Biometrika (2006), **93,** 3, *pp.* 601–611

Printed in Great Britain

On recovering a population covariance matrix in the presence of selection bias

By MANABU KUROKI

Division of Mathematical Science, Department of Systems Innovation, Graduate School of Engineering Science, Osaka University, 1-3, Machikaneyama-cho, Toyonaka, Osaka, 560-8531, Japan

mkuroki@sigmath.es.osaka-u.ac.jp

and ZHIHONG CAI

Department of Biostatistics, Graduate School of Public Health, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto, 606-8501, Japan

cai@pbh.med.kyoto-u.ac.jp

Summary

This paper considers the problem of using observational data in the presence of selection bias to identify causal effects in the framework of linear structural equation models. We propose a criterion for testing whether or not observed statistical dependencies among variables are generated by conditioning on a common response variable. When the answer is affirmative, we further provide formulations for recovering the covariance matrix of the whole population from that of the selected population. The results of this paper provide guidance for reliable causal inference, based on the recovered covariance matrix obtained from the statistical information with selection bias.

Some key words: Directed acyclic graph; Path diagram; Single factor model; Tetrad difference.

1. Introduction

The evaluation of causal effects from observational studies plays a significant role in many scientific domains. Although experimental studies are the most effective and reliable ways of evaluating causal effects, it is often too expensive, infeasible or unethical to employ such a study. Under such circumstances, evaluation of causal effects from observational data becomes an important problem. Compared with experimental studies, observational studies are more susceptible to bias because of the lack of randomisation. Confounding bias, selection bias and measurement error can all hinder the evaluation of causal effects from observational data.

In this paper, we consider identification of causal effects from observational data with selection bias, which typically leads to biased inference of causal effects. For example, if two variables C and D are independent in a given population, and a sample is chosen according to some value of S that is affected by both C and D, then C and D will be statistically dependent in this sample. Selection bias exists in many observational studies, a classical example being Berkson's bias in hospital-based studies (Berkson, 1946). In

recent years, the problem of selection bias has gained attention from epidemiologists (Greenland, 2003; Hernan et al., 2004).

Researchers in artificial intelligence have provided several methods for inferring causal relationships from observational data when selection bias may be at work (Spirtes et al., 1999; Cooper, 1995, 2000). However, their causal discovery algorithms are sometimes slightly weak in detecting causal relationships, since they use directly the statistical information based on selection bias. On the other hand, Heckman and other social scientists have done much work on selectivity problems in observational studies (Winship & Mare, 1992). In particular, Heckman's two-stage estimator is the most widely used approach to selection bias, using regression methods to estimate behavioural functions by least squares methods (Heckman, 1979). The present paper focuses on recovering the covariance matrix of the whole population from observational data with selection bias, in order to evaluate causal effects on the basis of the recovered covariance matrix in the framework of linear structural equation models. We assume that we know one possible causal structure that could have generated the data with selection bias. Observational equivalence in graphical models may exist (Dawid, 2002), but we do not consider this problem here. We use a continuous so-called selection variable S to represent selection. In general, it is more realistic to consider selection as occurring on a subset of the range of S, rather than on a specific value of S. Hence, we assume that a sample is selected from the population if $a \leqslant S \leqslant b$, where both a and b are possible values of S. First, we propose a criterion for testing whether statistical dependencies among variables are generated by conditioning on a common response variable. Next, when observed statistical dependencies among variables are judged to be generated by selection bias through our criterion, we propose a way of recovering the covariance matrix of the whole population, and thence identifying causal effects of the whole population.

2. PRELIMINARIES

2·1. *Graphs*

A directed graph is a pair $G = (V, E)$, where V is a finite set of vertices and the set E of arrows is a subset of ordered pairs of distinct vertices. An arrow pointing from a vertex a to a vertex b indicates that $(a, b) \in E$ and $(b, a) \notin E$. The arrow is said to emerge from a or to point to b. If there is an arrow pointing from a to b, a is said to be a parent of b, and b a child of a. The set of parents of b is denoted by $\mathrm{pa}(b)$, and the set of children of a by $\mathrm{ch}(a)$.

A path between a and b is a sequence $a = a_0, \ldots, b = a_n$ of distinct vertices such that $(a_{i-1}, a_i) \in E$ or $(a_i, a_{i-1}) \in E$, for all $i = 1, \ldots, n$. A directed path from a to b is a sequence $a = a_0, \ldots, b = a_n$ of distinct vertices such that $(a_{i-1}, a_i) \in E$ and $(a_i, a_{i-1}) \notin E$, for all $i = 1, \ldots, n$. If there exists a directed path from a to b, a is said to be an ancestor of b and b a descendant of a. If two arrows on a path point to a, then a is said to be a collider; otherwise, it is said to be a non-collider.

A directed cycle is a sequence $a_0, \ldots, a_n$ of distinct vertices such that $(a_{i-1}, a_i) \in E$ and $(a_i, a_{i-1}) \notin E$ for all $i = 1, \ldots, n$, and $(a_n, a_0) \in E$. If a directed graph has no directed cycle, then the graph is said to be a directed acyclic graph. A path is said to be blocked by a set Z, possibly empty, if the path contains at least one non-collider that is in Z, and/or the path contains at least one collider that is not in Z and has no descendant in Z. A set Z is said to d-separate a from b $(a, b \notin Z)$ in a directed acyclic graph G if Z blocks every path between a and b.

2·2. *Path diagram*

Suppose that a set $V = \{V_1, \ldots, V_n\}$ of variables and a directed acyclic graph $G = (V, E)$ are given. When each child-parent family in the graph G represents a linear structural equation model,

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

the graph G is called a path diagram, where $\varepsilon_{v_1}, \ldots, \varepsilon_{v_n}$ are assumed to be independent and normally distributed with mean zero. In addition, $\alpha_{v_j v_i}\,(\neq 0)$ is called a path coefficient. It is noted that neither two-headed-arrow relationships nor feedback relationships are featured in this paper since we are dealing with models representable by directed acyclic graphs. For further details of linear structural equation models, see Bollen (1989).

Here, we define some notation. For disjoint subsets W, Z and T of variables in V, let $\sigma_{v_i v_j \cdot z}$ and $\Sigma_{wz \cdot t}$ be the conditional covariance between V_i and V_j given Z and the conditional covariance matrix between W and Z given T, respectively. In addition, $\beta_{v_j v_i \cdot z}$ is the regression coefficient of v_i in the regression model of V_j on v_i and z. Furthermore, $B_{zw \cdot t}$ is the regression coefficient matrix of W in the regression model of each element of Z on W and T. Similar notation is used for other parameters.

For a set Z of variables not including descendants of V_j, if Z d-separates V_i from V_j in the graph obtained by deleting from a graph G an arrow pointing from V_i to V_j, then $\beta_{v_j v_i \cdot z} = \alpha_{v_j v_i}$ holds true. This criterion is called 'the single door criterion' (Pearl, 2000, p. 150). In addition, when W d-separates V_i from V_j in the graph G, V_i is conditionally independent of V_j given W in the corresponding distribution, denoted by $V_i \perp\!\!\!\perp V_j | W$ (Pearl, 2000, p. 18).

2·3. *Tetrad difference*

For any distinct variables $X_i, X_j, X_k, X_l \in X \subset V$, Spearman (1904, 1928) defined the tetrad difference as

$$\sigma_{x_i x_j} \sigma_{x_k x_l} - \sigma_{x_i x_k} \sigma_{x_j x_l}.$$

When the tetad difference is equal to zero, it is called a vanishing tetrad difference, and this implies that the observed statistical dependencies can be well explained by a single-factor model without correlated errors.

The tetrad difference has been studied by many researchers for decades. Bekker & de Leeuw (1987) summarised the previous results in a single comprehensive theorem. To present this theorem, the following preliminaries are required. When there exists a positive semidefinite diagonal matrix Ω such that $\Sigma_{xx} - \Omega$ is a positive semidefinite matrix and $\mathrm{rank}(\Sigma_{xx} - \Omega) = 1$, Σ_{xx} is said to be a Spearman matrix, which provides statistical evidence that the observed statistical dependencies can be well explained by a single-factor model (Bekker & Leeuw, 1987). In addition, any element of Σ_{xx} is assumed to be nonzero.

THEOREM 1 (*Bekker & de Leeuw*, 1987). *A covariance matrix $\Sigma_{xx} = (\sigma_{x_i x_j})$ of a set X that includes four or more variables is a Spearman matrix if and only if, after sign changes of rows and corresponding columns, all its elements are positive and such that*

$$\sigma_{x_i x_j} \sigma_{x_k x_l} - \sigma_{x_i x_k} \sigma_{x_j x_l} = 0, \quad \frac{\sigma_{x_i x_j} \sigma_{x_i x_k}}{\sigma_{x_j x_k}} \leqslant \sigma_{x_i x_i},$$

for any distinct variables $X_i, X_j, X_k, X_l \in X$.

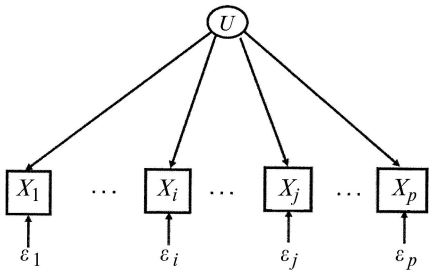

Fig. 1. Graphical representation of a single factor model without correlated errors.

We consider the above theorem in terms of the path diagram shown in Fig. 1. For a set $X = (X_1, \ldots, X_p)$ of observed variables, denoted by boxes in Fig. 1, and an unobserved variable U, denoted by a circle, the covariance structure corresponding to Fig. 1 can be described as

$$\Sigma_{xx} = \Sigma_{xx \cdot u} + \frac{1}{\sigma_{uu}} \Sigma_{xu} \Sigma'_{xu}. \tag{1}$$

The $\Sigma_{xx \cdot u}$ in equation (1) corresponds to the Ω described above which satisfies $\mathrm{rank}(\Sigma_{xx} - \Omega) = 1$. Then, $\Sigma_{xx} - \Omega$ is a positive semidefinite matrix. On the other hand, if we change the signs of the rows and corresponding columns of equation (1), all the elements of Σ_{xx} in equation (1) can become positive. In addition,

$$\sigma_{x_i x_j} \sigma_{x_k x_l} - \sigma_{x_i x_k} \sigma_{x_j x_l} = 0, \quad \sigma_{x_i x_i} - \frac{\sigma_{x_i x_j} \sigma_{x_i x_k}}{\sigma_{x_j x_k}} = \sigma_{x_i x_i \cdot u} \geqslant 0$$

hold true. For further discussion about the tetrad difference, see Bollen & Ting (1993) and Spirtes et al. (1993).

3. Recovering a covariance matrix

3·1. *Preliminaries*

Lemma 1 (*Johnson & Kotz*, 1972, p. 70). *When* $\{S\} \cup Y$ *are normally distributed*,

$$\Sigma^*_{yy} = \mathrm{var}(Y | a \leqslant S \leqslant b) = \Sigma_{yy \cdot s} + B_{ys} B'_{ys} \sigma^*_{ss} = \Sigma_{yy} - B_{ys} B'_{ys} \check{\sigma}_{ss}, \tag{2}$$

where $\sigma^*_{ss} = \mathrm{var}(S | a \leqslant S \leqslant b)$ *and* $\check{\sigma}_{ss} = \sigma_{ss} - \sigma^*_{ss}$.

Note that $\check{\sigma}_{ss} \geqslant 0$ since σ^*_{ss} is the variance of a doubly-truncated normal distribution. In particular, if $b = \infty$ in equation (2), equation (2) corresponds to the covariance matrix of the extended skew-normal distribution (Capitanio et al., 2003).

By the Sherman–Morrison–Woodbury formula for matrix inversion (Rao, 1973, p. 33), we can obtain

$$\Sigma^{*-1}_{yy} = \Sigma^{-1}_{yy \cdot s} - \frac{\Sigma^{-1}_{yy \cdot s} B_{ys} B'_{ys} \Sigma^{-1}_{yy \cdot s}}{1 + \sigma^*_{ss} B'_{ys} \Sigma^{-1}_{yy \cdot s} B_{ys}} \sigma^*_{ss}.$$

If we write

$$\begin{pmatrix} \Sigma_{yy} & \Sigma_{ys} \\ \Sigma'_{ys} & \sigma_{ss} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma^{yy} & \Sigma^{ys} \\ \Sigma^{ys\prime} & \sigma^{ss} \end{pmatrix},$$

and note that $\Sigma^{yy} = \Sigma_{yy\cdot s}^{-1}$ and $B_{ys} = -(\Sigma^{yy})^{-1}\Sigma^{ys}$, we can also obtain

$$\Sigma_{yy}^{*-1} = \Sigma_{yy\cdot s}^{-1} - \frac{\Sigma^{ys}\Sigma^{ys\prime}}{1 + \sigma_{ss}^{*}\Sigma^{ys\prime}\Sigma_{yy\cdot s}\Sigma^{ys}}\sigma_{ss}^{*}. \tag{3}$$

By partitioning Y into $X \cup Z$, we can also obtain

$$\Sigma_{xx\cdot z}^{*} = \mathrm{var}(X|a \leqslant S \leqslant b, Z) = (\sigma_{x_i x_j\cdot z}^{*}) = \Sigma_{xx\cdot z} - B_{xs\cdot z}B_{xs\cdot z}'\check{\sigma}_{ss\cdot z}, \tag{4}$$

where

$$\sigma_{x_i x_j\cdot z}^{*} = \mathrm{cov}(X_i, X_j|a \leqslant S \leqslant b, Z), \quad \sigma_{ss\cdot z}^{*} = \mathrm{var}(S|a \leqslant S \leqslant b, Z)$$

and $\check{\sigma}_{ss\cdot z} = \sigma_{ss\cdot z} - \sigma_{ss\cdot z}^{*} \geqslant 0$. Similar notation is used for other parameters.

3·2. *Tetrad difference under selection*

In this section, we propose a criterion for testing whether or not statistical dependencies among variables are generated by a common response variable.

The path diagram in Fig. 2 corresponds to the case in which any two elements of X are independent given Z in the population, but are dependent after conditioning on S, which indicates that sample selection is conducted according to a criterion $a \leqslant S \leqslant b$. For some examples of such a situation, see Greenland (2003) and Hernan et al. (2004). We give a detailed example in § 4. In addition, this situation is closely related to the problem of causal indicators which are variables affected by observed variables (Bollen & Lennox, 1991; Bollen & Ting, 2000; Edwards & Bagozzi, 2000).

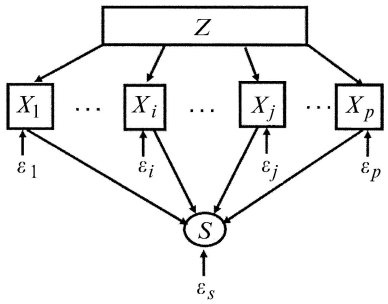

Fig. 2. Graphical representation in which selection bias occurs: Z d-separates any two elements of X but neither S nor $\{S\} \cup Z$ does.

For a set $X \cup Z$ of observed variables and a selection variable S, the covariance structure corresponding to Fig. 2 can be described by equation (4). Here, $\Sigma_{xx\cdot z}$ is a positive definite diagonal matrix such that $\Sigma_{xx\cdot z}^{*} - \Sigma_{xx\cdot z}$ is a negative semidefinite matrix and $\mathrm{rank}(\Sigma_{xx\cdot z}^{*} - \Sigma_{xx\cdot z}) = 1$. Thus, equation (4) is similar to equation (1).

On the other hand, if we change the signs of some rows and the corresponding columns, all the off-diagonal elements of $\Sigma_{xx\cdot z}^{*}$ in equation (4) can become negative. In addition, we can obtain

$$\sigma_{x_i x_j\cdot z}^{*}\sigma_{x_k x_l\cdot z}^{*} - \sigma_{x_i x_k\cdot z}^{*}\sigma_{x_j x_l\cdot z}^{*} = 0, \quad \sigma_{x_i x_i\cdot z}^{*} - \frac{\sigma_{x_i x_j\cdot z}^{*}\sigma_{x_i x_k\cdot z}^{*}}{\sigma_{x_j x_k\cdot z}^{*}} = \sigma_{x_i x_i\cdot z} \geqslant 0.$$

Hence, based on the same considerations as in Theorem 1, we can provide a statistical justification that the observed statistical dependencies among X given Z are generated by

a selection variable; that is, if there exists a positive definite diagonal matrix Ω such that $\Sigma^*_{xx\cdot z} - \Omega$ is a negative semidefinite matrix and $\mathrm{rank}(\Sigma^*_{xx\cdot z} - \Omega) = 1$, by considering the Ω as the conditional covariance matrix of X given Z, we can provide statistical justification that the observed statistical dependencies can be well explained by selection bias. Therefore, when any element of $\Sigma^*_{xx\cdot z}$ is assumed to be nonzero, the following theorem can be obtained.

THEOREM 2. *For a covariance matrix $\Sigma^*_{xx\cdot z}$ of a set X that includes four or more variables, there exists a positive definite diagonal matrix Ω such that $\Sigma^*_{xx\cdot z} - \Omega$ is a negative semidefinite matrix and* $\mathrm{rank}(\Sigma^*_{xx\cdot z} - \Omega) = 1$ *if and only if, after sign changes of rows and corresponding columns of $\Sigma^*_{xx\cdot z}$, all its off-diagonal elements are negative and such that*

$$\sigma^*_{x_i x_j\cdot z}\sigma^*_{x_k x_l\cdot z} = \sigma^*_{x_i x_k\cdot z}\sigma^*_{x_j x_l\cdot z}, \tag{5}$$

$$\frac{\sigma^*_{x_i x_k\cdot z}\sigma^*_{x_i x_j\cdot z}}{\sigma^*_{x_j x_k\cdot z}} \leqslant \sigma^*_{x_i x_i\cdot z}, \tag{6}$$

for any distinct variables $X_i, X_j, X_k, X_l \in X(\subset V)$.

Proof. First, suppose that there exists a positive definite diagonal matrix Ω such that $\Sigma^*_{xx\cdot z} - \Omega$ is a negative semidefinite matrix and $\mathrm{rank}(\Sigma^*_{xx\cdot z} - \Omega) = 1$. Then, letting $q' = (q_1, \ldots, q_p)$ be a p-dimensional vector, we can obtain

$$\Sigma^*_{xx\cdot z} - \Omega = -qq'$$

through the singular value decomposition. Therefore, on the basis of the discussion preceding Theorem 2, equations (5) and (6) can be obtained.

For sufficiency, as we can assume that all the off-diagonal elements of $\Sigma^*_{xx\cdot z}$ are negative, by changing the signs of rows and corresponding columns of $\Sigma^*_{xx\cdot z}$, we shall show that a p-dimensional vector $q' = (q_1, \ldots, q_p)$ exists such that $\Sigma^*_{xx\cdot z} + qq'$ is a positive definite diagonal matrix. From equation (5), each element of $\sigma^*_{x_i x_j\cdot z}$ $(X_i \neq X_j)$ can be described as

$$\sigma^*_{x_i x_j\cdot z} = \frac{\sigma^*_{x_i x_k\cdot z}\sigma^*_{x_j x_l\cdot z}}{\sigma^*_{x_k x_l\cdot z}} = -\left(\left|\frac{\sigma^*_{x_i x_k\cdot z}\sigma^*_{x_i x_j\cdot z}}{\sigma^*_{x_j x_k\cdot z}}\right|\right)^{\frac{1}{2}}\left(\left|\frac{\sigma^*_{x_j x_k\cdot z}\sigma^*_{x_i x_j\cdot z}}{\sigma^*_{x_i x_k\cdot z}}\right|\right)^{\frac{1}{2}} = -q_i q_j.$$

Here, $q_i = (|\sigma^*_{x_i x_k\cdot z}\sigma^*_{x_i x_j\cdot z}/\sigma^*_{x_j x_k\cdot z}|)^{\frac{1}{2}}$ takes the same value regardless of the choice of X_j and X_k, since

$$\frac{\sigma^*_{x_i x_k\cdot z}\sigma^*_{x_i x_j\cdot z}}{\sigma^*_{x_j x_k\cdot z}} = \frac{\sigma^*_{x_i x_l\cdot z}\sigma^*_{x_i x_j\cdot z}}{\sigma^*_{x_j x_l\cdot z}} = \frac{\sigma^*_{x_i x_l\cdot z}\sigma^*_{x_i x_k\cdot z}}{\sigma^*_{x_k x_l\cdot z}}.$$

On the other hand, equation (6) is automatically satisfied since all off-diagonal elements of $\Sigma^*_{xx\cdot z}$ are negative, and the structure of $\Sigma^*_{xx\cdot z}$ is

$$\Sigma^*_{xx\cdot z} = \begin{pmatrix} (\sigma^*_{x_1 x_1\cdot z} + q_1^2) - q_1^2 & -q_1 q_2 & \cdots & -q_1 q_p \\ -q_1 q_2 & (\sigma^*_{x_2 x_2\cdot z} + q_2^2) - q_2^2 & \ddots & -q_2 q_p \\ \vdots & \ddots & \ddots & \vdots \\ -q_1 q_p & -q_2 q_p & \cdots & (\sigma^*_{x_p x_p\cdot z} + q_p^2) - q_p^2 \end{pmatrix}. \tag{7}$$

Hence, $\Sigma^*_{xx\cdot z} + qq'$ can be written as a positive definite diagonal matrix Ω. □

From Theorem 2, when observed statistical dependencies are generated by a selection criterion such as $a \leqslant S \leqslant b$, we can judge whether or not $X_i \perp\!\!\!\perp X_j | Z$ holds true in the whole population for any distinct variables $X_i, X_j \in X(\subset V)$, on the basis of the observed covariance matrix of the selected population.

Theorem 2 can be used to test whether or not statistical dependencies among four or more variables are generated by selection bias. However, it cannot be applied to test whether or not statistical associations among three variables are due to a selection variable. In this case, we provide a simple necessary condition for testing for the existence of selection bias.

COROLLARY. *For a covariance matrix* $\Sigma^*_{xx\cdot z}$, *a necessary condition that there exist a positive definite diagonal matrix* Ω *such that* $\Sigma^*_{xx\cdot z} - \Omega$ *is a negative semidefinite matrix and* $\mathrm{rank}(\Sigma^*_{xx\cdot z} - \Omega) = 1$ *is that, after we change the signs of rows and corresponding columns of* $\Sigma^*_{xx\cdot z}$, *all its off-diagonal elements are negative.*

An intuitive interpretation of the corollary is that the correlation between two variables which have no association with each other tends to be negative because of selection bias.

3·3. *Recovering the covariance matrix*

Suppose that we can judge that $X_i \perp\!\!\!\perp X_j | Z$ in the whole population holds true for any distinct variables $X_i, X_j \in X(\subset V)$ from the observed covariance matrix of the selected population on the basis of Theorem 2. Then we will provide a procedure for recovering the covariance matrix of the whole population without selection bias.

We consider the case in which the variance of the selection variable in the whole population is unknown. In Fig. 2, since $Z \perp\!\!\!\perp S | X$ holds true, by partitioning Y into $X \cup Z$ in equation (3), we obtain $\Sigma^{zs} = 0$. Thus, if we write

$$\begin{pmatrix} \Sigma^*_{xx} & \Sigma^*_{xz} \\ \Sigma^{*\prime}_{xz} & \Sigma^*_{zz} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma^{*xx} & \Sigma^{*xz} \\ \Sigma^{*xz\prime} & \Sigma^{*zz} \end{pmatrix},$$

according to equation (3), neither $\Sigma^{*zz} = \Sigma^{-1}_{zz\cdot x}$ nor $\Sigma^{*xz} = -\Sigma^{-1}_{xx}\Sigma_{xz}\Sigma^{-1}_{zz\cdot x}$ is dependent on S but Σ^{*xx} is. Without loss of generality, it is assumed that all the off-diagonal elements of $\Sigma^*_{xx\cdot z}$ are negative. We then use $\Sigma^*_{xx\cdot z} = \Omega - qq'$ in the proof of Theorem 2 to evaluate the $\Sigma_{xx\cdot z}$. As seen from the proof of Theorem 2, the off-diagonal elements and the diagonal elements of qq' can be taken to be $-\sigma^*_{x_i x_j\cdot z}$ and $-\sigma^*_{x_i x_k\cdot z}\sigma^*_{x_i x_j\cdot z}/\sigma^*_{x_j x_k\cdot z}$, respectively. By considering qq' as $B_{xs\cdot z}B'_{xs\cdot z}\check{\sigma}_{xx\cdot z}$ in equation (4), we can obtain $\Omega = \Sigma^*_{xx\cdot z} + qq' = \Sigma_{xx\cdot z}$ from equations (4) and (7). Then, if we substitute the inverse of Ω for Σ^{*xx}, the covariance matrix of $X \cup Z$ can be recovered and is given by

$$\begin{pmatrix} \Omega^{-1} & -\Sigma^{-1}_{xx}\Sigma_{xz}\Sigma^{-1}_{zz\cdot x} \\ -\Sigma^{-1}_{zz\cdot x}\Sigma'_{xz}\Sigma^{-1}_{xx} & \Sigma^{-1}_{zz\cdot x} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma'_{xz} & \Sigma_{zz} \end{pmatrix}. \tag{8}$$

This procedure indicates that Σ_{xx}, Σ_{xz} and Σ_{zz} can be evaluated through the covariance matrix of the selected population regardless of the observation of S. In particular, if the values of σ_{ss} and $\mathrm{var}(S | a \leqslant S \leqslant b)\,(\neq 0)$ are known, since $E(X | a \leqslant S \leqslant b) = B_{xs}E(S | a \leqslant S \leqslant b)$ and $E(XS | a \leqslant S \leqslant b) = B_{xs}E(S^2 | a \leqslant S \leqslant b)$, we can obtain

$$B_{xs} = \frac{\mathrm{cov}(X, S | a \leqslant S \leqslant b)}{\mathrm{var}(S | a \leqslant S \leqslant b)}.$$

Therefore, Σ_{sx} and Σ_{sz} can be calculated through

$$\Sigma_{xs} = B_{xs}\sigma_{ss}, \quad \Sigma_{sz} = \Sigma_{sx}\Sigma_{xx}^{-1}\Sigma_{xz},$$

where Σ_{xx}^{-1} and Σ_{xz} can be obtained from equation (8).

On the other hand, when conditioning on a single value of S, which can be represented by the case $\mathrm{var}(S|a \leqslant S \leqslant b) = 0$, we can obtain

$$\Sigma_{xx}^{*} = \Sigma_{xx.s} = \Sigma_{xx} - \frac{\Sigma_{xs}\Sigma_{xs}'}{\sigma_{ss}}, \tag{9}$$

and qq' in Theorem 2 is consistent with $\Sigma_{xs\cdot z}\Sigma_{sx\cdot z}/\sigma_{ss\cdot z}$. Here, Σ_{xx} can be obtained from equation (8) and $\Sigma_{xx\cdot s}$ can be evaluated from observed data. Hence, if the value of σ_{ss} is known, Σ_{xs} can be provided as a solution of equation (9). Note that the solution of equation (9) exists but is not unique. In addition, Σ_{sz} can be evaluated from $\Sigma_{sz} = \Sigma_{sx}\Sigma_{xx}^{-1}\Sigma_{xz}$. This procedure indicates that the covariance matrix of $X \cup Z \cup \{S\}$ can be recovered when the values of σ_{ss} and $\mathrm{var}(S|a \leqslant S \leqslant b)$ are available.

4. EXAMPLE

In this section, we illustrate the procedure of our approach with a classroom environment study reported by Church et al. (2001). The data were collected with the purpose of examining the relationship between undergraduates' perceptions of their classroom environment, their adoption of achievement goals for the course, their graded performance and their intrinsic motivation. The size of the sample is 297 and the variables of interest are lecture engagement (X_1), evaluation focus (X_2), harsh evaluation (X_3), evaluation type (X_4), mastery of goals (X_5), performance approach to goals (X_6), performance avoidance goals (X_7), intrinsic motivation (X_8) and graded performance (X_9).

Regarding this study Church et al. (2001) provided the graph shown in Fig. 3, which is considered as the path diagram of interest in this paper.

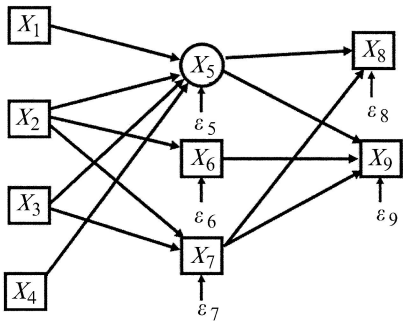

Fig. 3. Path diagram for the classroom environment study (Church et al., 2001).

We estimate the covariance matrix of $X_1, \ldots, X_9$ based on this path diagram and path coefficients given by Church et al. (2001), as shown in Table 1.

Throughout this section, it is assumed that the estimated covariance matrix in Table 1 and the data generating process shown in Fig. 3 represent the true causal relationships in the whole population. Then, for simplicity, our discussion starts by assuming that the conditional covariance matrix of $X = (X_1, \ldots, X_4)$ and $Z = X_6$ given $X_5 = x_5$ is observed; that is, X_5 is considered as a selection variable S.

Table 1: *Classroom environment study. The covariance matrix estimated from results in Church et al.* (2001)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1·000	0·000	0·000	0·000	0·370	0·000	0·000	0·141	0·074
X_2	0·000	1·000	0·000	0·000	−0·130	0·120	0·110	−0·077	−0·040
X_3	0·000	0·000	1·000	0·000	−0·230	0·000	0·210	−0·140	−0·105
X_4	0·000	0·000	0·000	1·000	0·230	0·000	0·000	0·087	0·046
X_5	0·370	−0·130	−0·230	0·230	1·000	−0·016	−0·063	0·396	0·215
X_6	0·000	0·120	0·000	0·000	−0·016	1·000	0·013	−0·009	0·133
X_7	0·000	0·110	0·210	0·000	−0·063	0·013	1·000	−0·274	−0·291
X_8	0·141	−0·077	−0·140	0·087	0·396	−0·009	−0·274	1·000	0·154
X_9	0·074	−0·040	−0·105	0·046	0·215	0·133	−0·291	0·154	1·000

First, we use Theorem 2 to judge whether or not statistical dependencies are generated by the selection variable S. With $X = (X_1, X_2, X_3, X_4)$ and $Z = X_6$, if we calculate the conditional covariance matrix $\Sigma^*_{xx\cdot z}$ based on Table 2, we can show that Theorem 2 holds true. Thus, we provide statistical justification for the assumption that statistical dependencies among X are indeed generated by a selection variable S.

Next, based on the result above, we recover the population's covariance matrix from Table 2. Since $q_1^2 = 0{\cdot}137$, $q_2^2 = 0{\cdot}016$, $q_3^2 = 0{\cdot}053$ and $q_4^2 = 0{\cdot}053$ from Theorem 2, we can obtain $\Omega = \mathrm{diag}(1{\cdot}000, 0{\cdot}986, 1{\cdot}000, 1{\cdot}000)$. Thus, by applying equation (8) to Table 2, we can obtain the covariance matrix of $X \cup Z$ for the whole population, which is consistent with the assumed true covariance matrix in Table 1. In addition, letting $W = X \cup Z$, since $\Sigma_{x_7 w} = \Sigma^*_{x_7 w}(\Sigma^*_{ww})^{-1}\Sigma_{ww}$ and $\sigma^*_{x_7 x_7 \cdot w} = \sigma_{x_7 x_7 \cdot w}$ from $\{S \cup Z\} \perp\!\!\!\perp X_7 | X$ (Wermuth, 1989), we can obtain the covariance matrix of $X \cup Z \cup \{X_7\}$ for the whole population, which is also consistent with the assumed true covariance matrix in Table 1. On the other hand, if we suppose that the variance of S is known to be 1·0, $\Sigma_{sx} = (-0{\cdot}370, 0{\cdot}130, 0{\cdot}230, -0{\cdot}230)$ can be obtained from equation (9). However, it is difficult to recover the covariance matrix of $\{X_8, X_9\}$ for the whole population.

Table 2: *Classroom environment study. The conditional covariance matrix given a single value of* X_5

	X_1	X_2	X_3	X_4	X_6	X_7	X_8	X_9
X_1	0·863	0·048	0·085	−0·085	0·006	0·023	−0·006	−0·006
X_2	0·048	0·983	−0·030	0·030	0·118	0·102	−0·025	−0·012
X_3	0·085	−0·030	0·947	0·053	−0·004	0·196	−0·049	−0·055
X_4	−0·085	0·030	0·053	0·947	0·004	0·014	−0·004	−0·004
X_6	0·006	0·118	−0·004	0·004	1·000	0·012	−0·003	0·137
X_7	0·023	0·102	0·196	0·014	0·012	0·996	−0·249	−0·277
X_8	−0·006	−0·025	−0·049	−0·004	−0·003	−0·249	0·843	0·069
X_9	−0·006	−0·012	−0·055	−0·004	0·137	−0·277	0·069	0·954

5. Discussion

Finally, we would like to point out some topics for further research. First, our criterion is based on the similarity between the covariance structure implied by factor models and that implied by selection variables. Therefore, just as observational equivalence problems may occur in factor analysis (Spirtes et al., 1993, Ch. 6), the same may happen in the

case of selection bias. Then, if criteria were developed for distinguishing observationally equivalent models regarding factor models, application of them to causal inference problems in the presence of selection bias would enable us to specify the plausible model of the whole population from observationally equivalent models. Secondly, this paper is motivated by the problem of causal discovery using the tetrad difference (Spirtes et al., 1993, Ch. 6). A parallel problem is to test whether or not a specific model can be identified by obervational data with selection bias. Some results regarding factor models, which are given by Davis (1993), Grzebyk et al. (2004), Stanghellini (1997) and Vicard (2000), help towards developing such identifiability criteria. Finally, Hipp & Bollen (2003) show that confirmatory tetrad analysis can handle dichotomous or ordinal variables by assuming that there are continuous underlying variables and the categorical ones represent discretised versions of them. Our results can be extended to the case where a selection variable is categorical under the same assumptions. Furthermore, often real data are not multivariate normal. Thus, it is necessary to extend our methods to more general statistical models, such as nonparametric models.

ACKNOWLEDGEMENT

The authors would like to thank Takaya Kojima of the Building Research Institute of Japan for his helpful discussion about the paper. We would also like to thank the editor and two anonymous referees whose comments significantly improved the presentation of the paper. This research was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Sumitomo Foundation, the Murata Overseas Scholarship Foundation, the Kayamori Foundation of Informational Science Advancement, the College Women's Association of Japan and the Japan Society for the Promotion of Science. Finally, we are grateful to Professor Judea Pearl for his encouragement and help.

REFERENCES

BEKKER, P. A. & DE LEEUW, J. (1987). The rank of reduced dispersion matrices. *Psychometrika* **52**, 125–35.

BERKSON, J. (1946). Limitation of the application of fourfold table analysis to hospital data. *Biomet. Bull.* **2**, 47–53.

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

BOLLEN, K. A. & LENNOX, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Phychol. Bull.* **110**, 305–14.

BOLLEN, K. A. & TING, K. (1993). Confirmatory tetrad analysis. *Sociol. Methodol.* **23**, 147–75.

BOLLEN, K. A. & TING, K. (2000). A tetrad test for causal indicators. *Psychol. Meth.* **5,** 3–22.

CAPITANIO, A., AZZALINI, A. & STANGHELLINI, E. (2003). Graphical models for skew-normal variates. *Scand. J. Statist.* **30**, 129–44.

CHURCH, M. A., ELLIOT, A. J. & GABLE, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *J. Educ. Psychol.* **93**, 43–54.

COOPER, G. F. (1995). Causal discovery from data in the presence of selection bias. In *Proc. Fifth Workshop Artific. Intell. Statist*., Ed. D. Fisher and H.-J. Lenz, pp. 140–50. Fort Lauderdale, FL: Society for Artificial Intelligence and Statistics.

COOPER, G. F. (2000). A Bayesian method for causal modeling and discovery under selection. In *Proc. 16th Conf. Uncertainty Artif. Intel*., Ed. C. Boutilier and M. Goldszmidt, pp. 98–106. San Francisco, CA: Morgan Kaufmann.

DAVIS, W. R. (1993). The FC1 rule of identification for confirmatory factor analysis: A general sufficient condition. *Sociol. Meth. Res.* **21**, 403–37.

DAWID, A. P. (2002). Influence diagrams for causal modeling and inference. *Int. Statist. Rev.* **70**, 161–89.

EDWARDS, J. R. & BAGOZZI, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychol. Meth.* **5**, 155–74.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* **14**, 300–6.
Grzebyk, M., Wild, P. & Chouanière, D. (2004). On identification of multi-factor models with correlated residuals. *Biometrika* **91**, 141–51.
Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–62.
Hernan, M. A., Hernandez-Diaz, S. & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology* **15**, 615–25.
Hipp, J. R. & Bollen, K. A. (2003). Model fit in structural equation models with censored, ordinal, and dichotomous variables: Testing vanishing tetrads. *Sociol. Meth.* **33**, 267–305.
Johnson, N. L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. New York: John Wiley & Sons.
Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. New York: John Wiley & Sons.
Spearman, C. (1904). General intelligence objectively determined and measured. *Am. J. Psychol.* **15**, 201–93.
Spearman, C. (1928). Pearson's contribution to the theory of two factors. *Br. J. Psychol.* **19**, 95–101.
Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction, and Search*. New York: Springer-Verlag.
Spirtes, P., Meek, C. & Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation, and Discovery*, Ed. C. Glymour and G. Cooper, pp. 211–52. Cambridge, MA: MIT/AAAI Press.
Stanghellini, E. (1997). Identification of a single-factor models using graphical Gaussian rules. *Biometrika* **84**, 241–4.
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.
Winship, C. & Mare, R. D. (1992). Models for sample selection bias. *Ann. Rev. Sociol.* **18**, 327–50.

[*Received June* 2004. *Revised February* 2006]