.5, and every one of the first four terms would be below the sample mean. In contrast, with a stationary distribution, such flip-flopping of status occurs only randomly and with a high probability only for values that are closer and closer to the population mean as sample size increases.

Sober ([2001], p. 334) claims that

> Pr[Higher than average sea level in year i & higher than average bread price in year i] > Pr[Higher than average sea level in year i]Pr[Higher than average bread price in year i].

He concludes from this that the levels of bread prices and sea levels are probabilistically dependent. Once stated correctly in terms of population means, the claim is false for trend-stationary data.

This last point is crucial. Sober rejects the principle of the common cause because two series that are not causally connected *ex hypothesi* are nevertheless correlated. Yet, the supposed correlation is a sample association. In the *most likely* probability model consistent with the data—a nonstationary model—such sample association does not necessarily correspond to probabilistic dependence.

Principle (P) would naturally be applied to the stationary components of a pair of trend-stationary time series. It will generally be misleading, as in Sober's example, if applied to series including their nonstationary trends. Equations (1) and (2) can be applied to stationary time series or to the

stationary components of trend-stationary time series with two modifications. First, we must account for the time order—notated by time subscripts on the variables and the residuals. Second, the residuals must be white noise.

> **Definition:** A series $\{\varepsilon_t\}$ is *(weak) white noise* if it is mean zero with a finite variance and $\text{cov}(\varepsilon_t\varepsilon_s) = E(\varepsilon_t\varepsilon_s) = 0$ for all $t \neq s$. (Hendry [1995], p. 39)

The definition says that residuals are *white noise* when there are no correlations between residuals at different times.

Sober ([2001], p. 335) also describes a process with a stochastic (i.e., a random, as opposed to a deterministic) trend. It is useful to look at this case as well because it prepares the ground for the next section. Sober's example illustrates the second major type of nonstationary time series, the 'integrated' time series. A series is *integrated* if its stochastic elements cumulate without loss over time. For example, in the series $X_t = X_{t-1} + \varepsilon_t$, where the ε_t are weakly stationary random shocks, the value k periods in the future can be written as $X_{t+k} = X_{t-1} + \varepsilon_t + \varepsilon_{t+1} + \varepsilon_{t+2} + \cdots + \varepsilon_{t+k}$. In such a process, a shock at $t = 1$, call it ε_1, will contribute its full value to X_t, not only at $t = 1$, but at $t = 100$ and $t = 100{,}000$, and every period that occurs after $t = 1$. As k goes to infinity, the effect of an individual shock never dwindles away. Similarly, the initial condition, X_{t-1}, fully contributes to the value of X_{t+k} at every period. An integrated time series 'remembers' its past. While one can always calculate the sample mean of an integrated series, the series does not tend to revert to its sample mean, but instead drifts about, depending on the particular realizations of the εs.

In contrast, consider the time series $X_t = 0.5X_{t-1} + \varepsilon_t$. The value k periods in the future can be written as $X_{t+k} = 0.5^k X_{t-1} + 0.5^{k-1}\varepsilon_t + 0.5^{k-2}\varepsilon_{t+1} + 0.5^{k-3}\varepsilon_{t+2} + \cdots + 0.5\varepsilon_{t+k-1} + \varepsilon_{t+k}$. A shock at $t = 1$ will contribute its full value to X_t only at $t = 1$. At $t = 100$, it contributes only $0.5^{99}\varepsilon_1 = (1.58 \times 10^{-30})\varepsilon_1$, and, as k goes to infinity, the effect of an individual shock dwindles away rapidly. Similarly, the contribution of the initial condition dwindles rapidly as k increases. Such a series is stationary, even though values of X close together in time may be correlated (i.e., $\{X_t\}$ is not white noise}. A stationary time series 'forgets' its past. And, while over a short enough sample it may display systematic deviation from its sample mean, over long samples it displays mean reversion.

The stochastic process $X_t = X_{t-1} + \varepsilon_t$ is one example of a nonstationary process. There are other processes that cumulate stochastic shocks in different patterns. These can be categorized by their order of integration. Define the difference operator: $\Delta X_t = X_t - X_{t-1}$. The difference operator transforms levels of variables into changes and can be applied to a series that has already been differenced, so that $\Delta(\Delta X_t) = \Delta^2 X_t$. More generally $\Delta^d X_t$ indicates that

the series has been differenced d times.

> **Definition:** Consider a time series $\{X_t\}$ that is neither stationary nor trend stationary. Let d be the minimum integer such that $\{\Delta^d X_t\}$ is weakly stationary. Then $\{X_t\}$ is said to be *integrated of order d*, which is notated I(d). (By convention, a stationary time series is notated as I(0).)

An integrated series is *not* weakly stationary. Weak stationarity can be tested using standard statistical procedures that amount to tests of subsample homogeneity. Most macroeconomic time series appear to be I(1)—that is, they are not stationary, but their first differences are. The series $X_t = X_{t-1} + \varepsilon_t$ is I(1). Some appear to be I(2). An example of an I(2) process is: $X_t = X_{t-1} + \Delta X_{t-1} + \varepsilon_t$. And only a very few appear to be I(0)—that is, their levels are stationary.

Sober's particular example is a (possibly asymmetric) two-state, discrete Markov process. We look instead at a more tractable, but closely related, continuous process. We have already seen a typical example of an I(1) nonstationary process:

$$X_t = X_{t-1} + \varepsilon_t \tag{4}$$

where ε_t is a draw from a weakly stationary stochastic process. Notice that the population mean—as we should expect in a nonstationary series—is not constant: $E(X_t) = X_{t-1}$. In fact, the series describes the famous 'random walk' in which the best expectation of today's value is yesterday's value.

Time order introduces considerable mathematical complexity into the characterization of the probability distributions of integrated time series, involving such notions as Wiener processes (Brownian motion) and functional central-limit theorems.[13] One critically important result of this analysis is relevant to the bread prices/sea levels example. Suppose that we have two series, $\{Y_t\}$ and $\{Z_t\}$, each fully described by an equation like (4) in which the two stationary time series of errors (the ε_t) are probabilistically independent. One realization of each series assigns random values to each ε_{Yt} and ε_{Zt} and generates each Y_t and Z_t according to an equation like (4). We could, of course, generate many such realizations with different random values for the εs, and different Ys and Zs. What happens if we calculate the sample correlation coefficient, r, for each realization? If the process had been stationary, the frequency distribution of r would tend to center on the population correlation coefficient, ρ, and would tend to become more concentrated on ρ (i.e., to have a smaller variance around it) as the number of realizations and the length of each sample realization became larger.

[13] Key articles in this literature include Dickey and Fuller ([1979]), Phillips ([1986]) and ([1987]), and Engle and Granger ([1987]). Textbook treatments are now available in, for example, Hamilton ([1994]) and Hendry ([1995]).

In contrast, with nonstationary series the sample correlation coefficient (r) between the two series does not get closer and closer to the population correlation (ρ). In the case of time series such as Y and Z, which are both I(1), the distribution of values of r is close to uniform over the interval -1 to $+1$ (see Hendry [1995], p. 128). This means that every value of r will show up with about the same probability in different realizations of the two processes, even though there is no connection whatsoever between them.

In general, unlike the case of stationary time series, the frequency distribution of the sample statistics of nonstationary time series does not center on a single population value, but displays a stable distribution of values. To take another example, if both series are I(2), the weight of the distribution of r is heavily concentrated in the tails, so that there is a very high and nearly equal probability of finding r to be near either -1 or $+1$ and a low probability of finding any value in between.

Despite the apparent correlation, there is once again no probabilistic dependence in these cases. The distribution of one series is not conditional on a realization of the other series. A striking result underscores the lack of probabilistic dependence. As the sample sizes grow, the variance of the difference between the two series approaches infinity. This last result says that, even though the series appear to be highly correlated, they do not tend toward parallel paths; indeed, they may drift infinitely far apart.

5 Probabilistic dependence in nonstationary time series

The strategy of concentrating on the changes, rather than the levels, of nonstationary time series might seem to gain considerable support from the analysis of the last section. The first difference of the integrated series X_t in (4) is $\Delta X_t = X_t - X_{t-1} = \varepsilon_t$, a stationary random process in which sample moments provide evidence for their straightforwardly analogous population moments. Similarly, the first differences of the trend-stationary series U and V would be two independent random terms. A lack of a statistically significant r statistic would provide evidence that ΔU and ΔV were not correlated in the population. But, as mentioned in Section 2, this application of (P*) may throw the baby out with the bathwater.

Recall the case of measuring the heights and weights of a class as it progressed from ages six to twelve. The connection between height and weight among the pupils at a particular time (i.e., in the cross-section) is likely to be genuine. Equally, the connection between the time series of the average heights and weights is also likely to be genuine. Yet, it does not follow from this last connection that there will be a close connection between the change in height and the change in weight. For example, suppose that the probabilistic process

governing them is:

$$\Delta W_t = \omega_t \tag{5}$$

$$\Delta H_t = -\gamma(H_{t-1} - W_{t-1}) + \nu_t \tag{6}$$

where ω and ν are independent, stationary random variables and $0 < \gamma < 1$.

Both W and H are I(1). W is a random walk. Its population mean drifts aimlessly about like a drunk wandering away from the tavern. Unlike (5), which describes the mean of ΔW_t as zero, (6) is like a random walk, but one with a mean that shifts in a particular deterministic manner. If $H_{t-1} > W_{t-1}$, then $-\gamma(H_{t-1} - W_{t-1})$ is negative, so that the negative realizations of ΔH_t become more common; while if $H_{t-1} < W_{t-1}$, positive realizations become more common. The gap between H_t and W_t tends probabilistically to be closed—the faster, the larger γ is. Hence, $-\gamma(H_{t-1} - W_{t-1})$ is often called an *error-correction mechanism*. While both series are I(1), the shifting mean implies that, unlike the general case mentioned in the last section in which the variance of the difference between two integrated series approaches infinity as the sample size grows, the variance of the difference between H and W is finite. So, as W wanders aimlessly like a drunk, H tends to follow like a faithful friend who does not want to lose sight of his inebriated companion. H and W are probabilistically dependent.

Nevertheless, ΔH_t and ΔW_t are not probabilistically dependent: conditional on the information available at $t-1$, the realization of ΔW_t conveys no information about the distribution of ΔH_t or vice versa.

Of course, it is also possible for series to be dependent in both changes and levels. For example, replace (5) with

$$\Delta W_t = \delta \Delta H_t + \omega_t \tag{7}$$

Now ΔH and ΔW are also probabilistically dependent.

Notice that in (6), if we could condition H_t on W_t, all that is left over is stationary.[14] We would have, in effect, decomposed (6) into a stationary and a nonstationary component. The nonstationary component represents the genuine relationship of probabilistic dependence between the levels of H and W known as 'cointegration'.

> **Definition:** Two time series $\{X_t\}$ and $\{Y_t\}$ are *cointegrated* if, and only if, each is I(1) and a linear combination $\{X_t - \beta_0 - \beta_1 Y_t\}$, where $\beta_1 \neq 0$, is I(0). (Hamilton [1994], p. 571)

In general, linear combinations of I(1) time series are also I(1). Cointegration is a particular feature not displayed between arbitrary pairs of time series.

[14] In practical statistics, such a conditioning is accomplished through regression. The residual errors from the regression of H_t on W_t is the stationary component of H_t.

If two time series are cointegrated, then the *cointegrating vector* ($[\beta_1]$) is unique (see the Appendix for a proof in a special case).[15]

It is easy to see heuristically that H and W are cointegrated in (5) and (6). (A formal demonstration that H and W are cointegrated is relegated to the Appendix.) W in (5) is clearly I(1). The error-correction mechanism in (6) (i.e., $-\gamma(H_{t-1} - W_{t-1})$ ensures that, in the long run, H follows W. So, if W is I(1), H must also be I(1). The gap between the two series cannot diverge in the long run. The error-correction mechanism also ensures that the bigger the gap between H_{t-1} and W_{t-1}, the faster H and W approach each other. Only I(0) realizations of ν_t can drive H and W apart. Since the variance of ν_t is finite, there is a vanishingly small probability of a string of realizations of ν_t large enough to overcome the error-correction mechanism.[16] Thus, while H and W may each have infinite variance, the variance of their difference is finite.

Like the correlation coefficient, cointegration is a symmetrical relationship. If X and Y are cointegrated with a linear combination as in the definition above, they are also cointegrated with another linear combination in which the roles of X and Y are reversed. There is a further analogy with the correlation coefficient. Notice that the cointegrating relationship in the definition has the same form as (1). Just as the regression equation (2) chooses the weights for this linear combination such that a pair of stationary series is reduced to a series of residual error terms uncorrelated with the right-hand side variables, a regression equation

$$X_t = b_0 + b_1 Y_t + e_t \tag{8}$$

chooses the weights of the cointegrating vector in such a way that a pair of nonstationary variables is reduced to a series of stationary residuals (e_t). Now, however, even if the underlying random error terms (e.g., ω and ν in (5) and (6)) are normal, the distribution of the estimated coefficients will not be normal, so that common test statistics used to check statistical significance (e.g., Student's t-test of whether b_1 is statistically significantly different from zero) no longer have the distributions familiar from stationary data. Determining valid distributions and appropriate test statistics is an important element in recent research on nonstationary time series.

The correspondence between facts of association based on familiar sample statistics (including 'eye-ball' measures) and the facts of probability implicit in

[15] Called a *vector* because in a multivariate case (see below) in which the linear combination involves more than two I(1) time series, there are additional coefficients for each additional time series. So, for j cointegrated I(1) time series the vector would be $[\beta_1\ \beta_2 \ldots \beta_{j-1}]$.

[16] It would be tedious to try to avoid terms such as 'correction mechanism', 'ensure', and 'drive apart' which might appear to be causal. In fact, we are appealing here only to descriptions of the nonstationary probability distributions and, with loss of economy, could avoid causally loaded language. No questions are begged.

Sober's counterexample fail in nonstationary settings. Yet evidence of cointegration does provide evidence of probabilistic dependence. In (5) and (6), H and W are cointegrated. The distribution of H differs with different realizations of the random terms ω_t in the W process and, hence, with the values of W. In contrast to the case of U and V in the last section, knowing W_t gives information about the likely values of H_t not derivable from knowing the H process alone. The joint probability distribution of H and W does not factor: $\mathrm{P}(H_t W_t) \neq \mathrm{P}(H_t)\mathrm{P}(W_t)$. This is the natural meaning of probabilistic dependence. It is also the sense needed to make sense of principle (P) and the sense implicit in correlation in Reichenbach's ([1956]) usage.

So far, we have considered pairwise cointegration, but two time series may not be cointegrated as a pair but, nonetheless, may be elements of a set of three or more I(1) variables for which there is a linear combination (e.g., $\{X_t - \beta_0 - \beta_1 Y_t - \beta_2 Z_t\}$) that is I(0). The statistical analysis generalizes from the pairwise to the multivariate case in straightforward ways. The multivariate case points up a disanalogy between cointegration and correlation when applied to the causal Markov condition or the principle of the common cause. If two I(0) series are not correlated ($\rho = 0$), then, except in special cases, they are not directly causally connected.[17] But if two I(1) series are not pairwise cointegrated, they may still be directly causally connected, since it may require a linear combination with one, two, or more additional I(1) variables to reduce the I(1) processes to I(0). This could, of course, affect practical causal investigations. With stationary series, one can start small with pairwise relationships and work out to more complex systems. With nonstationary series, one must start with a system and run the risk that it is not adequately large for the problem at hand.

6 Do Venetian sea levels and British bread prices violate the principle of the common cause?

Are Venetian sea levels and British bread prices a valid counterexample to the principle of the common cause?

It is worth recalling Reichenbach's ([1956], p. 156) original formulation of the principle quoted at the outset of this paper: '*If an improbable coincidence has occurred, there must exist a common cause*.' The antecedent presumes that coincidences can be either probable or improbable. As already noted, Reichenbach rules out *mere* coincidences as examples of the kind of improbability he has in mind. (It is in some sense highly unlikely that the space shuttle *Challenger* should have blown up on the day my daughter was born. It was a mere coincidence.) To apply Reichenbach's principle, the probability models

[17] See Spirtes *et al.* ([1993], p. 95) and Hoover ([2001], pp. 168–70).

that describe the stochastic occurrence of each event separately must place a small probability on their co-occurrence; and yet they do in fact co-occur. In that case, the common cause justifies a wider probability model which makes their co-occurrence follow as a matter of course.

Bread prices in England and the sea levels in Venice do not have a common cause *ex hypothesi*. But to establish that they violate the principle, the fact of their observable positive association would have to be interpretable as a true correlation and, therefore, as an improbable coincidence in Reichenbach's sense. As described in Sober's counterexample, the two time series are clearly nonstationary—either trend-stationary or integrated. The positive sample association between such nonstationary series is *not* improbable. Indeed, such observed associations between time series that lack any probabilistic dependence are the mathematical implication of nonstationary stochastic processes. They are highly predictable, and yet are mere coincidences. As a result, when, given the data, the most likely probability model is nonstationary, it is wrong to infer probabilistic dependence from Pearson's correlation coefficient or other formal or informal measures of association that would have provided good evidence for it in the stationary case. The time series are associated but not correlated in the sense required for the principle of the common cause. This is enough to justify the conclusion that Venetian sea levels and British bread prices do not provide a valid counterexample to the principle of the common cause.

Not all nonstationary time series are as probabilistically independent as sea levels and bread prices. Cointegration is the probabilistic dependence between the levels of nonstationary time series. The term 'correlation' in the sense relevant to Reichenbach's principle of the common cause encompasses cointegration. Far from needing to weaken (P*) to cope with Sober's counterexample, our analysis shows that principle (P) is fine as it is.

This conclusion may be thought to employ too elastic a definition of 'correlation'—one that does violence to common usage. Sober's counterexample might reasonably be held to highlight a need to clarify key terms. So, even though I argue that Sober's counterexample does not need to be replaced by any weaker principle, it might be clearer to restate (P) more expansively (with the clarifying phrases in italics):

> **(P**)** If events X and Y *are each stationary or trend-stationary and* are correlated with each other *or are each integrated and cointegrated with each other*, then either X caused Y, Y caused X, or X and Y are joint effects of a common cause (one that renders X and Y probabilistically independent).[18]

[18] Cointegration was defined in Section 4 as a linear combination that reduces an I(1) series to an I(0) series. But the notion generalizes to a linear combination that reduces an I(k) series to an I($k-1$) series. Such a generalization is adequate to defuse Sober's [(2001), Section 3) new counterexamples. These work through the construction of two time series that are not

Venetian sea levels and British bread prices are nonstationary, and they are not cointegrated. They fail to fulfill the antecedent of (P**), and so are not a counterexample to the principle of the common cause.[19]

The principle of the common cause may be open to valid criticism on many grounds. What we have shown here is that the case of Venetian sea levels and British bread prices is not among those grounds.

Acknowledgements

I am grateful to Paul Teller, Elliott Sober, Daniel Hausman, Daniel Steel, and an anonymous referee for valuable comments on an earlier draft, and to Oscar Jorda for helpful discussions of time-series econometrics.

Department of Economics
University of California
1 Shields Avenue
Davis, CA 95616
USA
kdhoover@ucdavis.edu

Appendix: Proof that the W and H in (5) and (6) are cointegrated

First, some useful facts: (i) linear combinations of I(0) series are themselves I(0); (ii) linear combinations of an I(1) series and an I(0) series are I(1).

Without further discussion, we start with the fact established informally in the text that W and H are I(1). The salient question, then, is whether there is a linear combination of W and H that is I(0). We prove that $(H_t - W_t)$ is in fact I(0). Rewrite (5) and (6) as

$$W_t = W_{t-1} + \omega_t \tag{5'}$$

$$H_t = (1-\gamma)H_{t-1} + \gamma W_{t-1} + \nu_t \tag{6'}$$

Subtracting (5′) from (6′) yields

$$(H_t - W_t) = (1-\gamma)(H_{t-1} - W_{t-1}) + \nu_t - \omega_t \tag{9}$$

causally connected *ex hypothesi* but in which *both* the levels *and* the changes of the series are correlated in sample in violation of (P*). Such series are not cointegrated, so that they do not violate (P**).

[19] Sober's ([2001]) artificial data (see note 2 above) conform to these claims. The levels trend up over time, which is reflected in $r = 0.99$. The changes are trendless and uncorrelated: $r = 0.47$ with a t-statistic = 1.19 (it would have to be greater than 2.57 to be significant at the conventional 5 percent level). Unfortunately, the two series are too short for most formal tests of cointegration.

Lagging (9) and repeatedly substituting out the lagged term k times yields

$$(H_t - W_t) = (1-\gamma)^k(H_{t-k} - W_{t-k}) + \sum_{j=0}^{k-1}(1-\gamma)^j(\nu_{t-j} - \omega_{t-j}) \qquad (10)$$

Since $\gamma < 1$, as $k \to \infty$, the first term on the right-hand side vanishes. The summation places weights that decline over time on a linear combination of I(0) error terms. The further a shock to W (i.e., ω) or to H (i.e., ν) lies in the past, the less it matters to the current value of $(H_t - W_t)$. This is the hallmark of an I(0) process. Since $H_t - W_t$ is I(0), H and W are cointegrated.

We can also prove that the cointegrating vector is unique. Let (5) and (6) describe the relationship between H and W. Now compute, an arbitrary linear combination

$$\begin{aligned}(H_t - \alpha W_t) &= (1-\gamma)H_{t-1} + \gamma W_{t-1} + \nu_t - \alpha W_{t-1} - \alpha\omega_t \\ &= (H_{t-1} - \alpha W_{t-1}) - \gamma(H_{t-1} - W_{t-1}) + \nu_t - \alpha\omega_t \qquad (11)\end{aligned}$$

If $\alpha = 1$, then (11) collapses to (9) and $(H_t - \alpha W_t)$ is I(0). If $\alpha \neq 1$, notice that (11) takes the form $X_t = X_{t-1} + \text{I(0) terms}$, which is an I(1) process. To see this another way, ask what happens if we repeatedly lag $(H_t - \alpha W_t)$ and substitute out the lagged values in (11) as we did in moving from (9) to (10). This process yields

$$(H_t - \alpha W_t) = (H_{t-k} - \alpha W_{t-k}) + \sum_{j=0}^{k-1}[-\gamma(H_{t-j} - W_{t-j}) + (\nu_{t-j} - \alpha\omega_{t-j})] \qquad (12)$$

Unlike in (10), as $k \to \infty$, the first term on the right-hand side does not vanish and the weights on the terms in the summation of I(0) terms do not decline. The process 'remembers' its initial conditions and the full value of every past shock, which is the hallmark of an I(1) process.

Since $\alpha = 1$ implies that $(H_t - \alpha W_t)$ is I(0) and $\alpha \neq 1$ implies that $(H_t - \alpha W_t)$ is I(1), the cointegrating vector $[\alpha]$ is unique.

References

Cartwright, N. [1999]: *The Dappled World*, Cambridge: Cambridge University Press.

Dickey, D. A. and Fuller, W. A. [1979]: 'Distributions of the Estimators for Autoregressive Time Series with a Unit Root', *Journal of the American Statistical Association*, **74**, pp. 427–31.

Engle, R. F. and Granger, C. W. J. [1987]: 'Cointegration and Error Correction: Representation, Estimation and Testing', *Econometrica*, **55**, pp. 251–76.

Everitt, B. S. [1998]: *The Cambridge Dictionary of Statistics*, Cambridge: Cambridge University Press.

Forster, M. [1988]: 'Sober's Principle of the Common Cause and the Problem of Comparing Incomplete Hypotheses', *The British Journal for the Philosophy of Science*, **55**, pp. 538–59.

Hacking, I. [1965]: *The Logic of Statistical Inference*, Cambridge: Cambridge University Press.

Hamilton, J. D. [1994]: *Time Series Analysis*, Princeton: Princeton University Press.

Hausman, D. M. and Woodward, J. [1999]: 'Independence, Invariance, and the Causal Markov Condition', *The British Journal for the Philosophy of Science*, **50**, pp. 521–83.

Hendry, D. F. [1980]: 'Econometrics: Alchemy or Science?' *Economica*, **47**, pp. 387–406.

Hendry, D. F. [1995]: *Dynamic Econometrics*, Oxford: Oxford University Press.

Hoover, K. D. [2001]: *Causality in Macroeconomics*, Cambridge: Cambridge University Press.

Lindgren, B. [1976]: *Statistical Theory*, 3rd edn, New York: Macmillan.

Meek, C. and Glymour, C. [1994]: 'Conditioning and Intervening', *The British Journal for the Philosophy of Science*, **45**, pp. 1001–21.

Mellor, D. H. [1971]: *The Matter of Chance*, Cambridge: Cambridge University Press.

Mood, A. M., Graybill, F. A. and Boes, D. C. [1974]: *Introduction to the Theory of Statistics*, 3rd edn, New York: McGraw-Hill.

Papineau, D. [1992]: 'Can We Reduce Causal Direction to Probabilities', in D. Hull and K. Okruhlik (*eds*), 1992, *PSA 1992*, Vol. 2, East Lansing, MI: Philosophy of Science Association, pp. 238–52.

Phillips, P. C. B. [1986]: 'Understanding Spurious Regressions in Econometrics', *Journal of Econometrics*, **33**, pp. 311–40.

Phillips, P. C. B. [1987]: 'Time Series Regression with a Unit Root', *Econometrica*, **55**, pp. 277–301.

Reichenbach, H. [1956]: *The Direction of Time*, Berkeley: University of California Press.

Sober, E. [1994]: 'The Principle of the Common Cause', in *From a Biological Point of View*, Cambridge: Cambridge University Press, pp. 158–74.

Sober, E. [2001]: 'Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause', *The British Journal for the Philosophy of Science*, **52**, pp. 331–46.

Spirtes, P., Glymour, C. and Scheines, P. [1993]: *Causation, Prediction, and Search*, New York: Springer Verlag.

Yule, G. U. [1926]: 'Why Do We Sometimes Get Nonsense Correlations Between Time Series? A Study of Sampling and the Nature of Time Series' (with discussion), *Journal of the Royal Statistical Society*, **89**, pp. 1–64.