Measurement Bias and Effect Restoration in Causal Inference

Manabu Kuroki†

The Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan

Judea Pearl

University of California, Los Angeles, Los Angeles, CA, USA

Summary.
This paper highlights several areas where graphical techniques can be harnessed to address the problem of measurement errors in causal inference. In particular, the paper discusses the control of partially observable confounders in parametric and non parametric models and the computational problem of obtaining bias-free effect estimates in such models.

Keywords: Causal diagram; Confounder; IV method; Proxy variable; Regression coefficient; Total effect

1. Introduction

This paper discusses methods of dealing with measurement errors in the context of graph-based causal inference. It is motivated by a powerful result reported in Greenland and Lash (2008) which is rooted in classical regression analysis (Greenland and Kleinbaum, 1983; Selén, 1986; Carroll et al., 2006), but has not been fully utilized in causal analysis or graphical models.

Let $\mathrm{pr}(\boldsymbol{v})$ be the joint distribution of $\boldsymbol{V} \stackrel{\triangle}{=} (V_1, \cdots, V_n) = (v_1, \cdots, v_n)$, $\mathrm{pr}(v_i|v_j)$ the conditional distribution of $V_i = v_i$ given $V_j = v_j$ and $\mathrm{pr}(v_i)$ the marginal distribution of $V_i = v_i$. Similar notation is used for other distributions. For the graph-theoretic terminology used in this paper, we refer readers to Pearl (1988, 2009).

Given a directed acyclic graph $G = (\boldsymbol{V}, \boldsymbol{E})$ with a set $\boldsymbol{V}$ of variables and a set $\boldsymbol{E}$ of arrows, a probability distribution $\mathrm{pr}(\boldsymbol{v})$ is said to be *compatible* with G if it can be factorized as:

$$\mathrm{pr}(\boldsymbol{v}) = \prod_{i=1}^{n} \mathrm{pr}\{v_i|\mathrm{pa}(v_i)\}, \tag{1}$$

†*Address for correspondence:* Manabu Kuroki, Department of Data Science, The Institute of Statistical Mathematics, 10-3, Midori-cho, Tachikawa, Tokyo, 190-8562, Japan
E-mail: mkuroki@ism.ac.jp

where $\mathrm{pa}(v_i)$ is a set of parents of V_i. When $\mathrm{pa}(v_i)$ is an empty set, $\mathrm{pr}\{v_i|\mathrm{pa}(v_i)\}$ is the marginal distribution $\mathrm{pr}(v_i)$. When equation (1) holds, we also say that G is a *Bayesian network* of $\mathrm{pr}(\boldsymbol{v})$ (Pearl, 2009,pp.13–16).

If a joint distribution is factorized recursively according to the graph G, then the conditional independencies implied by the factorization (1) can be obtained from the graph G according to the *d*-separation criterion (Pearl, 1988). That is, for any distinct subsets $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$, if $\boldsymbol{Z}$ *d*-separates $\boldsymbol{X}$ from $\boldsymbol{Y}$ in G, then $\boldsymbol{X}$ is conditionally independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$, denoted as $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$, in every distribution satisfying equation (1).

If every parent-child family in the graph G stands for an independent data-generating mechanism, the Bayesian network is called a *causal diagram* (see Pearl, 2009, p.24, for formal definition). Based on a causal diagram G, for any $X, Y \in \boldsymbol{V}$, the *causal effect* of X on Y is defined as

$$\mathrm{pr}\{y|\mathrm{do}(x)\} \triangleq \sum_{v\setminus\{x,y\}} \frac{\mathrm{pr}\{x,y,\boldsymbol{v}\setminus\{x,y\}\}}{\mathrm{pr}\{x|\mathrm{pa}(x)\}},$$

where $\mathrm{do}(x)$ indicates that X is fixed to x by an external intervention (Pearl, 2009). When the causal effect can be determined uniquely from a joint distribution of observed variables, the causal effect is said to be *identifiable*. The most common identifiability condition that can be obtained from the graph structure is the *back door criterion*. A set $\boldsymbol{S}$ of variables is said to satisfy the back door criterion relative to an ordered pair of variables (X, Y) if (i) no vertex in $\boldsymbol{S}$ is a descendant of X, and (ii) $\boldsymbol{S}$ *d*-separates X from Y in the graph obtained by deleting from a graph G all arrows emerging from X. If any such set can be measured, the causal effect of X on Y is identifiable and is given by the formula $\mathrm{pr}\{y|\mathrm{do}(x)\} = \sum_s \mathrm{pr}(y|x, \boldsymbol{s})\mathrm{pr}(\boldsymbol{s})$ (Pearl, 2009, pp.79–80); $\boldsymbol{S}$ is then called *sufficient*.

With the preparation above, we consider the problem of estimating the causal effect of X on Y when a sufficient confounder U is unobserved, and can only be measured with error (see Fig.1), via a proxy variable W. In Fig.1, U satisfies the back door criterion relative to an ordered pair of variables (X, Y), but its proxy variable W does not. Since U is sufficient, the causal effect is identifiable from measurement on X, Y and U, and can be written as

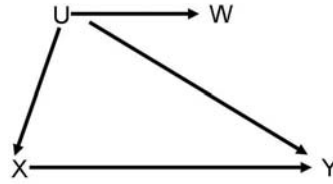$$\mathrm{pr}\{y|\mathrm{do}(x)\} = \sum_u \mathrm{pr}(y|x, u)\mathrm{pr}(u). \tag{2}$$

Fig.1: Needed the causal effect of X on Y when U is unobserved, and W provides a noisy measurement of U.

However, since U is unobserved and W is but a noisy measurement of U, d-separation tells us immediately that adjusting for W is inadequate, for it leaves the back door path(s) $X \leftarrow U \rightarrow Y$ unblocked. Therefore, regardless of sample size, the causal effect of X on Y cannot be estimated without bias. It turns out, however, that if we are given the conditional distribution $\mathrm{pr}(w|u)$ that governs the error mechanism, we can perform a modified-adjustment for W that, in the limit of very large sample, would amount to the same thing as observing and adjusting for U itself, thus rendering the causal effect identifiable. The possibility of removing bias by modified adjustment is far from obvious, because, although $\mathrm{pr}(w|u)$ is assumed to be given, the actual value of a confounder U remains uncertain for each measurement $W = w$, so one would expect to get either a distribution over causal effects, or bounds thereof. Instead, we can actually get a repaired point estimate of $\mathrm{pr}\{y|\mathrm{do}(x)\}$ that is asymptotically unbiased.

This result, which we will label *effect restoration*, has powerful consequences in practice because, when $\mathrm{pr}(w|u)$ is not given, one can resort to a Bayesian (or bounding) analysis and assume a prior distribution (or bounds) on the parameters of $\mathrm{pr}(w|u)$ which would yield a distribution (or bounds) over $\mathrm{pr}\{y|\mathrm{do}(x)\}$ (Greenland, 2005). Alternatively, if costs permit, one can estimate $\mathrm{pr}(w|u)$ by re-testing U in a sampled subpopulation (Carroll et al., 2006). This is normally done by re-calibration techniques (Greenland and Lash, 2008), called a *validation study*, in which U is measured without error in a subpopulation and used to calibrate the estimates in the main study (Selén, 1986).

On the surface, the possibility of correcting for measurement bias seems to undermine the importance of accurate measurements. It suggests that as long as we know how bad our measurements are there is no need to improve them because they can be corrected post-hoc by analytical means. This is not so. First, although an unbiased effect estimate can be recovered from noisy measure-

ments, sampling variability increases substantially with error. Second, even assuming unbounded sample size, the estimate will be biased if the postulated $\mathrm{pr}(w|u)$ is incorrect. In extreme cases, wrongly postulated $\mathrm{pr}(w|u)$ may even conflict with the data, and no estimate will be obtained. For example, if we postulate a non informative W, $\mathrm{pr}(w|u) = \mathrm{pr}(w)$, and we find that W strongly depends on X, a contradiction arises and no effect estimate will emerge.

Effect restoration can be analyzed from either a statistical or causal viewpoint. Taking the statistical view, one may argue that, once the causal effect is identified in terms of a latent variable U and given the estimand in equation (2), the problem is no longer one of causal inference, but rather of regression analysis, whereby the regressional expression $E_u\{\mathrm{pr}(y|x,u)\}$ need to be estimated from a noisy measurement of U, given by W. This is indeed the approach taken in the vast literature on measurement error (e.g., Selén, 1986; Carroll et al., 2006).

The causal analytic perspective is different; it maintains that the ultimate purpose of the analysis is not the statistics of X, Y, and U, as is normally assumed in the measurement-error literature, but the causal effect $\mathrm{pr}\{y|\mathrm{do}(x)\}$ that is mapped into regression vocabulary only when certain causal assumptions are deemed plausible. Awareness of these assumptions should shape the way we deal with measurement error. For example, the very idea of modeling the error mechanism $\mathrm{pr}(w|u)$ requires causal considerations; errors caused by noisy measurements are fundamentally different from those caused by noisy agitators. Likewise, the reason we seek an estimate $\mathrm{pr}(w|u)$ as opposed to $\mathrm{pr}(u|w)$, be it from scientific judgments or from pilot studies, is that we consider the former to be a more reliable and transportable parameter than the latter. Transportability (Pearl and Bareinboim, 2011) is a causal notion that is hardly touched upon in the measurement-error literature, where causal vocabulary is usually avoided and causal relations relegated to informal judgment (e.g., Carroll et al., 2006, pp.29-32).

Viewed from this perspective, the measurement-error literature appears to be engaged (unwittingly) in causal considerations that can benefit from making the causal framework explicit. The benefit can in fact be mutual; identifiability with partially specified causal parameters (as in Fig.1) is rarely discussed in the causal inference literature (notable exceptions are Goetghebeur and Vansteelandt (2005), Cai and Kuroki (2008), Hernán and Cole (2009) and Pearl (2010)), while graphical models are hardly used in the measurement-error literature.

In this paper, we will consider the mathematical aspects of effect restoration and will focus on asymptotic analysis. Our aims are to understand the conditions under which effect restoration is feasible, to assess the computational problems it presents, and to identify those features of $\mathrm{pr}(w|u)$ and $\mathrm{pr}(x, y, w)$ that are major contributors to measurement bias.

2.   Effect Restoration with External Studies

2.1.   Matrix Adjustment Method

Guided by the graph shown in Fig.1, we start with the joint probability $\mathrm{pr}(x, y, w, u)$ and assume that W depends only on U, i.e., $\mathrm{pr}(w|x, y, u) = \mathrm{pr}(w|u)$. This assumption is often called *non-differential error* (Carroll et al., 2006).

We further assume that:

(a) the distribution $\mathrm{pr}(w|u)$ of the error mechanism are available from external studies such as pilot studies or scientific judgments, and

(b) the confounder U is a discrete variable with a given finite number of categories, while X, Y and W may be continuous or discrete, as long as the number of categories of W is greater or equal to that of U.

The main idea of recovering $\mathrm{pr}(x, y, u)$ from both $\mathrm{pr}(x, y, w)$ and $\mathrm{pr}(w|u)$, adapted from Greenland and Lash (2008, p.360) and Pearl (2010), is as follows: for U and W such that $u \in \{u_1, ...u_k\}$ and $w \in \{w_1, ..., w_k\}$, we have

$$\mathrm{pr}(x, y, w) = \sum_{i=1}^{k} \mathrm{pr}(x, y, u_i)\mathrm{pr}(w|u_i), \tag{3}$$

Then, for any specific values x and y, letting

$$V_{xy}(u) = \begin{pmatrix} \mathrm{pr}(x, y, u_1) \\ \vdots \\ \mathrm{pr}(x, y, u_k) \end{pmatrix}, V_{xy}(w) = \begin{pmatrix} \mathrm{pr}(x, y, w_1) \\ \vdots \\ \mathrm{pr}(x, y, w_k) \end{pmatrix}, M(w, u) = \begin{pmatrix} \mathrm{pr}(w_1|u_1) & \cdots & \mathrm{pr}(w_1|u_k) \\ \vdots & \ddots & \vdots \\ \mathrm{pr}(w_k|u_1) & \cdots & \mathrm{pr}(w_k|u_k) \end{pmatrix},$$

equations (3) can be written as matrix multiplication:

$$V_{xy}(w) = M(w, u)V_{xy}(u). \tag{4}$$

Now, assuming that

(c) $M(w, u)$ is invertible,

the elements $\text{pr}(x, y, u)$ of $V_{xy}(u)$ are estimable and are given by

$$V_{xy}(u) = I(w, u)V_{xy}(w), \qquad (5)$$

where $I(w, u) = M(w, u)^{-1}$. Thus, equation (5) enables us to reconstruct $\text{pr}(x, y, u)$ from $\text{pr}(x, y, w)$. In other words, each term on the right hand side of equation (2) can be obtained from $\text{pr}(x, y, w)$ and $\text{pr}(w|u)$ through equation (5) and, assuming U is sufficient, the causal effect $\text{pr}\{y|\text{do}(x)\}$ is estimable from the available data. Explicitly, letting $i(w, u)$ be the corresponding element of $I(w, u)$ that corresponds to $(W, U) = (w, u)$, we have:

$$\text{pr}\{y|\text{do}(x)\} \quad = \quad \sum_{i=1}^{k} \frac{\text{pr}(x, y, u_i)\text{pr}(u_i)}{\text{pr}(x, u_i)} = \sum_{l=1}^{k} \frac{\sum_{j=1}^{k} i(w_j, u_l)\text{pr}(x, y, w_j) \sum_{j=1}^{k} i(w_j, u_l)\text{pr}(w_j)}{\sum_{j=1}^{k} i(w_j, u_l)\text{pr}(x, w_j)} \qquad (6)$$

Note that the same inverse matrix, $I(w, u)$ appears in all summations.

When we do not assume independent noise mechanisms, this will not be the case. In other words, if $\text{pr}(w|x, y, u) = \text{pr}(w|u)$ does not hold, we must write:

$$\text{pr}(x, y, w) = \sum_{i=1}^{k} \text{pr}(w|x, y, u_i)\text{pr}(x, y, u_i),$$

which can be transformed to matrix multiplication $V_{xy}(w) = M_{xy}(w, u)V_{xy}(u)$, where

$$M_{xy}(w, u) = \begin{pmatrix} \text{pr}(w_1|y, x, u_1) & \cdots & \text{pr}(w_1|y, x, u_k) \\ \vdots & \ddots & \vdots \\ \text{pr}(w_k|y, x, u_1) & \cdots & \text{pr}(w_k|y, x, u_k) \end{pmatrix},$$

and its inverse $I_{xy}(w, u)$ are both indexed by the specific values of x and y. Thus, when both X and Y are discrete variables with a given finite number of categories, we obtain:

$$V_{xy}(u) = I_{xy}(w, u)V_{xy}(w) \qquad (7)$$

which, again, permits the identification of the causal effect via equation (2) except that the expression becomes somewhat more complicated. It is also clear that errors in the measurement of X and Y can be absorbed into W, and do not present any conceptual problem.

Equation (6) demonstrates the feasibility of effect reconstruction and proves that, despite the uncertainty in the variables X, Y and U, the causal effect is identifiable once we know the statistics of the error mechanism.

This result is asymptotic, and presents practical challenges of computation and estimation. In particular, one needs to address the problem of empty cells which, owed to the high dimensionality of W and U would prevent us from getting reliable statistics of $\mathrm{pr}(x, y, w)$, as required by equation (6). When X is a binary variable, one can resort then to propensity score (PS) methods (Rosenbaum and Rubin, 1983), which map the cells of U onto a single scalar.

The error-free propensity score $L(u) = \mathrm{pr}(x_1|u)$ being a functional of $\mathrm{pr}(x, y, u)$ can be estimated consistently from samples of $\mathrm{pr}(x, y, w)$ using the transformation (5). Explicitly, we have:

$$L(u) = \mathrm{pr}(x_1|u) = \frac{\mathrm{pr}(x_1, u)}{\mathrm{pr}(u)},$$

where $\mathrm{pr}(x_1, u)$ and $\mathrm{pr}(u)$ are given in equations (5).

Using the decomposition in $\mathrm{pr}(w|x, y, u)=\mathrm{pr}(w|u)$, we can further write:

$$L(u) = \frac{\sum_{j=1}^{k} i(w_j, u)\mathrm{pr}(x_1, w_j)}{\sum_{j=1}^{k} i(w_j, u)\mathrm{pr}(w_j)} = \frac{\sum_{j=1}^{k} i(w_j, u)L(w)\mathrm{pr}(w_j)}{\sum_{j=1}^{k} i(w_j, u)\mathrm{pr}(w_j)}, \tag{8}$$

where $L(w)$ is the error-prone propensity score $L(w) = \mathrm{pr}(x_1|w)$. We see that $L(u)$ can be computed from $I(w, u)$, $L(w)$ and $\mathrm{pr}(w)$. Thus, if we succeed in estimating these three quantities in a parsimonious parametric form, the computation of $L(u)$ would be hindered only by the summations called for in equation (5). Once we estimate $L(w)$ parametrically for each conceivable w, equation (8) permits us to assign to each tuple u a bias-less score $L(u)$ that correctly represents the probability of $X = x_1$ given $U = u$. This, in turn, should permit us to estimate, for each stratum $L(u) = l$, the probability

$$\mathrm{pr}(l) = \sum_{u|L(u)=l} \mathrm{pr}(u),$$

then compute the causal effect using

$$\mathrm{pr}\{y|\mathrm{do}(x)\} = \sum_{l} \mathrm{pr}(y|x, l)\mathrm{pr}(l).$$

One technique for approximating $\mathrm{pr}(l)$ was proposed by Stürmer et al. (2005) and Schneeweiss et al. (2009), which did not make full use of the inversion in equation (8) or of graphical methods facilitating this inversion. A more promising approach would be to construct $\mathrm{pr}(l)$ and $\mathrm{pr}(y|x, l)$ directly from synthetic samples of $\mathrm{pr}(x, y, u)$ that can be created to mirror the empirical samples of $\mathrm{pr}(x, y, w)$. This is illustrated in the next subsection, using binary variables.

2.2. Matrix Adjustment in Binary Models

Let X, Y, W and U be binary variables, $w \in \{w_1, w_2\}$, $u \in \{u_1, u_2\}$, and let the noise mechanism be characterizes by $\mathrm{pr}(w_2|u_1)$ and $\mathrm{pr}(w_1|u_2)$. Then, equation (5) translates to

$$\left.\begin{aligned}
\mathrm{pr}(x, y, u_1) &= \frac{\{1 - \mathrm{pr}(w_1|u_2)\}\mathrm{pr}(x, y, w_1) - \mathrm{pr}(w_1|u_2)\mathrm{pr}(x, y, w_2)}{1 - \mathrm{pr}(w_2|u_1) - \mathrm{pr}(w_1|u_2)} \\
\mathrm{pr}(x, y, u_2) &= \frac{\{1 - \mathrm{pr}(w_2|u_1)\}\mathrm{pr}(x, y, w_2) - \mathrm{pr}(w_2|u_1)P(x, y, w_1)}{1 - \mathrm{pr}(w_2|u_1) - \mathrm{pr}(w_1|u_2)},
\end{aligned}\right\} \tag{9}$$

which represents the inverse matrix

$$I(w, u) \quad = \quad \frac{1}{1 - \mathrm{pr}(w_2|u_1) - \mathrm{pr}(w_1|u_2)} \begin{pmatrix} 1 - \mathrm{pr}(w_1|u_2) & -\mathrm{pr}(w_1|u_2) \\ -\mathrm{pr}(w_2|u_1) & 1 - \mathrm{pr}(w_2|u_1) \end{pmatrix}$$

Metaphorically, the transformation in equation (9) can be described as a mass re-assignment process, as if every two cells, (x, y, w_1) and (x, y, w_2), compete on how to split their combined weight $\mathrm{pr}(x, y)$ between the two latent cells (x, y, u_1) and (x, y, u_2) thus creating a synthetic population $\mathrm{pr}(x, y, u)$ governed by equation (6). Fig.2 describes how $\mathrm{pr}(w_1|x, y)$, the fraction of the weight held by the (x, y, w_1) cell determines the ratio $\mathrm{pr}(u_1|x, y)/\mathrm{pr}(u_2|x, y)$ that is eventually received by cells (x, y, u_1) and (x, y, u_0) respectively. We see that when $\mathrm{pr}(w_1|x, y)$ approaches $1 - \mathrm{pr}(w_2|u_1)$, most of the $\mathrm{pr}(x, y)$ weight goes to cell (x, y, u_1), whereas when $\mathrm{pr}(w_1|x, y)$ approaches $\mathrm{pr}(w_1|u_2)$, most of that weight goes to cell (x, y, u_2).

Clearly, when $\mathrm{pr}(w_2|u_1) + \mathrm{pr}(w_1|u_2) = 1$, or $U \perp\!\!\!\perp W$, W provides no information about U and the inverse does not exist. Likewise, whenever any of the synthetic distributions $\mathrm{pr}(x, y, u)$ falls outside the $(0, 1)$ interval, a modeling constraint is violated (see Pearl (1988, Chapter 8)) meaning that the observed distribution $\mathrm{pr}(x, y, w)$ and the postulated error mechanism $\mathrm{pr}(w|u)$ are incompatible with the structure of Fig.1. If we assign reasonable priors to $\mathrm{pr}(w_2|u_1)$ and $\mathrm{pr}(w_1|u_2)$, the linear function in Fig.2 will become an S-shaped curve over the entire $[0, 1]$ interval, and each sample (x, y, w) can then be used to update the relative weight $\mathrm{pr}(x, y, u_1)/\mathrm{pr}(x, y, u_2)$.
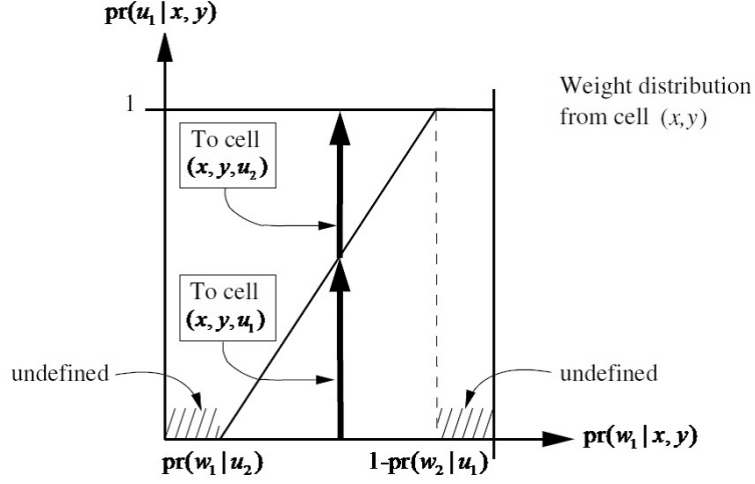
Fig.2: A curve describing how the weight $\mathrm{pr}(x, y)$ is distributed to cells (x, y, u_1) and (x, y, u_2), as a function of $\mathrm{pr}(w_1|x, y)$.

To compute the causal effect $\mathrm{pr}\{y|\mathrm{do}(x)\}$, we need only substitute $\mathrm{pr}(x, y, u)$ from equation (9) into equation (2) or (6), which gives

$$
\begin{aligned}
\mathrm{pr}\{y|\mathrm{do}(x)\} \quad = \quad & \frac{\mathrm{pr}(x, y, w_1)}{\mathrm{pr}(x|w_1)} \frac{\left(1 - \dfrac{\mathrm{pr}(w_1|u_2)}{\mathrm{pr}(w_1|x, y)}\right)\left(1 - \dfrac{\mathrm{pr}(w_1|u_2)}{\mathrm{pr}(w_1)}\right)}{1 - \dfrac{\mathrm{pr}(w_1|u_2)\mathrm{pr}(x)}{\mathrm{pr}(w_1)}} \\[2em]
& + \frac{\mathrm{pr}(x, y, w_2)}{\mathrm{pr}(x|w_2)} \frac{\left(1 - \dfrac{\mathrm{pr}(w_2|u_1)}{\mathrm{pr}(w_2|x, y)}\right)\left(1 - \dfrac{\mathrm{pr}(w_2|u_1)}{\mathrm{pr}(w_2)}\right)}{1 - \dfrac{\mathrm{pr}(w_2|u_1)\mathrm{pr}(x)}{\mathrm{pr}(w_2)}} \quad (10)
\end{aligned}
$$

This expression highlights the difference between the standard and modified adjustment for W; the former (equation (2)), which is valid if $W = U$, is given by the standard inverse probability weighting (e.g., Pearl, 2009, equation (3.11)):

$$
\mathrm{pr}\{y|\mathrm{do}(x)\} = \frac{\mathrm{pr}(x, y, w_1)}{\mathrm{pr}(x|w_1)} + \frac{\mathrm{pr}(x, y, w_2)}{\mathrm{pr}(x|w_2)}.
$$

The extra factors in equation (10) can be viewed as modifiers of the inverse probability weight needed for a bias-free estimate. Alternatively, these terms can be used to assess, given $\mathrm{pr}(w_2|u_1)$ and $\mathrm{pr}(w_1|u_2)$, what bias would be introduced if we ignore errors altogether and treat W as a faithful representation of U. When both $\mathrm{pr}(w_2|u_1) \ll 1$ and $\mathrm{pr}(w_1|u_2) \ll 1$ hold, the first-order
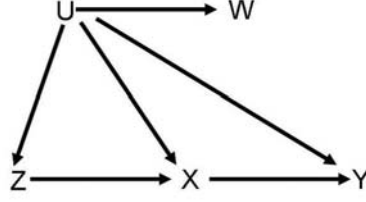
Fig.3: A causal model with two proxy variables of U, permitting the identification of $\text{pr}\{y|\text{do}(x)\}$.

approximation of equation (10) reads:

$$
\begin{aligned}
\text{pr}\{y|\text{do}(x)\} \quad \simeq \quad & \frac{\text{pr}(x,y,w_1)}{\text{pr}(x|w_1)} \left\{ 1 - \text{pr}(w_1|u_2) \left(\frac{1}{\text{pr}(w_1|x,y)} - \frac{1-\text{pr}(x)}{\text{pr}(w_1)} \right) \right\} \\
& + \frac{\text{pr}(x,y,w_2)}{\text{pr}(x|w_2)} \left\{ 1 - \text{pr}(w_2|u_1) \left(\frac{1}{\text{pr}(w_2|x,y)} - \frac{1-\text{pr}(x)}{\text{pr}(w_2)} \right) \right\}.
\end{aligned}
$$

We see that, even with two error parameters (i.e. $\text{pr}(w_2|u_1)$ and $\text{pr}(w_1|u_2)$), and eight cells, the expression for $\text{pr}\{y|\text{do}(x)\}$ does not simplify to provide an intuitive understanding of the effect of $\text{pr}(w_2|u_1)$ and $\text{pr}(w_1|u_2)$ on the estimand. Such evaluation will be facilitated in linear models (Section 4).

Assuming now that U is a sufficient set of K binary variables and, similarly, W is a set of K local indicators of U satisfying equation $\text{pr}(w|u) = \text{pr}(w|x,y,u)$. Each samples (x,y,w) should give rise to a synthetic distribution over the 2^K cells of (x,y,u) given by a product of K local distributions in the form of equation (9). This synthetic distribution can be sampled to generate synthetic (x,y,u) samples, from which the true propensity score $L(u) = \text{pr}(x_1|u)$ as well as the causal effect $\text{pr}\{y|\text{do}(x)\}$ can be estimated, as discussed in Section 2.1.

3. Effect Restoration without External Studies

In this section, we will tackle the more difficult problem of estimating causal effects without prior knowledge of the noise statistics. We will show that, under certain conditions, causal effects can be restored from proxy measurements alone.

Consider a causal diagram shown in Fig. 3 which is obtained by adding an observed variable Z to Fig.1. We will first show that $\text{pr}(x,y,u)$ can be recovered from $\text{pr}(x,y,z,w)$ under the following conditions:

(a) two proxy variables of U which are conditionally independent of each other given U can be

observed (e.g. W and Z in Fig.3), and U satisfies both $W \perp\!\!\!\perp \{X, Y, Z\}|U$ and $Y \perp\!\!\!\perp \{W, Z\}|\{U, X\}$, as in Fig.3.

(b) the confounder U is a discrete variable with a given finite number of categories, while X, Y, Z, and W may be continuous or discrete, as long as the number of categories of W and Z is greater or equal to that of U.

To show that, we first rearrange $\mathrm{pr}(y|x, u_1), ..., \mathrm{pr}(y|x, u_k)$ in decreasing order and relabel $\{u_1, ..., u_k\}$ as $\{u_{(1)}, ..., u_{(k)}\}$ such that $\mathrm{pr}(y|x, u_{(1)}) \geq ... \geq \mathrm{pr}(y|x, u_{(k)})$ for a given x and y, and, then, we recover $\mathrm{pr}(x, y, u)$ from $\mathrm{pr}(x, y, z, w)$ using eigenvalue analysis.

From Fig.3, with U, W and Z taking on values, $u \in \{u_1, , ... u_k\} = \{u_{(1)}, , ... u_{(k)}\}$, $w \in \{w_1, ..., w_k\}$ and $z \in \{z_1, ..., z_k\}$ respectively, we have

$$\mathrm{pr}(z, w|x) = \sum_{i=1}^{k} \mathrm{pr}(w|u_i)\mathrm{pr}(z|x, u_i)\mathrm{pr}(u_i|x) = \sum_{i=1}^{k} \mathrm{pr}(w|u_{(i)})\mathrm{pr}(z|x, u_{(i)})\mathrm{pr}(u_{(i)}|x),$$

$$\mathrm{pr}(y, w|x) = \sum_{i=1}^{k} \mathrm{pr}(w|u_i)\mathrm{pr}(y|x, u_i)\mathrm{pr}(u_i|x) = \sum_{i=1}^{k} \mathrm{pr}(w|u_{(i)})\mathrm{pr}(y|x, u_{(i)})\mathrm{pr}(u_{(i)}|x),$$

$$\mathrm{pr}(y, z|x) = \sum_{i=1}^{k} \mathrm{pr}(y|x, u_i)\mathrm{pr}(z|x, u_i)\mathrm{pr}(u_i|x) = \sum_{i=1}^{k} \mathrm{pr}(y|x, u_{(i)})\mathrm{pr}(z|x, u_{(i)})\mathrm{pr}(u_{(i)}|x),$$

$$\mathrm{pr}(y, z, w|x) = \sum_{i=1}^{k} \mathrm{pr}(w|u_i)\mathrm{pr}(z|x, u_i)\mathrm{pr}(y|x, u_i)\mathrm{pr}(u_i|x)$$

$$= \sum_{i=1}^{k} \mathrm{pr}(w|u_{(i)})\mathrm{pr}(z|x, u_{(i)})\mathrm{pr}(y|x, u_{(i)})\mathrm{pr}(u_{(i)}|x).$$

Denote by $P(z, w)$, $Q(z, w)$, $U(w, u)$ and $R(z, u)$ the following matrices:

$$P(z, w) = \begin{pmatrix} 1 & \mathrm{pr}(w_1|x) & \cdots & \mathrm{pr}(w_{k-1}|x) \\ \mathrm{pr}(z_1|x) & \mathrm{pr}(z_1, w_1|x) & \cdots & \mathrm{pr}(z_1, w_{k-1}|x) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{pr}(z_{k-1}|x) & \mathrm{pr}(z_{k-1}, w_1|x) & \cdots & \mathrm{pr}(z_{k-1}, w_{k-1}|x) \end{pmatrix},$$

$$Q(z, w) = \begin{pmatrix} \mathrm{pr}(y|x) & \mathrm{pr}(y, w_1|x) & \cdots & \mathrm{pr}(y, w_{k-1}|x) \\ \mathrm{pr}(y, z_1|x) & \mathrm{pr}(y, z_1, w_1|x) & \cdots & \mathrm{pr}(y, z_1, w_{k-1}|x) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{pr}(y, z_{k-1}|x) & \mathrm{pr}(y, z_{k-1}, w_1|x) & \cdots & \mathrm{pr}(y, z_{k-1}, w_{k-1}|x) \end{pmatrix},$$

$$U(w, u) = \begin{pmatrix} 1 & \mathrm{pr}(w_1|u_{(1)}) & \cdots & \mathrm{pr}(w_{k-1}|u_{(1)}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{pr}(w_1|u_{(k)}) & \cdots & \mathrm{pr}(w_{k-1}|u_{(k)}) \end{pmatrix},$$

$$R(z, u) = \begin{pmatrix} 1 & \mathrm{pr}(z_1|x, u_{(1)}) & \cdots & \mathrm{pr}(z_{k-1}|x, u_{(1)}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{pr}(z_1|x, u_{(k)}) & \cdots & \mathrm{pr}(z_{k-1}|x, u_{(k)}) \end{pmatrix},$$

and let $\Delta(u) = \mathrm{diag}\{\mathrm{pr}(y|x, u_{(1)}), \cdots, \mathrm{pr}(y|x, u_{(k)})\}$ and $M(u) = \mathrm{diag}\{\mathrm{pr}(u_{(1)}|x), \cdots, \mathrm{pr}(u_{(k)}|x)\}$, where $\mathrm{diag}\{a_1, ..., a_k\}$ is a $k \times k$ dimensional diagonal matrix whose diagonal entries starting in the upper left corner are $a_1, ..., a_k$.

Assume further that

(c) both $P(z, w)$ and $Q(z, w)$ are invertible, and

(d) $\mathrm{pr}(y|x, u_1), \cdots, \mathrm{pr}(y|x, u_k)$ take on distinct values

for any x and y. Then, writing $P(z, w) = R(z, u)'M(u)U(w, u)$ and $Q(z, w) = R(z, u)'M(u)\Delta(u)U(w, u)$, we have

$$\begin{aligned} P(z, w)^{-1}Q(z, w) &= \{R(z, u)'M(u)U(w, u)\}^{-1}\{R(z, u)'M(u)\Delta(u)U(w, u)\} \\ &= U(w, u)^{-1}M(u)^{-1}R(z, u)'^{-1}R(z, u)'M(u)\Delta(u)U(w, u) \\ &= U(w, u)^{-1}M(u)^{-1}M(u)\Delta(u)U(w, u) = U(w, u)^{-1}\Delta(u)U(w, u), \end{aligned}$$

where a prime notation ($'$) indicates that a vector/matrix is transposed. Thus, the recovery problem of $\mathrm{pr}(w|u)$ from $U(w, u)$ rests on solving the eigenvalue problem of $P(z, w)^{-1}Q(z, w)$. Once $\mathrm{pr}(w|u)$ is known, we can evaluate causal effects by using the matrix adjustment method in Section 2.1. Based on this consideration, the following theorem can be obtained:

Theorem 1: *Under conditions (a),(b), (c) and (d), if U is a sufficient confounder relative to an ordered pair of variables (X, Y), then the causal effect $\mathrm{pr}\{y|\mathrm{do}(x)\}$ of X on Y is identifiable.*

The proof is provided in Appendix.

Here, it should be noted that $\mathrm{pr}(x, y, u)$ is not identifiable because we do not know whether $\mathrm{pr}(x, y, u_i) = \mathrm{pr}(x, y, u_{(i)})$ holds for $i = 1, ..., k$. That is, letting $\{\lambda_1, ..., \lambda_k\}$ be a set of eigenvalues of $P(z, w)^{-1}Q(z, w)$, we know that a set $\{\lambda_1, ..., \lambda_k\}$ of solutions of $|P(z, w)^{-1}Q(z, w) - \lambda I_k| = 0$ is consistent with a set $\{\mathrm{pr}(y|x, u_1), ..., \mathrm{pr}(y|x, u_k)\}$ of distributions, but we do not know which solution of $|P(z, w)^{-1}Q(z, w) - \lambda I_k| = 0$ corresponds to each $\mathrm{pr}(y|x, u_i)(i = 1, ..., k)$. The causal effect is nevertheless identifiable because it involves the summation over $U = u$, not the individual solutions of $|P(z, w)^{-1}Q(z, w) - \lambda I_k| = 0$.

If, on the other hand, we have knowledge of the correspondence, e.g., by establishing the orders $\lambda_1 > ... > \lambda_k$ and $\mathrm{pr}(y|x,u_1) > ... > \mathrm{pr}(y|x,u_k)$ and X is a discrete variable with a given finite number of categories, then the condition $W \perp\!\!\!\perp \{Z,Y\}|U$ can be relaxed to $W \perp\!\!\!\perp \{Z,Y\}|\{U,X\}$. To see this, when we replace $U(w,u)$ by a k dimensional matrix

$$U_x(w,u) = \begin{pmatrix} 1 & \mathrm{pr}(w_1|x,u_{(1)}) & \cdots & \mathrm{pr}(w_{k-1}|x,u_{(1)}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{pr}(w_1|x,u_{(k)}) & \cdots & \mathrm{pr}(w_{k-1}|x,u_{(k)}) \end{pmatrix},$$

both $P(z,w) = R(z,u)'M(u)U_x(w,u)$ and $Q(z,w) = R(z,u)'M(u)\Delta(u)U_x(w,u)$ hold. If both $P(z,w)$ and $Q(z,w)$ are invertible, then using the steps in Appendix, the causal effect is identifiable.

This deviation demonstrates that, whenever we observe two independent proxy variables associated with an unmeasured confounder, the distribution of the latter can be constructed from the proxies, which renders the causal effect identifiable. Thus, our result extends the range of solvable identification problems (Pearl, 2009, Chapters. 3 and 4; Shpitser and Pearl, 2006; Tian and Pearl, 2007) to cases where discrete confounders are measured with error. However, it should be noted that the identifiability criteria developed in (Pearl, 2009; Shpitser and Pearl, 2006; Tian and Pearl, 2003) apply to nonparametric models where the dimensionality of the variables is assumed arbitrary, while our result applies to causal models with finitely discretized confounder. Our method also provides guidance on how to choose proxy variables so as to construct the distribution of the unmeasured confounders from the proxies.

4. Effect Restoration in Linear Structural Equation Models

4.1. Linear Structural Equation Model

In this section, we assume each child-parent family in the graph G represents a linear structural equation model (SEM)

$$V_i = \sum_{V_j \in \mathrm{pa}(V_i)} \alpha_{v_i v_j} V_j + \epsilon_{v_i}, \qquad i = 1, 2, \ldots, n, \tag{11}$$

where normal random disturbances $\epsilon_{v_1}, \epsilon_{v_2}, \ldots, \epsilon_{v_n}$ are assumed to be independent of each other and have mean 0. In addition, $\alpha_{v_i v_j}$ is a constant value, and $\alpha_{v_i v_j}(\neq 0)$ is called a path coefficient or a direct effect. For the details on linear structural equation models, see Bollen (1989).

The following notation will be used in our discussion. For univariates X and Y and a set $\boldsymbol{Z}$ of variables, let $\sigma_{xy \cdot z} = \text{cov}(X, Y | \boldsymbol{Z} = \boldsymbol{z})$ and $\sigma_{xx \cdot z} = \text{var}(X | \boldsymbol{Z} = \boldsymbol{z})$ and $\beta_{yx \cdot z} = \sigma_{xy \cdot z} / \sigma_{xx \cdot z}$. For disjoint sets $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$, let $\Sigma_{xy \cdot z}$ be the conditional covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z} = \boldsymbol{z}$. In addition, let $\Sigma_{xx \cdot z}$ be the conditional covariance matrix of $\boldsymbol{X}$ given $\boldsymbol{Z} = \boldsymbol{z}$, and let $B_{yx \cdot z} = \Sigma_{yx \cdot z} \Sigma_{xx \cdot z}^{-1}$ be the regression coefficient matrix of $\boldsymbol{x}$ in the regression of $\boldsymbol{Y}$ on $\boldsymbol{X} \cup \boldsymbol{Z}$. We use the same notation in the case where either $\boldsymbol{X}$ or $\boldsymbol{Y}$ is univariate. When $\boldsymbol{Z}$ is an empty set, $\boldsymbol{Z}$ will be omitted from the expressions above. Similar notation is used for other parameters. Note the critical distinction between α_{yx} and β_{yx}. The former are structural coefficients that convey causal information, the latter are regression coefficients which are purely statistical.

The total effect τ_{yx} of X on Y is defined as the total sum of the products of the path coefficients on the sequence of arrows along all directed paths from X to Y. τ_{yx} can often be identified from graphs using the *back door criterion*. That is, if a set $\boldsymbol{S}$ of observed variables satisfies the back door criterion relative to an ordered pair of variables (X, Y), then the total effect τ_{yx} of X on Y is identifiable, and is given by the regression coefficient $\beta_{yx \cdot s}$ (Pearl, 2009).

Another identification condition invokes an instrumental variable (IV) (Brito and Pearl, 2002). Let $\{X, Y, Z\}$ and $\boldsymbol{S}$ be disjoint subsets of $\boldsymbol{V}$ in a directed acyclic graph G. If a set $\boldsymbol{S} \cup \{Z\}$ of variables satisfies (i) $\boldsymbol{S}$ contains no descendants of X or Y in G, and (ii) $\boldsymbol{S}$ d-separates Z from Y but not from X in the graph obtained by deleting all arrows emerging from X, then Z is said to be a *conditional instrumental variable* (CIV) given $\boldsymbol{S}$ relative to an ordered pair of variables (X, Y) (Pearl, 2009, p.366; see also Brito and Pearl, 2002). By CIV, we mean a variable that becomes an instrument relative to the target effect upon conditioning on a set $\boldsymbol{S}$ of variables. If an observed variable Z is a CIV given $\boldsymbol{S}$ relative to an ordered pair of variables (X, Y), then the total effect τ_{yx} of X on Y is identifiable, and is given by $\sigma_{yz \cdot s} / \sigma_{xz \cdot s}$ (Brito and Pearl, 2002). Especially, when $\boldsymbol{S}$ is an empty set, Z is called an instrumental variable (IV) (Bowden and Turkington, 1984).

To derive a new graphical identification condition for total effects, we review some properties of the regression coefficients. First, when $\{X, Y\} \cup \boldsymbol{S} \cup \boldsymbol{T}$ are normally distributed, we have the identity $\beta_{yx \cdot s} = \beta_{yx \cdot st} + B_{yt \cdot xs} B_{tx \cdot s}$ (Cochran, 1938). Second, if $\boldsymbol{T}$ is conditionally independent of X given $\boldsymbol{S}$ or Y is conditionally independent of $\boldsymbol{T}$ given $\{X\} \cup \boldsymbol{S}$, then $\beta_{yx \cdot st} = \beta_{yx \cdot s}$ (Wermuth, 1989). Third, $\beta_{yx \cdot z}$ is given by $\beta_{yx \cdot z} = (\sigma_{xy} - \Sigma_{xz} \Sigma_{zz}^{-1} \Sigma_{zy}) / (\sigma_{xx} - \Sigma_{xz} \Sigma_{zz}^{-1} \Sigma_{zx})$ because we
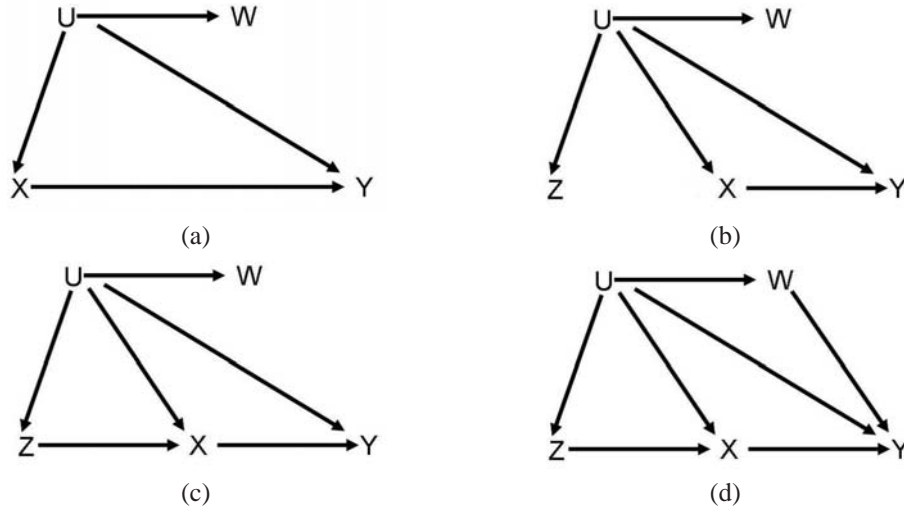
Fig.4: Linear SEMs with proxy variables of U, for the identification of τ_{yx}. (a) requires knowledge of $\alpha_{wu}^2 \sigma_{uu}$, while (b), (c) and (d) identify τ_{yx} from data.

have $\sigma_{xy \cdot z} = \sigma_{xy} - \Sigma_{xz} \Sigma_{zz}^{-1} \Sigma_{zy}$ and $\sigma_{xx \cdot z} = \sigma_{xx} - \Sigma_{xz} \Sigma_{zz}^{-1} \Sigma_{zx}$.

4.2.   Identification using proxy variables

In this section, we consider the linear version of the problem discussed in Section 3, i.e., estimating the total effect of X on Y when a sufficient covariate U is measured via proxy variables, as shown in Fig. 4.

The linear SEM offers two advantages in handling measurement errors. First, it provides a more transparent picture into the role of each factor in the model. Second, there are quite a few graphical structures in which the causal effect is identifiable in linear models but not in nonparametric models.

To see this, consider the causal diagrams shown in Fig.4. Since U is sufficient in Fig.4, the total effect is identifiable from the measurement on X, Y and U, and is given by $\tau_{yx} = \beta_{yx \cdot u}$, the regression coefficient of Y on X and U. However, if U is unobserved and W is but a noisy measurement of U, as in Fig.4 (a), knowledge of the error mechanism $W = \alpha_{wu} U + \epsilon_w$ is needed in order to identify $\tau_{yx} = \beta_{yx \cdot u}$. We note, however, that knowledge of both α_{wu} and σ_{uu} is not

necessary; the product $\alpha_{wu}^2 \sigma_{uu}$ is sufficient. To see this, we write

$$\tau_{yx} = \beta_{yx \cdot u} \quad = \quad \frac{\sigma_{xy} - \dfrac{\sigma_{xu}\sigma_{yu}}{\sigma_{uu}}}{\sigma_{xx} - \dfrac{\sigma_{xu}^2}{\sigma_{uu}}} = \frac{\sigma_{xy} - \dfrac{\alpha_{wu}^2 \sigma_{xu}\sigma_{yu}}{\alpha_{wu}^2 \sigma_{uu}}}{\sigma_{xx} - \dfrac{\alpha_{wu}^2 \sigma_{xu}^2}{\alpha_{wu}^2 \sigma_{uu}}}$$

and, from $\sigma_{xw} = \sigma_{xu}\alpha_{wu}$ and $\sigma_{yw} = \sigma_{yu}\alpha_{wu}$, we have

$$\tau_{yx} \quad = \quad \frac{\sigma_{xy} - \dfrac{\sigma_{xw}\sigma_{yw}}{\alpha_{wu}^2 \sigma_{uu}}}{\sigma_{xx} - \dfrac{\sigma_{xw}^2}{\alpha_{wu}^2 \sigma_{uu}}} \tag{12}$$

We see that, if $\alpha_{wu}^2 \sigma_{uu}$ is given, τ_{yx} is identifiable.

Next, we consider the identification of τ_{yx} without external information. We first show that if U possesses two independent proxy variables, say W and Z (as in Fig. 4(b)) then $\alpha_{wu}^2 \sigma_{uu}$ is identifiable. Indeed, writing $\sigma_{xw} = \alpha_{wu}\alpha_{xu}\sigma_{uu}$, $\sigma_{wz} = \alpha_{wu}\alpha_{zu}\sigma_{uu}$ and $\sigma_{xz} = \alpha_{xu}\alpha_{zu}\sigma_{uu}$, we have

$$\frac{\sigma_{xw}\sigma_{wz}}{\sigma_{xz}} = \frac{(\alpha_{wu}\alpha_{xu}\sigma_{uu})(\alpha_{wu}\alpha_{zu}\sigma_{uu})}{\alpha_{xu}\alpha_{zu}\sigma_{uu}} = \alpha_{wu}^2 \sigma_{uu}. \tag{13}$$

By substituting equation (13) into equation (12), we can see that $\tau_{yx}(= \alpha_{yx})$ is identifiable and is given by

$$\tau_{yx} = \beta_{yx \cdot u} = \frac{\sigma_{xy}\sigma_{wz} - \sigma_{xz}\sigma_{yw}}{\sigma_{xx}\sigma_{wz} - \sigma_{xw}\sigma_{xz}}. \tag{14}$$

This result reflects the well known fact (e.g., (Bollen, 1989, p. 224)) that, in linear SEMs, structural parameters are identifiable, up to a constant σ_{uu}, whenever each latent variable (in our case U) has three independent proxies (in our case X, W and Z). We see that the non-identifiability of σ_{uu} is not an impediment for the identification of α_{yx}.

We next relax the requirement that U possesses three independent proxies (as in Fig. 4(b)) and consider a situation as in Fig. 4(c), where two of these proxies (X and Z) are dependent. Here we note that $\{X, U\}$ d-separates Y, Z and W from each other. Therefore, given X, the tuple Y, Z and W work as three independent indicators of U (i.e., Y, Z and W are conditionally independent of each other given $\{X, U\}$). This will permit us to identify the key factor, $\alpha_{wu}^2 \sigma_{uu}$ from the measurement of X, Y, Z and W, and obtain:

$$\alpha_{wu}^2 \sigma_{uu} = \frac{\sigma_{yw \cdot x}\sigma_{wz \cdot x}}{\sigma_{yz \cdot x}} + \frac{\sigma_{xw}^2}{\sigma_{xx}}. \tag{15}$$

The derivation is as follows. Since $\sigma_{yw\cdot x} = \sigma_{wu\cdot x}\sigma_{yu\cdot x}/\sigma_{uu\cdot x}$, $\sigma_{wz\cdot x} = \sigma_{wu\cdot x}\sigma_{zu\cdot x}/\sigma_{uu\cdot x}$ and

$\sigma_{yz\cdot x} = \sigma_{yu\cdot x}\sigma_{zu\cdot x}/\sigma_{uu\cdot x}$, we have

$$\frac{\sigma_{yw\cdot x}\sigma_{wz\cdot x}}{\sigma_{yz\cdot x}} = \frac{(\sigma_{wu\cdot x}\sigma_{yu\cdot x}/\sigma_{uu\cdot x})(\sigma_{wu\cdot x}\sigma_{zu\cdot x}/\sigma_{uu\cdot x})}{\sigma_{yu\cdot x}\sigma_{zu\cdot x}/\sigma_{uu\cdot x}} = \frac{\sigma_{wu\cdot x}^2}{\sigma_{uu\cdot x}} = \beta_{wu\cdot x}^2\sigma_{uu\cdot x} = \beta_{wu}^2\sigma_{uu\cdot x}.$$

Further, noting that $\beta_{wu} = \alpha_{wu}$ and $\sigma_{xw} = \beta_{wu}\sigma_{xu} = \alpha_{wu}\sigma_{xu}$, we have

$$\alpha_{wu}^2\sigma_{uu\cdot x} = \alpha_{wu}^2\sigma_{uu} - \frac{\alpha_{wu}^2\sigma_{xu}^2}{\sigma_{xx}} = \alpha_{wu}^2\sigma_{uu} - \frac{\sigma_{xw}^2}{\sigma_{xx}}.$$

Using these results, equation (15) is obtained. The first term of equation (15) can be interpreted as the conditional modified-adjustment of U through the proxy variable W given X, and the second is a correction term, which transforms the conditional modified-adjustment of U through W given X to the unconditional modified-adjustment of U through W.

To derive an explicit expression for α_{yx}, we substitute equation (15) into equation (12), and using $\sigma_{yw\cdot x} = \sigma_{yw} - \sigma_{xy}\sigma_{xw}/\sigma_{xx}$ we have

$$
\begin{aligned}
\alpha_{yx} &= \frac{\sigma_{xy} - \dfrac{\sigma_{xw}\sigma_{yw}}{\alpha_{wu}^2\sigma_{uu}}}{\sigma_{xx} - \dfrac{\sigma_{xw}^2}{\alpha_{wu}^2\sigma_{uu}}} = \frac{\sigma_{xy}(\sigma_{yw\cdot x}\sigma_{wz\cdot x}\sigma_{xx} + \sigma_{yz\cdot x}\sigma_{xw}^2) - \sigma_{xw}\sigma_{yw}\sigma_{xx}\sigma_{yz\cdot x}}{\sigma_{xx}(\sigma_{yw\cdot x}\sigma_{wz\cdot x}\sigma_{xx} + \sigma_{yz\cdot x}\sigma_{xw}^2) - \sigma_{xw}^2\sigma_{xx}\sigma_{yz\cdot x}} \\[2mm]
&= \frac{\sigma_{xy}\sigma_{yw\cdot x}\sigma_{wz\cdot x}\sigma_{xx} + \sigma_{yz\cdot x}\sigma_{xw}(\sigma_{xy}\sigma_{xw} - \sigma_{yw}\sigma_{xx})}{\sigma_{yw\cdot x}\sigma_{wz\cdot x}\sigma_{xx}^2} \\[2mm]
&= \frac{\sigma_{xy}\sigma_{yw\cdot x}\sigma_{wz\cdot x}\sigma_{xx} - \sigma_{xx}\sigma_{yz\cdot x}\sigma_{xw}\sigma_{yw\cdot x}}{\sigma_{yw\cdot x}\sigma_{wz\cdot x}\sigma_{xx}^2} = \frac{\sigma_{xy}\sigma_{wz\cdot x} - \sigma_{yz\cdot x}\sigma_{xw}}{\sigma_{wz\cdot x}\sigma_{xx}}.
\end{aligned}
\tag{16}
$$

We see that $\alpha_{yx} = \beta_{yx\cdot u}$ is identifiable and is given by equation (14).

From Fig. 4 (a), (b) and (c), we see that the pivotal quantity needed for the identification of α_{yx} is the product

$$\alpha_{wu}^2\sigma_{uu} = \sigma_{ww} - \sigma_{\epsilon_w\epsilon_w}, \tag{17}$$

which stands for the portion of σ_{ww} that is contributed by variations of U. As seen from the consideration above, if we are in possession of several proxies for U, then $\alpha_{wu}^2\sigma_{uu}$ can be estimated from the data as in equation (13) or (15), yielding equation (12). If however U has only one proxy W, as Fig.4 (a), the product $\alpha_{wu}^2\sigma_{uu}$ must be estimated externally, using either a pilot study or judgmental assessment.

Judgmental assessment of the product $\alpha_{wu}^2\sigma_{uu}$ can be made more meaningful through the decomposition on the right hand side of equation (17), since both α_{wu} and ϵ_w are causal parameters

of the error mechanism $W = \alpha_{wu}U + \epsilon_w$, $\alpha_{wu} = E(W|u)/u$ measures the slope with which the average of W tracks the value of U, while $\sigma_{\epsilon_w \epsilon_w}$ measures the dispersion of W around that average. σ_{ww} can, of course be estimated from the data.

Under a Gaussian distribution assumption, α_{wu} and $\sigma_{\epsilon_w \epsilon_w}$ fully characterize the conditional density $f(w|u)$ which, according to Section 2, is sufficient for restoring the joint distribution of x, y and u, and thus secure the identification of the causal effect, through equation (2). This explains why the estimation of α_{wu} alone, be it from experimental data or our understanding of the physics behind the error process, is not sufficient for neutralizing the confounder U. It also explains why the technique of *latent factor* analysis (Bollen, 1989) is sufficient for identifying causal effects, even though it fails to identify the *factor loading* α_{wu} separately of σ_{uu}.

In the noiseless case, i.e., $\sigma_{\epsilon_w \epsilon_w} = 0$, we have $\sigma_{uu} = \sigma_{ww}/\alpha_{wu}^2$ and equation (12) reduces to:

$$\alpha_{yx} = \frac{\sigma_{xy} - \dfrac{\sigma_{xw}\sigma_{yw}}{\sigma_{ww}}}{\sigma_{xx} - \dfrac{\sigma_{xw}^2}{\sigma_{ww}}} = \frac{\sigma_{yx \cdot w}}{\sigma_{xx \cdot w}} = \beta_{yx \cdot w}, \qquad (18)$$

where $\beta_{yx \cdot w}$ is the regression coefficient of x in the regression model of Y on X and W, or:

$$\beta_{yx \cdot w} = \frac{\partial}{\partial x} E(Y|x, w).$$

As expected, the equality $\alpha_{yx} = \beta_{yx \cdot u} = \beta_{yx \cdot w}$ assures a bias-free estimate of α_{yx} through adjustment for W, instead of U; α_{wu} plays no role in this adjustment.

In the error-prone case, α_{yx} can be written as

$$\alpha_{yx} = \frac{\beta_{yx} - \dfrac{\beta_{yw}\beta_{wx}}{k}}{1 - \dfrac{\beta_{xw}\beta_{wx}}{k}},$$

where $k = 1 - \sigma_{\epsilon_w \epsilon_w}/\sigma_{ww}$ and, as the formula reveals, α_{yx} cannot be interpreted in terms of an adjustment for a proxy variable W.

The strategy of adjusting for a proxy variable has served as an organizing principle for many studies in traditional measurement error analysis (Carroll et al., 2006). For example, if one seeks to estimate the regression coefficient $\beta_{xu} = E(X|u)/u$ through a proxy W of U, one can always choose to regress X on another variable, V, such that the slope of X on V, $E(X|v)/v$, would yield an unbiased estimate of β_{xu}. Choosing V to be the best linear estimate of U, given W would permit

such regression. In our example of Fig.4 (b), one should choose $V = \gamma W$, where

$$\gamma = \frac{\sigma_{vw}}{\sigma_{ww}} = \frac{\alpha_{wu}\sigma_{uu}}{\sigma_{ww}}$$

is to be estimated separately, from a pilot study. However, this Two Stage Least Square strategy is not applicable in adjusting for latent confounders; i.e., there is no variable $V(W)$ such that $\alpha_{yx} = \beta_{yx \cdot w}$.

Fig. 4 (d) represents a new challenge; although $\alpha_{wu}^2 \sigma_{uu}$ is not identifiable, the total effect α_{yx} is nevertheless identifiable without external studies. In the next section, we will discuss this identification strategy.

4.3. Instrumental Variable (IV) method with a proxy variable

In Fig.4 (d), if U can be observed, then both the CIV condition and the back door criterion can be applied to evaluating the total effect simultaneously, giving $\tau_{yx} = \beta_{yx \cdot u}$ and $\tau_{yx} = \sigma_{yz \cdot u}/\sigma_{xz \cdot u}$, respectively. We shall now show that equating these two expressions to each other, together with the independence condition $\{X, Z\} \perp\!\!\!\perp W | U$ will allow us to remove all terms involving u as a subscript. Indeed, starting with $\sigma_{xw} = \sigma_{xu}\sigma_{wu}/\sigma_{uu}$ and $\sigma_{wz} = \sigma_{zu}\sigma_{wu}/\sigma_{uu}$, we have $\sigma_{zu} = \sigma_{xu}\sigma_{wz}/\sigma_{xw}$. Then, using

$$\tau_{yx} = \beta_{yx \cdot u} = \frac{\sigma_{xy \cdot u}}{\sigma_{xx \cdot u}} = \frac{\sigma_{xy} - \dfrac{\sigma_{xu}\sigma_{yu}}{\sigma_{uu}}}{\sigma_{xx} - \dfrac{\sigma_{xu}^2}{\sigma_{uu}}},$$

we have

$$\left(\sigma_{xx} - \frac{\sigma_{xu}^2}{\sigma_{uu}}\right)\tau_{yx} = \sigma_{xy} - \frac{\sigma_{xu}\sigma_{yu}}{\sigma_{uu}},$$

and, from $\tau_{yx} = \sigma_{yz \cdot u}/\sigma_{xz \cdot u}$ and $\sigma_{zu} = \sigma_{xu}\sigma_{wz}/\sigma_{xw}$, we have

$$\left(\sigma_{xz} - \frac{\sigma_{xu}\sigma_{zu}}{\sigma_{uu}}\right)\tau_{yx} = \sigma_{yz} - \frac{\sigma_{zu}\sigma_{yu}}{\sigma_{uu}}, \quad \text{that is,} \quad \left(\sigma_{xz} - \frac{\sigma_{wz}}{\sigma_{xw}}\frac{\sigma_{xu}^2}{\sigma_{uu}}\right)\tau_{yx} = \sigma_{yz} - \frac{\sigma_{wz}}{\sigma_{xw}}\frac{\sigma_{xu}\sigma_{yu}}{\sigma_{uu}}.$$

By solving these equations for τ_{yx}, we obtain

$$\tau_{yx} = \frac{\sigma_{yz} - \sigma_{xy}\dfrac{\sigma_{wz}}{\sigma_{xw}}}{\sigma_{xz} - \sigma_{xx}\dfrac{\sigma_{wz}}{\sigma_{xw}}},$$

which is consistent with equation (12). This derivation demonstrates a more general approach that differs from Cai and Kuroki (2007) which was based on latent factor analysis (e.g. Bollen, 1989;
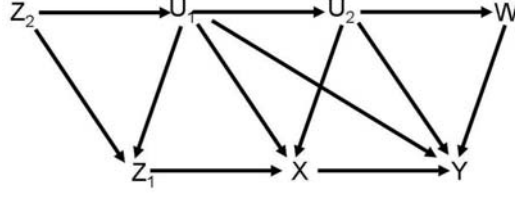
Fig.5: Causal diagram with unmeasured confounders

Stanghellini, 2004; Stanghellini and Wermuth, 2005; Vicard, 2000). Our approach extends the identification conditions to cases where the total effect can not be identified by any single strategy but by a combination of several strategies (e.g., the back door criterion and the CIV condition in this case). In addition, unlike the discussion in Section 4.2, the identification of $\alpha_{wu}^2 \sigma_{uu}$ is not required; instead, we will require a proxy variable W such that U d-separates W from $\{X, Z\}$.

The power of this approach can be demonstrated in the model of Fig.5 where a sufficient set $\{U_1\} \cup \boldsymbol{U}_2$ of variables is unobserved. Here, U_1 is univariate but the number of variables in $\boldsymbol{U}_2$ can be uncertain. In this situation, the back door criterion can not be used to identify the total effect of X on Y, and the uncertain number of variables in $\boldsymbol{U}_2$ prevents us from identifying the total effect based on latent factor analysis in which we need to know the number of unobserved variables. In addition, because neither Z_1 nor Z_2 is (conditionally) independent of $\{U_1\} \cup \boldsymbol{U}_2$, they can not be used as the CIVs. Nevertheless, we will show that the total effect is identifiable as follows: Since both Z_1 and Z_2 are CIV given U_1 relative to an ordered pair of variables (X, Y), the total effect is given by

$$\tau_{yx} = \frac{\sigma_{yz_1 \cdot u_1}}{\sigma_{xz_1 \cdot u_1}} = \frac{\sigma_{yz_2 \cdot u_1}}{\sigma_{xz_2 \cdot u_1}}$$

Moreover, since $\{Z_1, Z_2\} \perp\!\!\!\perp W | U_1$ holds in the model, we have $\sigma_{z_i w} = \sigma_{z_i u_1} \sigma_{wu_1} / \sigma_{u_1 u_1}$ ($i = 1, 2$), and we can write

$$\sigma_{yz_1} - \frac{\sigma_{yu_1} \sigma_{z_2 u_1}}{\sigma_{u_1 u_1}} \frac{\sigma_{z_1 w}}{\sigma_{z_2 w}} = \tau_{yx} \left(\sigma_{xz_1} - \frac{\sigma_{xu_1} \sigma_{z_2 u_1}}{\sigma_{u_1 u_1}} \frac{\sigma_{z_1 w}}{\sigma_{z_2 w}} \right), \quad \text{and,} \, \sigma_{yz_2} - \frac{\sigma_{yu_1} \sigma_{z_2 u_1}}{\sigma_{u_1 u_1}} = \tau_{yx} \left(\sigma_{xz_2} - \frac{\sigma_{xu_1} \sigma_{z_2 u_1}}{\sigma_{u_1 u_1}} \right).$$

By solving these equations for τ_{yx}, we have

$$\tau_{yx} = \frac{\sigma_{z_1 y} \sigma_{z_2 w} - \sigma_{z_2 y} \sigma_{z_1 w}}{\sigma_{z_1 x} \sigma_{z_2 w} - \sigma_{z_2 x} \sigma_{z_1 w}}. \tag{19}$$

We now summarize these considerations in a theorem.

Theorem 2: *Suppose that*

(i) a non-empty set $\{Z_1, Z_2\}$ of distinct variables satisfies one of the following conditions: (i-a) both Z_1 and Z_2 are CIVs given a univariate U relative to an ordered pair of variables (X, Y), (i-b) Z_1 is a CIV given U relative to an ordered pair of variables (X, Y), and $Z_2 = X$ and U satisfies the back door criterion relative to an ordered pair of variables (X, Y).

(ii) U d-separates $\{Z_1, Z_2\}$ from an observed variable W.

Then, the total effect τ_{yx} of X on Y is identifiable and is given by the formula (19). □

5. Conclusion

The paper discusses computational and representational problems connected with effect restoration when confounders are mismeasured or misclassified. In particular, we have explicated how measurement bias can be removed by creating synthetic samples from empirical samples, and how inverse-probability weighting can be modified to account for measurement error. These techniques required an estimate of the noise mechanism, which can be obtained from external studies or assessed judgmentally. Subsequently, we have derived conditions under which causal effects can be restored without resorting to external studies, provided the confounder is discrete and is measured through proxies of sufficiently high cardinality. Finally, we have analyzed measurement bias in linear systems and explicated graphical conditions under which such bias can be removed.

Appendix: The proof of Theorem 1

The proof of Theorem 1 is based on the following two-step procedure which recovers $\mathrm{pr}(x, y, u)$ from $\mathrm{pr}(x, y, z, w)$.

Step 1: Solve an eigenvalue problem of $P(z, w)^{-1}Q(z, w)$ to recover $\mathrm{pr}(w|u)$ from $U(w, u)$.

Step 2: Recover $\mathrm{pr}(x, y, u)$ using the matrix adjustment method introduced in Section 2.1.

Step 1: To find $\mathrm{pr}(w|u)$ encoded in $U(w, u)$, in terms of observed probabilities, let us consider the eigenvalue problem of $P(z, w)^{-1}Q(z, w)$. First, noting that $|U(w, u)^{-1}| = 1/|U(w, u)|$, we solve $|P(z, w)^{-1}Q(z, w) - \lambda I_k| = 0$ for λ to obtain the set of eigenvalues of $P(z, w)^{-1}Q(z, w)$. In other words, λ should satisfy

$$|P(z, w)^{-1}Q(z, w) - \lambda I_k| = |U(w, u)^{-1}\Delta(u)U(w, u) - \lambda U(w, u)^{-1}U(w, u)|$$

$$= |U(w,u)^{-1}||\Delta(u) - \lambda I_k||U(w,u)| = |\Delta(u) - \lambda I_k|$$

$$= (\mathrm{pr}(y|x, u_{(1)}) - \lambda)...(\mathrm{pr}(y|x, u_{(k)}) - \lambda) = 0. \tag{20}$$

From this equation, letting $\lambda_1 > ... > \lambda_k$ for eigenvalues of $P(z,w)^{-1}Q(z,w)$, we have $\lambda_i = \mathrm{pr}(y|x, u_{(i)})$ $(i = 1,...,k)$, thus the elements of $\Delta(u)$ are estimable. In order to obtain the eigenvector $\boldsymbol{\eta}_i$ for λ_i, letting $H = (\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_k)$, we solve the following simultaneous linear equations

$$(P(z,w)^{-1}Q(z,w) - \lambda_i I_k)\boldsymbol{\eta}_i = \mathbf{0}, \quad i = 1,...,k \tag{21}$$

or, equivalently,

$$P(z,w)^{-1}Q(z,w)H = (\lambda_1\boldsymbol{\eta}_1, ..., \lambda_k\boldsymbol{\eta}_k) = (\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_k)\begin{pmatrix} \lambda_1 & 0 & \cdots \\ 0 & \ddots & \ddots \\ \vdots & 0 & \lambda_k \end{pmatrix} = H\Delta(u).$$

Here, it is noted that $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_k$ are uniquely determined except for a multiplicative constant because $\lambda_1, ..., \lambda_k$ take different values according to assumption (d). On the other hand, letting $A = U(w,u)^{-1}E$ and $E = \mathrm{diag}(\alpha_1, ..., \alpha_k)$ for any non-zero values of $\alpha_1, ..., \alpha_k$, we have

$$\begin{aligned} \left\{P(z,w)^{-1}Q(z,w)\right\}A &= \left\{U(w,u)^{-1}\Delta(u)U(w,u)\right\}\left\{U(w,u)^{-1}E\right\} \\ &= U(w,u)^{-1}\Delta(u)E = U(w,u)^{-1}E\Delta(u) = A\Delta(u), \end{aligned}$$

This means that A is also a matrix from eigenvectors of $P(x,z)^{-1}Q(x,z)$ and we have $A(= U(w,u)^{-1}E) = H$ by taking certain values of $\alpha_1, ..., \alpha_k$. Then, for the inverse $H^{-1} = (h_{ij})$ of the estimable matrix H, we have using $U(w,u)^{-1}E = H$,

$$U(w,u) = \begin{pmatrix} 1 & \mathrm{pr}(w_1|u_{(1)}) & ... & \mathrm{pr}(w_{k-1}|u_{(1)}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{pr}(w_1|u_{(k)}) & ... & \mathrm{pr}(w_{k-1}|u_{(k)}) \end{pmatrix} = EH^{-1} = \begin{pmatrix} \alpha_1 h_{11} & ... & \alpha_1 h_{1k} \\ \vdots & \vdots & \vdots \\ \alpha_k h_{k1} & & ...\alpha_k h_{kk} \end{pmatrix}.$$

Equating the first column of both hand sides of the equation, the diagonal element $\alpha_1 = 1/h_{11}, ..., \alpha_k = 1/h_{k1}$ of E can be obtained, which indicates that $U(w,u)$ is identifiable from EH^{-1}, since H^{-1} is estimable. Thus, every element $\mathrm{pr}(w|u)$ of $U(w,u)$ can be obtained.

Step 2: To express $\mathrm{pr}(x,y,u)$ in terms of observed probabilities, we use the matrix adjustment method introduced in Section 2.1. Since we have

$$\mathrm{pr}(x,y,w) = \sum_{i=1}^{k} \mathrm{pr}(x,y,u_i)\mathrm{pr}(w|u_i) = \sum_{i=1}^{k} \mathrm{pr}(x,y,u_{(i)})\mathrm{pr}(w|u_{(i)}),$$

substitute elements of $\mathrm{pr}(w_i|u_{(j)})$ $(i, j = 1, ..., k)$ obtained in Step 1 for $M(w, u)$ in equation (5). Then, if $M(w, u)$ is invertible, we can obtain elements of $V_{xy}(u)$. Thus, the causal effect

$$\mathrm{pr}\{y|\mathrm{do}(x)\} = \sum_{i=1}^{k} \mathrm{pr}(y|x, u_i)\mathrm{pr}(u_i) = \sum_{i=1}^{k} \mathrm{pr}(y|x, u_{(i)})\mathrm{pr}(u_{(i)}) = \sum_{i=1}^{k} \frac{\mathrm{pr}(x, y, u_{(i)})}{\mathrm{pr}(x, u_{(i)})}\mathrm{pr}(u_{(i)})$$

is identifiable.

Acknowledgement

References

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Bowden, R. J., and Turkington, D. A. (1984). *Instrumental variables*. Cambridge University Press.

Brito, C. and Pearl, J. (2002). Generalized instrumental variables. *Proceeding of the 18th Conference on Uncertainty in Artificial Intelligence*, 85-93.

Cai, Z. and Kuroki, M. (2008). On identifying total effects in the presence of latent variables and selection bias. *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence*, 62-69.

Carroll, R., Ruppert, D., Stefanski, L. and Crainiceanu, C. (2006). *Measurement error in nonlinear Models: A modern perspective. 2nd ed.* Chapman & Hall/CRC, Boca Raton, FL.

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society*, **5**, 171-176.

Goetghebeur, E. and Vansteelandt, S. (2005). Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research*, **14**, 397-415.

Greenland, S. (2005). Multiple-bias modeling for analysis of observational data. *Journal of the Royal Statistical Society: Series A*, **168**, 267-306.

Greenland, S. and Kleinbaum, D. (1983). Correcting for misclassification in two-way tables and matched-pair studies. *International Journal of Epidemiology*, **12**, 93-97.

Greenland, S. and Lash, T. (2008). Bias analysis. In *Modern epidemiology* (K. Rothman, S. Greenland and T. Lash, eds.), 3rd ed. Lippincott Williams and Wilkins, Philadelphia, PA, 345-380.

Hernán, M. and Cole, S. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, **170**, 959-962.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (2009). *Causality: Models, reasoning, and inference. 2nd ed.*. Cambridge University Press.

Pearl, J. (2010). On measurement bias in causal inference. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, 425–432.

Pearl, J. and Bareinboim, E. (2011). Transportability across studies: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 247–254.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Schneeweiss, S., Rassen, J., Glynn, R., Avorn, J., Mogun, H. and Brookhart, M. (2009). Highdimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, **20**, 512-522.

Selén, J. (1986). Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *Journal of the American Statistical Association*, **81**, 75-81.

Shpitser,I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. *Proceedings of the National Conference on Artificial Intelligence*, **21**, 1219-1226.

Stanghellini, E. (2004). Instrumental variables in Gaussian directed acyclic graph models with an unobserved confounder. *Environmetrics*, **15**, 463-469.

Stanghellini, E. and Wermuth, N. (2005). On the identification of directed acyclic graph models with one hidden variable. *Biometrika*, **92**, 337-350.

Stürmer, T., Schneeweiss, S., Avorn, J. and Glynn, R. (2005). Adjusting effect estimates for unmeasured confounding in cohort studies with validation studies using propensity score calibration.

American Journal of Epidemiology, **162**, 279-289.

Tian, J. and Pearl, J. (2003).   On the identification of causal effects.   UCLA Cognitive Systems Laboratory, Technical Report (R-290-L).

Vicard, P. (2000). On the identification of a single factor model with correlated residuals. *Biometrika*, **87**, 199-205.

Wermuth, N. (1989). Moderating effects in multivariate normal distributions. *Methodika*, **3**, 74-93.