On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection

Kun Zhang[*] kunz1@cmu.edu
Jiji Zhang[♯] jijizhang@ln.edu.hk
Biwei Huang[‡*] biweih@andrew.cmu.edu
Bernhard Schölkopf[‡] bs@tuebingen.mpg.de
Clark Glymour[*] cg09@andrew.cmu.edu

[*]Department of philosophy, Carnegie Mellon University
[♯]Lingnan University, Hong Kong
[‡]Max Planck Institute for Intelligent Systems, Germany

Abstract

We study the identifiability and estimation of functional causal models under selection bias, with a focus on the situation where the selection depends solely on the effect variable, which is known as outcome-dependent selection. We address two questions of identifiability: the identifiability of the causal direction between two variables in the presence of selection bias, and, given the causal direction, the identifiability of the model with outcome-dependent selection. Regarding the first, we show that in the framework of post-nonlinear causal models, once outcome-dependent selection is properly modeled, the causal direction between two variables is generically identifiable; regarding the second, we identify some mild conditions under which an additive noise causal model with outcome-dependent selection is to a large extent identifiable. We also propose two methods for estimating an additive noise model from data that are generated with outcome-dependent selection.

1 Introduction

Selection bias is an important issue in statistical inference. Ideally, samples should be drawn randomly from the population of interest. In reality, however, it is commonplace that the probability of including a unit in the sample depends on some attributes of the unit. Such selection bias, if not corrected, often distorts the results of statistical analysis. For example, it is well known that in a regression analysis, if there is selection on the dependent variable, the ordinary least squares estimation of the regression coefficients will be biased and inconsistent (Heckman, 1979). The challenge is even bigger in causal inference; both the task of learning causal structures from data and the task of estimating causal mechanisms or parameters given a causal structure are usually rendered more difficult by the presence of selection bias.

In this paper, we are concerned with the approach to causal inference based on (restricted) functional causal models (Shimizu et al., 2006; Hoyer et al., 2009; Zhang & Hyvärinen, 2009), and aim to investigate the extent to which selection bias can be handled within this approach. Specifically, we mainly focus on the outcome-dependent selection bias, where the selection mechanism depends only on the effect, and are interested in the following two questions:

- Is the causal direction between two random variables identifiable in the presence of selection bias?
- Is the causal mechanism as represented by a functional causal model identifiable in the presence of selection bias?

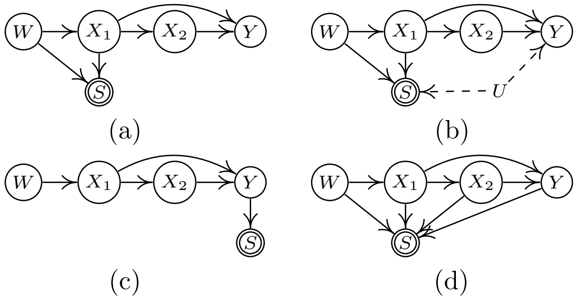
These two questions have to do with the two main aspects of causal inference, respectively. The former question is about the inference of causal structure. In the traditional conditional-independence-constraint-based approach to learning causal structures (Spirtes et al., 2001; Pearl, 2000), some methods have been developed to handle selection bias (Spirtes et al., 1999; Zhang, 2008; Borboudakis & Tsamardinos, 2015). However, the structural information that can be learned via the constraint-based approach is typically limited to a Markov equivalence class. In particular, the approach cannot distinguish cause from effect with just two variables. In contrast, a distinctive virtue of the approach based on functional causal models is that Markov equivalent causal structures can usually be distinguished. In particular, the direction between two random variables is generically identifiable. Whether this virtue survives the challenge posed by selection bias is therefore worth investigating.

The latter question is related to the inference of causal parameters (i.e., parameters or quantities that have causal interpretations), including intervention effects. In addition to the work on various selection models in econometrics and social science (Heckman, 1979; Winship & Mare, 1992), recent literature has seen interesting work on the recoverability of causal parameters based on graphical models (Didelez et al., 2010; Bareinboim & Pearl, 2012; Bareinboim et al., 2014; Evans & Didelez, 2015). Much of this work, however, deals with linear models or discrete variables, whereas we are concerned in this paper with continuous variables that may bear a nonlinear relationship.

We will proceed as follows. In Section 2, we introduce the general setup and briefly discuss several types of selection, before focusing our attention on the situation where the selection depends on the effect variable, known as outcome-dependent selection. In Section 3, we show that in the framework of post-nonlinear causal models, once outcome-dependent selection is properly modeled, the causal direction between two variables is generically identifiable. In Section 4, we identify some mild conditions under which an additive noise causal model with outcome-dependent selection is to a large extent identifiable. We then propose, in Section 5, two methods for estimating an additive noise model from data that are generated with outcome-dependent selection. Some experiments are reported in Section 6.

2 Outcome-Dependent Selection Bias

A common way to represent selection bias is to use a binary selection variable S encoding whether or not a unit is included in the sample. Suppose we are interested in the relationship between X and Y, where X has a causal influence on Y. Let p_{XY} denote the joint distribution of X and Y in the population. The selected sample follows $p_{XY|S=1}$ instead of p_{XY}. In general, $p_{XY|S=1} \neq p_{XY}$, and that is how selection may distort statistical and causal inference. However, different kinds of selection engender different levels of difficulty. In general, S may depend on any number of substantive variables, as illustrated in Figure 1, where $X = (X_1, X_2)$. [1]



(a) (b)

(c) (d)

Figure 1: Illustration of different situations with sample selection bias. (a) S depends on $X = (X_1, X_2)$ but not on Y. (b) S depends on X and is also statistically dependent on Y given X due to a confounder U. (c) S directly depends solely on Y (outcome-dependent selection). (d) S depends on both X and Y.

Selection Bias on the Cause For the purpose of causal inference, the least problematic kind of situation is depicted in Figure 1(a), in which S is independent of the effect variable Y given the cause variable X. It follows that $p_{Y|X,S=1} = p_{Y|X}$. That is, the selection bias does not distort the conditional distribution of the effect Y given the cause X or the structural equation model for the causal process. In such a situation, causal inference can essentially proceed as usual. However, if there is a (latent) confounder for Y and S, as illustrated in Figure 1(b), S and Y are not conditionally independent given X any more, that is, $p_{Y|X,S=1} \neq p_{Y|X}$. Such a distortion may be corrected under rather restrictive assumptions; see, e.g., Heckman's correction (Heckman, 1979).

Selection Bias on the Effect If the selection depends solely on the effect, as depicted in Figure 1(c), then $p_{Y|X,S=1} \neq p_{Y|X}$, and the selection bias, if not corrected, will mislead inference. Consider, for example, a standard assumption in functional causal modeling that the effect Y is a function of the cause variable X and an noise variable E that is independent of X. Suppose this assumption holds in the population. With the outcome-dependent selection, X and E are typically not independent in the selected sample, as they are typically not independent conditional on S (which is a descendant of a collider between X and E, i.e., Y). Furthermore, even if one fits a regression model on selected sample, the estimated residual (which is usually different from the true noise term in the causal process) is usually not independent from X; we will get back to this issue in Section 4.1.

This kind of selection is known as *outcome-dependent selection bias* (OSB) (Didelez et al., 2010; Bareinboim et al., 2014), and will be our focus in this paper. We will show that although outcome-dependent selection seriously complicates analysis, it can be handled in the identification and estimation of functional causal

[1] In this paper, we assume that we only know which variables the selection variable S depends on, but the selection mechanism is unknown, i.e., the probability of $S = 1$ given those variables is unknown. Notice that we do not have access to the data points that were not selected. This is very different from Heckman's framework to correct the bias caused by a censored sample (Heckman, 1979), which assumes access to an i.i.d. sample from the whole population, on which the Y values are observable only for the data points that satisfy the selection criterion (implied by the selection equation), but other attributes of the "censored" points are still available, enabling one to directly identify the selection mechanism.

models. Note that in the case of outcome-dependent selection, X is independent of S given Y, and so we can model the distribution of the observed sample as:

$$p^{\beta}_{XY} \triangleq p_{XY|S=1} = \frac{p_{X,Y,S=1}}{P(S=1)} = p_{XY} \cdot \frac{P(S=1|X,Y)}{P(S=1)} \\ = p_{XY} \cdot \frac{P(S=1|Y)}{P(S=1)} = \beta(y) p_{XY}, \qquad (1)$$

where the nonnegative function $\beta(y) \triangleq P(S = 1|Y)/P(S=1)$ is a density ratio for biased sampling that only depends on Y. We will adopt this representation of outcome-dependent selection in what follows.

Selection Bias on Both the Cause and the Effect An even more general situation is depicted in Figure 1(d), where the selection depends on both X and Y (and probably others). In such a situation, the density ratio function β will depend on both X and Y. The selected sample follows the distribution $p^{\beta}_{XY} \propto p_{XY}\beta(x,y,w)$. Roughly speaking, the selection procedure is so flexible that without further constraints on $\beta(x,y,w)$, we cannot see much information about the population p_{XY}: if p_{XY} is positive on $(-\infty,+\infty)$, the same p^{β}_{XY} can be generated from a large class of distributions p_{XY} with a suitably chosen $\beta(x,y,w)$. Moreover, the causal direction is generally not identifiable, for with a sufficiently flexible $\beta(x,y,w)$, either direction can be made compatible with whatever distribution. Interestingly, when β depends only on Y, as is the case under outcome-dependent selection, the causal direction according to a restricted functional causal model is still generically identifiable, without any substantial restriction on β. To this result we now turn.

3 Identifiability of Causal Direction

In this section we investigate whether it is possible to successfully recover the causal direction between two variables when the data are generated according to a functional causal model, but with outcome-dependent selection. Here we assume that both X and Y are scalar variables.

3.1 Identifiability Without Selection Bias

The traditional approaches to inferring causal structure from data, such as the constraint-based approach (Spirtes et al., 2001; Pearl, 2000) and the score-based approach (Chickering, 2002; Heckerman et al., 1995) cannot distinguish Markov equivalent causal structures without background knowledge. In particular, with only two variables, those methods cannot distinguish cause from effect. The more recent approach based on restricted functional causal models is usually more powerful in this respect. In a functional causal model, the effect is taken to be a function of the direct causes together with an noise term that is independent of the direct causes (Pearl, 2000). When the class of functions is constrained, the causal direction is usually identifiable in that only one direction can satisfy the model assumptions, such as the assumed independence between the noise term and the direct causes. Available identifiability results include those on linear, non-Gaussian, acyclic Model (LiNGAM) (Shimizu et al., 2006)), additive noise model (ANM) (Hoyer et al., 2009), and post-nonlinear (PNL) causal model (Zhang & Hyvärinen, 2009). In this section, we will establish a main result for the PNL causal model. The result also applies to linear models and additive noise models, as they are special cases of PNL models.

A PNL model for $X \to Y$ is specified as follows:

$$Y = f_2(f_1(X) + E), \qquad (2)$$

where X and E are statistically independent, f_1 is a non-constant smooth function, f_2 is an invertible smooth function, and $f_2' \neq 0$. This model is sufficiently flexible to represent or approximate many causal processes in reality (Zhang & Hyvärinen, 2009).

Similarly, for the reverse direction $Y \to X$, a PNL model would take the following form:

$$X = g_2(g_1(Y) + \tilde{E}), \qquad (3)$$

where Y and $\tilde{E}$ are independent, g_1 is non-constant and smooth, g_2 is invertible and smooth, and $g_2' \neq 0$.

As shown in (Zhang & Hyvärinen, 2009), (2) and (3) can generate the same distribution of X and Y only for very special configurations of the functions and distributions. In generic cases, if data are generated according to a model of form (2), there is no model of form (3) that generates the same distribution. Hence the causal direction is generically identifiable.

3.2 Identifiability of Causal Direction in PNL-OSB

We now show that the generic identifiability of causal direction based on PNL models still holds even if we allow the possibilty of outcome-dependent selection.

Suppose the data distribution is generated by a PNL causal model from X to Y in the form of (2), denoted by $\mathcal{F}_{\rightarrow}$, followed by an outcome-dependent selection with an density ratio $\beta(y)$, as in (1). Call $(\mathcal{F}_{\rightarrow}, \beta(y))$ a PNL-OSB model, and let $p^{\rightarrow}_{XY}$ denote the joint density of X and Y resulting from $(\mathcal{F}_{\rightarrow}, \beta(y))$. We are interested in whether there is a PNL-OSB model in the reverse direction that can generate the same data distribution. That is, consider $(\mathcal{F}_{\leftarrow}, v(x))$, where $\mathcal{F}_{\leftarrow}$

is a PNL causal model from Y to X in the form of (3), and $v(x)$ is an density ratio function that depends on X. Let $p^{\leftarrow}_{XY}$ denote the joint density of X and Y resulting from $(\mathcal{F}_{\leftarrow}, v(x))$. When is it the case that $p^{\rightarrow}_{XY} = p^{\leftarrow}_{XY}$?

To simplify the presentation, we define random variables $T \triangleq g_2^{-1}(X)$, $Z \triangleq f_2^{-1}(Y)$, and function $h \triangleq f_1 \circ g_2$. That is, $h(t) = f_1(g_2(t)) = f_1(x)$. Similarly, $h_1 \triangleq g_1 \circ f_2$ is a function of Z. Moreover, we let $\eta_1(t) \triangleq \log p_T(t) = \log p_X(x) + \log |g_2'(t)|$, and $\eta_2(e) \triangleq \log p_E(e)$.

Note that T and E are independent (for X and E are assumed to be independent), and Z and $\tilde{E}$ are independent (for Y and $\tilde{E}$ are assumed to be independent). It follows that

$$p^{\rightarrow}_{XY} = \beta(y) p^{\mathcal{F}_{\rightarrow}}_{XY} = \beta(y) p_{XE}/|f_2'| = \beta_{f_2}(z) p_T p_E / |f_2' g_2'|,$$
$$p^{\leftarrow}_{XY} = v(x) p^{\mathcal{F}_{\leftarrow}}_{XY} = v(x) p_{Y\tilde{E}}/|g_2'| = v_{g_2}(t) p_{Z\tilde{E}} / |f_2' g_2'|,$$

where $\beta_{f_2} = \beta \circ f_2$, and $v_{g_2} = v \circ g_2$.

Now suppose

$$p^{\rightarrow}_{XY} = p^{\leftarrow}_{XY} \tag{4}$$

This implies

$$p_{Z\tilde{E}} = \frac{\beta_{f_2}(z)}{v_{g_2}(t)} p_T p_E,$$

or equivalently

$$\begin{aligned} \log p_{Z\tilde{E}} &= \log \beta_{f_2}(z) - \log v_{g_2}(t) + \log p_T + \log p_E \\ &= \log \beta_{f_2}(z) + \tilde{\eta}_1(t) + \eta_2(e), \end{aligned} \tag{5}$$

where $\tilde{\eta}_1(t) \triangleq \log p_T - \log v_{g_2}(t) = \eta_1(t) - \log v_{g_2}(t)$. Since Z and $\tilde{E}$ are independent, we have

$$\frac{\partial^2 \log p_{Z\tilde{E}}}{\partial z \partial \tilde{e}} \equiv 0. \tag{6}$$

(5) and (6) entail very strong constraints on the distribution of E, as stated in the following theorem.

Theorem 1 *Suppose that the densities of E and T and the functions f_1, f_2, g_1, g_2, and $v(x)$ are third-order differentiable and that p_E is positive on $(-\infty, +\infty)$. The condition (4) implies that for every point of (X, Y) satisfying $\eta_2'' h' \neq 0$:*

$$\begin{aligned} \tilde{\eta}_1''' - \frac{\tilde{\eta}_1'' h''}{h'} \quad = \quad & \Big(\frac{\eta_2' \eta_2'''}{\eta_2''} - 2\eta_2''\Big) \cdot h' h'' - \frac{\eta_2'''}{\eta_2''} \cdot h' \tilde{\eta}_1'' \\ & + \eta_2' \cdot \Big(h''' - \frac{h''^2}{h'}\Big), \end{aligned} \tag{7}$$

and h_1 depends on $\tilde{\eta}_1$, η_2, and h in the following way:

$$\frac{1}{h_1'} = \frac{\tilde{\eta}_1'' + \eta_2'' h'^2 - \eta_2' h''}{\eta_2'' h'}. \tag{8}$$

Further assume that $\eta_2'' h' \neq 0$ almost everywhere. Then in order for (7) to hold, p_E and h must satisfy one of the five conditions listed in Table 1.

Table 1: All situations in which the causal direction implied by the PNL-OSB model may be unidentifiable.

	p_E	$h = f_1 \circ g_2$
1	Gaussian	linear
2	log-mix-lin-exp	linear
3	log-mix-lin-exp	h strictly monotonic, and $h' \to 0$, as $t_1 \to +\infty$ or as $t_1 \to -\infty$
4	generalized mixture of two exponentials	Same as above

All proofs are given in the Supplementary material. In the five situations given in Table 1, the causal direction may not be identifiable according to the PNL-OSB model, and the involved distribution p_E is very specific. For the definition of distributions of the form `log-mix-lin-exp` or `generalized mixture of two exponentials`, see (Zhang & Hyvärinen, 2009). As a consequence, generally speaking, the causal direction implied by PNL-OSB is identifiable.

This identifiability result regarding the causal direction implied by PNL-OSB is similar to the original result on PNL, which was given in (Zhang & Hyvärinen, 2009). The difference is that $\eta_1(t) = \log p_T(t)$ in the original identifiability result on PNL is replaced by $\tilde{\eta}_1(t) = \log \frac{p_T(t)}{v_{g_2}(t)}$. Recall that $v_{g_2}(t)$ can be any valid density ratio; if $p_T(t)$ is positive on $(-\infty, +\infty)$, one can always adjust $v_{g_2}(t)$ so that $\frac{p_T(t)}{v_{g_2}(t)}$ meets the constraint on η_1 in (Zhang & Hyvärinen, 2009). That is, in our result any $p_T(t)$ that is positive on $(-\infty, +\infty)$ is allowed. Therefore, our non-identifiable situations (Table 1) do not contain any constraints on p_T, but still have very strong constraints on P_E and $h = f_1 \circ g_2$.

4 Identifiability of ANM-OSB Model

Given the causal direction, a further important question is whether the causal mechanism, represented by the functional causal model, and the selection procedure, represented by $\beta(y)$, can be recovered from data.

For simplicity of the derivation and presentation, we shall consider the ANM for the causal mechanism (not a PNL one in this section):

$$Y = f^{AN}(X) + E, \tag{9}$$

where $E \perp\!\!\!\perp X$. Here we further assume that f^{AN} is smooth. The observed data are generated by applying the selection bias on Y, i.e., they were drawn from the distribution

$$p^{\beta}_{XY} = \beta(y) p^{\mathcal{F}}_X p^{\mathcal{F}}_{Y|X}, \tag{10}$$

where $p^{\mathcal{F}}_{Y|X}$ is specified by the causal model (9) and $p^{\mathcal{F}}_X$ denotes the distribution of X before applying the selection procedure. Note that generally speaking, $p_X^{\mathcal{F}}$ is not identical to p_X^{β}. Call the model $(\mathcal{F}, \beta(y))$ an ANM-OSB model.

Suppose that the observed data are generated from an ANM-OSB $(\mathcal{F}_1, \beta_1(y))$. We are interested in whether another ANM-OSB $(\mathcal{F}_2, \beta_2(y))$ can generate the same data distribution. Suppose it does. The observed data distribution is then

$$p_{XY}^{\beta} = p_{XY}^{\mathcal{F}_1}\beta_1(y) = p_{XY}^{\mathcal{F}_2}\beta_2(y). \tag{11}$$

Let $\beta_r(y) \triangleq \frac{\beta_2(y)}{\beta_1(y)}$. Bear in mind that $p_{XY}^{\mathcal{F}_1} = p_X^{(1)} p_{E_1}(Y - f^{(1)}(X))$ and $p_{XY}^{\mathcal{F}_2} = p_X^{(2)} p_{E_2}(Y - f^{(2)}(X))$. If (11) holds, we have

$$\beta_r^{-1}(y) p_X^{(1)}(x) p_{E_1}(e_1) = p_X^{(2)}(x) p_{E_2}(e_2). \tag{12}$$

Taking the logarithm of both sides gives

$$-\log\beta_r(y) + \log p_X^{(1)} + \log p_{E_1}(e_1) = \log p_X^{(2)} + \log p_{E_2}(e_2). \tag{13}$$

Now let us see whether it is possible for (13) to hold and, if yes, what constraints the functions $\beta_r(y)$, $\log p_X^{(1)}$, and $\log p_{E_1}(e_1)$ must satisfy. Denote by $J_{AN} \triangleq \log p_X^{(2)} + \log p_{E_2}(e_2)$. As seen from the RHS, (13) implies

$$\frac{\partial^2 J_{AN}}{\partial x \partial e_2} \equiv 0. \tag{14}$$

Let $l_\beta(y) = \log \beta_r(y)$, $\eta_{X_1}(x) \triangleq \log p_X^{(1)}$ and $\eta_{E_1}(e_1) \triangleq \log p_{E_1}(e_1)$. By solving (14), we can establish the relationship between the two ANM-OSB models.

4.1 General Results

Interestingly, as stated in the following theorem, if the noise E_1 is non-Gaussian, then $f^{(2)}(x)$ must be a shifted version of $f^{(1)}(x)$; in other words, the underlying function f^{AN} is identifiable up to a constant. Furthermore, if E_1 is non-Gaussian, the selection weight $\beta(y)$ can be recovered up to a factor which is an exponential function of y, i.e., $\beta_2(y) \propto \beta_2(y) \cdot e^{c_2 y}$, where c_2 is a constant; accordingly, $p_{E_2} \propto p_{E_1} \cdot e^{-c_2 e_1}$.

Theorem 2 *Let Assumptions A_1 and A_2 hold true:*
A_1. $p_X^{(1)}$ and p_{E_1} are positive on $(-\infty, +\infty)$.
A_2. $\eta''_{E_1}(e_1) f^{(1)'}(x) = 0$ only at finite points.[2]
Then if (11) is true, the following statements hold.

a) If E_1 is not Gaussian, then $f^{(2)}(x) = f^{(1)}(x) + c_1$, and $\beta_2(y) = \beta_1(y)\beta_r(y)$, where $\beta_r(y) = e^{c_2 y + d_1} = e^{d_1} \cdot e^{c_2 f^{(1)}(x)} \cdot e^{c_2 e_1}$. Accordingly, $p_X^{(2)} \propto p_X^{(1)} \cdot e^{-c_2 f_1(x)}$, and $p_{E_2} \propto p_{E_1} \cdot e^{-c_2 e_1}$. Here c_1, c_2, and d_1 are constants, and d_1 guarantees that $\beta_2(y)$ is a valid density ratio.

That is, $f^{(2)}(x)$ is equal to $f^{(1)}(x)$ (up to a constant), and with proper scaling, $\beta_2(y)$ equals $\beta_1(y)$ times an exponential function of y.

b) If E_1 is Gaussian, then $\beta_2(y) = \beta_1(y)\beta_r(y)$, where $\beta_r(y) = e^{\frac{-ab}{2}y^2 + c_4 y + d_4}$, and $f^{(2)}(x) = \frac{1}{1+b} f^{(1)}(x) + d_3$. Here a, b, c_4, d_3, and d_4 are constants, d_4 guarantees that $\beta_2(y)$ is a valid density ratio, and $a \neq 0$.

That is, with proper scaling, $\beta_2(y)$ equals $\beta_1(y)$ times a Gaussian function of y (which includes the exponential function of y as a special case by setting $b = 0$).

An interesting implication of Theorem 2 is that generally speaking, fitting an ordinary ANM on the data that were generated by an ANM-OSB will not produce an independent error term. That is, under mild assumptions, if one sets $\beta_2(y) \equiv 1$, $(\mathcal{F}_2, \beta_2(y))$ cannot produce the same distribution over (X, Y) as $(\mathcal{F}_1, \beta_1(y))$ does, as is given in the following corollary.

Corollary 3 *Let Assumptions A1 and A2 hold. Then under either of the following conditions, there does not exist an ANM, specified by (9), to generate the same distribution over (X, Y) as $(\mathcal{F}_1, \beta_1(y))$ does.*

a) E_1 is not Gaussian and $\beta_1(y)$ is not proportional to $e^{c'y}$ for any c'.

b) E_1 is Gaussian and $\beta_1(y)$ is not proportional to $e^{a'y^2 + c'y}$ for any a' and c' (i.e., $\beta_1(y)$ is not proportional to an exponential function of any polynomial of y of degree 1 or 2).

4.2 When the Noise is Gaussian

When p_{E_1} is Gaussian, as stated in b) of Theorem 2, the function f^{AN} is not identifiable any more: $f^{(2)}(x)$ and $f^{(1)}(x)$can differ by an affine transformation (not simply a shift). Accordingly, $\beta_2(y)$ can differ from $\beta_1(y)$ in a Gaussian function of y. Compared to the case where E_1 is non-Gaussian, the Gaussian case suffers from more indeterminacies because the product of two Gaussian functions is still a Gaussian function. In particular, since $\beta_r^{-1}(y)$ and $p_{E_1}(y - f^{(1)}(x))$ (or $p_{Y|X}^{\mathcal{F}_1}(y|x)$) are both Gaussian functions in y, their product is still a Gaussian function in y. Accordingly, (12) will hold true, by setting p_{E_2} to another appropriate Gaussian density; in other words, in this case two additive noise models $\mathcal{F}_1$ and $\mathcal{F}_2$ can both generate

[2] This excludes the special case where $f^{(1)'} \equiv 0$, i.e., where X and Y are independent; in this case clearly the selection procedure is not identifiable.

the same observed data, by applying the bias selection procedures $\beta_1(y)$ and $\beta_2(y) = \beta_1(y)\beta_r(y)$, respectively.

More specifically, we can derive the function $f^{(2)(x)}$ and noise distribution $p_{E_2}(e_2)$ for the model $\mathcal{F}_2$. As shown above, $f^{(2)(x)} = \frac{1}{1+b}f^{(1)}(x) + d_3$. Eq. 12, combined with (25) and (26), implies that $p_{E_2}(e_2) \propto e^{\frac{a}{2}e_2^2 + \frac{ab}{2}e_2^2} = e^{\frac{a(1+b)}{2}e_2^2}$, while $p_{E_1}(e_1) \propto e^{\frac{a}{2}e_1^2}$.

Figure 2 gives an illustration of this result. Notice that the identifiability results imply some constraints on $\beta_r(y) = \beta_2(y)/\beta_1(y)$, not on $\beta_1(y)$, so without loss of generality, we set $\beta_1(y) \equiv 1$, leading to $\beta_2(y) = \beta_r(y)$. The circles denote the data points generated by applying the density ratio $\beta_1(y) \equiv 1$ on additive noise model $\mathcal{F}_1$: the dash line shows the nonlinear function $f^{(1)}(x)$, and the red solid line shows the shape of p_{E_1}. In contrast, additive noise model $\mathcal{F}_2$ uses nonlinear function $f^{(2)}(x)$, which is different from $f^{(1)}(x)$, and its noise variance is slightly larger than that in $\mathcal{F}_1$. The crosses denote the data points generated by $\mathcal{F}_2$. Although $\mathcal{F}_1$ and $\mathcal{F}_2$ are not the same in this case, applying the density ratio function $\beta_r(y)$ on $p^{\mathcal{F}_2}_{XY}$ gives the same joint distribution as $p^{\mathcal{F}_1}_{XY}$, i.e., $\beta_1(y)p^{\mathcal{F}_1}_{XY} = p^{\mathcal{F}_1}_{XY} = \beta_r(y)p^{\mathcal{F}_2}_{XY} = \beta_2(y)p^{\mathcal{F}_2}_{XY}$.

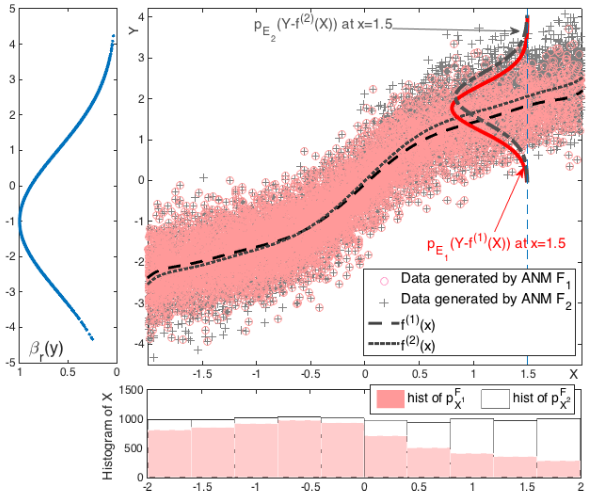


Figure 2: Illustration of the non-identifiability of the additive noise model, especially f^{AN}, when the noise is Gaussian. Red circles denote data points generated by the ANM $\mathcal{F}_1$, or by the ANM-OSB $(\mathcal{F}_1, \beta_1(y) \equiv 1)$. The gray crosses denote data generated by the ANM $\mathcal{F}_2$. The two ANM-OSB models, $(\mathcal{F}_2, \beta_r(y))$ and $(\mathcal{F}_1, 1)$, produce the same distribution of (X, Y).

4.3 With Further Constraints

Not surprisingly, if we have more knowledge about the noise distribution p_E or the density ratio function $\beta(y)$, the ANM model, including the function, the noise distribution, and the density ratio, can be fully identifiable. Below is an example showing that this is the case if we know that p_E is symmetric and non-Gaussian.

Corollary 4 *Let the assumptions made in Theorem 2 A_1 and A_2 hold. Suppose E_1 is not Gaussian. Then If both p_{E_1} and p_{E_2} are symmetric about the origin, then $f^{(2)}(x) = f^{(1)}(x)$, $E_1 = E_2$, $p_{E_1}(e_1) = p_{E_2}(e_2)$, and $\beta_r(y) \equiv 1$, i.e., $\beta_1(y) = \beta_2(y)$.*

5 Estimation of ANM-OSB

Eq. 10 gives the distribution for the observed data. In theory, we can then estimate the parameters involved in $\beta(y)$, $p^{\mathcal{F}}_{Y|X}$, as well as p_X, by maximum likelihood. However, when using maximum likelihood, we have to guarantee that the quantity on the right hand side of (10) is a valid density. This constraint is notoriously difficult to enforce in the optimization procedure. Below we propose two methods to estimate the underlying additive noise model and $\beta(y)$; one is maximum likelihood with the above constraint enforced approximately, and the other makes use of the score matching technique.

5.1 Maximum Likelihood Estimation with a Sample Approximation Constraint

To estimate the involved functions $\beta(y)$, $p^{\mathcal{F}}_X$, and $p^{\mathcal{F}}_{Y|X}$, we can maximize the data likelihood:

$$\begin{aligned}\mathcal{L} &= \sum_{k=1}^{n} \log p^{\beta}_{XY}(x_k, y_k)\\ &= \sum_{k=1}^{n} \left[\log \beta(y_k) + \log p^{\mathcal{F}}_X(x_k) + \log p^{\mathcal{F}}_{Y|X}(y_k|x_k)\right]. \quad (15)\end{aligned}$$

According to the theory shown in Section 4, the solution to βy, $p^{\mathcal{F}}_X$, and $p^{\mathcal{F}}_{Y|X}$ suffers from some indeterminacies, e.g., the solution to βy may differ from the true one by an exponential transformation. To find the solution for which the biased selection procedure is as weak as possible, we regularize the likelihood function with the constraint that $\log \beta(y)$ is close to 0. That is, we maximize

$$\mathcal{L}^r = \mathcal{L} - \lambda_r \sum_{k=1}^{n} \sqrt{(\log \beta(y_k))^2 + r}, \quad (16)$$

where the regularization parameter λ_r was set to 10^{-3} in our experiments, and r is a small positive number and was set to 0.02.

Now we have two issues to consider. One is how to parameterize the involved functions. The other is how to enforce that p^{β}_{XY}, specified in (10), corresponds to a valid density. More specifically, the constraint is

$$\beta(y)p_X^{\mathcal{F}}p_{Y|X}^{\mathcal{F}} > 0, \text{ or equivalently } \beta(y) > 0, \text{ and} \tag{17}$$

$$\int \beta(y)p_X^{\mathcal{F}}p_{Y|X}^{\mathcal{F}}dxdy = 1. \tag{18}$$

Without constraint (18), the scale of p_{XY}^{β} will go to infinity during the process of maximizing (15).

Parameterization The additive noise model for the data-generating process, (9), implies that $p_{Y|X}^{\mathcal{F}} = p_E(y - f^{AN}(x))$. We parameterize $\beta(y)$ as the exponential transformation of a nonlinear function represented by MLP's (with the tanh activation function); this automatically guarantees the nonnegativity constraint of $\beta(y)$, as required in (17). Furthermore, we represent $p_X^{\mathcal{F}}$ with a mixture of Gaussians, the nonlinear function f^{AN} with MLP's (with the tanh activation function), and p_E with another mixture of Gaussians.

Enforcing p_{XY}^{β} to Be a Valid Density We present a sample-average approximation scheme to approximately enforce the condition that the right hand side of (10) corresponds to a valid distribution, or more specifically, to impose the constraint (18). Notice that the given data points $\{x_k, y_k\}_{k=1}^n$ were drawn from p_{XY}^{β}. As a matter of fact, we have

$$\int \beta(y)p_X^{\mathcal{F}}p_{Y|X}^{\mathcal{F}} = \int p_{XY}^o \frac{\beta(y)p_X^{\mathcal{F}}p_{Y|X}^{\mathcal{F}}}{p_{XY}^o}dxdy \tag{19}$$

$$\approx \frac{1}{n}\sum_{k=1}^{n}\frac{\beta(y_k)p_X^{\mathcal{F}}(x_k)p_{Y|X}^{\mathcal{F}}(y_k|x_k)}{p_{XY}^o(x_k, y_k)} \tag{20}$$

$$\approx \frac{1}{n}\sum_{k=1}^{n}\frac{\beta(y_k)p_X^{\mathcal{F}}(x_k)p_{Y|X}^{\mathcal{F}}(y_k|x_k)}{\hat{p}_{XY}^o(x_k, y_k)}, \tag{21}$$

where p_{XY}^o denotes the data distribution of (X, Y), and $\hat{p}_{XY}^o(x_k, y_k)$ denotes its estimate at point (x_k, y_k). Here the expression in (20) is an empirical estimate of (19) on the sample drawn from the distribution p_{XY}^o; furthermore, (20) replaces the density $p_{XY}^o(x_k, y_k)$ with its empirical estimate $\hat{p}_{XY}^o(x_k, y_k)$. As a consequence, the constraint (18) can be (approximately) achieved by enforcing

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\beta(y_k)p_X^{\mathcal{F}}(x_k)p_{Y|X}^{\mathcal{F}}(y_k|x_k)}{\hat{p}_{XY}^o(x_k, y_k)} = 1. \tag{22}$$

In our experiments, we used kernel density estimation with a Gaussian kernel for $\hat{p}_{XY}^o(x_k, y_k)$; for each dimension, we set the kernel width to the median distance between points in the sample, as in (Gretton et al., 2007).

Under the parameterization given in (27) and with the above approach to guarantee that p_{XY}^o is (approximately) a valid density, one can then maximize the likelihood function given in (15) to estimate the function f^{AN}, the noise distribution, and $\beta(y)$.

5.2 With Score Matching

Alternatively, we can estimate the parameters by score matching (Hyvärinen, 2005), i.e., by minimizing the expected squared distance between the gradient of the log-density given by the model and the gradient of the log-density of the observed data. This procedure aims to match the *shape* of the density given by the model and that of the empirical density of the observed data, and is invariant to the scaling factor of the model density. As a clear advantage, in the optimization procedure one does not need to guarantee that p_{XY}^{β} is a valid density.

Given any model density $p_Z(z;\theta)$ of a m-dimensional random vector Z, the score function is the gradient of the log-density w.r.t. the data vector, i.e., $\psi(z;\theta) = (\psi_1(z;\theta), ..., \psi_m(z;\theta))^\intercal = (\frac{\partial \log p_Z(z;\theta)}{\partial z_1}, ..., \frac{\partial \log p_Z(z;\theta)}{\partial z_m})^\intercal$. Note that the score function is invariant to scale transformations in $p_Z(z)$, i.e., it is invariant to the normalization constant for a valid density. One can then estimate model parameters by minimize the expected squared distance between the model score function $\psi(\cdot;\theta)$ and the data score function $\psi_Z(\cdot;\theta)$, i.e., minimize $\frac{1}{2}\int_{z\in\mathbb{R}^m} p_Z(z)||\psi(z;\theta) - \psi_Z(z)||^2 dz$. It has been shown in (Hyvärinen, 2005) that minimizing the above squared distance is equivalent to minimizing

$$J^{SM}(\theta) = \int_{z\in\mathbb{R}^m} p_Z(z)\sum_{i=1}^{m}\big[\tilde{\psi_i}(z;\theta) + \frac{1}{2}\psi_i^2(z;\theta)\big]dz,$$

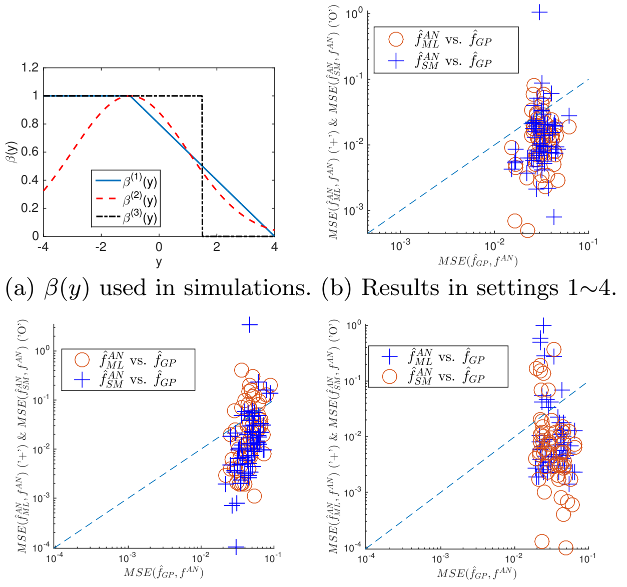
where $\tilde{\psi_i}(z;\theta) = \frac{\partial \psi_i(z;\theta)}{\partial z_i}$. The sample version of $J^{SM}(\theta)$ over the sample $\mathbf{z}_1, ..., \mathbf{z}_n$ is

$$\hat{J}^{SM}(\theta) = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{m}\big[\tilde{\psi_i}(\mathbf{z}_k;\theta) + \frac{1}{2}\psi_i^2(\mathbf{z}_k;\theta)\big]. \tag{23}$$

In particular, here we have $\psi_1 = \psi_X$ and $\psi_2 = \psi_Y$; noting that $p_{Y|X}^{\mathcal{F}} = p_E(y - f^{AN}(x))$, we can write down the involved derivatives involved in (28), and then minimize the regularized score function (with the same regularization term as in Eq. 16) to estimate the involved parameter.

6 Experiments

Simulations The simulated data are generated by applying the biased selection procedure on the data generated by a additive noise model with function f^{AN}, i.e., by (9) and (10). As shown in Section 4, the function f^{AN} is identifiable up to some shift when the noise is non-Gaussian. We shall study the estimation quality of the regression function f^{AN} under different settings.



(a) $\beta(y)$ used in simulations. (b) Results in settings 1∼4.

(c) Results in settings 5∼8. (d) Results in settings 9∼12.

Figure 3: Simulation settings and results. (a) shows the density ratio functions for OSB, $\beta_1(y)$, $\beta_2(y)$, and $\beta_3(y)$, which are used in settings 1-4, 5-8, and 9-12, respectively. (b), (c), and (d) show the pairwise MSE of the estimated function for the proposed methods against GP regression on the given sample, in settings 1-4, 5-8, and 9-12, respectively. The dashed line marks the threshold where the proposed methods and GP regression on the given sample perform equally well.

We consider three settings for OSB, by setting $\beta(y)$ (see Eq. 10) to different functions. As shown in Figure 3(a), $\beta^{(1)}(y)$ is a piecewise linear function, $\beta^{(2)}(y)$ is a (scaled) Gaussian function with mean -1 and standard deviation 2, and $\beta^{(3)}(y)$ corresponding to a hard biased selection procedure: it drops all data points corresponding to the 10% largest values of Y.

We use two ways to generate the distributions of X and E; one is the uniform distribution, and the other the mixture of three Gaussians with random coefficients. The function f^{AN} is a mixture of the linear, tanh, and cubic function with random coefficients (the coefficient for the cubic function is constrained to be small to avoid extreme values in Y).

In total there are $2 \times 2 \times 3$ simulation settings. For each setting we repeat the experiment with 15 random replications. We use the methods proposed in Section 5 to recover this function. Denote by $\hat f^{AN}_{ML}$ the estimate given by the (approximate) maximum likelihood method, and by $\hat f^{AN}_{SM}$ that given by the score matching method. The estimation performance is evaluated by the mean square error (MSE) between the estimate and the true function, $\frac{1}{n}\sum_{i=1}^{n}(\hat f^{AN}(x_i) - f^{AN}(x_i))^2$. We compare the estimates produced by our methods with that estimated by Gaussian process (GP) regression on the given sample, denoted by $\hat f_{GP}$. Figure 3(b-d) compares the estimation quality of $\hat f^{AN}_{ML}$ and $\hat f^{AN}_{SM}$ against $\hat f_{GP}$; note that they are plotted on a log scale.

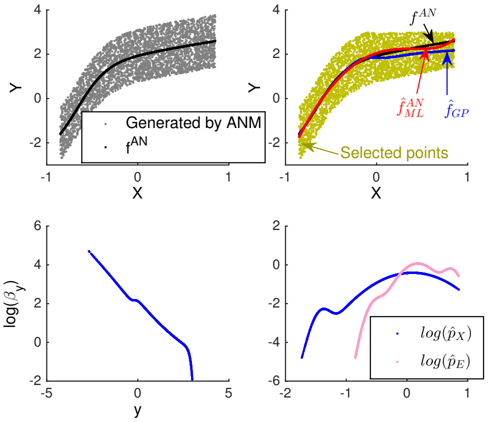


Figure 4: Results of a typical run estimated by the maximum likelihood approach. The four subfigures show the data produced by the ANM, the selected sample and the estimates of f^{AN}, the estimate of $\beta(y)$, and the estimates of p_X and p_E, respectively.

As one can see from Figure 3(b-d), the proposed methods may converge to unwanted solutions, as shown by the few points above the dashed lines. However, in most cases the proposed method provides a better estimate of the function f^{AN}. As suggested by (Demšar, 2006), we use the Wilcoxon signed ranks test to check whether $MSE(\hat f^{AN}, f^{AN})$ is significantly better than $MSE(\hat f_{GP}, f^{AN})$ under all the three settings for $\beta(y)$. It is a nonparametric test to detect shifts in populations given a number of paired samples. Under the null hypothesis the distribution of differences between the two populations is symmetric about 0. We find that under all the three sets of settings, for the score matching method, the null hypothesis is always rejected at the 0.01 level (the p-values are 6×10^{-7}, 2×10^{-6}, and 8×10^{-5}, respectively). For the maximum likelihood method, the null is rejected in settings 1-4 and 5-9 (the p-values are 2×10^{-7}, and 8×10^{-5}, respectively); we fail to reject the null in settings 5-8 (with the p-value 0.22): this seems to be caused by local optima, which will be discussed below. This means that the proposed method outperforms the method that fits GP regression on the observed data, in terms of the estimation quality of the true function.

Figure 4 gives the result of a typical run (with $\beta_3(y)$) produced by maximum likelihood. Interestingly, one can see that compared to the true $\beta(t)$, which is $\beta_3(y)$ in Figure 3(a), $\hat{\beta}(y)$ contains an additional factor of the form $e^{c_2 y}$ with some constant c_2. The estimates of p_X and p_E are also skewed accordingly. This verifies the statements given in Theorem 2(a).

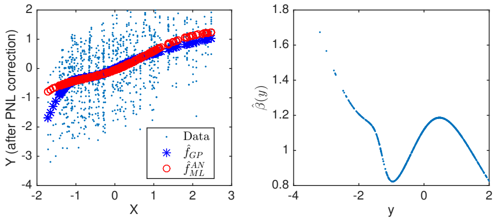
As seen from Figure 3, both algorithms may get stuck in local optima. Let us have a closer look at the results given by maximum likelihood. We found that for each simulation setting, in all runs where the function f^{AN} was not accurately recovered, or more specifically, where $MSE(\hat{f}^{AN}_{ML}, f^{AN}) > MSE(\hat{f}_{GP}, f^{AN})$, the corresponding likelihood values are among the lowest across all 15 runs. That is, the attained likelihood value suggests whether the algorithm converges to a local optimum. Therefore, in practice one may run the algorithms multiple times with random initializations and choose the one which gives the best model fit (e.g., the highest likelihood). However, this is not the case for the score matching-based approach: we did not find that the unwanted solutions always correspond to relatively large score distances. The precise reason for this phenomenon is still under investigation. Hence, below we only report the results given by the (approximate) maximum likelihood approach.

Experiments on Real Data We went through the cause-effect pairs (`http://webdav.tuebingen.mpg.de/cause-effect/`) to find data sets which are likely to suffer the OSB issue according to *commonsense or background knowledge*. We selected Pairs 25, 40, and 41. Here to save space, we only report the results on Pair 25; it is about the relationship between the age (X) and the concrete compressive strength (Y) of different samples of concrete.

The empirical distribution of the data in Pair 25 suggests that it is very likely for the effect to suffer from a PNL distortion. We use a rough way to take into account both the PNL distortion in the causal process and the OSB. We first fit the PNL causal model (Zhang & Hyvärinen, 2009) on the data and correct the data with the estimated PNL transformation on Y. We then fit the ANM-OSB procedure on the corrected data. To avoid local optima, we run the (approximate) maximum likelihood algorithm presented in Section 5.1 five times with random initializations and choose the one with the highest likelihood. Figure 7 shows the result on Pair 25. As seen from $\hat{\beta}(y)$, it seems for some reason, the samples whose compressive strength is very high were not selected. The estimated function $\hat{f}^{GP}_{ML}$ seems to address this issue.

7 Conclusion and Discussions

As we have shown, in the presence of outcome-dependent selection, the causal direction is still generically identifiable if the causal relationship can be modeled by a post-nonlinear causal model. Moreover, in the case of an additive noise model, the causal mechanism as represented by the function in the model is identifiable up to a constant if the noise term is non-Gaussian (and completely identified if the noise term follows a symmetric, non-Gaussian distribution). However, due to the selection bias, the estimation requires more care than standard methods for fitting such models, and we developed two estimation procedures in this paper.



(a) Data & estimated functions. (b) $\hat{\beta}(y)$.

Figure 5: Results on pair 25 of the cause-effect pairs. (a) The scatterplot of the data (after correcting the nonlinear distortion in Y with the PNL causal model), the nonlinear regression function $\hat{f}_{GP}$ on the data, and the estimated function $\hat{f}^{AN}_{ML}$ by the maximum likelihood approach. (b) The estimated density ratio $\beta(y)$.

This is a first step towards a better understanding of the bearing of selection bias on the identifiability and estimation of functional causal models. There are several interesting problems for future work. First, the identifiability result on additive noise models can be generalized to post-nonlinear models, but it will take more work to put the more general result in a sufficiently simple form. Second, our positive results here are confined to outcome-dependent selection. Thanks to this restriction, our results do not rely on any substantial assumption on the selection mechanism. For more complex structures of selection, such as when the selection depends on both cause and effect, identifiability will require more specifications of the selection model. Third, our result on the identification of causal direction is confined to the two-variable case without latent confounders; how to handle selection in multi-variable structural learning, with or without latent confounders, remains an open problem.

Acknowledgement: Research reported in this publication was supported by the National Institutes of Health (NIH) under Award Number U54HG008540. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The research of J. Zhang was supported in part by the Research Grants Council of Hong Kong under the General Research Fund LU342213.

References

Bareinboim, E. and Pearl, J. Controlling selection bias in causal inference. In *Proceedings of AAAI*, pp. 100–108, 2012.

Bareinboim, E., Tian, J., and Pearl, J. Recovering from selection bias in causal and statistical inference. In *Proceedings of AAAI*, 2014.

Borboudakis, G. and Tsamardinos, I. Bayesian network learning with discrete case-control data. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, 2015.

Castillo, E. *Functional Equations and Modelling in Science and Engineering.* CRC Press, 1992.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Didelez, V., Kreiner, S., and Keiding, N. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25:368–387, 2010.

Evans, R. J. and Didelez, V. Recovering from selection bias using marginal structure in discrete models. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, 2015.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. In *NIPS 19*, pp. 513–520, Cambridge, MA, 2007. MIT Press.

Heckerman, D., Geiger, D., and Chickering, D. M. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

Heckman, J. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979.

Hoyer, P.O., Janzing, D., Mooji, J., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, Vancouver, B.C., Canada, 2009.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2005.

Kagan, A. M., Linnik, Y. V., and Rao, C. R. *Characterization Problems in Mathematical Statistics.* Wiley, New York, 1973.

Pearl, J. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, 2000.

Shimizu, S., Hoyer, P.O., Hyvärinen, A., and Kerminen, A.J. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Spirtes, P., Meek, C., and Richardson, T. An algorithm for causal inference in the presence of latent variables and selection bias. In Glymour, C. and Cooper, G. (eds.), *Computation, Causation, and Discovery*, pp. 211–252. MIT Press, 1999.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, 2nd edition, 2001.

Winship, C. and Mare, R. D. Models for sample selection bias. *Annual Review of Sociology*, 18:327–350, 1992.

Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172:1873–1896, 2008.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.