Likelihood and Consilience: On Forster's Counterexamples to the Likelihood Theory of Evidence

Jiji Zhang

Department of Philosophy, Lingnan University

Kun Zhang

Max Planck Institute for Intelligent Systems

Abstract

Forster presented some interesting examples having to do with distinguishing the direction of causal influence between two variables, which he argued are counterexamples to the likelihood theory of evidence (LTE). In this paper, we refute Forster's arguments by carefully examining one of the alleged counterexamples. We argue that the example is not convincing as it relies on dubious intuitions that likelihoodists have forcefully criticized. More importantly, we show that contrary to Forster's contention, the consilience-based methodology he favored is accountable within the framework of the LTE.

1 Introduction

Forster (2006) presented some putative counterexamples to what he called a/the likelihood theory of evidence (LTE):

> "The Likelihood Theory of Evidence (LTE): The observed data are relevant to the comparison of simple hypotheses (or models) only via the likelihoods of the simple hypotheses being compared (or the likelihood functions of the models under comparison)." (321)

The LTE entails that if the likelihood of one hypothesis relative to a given body of data — i.e., the probability of obtaining the data given the hypothesis — is the same as that of another hypothesis, then the hypotheses cannot be distinguished based on the data alone. Forster challenged this consequence with examples in which the data, he argued, favor one hypothesis over another even though the two have the same likelihood.

For those concerned with causal inference, Forster's examples are particularly interesting because they have to do with distinguishing the direction of causal influence between two random variables. Forster contended that his examples demonstrate a distinctive methodology based on Whewell's notion of "consilience of inductions" (Whewell 1858; Forster 1988), which cannot be captured by a likelihoodist or Bayesian philosophy of science that subscribes to the LTE.

Our purpose in this paper is twofold, one critical and one positive. First, in section 2, we argue that Forster's challenge to the LTE is based on denying a basic, well-argued thesis of likelihoodism. The thesis is that the evidential bearing of a body of data on a given statistical hypothesis is essentially relative, depending on the alternative against which the hypothesis is assessed. The apparent force of Forster's counterexamples, we argue, relies upon embracing an intuition that likelihoodists (e.g., Hacking 1965; Royall 1997; Sober 2008) have forcefully criticized — the intuition that a statistical hypothesis, taken alone, can be rejected or shown to be false by data. At best, therefore, Forster's argument begs an important question against the likelihoodist.

Second, and more importantly, we aim to vindicate Forster's preferred methodology using likelihoods. We show in section 3 that there is a systematic connection between likelihood and the kind of consilience Forster emphasized. Forster is right that considerations of consilience are evidentially relevant. However, such relevance, we contend, is

reflected in likelihoods.

Due to the space limit, we will focus on Forster's example featuring discrete variables, but our points extend straightforwardly to his example with continuous variables, as we will briefly comment in section 4.

2 On Forster's challenge to the LTE

For the likelihoodist, a thesis of fundamental importance is what Royall (1997) called the "relativity of evidence". A body of data constitutes evidence for or against a statistical hypothesis *only* relative to some alternative hypothesis. For example, getting ten heads straight in tossing a coin is not evidence against the coin being fair *simpliciter*. It disconfirms the fair-coin hypothesis in reference to some alternative hypothesis, e.g., the hypothesis that the coin is a trick coin with heads on both sides, or that the coin is so biased towards one side that the chance of landing head in each flip is 0.9. But when compared to certain other alternatives, e.g., the hypothesis that the coin is a trick coin with tails on both sides, the observations favor the fair-coin hypothesis. The evidential bearing of the data on the fair-coin hypothesis is thus relative to the alternative being considered; evidential statements are essentially contrastive in form.

Detailed and compelling arguments for this view were elegantly presented by, among others, Royall (1997, 65-68) and Sober (2008, 48-52), and we shall not repeat them here. Suffice it to say that objections to the likelihood account of evidence that rely on denying the relativity of evidence begs an important question. We will argue that Forster's challenge ends up question-begging in this way.

It is not obvious that Forster ran afoul of the relativity of evidence. His counterexamples apparently pit a hypothesis against another. Here is the first version of the example

we will focus on in this paper. Suppose that two variables X and Y are related by a simple law: $Y = X + U$, where X is a variable taking positive integer values, and U is an unobserved variable — error term — taking one of two values: 0.5 or -0.5, with equal probability. Suppose also that data are generated by twenty independent trials, with $X = 4$ in each trial. As it happens, in ten of the twenty trials, Y is observed to be equal to 3.5 (i.e., the values of U in those trials are -0.5), and in the other ten trials, Y is observed to be equal to 4.5 (i.e., the values of U in those trials are 0.5).

Let us use X_i, Y_i, etc. to model the ith trial. Consider now two hypotheses. One is the true hypothesis, which Forster referred to as Hypothesis A: $Y_i = X_i + U_i$ $(i = 1, ..., 20)$, and the error terms U_i's are independently and identically distributed (i.i.d.) such that $P(U_i = -0.5) = P(U_i = 0.5) = 1/2$.

The other hypothesis is referred to as Hypothesis B (for Backwards): $X_i = Y_i + V_i$ $(i = 1, ..., 20)$, and the error terms V_i's are i.i.d. such that $P(V_i = -0.5) = P(V_i = 0.5) = 1/2$.[1]

In the first version of the example, Forster considered these two hypotheses as such, and treated the exogenous variable in each hypothesis as non-random or given. Specifically, in A, X_i's are not treated as random variables, but Y_i's are (because U_i's are); in B, Y_i's are not treated as random variables, but X_i's are (because V_i's are). For these hypotheses, as Forster pointed out, only *conditional* likelihoods are well defined. For A, the conditional likelihood is the probability of obtaining the observed values of Y_i under hypothesis A, given the values of X_i, which is $(1/2)^{20}$; for B, the conditional likelihood is the probability of obtaining the observed values of X_i under hypothesis B, given the

[1]Forster used the same symbol U to denote the error terms in both hypotheses, which is potentially misleading. To avoid confusions, we use V to denote the error term postulated by the backwards hypothesis.

values of Y_i, which is also $(1/2)^{20}$.

According to Forster (2006),

> "The example is already a counterexample to LTE in the following sense: We are told that either A or B is true, and we can tell from the data that A is true and B is false. But there is nothing in the *likelihoods* that distinguishes between them." (328, original emphasis)

We will return to Forster's claim that one can tell from the data that A is true and B is false. For now let us focus on a more basic problem with this version of the example. The problem is that the two hypotheses concern *different* random variables: the random variable in A is Y (or more accurately, $\langle Y_1, ..., Y_{20} \rangle$), and the random variable in B is X (or $\langle X_1, ..., X_{20} \rangle$). However, a presupposition of LTE is that the hypotheses in question concern a *common* set of random variables: the hypotheses imply probability distributions over these variables, and the data are observations of their values. Royall, for example, made it explicit in his influential formulation of the law of likelihood:

> "If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$ ... ." (Royall 1997, 3)[2]

Clearly the present version of the example does not satisfy the presupposition. Thus, for likelihoodists like Royall, it does not make sense to talk about the evidential support of A versus B.

[2]Royall seemed to attribute this formulation to Hacking (1965), but as far as we can see, Hacking did not formulate his law of likelihood in precisely these terms.

To be fair, Forster was quick to acknowledge that a subscriber to LTE could easily respond to this version of the example by denying that LTE should apply to such "incomplete" hypotheses. He put the subscriber's complaint in the following terms:

> "They might insist that the example violates the principle of total evidence because the likelihoods are not relative to the full data, even though there are no data "hidden from view", or withheld in any way." (Forster 2006, 328)

This, in our view, is a misdiagnosis on behalf of the the friends of LTE. The principle of total evidence is about *what* evidence to take into account, but the LTE is about the evidential bearing of *given* evidence on the comparison of hypotheses. It is perfectly sensible to ask whether a certain *part* of the data supports one hypothesis against another (though one should take total evidence into account, if possible, when updating beliefs or judgements). In the present case, for example, there is no problem comparing, based on conditional likelihoods, hypothesis A with, say, A^*: $Y_i = X_i + U_i, i = 1, ..., 20$, and $P(U_i = -0.5) = 1/4$ and $P(U_i = 0.5) = 3/4$. A and A^* are as "incomplete" as A and B are, but they are about the same random variables, and hence are comparable given the data. By contrast, A and B as such are incomparable[3] because they concern entirely different random variables.

Why are hypotheses incomparable if they are about different random variables? This is connected to the thesis of evidential relativity. To see the matter clearly, it helps to consider a simpler case. Suppose two coins are each flipped independently for twenty times. Of the first coin, all of the twenty flips turn up heads; of the second coin, half of the flips turn up heads and half turn up tails. Consider two hypotheses: (1) the first

[3]By "incomparable" we mean only that the hypotheses are not subject to *evidential* comparison. They may still be comparable in terms of rational credences or someone's personal credences.

coin is fair, and (2) the second coin is fair. The observations on the first coin — call them D_1 — are irrelevant to hypothesis (2) (in the absence of any background knowledge or assumption linking the two coins). So it does not make sense to say that D_1 provide evidence for or against (1) versus (2). Similarly, we cannot say that D_2, the data on the second coin, provide evidence for or against (2) versus (1).

However, it may be tempting to think that the degree to which D_2 support (2) is greater than the degree to which D_1 support (1). After all, it seems intuitive that (1) fits D_1 very poorly while (2) fits D_2 rather well. If so, it would be fair to say that (1) and (2) are comparable after all, given the combined data $D = \langle D_1, D_2 \rangle$. But according to the relativity of evidence, there is no such thing as the degree to which D_2 support (2) *simpliciter* or that to which D_1 support (1) *simpliciter*. D_1 confirm or disconfirm (1) only in contrast to some other hypothesis concerning the outcomes of flipping coin 1, and D_2 confirm or disconfirm (2) only in contrast to some other hypothesis concerning the outcomes of flipping coin 2. Hence, it does not make sense to say that D_2 support (2) better than D_1 support (1).

Therefore, from the likelihoodist point of view, the basic problem with the present 'counterexample' is not so much a violation of the principle of total evidence as a juxtaposition of incomparable hypotheses, and the incomparability is closely related to the relativity of evidence. Forster's neglect of this point signals his denial of the relativity of evidence.

In any case, Forster did develop the example into one with comparable hypotheses. Treat both $X_i's$ and $Y_i's$ as random variables. To A add the assumption that X_i and U_i are statistically independent, and that $P(X_i = x_i) = 1$, where x_i is the observed value of X on the ith trial. To B add the assumption that Y_i and V_i are statistically independent,

and that $P(Y_i = y_i) = 1$, where y_i is the observed value of Y on the ith trial. That is, the marginal distributions are specified in an *ad hoc* way to the effect that whatever values the hypothesized exogenous variables actually take, the (constructed) hypotheses entail that they take those values with probability 1. Such marginals are objectionable and useless in practice for a number of reasons, but we will leave them aside. Following Forster, call the resulting hypotheses A' and B'. They have the same likelihood.[4]

$$\begin{aligned} L(A') &= \prod_i P_{A'}(X_i = x_i, Y_i = y_i) \\ &= \prod_i P_{A'}(Y_i = y_i | X_i = x_i) P_{A'}(X_i = x_i) = (1/2)^{20} \\ L(B') &= \prod_i P_{B'}(X_i = x_i, Y_i = y_i) \\ &= \prod_i P_{B'}(X_i = x_i | Y_i = y_i) P_{B'}(Y_i = y_i) = (1/2)^{20} \end{aligned}$$

According to Forster, this constitutes a counterexample to the LTE because despite the equality of likelihoods, one can tell from the data that B' is false and A' is true. Here is his argument. B' entails that V_i and Y_i are independent (for every i):

$$P(V_i = 0.5 | Y_i = 3.5) = P(V_i = 0.5 | Y_i = 4.5) = 1/2$$

or equivalently, $P(X_i = 4 | Y_i = 3.5) = P(X_i = 5 | Y_i = 4.5) = 1/2$. Call this consequence B'_1. However, from the data we see that the relative frequency of $X = 4$ in the trials in which $Y = 3.5$ is 1, and the relative frequency of $X = 5$ in the trials in which $Y = 4.5$ is 0. Hence the data show that B'_1 is false, and so B' is false.

Formulated this way, the argument is clearly not contrastive, and seems akin to the *probabilistic modus tollens* that has been resolutely refuted by likelihoodists (Sober 2008, 51-53). It is not impossible, just very improbable, to obtain the data as they are even if

[4]Throughout the paper, we use upper case letters to denote variables and the corresponding lower case letters to denote values of the variables.

B_1' is true. A more charitable reading is that Forster did not literally mean that B_1' is shown to be false, but that the data overwhelmingly disconfirm B_1' relative to Not-B_1'. However, Not-B_1' is a complex class of hypotheses. Relative to some members in the class, the data are evidence against B_1', but relative to others, e.g. that $P(X_i = 4|Y_i = 3.5) = 0 \neq P(X_i = 5|Y_i = 4.5) = 1$, the data are arguably evidence for B_1'. In the absence of a well-grounded prior over these members, it is hard to make sense of the sweeping claim that the data seriously disconfirm B'-1 in favor of its logical negation.

Therefore, if we take the relativity of evidence seriously, the right way to state Forster's intuition is that the data provide evidence against B'-1 in reference to the given alternative A'. More accurately, the data disconfirm B_1' relative to A_1': $P(X_i = 4|Y_i = 3.5) = 1 \neq P(X_i = 5|Y_i = 4.5) = 0$, which is entailed by A'. Indeed, the evidence against B_1' versus A_1' is overwhelming, judging either intuitively or by a formal measure such as likelihood ratio.

But the fact that the data constitute weighty evidence against B_1' versus A_1' does not entail that the data are weighty evidence against B' versus A'. A_1' and B_1' are just parts of what A' and B' have to say about the data at hand; they are about the conditional probability of X_i given Y_i. But A' and B' also have implications for the marginal probability of Y_i. A' entails A_2': $P(Y_i = 3.5) = P(Y_i = 4.5) = 1/2$ (for every i), whereas B' entails B_2': $P(Y_i = y_i) = 1$, where y_i is the actual observed value of Y_i. How do the data bear on A_2' versus B_2'?

Essentially the same question was addressed very early on in Royall (1997)'s elaborate defense of likelihoodism. Shortly after he described the law of likelihood, he considered and refuted a putative counterexample (13-15). Suppose an ordinary-looking deck of 52 cards is well-shuffled. We turn over the top card and find it to be the ace of diamonds.

According to the law of likelihood, the observation supports the hypothesis that it is a trick deck consisting of 52 aces of diamonds against the hypothesis that the deck is ordinary. This may sound counterintuitive; intuitively the trick-deck hypothesis is not rendered more probable or believable than the ordinary-deck hypothesis based on the observation. But the evidential judgment is perfectly consistent with the intuition, for the question of credence is different from that of evidence. Even though the observation supports the trick-deck hypothesis against the ordinary-deck hypothesis, the former, in ordinary circumstances, is much less credible prior to the observation and may well end up less credible overall despite the positive evidence.

By the same token, for every trial in Forster's example, the observation of $Y_i = y_i$ supports the hypothesis that $P(Y_i = y_i) = 1$ against the hypothesis that $P(Y_i = 3.5) = P(Y_i = 4.5) = 1/2$, and overall the data favor B_2' over A_2'. Again, this evidential judgment should not be conflated with the judgment that the data render B_2' more credible than A_2'. In normal circumstances, there are a number of reasons to regard B_2' as much less plausible than A_2', prior to considering the evidence, and the evidential support may well be insufficient to overcome the initial implausibility.

The upshot is that the data are evidence for A_1' versus B_1', but also constitute evidence against A_2' versus B_2'. There is no compelling reason to think that the data *alone* favor the conjunction of A_1' and A_2' over that of B_1' and B_2' (or the other way around). The LTE, we conclude, is not threatened by the example.

3 Likelihood and consilience

Forster's positive insight, however, is not to be ignored. As he put it, Hypothesis B' suffers from a lack of "consilience". Given B', the probability distribution of the error

term V can be measured or estimated under two conditions: when $Y = 3.5$ and when $Y = 4.5$, but the two estimates do not "jump together": the empirical distribution of V estimated from the group of $Y = 3.5$ is very different from that of V estimated from the group of $Y = 4.5$. In contrast, Hypothesis A' does not have this problem (though in this case it does not display interesting consilience due to the lack of variation in X). We agree with Forster that this kind of consilience or lack thereof is evidentially significant, but we submit that the contrast is reflected in the comparison of likelihoods.

To show this, it helps to consider yet another version of the example Forster discussed. This version specifies the two hypotheses in the standard and perhaps most natural way, which assumes that X_i's and Y_i's are i.i.d. Under the i.i.d assumption, the best fitting marginal of X is $P(X = 4) = 1$. Add this marginal of X, together with the i.i.d assumption and that of the independence between X and U, to A, and call the resulting hypothesis A''. Similarly, the best fitting marginal of Y is $P(Y = 3.5) = P(Y = 4.5) = 1/2$. Add this marginal of Y, together with the i.i.d assumption and that of the independence between Y and V, to B, and call the resulting hypothesis B''. In this example, A'' happens to be the same as A', so $L(A'') = (1/2)^{20}$. But B'' is different from B', and has a much lower likelihood:

$$\begin{aligned} L(B'') &= \prod_i P_{B''}(X_i = x_i, Y_i = y_i) \\ &= \prod_i P_{B''}(X_i = x_i | Y_i = y_i) P_{B''}(Y_i = y_i) = (1/2)^{40} \end{aligned}$$

The difference in likelihoods accords well with the intuition that the data favor A'' over B''. But there is a "mystery" according to Forster. The likelihoods seem to differ just because of the difference in the parts contributed by the added marginals, but why should that matter? Intuitively, "the generation of the independent or exogenous variable [is] an inessential part of the causal hypothesis."

We share the latter intuition. In particular, we are sympathetic with the view that a defining feature of a *causal* relationship is that the relationship remains invariant under suitable interventions of an exogenous cause that change its marginal distribution (Woodward 2003). But it does not follow that marginals are irrelevant in causal inference. They are especially relevant to the kind of problems under discussion: distinguishing the direction of causal influence. In the present case, for example, X is hypothesized as the cause in only one of the hypotheses; in the other hypothesis it is modeled as the effect. The marginal distribution of X is relevant to judging, for example, how well the other hypothesis, by treating X as endogenous, fits the observations on X, compared to the former hypothesis that treats X as exogenous.

The right explanation in our view of the difference between the likelihoods actually agrees nicely with Foster's consideration of consilience. Notice that the lack of consilience under B highlighted by Forster corresponds to the statistical dependence of V on Y as shown in the data. A convenient measure of statistical dependence between random variables is known as *mutual information* (Cover and Thomas 1991, 18). The mutual information between two random variables Z and W is defined as:

$$\mathbf{I}(Z, W) = \mathbb{E} \log \frac{P(Z, W)}{P(Z)P(W)} = \mathbb{E}(\log P(Z, W) - \log P(Z) - \log P(W))$$

where $\mathbb{E}(\cdot)$ denotes the expectation (with respect to $P(Z, W)$). The mutual information is a way to measure the difference between the joint distribution $P(Z, W)$ and the product of the marginals $P(Z)P(W)$, and hence a way to measure the statistical dependence between Z and W. It is non-negative and is equal to zero just in case Z and W are independent.

Given i.i.d samples from the joint distribution $P(Z, W)$, we can use the following sample approximation of the mutual information, by replacing the expectation with the

sample mean, and P with the corresponding empirical distribution $\hat{P}$ (where probabilities are estimated by sample frequencies):

$$\hat{\mathbf{I}}(Z, W) = \frac{1}{n}\sum_i (\log \hat{P}(z_i, w_i) - \log \hat{P}(z_i) - \log \hat{P}(w_i))$$

where n is the sample size. This provides a measure of dependence as shown in the samples. Accordingly, in Forster's example, $\hat{\mathbf{I}}(X, U)$ and $\hat{\mathbf{I}}(Y, V)$ can be regarded as plausible measures of the lack of consilience in A'' and B'', respectively.

Some calculations are in order. Given the data in Forster's example, the empirical joint distribution of X and Y puts half of the mass on $\langle X = 4, Y = 3.5\rangle$ and half of the mass on $\langle X = 4, Y = 4.5\rangle$. It follows that the empirical joint distribution of X and $U = Y - X$ puts half of the mass on $\langle X = 4, U = -0.5\rangle$ and half of the mass on $\langle X = 4, U = 0.5\rangle$; and the empirical joint distribution of Y and $V = X - Y$ puts half of the mass on $\langle Y = 3.5, U = -0.5\rangle$ and half of the mass on $\langle Y = 4.5, U = 0.5\rangle$. From these it is easy to calculate, taking 2 as the base of the logarithm to simplify the numbers, that $\hat{\mathbf{I}}(X, U) = 0$ and $\hat{\mathbf{I}}(Y, V) = 1$. These, we repeat, are plausible measures of the lack of consilience in Forster's sense.

Consider now the log-likelihoods of A'' and B'', taking again 2 as the base of the logarithm: $l(A'') = -20$ and $l(B'') = -40$. The difference between them per datum is $20/20 = 1$, which is precisely the difference between $\hat{\mathbf{I}}(Y, V)$ and $\hat{\mathbf{I}}(X, U)$.

This is not a numerical accident. The log-likelihood of A'' can be written as:

$$l(A'') = \sum_i \log P_{A''}(x_i, y_i) = \sum_i \log P_{A''}(x_i, u_i) = \sum_i (\log P_{A''}(x_i) + \log P_{A''}(u_i))$$

Because for each i, $\langle X_i = x_i, Y_i = y_i\rangle$ and $\langle X_i = x_i, U_i = u_i = y_i - x_i\rangle$ are descriptions of the same event. Since the marginals in A'' are specified as the corresponding empirical

distributions — $P_{A''}(X) = \hat{P}(X)$ and $P_{A''}(U) = \hat{P}(U)$ — we have

$$\begin{aligned} l(A'') &= \sum_i (\log \hat{P}(x_i) + \log \hat{P}(u_i)) \\ &= \sum_i \log \hat{P}(x_i, u_i) - \sum_i (\log \hat{P}(x_i, u_i) - \log \hat{P}(x_i) - \log \hat{P}(u_i)) \\ &= \sum_i \log \hat{P}(x_i, u_i) - n\hat{\mathbf{I}}(X, U) \end{aligned}$$

Similarly,

$$l(B'') = \sum_i \log \hat{P}(y_i, v_i) - n\hat{\mathbf{I}}(Y, V)$$

Note further that for every i, $\hat{P}(x_i, u_i) = \hat{P}(y_i, v_i) = \hat{P}(x_i, y_i)$. Hence,

$$l(A'') - l(B'') = n(\hat{\mathbf{I}}(Y, V) - \hat{\mathbf{I}}(X, U))$$

Therefore, there is here a systematic connection between the comparison of likelihoods and the comparison of how hypotheses fare in terms of "consilience of inductions" highlighted by Forster. The evidential significance of consilience, or at least one plausible interpretation of it, is not beyond the grip of the LTE.

4 Conclusion

Whether or not the LTE can survive other challenges, Forster's examples, we conclude, are not convincing counterexamples. We have only examined one of his examples in this paper, but the other example, set up in linear models with continuous variables, employs parallel devices and arguments, to which our points in section 2, *mutatis mutandis*, carry over. The apparent force of his examples hinges on an (implicit) denial of the basic tenet of likelihoodism, i.e., the thesis of the relativity of evidence. Since the denial is based on no argument but dubious intuitions that have been forcefully criticized by likelihoodists, Forster's criticism is at best question begging.

More interestingly, we showed a way to vindicate Forster's preferred consilience-based methodology within the framework of the LTE, by establishing a systematic connection between likelihood and (one plausible interpretation of) consilience. The connection holds much more generally for such causal models than we can show in this paper (Hyvärinen and Smith 2013; Zhang et al. forthcoming). In particular, Hyvärinen and Smith (2013, 115) presented a similar result on linear models, which is applicable to Forster's example with continuous variables. Whether similar connections exist in contexts other than this sort of causal inference problems is worth exploring.

References

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York: John Wiley and Sons.

Forster, M. (1998). Unification, explanation, and the composition of causes in Newtonian mechanics. *Studies in the History and Philosophy of Science*, **19**: 55–101.

Forster, M. (2006). Counterexamples to a likelihood theory of evidence. *Minds and Machines*, **16**(3): 319–338.

Hacking, I. (1865). *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, **14**: 111–152.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, Fla.: Chapman and Hall.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Whewell, W. (1858). *Novum Organon Renovatum*. London: John W. Parker.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford and New York: Oxford University Press.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. (forthcoming). On estimation of functional causal models: Post-nonlinear causal model as an example. *ACM Transactions on Intelligent Systems and Technologies*.