Biometrika, 101(2), 423-437, 2014.
Biometrika Advance Access published March 18, 2014
TECHNICAL REPORT
R-366
March 2014

Biometrika (2014), *pp.* 1–15
Printed in Great Britain
doi: 10.1093/biomet/ast066

Measurement bias and effect restoration in causal inference

BY MANABU KUROKI

Department of Data Science, The Institute of Statistical Mathematics, 10-3, Midori-cho, Tachikawa, Tokyo, 190-8562, Japan
mkuroki@ism.ac.jp

AND JUDEA PEARL

Department of Computer Science, University of California, Los Angeles, 4532 Boelter Hall, Los Angeles, California, 90095-1596, U.S.A.
judea@cs.ucla.edu

SUMMARY

This paper highlights several areas where graphical techniques can be harnessed to address the problem of measurement errors in causal inference. In particular, it discusses the control of unmeasured confounders in parametric and nonparametric models and the computational problem of obtaining bias-free effect estimates in such models. We derive new conditions under which causal effects can be restored by observing proxy variables of unmeasured confounders with/without external studies.

Some key words: Causal diagram; Confounder; Instrumental variable method; Proxy variable; Regression coefficient; Total effect.

1. INTRODUCTION

1·1. *Preliminaries*

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 & Lash (2008) that is rooted in classical regression analysis (Greenland & Kleinbaum, 1983; Selén, 1986; Carroll et al., 2006), but has not been fully utilized in causal analysis or graphical models.

For $V = (V_1, \ldots, V_n)$, let $\mathrm{pr}(v) = \mathrm{pr}(v_1, \ldots, v_n)$ be the joint distribution of $V = (V_1, \ldots, V_n) = (v_1, \ldots, v_n)$, $\mathrm{pr}(v_i \mid v_j)$ 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 $G = (V, E)$ with a set V of variables and a set E of arrows, a probability distribution $\mathrm{pr}(v)$ is said to be compatible with G if it can be factorized as

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

where $\mathrm{pa}(v_i)$ is a set of parents of V_i in G. When $\mathrm{pa}(v_i)$ is an empty set, $\mathrm{pr}\{v_i \mid \mathrm{pa}(v_i)\}$ is the marginal distribution $\mathrm{pr}(v_i)$. When equation (1) holds, we also say that G is a Bayesian network of $\mathrm{pr}(v)$ (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 X, Y and Z, if Z d-separates X from Y in G, then X is conditionally independent of Y given Z, denoted as $X \perp\!\!\!\perp Y \mid Z$, 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 (Pearl, 2009, p. 24). Based on a causal diagram G, for any variables $X, Y \in V$, the causal effect of X on Y is defined as

$$\mathrm{pr}\{y \mid \mathrm{do}(x)\} = \sum_{v \backslash \{x,y\}} \frac{\mathrm{pr}\{x, y, v \backslash \{x, y\}\}}{\mathrm{pr}\{x \mid \mathrm{pa}(x)\}},$$

for all x for which $\mathrm{pr}\{x \mid \mathrm{pa}(x)\} > 0$. The symbol $\mathrm{do}(x)$ 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 S of variables is said to satisfy the back door criterion relative to (X, Y) if it satisfies the following conditions:

(i) no vertex in S is a descendant of X, and
(ii) 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 $\mathrm{pr}\{y \mid \mathrm{do}(x)\} = \sum_s \mathrm{pr}(y \mid x, s)\mathrm{pr}(s)$ (Pearl, 2009, pp. 79–80); S is then called sufficient.

1·2. *Motivation*

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 via one or several proxy variables shown in Fig. 1.

Our motivation can be illustrated through Magidson's analysis of Head Start Program (Magidson, 1977), which is a program of the United States Department of Health and Human Services that provides comprehensive education, health, nutrition, and parent involvement services to low-income children and their families. The programme's services and resources are designed to foster stable family relationships, enhance children's physical and emotional well-being, and establish an environment to develop strong cognitive skills. Magidson's sample consists of 148 children who received the summer Head Start Program and 155 control children. Let X be a dummy variable indicating attendance in the Head Start Program, and let Y be the outcome variable of the Metropolitan Readiness Test, which is a measure of children cognitive ability. Let U represent the socio-economic status, which was unobserved but can be considered as a sufficient confounder based on the discussion in Magidson (1977).

Figure 1 shows several situations that may be realized of this study. Figure 1(a) depicts a situation where father's occupation W is measured as a proxy variable of U, and Fig. 1(b) depicts a situation where father's occupation W and family income Z are measured as proxy variables of U. Figure 1(c) also depicts a situation where both W and Z are measured as proxy variables of U. Different from Fig. 1(b), since family income Z is an important factor that determines whether or not a child attended the Head Start Program, it is reasonable to model Z as having a

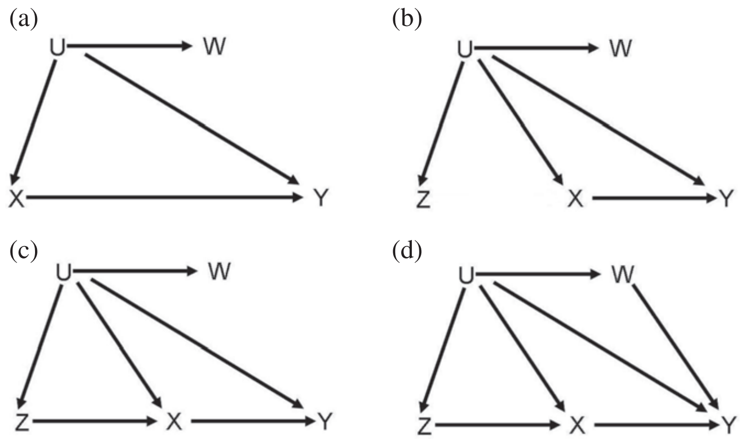

Fig. 1. Causal diagrams with proxy variables of U, for the identification of causal effects. (a) requires knowledge of $\text{pr}(w \mid u)$, while (b) and (c) identify causal effects from data. (d) identifies causal effects in linear models but not in nonparametric models from data.

direct effect on X. Finally, Fig. 1(d) depicts a situation where father's occupation, W, is assumed to have a direct effect on Y. Figures 1(b), (c) and (d) assume that father's occupation and family income are conditionally independent given U, which may be unrealistic and is relaxed later in § 4·3.

In Fig. 1, U satisfies the back door criterion relative to (X, Y), but its proxy variables W and Z do not. Since U is sufficient, the causal effect is identifiable from measurement on X, Y and U, and can be written as

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

However, since U is unobserved and both W and Z are noisy measurements of U, d-separation tells us immediately that adjusting for W and/or Z is inadequate, as 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 or estimate the conditional distribution $\text{pr}(w \mid u)$ that governs the proxy mechanism, we can perform a modified adjustment for W and Z that, in the limit of a 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 the actual value of a confounder U remains uncertain for each measurement $W = w$ and $Z = z$, so one would expect to get either a distribution over causal effects, or bounds thereof (MacLehose et al., 2005; Cai et al., 2008; Kuroki & Cai, 2008). Instead, we can actually get a repaired point estimate of $\text{pr}\{y \mid \text{do}(x)\}$ that is asymptotically unbiased.

This result, which we will label effect restoration, has powerful consequences in practice because, when $\text{pr}(w \mid u)$ is not given or estimated, one can resort to a Bayesian or bounding analysis and assume a prior distribution or bounds on the parameters of $\text{pr}(w \mid u)$, which would yield a distribution or bounds over $\mathrm{pr}\{y \mid \mathrm{do}(x)\}$ (Greenland, 2005). Alternatively, if costs permit, one can estimate $\mathrm{pr}(w \mid u)$ by re-testing U in a sampled subpopulation (Carroll et al., 2006). This is normally done by re-calibration techniques (Greenland & 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). In our example of the Head Start Program, an estimate of $\mathrm{pr}(w \mid u)$ could be obtained by measuring dozens of indicators which, according to professional consensus, provide an accurate assessment of U.

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 measurements, sampling variability increases substantially with error. Second, even assuming unbounded sample size, the estimate will be biased if the postulated $\mathrm{pr}(w \mid u)$ is incorrect. In extreme cases, wrongly postulated $\mathrm{pr}(w \mid u)$ may even conflict with the data, and no estimate will be obtained. For example, if we postulate a noninformative W, $\mathrm{pr}(w \mid u) = \mathrm{pr}(w)$, and we find that W strongly depends on X, a contradiction arises and no effect estimate will emerge (Pearl, 2010).

Effect restoration can be analysed 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 is given the estimand in equation (2), the problem is no longer one of causal inference, but rather of regression analysis, whereby the expression $E_U\{\mathrm{pr}(y \mid x, U)\}$ must be estimated from a noisy measurement of U, given by W or $\{W, Z\}$. 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 commonly assumed in the measurement-error literature, but the causal effect $\mathrm{pr}\{y \mid \mathrm{do}(x)\}$ 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 modelling the error mechanism $\mathrm{pr}(w \mid u)$ requires causal considerations; errors caused by noisy measurements of U are fundamentally different from those caused by random factors affecting U. Likewise, the reason we seek an estimate of $\mathrm{pr}(w \mid u)$ as opposed to $\mathrm{pr}(u \mid w)$, 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 & Bareinboim, 2011; 2013 technical report available from author) 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, and 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 & Vansteelandt (2005), Cai & Kuroki (2008), Hernán & Cole (2009) and Pearl (2010), while graphical models are hardly used at all 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 $\mathrm{pr}(w \mid u)$ and $\mathrm{pr}(x, y, w)$, or $\mathrm{pr}(x, y, w, z)$, that are major contributors to measurement bias. We derive new conditions under which causal effects can be restored by observing proxy variables of unmeasured confounders with/without external studies.

2. Effect restoration with external studies

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

(a) the distribution $\mathrm{pr}(w \mid u)$ of the error mechanism is available from external studies such as pilot studies or scientific judgments, and
(b) W and the confounder U are discrete variables with a given finite number of categories k.

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

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

Then, for any specific values x and y, letting

$$V_{xy}(u) = \begin{pmatrix} \mathrm{pr}(y, u_1 \mid x) \\ \vdots \\ \mathrm{pr}(y, u_k \mid x) \end{pmatrix}, \quad V_{xy}(w) = \begin{pmatrix} \mathrm{pr}(y, w_1 \mid x) \\ \vdots \\ \mathrm{pr}(y, w_k \mid x) \end{pmatrix},$$

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

equations (3) can be written as matrix multiplication: $V_{xy}(w) = M(w, u)V_{xy}(u)$. Now, assuming that

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

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

$$V_{xy}(u) = M(w, u)^{-1} V_{xy}(w). \tag{4}$$

Similarly, the estimation of $\mathrm{pr}(x, u) = \mathrm{pr}(u \mid x)\mathrm{p}(x)$ can be achieved by replacing $V_{xy}(u)$ and $V_{xy}(w)$ with $V_x(u) = \{\mathrm{pr}(u_1 \mid x), \ldots, \mathrm{pr}(u_k \mid x)\}'$ and $V_x(w) = \{\mathrm{pr}(w_1 \mid x), \ldots, \mathrm{pr}(w_k \mid x)\}'$ respectively, and the estimation of $\mathrm{pr}(u)$ can be achieved by replacing $V_{xy}(u)$ and $V_{xy}(w)$ with $V(u) = \{\mathrm{pr}(u_1), \ldots, \mathrm{pr}(u_k)\}'$ and $V(w) = \{\mathrm{pr}(w_1), \ldots, \mathrm{pr}(w_k)\}'$ respectively, where primes stand for a transposed vector/matrix. Thus, equation (4) enables us to reconstruct $\mathrm{pr}(y, u \mid x)$, $\mathrm{pr}(u \mid x)$ and $\mathrm{pr}(u)$ from $\mathrm{pr}(x, y, w)$ and $\mathrm{pr}(w \mid u)$. In other words, each term on the right-hand side of equation (2) can be obtained from $\mathrm{pr}(x, y, w)$ and $\mathrm{pr}(w \mid u)$ through equation (4) and, assuming U is sufficient, the causal effect $\mathrm{pr}\{y \mid \mathrm{do}(x)\}$ is estimable from the available data. Explicitly, letting $i(w, u)$ be the corresponding element of $M(w, u)^{-1}$ that corresponds to $(W, U) = (w, u)$, we have

$$\mathrm{pr}\{y \mid \mathrm{do}(x)\} = \sum_{i=1}^{k} \frac{\mathrm{pr}(y, u_i \mid x)\mathrm{pr}(u_i)}{\mathrm{pr}(u_i \mid x)} = \sum_{l=1}^{k} \frac{\sum_{j=1}^{k} i(w_j, u_l)\mathrm{pr}(y, w_j \mid x) \sum_{j=1}^{k} i(w_j, u_l)\mathrm{pr}(w_j)}{\sum_{j=1}^{k} i(w_j, u_l)\mathrm{pr}(w_j \mid x)}.$$

The same inverse matrix, $M(w, u)^{-1}$, appears in all summations.

For full discussion of the proposed method in the case of a binary U and the application of the propensity score method (Rosenbaum & Rubin, 1983) to multivariate U, see Pearl (2010).

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 the causal diagrams shown in Figs. 1(b) and (c) that are obtained by adding an observed variable Z to Fig. 1(a). We will show that $\mathrm{pr}(y, u \mid x)$, $\mathrm{pr}(u \mid x)$ and $\mathrm{pr}(u)$ can be recovered from $\mathrm{pr}(x, y, z, w)$ under the following conditions:

Condition 1. Two proxy variables of U that are conditionally independent of each other given U can be observed, e.g., W and Z in Fig. 1(b) and (c) and U satisfies both $W \perp\!\!\!\perp \{X, Y, Z\} \mid U$ and $Y \perp\!\!\!\perp \{W, Z\} \mid \{U, X\}$, as in Fig. 1(b) and (c);

Condition 2. W, Z and the confounder U are discrete variables with a given finite number of categories, k.

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

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

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

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

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

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

Let

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

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

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

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

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

Assume further that

Condition 3. Both $P(z, w)$ and $Q(z, w)$ are invertible.

Condition 4. The probabilities $\mathrm{pr}(y \mid x, u_1), \ldots, \mathrm{pr}(y \mid x, u_k)$ take distinct values for given 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 $P(z, w)^{-1} Q(z, w) = U(w, u)^{-1} \Delta(u) U(w, u)$, Thus, the recovery problem of $\mathrm{pr}(w \mid u)$ from $U(w, u)$ rests on the eigenvalue decomposition of $P(z, w)^{-1} Q(z, w)$. Once $\mathrm{pr}(w \mid u)$ is known, we can evaluate causal effects by using the matrix adjustment method in § 2. Based on this consideration, the following theorem can be obtained.

THEOREM 1. *Under Conditions* 1, 2, 3, *and* 4, *if U is a sufficient confounder relative to* (X, Y), *then the causal effect* $\mathrm{pr}\{y \mid \mathrm{do}(x)\}$ *of X on Y is identifiable.*

The proof is provided in the Appendix.

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

This derivation 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 (Tian & Pearl, 2002; Kuroki, 2007; Pearl, 2009, Chs. 3 and 4; Shpitser & Pearl, 2006) to cases where discrete confounders are measured with error. However, it should be noted that the identifiability criteria developed in Pearl (2009), Shpitser & Pearl (2006) and Tian & Pearl (2002) apply to nonparametric models where the dimensionality of the variables is assumed arbitrary, while our result applies to finite and discrete confounders, that is, confounders such as gender and race that have known categories but remain unmeasured.

4. Effect restoration in linear structural equation models

4·1. *Linear structural equation model*

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

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

where the normal random disturbances $\epsilon_{v_1}, \ldots, \epsilon_{v_n}$ are assumed to be independent of each other and have mean zero. In addition, $\alpha_{v_i v_j}$ is a constant, and if nonzero 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, Y and a set Z of variables, let $\sigma_{xy\cdot z} = \mathrm{cov}(X, Y \mid Z = z)$ and $\sigma_{xx\cdot z} = \mathrm{var}(X \mid Z = z)$. In addition, let $\beta_{yx\cdot z} = \sigma_{xy\cdot z}/\sigma_{xx\cdot z}$ be the regression coefficient of x in the regression model of Y on X and Z. For disjoint sets X, Y and Z, let $\Sigma_{xy\cdot z}$ and $\Sigma_{xx\cdot z}$ be the conditional covariance matrices of X and Y given $Z = z$ and the covariance matrix of X given Z, respectively. When Z is an empty set, it will be omitted from the expressions above. Similar notation is used for other parameters. Then, we have $\sigma_{xy\cdot z} = \sigma_{xy} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zy}$ and $\sigma_{xx\cdot z} = \sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}$. Note the critical distinction between path coefficients and regression coefficients. The former convey causal information, the latter are regression coefficients, which are purely statistical measures.

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, and can often be identified from graphs using the back door criterion. That is, if a set S of observed variables satisfies the back door criterion relative to (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 (Brito & Pearl, 2002). Let $\{X, Y, Z\}$ and S be disjoint subsets of V in a directed acyclic graph G. If a set $S \cup \{Z\}$ of variables satisfies the following conditions, then Z is said to be a conditional instrumental variable given S relative to (X, Y) (Pearl, 2009, p. 366; Brito & Pearl, 2002):

(i) S contains no descendants of X or Y in G, and
(ii) S d-separates Z from Y but not from X in the graph obtained by deleting all arrows emerging from X.

By a conditional instrumental variable, we mean a variable that becomes an instrument relative to the target effect upon conditioning on a set S of variables. If an observed variable Z is a conditional instrumental variable given S relative to (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 & Pearl, 2002). Especially, when S is an empty set, Z is called an instrumental variable (Bowden & Turkington, 1984).

4·2. *Identification using proxy variables*

In this section, we consider the linear version of the problem discussed in § 3, i.e., estimating the total effect of X on Y when a sufficient confounder U is measured via proxy variables, as in Fig. 1.

The linear structural equation model offers two advantages in handling measurement errors. First, it provides a more transparent picture of 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. 1(a).

Since U is sufficient in Fig. 1(a), the total effect is identifiable from the measurement on X, Y and U, and is given by $\tau_{yx} = \beta_{yx\cdot u}$. However, if U is unobserved and W is but a noisy measurement of U, as in Fig. 1(a), knowledge of the error mechanism $W = \alpha_{wu}U + \epsilon_w$ from an external study is needed in order to identify $\tau_{yx} = \beta_{yx\cdot u}$. However, 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

$$\tau_{yx} = \beta_{yx\cdot u} = \left(\sigma_{xy} - \frac{\alpha_{wu}^2\sigma_{xu}\sigma_{yu}}{\alpha_{wu}^2\sigma_{uu}}\right) \Big/ \left(\sigma_{xx} - \frac{\alpha_{wu}^2\sigma_{xu}^2}{\alpha_{wu}^2\sigma_{uu}}\right)$$

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

$$\tau_{yx} = \left(\sigma_{xy} - \frac{\sigma_{xw}\sigma_{yw}}{\alpha_{wu}^2\sigma_{uu}}\right) \Big/ \left(\sigma_{xx} - \frac{\sigma_{xw}^2}{\alpha_{wu}^2\sigma_{uu}}\right). \tag{5}$$

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. 1(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 $\sigma_{xw}\sigma_{wz}/\sigma_{xz} = \alpha_{wu}^2\sigma_{uu}$. By substituting this equation into equation (5), we can see that τ_{yx} is identifiable and is given by

$$\tau_{yx} = \beta_{yx\cdot u} = \frac{\sigma_{xy}\sigma_{wz} - \sigma_{xz}\sigma_{yw}}{\sigma_{xx}\sigma_{wz} - \sigma_{xw}\sigma_{xz}} = \frac{\sigma_{xy}\sigma_{wz} - \sigma_{xw}\sigma_{yz}}{\sigma_{xx}\sigma_{wz} - \sigma_{xw}\sigma_{xz}}, \tag{6}$$

using $\sigma_{xz}\sigma_{yw} = \sigma_{xw}\sigma_{yz}$ from the fact that $\{X, Y\}$, Z and W are independent of each other given U. This result reflects the well-known fact (e.g., Bollen, 1989, p. 224) that, in linear structural equation models, structural parameters are identifiable, up to a constant σ_{uu}, whenever each latent variable has three independent proxies. We see that the nonidentifiability of σ_{uu} is not an impediment for the identification of τ_{yx}.

We next relax the requirement that U possesses three independent proxies as in Fig. 1(b) and consider a situation as in Fig. 1(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

$$\alpha_{wu}^2\sigma_{uu} = \frac{\sigma_{yw\cdot x}\sigma_{wz\cdot x}}{\sigma_{yz\cdot x}} + \frac{\sigma_{xw}^2}{\sigma_{xx}}. \tag{7}$$

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 $\sigma_{yw\cdot x}\sigma_{wz\cdot x}/\sigma_{yz\cdot x} = \beta_{wu\cdot x}^2\sigma_{uu\cdot x} = \beta_{wu}^2\sigma_{uu\cdot x}$ from $X \perp\!\!\!\perp W \mid U$. In addition, noting that $\beta_{wu} = \alpha_{wu}$ and $\sigma_{xw} = \beta_{wu}\sigma_{xu} = \alpha_{wu}\sigma_{xu}$, we have $\alpha_{wu}^2\sigma_{uu\cdot x} = \alpha_{wu}^2\sigma_{uu} - \sigma_{xw}^2/\sigma_{xx}$. Using these results, equation (7) is obtained. The first term of equation (7) 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 τ_{yx}, we substitute equation (7) into equation (5), and using $\sigma_{yw\cdot x} = \sigma_{yw} - \sigma_{xy}\sigma_{xw}/\sigma_{xx}$ we have

$$\tau_{yx} = \frac{\sigma_{xy}\sigma_{yw\cdot x}\sigma_{wz\cdot x}\sigma_{xx} + \sigma_{yz\cdot x}\sigma_{xw}(\sigma_{xy}\sigma_{xw} - \sigma_{yw}\sigma_{xx})}{\sigma_{yw\cdot x}\sigma_{wz\cdot x}\sigma_{xx}^2} = \frac{\sigma_{xy}\sigma_{wz\cdot x} - \sigma_{yz\cdot x}\sigma_{xw}}{\sigma_{wz\cdot x}\sigma_{xx}}. \tag{8}$$

We see that $\tau_{yx} = \beta_{yx\cdot u}$ is identifiable and is given by equation (6) or (8).

From Fig. 1(a), (b) and (c), we see that the pivotal quantity needed for the identification of τ_{yx} is the product

$$\alpha_{wu}^2\sigma_{uu} = \sigma_{ww} - \sigma_{\epsilon_w\epsilon_w}, \tag{9}$$

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 (6) or (8), yielding equation (5). If however U has only one proxy W, as in Fig. 1(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 (9), since both α_{wu} and ϵ_w are causal parameters of the error mechanism $W = \alpha_{wu}U + \epsilon_w$, $\alpha_{wu} = dE(W \mid u)/du$ 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.

In the noiseless case, i.e., $\sigma_{\epsilon_w\epsilon_w} = 0$, we have $\sigma_{uu} = \sigma_{ww}/\alpha_{wu}^2$ and equation (5) reduces to

$$\tau_{yx} = \frac{\sigma_{yx\cdot w}}{\sigma_{xx\cdot w}} = \beta_{yx\cdot w},$$

where $\beta_{yx\cdot w}$ is the regression coefficient of x in the regression model of Y on X and W, or $\beta_{yx\cdot w} = \partial E(Y \mid x, w)/\partial x$. As expected, the equality $\tau_{yx} = \beta_{yx\cdot u} = \beta_{yx\cdot w}$ assures a bias-free estimate of τ_{yx} through adjustment for W, instead of U; α_{wu} plays no role in this adjustment.

Figure 1(d) represents a new challenge; although $\alpha_{wu}^2\sigma_{uu}$ is not identifiable, the total effect τ_{yx} is nevertheless identifiable without external studies. In the next section, we will discuss this identification strategy.

4·3. *Instrumental variable method with a proxy variable*

In Fig. 1(d), if U can be observed, then both the condition of the conditional instrumental variable 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 $\{X, Z\} \perp\!\!\!\perp W \mid U$, 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

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

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

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

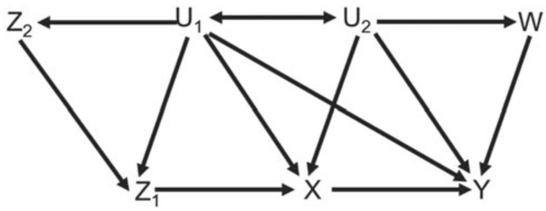

Fig. 2. Causal diagram with unmeasured confounders.

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

$$\tau_{yx} = \left(\sigma_{yz} - \sigma_{xy}\frac{\sigma_{wz}}{\sigma_{xw}}\right) \Big/ \left(\sigma_{xz} - \sigma_{xx}\frac{\sigma_{wz}}{\sigma_{xw}}\right),$$

which is consistent with equation (6). This derivation demonstrates a more general approach that differs from that of Cai & Kuroki (2008), which was based on latent factor analysis (e.g., Bollen, 1989; Stanghellini, 2004; Stanghellini & Wermuth, 2005). Our approach extends the identification conditions to cases where the total effect cannot be identified by any single strategy but by a combination of several strategies, in our example, the back door criterion combined with the condition for conditional instrumental variables. In addition, unlike the discussion in § 4·2, the identification of $\alpha^2_{wu}\sigma_{uu}$ 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. 2 where the sufficient set $\{U_1\} \cup U_2$ is unobserved. Here, U_1 is univariate but the number of variables in U_2 can be arbitrary. In this situation, the back door criterion cannot be used to identify the total effect of X on Y, and the uncertain number of variables in 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 independent of $\{U_1\} \cup U_2$, they cannot be used as the conditional instrumental variables. Nevertheless, we will show that the total effect is identifiable, as follows: since both Z_1 and Z_2 are conditional instrumental variables given U_1 relative to (X, Y), the total effect is given by $\tau_{yx} = \sigma_{yz_1\cdot u_1}/\sigma_{xz_1\cdot u_1} = \sigma_{yz_2\cdot u_1}/\sigma_{xz_2\cdot u_1}$. Moreover, since $\{Z_1, Z_2\} \perp\!\!\!\perp W \mid U_1$ 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

$$\sigma_{yz_1} - \frac{\sigma_{yu_1}\sigma_{z_2u_1}}{\sigma_{u_1u_1}}\frac{\sigma_{z_1w}}{\sigma_{z_2w}} = \tau_{yx}\left(\sigma_{xz_1} - \frac{\sigma_{xu_1}\sigma_{z_2u_1}}{\sigma_{u_1u_1}}\frac{\sigma_{z_1w}}{\sigma_{z_2w}}\right) \tag{10}$$

and

$$\sigma_{yz_2} - \frac{\sigma_{yu_1}\sigma_{z_2u_1}}{\sigma_{u_1u_1}} = \tau_{yx}\left(\sigma_{xz_2} - \frac{\sigma_{xu_1}\sigma_{z_2u_1}}{\sigma_{u_1u_1}}\right).$$

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

$$\tau_{yx} = \frac{\sigma_{z_1y}\sigma_{z_2w} - \sigma_{z_2y}\sigma_{z_1w}}{\sigma_{z_1x}\sigma_{z_2w} - \sigma_{z_2x}\sigma_{z_1w}}. \tag{11}$$

Even if Z_2 is not a conditional instrumental variable given U_1 relative to (X, Y), when $Z_2 \perp\!\!\!\perp \{X, Y\} \mid U_1$ holds, the total effect τ_{yx} of X on Y is still identifiable and is given by equation (11) by substituting $\sigma_{xz_2} = \sigma_{xu_1}\sigma_{z_2u_1}/\sigma_{u_1u_1}$ and $\sigma_{yz_2} = \sigma_{yu_1}\sigma_{z_2u_1}/\sigma_{u_1u_1}$ into equation (10).

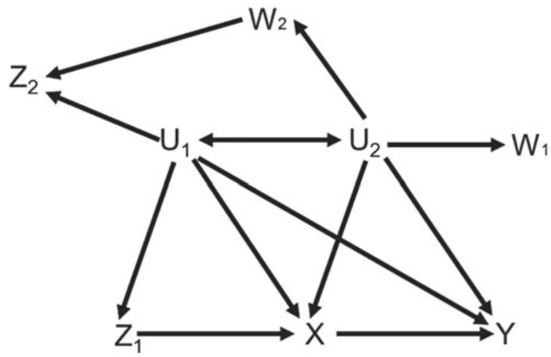

Fig. 3. Causal diagram of a study on the Head Start Program.

We now summarize these considerations in a theorem.

Theorem 2. *Suppose that:*

(i) *a nonempty set* $\{Z_1, Z_2\}$ *of distinct variables satisfies one of the following conditions:* (i-a) *both* Z_1 *and* Z_2 *are conditional instrumental variables given a univariate* U *relative to* (X, Y)*,* (i-b) Z_1 *is a conditional instrumental variable given* U *relative to* (X, Y)*, and* $Z_2 = X$ *and* U *satisfies the back door criterion relative to* (X, Y)*,* (i-c) Z_1 *is a conditional instrumental variable given* U *relative to* (X, Y)*, and* U *d-separates* Z_2 *from* $\{X, Y\}$*;*
(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* (*11*).

Drton et al. (2011) proved that the linear structural equation model of interest is globally identifiable if and only if the graph lacks a convergent arborescence or a C-tree, in the nomenclature of Shpitser & Pearl (2006), where 'globally identifiable' means that all path coefficients can be estimated uniquely for any values taken by covariance parameters. Our results are applicable to situations where C-trees exist. For example, although a confounding path ($X \leftarrow U_2 \rightarrow Y$) constitutes a C-tree in Fig. 2, the total effect τ_{yx} is identifiable according to Theorem 2. Here, it should be noted that the total effect cannot be estimated by our result when the denominator of equation (11) is zero.

4·4. *Example*

In this section, as an example of Theorem 2, we consider again the Head Start Program described in § 1. First, we constructed a causal diagram shown in Fig. 3 based on Kenny (1979). Let U_1 and U_2 represent a socio-economic factor and an educational background factor, respectively. In addition, let Z_1, Z_2, W_1 and W_2 be family income, father's occupation, mother's education and father's education, respectively. Z_1 and Z_2 are observed proxy variables of U_1, and W_1 and W_2 are observed proxy variables of U_2. Figure 3 differs from the model in Kenny (1979), which did not consider the direct effects from W_2 to Z_2 and from Z_1 to X. Although Kenny (1979) stated that 'no doubt father's education causes father's occupation,' he did not consider $W_2 \rightarrow Z_2$, because he focused on the application of classical latent structure analysis. In addition, as stated in § 1, since family income is an important factor that determines whether or not a child attended the Head Start Program, we added a direct effect from Z_1 to X. Based on the correlation matrix in Kenny (1979), the model chi-square test for Fig. 3 yields $\chi^2 = 3{\cdot}874$ with

Table 1. *The estimated correlation matrix based on Fig. 3*

	W_1	W_2	Z_1	Z_2	X	Y
W_1	1·000	0·468	0·278	0·270	−0·117	0·276
W_2	0·468	1·000	0·286	0·209	−0·091	0·214
Z_1	0·278	0·286	1·000	0·407	−0·219	0·209
Z_2	0·270	0·209	0·407	1·000	−0·180	0·222
X	−0·117	−0·091	−0·219	−0·180	1·000	−0·094
Y	0·276	0·214	0·209	0·222	−0·094	1·000

3 degrees of freedom, which is not statistically significant; the p-value is 0·275. Thus, in this section, we assume that the graph shown in Fig. 3 represents a causal diagram of this case study. Based on the correlation matrix in Kenny (1979), we estimated the correlation matrix based on Fig. 3, which is shown in Table 1. Here, in order to avoid discussion of sampling variability, we assume that Table 1 is the correlation matrix from Fig. 3. Since Z_1 is a conditional instrumental variable given $\{U_1, W_2\}$, and $\{U_1, W_2\}$ d-separates Z_2 from $\{X, Y\}$ and $\{Z_1, Z_2\}$ from W_1 in Fig. 3, when W_2 is given, both conditions (i-c) and (ii) of Theorem 2 hold. Then, the total effect τ_{yx} of X on Y is evaluated by a positive value 0·183 from equation (11). This shows that the Head Start Program had a positive effect on child's cognitive skills, which is consistent with the results in Magidson (1977) and Kenny (1979).

On the other hand, since Magidson (1977) stated that a 'various multifactor model could be formulated' for this case study, one may consider a causal diagram by adding a direct effect from father's occupation Z_2 to family income Z_1 and a direct effect from mother's education W_1 to the Metropolitan Readiness Test Y in Fig. 3. Letting W_1 in Fig. 3 be W in Fig. 2, Fig. 2 is a subgraph of this causal diagram. Based on the correlation matrix in Kenny (1979), the model chi-square test for Fig. 2 yields $\chi^2 = 3{\cdot}851$ with one degree of freedom, which is statistically significant; the p-value is 0·050. However, if we assume that this causal diagram reflects the true data generating process, since both conditions (i-a) and (ii) in Theorem 2 hold given W, we obtain $\tau_{yx} = 0{\cdot}001$ by the similar procedure above. This shows that the Head Start Program had only a minor effect on cognitive skills.

5. Conclusion

This paper focuses on the identification problem and stops short of dealing with the problem of estimating the derived estimands from finite samples. Since the effect restoration method presented in § 3 can be considered an extension of the identification of latent structure models, known methods of estimation and variance analysis used in latent structure analysis (e.g., Bartholomew et al., 2011; Hu, 2008) are applicable to the estimands derived in this paper under most situations.

In the nonparametric case, we assumed that the proxy variables and the unmeasured confounder are discrete variables with a given finite number of categories, k in § 2 and § 3. This assumption can be relaxed by allowing the proxy variables to have more than k categories, or even continuous support. In such cases, as Pearl (2010) pointed out, it may be difficult to obtain reliable statistics of the recovered probabilities $\mathrm{pr}(x, y, u)$ due to data sparseness, and the use of propensity score methods (Rosenbaum & Rubin, 1983) may be crucial.

Acknowledgement

We thank the editor, associate editor and two referees whose comments significantly improved the presentation of the paper. This research was funded by the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Asahi Glass Foundation, Office of Naval Research, National Institutes of Health, and National Science Foundation.

APPENDIX

Proof of Theorem 1

The proof of Theorem 1 is based on the following two-step procedure that 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 $\mathrm{pr}(w \mid u)$ from $U(w, u)$.
Step 2. Recover $\mathrm{pr}(x, y, u)$ using the matrix adjustment method introduced in § 2.

Step 1. To find $\mathrm{pr}(w \mid u)$ 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, we solve $\mid P(z, w)^{-1}Q(z, w) - \lambda I_k \mid = 0$ for λ to obtain the set of eigenvalues of $P(z, w)^{-1}Q(z, w)$. In other words, λ should satisfy $|P(z, w)^{-1}Q(z, w) - \lambda I_k \mid = \mid \Delta(u) - \lambda I_k \mid = 0$. From this equation, letting $\lambda_1 > \cdots > \lambda_k$ for eigenvalues of $P(z, w)^{-1}Q(z, w)$, we have $\lambda_i = \mathrm{pr}(y \mid x, u_{(i)})$ $(i = 1, \ldots, k)$, thus the elements of $\Delta(u)$ are estimable. In order to obtain the eigenvector η_i for λ_i, letting $H = (\eta_1, \ldots, \eta_k)$, we solve the following simultaneous linear equations $P(z, w)^{-1}Q(z, w)\eta_i = \lambda_i \eta_i$ $(i = 1, \ldots, k)$ or, equivalently, $P(z, w)^{-1}Q(z, w)H = H\Delta(u)$. Here, $\eta_1, \ldots, \eta_k$ are uniquely determined except for a multiplicative constant because $\lambda_1, \ldots, \lambda_k$ take different values from condition (4) in § 3. On the other hand, letting $A = U(w, u)^{-1}E$ and $E = \mathrm{diag}(\alpha_1, \ldots, \alpha_k)$ for any nonzero values of $\alpha_1, \ldots, \alpha_k$, we have $\{P(z, w)^{-1}Q(z, w)\}A = U(w, u)^{-1}\Delta(u)E = U(w, u)^{-1}E\Delta(u) = A\Delta(u)$. 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 $\alpha_1, \ldots, \alpha_k$. Then, for the inverse $H^{-1} = (h_{ij})$ of the estimable matrix H, we have using $U(w, u)^{-1}E = H$,

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

Equating the first column of both sides of the equation, the diagonal element $\alpha_1 = 1/h_{11}, \ldots, \alpha_k = 1/h_{k1}$ of E can be obtained, which indicates that $U(w, u)$ is identifiable from EH^{-1}, since H^{-1} is estimable. Thus, every element $\mathrm{pr}(w \mid u)$ 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 § 2. Since we have

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

substitute elements of $\mathrm{pr}(w_i \mid u_{(j)})$ $(i, j = 1, \ldots, k)$ obtained in Step 1 for $M(w, u)$ in equation (4). Then, since $M(w, u)$ is invertible, we can obtain elements of $V_{xy}(u)$. Thus, the causal effect

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

is identifiable.

REFERENCES

BARTHOLOMEW, D. J., KNOTT, M. & MOUSTAKI, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed. Chichester: John Wiley & Sons.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
Bowden, R. J. & Turkington, D. A. (1984). *Instrumental Variables*. Cambridge, MA: Cambridge University Press.
Brito, C. & Pearl, J. (2002). Generalized instrumental variables. In *Proc. 18th Conf. Uncertainty Artificial Intel.*, pp. 85–93, Alberta, Canada.
Cai, Z. and Kuroki, M. (2008). On identifying total effects in the presence of latent variables and selection bias. *Proc. 24th Annu. Conf. Uncertainty Artificial Intel.*, pp. 62–9, Helsinki, Finland.
Cai, Z, Kuroki, M., Pearl, J. & Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* **64**, 695–701.
Carroll, R., Ruppert, D., Stefanski, L. & Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. Boca Raton: Chapman & Hall/CRC.
Drton, M., Foygel, R. & Sullivant, S. (2011), Global identifiability of linear structural equation models. *Ann. Statist.* **39**, 865–86.
Goetghebeur, E. & Vansteelandt, S. (2005). Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statist. Meth. Med. Res.* **14**, 397–415.
Greenland, S. (2005). Multiple-bias modeling for analysis of observational data. *J. R. Statist. Soc.* A **168**, 267–306.
Greenland, S. & Kleinbaum, D. (1983). Correcting for misclassification in two-way tables and matched-pair studies. *Int. J. Epidemiol.* **12**, 93–7.
Greenland, S. & Lash, T. (2008). Bias analysis. In *Modern Epidemiology*, 3rd ed., Ed. K. Rothman, S. Greenland and T. Lash, pp. 345–80. Philadelphia: Lippincott Williams and Wilkins.
Hernán, M. & Cole, S. (2009). Invited commentary: Causal diagrams and measurement bias. *Am. J. Epidemiol.* **170**, 959–62.
Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *J. Economet.* **144**, 27–61.
Kenny, D. A. (1979). *Correlation and Causality*. New York: John Wiley & Sons.
Kuroki, M. (2007). Graphical identifiability criteria for average causal effects in studies with an unobserved treatment/response variable. *Biometrika* **94**, 37–47.
Kuroki, M. & Cai, Z. (2008). Formulating tightest bounds on causal effects in studies with unmeasured confounders. *Statist. Med.* **27**, 6597–611.
MacLehose, R. F., Kaufman, S., Kaufman, J. S. & Poole, C. (2005). Bounding causal effects under uncontrolled confounding using counterfactuals. *Epidemiology* **16**, 548–55.
Magidson, J. (1977). Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation: A general alternative to ANCOVA. *Eval. Rev.* **1**, 399–420.
Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann.
Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, MA: Cambridge University Press.
Pearl, J. (2010). On measurement bias in causal inference. In *Proc. 26th Conf. Uncertainty Artificial Intel.*, pp. 425–32, Catalina Island. Available at: http://ftp.cs.ucla.edu/pub/stat_ser/r357.pdf.
Pearl, J. & Bareinboim, E. (2011). Transportability across studies: A formal approach. In *Proc. 25th AAAI Conf. Artificial Intel.* pp. 247–54, San Francisco.
Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
Selén, J. (1986). Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *J. Am. Statist. Assoc.* **81**, 75–81.
Shpitser, I. & Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proc. 21st Nat. Conf. Artificial Intel.* pp. 1219–26, Boston.
Stanghellini, E. (2004). Instrumental variables in Gaussian directed acyclic graph models with an unobserved confounder. *Environmetrics* **15**, 463–69.
Stanghellini, E. & Wermuth, N. (2005). On the identification of directed acyclic graph models with one hidden variable. *Biometrika* **92**, 337–50.
Tian, J. & Pearl, J. (2002). A general identification condition for causal effects. In *Proc. 18th Nat. Conf. Artificial Intel.*, pp. 567–73, Edmonton, Canada.

[*Received June* 2012. *Revised October* 2013]