On indirect confounding

Nanny Wermuth, Mathematical Statistics, Chalmers/Göteborgs Universitet, Sweden*
D.R. Cox, Nuffield College, Oxford, UK

Summary

Unnoticed confounding may severely distort the direction and strength of the dependence of a response on some of its explanatory variables, given from a stepwise data generating process. This holds both for direct confounding connected mainly to observational studies and for indirect confounding which we define and study in this paper and which may also be present in intervention studies. We provide graphical and matrix criteria to decide on the absence or presence of indirect confounding. For linear systems with indirect confounding, we derive the corrections for linear least squares regression coefficients that are needed to recover the coefficients of the generating process.
Key words: Graphical Markov models, identification, independence equivalence, independence graphs, linear least squares regressions, parameter equivalence, recursive regression graphs, structural equation models, triangular systems

1 Introduction

The paper concerns consequences of stepwise data generating processes for a sequence of random variables $(Y_1, \ldots, Y_d)$, when some background variables are unobserved. Of the generating dependencies, i.e. those that are part of the process for all d variables, some remain for the observed variables but may become confounded, that is mixed with effects of paths containing unobserved variables.

Confounding arises when for a response variable Y_i with a generating dependence on Y_j, there is in the process also a (partly) unobserved path of a special type connecting this variable pair (Y_i, Y_j). For example in linear systems, such a confounding path generates a residual correlation that destroys the interpretation of least squares regression coefficients as measuring a generating dependence. We speak of direct confounding when the confounding path concerns exclusively variables omitted from the generating process and of indirect confounding when the path contains, in addition, some observed variables of a special kind.

* *Address:* Chalmers, Eklandagatan 86, 41296 Göteborg; e-mail: wermuth@math.chalmers.se

When direct confounding is present, it may not be possible to recover the generating dependence from the remaining observed variables alone. Possibilities and limitations for the estimation of a generating dependence in the presence of so-called instrumental variables, have for such cases been discussed in an extensive literature which builds on early work by Sargan (1958). We show in this paper that even if there is no direct confounding, similar distortions of generating dependencies may occur by indirect confounding and we derive corrections on the basis of the observed variables, that are needed for linear least squares coefficients.

We assume that the structure of the generating process is captured by a graph of nodes and edges, in which node i corresponds to variable Y_i and an ij-edge for $i < j$ is an arrow starting at node j and pointing to node i. There are some potentially explanatory variables for Y_i, denoted by $Y_{r(i)}$ with $r(i) = (i+1, \ldots, d)$. From each node in a subset of these, called the parents of i and denoted by $\mathrm{par}(i)$, an arrow points to i; the parent nodes identify which of the potentially explanatory variables are directly explanatory for Y_i. Then, either a joint density f is generated, written in a condensed notation of nodes, as

$$f = \prod_{i=1}^{d} f_{i|r(i)} = \prod_{i=1}^{d} f_{i|\mathrm{par}(i)}, \tag{1}$$

or a linear system of recursive equations is generated with regressions of Y_i on vector variables Y_a, so that conditional expectations in mean-centered variables are expressable with row vectors $\Pi_{i|a}$ of least squares regression coefficients (Cramér, 1946, p. 302) as

$$\mathrm{E}(Y_i|Y_{r(i)}) = \Pi_{i|\mathrm{par}(i)} Y_{\mathrm{par}(i)}. \tag{2}$$

Processes given by equations (1) and (2) form a subclass of graphical Markov models, called triangular systems; see Wermuth and Cox (2004) for derivations and discussions of different types of properties and consequences.

By construction, the corresponding directed graph in ordered nodes $V = (1, \ldots, d)$, called the parent graph, G^V_{par}, is acyclic, i.e. it contains no directed cycles. Densities of arbitrary form (1) and least squares equations (2) are said to be generated over a given parent graph, by starting with the last background variable, Y_d, continuing with Y_{d-1}, up to Y_1, a response of primary interest.

We speak of a generating dependence of Y_i on Y_j in the system, if and only if j is a parent of offspring i, i.e. for $i \prec\!\!-\!\!- j$ and $i < j$. A path in the graph is a sequence of edges. A node j is called an ancestor of node i if there is a direction-preserving path leading from j to node i, i.e. for $i \prec\!\!-\!\!- \circ \prec\!\!-\!\!- \ldots \prec\!\!-\!\!- \circ \prec\!\!-\!\!- j$. Then, one says alternatively that node i is a descendant of j and that the sequence of edges is a descendant-ancestor path. An edge is viewed as path of length one, so that a parent is also an ancestor. There is an only indirect dependence of Y_i on Y_j if and only if node j is a forefather of i, i.e. when j is an ancestor but not a parent in G^V_{par}.

We suppose further that the purpose of an empirical study is to investigate the generating dependence of the primary response Y_1 on one or several of its directly explanatory variables

Y_j, when only the first d_N variables are observed, i.e. when we have an ordered split

$$V = (N, M), \quad N = (1, \ldots, d_N), \quad M = (d_N + 1, \ldots, d)$$

and variables Y_M are unobserved since marginalized over in the generating process.

A whole set of background variables may be unobserved, for instance in a controlled clinical trial when some information on the health status of the patients is unavailable because it would require previous records that were never obtained in the country of study, or that would need tests which are considered to be too time-consuming before treatment. In a marketing study in a particular region, information on previous buying behaviour may be unavailable for a whole range of products as well as on related advertisement campaigns. In a study of the performance of used cars, some information may no longer be accessible which only first owners can provide.

Then some of the generating dependencies may remain unchanged, others may appear modified. Such changes are essential for an understanding of processes, especially but not exclusively when they are to be interpreted causally (Cox, 1992; Cox and Wermuth, 2001, 2004) or they are used for comparing results of studies with smaller sets of background variables to those of the generating process. Related issues have been discussed as so-called over-conditioning (Edwards, 2000) or as changes in probability distributions when intervening on instead of observing a stepwise generating process for densities (Lindley, 2002).

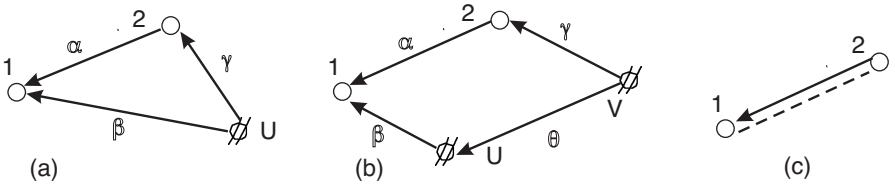


Figure 1: Two examples of direct confounding. a) In the parent graph with Y_1 dependent on both Y_2 and U, with Y_2 dependent on U alone, the generating coefficient α of Y_1 on Y_2 becomes directly confounded by $\beta\gamma$ since the common parent path connecting (Y_1, Y_2) via U is unobserved; b) the generating coefficient α becomes confounded by $\beta\gamma\theta$ since the common ancestor path connecting (Y_1, Y_2) via U, V is unobserved; c) the graph generated by marginalizing over the unobserved nodes in both, Figures 1a) and 1b), the double edge points to direct confounding

It is well known that an unobserved variable which affects both a response Y_1 and one of its directly explanatory variables Y_2 may severely distort a generating dependence. This case, shown in G^V_{par} of Figure 1a), is an example of direct confounding. In both Figures 1a) and 1b), the two parents graphs include some unobserved variables and standardized least squares regression coefficients are attached to each arrow. Both lead to the same graph 1c) for the remaining observed variables. Dashed lines indicate in linear systems correlated residuals. The double edge, an arrow and a dashed line, points to a directly confounded dependence.

Graphical criterion 1: *Detecing direct confounding in the parent graph.* An offspring-parent pair (i, j) (a pair with $i \prec\!\!\leftarrow j$) is directly confounded if it is connected in $G^V_{\rm par}$ by an unobserved common-ancestor path, i.e. by a path like

$$i \prec\!\!\leftarrow \not\emptyset \rightarrow\!\!\succ j \quad \text{or} \quad i \prec\!\!\leftarrow \not\emptyset \ldots \prec\!\!\leftarrow \not\emptyset \rightarrow\!\!\succ \ldots \not\emptyset \rightarrow\!\!\succ j,$$

where the unobserved variables are shown as circles crossed out.

For linear systems of equations, the amount of confounding is compactly described in terms of regression coefficients for variables standardized to have mean zero and variance one. For Figure 1a) the generating equations are

$$\begin{aligned} Y_1 &= \alpha Y_2 + \beta U + \varepsilon_1 \\ Y_2 &= \gamma U + \varepsilon_2 \\ U &= \varepsilon_u \end{aligned} \tag{3}$$

with each residual, ε_i, being uncorrelated with the explanatory variables in the equation and hence with each other. When the unobserved variable U is regarded as part of a new residual the following two linear equations in observed variables result, which are an example of so-called recursive regressions with correlated residuals (Goldberger, 1964)

$$\begin{aligned} Y_1 &= \alpha Y_2 + \eta_1 \\ Y_2 &= \eta_2 \end{aligned} \tag{4}$$

where the residuals, η, are correlated since both contain the unobserved variable U of the parent graph in Figure 1a),

$$\eta_1 = \beta U + \varepsilon_1, \quad \eta_2 = \gamma U + \varepsilon_2.$$

In linear equations generated over any parent graph, each coefficient of dependence is a least squares regression coefficient. For standardized variables in equations (3) and the parent graph in Figure 1a), we have in particular $\alpha = (\rho_{12} - \rho_{1u}\rho_{2u})/(1 - \rho^2_{2u})$, where ρ denotes a correlation coefficient. The generating coefficient α is preserved in the structural equations (4), obtained by marginalizing in the generating equations (3) over U. But with α and the residual correlation, there are more parameters than can be estimated, given observations for Y_1 and Y_2 alone.

By contrast, in the linear least squares equation for the dependence of Y_1 on Y_2

$$Y_1 = \rho_{12} Y_2 + \varepsilon,$$

the standardized regression coefficient, ρ_{12}, may be estimated, but is a confounded measure of the generating coefficient α, since from equations (3) or (4) $\rho_{12} = \alpha + \beta\gamma$. For the linear system

to Figure 1b), one finds in a similar way that $\rho_{12} = \alpha + \beta\gamma\theta$, so that the amount of confounding of the generating coefficient α for this standardized least squares coefficient given is by $\beta\gamma\theta$.

A least squares regression coefficient may be substantially changed in magnitude or even in sign compared to the generating coefficient. Whether such changes are of qualitative importance for interpretation depends on the strengths of the unobserved parts of the generating process and could be studied by sensitivity analysis (Rosenbaum, 2002).

An example of a generating system, in which marginalizing over variables U, V does not lead to direct confounding, is described by the parent graph in Figure 2. It is for five quantitative variables and one binary variable A which captures whether a bladder substitute leads to continent or incontinent urine diversion.

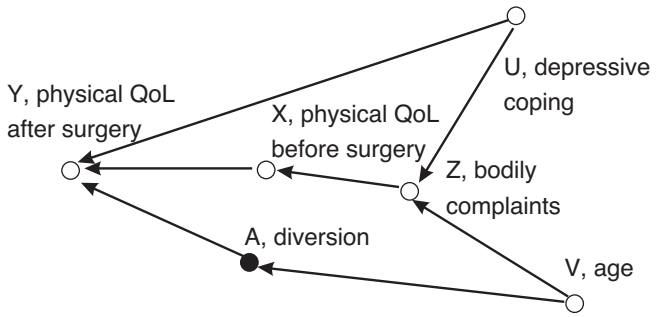


Figure 2: A potential generating process for physical quality of life after surgical removal of the bladder as given for a subgroup of male patients with a bladder tumor; data by Hardt et al. (2004)

When both U and V are unobserved, there is no direct confounding of the generating dependencies for pairs (Y, X) and (Y, A); see the graphical criterion 1. In the graph of the remaining observed variables, no double edge is generated. Instead, two pairs with a missing edge in the parent graph, (Y, Z) and (A, Z), become coupled, i.e. joined by an edge. The new path from A to Y via Z causes confounding for (Y, A) in linear least squares regression of Y on both A, X. Expressed differently, marginalizing over U, V and conditioning Y on A, X generates indirect confounding of the generating dependence of Y on A. The dependence of main substantive interest, of quality of life, Y, on the type of bladder restoration, A, is confounded even though there is no direct confounding. It is an example of what we study in this paper as indirect confounding, to be defined formally in Section 3.

The plan of the paper is to introduce some general notation and previous results in Section 2. In Section 3 linear systems of equations generated by unobserved background variables are studied further to obtain least squares regression equations that are parameter equivalent, to find corrections for indirectly confounded coefficients, and to derive the graphical criteria to decide on the absence or presence of indirect confounding. In Section 4 we discuss briefly some

related results. In an appendix we give a numerical example in which a least squares regression coefficient is indirectly confounded and has similar strength, but a reversal in direction, compared to the generating dependence.

2 Some notation and previous results

2.1 Linear triangular systems

For a linear triangular system we take without loss of generality the $d \times 1$ vector random variable Y to be mean-centered. The system of equations, corresponding to the conditional expectations (2) and written in matrix form, is

$$AY = \varepsilon, \tag{5}$$

where A is an upper-triangular matrix with unit diagonal elements and ε is a vector of zero-mean, uncorrelated random variables, called residuals. The diagonal form of the residual covariance matrix $\mathrm{cov}(\varepsilon) = \Delta$ is equivalent to specifying that each row of A in (5) defines a linear least squares regression equation.

With $\beta_{i|j.c}$ denoting the least squares regression coefficient of Y_j when response Y_i is regressed on Y_j and Y_c, i.e. on all variables with indices listed after the conditioning sign, we build on the Yule-Cochran notation for such coefficients. Then for instance, the matrix version of the complete system of equation (3) generated over the graph in Figure 1a) for mean-centered variables, is

$$AY = \begin{pmatrix} 1 & -\beta_{1|2.3} & -\beta_{1|3.2} \\ 0 & 1 & -\beta_{2|3} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ U \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_u \end{pmatrix},$$

The nonzero elements of A for $i < j$ are in general

$$-a_{ij} = \beta_{i|j.r(i)\setminus j} = \beta_{i|j.\mathrm{par(i)}\setminus j}, \tag{6}$$

Zero values a_{ij} in the upper triangular part of A represent the vanishing contributions to the regression of Y_i on $Y_{r(i)}$ of all variables that do not correspond to parent nodes of i. Individual regression coefficients may be collected in the row vector denoted by $\Pi_{i|a}$. For instance in the case $a = (3, 5)$, the individual components of $\Pi_{i|a}$ are $\Pi_{i|a} = (\beta_{i|3.5} \;\; \beta_{i|5.3})$. For a split of a into two components, each with possibly more than one element, $a = b \dot{\cup} c$, we write alternatively

$$\Pi_{i|a} = \Pi_{i|b,c} = (\Pi_{i|b.c} \;\; \Pi_{i|c.b}).$$

In this notation the vector version of Cochran's (1938) recursion relation for linear least squares regression coefficients (for a proof see Wermuth and Cox, 2004), becomes

$$\Pi_{i|c} = \Pi_{i|c.b} + \Pi_{i|b.c}\Pi_{b|c}. \tag{7}$$

It shows in particular that $\Pi_{i|c} = \Pi_{i|c.b}$ if $\Pi_{i|b.c} = 0$ and thereby proves equation (6).

The matrix pair (A, Δ^{-1}) may be obtained by successive orthogonalization (Gram, 1883; Schmidt, 1907; Dempster, 1969, chapter 4) of Σ and the pair defines a triangular decomposition of the concentration matrix with $\Sigma^{-1} = A^{\mathrm{T}}\Delta^{-1}A$. This decomposition is unique for the given fixed order $(d, d-1, \ldots, 1)$.

Linear triangular systems have been introduced as path analyses in genetics (Wright, 1923; 1934) and as linear recursive equations with uncorrelated residuals in econometrics (Wold, 1954). They form a subclass of linear structural equations, see e.g. Goldberger (1964). Early studies of their properties include Tukey (1954), Wermuth (1980), Kiiveri et al. (1984).

A matrix representation, $\mathcal{A}$, of the parent graph G^V_{par}, associated with any given linear triangular system, is binary, i.e. has zero-one elements, and is of upper-triangular form. It is obtained as the indicator matrix of A in equation (5)

$$\mathcal{A} = \mathrm{In}(A).$$

It has been called the edge matrix of the parent graph, sometimes denoted by $\mathrm{Ed}(G^V_{\mathrm{par}})$, and it gives the defining structural zeros of the family of real-valued matrices of which A is a member.

2.2 Generating processes for linear recursive regressions with correlated residuals

If in a linear triangular system (5) common background variables are unobserved, then linear recursive equations with correlated residuals are generated, another subclass of linear structural equations. More generally, let (N, M) be an ordered split of V so that Y_N denotes the first d_N variables of the generating process that remain observed and Y_M denotes the variables marginalized over in the generating process so that they are now unobserved. Then, the recursive equations induced by the generating system (3), are

$$A_{NN}Y_N = \eta_N, \tag{8}$$

where

$$\eta_N = \varepsilon_N - A_{NM}A^{-1}_{MM}\varepsilon_M, \quad \kappa = \mathrm{cov}(\eta_N) = \Delta_N + (A_{NM}A^{-1}_{MM})\Delta_M(A_{NM}A^{-1}_{MM})^{\mathrm{T}},$$

and $A_{NM} = [A]_{N,M}$ denotes the submatrix of A of rows N and columns M, both A_{NN} and A_{MM} are upper-triangular submatrices of A. Orthogonality, i.e. uncorrelatedness, of the parameters in equations (8) to those in Y_M, may be recognized by direct computations or as a special case of Corollary 1 in Wermuth and Cox (2004).

We denote the graph in nodes N of these induced recursive regressions by G^N_{rec}. It has two types of edge, arrows for dependencies, corresponding to A_{NN} and dashed lines for associations, corresponding to κ, both defined in terms of submatrices of the generating edge matrix $\mathcal{A}$, as

$$\mathrm{Ed}(G^N_{\mathrm{rec}}) = \{\mathcal{A}_{NN}, \quad \mathcal{K} = \mathrm{In}[I_{NN} + (\mathcal{A}_{NM}\mathcal{A}^{MM})(\mathcal{A}_{NM}\mathcal{A}^{MM})^{\mathrm{T}}]\}, \tag{9}$$

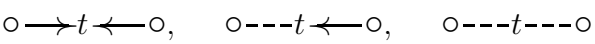
where

$$\mathcal{A}^{MM} = \mathrm{In}[(2I_{MM} - \mathcal{A}_{MM})^{-1}],$$

and e.g. I_{NN} denotes the identity matrix of size d_N; for a proof see Lemma 4b) and equation (33) of Wermuth and Cox (2004).

The upper triangular edge matrix for the arrows, $\mathcal{A}_{NN}$, gives the subgraph induced by nodes N in the parent graph so that there is an ij-arrow in G^N_{rec} if and only if there is an ij-arrow in G^V_{par}. The symmetric edge matrix, $\mathcal{K}$, has an off-diagonal ij-one, i.e. gives a dashed line edge for nodes i, j, if and only if an unobserved common-ancestor path connects nodes i and j in G^V_{par}. To see this, note equation (9), where an ik-one in $\mathcal{A}_{NM}$ points to an ik-arrow in the parent graph with i in N and k in M and a kl-one in $\mathcal{A}^{MM}$ points to a descendant-ancestor path connecting k to l with all nodes in M, so that there is a nonzero ij-entry in $(\mathcal{A}_{NM}\mathcal{A}^{MM})(\mathcal{A}_{NM}\mathcal{A}^{MM})^{\mathrm{T}}$ for a common-ancestor path with all nodes along it in M.

Recursive regression graphs generated in this way by equations (9) may have double edges; see equations (3) as the simplest example. For graphs associated with general Gaussian structural equation models, Koster (1999) has shown how to read off all independencies. Also Smith (1989) has pointed out that criteria for reading probabilistic independencies off graphs may, in addition, be used for reading linear independencies off the same graph for corresponding linear equations, in which no distributional form is specified for the residuals. Essential for implied independencies are configurations defining so-called collision nodes t



Separation criterion for G^N_{rec} (Koster, 1999). The recursive regression graph (9) implies for all Gaussian equations (8) that Y_i independent of Y_j given another (vector) variable Y_C, if and only if i and j are separated by C in G^N_{rec}, i.e. along every path from i to j there is either a collision node, which is together with all its descendants outside C, or there is a non-collision node within C (or both).

A pure collision path in G^N_{rec} is a path for which all nodes along it, i.e. all nodes except the path endpoints, are collision nodes. One consequence of the separation criterion is that nodes i and j are not separated by nodes in C when there is a pure collision path connecting (i, j), with each collision node along it having a descendent in C. For linear systems, where each edge present in G^N_{rec} corresponds to a nonzero parameter in equations (8), such a path connecting (i, j) means a nonzero contribution to the partial correlation of Y_i, Y_j given Y_C.

3 Indirect confounding in the absence of direct confounding

In this Section we can now study linear recursive regression equations (8) for which the graph G^N_{rec}, with edge matrix given by equation (9), has at most one edge for each node pair so that there is no direct confounding of generating dependencies.

3.1 The amount of indirect confounding

To simplify notation we write the induced recursive regressions (8), in observed variables Y and without direct confounding, as

$$HY = \eta, \tag{10}$$

with $\mathrm{cov}(\eta) = \kappa$ and $H = A_{NN}$ so that the observed covariance matrix Σ_{NN} and its inverse, the overall concentration matrix, are

$$\Sigma_{NN} = H^{-1}\kappa H^{-\mathrm{T}}, \qquad \Sigma_{NN}^{-1} = H^{\mathrm{T}}\kappa^{-1}H,$$

with elements σ_{ij} of Σ_{NN} and κ_{ij} of κ relating to nodes i and j of the generating process. Lack of direct confounding implies that for each pair (i, j) we can have $h_{ij} \neq 0$ or $\kappa_{ij} \neq 0$ but not both.

Graphical criterion 2. *Detecting indirect confounding in a recursive regression graph without double edges.* An offspring-parent pair (i, j) (a pair with $i \leftarrow\!\!- j$) is indirectly confounded if it is connected in G^N_{rec} by a collision-forefather path, i.e. by

$$i\text{---}\circ\text{---}\circ \ldots \circ\text{---}j, \quad \text{or} \quad i\text{---}\circ \ldots \circ\text{---}\circ \prec\!\!\!- j,$$

where all nodes along the path are forefather nodes of i, i.e. ancestors but not parents.

For equations (10), where each edge in the corresponding recursive regression graph corresponds to a nonzero parameter, the two types of paths shown above are association-inducing. There is only one further type of a pure collision path. It has the configuration $i\text{---}\circ$, in the path above on the right-hand side, replaced by $i \rightarrow\!\!\!\circ$, but then not all nodes along the path could be forefathers of node i. Therefore, the two types of path of the graphical criterion 2, are the only types of path which can lead to indirect confounding of the generating dependence for (i, j) by marginalizing over M and conditioning on the parents or the ancestors of i, present in the recursive regression graph.

To derive the amount of confounding, we note first that for $i = d_N$, we have by the triangularity of H that $\sigma_{ii} = \kappa_{ii}$. For $i = d_N - 1$, the relations between $\sigma_{ii}, \beta_{i|i+1}$ and $\kappa_{ii}, h_{i,i+1}, \kappa_{i,i+1}$ depend on the zero constraints, i.e. on whether $h_{i,i+1} = 0$ or $\kappa_{i,i+1} = 0$ or both, and are readily obtained. Next, we assume that for some other node i the submatrices of Σ_{NN}, H and κ are given which correspond to nodes in N larger than i, so that we can consider the relations for i together with its ancestor nodes in G^N_{rec}.

For this, we denote by $a = \mathrm{anc}(i)$ the set of ancestor nodes of node i in the recursive regression graph, by $S = (i, \mathrm{anc}(i))$ of size d_S, the ordered set that we call its stem family, and by O the set of nodes outside this family, $O = N \setminus S$. Since no directed path leads from any node in O to S, it is possible to omit Y_O from the system without affecting relations within

Y_S. This follows by orthogonalizing the two systems in O and S, which have $H_{SO} = 0$, and are written in the order (O, S) as

$$\begin{pmatrix} H_{OO} & H_{OS} \\ 0 & H_{SS} \end{pmatrix} \begin{pmatrix} Y_O \\ Y_S \end{pmatrix} = \begin{pmatrix} \eta_O \\ \eta_S \end{pmatrix}.$$

If we denote by $(G_{SS}, \Delta_{SS}^{-1})$ the triangular decomposition of κ_{SS}^{-1}, obtained by starting with the last component in S and proceeding to the first, then the two representations

$$\kappa_{SS}^{-1} = G_{SS}^{\mathrm{T}} \Delta_{SS}^{-1} G_{SS}, \quad \Sigma_{SS}^{-1} = H_{SS}^{\mathrm{T}} \kappa_{SS}^{-1} H_{SS}$$

can be combined to give a triangular decomposition of the concentration matrix of Y_S with $(P_{SS}, \Delta_{SS}^{-1})$ for the same fixed order. By the uniqueness of these decompositions

$$P_{SS} = G_{SS} H_{SS}. \tag{11}$$

with

$$G_{SS} = \begin{pmatrix} 1 & -\gamma_{i|a} \\ 0 & G_{aa} \end{pmatrix}, \quad H_{SS} = \begin{pmatrix} 1 & h_{ia} \\ 0 & H_{aa} \end{pmatrix}, \quad P_{SS} = \begin{pmatrix} 1 & -\Pi_{i|a} \\ 0 & P_{aa} \end{pmatrix},$$

where $\gamma_{i|a} = \kappa_{ia} \kappa_{aa}^{-1}$ denotes the least squares regression coefficient vector of η_a when regressing the residual η_i on η_a.

The key relation between the least-squares regression coefficients in $\Pi_{i|a}$ and the equation parameters in h_{ia} is then from (11)

$$-\Pi_{i|a} = h_{ia} - \gamma_{i|a} H_{aa}. \tag{12}$$

This quantifies the amount by which the least squares regression coefficients are indirectly confounded. One component of $\Pi_{i|a}$ is the vector of generating dependencies h_{ia}, the other contains contributions of confounding paths. For purposes of estimation it is useful to study parameter equivalence of the least squares and the recursive regression equations.

3.2 Parameter equivalent equations

In the case of parameter equivalence of two sets of parameters, each parameter of the first set can be obtained in terms of those in the second set and vice versa. Here we consider a linear system with constraints specified by the graph G_{rec}^N without double edges and show parameter equivalence of the ith least squares equation involving $(\Pi_{i|a},\ \delta_{ii})$ and the i-th recursive regression equation with parameters $(h_{ia},\ \kappa_{ia},\ \kappa_{ii})$, both given Σ_{aa} and its decompositions. If we start with the recursive regression equation, then the parameters of the least squares regression equation are given by equation (12) and $\delta_{ii} = [\Delta]_{i,i}$.

Conversely, given the parameters of the i'th least squares equation each parameter in the i'th recursive regression equation may also be obtained. To see this, we define h_{ia} and k_{ia} using the matrix $Q_{aa} = \kappa_{aa}^{-1} H_{aa}$, possibly in reordered form. We consider two cases separately.

First, let G^N_{rec} have no missing edge for node i and nodes in $a = b\dot{\cup}c$, where b denotes nodes with arrows pointing to node i and c denote nodes with dashed lines connecting node i to ancestors of i in G^N_{rec}, then $\kappa_{ib} = 0$, $h_{ic} = 0$ and

$$\begin{aligned} \kappa_{ic} &= \Pi_{i|c.b} Q_{cc}^{-1} \\ -h_{ib} &= \Pi_{i|b.c} + \Pi_{i|c.b} Q_{cc}^{-1} Q_{cb} \\ \kappa_{ii} &= \sigma_{ii} + \sigma_{ia} h_{ia}^T + h_{ia} H_{aa}^{-1} \kappa_{ai}. \end{aligned} \tag{13}$$

The result follows by using the zero constraints on equation parameters and residual covariances and rewriting equation (12) as

$$-(\Pi_{i|b.c}\ \Pi_{i|c.b}) = (h_{ib}\ 0) - (0\ k_{ic}) Q_{aa}.$$

This gives

$$-\Pi_{i|b.c} = h_{ib} - k_{ic} Q_{cb}, \quad -\Pi_{i|c.b} = -k_{ic} Q_{cc},$$

and therefore the first two equations. The last equation results from $k_{Si} = [H_{SS} \Sigma_{SS} H_{SS}^T]_{S,i}$.

Second, let there be some ancestor nodes of i not coupled to node i, i.e. there be a a subset d of a with ij-missing edges in G^N_{rec}, so that we can take $a = b\dot{\cup}c\dot{\cup}d$ with b and c as in the first case. Then, with

$$-(\Pi_{i|b.cd}\ \Pi_{i|c.bd}\ \Pi_{i|d.cb}) = (h_{ib}\ 0\ 0) - (0\ k_{ic}\ 0) Q_{aa}$$

we get $\Pi_{i|d.cb} = -\kappa_{ic} Q_{cd}$, in the first two equations of (13) the coefficients $\Pi_{i|c.b}$ are replaced by $\Pi_{i|c.bd}$ and $\Pi_{i|b.c}$ by $\Pi_{i|b.cd}$, while the last equation remains unchanged.

Parameter equivalence implies that maximum-likelihood estimates are in the same one-to-one correspondence (Fisher, 1922; p. 327). This applies here to the two types of equation parameters given Σ_{aa} in Gaussian systems and it is possible to justify these estimates also for non-Gaussian linear systems.

If we define next an upper-triangular matrix F by adding $-\Pi_{i|a}$ from equation (12) to row i of the identity matrix I_{NN} for $i = 1, \ldots, N-1$, then, in general, this matrix does not give the equation parameters in a triangular decomposition of the overall observed concentration matrix Σ_{NN}^{-1}. Instead,

$$F\ \Sigma_{NN} F^T = \tau \tag{14}$$

may contain some nonzero residual covariances τ_{ij} and therefore defines a subclass of recursive regression models without direct confounding, in which each equation parameter coincides with a least squares coefficient in the observed variables.

It follows by the separation criterion given in Section 2.2 and by construction that linear equations with F from (14) preserve the independencies of the corresponding system of recursive equations (10). The key relation (12) between rows in F and in H implies that the intepretation of equation parameters in F may be drastically changed compared to the generating dependencies in H (see the numerical example in the appendix), but that in linear systems it is always possible to correct for this indirect confounding to recover the generating dependencies.

3.3 Illustrations for the correction of indirect confounding

We give two small examples as illustrations. The first is a linear system to a parent graph as in Figure 2, the second is an example for indirect confounding as it may occur in an intervention study.

Illustration 1. In the following equations in four observed variables and two uncorrelated unobserved variables U and V, all variables are mean-centred

$$Y_1 = \alpha Y_2 + \gamma Y_3 + \beta U + \varepsilon_1, \quad Y_2 = \delta Y_4 + \varepsilon_2, \quad Y_3 = \theta V + \varepsilon_3, \quad Y_4 = \psi U + \xi V + \varepsilon_4. \quad (15)$$

An interpretation of equation parameter α, for instance, is $\alpha = \beta_{1|2.3U} = \beta_{1|2.34U}$, since Y_2, Y_3 and U are the directly explanatory variables of Y_1 and this response is generated without a contribution of variable Y_4. Figure 3a) shows the corresponding parent graph.

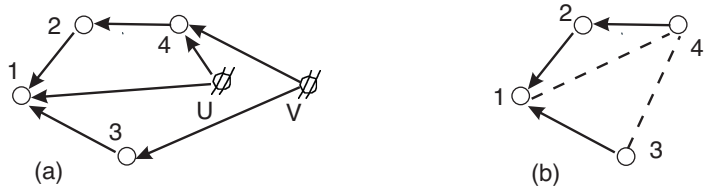


Figure 3: (a) Parent graph of the generating equations (15) in four observed and two unobserved variables (U, V). (b) Recursive regression graph of equations (16) in observed variables, derived from parent graph in (a) by marginalizing over (U, V); indirect confounding, present for linear least squares regression coefficients in observed variables measuring dependence of Y_1 on Y_3 given Y_3, Y_4, can be corrected to recover the generating dependence γ.

The corresponding system of recursive equations are obtained from the generating equations (15) by using

$$\eta_1 = (\beta U + \varepsilon_1), \quad \eta_2 = \varepsilon_2, \quad \eta_3 = (\theta V + \varepsilon_3), \quad \eta_4 = (\psi U + \xi V + \varepsilon_4)$$

to give

$$Y_1 = \alpha Y_2 + \gamma Y_3 + \eta_1, \quad Y_2 = \delta Y_4 + \eta_2, \quad Y_3 = \eta_3, \quad Y_4 = \eta_4, \quad (16)$$

so that $\mathrm{cov}(\eta) = \kappa$ is not a diagonal matrix, there are nonzero residual correlations κ_{14} and κ_{34}. In this example we have $A_{MM} = I_{MM}$ so that equation (9) reduces to

$$\mathcal{K} = \mathrm{In}[I_{NN} + \mathcal{A}_{NM}^{\mathrm{T}} \mathcal{A}_{NM}^{\mathrm{T}}],$$

and indicates unobserved common-parent paths.

The ancestors of node $i = 1$ are $\mathrm{anc}(1) = (2, 3, 4)$, the coefficient matrix H_{SS} for the recursive regressions to graph 3b) and the triangular decomposition matrix G_{SS} of κ_{SS}^{-1} (see e.g. Wermuth et al., 2005) are

$$G_{SS} = \begin{pmatrix} 1 & 0 & \kappa_{14}\kappa_{34}/D_{34} & -\kappa_{14}D_3/D_{34} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -\kappa_{34}/D_4 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad H_{SS} = \begin{pmatrix} 1 & -\alpha & -\gamma & 0 \\ 0 & 1 & 0 & -\delta \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where D denotes determinants of submatrices of κ; $D_i = \kappa_{ii}$, $D_{ij} = \kappa_{ii}\kappa_{jj} - \kappa_{ij}^2$. Thus, from element (1,3) of $P_{SS} = G_{SS}H_{SS}$ or from equation (12), we get the required correction of $\beta_{1|3.24}$ to recover $\gamma = -h_{13}$ as

$$\gamma = \beta_{1|3.24} + \kappa_{14}\kappa_{34}/D_{34},$$

while it is seen from element (1,2) of P_{SS} that $\beta_{1|2.34}$ is an unconfounded measure of α. Since Figure 3b) implies $\beta_{2|3.4} = 0$, it follows, in addition, from equation (7) that $\beta_{1|3.24} = \beta_{1|3.4}$.

Illustration 2. As a second illustration we use in Figure 4 an example due to Robins and Wasserman (1997).

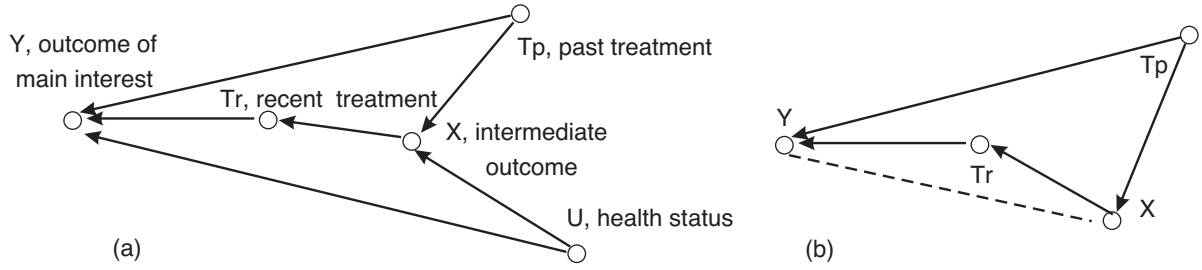


Figure 4: (a) Generating directed graph of *Illustration* 2 in four observed variables and U. (b) Recursive regression graph in observed variables, derived from graph (a) by marginalizing over U; indirect confounding, present in the least squares regression coefficient in observed variables measuring dependence of Y on T_{p} given T_r and X can be corrected to recover the generating dependence γ.

The authors introduced it to show that the coefficient of dependence of the main outcome variable, Y, on past treatment, T_{p}, given a more recent treatment, T_{r}, and the health status

of a patient, U, cannot be consistently estimated by any least squares regression coefficient in observed variables, that is for U unobserved, in spite of using randomization when administering the two treatments.

The past treatment $T_{\rm p}$ is decoupled from U due to full randomized allocation of treatments to patients, there is an intermediate outcome, X. The recent treatment $T_{\rm r}$ is decoupled from $T_{\rm p}$ and U since allocation of treatments to patients is randomized conditional on the level of intermediate outcome X. The purpose is to estimate treatment effects as present in the data generating process, i.e. given the health status U.

For $(Y, T_{\rm r}, X, T_{\rm p}) = (1, 2, 3, 4)$ and $i = 1$, we have $\mathrm{anc}(i) = (2, 3, 4)$, the coefficient matrix H_{SS} for the recursive regressions to graph 2b) and the triangular decomposition matrix G_{SS} of κ_{SS}^{-1} are

$$G_{SS} = \begin{pmatrix} 1 & 0 & -\kappa_{13}/D_3 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad H_{SS} = \begin{pmatrix} 1 & -\alpha & 0 & -\gamma \\ 0 & 1 & -\delta & 0 \\ 0 & 0 & 1 & -\theta \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

so that from element (1,4) of $G_{SS}H_{SS}$ we get the required correction to obtain $\gamma = -h_{14}$ as

$$\gamma = \beta_{1|4.23} + \kappa_{13}\theta/D_3.$$

Estimates are obtained by least squares regressions, using equations (13).

3.4 Matrix criteria for paths of indirect confounding

The key relation (12) for the amount of indirect confounding in linear systems without direct confounding specifies the vector of regression coefficients $\Pi_{i|a}$ in a univariate linear least squares regression of Y_i on its directly and indirectly explanatory variables Y_a. A corresponding edge matrix $\mathcal{P}_{i|a}$ can be viewed as the part of a univariate regression graph, connecting offspring node i to its parent and forefather nodes.

From the edge matrix results in Wermuth and Cox (2004) we denote by $\mathrm{clos}(\mathcal{K}_{aa})$, the edge matrix of an undirected graph, obtained by closing each path in the graph with edge matrix $\mathcal{K}_{aa}$ by an edge, and get for $-\Pi_{i|a} = h_{ia} - \kappa_{ia}\kappa_{aa}^{-1}H_{aa}$ the corresponding edge matrix as

$$\mathcal{P}_{i|a} = \mathrm{In}[\mathcal{H}_{ia} + \mathcal{K}_{ia}\,\mathrm{clos}(\mathcal{K}_{aa})\mathcal{H}_{aa}]. \tag{17}$$

Here, $\mathcal{H}_{ia}$ and $\mathcal{H}_{aa}$ point to arrows present in the parent graph as well as in the recursive regression graph, i.e. to generating dependencies. The added matrix product can be translated into paths of indirect confounding. The first two components, $\mathcal{K}_{ia}\,\mathrm{clos}(\mathcal{K}_{aa})$ give an undirected path from a node in a to i in the graph with edge matrix $\mathcal{K}$ which generates a nonzero regression coefficient when regressing residual η_i on η_a. The edge matrix $\mathcal{H}_{aa}$ points to arrows present

in the recursive regression graph. Then for j in a, the matrix product are seen to define the following two types of path

$$i - - - a - - - a \ldots a - - - j, \qquad i - - - a - - - a \ldots a - - - a \prec\!\!\!\longleftarrow j,$$

which are the two types of paths derived before in Section 3.1 using the separation criterion.

Since these paths apply to all linear systems generated over a given parent graph and leading to the recursive regression graphs of equations (10), they also apply to distributions of arbitrary form when generated over the same graph (see Wermuth and Cox, 2004; Theorem 4). Such paths are association inducing only if the bivariate densities along the path belong to, what are called, complete families of distributions.

4 Discussion

Special cases of what we call indirect confounding of least squares coefficient have been recognized early, see Haavelmo (1943), van de Geer (1971), and Robins and Wasserman (1997). In a more recent discussion of confounding by Greenland, Robins and Pearl (1999) this is not considered in detail, its existence being, however, mentioned in the conclusion. The results in this paper provide general path criteria to decide on the presence of confounding.

The result in this paper relate to the ancestral graphs and models of Richardson and Spirtes (2002), in that they provide recursive regression graphs, which are independence equivalent to ancestral graphs under appropriate conditions, as well as the means of adjusting linear least squares regression coefficients for indirect confounding.

Since identification is another necessary but not sufficient condition for parameter equivalence, our results on parameter equivalence supplement also graphical conditions for the identification of linear recursive equations with correlated residuals; see Brito and Pearl (2002) and Stanghellini and Wermuth (2005).

The recursive regression graphs to equations (10) are special graphs obtainable by marginalizing and conditioning. Corresponding computational tools to construct induced graphs have been made available as open source software for the R Project by Marchetti (2005), see also Marchetti and Drton (2003).

The graphical criteria developed here for detecting indirect confounding apply to distributions of any form generated over a given parent graph. However for other than linear relations, corrections of observed dependencies, needed to recover the generating dependencies, still have to be derived. First results in this direction involve generalizations of Cochran's recursion relation for regression coefficients (Cox and Wermuth, 2003).

Acknowledgement

We are grateful to Giovanni Marchetti for most insightful and constructive comments; the referees and Ayesha Ali for their helpful comments to improve the presentation of the results. We thank the Swedish research society for supporting our cooperation.

References

Brito, C. & Pearl, J. (2002). A new identification condition for recursive models with correlated errors *Structural equation modeling* **9**, 459-474.

Cochran, W.G. (1938). The omission or addition of an independent variate in multiple linear regression. *Suppl. J. R. Statist. Soc.* **5**, 171–176.

Cox, D.R. (1992). Causality: some statistical aspects. *J.R. Statist. Soc.* A **155**, 291-301.

Cox, D.R. & Wermuth, N. (2001). Causal inference and statistical fallacies. *International Encyclopedia of the Social and Behavioral Sciences* P.B. Baltes and N.J. Smelser (eds). Elsevier, Amsterdam, **3**, 1554-1661.

Cox, D.R. & Wermuth, N. (2003). A general condition for avoiding effect reversal after marginalization. *J. R. Statist. Soc. B.* **56**, 934-940.

Cox, D.R. & Wermuth, N. (2004). Causality: a statistical view. *Intern. Statist. Review*, **72**, 285-305.

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton, N.J.: Princeton University Press.

Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis.* Reading: Addison-Wesley.

Edwards, D. (2000). Introduction to Graphical Modelling. 2nd ed. Springer, New York.

Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Transactions Roy. Soc. A*, **222**, 309-368.

Goldberger, A. S. (1964). *Econometric Theory.* New York: Wiley.

Gram, J. P. (1883). Über die Entwickling reeller Funktionen in Reihen mittelst der Methode der kleinsten Quadrate. *Journal für die reine und angewandte Mathematik*, **94**, 41-73.

Greenland, S., Robins, J. M. and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, **46**, 29-46.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12.

Hardt, J., Petrak, F., Filipas, D. and Egle, U.T. (2004). Adaption to life after surgical removal of the bladder – an application of graphical Markov models for analysing longitudinal datat. *Statistics in Medicine* **23**, 649-666.

Kiiveri, H. T., Speed, T. P. and Carlin, J. B. (1984). Recursive causal models. *J. Austral. Math. Soc. A* **36**, 30-52.

Koster, J.T.A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scand. J. Statist.*, **26**, 413–431.

Lindley, D.V. (2002). Seeing and doing: the concept of causation. *Int. Statist. Rev.* **70**, 191-214.

Marchetti, G. M. (2005). R functions for computing graphs induced from a DAG after marginalization and conditioning. Submitted.

Marchetti, G. M. & Drton, M. (2003). GGM: an R package for Gaussian graphical models. URL:http://cran.r-project.org.

Richardson, T.S. & Spirtes, P. (2002). Ancestral Markov graphical models. *Ann. Statist.* **30**, 962–1030.

Robins, J. & Wasserman, L. (1997). Estimation of effects of sequential treatments by reparametrizing directed acyclic graphs. In: D. Geiger and O. Shenoy (eds.) *Proccedings, 13th Annual Conference on Uncertainty in Artificial Intelligence.* 409-420. San Francisco: Morgan and Kaufmann.

Rosenbaum, P.R. (2002). *Observational studies.* Second ed. New York: Springer.

Sargan, J.D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* **26**, 393–415.

Schmidt, E. (1907). Entwicklungen willkürlicher Funktionen. *Mathematische Annalen*, **63**, 433-476.

Smith, J.Q. (1989). Influence diagrams for statistical modelling. *Ann. Statist.* **17**, 654–672.

Stanghellini, E. & Wermuth, N. (2005). On the identification of path analysis models with one hidden variable. *Biometrika* **92**. To appear.

Tukey, J. W. (1954). Causation, regression, and path analysis. In: O. Kempthorne, T. A. Bancroft, J. W. Gowen, and J. L. Lush (eds.). *Statistics and Mathematics in Biology*, 35-66. Ames: The Iowa State College Press.

Van de Geer, J.P. (1971). *Introduction to Multivariate Analysis for the Social Sciences.* San Francisco: Freeman.

Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.*, **75**, 963-97.

Wermuth, N. & Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J. Roy. Statist. Soc. B.* **66**, 687-717.

Wermuth, N., Cox, D.R. & Marchetti, G. (2005). Covariance chains. Submitted.

Wold, H. O. (1954). Causality and econometrics, *Econometrica* **22**, 162-177.

Wright, S. (1923). The theory of path coefficients: a reply to Niles' Criticism. *Genetics* **8**, 239-255.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5**, 161-215.

Appendix. Illustration of effect reversal due to indirect confounding

The following numerical example to *Illustration* 1 shows a case of effect reversal for standardized variables. The negative values of the linear least squares coefficients in the generating system to Figure 3a) are in A, where $\Sigma^{-1} = A^T \Delta^{-1} A$ is the inverse of a correlation matrix.

$$A = \begin{pmatrix} 1 & -.30 & -.36 & 0 & -.90 & 0 \\ 0 & 1 & 0 & -.60 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -.90 \\ 0 & 0 & 0 & 1 & .65 & .75 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\text{diag}(\Delta) = (.2685,\ .6400,\ .1900,\ .0150,\ 1,\ 1).$$

The observed variables correspond to rows and columns 1 to 4 of A, variable U to column 5 and variable V to column 6. The matrix H of equation parameters to Figure 3b) is the submatrix of A for the observed variables and for $i = 1$ with $a = \text{anc}(i) = (2, 3, 4)$ we also have $H_{SS} = H$.

The correlation matrix Σ_{NN} to Figure 3b) of the four observed variables and the coefficient matrix F of the triangular decomposition of Σ_{NN}^{-1}, which coincides in the example with P_{SS} for $i = 1$, are

$$\Sigma_{NN} = \begin{pmatrix} 1 & -.1968 & .2385 & -.6480 \\ . & 1 & -.4050 & .6000 \\ . & . & 1 & -.6750 \\ . & . & . & 1 \end{pmatrix}, \quad P_{SS} = \begin{pmatrix} 1 & -.3000 & .3654 & 1.0746 \\ 0 & 1 & .000 & -.6000 \\ 0 & 0 & 1 & .6750 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The diagonal elements of Δ_{SS} are

$$\delta_{11} = .4498,\ \delta_{22} = .6400,\quad \delta_{33} = .5444,\quad \delta_{44} = 1.$$

Nothing peculiar can be detected in the correlation matrix of the observed variables: there are no very high individual correlations and there is no strong multicollinearity.

The covariance matrix of residual covariances, $\kappa = H\Sigma_{NN}H^T$, and the matrix G_{SS} of the triangular decomposition of κ^{-1} are

$$\kappa = \begin{pmatrix} 1.0785 & 0 & 0 & -.5850 \\ . & .6400 & 0 & 0 \\ . & . & 1 & -.6750 \\ . & . & . & 1 \end{pmatrix}, \quad G_{SS} = \begin{pmatrix} 1 & 0 & .7254 & 1.0746 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & .6750 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The generating coefficients of dependence of Y_1 on Y_2 and Y_3 given U are, respectively, $\beta_{1|2.3U} = .3000$ and $\beta_{1|3.2U} = .3600$. The least squares regression coefficient of Y_3, when regressing Y_1 on Y_a, is, from equation (12) or from element (1,3) of $P_{SS} = G_{SS}H_{SS}$, instead $\beta_{1|3.24} = -.3654$, a reversal in sign and similar in strength compared to the generating dependence, whereas $\beta_{1|2.34}$ measures $\beta_{1|2.3U}$ without any confounding.