Biometrika (2005), **92,** 2, *pp.* 337–350

Printed in Great Britain

On the identification of path analysis models with one hidden variable

BY ELENA STANGHELLINI

Dipartimento di Scienze Statistiche, Università di Perugia, Via A. Pascoli 1, C. P. 1315 Succ.1, 06100 Perugia, Italy

elena.stanghellini@stat.unipg.it

AND NANNY WERMUTH

Department of Mathematical Statistics, Chalmers/Gothenborg University of Technology, S-412 96 Göteborg, Sweden

wermuth@math.chalmers.se

SUMMARY

We study criteria for identifiability of path analysis models with one hidden variable. We first derive sufficient criteria for identification of models in which marginalisation is carried out over the hidden variable. The sufficient criteria are based on the structure of the directed acyclic graph associated with the path analysis model and can be derived from the graph. We treat further the identification of models when the hidden variable is conditioned on and establish connections with the extended skew-normal distribution. Finally it is shown that the derived conditions extend the existing graphical criteria for identification.

Some key words: Conditional independence model; Directed acyclic graph; Identification; Latent variable; Linear system; Unobserved confounder.

1. INTRODUCTION

The combination of ideas from the area of graphical models with those from path analysis and, more generally, from structural equation modelling, has led to reinterpretation and enlargement of existing results. Examples are the implementation of a unifying language, based on graphs, for establishing testable implications contained in a model, and therefore permitting a listing of equivalent models (Frydenberg, 1990), and evaluation of the state of identifiability of models with hidden variables (Stanghellini, 1997; Pearl, 1998; Vicard, 2000). In particular, a graphical criterion has been given by Stanghellini (1997) and Vicard (2000) for assessing the identifiability of single-factor models with correlated residuals. The extension to models with more than one factor has been addressed by Giudici & Stanghellini (2001) and Grzebyk et al. (2004). Pearl (1998) treats identifiability of subsets of the parameters of a path analysis model with correlated residuals giving a sufficient condition based on the graph called the back-door criterion.

In this paper we focus on path analysis models with uncorrelated residuals. The simplifying structure of these models can be represented by a directed graph introduced in § 2 as

a parent graph. We derive criteria for the identification of all parameters of the model when one variable is hidden. We treat two types of hidden variable: either it is a variable over which we marginalise or it is a variable on which we condition. There are various notions of identification in the literature. Here we shall refer to the notion of global identification of a model (Rothenberg, 1971; Bowden, 1973). Models with one unobserved variable arise for instance in contexts with one variable measured with error, such as air pollution, or with an unmeasured confounder. For example the knowledge that exposure to asbestos causes leukaemia is only recent and therefore older studies of potential causes of leukaemia did not include a measure of this exposure. When such data are reanalysed this unmeasured variable should be taken into account.

2. KNOWN RESULTS

2·1. *Covariance and concentration graphs*

For the definitions of the various types of graph and for aspects of their interpretation, we refer the reader to Cox & Wermuth (1996) or Edwards (2000). Here only the notions necessary for our results will be restated. Let $Y = \{Y_1, \ldots, Y_k\}$ be a vector of random variables. A covariance graph $G^V_{\mathrm{cov}} = (V, E^V_{\mathrm{cov}})$ for linear relationships is the pair of a set V of nodes associated with Y and a set E^V_{cov} of undirected edges such that there is no edge joining nodes j and i whenever Y_j and Y_i are marginally uncorrelated. Edges in a covariance graph are represented here by dashed lines. A concentration graph $G^V_{\mathrm{con}} = (V, E^V_{\mathrm{con}})$ for linear relationships is the pair of a set V of nodes associated with Y and a set E^V_{con} of undirected edges such that there is edge joining nodes j and i whenever Y_j and Y_i are uncorrelated given all other variables. Edges in a concentration graph are represented here by full lines.

Let $\Sigma = (\sigma_{ij})$ be the covariance matrix and $\Sigma^{-1} = (\sigma^{ij})$ the concentration matrix of Y. The following results hold (Wermuth, 1976):

$$\sigma_{ij} = \rho_{ij}\sqrt{(\sigma_{ii}\sigma_{jj})}, \quad \sigma^{ij} = -\rho_{ij.V\setminus\{i,j\}}\sqrt{(\sigma^{ii}\sigma^{jj})},$$

where ρ_{ij} is the correlation coefficient between Y_i and Y_j and $\rho_{ij.V\setminus\{i,j\}}$ is the partial correlation coefficient between Y_i and Y_j given all other variables. It follows that, if Y has a joint Gaussian distribution, then

$$\sigma_{ij} = 0 \text{ if and only if } Y_i \perp\!\!\!\perp Y_j,\ \sigma^{ij} = 0 \text{ if and only if } Y_i \perp\!\!\!\perp Y_j | Y_{V\setminus\{i,j\}},$$

in which $Y_a \perp\!\!\!\perp Y_b | Y_c$ is the notation for Y_a and Y_b to be conditionally independent given Y_c (Dawid, 1979). In this paper, the probabilistic independence interpretation applies only to Gaussian distributions; missing edges in graphs mean linear independencies otherwise.

We partition $Y = \{Y_a, Y_b, Y_c\}$ and the node set $V = \{a, b, c\}$ is partitioned accordingly. In the following we indicate by M_{ab} the submatrix $(M)_{a,b}$ of a matrix M and by M^{ab} the submatrix $(M^{-1})_{a,b}$ of its inverse. The covariance matrix Σ and the concentration matrix Σ^{-1} of Y are then written as

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} \Sigma^{aa} & \Sigma^{ab} & \Sigma^{ac} \\ \Sigma^{ba} & \Sigma^{bb} & \Sigma^{bc} \\ \Sigma^{ca} & \Sigma^{cb} & \Sigma^{cc} \end{pmatrix}. \tag{1}$$

We will make use of the following well-known results for the inverse of partitioned matrices, using $d = \{a, b\}$:

$$\Sigma_{dd} = (\Sigma^{dd})^{-1} + \Sigma_{dc}\Sigma_{cc}^{-1}\Sigma_{cd}, \tag{2}$$

$$\Sigma_{dd}^{-1} = \Sigma^{dd} - \Sigma^{dc}(\Sigma^{cc})^{-1}\Sigma^{cd}, \tag{3}$$

$$\Sigma_{dc}\Sigma_{cc}^{-1} = -(\Sigma^{dd})^{-1}\Sigma^{dc}. \tag{4}$$

As Dempster (1969, p. 58) noted, $\Sigma^{dd} = \Sigma_{dd.c}^{-1}$, where $\Sigma_{dd.c}$ is the covariance matrix of Y_d given Y_c, that is after conditioning on Y_c. Analogously, $\Sigma^{cc} = \Sigma_{cc.d}^{-1}$. It follows that, if $\Sigma_{bc} = 0$, then $\Sigma_{bb.c} = \Sigma_{bb}$ and the covariance matrix of Y_a after conditioning on Y_b and Y_c is

$$\Sigma_{aa.bc} = \Sigma_{aa.b} - \Sigma_{ac}\Sigma_{cc}^{-1}\Sigma_{ca}. \tag{5}$$

In the rest of the paper, the variance of a single random variable Y_a will be denoted by σ_{aa} and the partial variance of Y_a given Y_b will be denoted by $\sigma_{aa.b}$. The following lemma can be stated.

Lemma 1. *Let Y be a vector of random variables with covariance matrix Σ and concentration matrix Σ^{-1}. We partition $Y = \{Y_d, Y_c\}$, with c being a single random variable, and Σ and Σ^{-1} accordingly. Let $\alpha = \Sigma^{dc}\sqrt{\sigma_{cc.d}}$. Then*

$$\sigma_{cc.d} = \sigma_{cc}(1 + \alpha^{\mathrm{T}}\Sigma_{dd}\alpha)^{-1}.$$

Proof. From (2) and (4) we find that

$$\sigma_{cc.d} = \sigma_{cc} - \sigma_{cc.d}\Sigma^{cd}\Sigma_{dd}\Sigma^{dc}\sigma_{cc.d} = \sigma_{cc} - \sigma_{cc.d}\alpha^{\mathrm{T}}\Sigma_{dd}\alpha$$

and the result follows. □

2·2. *Univariate generating process and graphs*

A univariate generating process determines a full ordering of the variables in Y such that each variable in the ordering is potentially a response variable for the preceding ones and an explanatory variable for the following ones. The joint density of the variables in Y can then be factorised accordingly into k univariate densities:

$$f_{1,\ldots,k}(y) = f_k(y_k)\prod_{i=1}^{k-1} f_i(y_i|y_{\mathrm{par}(i)}), \tag{6}$$

where $\mathrm{par}(i)$ is the subset of $\{i+1, \ldots, k\}$ containing the variables that Y_i still depends on, given all other preceding variables. These could be regarded as 'direct influences'. Elements of $\mathrm{par}(i)$ are called the parents of node i. A univariate generating process is represented by the distribution generated over a graph $G_{\mathrm{par}}^V = (V, E_{\mathrm{par}}^V)$, where V is the set of vertices or nodes corresponding to the variables $Y_1, \ldots, Y_k$ and E_{par}^V is the set of directed edges drawn as arrows pointing from j to i whenever $j \in \mathrm{par}(i)$. As a result of this structure, we call this graph the parent graph. The set of nodes with a directed edge originating from j contains the children of j and is denoted by $\mathrm{chl}(j)$. The defining independence structure in (6) can be equivalently formulated in a condensed notation

of nodes:

$$\{i \perp\!\!\!\perp \text{potential ancestor of } i \text{ excluding par}(i) | \text{par}(i)\}$$

for all $i \in V$.

In Fig. 1 three possible parent graphs involving three nodes are presented. In Fig. 1(a) node t acts as a transition node, whereas in Fig. 1(b) node t acts as a source node and in Fig. 1(c) node t acts as a collision node.

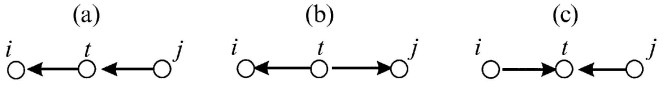

Fig. 1. Three parent graphs with node t as (a) a transition node, (b) a source node and (c) a collision node.

Given a graph G, a path of length n is a succession of $n > 1$ edges connecting nodes $i_0, \ldots, i_n$ irrespective of the orientation of the edges. A cycle is a path in which i_0 and i_n coincide. An odd cycle is a cycle involving an odd number of nodes. For a subset $S \subseteq V$ we define the boundary of S in a given graph G, bd(S, G), as the set of all nodes connected by an edge with a node in S. A subgraph G_S induced by a subset $S \subseteq V$ in G consists of the nodes S and the edges having both endpoints in S. A graph or a subgraph is connected if every two nodes are connected by a path. If all edges are present a graph is called complete. We define the complementary graph of a graph G as the graph $\bar{G}$ with the same set of nodes, and an undirected edge connecting i and j whenever an ij-edge is missing in G. By connectivity component we mean a maximal connected subgraph. A node i is called a descendant of j in a G^V_{par} if there is a direction-preserving path from j to i. In this case j is called an ancestor of i. A parent graph is such that a node cannot be ancestor of itself. For this reason a parent graph is also called a directed acyclic graph.

Let S and C be two subsets of V with $S \cap C = \varnothing$. We will make use of the undirected graphs $G^{S|C}_{\text{con}}$ and $G^{S|C}_{\text{cov}}$ induced by a parent graph. The former shows the independencies of variable-pairs in S induced after conditioning on C and all remaining variables in S, and the latter shows the marginal pairwise independencies of variable-pairs in S induced after conditioning on C. More precisely, the graph $G^{S|C}_{\text{con}}$ has S as a set of nodes and E_S as a set of edges such that the undirected ij-edge is not in E_S whenever $i \perp\!\!\!\perp j | C \cup S \setminus \{i, j\}$ follows from (6). Similarly, the graph $G^{S|C}_{\text{cov}}$ has S as a set of nodes and E_S as a set of edges such that the undirected ij-edge is not in E whenever $i \perp\!\!\!\perp j | C$ follows from (6).

The above independencies may be derived by combining directly probability statements (Dawid, 1979) or by using a separation criterion for directed acyclic graphs (Pearl, 1988, p. 117). One formulation of the criterion (Wermuth & Cox, 1998) is that i and j are independent when conditioning on $C \subseteq V \setminus \{i, j\}$ if, along every path from i to j, either there is a source node or a transition node in C or outside C there is a collision node together with all its descendants.

A direct implication of the above criterion is that, when C is the empty set, the overall concentration graph induced by a parent graph, G^V_{con}, has an undirected, full-line, ij-edge whenever there is an ij-arrow in the parent graph or i and j have a common child. Analogously, the overall covariance graph induced by a parent graph, G^V_{cov}, has an undirected, dashed-line, ij-edge if and only if there is a path connecting i to j which does not contain a collision node. Figure 2 shows a parent graph together with its overall induced concentration and covariance graphs.

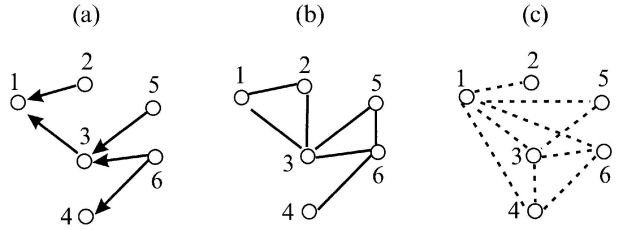

Fig. 2. (a) A parent graph G^V_{par}, (b) the overall induced concentration graph G^V_{con} and (c) the overall induced covariance graph G^V_{cov}.

2·3. *Identification of a single-factor model*

Let Y be a vector of k mean-centred Gaussian random variables. We partition Y into Y_O, a set of observable variables, and Y_L a single unobserved random variable. A single-factor model is constructed as follows:

$$Y_O = \lambda Y_L + \eta, \tag{7}$$

where λ is the vector of so-called factor loadings and η is a vector of residuals such that $E(\eta Y_L) = 0$. It then follows that the covariance matrix implied by a single-factor model on the observable variables is

$$\Sigma_{OO} = \lambda\lambda^{\mathrm{T}}\sigma_{LL} + \Sigma_{OO.L}, \tag{8}$$

in which $\Sigma_{OO.L}$ is the covariance matrix of η.

If system (8) admits a unique solution or a finite number of solutions in $\lambda\lambda^{\mathrm{T}}\sigma_{LL}$ and $\Sigma_{OO.L}$ then the single-factor model is identified. Stanghellini (1997) gave a graphical rule for solving system (8) as arising from a single-factor model with correlated residuals. The sufficient rule is based on the structure of zeros in $\Sigma^{-1}_{OO.L}$ when $\lambda \neq 0$. This rule was later proved to be necessary by Vicard (2000). The derivation hinges on equation (3), that is the fact that the concentration matrix of the observable variables Σ^{-1}_{OO} has a structure similar to that of the covariance matrix in (8); in fact,

$$\Sigma^{-1}_{OO} = -\delta\delta^{\mathrm{T}} + \Sigma^{-1}_{OO.L}, \tag{9}$$

where $\delta = \Sigma^{OL}\sqrt{\sigma_{LL.O}}$. The following lemma is a direct consequence.

Lemma 2. *System* (8) *can be solved with respect to* $\lambda\lambda^{\mathrm{T}}\sigma_{LL}$ *and* $\Sigma_{OO.L}$ *if and only if one of the following conditions holds*:

(i) $\lambda \neq 0$ *and the structure of zeros in* $\Sigma_{OO.L}$ *is such that every connectivity component of the complementary graph of* $G^{O|L}_{\text{cov}}$ *contains an odd cycle*;

(ii) $\delta \neq 0$ *and the structure of zeros in* $\Sigma^{-1}_{OO.L}$ *is such that every connectivity component of the complementary graph of* $G^{O|L}_{\text{con}}$ *contains an odd cycle.*

Proof. If (i) holds then, from a parallel argument to that of Stanghellini (1997) applied to the covariance matrix $\Sigma_{OO.L}$, system (8) can be solved. If (ii) holds then, from Stanghellini (1997), system (9) can be solved for $\Sigma^{-1}_{OO.L}$. If we invert $\Sigma^{-1}_{OO.L}$, system (8) can be solved as well. The necessity follows if we note that if none of the two conditions holds then system (8) has infinitely many solutions or no solution (Vicard, 2000). □

Note that only $\lambda\lambda^{\mathrm{T}}\sigma_{LL}$ is uniquely identified. This implies that λ is only identified up to the sign and the positive constant $\sqrt{\sigma_{LL}}$. Usually, in factor analysis modelling the second problem is solved via the assumption that $\sigma_{LL} = 1$.

Note that an ij-edge in the complementary graph of $G^{O|L}_{\mathrm{cov}}$ implies a zero entry in $\Sigma_{OO.L}$ and thus, from (8), an ij-entry of Σ_{OO} that is equal to $\lambda_i\lambda_j$. When the complementary graph contains a three-cycle, then there is a triple of observed variables, i, j and k say, such that

$$\frac{\sigma_{ik}\sigma_{jk}}{\sigma_{ij}} = \frac{\lambda_i\lambda_k\lambda_j\lambda_k}{\lambda_i\lambda_j} = \lambda_k^2,$$

leading to the identification of λ_k^2. When the cycle involves an odd number greater than three, an extended version of the above relationship is implied. Analogously, an ij-edge in the complementary graph of $G^{O|L}_{\mathrm{con}}$ defines an ij-entry of Σ^{-1}_{OO} that is equal to $\delta_i\delta_j$.

3. PATH ANALYSIS AND RELATED GRAPHS

In this paper we will assume Y to be a vector of k mean-centred random variables such that

$$AY = \varepsilon, \tag{10}$$

where $A = (-\alpha_{ij})$ is an upper triangular matrix with ones along the diagonal and the errors ε have zero means and are uncorrelated so that $\mathrm{cov}(\varepsilon) = \Delta$ is a diagonal matrix. The linear system (10), with some elements α_{ij} restricted to be zero, is known as path analysis, from the work of Wright (1923, 1934).

From (10) the covariance matrix Σ and the concentration matrix Σ^{-1} of Y are

$$\Sigma = B\Delta B^{\mathrm{T}}, \quad \Sigma^{-1} = A^{\mathrm{T}}\Delta^{-1}A, \tag{11}$$

where $B = A^{-1}$. Therefore, given A and Δ the matrix Σ, or equivalently Σ^{-1}, is uniquely determined. When the full ordering of the variables is given the converse also holds.

In what follows, associated with (10) is a parent graph, $G^{V}_{\mathrm{par}} = (V, E_{\mathrm{par}})$, such that Y_i corresponds to node i and an arrow points from j to i whenever α_{ij} is a nonzero coefficient. If all the α_{ij} in (10) are different from zero the model is saturated and the corresponding parent graph is complete. The parent graph so constructed is not different from the usual path analysis diagram introduced by Wright (1923, 1934), and it coincides with the parent graph of § 2·2 in the case of a Gaussian distribution for Y.

Let S and C be two subsets of V with $S \cap C = \varnothing$. Then we will also make use of the covariance and concentration graphs $G^{S|C}_{\mathrm{cov}}$ and $G^{S|C}_{\mathrm{con}}$, induced by (10), such that $G^{S|C}_{\mathrm{cov}}$ has a missing ij-edge whenever $\rho_{ij.C} = 0$ is implied by (10). Analogously, $G^{S|C}_{\mathrm{con}}$ has a missing ij-edge whenever $\rho_{ij.C\cup S\setminus\{i,j\}} = 0$ is implied by (10). Such graphs may be constructed using a separation criterion for directed graphs, as stated in § 2·2.

A univariate process for generating densities as in (6) or equations as in (10) may contain latent variables, where latent means hidden or unobserved. Such variables act either as variables that are marginalised over or as variables defining a selected subpopulation by fixing some of their levels; the latter are the variables conditioned upon. The corresponding parent graph then contains hidden nodes. Questions about what can be learnt from the distribution of the observable variables about the joint distribution specified by the parent graph should therefore be addressed. This problem is closely related to the problem of identifiability.

We partition $Y = \{Y_O, Y_L\}$. When Y_L is marginalised over, the observable variables are Y_O. When Y_L acts as a conditioning node, we distinguish between two types of conditioning. In the first, which we call conditioning on a point, the observable variables are Y_O given

$Y_L = y_L$. If Y has a Gaussian joint distribution then the same is true of the observable variables. We refer to the second type as conditioning on an open interval. In this situation the observable variables are Y_O given $Y_L \geqslant b$. Then, if Y has a Gaussian distribution the observable variables have an extended skew-normal joint distribution (Capitanio et al., 2003).

Model (10) is globally identified if the elements of A and Δ can be uniquely reconstructed from the parameters of the joint distribution of the observable variables. If we denote by θ_{obs} the vector of the parameters of the distribution of the observable variables, model (10) is globally identified if θ_{obs} has a unique solution, or at most a finite number of solutions, in A and Δ. A common assumption of latent variable models, which we will adopt here, is that $E(Y_L) = 0$. This constraint is required in order to solve the non-identifiability problem of the expected value of the latent variable.

Global identification of (10), when all variables are observed, has been established by Wold (1960); see also Goldberger (1964, p. 383). The problem of identification of structural equation models with latent variables has also been studied in the econometric literature, see Bollen (1989) for a review, but general results are not yet available.

Here we develop criteria based on the properties of the graph for assessing whether or not a path analysis model with one latent variable is identified. Nodes over which we marginalise are denoted in the graph by a double crossing over the nodes. Nodes on which we condition are put in a square. In Fig. 3 two parent graphs are presented, both of which we later prove to be identified. In Fig. 3(a) Y_4 acts as a node over which we marginalise, while in Fig. 3(b) Y_2 acts as a node on which we condition.

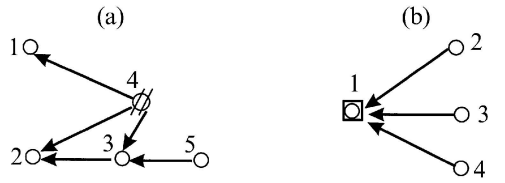

Fig. 3. Two parent graphs with (a) marginalisation over node 4 and (b) conditioning on node 1.

To derive our sufficient criteria we now define a particular class of graphs.

DEFINITION. *An undirected graph G is G-identifiable if every connectivity component of the complementary graph* $\bar{G}$ *contains an odd cycle.*

In the following sections we apply marginalisation or conditioning on the observable variables in order to reduce the identification problem to the solution of systems of equations such as (8) and (9).

4. IDENTIFICATION WHEN MARGINALISING OVER A NODE

For Y_L not observed and marginalised over, the relevant parameters of the observable variables are $\theta_{\text{obs}} = \Sigma_{OO}$. The covariance matrix implied by the model coincides with (8), where $\lambda = \Sigma_{OL}/\sigma_{LL}$ and $\Sigma_{OO.L}$ is the possibly non-diagonal covariance matrix of Y_O given Y_L. Also, the implied concentration matrix coincides with (9), where $\delta = \Sigma^{OL}\sqrt{\sigma_{LL.O}}$ and $\Sigma^{-1}_{OO.L}$ is the possibly non-diagonal concentration matrix of Y_O given Y_L.

The relationship between λ and δ can be made explicit as a function of Σ_{OO} by the use of (4) and the definition of δ:

$$\lambda = -\Sigma_{OO}\delta\sqrt{(\sigma_{LL.O})}/\sigma_{LL}. \tag{12}$$

Note that equations (8) and (9) now arise from model (10). While the single-factor model is a model of the conditional distribution of the observed variables given the unobserved one, and does not imply any other ordering, model (10) implies a full ordering of the variables. Therefore, in contrast to the single-factor model, in which $\lambda \neq 0$ is an assumption of the model, an i-element of λ equals zero whenever model (10) implies that element Y_i is marginally independent of Y_L. Therefore $\lambda \neq 0$ or, analogously, $\delta \neq 0$ is not implied by the model. For that reason, the previous results cannot be applied directly to the present situation. On the other hand, model (10) could imply a relationship between different elements of λ and δ that, if taken into account, may enlarge the class of identified models; see the Proof of Theorem 2. Note also that, if λ and σ_{LL} are identified, then, from (8), $\Sigma_{OO.L}$ can be uniquely reconstructed as the difference between Σ_{OO} and $\lambda\lambda^{\mathrm{T}}\sigma_{LL}$, so that the model is identified. Therefore, the problem reduces to finding conditions under which λ and σ_{LL} are identified. As we shall see, the conditions in this paragraph lead us to identify uniquely $\lambda\lambda^{\mathrm{T}}\sigma_{LL}$ and therefore $\Sigma_{OO.L}$, but, as in the single-factor model, λ is only identified up to the sign and the positive constant $\sqrt{\sigma_{LL}}$. Note also that, if δ is known up to the sign, then, from (12), the same is true of λ.

In what follows we define $m = \mathrm{bd}(L, G^V_{\mathrm{cov}})$, $c = \mathrm{bd}(L, G^V_{\mathrm{con}})$ and $a = \mathrm{bd}(L, G^V_{\mathrm{par}})$ as subsets of the observed nodes O which need not coincide. Note that $a = \{\mathrm{par}(L) \cup \mathrm{chl}(L)\}$ in the parent graph. Recall that the set V is partitioned into $V = \{O, L\}$. Moreover, whenever the set O is partitioned further, then so are the matrices Σ and Σ^{-1}. In the following theorem we assume that no parametric cancellation occurs; that is, each iL-edge present in G^V_{cov} and in G^V_{con}, respectively, corresponds to nonvanishing λ_i and δ_i.

THEOREM 1. *Let $Y = \{Y_O, Y_L\}$ with marginalisation over Y_L and $\sigma_{LL} = 1$. Then a path analysis model* (10) *is identified if one of the following conditions holds*:

(i) *the boundary of the latent variable, m, in the covariance graph G^V_{cov} contains at least three nodes and $G^{m|L}_{\mathrm{cov}}$, the subgraph induced by m in $G^{O|L}_{\mathrm{cov}}$, is G-identifiable*;

(ii) *the boundary of the latent variable, c, in the concentration graph G^V_{con} contains at least three nodes and $G^{c|V\setminus c}_{\mathrm{con}}$, the subgraph induced by c in G^V_{con}, is G-identifiable.*

Proof. For (i) we partition $O = \{m, O\setminus m\}$ and $\lambda = \{\lambda_m, \lambda_{O\setminus m}\}$ with $\lambda_m \neq 0$ and $\lambda_{O\setminus m} = 0$, so the problem reduces to that of identifying λ_m after imposing $\sigma_{LL} = 1$. Note that

$$\Sigma_{mm} = \sigma_{LL}\lambda_m\lambda_m^{\mathrm{T}} + \Sigma_{mm.L}. \tag{13}$$

From Lemma 2 we see that (i) is sufficient for solving (13) with respect to $\lambda_m\lambda_m^{\mathrm{T}}$ and $\Sigma_{mm.L}$.

For (ii) we partition $O = \{c, O\setminus c\}$ and $\delta = \{\delta_c, \delta_{O\setminus c}\}$ with $\delta_c \neq 0$ and $\delta_{O\setminus c} = 0$. From (9), for $i \in O$ and $j \in O\setminus c$, an (i, j)-element of Σ^{-1}_{OO} is equal to the corresponding (i, j)-element of $\Sigma^{-1}_{OO.L}$. Moreover,

$$(\Sigma^{-1}_{OO})_{c,c} = -\delta_c\delta_c^{\mathrm{T}} + \Sigma^{cc} \tag{14}$$

as $(\Sigma^{-1}_{OO.L})_{c,c}$ is equal to Σ^{cc}. From Lemma 2 we see that (ii) is sufficient for solving (14) with respect to $\delta_c\delta_c^{\mathrm{T}}$ and Σ^{cc}. By imposing $\sigma_{LL} = 1$ and noting that $\Sigma^{O\setminus c, L} = 0$, from Lemma 1 we derive $\sigma_{LL.O}$ and Σ^{cL}. The matrix Σ^{-1} is then identified and λ_m can be derived as the matrix Σ_{mL}. □

The graph in Fig. 4(a) corresponds to an identified model. In this case $L=3$. From Fig. 4(b) we see that $c=\{1, 2, 4, 5\}$. In Fig. 4(c) the subgraph induced by c in G^V_{con} is shown and in Fig. 4(d) the complementary graph of this subgraph is presented. As this graph has just one connectivity component which contains the odd cycle formed by $\{1, 2, 4\}$, condition (ii) of Theorem 1 is satisfied. Note that condition (i) of Theorem 1 is not satisfied. In fact, $m=\{1, 2, 4, 5, 6\}$ and $G^{m|L}_{\text{cov}}$ is not G-identifiable. An instance of a model identified by condition (i) of Theorem 1 is presented in § 7.

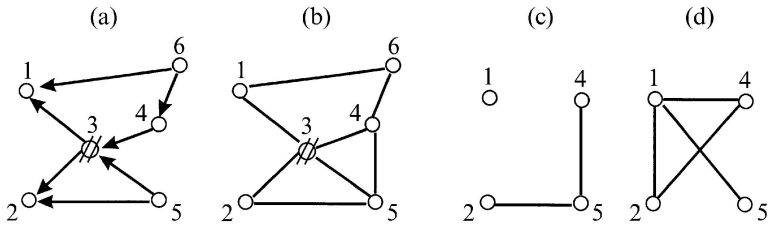

Fig. 4. (a) A parent graph with $L=3$ as a node to be marginalised over, (b) the induced overall concentration graph G^V_{con} with $c=\text{bd}(3)=O\backslash\{6\}$, (c) $G^{c|V\backslash c}_{\text{con}}$, the subgraph induced by c in G^V_{con} and (d) its complementary graph.

The following theorem enlarges the class of identified models, where again $m=\text{bd}(L, G^V_{\text{cov}})$, $c=\text{bd}(L, G^V_{\text{con}})$ and $a=\text{bd}(L, G^V_{\text{par}})$.

THEOREM 2. *Let* $Y=\{Y_O, Y_L\}$ *with marginalisation over* Y_L *and* $\sigma_{LL}=1$. *Then a path analysis model* (10) *is identified if the subset* a *contains at least three nodes and there exists a subset* $h\subseteq\{O\backslash a\}$ *such that* $G^{a|L\cup h}_{\text{cov}}$ *or* $G^{a|L\cup h}_{\text{con}}$ *is G-identifiable and one of the following conditions holds*:

(i) $\Sigma_{hL.a}=0$;

(ii) $\Sigma_{hL}=0$.

Proof. We partition $O=\{a, m\backslash a, O\backslash m\}$, $\lambda=\{\lambda_a, \lambda_{m\backslash a}, \lambda_{O\backslash m}\}$ and note that $\lambda_{O\backslash m}=0$ and $\lambda_{m\backslash a}=K\lambda_a$, in which $K=\Sigma_{m\backslash a,a}\Sigma^{-1}_{aa}$ is a function of elements of Σ_{OO}. The problem therefore reduces to that of identifying λ_a.

For (i) let $J=\{a, h, L\}$. If there exists an h such that Σ_{JJ} or Σ^{-1}_{JJ} can be uniquely reconstructed from Σ_{OO} then the model is identified. We denote by Ω^{aL} the matrix $(\Sigma^{-1}_{JJ})_{a,L}$ and by Ω^{hL} the matrix $(\Sigma^{-1}_{JJ})_{h,L}$. First note that, from (3) and (4),

$$\Sigma^{-1}_{aa.h}=\Sigma^{-1}_{aa.hL}-\beta\beta^{\text{T}}, \tag{15}$$

in which $\beta=\Omega^{aL}\sqrt{\sigma_{LL.ha}}$. Moreover, from (2) and (4) we have

$$\Sigma_{aa.h}=\Sigma_{aa.hL}+\gamma\gamma^{\text{T}}\sigma_{LL.h}, \tag{16}$$

in which $\gamma=\Sigma_{aa.hL}\Omega^{aL}$. Therefore, if $G^{a|L\cup h}_{\text{cov}}$ and $G^{a|L\cup h}_{\text{con}}$ are G-identifiable then, by Lemma 2, $\beta\beta^{\text{T}}$ is identified. As $\sigma_{LL}=1$ and $\Omega^{hL}=0$, Lemma 1 can be applied to elements of Σ^{-1}_{JJ} leading us to identify Ω^{aL}. Thus Σ_{JJ} is also identified and (i) is sufficient.

For (ii) if $\Sigma_{hL}=0$, from (5), expression (16) simplifies to

$$\Sigma_{aa.h}=\lambda_a\lambda^{\text{T}}_a\sigma_{LL}+\Sigma_{aa.Lh}, \tag{17}$$

with the inverse having the same expression as (15). Therefore $\lambda_a\lambda^{\text{T}}_a$ is identified and λ_a is identified up to its sign and (ii) is sufficient. □

Note that conditions (i) and (ii) of Theorem 2 can be checked on the graph by separation criteria. Furthermore, condition (ii) of Theorem 2 implies that $h \perp\!\!\!\perp L$ and can also be expressed as $h \subseteq O \backslash m$. Note that h could be empty, in which case conditions (i) and (ii) are trivially satisfied. The model corresponding to the graph of Fig. 5(a), with $L = 5$, provides an example of a model identified according to Theorem 2 with $h = \varnothing$. In this case, $O = \{1, 2, 3, 4, 6\}$, $a = \{2, 3, 4, 6\}$ and $G^{a|L}_{\text{con}}$ is G-identifiable. The graph in Fig. 6(a) with $L = 4$ corresponds to a model that is identified according to Theorem 2(ii). In this case $O = \{1, 2, 3, 5\}$, $a = \{1, 2, 3\}$ and $c = O$. By choosing $h = \{c \backslash a\} = \{5\}$ we see that $\Sigma_{hL} = 0$ and $G^{a|L \cup h}_{\text{con}}$ is G-identifiable. Note that the graph does not meet any of the conditions of Theorem 1. The graph in Fig. 3(a) with $L = 4$ corresponds to a model identified according to Theorem 1(ii). In this case, $O = \{1, 2, 3, 5\}$, $a = \{1, 2, 3\}$ and $c = O$. The graph $G^{O|L}_{\text{con}}$ is G-identifiable. Note that this graph does not meet any of the conditions of Theorem 2. In fact, $O \backslash a = \{5\}$ and, if we choose $h = \varnothing$, neither $G^{a|L}_{\text{con}}$ nor $G^{a|L}_{\text{cov}}$ is G-identifiable; also, if we choose $h = \{5\}$, then neither $G^{a|h \cup L}_{\text{con}}$ nor $G^{a|h \cup L}_{\text{cov}}$ is G-identifiable.

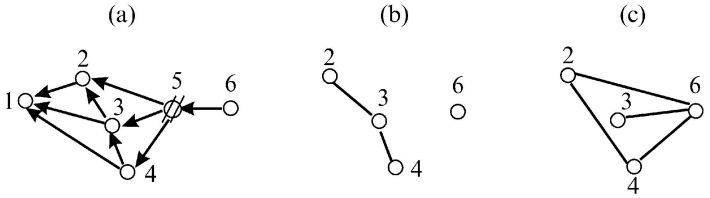

Fig. 5. (a) A parent graph with node $L = 5$ as a node to be marginalised over, (b) the graph $G^{a|L}_{\text{con}}$ and (c) the complementary graph of $G^{a|L}_{\text{con}}$.

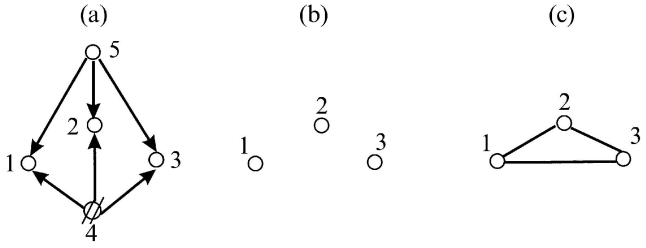

Fig. 6. (a) A parent graph with node $L = 4$ as a node to be marginalised over and $h = \{5\}$ such that $\Sigma_{hL} = 0$, (b) the graph $G^{a|L \cup h}_{\text{con}}$ and (c) the complementary graph of $G^{a|L \cup h}_{\text{con}}$.

5. IDENTIFICATION WHEN CONDITIONING ON A NODE

In this section we establish criteria for identification of path analysis models (10) when conditioning on a single value of the hidden variable, that is $Y_L = y_L$. In this situation the relevant parameters of the observable variables are $\theta_{\text{obs}} = \Sigma_{OO.L}$, that is the covariance matrix of Y_O given Y_L. Note that, from (8) and (9),

$$\Sigma_{OO.L} = \Sigma_{OO} - \sigma_{LL} \lambda \lambda^{\mathrm{T}}, \quad \Sigma^{-1}_{OO.L} = \Sigma^{-1}_{OO} + \delta \delta^{\mathrm{T}}, \tag{18}$$

with λ and δ as previously defined. Again, we impose $\sigma_{LL} = 1$. The relationship between λ and δ can be made explicit as a function of $\Sigma_{OO.L}$; that is, from (4),

$$\lambda = -\Sigma_{OO.L} \delta / \sqrt{\sigma_{LL.O}}. \tag{19}$$

The arguments in this section follow closely those of § 4. Systems in (18) can be solved, depending on the structure of zeros in Σ_{OO} or Σ_{OO}^{-1}. For the following theorem, stated without proof, we exclude parametric cancellation; that is, each *iL*-edge present, in order, in G_{cov}^{V} and in G_{con}^{V} corresponds to nonvanishing λ_i and δ_i.

THEOREM 3. *Let* $Y = \{Y_O, Y_L\}$ *with* Y_L *a conditioning node and* $\sigma_{LL} = 1$. *Then a path analysis model* (10) *is identified if one of the following conditions holds*:

(i) *the boundary of the latent variable*, *m*, *in the covariance graph* G_{cov}^{V} *contains at least three nodes and* G_{cov}^{m}, *the subgraph induced by m in* G_{cov}^{V}, *is G-identifiable*;

(ii) *the boundary of the latent variable*, *c*, *in the concentration graph* G_{con}^{V} *contains at least three nodes and* $G_{\text{con}}^{c|O\setminus c}$, *the subgraph induced by c in* G_{con}^{O}, *is G-identifiable.*

Conditions (i) and (ii) are never met in models in which L has children, as the following Corollary proves.

COROLLARY 1. *Conditions* (i) *and* (ii) *of Theorem* 3 *are never met in models with* $\text{chl}(L) \neq \varnothing$.

Proof. We first consider the case with $O = \{\text{chl}(L) \cup \text{par}(L)\}$. We denote this set by a and assume that it contains at least three nodes; otherwise conditions (i) and (ii) of Theorem 3 are trivially violated. The subgraph induced by $\text{chl}(L)$ in G_{cov}^{m} is complete and every element of $\text{chl}(L)$ is connected to every element of $\text{par}(L)$, so that G_{cov}^{m} is not G-identifiable. Furthermore, in this case $O\setminus c$ is empty, the subgraph induced by $\text{par}(L)$ in G_{con}^{c} is complete and every element in $\text{par}(L)$ is connected to every element of $\text{chl}(L)$, so that G_{con}^{c} is not G-identifiable. We now consider the case with $a \subset O$. The set $m\setminus a$ includes either ancestors of L or descendants of L or both. The subgraph induced by $\text{chl}(L)$ in G_{cov}^{m} is complete and every element of $\text{chl}(L)$ is connected to all the other elements, so that G_{cov}^{m} is not G-identifiable. Furthermore, the set $c\setminus a$ includes $\text{par}\{\text{chl}(L)\}$ and the subgraph induced by $\{\text{par}(L) \cup \text{par}\{\text{chl}(L)\}\}$ in $G_{\text{con}}^{c|O\setminus c}$ is complete and every element in this subset is connected to every element of $\text{chl}(L)$. Therefore $G_{\text{con}}^{c|O\setminus c}$ is not G-identifiable. □

In Fig. 7(a) we present a parent graph with $L = 1$ acting as a conditioning node. From Theorem 3(i) we can see that the associated model is identified. In this case $m = O\setminus\{6\}$ and the complementary graph of G_{cov}^{m} contains one connectivity component formed by two odd cycles. Note that this model also satisfies condition (ii) of Theorem 3. In fact, $c = \{3, 4, 5\}$ and $G_{\text{con}}^{c|O\setminus c}$ is G-identifiable.

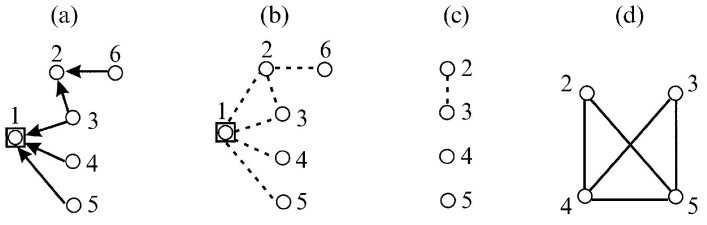

Fig. 7. (a) A parent graph with node $L = 1$ as a conditioning node, (b) the induced overall covariance graph G_{cov}^{V} with $m = \text{bd}(1) = O\setminus\{6\}$, (c) G_{cov}^{m}, the subgraph induced by m in G_{cov}^{V} and (d) the complementary graph of G_{cov}^{m}.

The parent graph in Fig. 3(b) with $L=1$ corresponds to a model that is identified according to Theorem 3(ii), as $c=O$ and the complementary graph of G^O_{con} is G-identifiable. Note that this model also satisfies condition (i) of Theorem 3.

6. Conditioning on a hidden open interval

In this section we consider models in which the joint distribution of the observable variables arises from conditioning on an open interval of a hidden node of the form $Y_L \geqslant b$. Let $Y=\{Y_O, Y_L\}$ be a k-dimensional Gaussian random vector with $E(Y)=\{\xi, 0\}$ and covariance matrix Σ partitioned as in (1) with $\sigma_{LL}=1$. Let $Z=Y_O|Y_L \geqslant -\tau$ be the observable variables. The joint distribution of Z is an extended skew-normal distribution (Capitanio et al., 2003) with parameters α, Σ_{OO}, ξ and τ, and density function

$$f(z)=\phi_{k-1}(z-\xi;\Sigma_{OO})\Phi\{\alpha_0+\alpha^{\mathrm{T}}(z-\xi)\}/\Phi(\tau), \tag{20}$$

where $\phi(.)$ and $\Phi(.)$ denote the standard normal density function and its integral. Note that $\alpha=\Sigma_{OO}^{-1}\Sigma_{OL}/\sqrt{\sigma_{LL.O}}$ and α_0 is a function of the other parameters, as $\alpha_0=\tau(1+\alpha^{\mathrm{T}}\Sigma_{OO}\alpha)^{\frac{1}{2}}$. Also

$$\Sigma_{OL}=\frac{\tau}{\alpha_0}\Sigma_{OO}\alpha$$

and therefore every path analysis model is identified. Note that in this particular situation the sign of Σ_{OL} is identified, as it corresponds to the sign of the parameter α.

7. Some implications

We use the previous results to investigate the identifiability of models for the effect of sequentially administered treatments in randomised clinical trials with an unobserved confounder. Robins & Wasserman (1997) describe the following hypothetical clinical trial in which AIDS patients receive two AZT treatments in sequence. At both times the treatment dose is assigned at random. Randomisation probabilities of the recent dose T_r depend on the previous treatment dose T_p and on an intermediate variable I, a measure of the anaemia of the patients. The overall outcome W is the measure of the HIV-viral load at the end of a follow-up period. There is an unobserved confounder L representing the patient's underlying immune function prior to the treatment. It affects both the intermediate variable I and the outcome variable W. The parent graph under the null hypothesis of no treatment effect is represented in Fig. 8(a). We assume the graph to be a conditional independence graph and to reflect the factorisation of a joint Gaussian distribution as in (6).

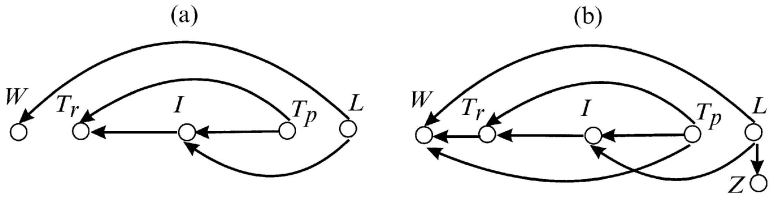

Fig. 8. (a) Parent graph in a randomised trial for sequentially administered treatments showing no treatment effect and (b) the parent graph with both treatment effects and the instrument Z.

By applying the separation criterion of § 2·2 we see that the only independencies implied by the model in the marginal distribution of the observable variables are $W \perp\!\!\!\perp T_p$ and $W \perp\!\!\!\perp T_r | T_p, I$. These independencies can therefore be tested, and acceptance of both hypotheses leads one to conclude that there is no treatment effect. Suppose now that there is a direct effect of the second treatment T_r on W. Then, no independency is implied in the marginal distribution of the observable variables. By application of the criteria derived in this paper, the model under the assumption that both treatments have an effect can be made identifiable and both hypotheses can be independently tested. If we assume that an auxiliary observed variable Z exists which is influenced by the unobserved variable L and that a path analysis system (10) on $V = \{W, T_r, I, T_p, Z, L\}$ holds, then the model under the assumption that both treatments affect the outcome variable, as described in Fig. 8(b), is identified. In fact, condition (ii) of Theorem 1 is satisfied, as $c = \{W, I, T_p, Z\}$ and the subgraph induced by c in G^V_{con} is G-identifiable. Note that Z does not meet the back-door criterion (Pearl, 1998) relative to any of the treatment effects. Moreover, removal of one arrow at the time between T_r and T_p and W also leads to identified models, and the two hypotheses can be independently tested. The model with no treatment effect is identified also according to condition (i) of Theorem 1.

Alternatively, if we may assume that only patients with the underlying immune function above a threshold enter the trial, then the model can be made identified using the results of § 6, as Y_L acts as a conditioning node of the kind $Y_L \geqslant b$. In this second situation all possible models are identified.

8. Discussion

Since all models obtained as a reparametrisation of an identified model are also identified, the criteria shown in this paper can be applied to the class of linear models corresponding to graphs that are independence-equivalent to a parent graph.

To fit the models described in § 4, standard latent variable software can be used. When a joint Gaussian distribution is assumed, maximum likelihood estimation of these models can be performed via the EM algorithm. Routines for estimation and testing the state of identification according to the criteria presented in this paper have been implemented in R by G. M. Marchetti and M. Drton in the package `ggm` obtainable from http://cran.r-project.org/. In this case, standard errors of the relevant parameters can be computed by extending the work of H. T. Kiiveri's 1982 Ph.D. thesis from the University of Western Australia. This has been done in a 2004 Università di Firenze Ph.D. thesis by F. Pennoni. Likelihood factorisation conditions based on graphs have been given for the skew-normal distribution in Capitanio et al. (2003), where algorithms for maximum likelihood estimation of the extended skew-normal distribution have also been presented. Therefore, the models discussed as identified lead to estimation and test problems that now have known solutions.

Acknowledgement

We thank a referee and Giovanni M. Marchetti for their constructive criticism on a previous version of the paper.

REFERENCES

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.
BOWDEN, R. (1973). The theory of parametric identification. *Econometrica* **41**, 1069–74.
CAPITANIO, A., AZZALINI, A. & STANGHELLINI, E. (2003). Graphical models for skew-normal variates. *Scand. J. Statist*. **30**, 129–44.
COX, D. R. & WERMUTH, N. (1996). *Multivariate Dependencies—Models, Analysis and Interpretation*. London: Chapman and Hall.
DAWID, A. P. (1979). Conditional independence in statistical theory (with Discussion). *J. R. Statist. Soc*. B **41**, 1–31.
DEMPSTER, A. P. (1969). *Elements of Continuous Multivariate Analysis*. Reading, MA: Addison-Wesley.
EDWARDS, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. New York: Springer.
FRYDENBERG, M. (1990). The chain graph Markov property. *Scand. J. Statist*. **17**, 333–53.
GIUDICI, P. & STANGHELLINI, E. (2001). Graphical Gaussian factor analysis models. *Psychometrika* **66**, 577–92.
GOLDBERGER, A. A. (1964). *Econometric Theory*. London: John Wiley & Sons.
GRZEBYK, M., WILD, P. & CHOUANIÈRE, D. (2004). On identification of multifactor models with correlated residuals. *Biometrika* **91**, 141–51.
PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.
PEARL, J. (1998). Graph, causality, and structural equation models. *Sociol. Meth. Res*. **27**, 226–84.
ROBINS, J. & WASSERMAN, L. (1997). Estimation of effects of sequential treatments by reparametrizing directed acyclic graphs. In *Proc. 13th Annual Conference on Uncertainty in Artificial Intelligence*, Ed. D. Geiger, O. Shenoy and P. Pundaly, pp. 409–20. San Francisco, CA: Morgan Kaufmann.
ROTHENBERG, T. (1971). Identification in parametric models. *Econometrica* **39**, 577–91.
STANGHELLINI, E. (1997). Identification of a single-factor model 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. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108.
WERMUTH, N. & COX, D. R. (1998). On association models defined over independence graphs. *Bernoulli* **4**, 477–95.
WOLD, H. (1960). A generalization of causal chain models. *Econometrica* **28**, 443–63.
WRIGHT, S. (1923). The theory of path coefficients: a reply to Niles' criticism. *Genetics* **8**, 239–55.
WRIGHT, S. (1934). The method of path coefficients. *Ann. Statist*. **5**, 161–215.

[*Received December* 2002. *Revised September* 2004]