In: (Balakrishnan, N. et al., eds) *Wiley StatsRef: Statistics Reference Online* (2015); to appear.

Graphical Markov models, unifying results and their interpretation

Nanny Wermuth

Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden and of Medical Psychology and Medical Sociology, Gutenberg-University, Mainz, Germany

Abstract Graphical Markov models combine conditional independence constraints with graphical representations of stepwise data generating processes. The models started to be formulated about 40 years ago and vigorous development is ongoing. Longitudinal observational studies as well as intervention studies are best modeled via a subclass called regression graph models and, especially traceable regressions. Regression graphs include two types of undirected graph and directed acyclic graphs in ordered sequences of joint responses. Response components may correspond to discrete or continuous random variables and may depend exclusively on variables which have been generated earlier. These aspects are essential when causal hypothesis are the motivation for the planning of empirical studies.

To turn the graphs into useful tools for tracing developmental pathways and for predicting structure in alternative models, the generated distributions have to mimic some properties of joint Gaussian distributions. Here, relevant results concerning these aspects are spelled out and illustrated by examples. With regression graph models, it becomes feasible, for the first time, to derive structural effects of (1) ignoring some of the variables, of (2) selecting subpopulations via fixed levels of some other variables or of (3) changing the order in which the variables might get generated. Thus, the most important future applications of these models will aim at the best possible integration of knowledge from related studies.

Keywords *Composition property, Conditional dependence, Conditional Independence, Connector transitivity, Directed acyclic graphs, Intersection property, Partial Closure, Partial Inversion, Regression graphs, Singleton transitivity, Traceable regressions, Undirected graphs.*

Some historical remarks and overview

Graphical Markov models provide the most flexible tool for formulating, analyzing, and interpreting relations among many variables. The models combine and generalize three different concepts developed about a century ago: (1) directed graphs, in which variables are represented by nodes, used to study linear processes by which joint distributions may have been generated (Sewell Wright, [119, 120]; [89]), (2) simplification of a joint distribution with the help of conditional independences (Andrei A. Markov, [56]), and (3) specification of associations only for variable pairs which are in some sense strongly related and are turned into nearest neighbors in an undirected graph (Willard Gibbs, [34]; [82]).

First formulations of graphical Markov models started about 40 years ago, [96, 97, 98], [19], [108],, several books with differing emphases have appeared since then, for

instance, [116], [62], [45], [18], [25], [37], [87], [94], [39]. Vigorous development is ongoing. These multivariate statistical models combine the above simple but most powerful notions: data generating processes in sequences of single or of joint responses and conditional independences and dependences captured by graphs. Arguably, the most outstanding feature of these types of models is that many of their implications can be derived using the graphs. Some of this will be outlined and illustrated below.

The generating processes concern no longer only linear relations, as a century ago, but they include, among others, linear regressions, [95], generalized linear models, [58], [2], exponential response models, [38], [7], subclasses of structural equations for longitudinal studies, [41], [9], models for planned interventions such as controlled clinical trials with randomized allocation of individuals to treatments, and models for only virtual interventions, [84], [66], [91]. In particular, response variables may in general be vector variables that contain discrete or continuous variables or both types as components.

We concentrate here on ordered series of regressions for which the responses have as regressors exclusively variables in their past. Throughout, we use the terms regression and conditional distribution interchangeably. The generated distributions are called traceable regressions, [100], when different pathways of development can be traced in a corresponding graph, called their regression graph, [113]. Regression graphs extend graphs for multivariate regression, [17], which are one of four different types of the so-called chain graphs introduced in the literature, [24], [47], [32], [5].

Each such graph may represent a research hypothesis on how data could have been generated, [109] so that we speak of the starting or the '**generating graph**'. When one starts with such a general type of graph, one ordering of the joint responses is taken as fixed and the properties of regression graphs, stated here in Propositions 9 and 10, assure that their graphical structures have an interpretation in terms of probability distributions.

Often the objective is to uncover graphical representations that lead to an understanding of the generating process for appropriately collected data. Then for each such study, the starting point is the available substantive knowledge. It is used to decide on variables that are relevant in a given context and on their ordering into responses, intermediate and explanatory variables. Explanatory variables or regressors may for instance be treatments, intermediate outcomes, risks or variables available at baseline, that is at the start of the study. The last are named context variables since they capture features that are taken as given, of the study or of the study individuals.

Well-fitting graphs are derived by using a combination of information from the study design, from statistical analyses that are used to decide on conditional dependences and independences, from past empirical evidence and from theoretically postulated relations. For detailed analyses in some studies, see [113], [105]; links to further sizeable empirical studies are in an overview, [106].

In the following, we do not discuss fitting- or model search-procedures in detail. Instead, we describe first models and graphs for few variables; especially graphs that are fully directed or that are undirected, since they had been developed first and are now still intensively studied, mainly in the context of Bayesian inference or in computer science. We then proceed to regression graphs and models, to special binary distributions, to a summary and some open problems. The main purpose here is to introduce concepts, especially the interplay between generating processes, graphs, factorizations of densities,

edge matrices and matrix operators to modify graphs. Simple examples illustrate some of the now available, unifying results.

Directed acyclic graphs and three Vs

We start by introducing some terms commonly used for graphs in order to discuss the three key situations for directed graphs. A '**graph**' consists of a node set, $N = \{1, \ldots, d\}$, and one or more edge sets. Nodes are also called vertices. Two distinct nodes are said to be '**coupled**', or to be adjacent, if they are directly linked in the graph. Such a link is named an '**edge**'. A '**simple graph**' has at most one edge for each node pair and has no node linked to itself. A graph is '**complete**' if all its node pairs are coupled.

A sequence of edges connecting distinct nodes is a '**path**'. By convention, the shortest type of path is an edge. A '**directed graph**' has exclusively arrows as edges; it is '**acyclic**' if it is impossible to return to any starting node by following a '**direction-preserving path**' that is a sequence of arrows pointing in the same direction. Directed acyclic graphs are simple graphs and each ij-arrow, $i \prec\!\!\!- j$, points from a regressor node j to its response node i; or are said to point from a parent j to its child i. We shorten the name 'subgraph induced by a set of nodes', to '**subgraph of nodes**', which just keeps those nodes and the edges present among them in a given graph.

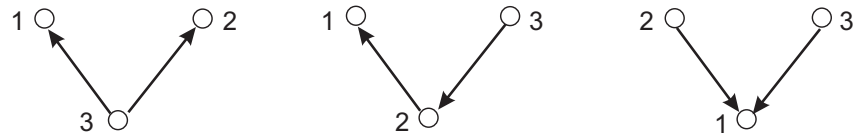


Figure 1: *The three types of V in directed acyclic graphs; left: source V, middle: transition V, right: sink V, called in the literature also a collision V, an unshielded collider or unmarried parents having a common child.*

Fig.1 shows the possible three types of V in directed acyclic graphs. A subgraph of three nodes is called '**a V**' if it has two edges. In each V, there are two '**outer nodes**' that are both coupled to one common neighbor, the '**inner node**' of the V. The name of a V stems from its type of inner node. In a self-explanatory way, the Vs in Fig.1 are called, a '**source V**' on the left, a '**transition V**' in the middle and a '**sink V**' on the right.The notion of inner nodes extends to ij-paths.

For just three variables and in a condensed notation for the generated probability density functions, the factorizations corresponding to Fig.1 are

$$f_{123} = f_{1|3} f_{2|3} f_3, \quad f_{123} = f_{1|2} f_{2|3} f_3, \quad f_{123} = f_{1|23} f_2 f_3.$$

The implied constraints are conditional independence of the outer node pair given the inner node, both on the left and in the middle, and marginal independence of the outer node pair, on the right. In the notation introduced by Dawid, [20], one writes these constraints equivalently as

$$(f_{1|23} = f_{1|3}) \Leftrightarrow 1 \perp\!\!\!\perp 2|3, \quad (f_{1|23} = f_{1|2}) \Leftrightarrow 1 \perp\!\!\!\perp 3|2, \quad (f_{23} = f_2 f_3) \Leftrightarrow 2 \perp\!\!\!\perp 3,$$

again in a condensed notation in which each node denotes also a variable.

Only the generating process in the middle of Fig.1, specifies a full ordering of all three variables as $(1, 2, 3)$, while one cannot distinguish with the graph alone between $(1, 2, 3)$ and $(2, 1, 3)$ for the source V and between $(1, 2, 3)$ and $(1, 3, 2)$ for the sink V. More generally, a directed acyclic graph may be '**compatible with several orderings**' of the variables such that the set of all independences, that is the '**independences structure**' of a graph, remains unchanged. This poses problems for some machine-learning strategies. In many applications however, one compatible ordering can be taken as fixed; substantive knowledge may even give a full ordering of all variables.

Parent graphs and three Vs

A graph is said to form a '**dependence base**' if a full ordering of the nodes is fixed and each edge present in the graph means the lack of a conditional independence, typically a dependence that is considered to be strong in a given context. General properties of the graphs are also used. For regression graphs, these are stated here in Propositions 9 and 10. Directed acyclic graphs that form a dependence base have been named '**parent graphs**', [55], denoted by G^N_{par}. Their defining pairwise relations are in equation (1).

For each node i in the ordered node set, $N = (1, \ldots, d)$, of a parent graph, one knows which nodes are in '**the past of node i**', that is in set $\{> i\} = (i+1, \ldots, d)$. The subset of nodes in $\{> i\}$ from which arrows start and point to node i is the set of '**parents of node i**', denoted by par_i. In G^N_{par}, we have a dependence of each node i on all nodes in par_i and independence of i on all other nodes in the past of i. Expressed by using the $\Cap$-notation introduced for non-vanishing dependences by Wermuth and Sadeghi, [113], we have for $j > i$ in G^N_{par}:

$$i \Cap j|\text{par}_i \setminus \{j\} \text{ for } j \in \text{par}_i \text{ and } i \perp\!\!\!\perp j|\text{par}_i \text{ for } j \in \{> i\} \setminus \text{par}_i. \tag{1}$$

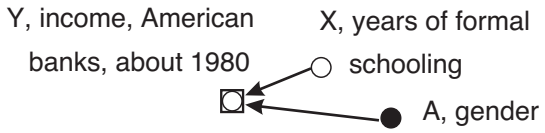
As mentioned before, one outstanding feature of a graphical Markov model is that its consequences can be derived, for instance for marginal or for conditional distributions. To illustrate this first for the graphs in Fig.1, we use a special notation. A '**boxed-in node**', $\boxed{\bigcirc}$, indicates conditioning on the levels of the variable at this node, and a '**crossed-out**' node, $\not\bigcirc$, means marginalizing over the variable, [104].

As justified later, we take sink Vs in G^N_{par} to be **edge-inducing** by conditioning and the source and transition Vs, to be edge-inducing by marginalizing; each of the Vs of Fig. 1 introduces a different type of edge. The Vs with the edge-inducing operation on the inner node is shown in the following line and the induced edges in the line thereafter.

$$\begin{array}{ccc} i \prec\!\!- \not\bigcirc \!\!-\!\!\succ j, & i \prec\!\!- \not\bigcirc \!\!-\!\!\prec j, & i \!-\!\!\succ \boxed{\bigcirc} \prec\!\!- j \\ i\text{---}j, & i \prec\!\!- j, & i \text{---} j\,. \end{array} \tag{2}$$

The induced edges 'remember at first' the type of path ends at i, j of the generating V, but then each $\longleftrightarrow$ is replaced by ---, since no direction is implied after ignoring a common source and, as explained below, the two types of undirected dependence can be readily distinguished.

The following example is derived from information on a social survey, [93]. It shows how conditioning on the inner node of a sink V induces a conditional dependence. To distinguish underlying continuous variables from discrete ones. The former are drawn with a circle, the latter with a dot.



In American banks in the nineteeneighties, salaries, Y, increased with higher levels of formal education, X, for both women and men, that is $Y \not\perp\!\!\!\perp X|A$, with A denoting gender. Men received a clearly higher salary than women at given levels of X, so that $Y \not\perp\!\!\!\perp A|X$. Furthermore, men and women had had equal chances to obtain higher levels of formal education, $X \perp\!\!\!\perp A$. This implies for $X \not\perp\!\!\!\perp A|Y$: for any given level of the salaries, women had a higher level of formal education than men.

We show in the next section how the above edge-inducing rules mimic the effects of marginalizing and conditioning in non-degenerate Gaussian distributions, those that have invertible covariance matrices.

Gaussian distributions generated over parent graphs

For linear relations in d mean-centered variables X_i, a non-degenerate Gaussian distribution is generated with

$$\boldsymbol{AX} = \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = 0, \quad \mathrm{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Delta} \text{ diagonal}, \tag{3}$$

where zero-mean, uncorrelated Gaussian residuals, ε_i, have positive variances $\sigma_{ii|>i}$ and are in the $d \times 1$ vector $\boldsymbol{\varepsilon}$. Vector $\boldsymbol{X}$ contains the variables X_i, and matrix $\boldsymbol{A}$ is '**unit upper-triangular**, that is it has ones along the diagonal and zeros below the diagonal. In row i, it has minus the values of linear regression coefficients resulting with response X_i regressed on $X_{>i}$, [95], [98].

In the early literature of econometrics, such linear relations have been discussed as recursive equations, [117] and were written in triangular form; for three variables as:

$$\begin{aligned} X_1 + a_{12}X_2 + a_{13}X_3 &= \varepsilon_1\,, \\ X_2 + a_{23}X_3 &= \varepsilon_2\,, \\ X_3 &= \varepsilon_3\,. \end{aligned}$$

Note that in a Gaussian distribution generated over a complete parent graph, none of the regression coefficients vanishes when each response X_i is regressed on all variables in its past, that is on $X_{>i}$.

By equation (1), missing edges in the starting graph define the '**independence constraints**'. For Gaussian distributions generated over parent graphs, these are reflected in vanishing regression coefficients and as zeros in '**matrices of equation parameters**'. For example, in the first and third generated distribution of Fig.1, we can write:

$$\begin{pmatrix} 1 & 0 & a_{13} \\ 0 & 1 & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}, \qquad \begin{pmatrix} 1 & a_{12} & a_{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

while for the second case in Fig.1, a_{12} and a_{23} are nonzero but $a_{13} = 0$.

For an explicit distinction between conditional and marginal dependences, we switch to a more detailed notation for trivariate Gaussian distributions. For instance, $\beta_{1|3.2} =$

$-a_{13}$ is the coefficient of X_3 in the linear regression of X_1 on X_2 and X_3, while $\beta_{2|3} = -a_{23}$ is the coefficient of X_3 in the linear regression of X_2 on X_3 alone.

The relation between marginal and conditional linear least-squares regression coefficients, due to William Cochran, [13], is called the recursion relation of these regression coefficients:

$$\beta_{1|3} = \beta_{1|3.2} + \beta_{1|2.3}\beta_{2|3}\,. \tag{4}$$

Thus, for a Gaussian distribution generated over the parent graph of Fig.1, which is a transition V, the conditional independence $1 \perp\!\!\!\perp 3|2$, ($\beta_{1|3.2} = 0$), implies the marginal dependence $1 \not\perp\!\!\!\perp 3$, ($\beta_{1|3} \neq 0$). Similarly, the marginal independence $1 \perp\!\!\!\perp 3$ implies the conditional dependence $1 \not\perp\!\!\!\perp 3|2$ since the edges present in the transition V mean $\beta_{1|2.3} \neq 0$ and $\beta_{2|3} \neq 0$. This property is shared by trivariate binary distributions, [81]. Joint distributions with this property in its generalized form, given here in equation (30), are said to be dependence inducing, [102], or to satisfy singleton transitivity, [100].

For a Gaussian distribution generated over a $G^N_{\rm par}$ of Fig.1, which is a sink V, the marginal independence $2 \perp\!\!\!\perp 3$ implies the conditional dependence $2 \not\perp\!\!\!\perp 3|1$. These features may be recognized with equation (5) below, after introducing correlations and their relations to other types of parameter.

With the covariance matrix denoted by $\mathbf{\Sigma}$ and its inverse, the concentration matrix, by $\mathbf{\Sigma}^{-1}$, we write explicitly

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ . & \sigma_{22} & \sigma_{23} \\ . & . & \sigma_{33} \end{pmatrix}, \qquad \mathbf{\Sigma}^{-1} = \begin{pmatrix} \sigma^{11} & \sigma^{12} & \sigma^{13} \\ . & \sigma^{22} & \sigma^{23} \\ . & . & \sigma^{33} \end{pmatrix}.$$

The .-notation indicates symmetric entries, the diagonal elements of $\mathbf{\Sigma}$ are the '**variances**', $\sigma_{ii} = E(X_i^2)$, and the off-diagonal elements are the '**covariances**', $\sigma_{ij} = E(X_i, X_j)$, of the mean-centered X_i. The diagonal elements of $\mathbf{\Sigma}^{-1}$ are the '**precisions**', σ^{ii}, the off-diagonal elements are the '**concentrations**', σ^{ij}.

The '**correlation coefficient**', ρ_{23}, and the '**partial correlation coefficient**', $\rho_{23|1}$, relate to the other parameters and to each other via

$$\rho_{23} = \sigma_{23}/\sqrt{\sigma_{22}\sigma_{33}},\;\; \rho_{23|1} = -\sigma^{23}/\sqrt{\sigma^{22}\sigma^{33}} = (\rho_{23} - \rho_{12}\rho_{13})/\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)},$$

$$\beta_{2|3} = \sigma_{23}/\sigma_{33} = -\sigma^{23.1}/\sigma^{22.1}, \quad \beta_{1|3.2} = \sigma_{13|2}/\sigma_{33|2} = -\sigma^{13}/\sigma^{11}.$$

In this notation, $\sigma^{23.1}$ is the concentration of $(2,3)$ after marginalizing over X_1 and $\sigma_{13|2}$ is the covariance of $(1,3)$ conditionally given $X_2 = x_2$.

Correlations are best suited to reflect the strength of linear dependences, here those induced by the independence constraints. With $2 \perp\!\!\!\perp 3$ and with $1 \perp\!\!\!\perp 3 \mid 2$, the induced conditional and marginal dependences are, respectively,

$$\rho_{23|1}* = -\rho_{12|3}\rho_{13|2}\,, \qquad \rho_{13}* = \rho_{12}\rho_{23}\,. \tag{5}$$

Thus, the induced linear dependence can be considerably stronger for a marginal than for a conditional independence. For instance with $2 \perp\!\!\!\perp 3$, there is $-\rho_{23|1}* > 0.96$ if $\rho_{12} = \rho_{13} = 0.7$ and $\mathbf{\Sigma}^{-1}$ does not exist if $\rho_{12} = \rho_{13} \geq \sqrt{0.5}$. By contrast if $\rho_{12} = \rho_{23} = 0.7$ and $1 \perp\!\!\!\perp 3|2$, the induced marginal correlation is only $\rho_{13}* = 0.49$.

Some properties of Gaussian distributions

There are recursions also for concentrations, [23], and for covariances, [3]:

$$\sigma^{23.1} = \sigma^{23} - \sigma^{12}\sigma^{13}/\sigma^{11}, \qquad \sigma_{13|2} = \sigma_{13} - \sigma_{12}\sigma_{23}/\sigma_{22}\,. \tag{6}$$

The first recursion shows that $0 = \sigma^{23.1} = \sigma^{23}$, that is both of $(2 \perp\!\!\!\perp 3$ and $(1 \perp\!\!\!\perp 2|3)$ hold, if $(0 = \sigma^{12}$ or $= \sigma^{13})$ in addition. Similarly, the second recursion shows that both of $(1 \perp\!\!\!\perp 3|2$ and $1 \perp\!\!\!\perp 3)$ hold if $(\sigma_{12} = 0$ or $\sigma_{23} = 0)$ in addition. Thus, an independence statement involving the third variable is needed for a variable pair to be both marginally and conditionally independent. This is the simplest case of inducing dependences, that is of '**singleton transitivity**'; see [100] and here equation (30).

Recursion relations such as in equations (4) and (6) and their connection to the elements of the above matrices $\boldsymbol{A}$ show also that in trivariate Gaussian distributions '**conditional independences combine downwards**' as:

$$(1 \perp\!\!\!\perp 2 \mid 3 \text{ and } 1 \perp\!\!\!\perp 3 \mid 2) \Longrightarrow \{1 \perp\!\!\!\perp (2,3) \Leftrightarrow f_{123} = f_1 f_{23}\} \Longrightarrow (1 \perp\!\!\!\perp 2 \text{ and } 1 \perp\!\!\!\perp 3),$$

that is they satisfy what is also called the '**intersection property**'. Furthermore, in these distributions '**conditional independences combine upwards**' as:

$$(2 \perp\!\!\!\perp 3 \text{ and } 1 \perp\!\!\!\perp 3) \Longrightarrow \{3 \perp\!\!\!\perp (1,2) \Leftrightarrow f_{123} = f_{12} f_3\} \Longrightarrow (2 \perp\!\!\!\perp 3 \mid 1 \text{ and } 1 \perp\!\!\!\perp 3 \mid 2),$$

that is they satisfy what is also called the '**composition property**'.

In the information theory literature, non-degenerate Gaussian distributions have been characterized by the above properties in terms of graphoids; these structures satisfy the properties common to all probability distributions plus intersection, [67], [87]:

Proposition 1 Lněnička and Matúš, [50]. *Gaussian distributions are singleton-transitive, compositional graphoids.*

To make graphs useful tools for empirical studies, the distributions generated over graphs have to share the properties of Prop.1 and are then called '**traceable regressions**', [100]; their graphs can be used to trace developmental pathways; see Example 2 below.

Families of discrete distributions which violate singleton transitivity, the intersection or the composition property require very special types of parametrizations, [100]. For the combination of independence statements of regression graphs, the intersection and the composition property are always used, [75]. These two properties also hold in distributions generated over parent graphs; see [55], discussion of Lemma 1, provided the ordering and the dependences are indeed as given with equation (1).

The relations between linear parameters, discussed above, generalize to more than three variables, but switching to a matrix notation and to edge matrix representations of graphs becomes useful for discussing most independence properties in general; for joint Gaussian distributions, see for instance [55], Appendix 2. Here, we start again with the simplest type of edge matrices, those to the graphs of Fig.1.

Structural versus parametric implications

An edge matrix $\boldsymbol{\mathcal{A}}$ can be viewed as the sum of an identity matrix, $\boldsymbol{I}$, and what has been named the adjacency matrix in graph theory; a square binary matrix with an iij-one if

there is a directed edge in the graph and an additional ji–one for an undirected edge. The small change of adding $\boldsymbol{I}$ leads to well-defined matrix products which can be used to derive structural consequences of a given generating graph. As we shall see, such structural consequences may differ from those of a given generating set of parameters.

The edge matrices $\boldsymbol{\mathcal{A}}$ in Table 1 share the unit upper-triangular form with the linear equation parameter matrices $\boldsymbol{A}$ given with the generating equations (3).

Table 1: *Edge matrices for the three Vs of Fig.1*

Edge matrices $\boldsymbol{\mathcal{A}}$ of	$1 \prec\!\!-\!\!3 \succ\!\!- 2$	$1 \prec\!\!-\!\!2 \prec\!\!- 3$	$2 \succ\!\!-\!\!1 \prec\!\!- 3$
$\boldsymbol{\mathcal{A}}$:	$\begin{pmatrix} 1 & 0 & 1 \\ & 1 & 1 \\ \mathbf{0} & & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 0 \\ & 1 & 1 \\ \mathbf{0} & & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ & 1 & 0 \\ \mathbf{0} & & 1 \end{pmatrix}$

For instance with $\boldsymbol{M}^{\mathrm{T}}$ denoting the transpose of a matrix $\boldsymbol{M}$, Gaussian systems, $\boldsymbol{AX} = \boldsymbol{\varepsilon}$ of equation (3), imply as covariance and concentration matrices,

$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}\boldsymbol{\Delta}\,(\boldsymbol{A}^{-1})^{\mathrm{T}}, \qquad \boldsymbol{\Sigma}^{-1} = \boldsymbol{A}^{\mathrm{T}}\,\boldsymbol{\Delta}^{-1}\boldsymbol{A}\,, \tag{7}$$

where the matrix pairs $(\boldsymbol{A}, \boldsymbol{\Delta}^{-1})$ and $(\boldsymbol{A}^{-1}, \boldsymbol{\Delta})$ are ordered Cholesky decompositions or '**triangular decompositions**' of $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}$, respectively, [98].

For Gaussian distributions, zero elements in $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ coincide with those independences that hold, more generally, in **covariance and concentration graphs**, respectively, of other types of distribution:

$$(\sigma_{ij} = 0) \Leftrightarrow i \perp\!\!\!\perp j\,, \qquad (\sigma^{ij} = 0) \Leftrightarrow (i \perp\!\!\!\perp j \mid N \setminus \{i, j\})\,.$$

For dependence base Vs of Fig.1, it may be checked directly that with $\Delta_{ii} = \sigma_{ii|\mathrm{par}_i} > 0$, a nonzero element is induced in different positions in row one of $\boldsymbol{\Sigma}^{-1}$ for the source V and for the transition V, while a nonzero element is induced in position $(2, 3)$ of $\boldsymbol{\Sigma}$ for the sink V; see also equation (5).

In general, implications of a graph result via transformations of edge matrices. The edge matrix $\boldsymbol{\mathcal{A}}$ of G^N_{par}, for node set N of size d, is the $d \times d$ unit upper-triangular matrix $\boldsymbol{\mathcal{A}} = (\mathcal{A}_{ij})$ such that

$$\mathcal{A}_{ij} = \begin{cases} 1 & \text{if and only if } i \prec\!\!- j \text{ in } G^N_{\mathrm{par}} \text{ or } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

For path interpretations, a definition of node j being an '**ancestor**' of '**descendant**' i is needed: there starts a direction-preserving path at j leading to node i. We will now derive the edge matrix transformation that turns every ancestor in G^N_{par} into a parent.

The k'th power of the adjacency matrix $(\boldsymbol{\mathcal{A}} - \boldsymbol{I})$ is known to count for each $i < j$ in G^N_{par} the number of direction-preserving paths of length k connecting nodes i and j. Since the longest of such paths has $d - 1$ edges, zero matrices $(\boldsymbol{\mathcal{A}} - \boldsymbol{I})^k$ result for all $k > d-1$. Thus, the edge matrix of the ancestor graph of G^N_{par}, denoted by $\boldsymbol{\mathcal{A}}^-$, becomes

$$\boldsymbol{\mathcal{A}}^- = \mathrm{In}[(2\boldsymbol{I} - \boldsymbol{\mathcal{A}})^{-1}],\ (2\boldsymbol{I} - \boldsymbol{\mathcal{A}})^{-1} = \boldsymbol{I} + (\boldsymbol{\mathcal{A}} - \boldsymbol{I}) + (\boldsymbol{\mathcal{A}} - \boldsymbol{I})^2 + \ldots + (\boldsymbol{\mathcal{A}} - \boldsymbol{I})^{\{d-1\}},$$

where 'In' is the indicator function that replaces every positive entry of a nonnegative matrix by a one. The above sum is the matrix analogue to the sum of an infinite geometric series, where for $\mid a \mid < 1$, one obtains $(1-a)^{-1} = 1+a+a^2+\ldots$, ([61], p. 29, [53]). This is generalized here in equation (13). The edge matrix analogue to equation (7) is introduced next.

With the edge matrices $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{A}}^{-}$, the consequences of the starting graph, G^N_{par}, for pairwise marginal and for conditional independences given all remaining variables, can be directly given. An implied independence $i \perp\!\!\!\perp j$ and $i \perp\!\!\!\perp j \mid N \setminus \{i,j\}$, respectively, is indicated by a zero in positions (i,j) of

$$\boldsymbol{\mathcal{N}}_{NN} = \text{In}[\boldsymbol{\mathcal{A}}^{-}(\boldsymbol{\mathcal{A}}^{-})^{\text{T}}], \qquad \boldsymbol{\mathcal{N}}^{NN} = \text{In}[\boldsymbol{\mathcal{A}}^{\text{T}}\boldsymbol{\mathcal{A}}], \tag{9}$$

that is in the edge matrices of the '**overall covariance and concentration graph induced by G^N_{par}**'; see also equation (7) and the next section.

Such zeros are said to be '**structurally induced**' since they result for all distributions that factorize as prescribed by a given generating graph. With the examples in Fig. 2 and Fig. 3 in the next section, the types of path are identified which induce more edges than there are present in a starting parent graph and therefore lead to more complex structures, captured in one or both of the two induced undirected graphs.

These edge-inducing paths introduce an additional dependence in Gaussian distributions generated over parent graphs, provided no other constraints apply than the pairwise independences and dependences defining their generating graph; see equation (1) so that contributions of several paths may get cancelled.

If an independence is not structurally induced, then it may still get generated by particular constellations of the parameters. Such cases have been called '**parametric cancellation**', [102] or '**lack of faithfulness to the graph**', [84]. For instance, a parametric cancellation occurs if in equation (3), one has $\beta_{1|3.2} = -\beta_{1|2.3}\beta_{2|3}$. This leads to a zero in position $(1,3)$ of $\boldsymbol{\Sigma}$ even when $1 \not\perp\!\!\!\perp 2|3$ and $2 \not\perp\!\!\!\perp 3$ and hence to a non-structural independence, $1 \perp\!\!\!\perp 3$, for Gaussian distributions.

Some consequences of a five-node parent graph

For five ordered nodes, $N = (1,\ldots,5)$, Fig.2 shows a parent graph, G^N_{par}, which contains the three types of V of Fig.1. Edges present and edges missing are defined by equation (1). The factorization of f_N can be read directly off the graph:

$$f_N = f_{1|23} f_2 f_{3|5} f_{4|5} f_5.$$

Also, the graph can be drawn using the given order of the nodes and this factorization.

To generate the joint distribution over G^N_{par}, one starts with f_5, generates $f_{4|5}$ next, then $f_{3|5}$, then f_2 and finally $f_{1|23}$. The defining pairwise dependences in equation (1) give

$$1 \not\perp\!\!\!\perp \{2,3\}\,, \qquad 3 \not\perp\!\!\!\perp 5\,, \qquad 4 \not\perp\!\!\!\perp 5\,,$$

so that no simpler factorization holds in distributions generated over this parent graph. From the pairwise independences in equation (1) or from the factorization of f_N, one obtains the defining independence structure of G^N_{par} as:

$$1 \perp\!\!\!\perp \{4,5\} \mid \{2,3\}\,, \qquad 2 \perp\!\!\!\perp \{3,4,5\}\,, \qquad 3 \perp\!\!\!\perp 4 \mid 5.$$

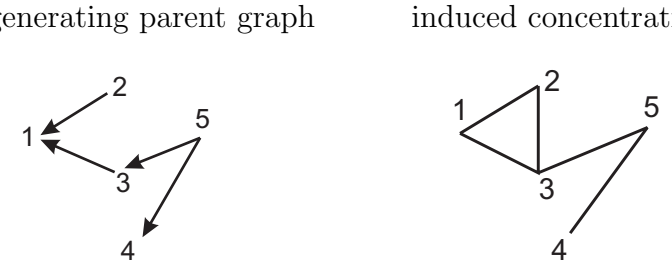


Figure 2: *left: a small $G^N_{\rm par}$ with the three types of* V*, with source node 5, transition node 3 and sink node 1; right: the induced concentration graph with edge for* $(2,3)$ *due to conditioning on the common sink node 1 of (2,3).*

All further implied independences may, in principle, be derived directly from such a list of independences by using the properties of the starting graph. We turn to these properties in Prop. 9, 10. Similarly, further implied dependences can be obtained by using the factorization of f_N and the information that the factorization cannot be further simplified. But, one may instead use the edge-inducing properties, [68], of Vs in parent graphs, extending the discussion above for three-node graphs. Proofs may be based on Prop. 4 below. We start with consequences of the sink V in Fig.2.

It can be derived that for every sink V with outer nodes i, j, all independence statements for i and j implied by $G^N_{\rm par}$ exclude the inner sink node 'o'. Here we have for instance, $2 \perp\!\!\!\perp 3$, $2 \perp\!\!\!\perp 3|4$ and $2 \perp\!\!\!\perp 3|\{4,5\}$ so that there are several subsets c of $N \setminus \{i, o, j\}$ for which $i \perp\!\!\!\perp j|c$ is implied by the parent graph, here e.g. $c = \emptyset$, $c = \{4\}$, $c = \{4,5\}$. Thus, given a sink V, there are $c \subset N \setminus \{i, o, j\}$ such that $i \perp\!\!\!\perp k|c$ is implied by the graph. For each such c,

$$\text{nodes } (i, \mathrm{o}, j) \text{ forming a sink V in } G^N_{\rm par} \Leftrightarrow (i \perp\!\!\!\perp j|c \implies i \pitchfork j|\mathrm{o}c)\,. \tag{10}$$

In Fig. 2, for instance, $2 \pitchfork 3|1$, $2 \pitchfork 3|\{1,4\}$ and $2 \pitchfork 3|\{1,4,5\}$ are induced. For Gaussian distributions, the size of such dependences can be expressed in terms of induced partial correlations, in a similar way as in equation (5).

The concentration graph induced by $G^N_{\rm par}$, involves conditioning on all nodes. The additional edges result by closing sink Vs, as captured by $\mathcal{N}^{NN}$ in equation (9). More edges represent in general a more complex structure and in cases with complete, undirected subgraphs of three or more nodes, it cannot be recognized from a concentration graph alone which edges are due to conditioning on sink Vs in $G^N_{\rm par}$.

We turn next to consequences of transition and source Vs by using Fig.3. It can be derived that for every transition V with outer nodes i, j, all independence statements for i and j implied by $G^N_{\rm par}$ include the inner node 'o';. In Fig. 3, we have for instance $1 \perp\!\!\!\perp 5|3$, $1 \perp\!\!\!\perp 5|\{2,3\}$, so that there are several subsets c of $N \setminus \{i, o, j\}$ for which $i \perp\!\!\!\perp j|oc$ is implied by the parent graph, here such as $c = \emptyset$, $c = \{2\}$.

Thus, given a transition V, there are $c \subset N \setminus \{i, o, j\}$ such that $i \perp\!\!\!\perp j|oc$ is implied by the graph. For each such c,

$$\text{nodes } (i, \mathrm{o}, j) \text{ forming a transition V in } G^N_{\rm par} \Leftrightarrow (i \perp\!\!\!\perp j|oc \implies i \pitchfork j|c)\,. \tag{11}$$

In Fig. 3, for instance, $1 \pitchfork 5$ and $1 \pitchfork 5|2$ are induced. A fully analogous statement results by replacing in the previous paragraphs each time 'transition node' by 'source node'. Just the examples relating to Fig.3 change.

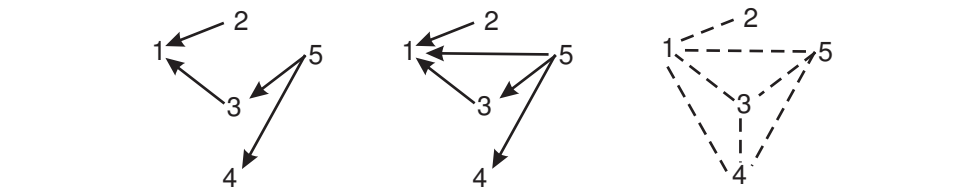


Figure 3: *left: the same generating* $G^N_{\rm par}$ *as in Fig.2; middle: the corresponding ancestor graph, also called the transitive closure of* $G^N_{\rm par}$ *; right: the induced covariance graph with new edges for* $(1,4)$*,* $(1,5)$*,* $(3,4)$ *compared to* $G^N_{\rm par}$ *.*

The edge matrix $\mathcal{N}_{NN}$ in equation (9) shows that by moving from the ancestor graph to the induced covariance graph, every source V in the former is closed by an edge. Then, in the overall covariance graph induced by $G^N_{\rm par}$, there is an additional ij-edge if either j is an ancestor of i or i and j have a common ancestor.

Unless $G^N_{\rm par}$ contains exclusively sink Vs, there will be more edges in the induced covariance graph. And again, whenever three or more nodes are contained in some of its complete subgraphs, it is impossible to see from the graph alone, whether additional dependences have been generated. Therefore, the same type of general warning as above applies to using the class of covariance graphs for model selection. But furthermore, when a learning strategy is based on only the relations among variable pairs, no joint distribution may exist for such a given set of two-way margins that results from joint distributions with higher-order interactions; for an example see [110].

To summarize, with information on the ordering of the variables, simpler structures will typically be uncovered, unless no additional edges are introduced, so that, say, a starting $G^N_{\rm par}$ and a concentration graph have the same edge and node sets but different types of edge. In such important special situations, there is '**Markov equivalence**'; the same independence structure is captured by two different graphs; see Prop. 5 below.

Undirected generating graphs

Suppose now that variables are unordered, that is arising at the same time, like several symptoms of a disease or local consequences of a global economic shock. Their joint distribution could then have a generating concentration graph, $G^N_{\rm con}$, or a generating covariance graph, $G^N_{\rm cov}$. The defining pairwise independences for $G^N_{\rm con}$ are $i \perp\!\!\!\perp j | N \setminus \{i,j\}$ and those for $G^N_{\rm cov}$ are $i \perp\!\!\!\perp j$. For dependence base undirected graphs, each ij-edge present means:

$$i \text{ —— } j \Leftrightarrow i \pitchfork j | N \setminus \{i,j\} \text{ in } G^N_{\rm con} \quad \text{and} \quad i \text{- - - } j \Leftrightarrow i \pitchfork j \text{ in } G^N_{\rm cov} \,.$$

To read all implied independences off their graphs, a standard separation criterion from graph theory can be applied. For this, one says '**a path intersects a subset** $\boldsymbol{g}$' of node set N if it has an inner node in g. We let next $\{\alpha, \beta, c, m\}$ partition node set N, where only m or c may be empty sets. This notation is to remind one that with any independence statement $\alpha \perp\!\!\!\perp \beta | c$, one implicitly has marginalised over the remaining nodes in $m = N \setminus \{\alpha \cup \beta \cup c\}$; one considers the joint distribution of Y_α, Y_β given Y_c.

Proposition 2 Darroch et al., [19]. *A generating concentration graph, $G^N_{\rm con}$, implies $\alpha \perp\!\!\!\perp \beta | c$ if every path between α and β intersects c.*

Proposition 3 Kauermann, [42]. *A generating covariance graph, $G^N_{\rm cov}$, implies $\alpha \perp\!\!\!\perp \beta | c$ if every path between α and β intersects m.*

Whenever these undirected generating graphs also form dependence bases, the converse holds as well: if a node in α is connected to one in β by a path that does not intersect c in $G^N_{\rm con}$ or that does not intersect m in $G^N_{\rm cov}$, then $\alpha \pitchfork \beta \mid c$ is implied. Some additional effects of Prop. 1 and 2 result by considering '***a*-line ij-paths**': those which connect node pair i, j and have all inner nodes in $a \subset N$.

Corollary 1 *By marginalizing over any subset a of N in $G^N_{\rm con}$, all a-line paths are closed while by conditioning on a, its subgraph of $N \setminus a$ is induced for $N \setminus a$. By conditioning on subset a in $G^N_{\rm cov}$, all a-line paths are closed while by marginalizing over a, its subgraph of $N \setminus a$ is induced for $N \setminus a$.*

Prop. 1 and 2 imply more for '**connected graphs**', that is when the nodes of every node pair can be reached via some path.

Corollary 2 *A connected $G^N_{\rm con}$ induces for node set N a complete covariance graph and a connected $G^N_{\rm cov}$ induces for N a complete concentration graph.*

To summarize, in $G^N_{\rm con}$, each full-line V is edge-inducing by marginalizing and in $G^N_{\rm cov}$, each dashed-line V is edge-inducing by conditioning, where we take again the induced edges to remember the edge-ends of the starting V. Note again that for Gaussian distributions generated over a dependence base $G^N_{\rm con}$ or $G^N_{\rm cov}$, an induced edge coincides always with an induced dependence:

$$i \text{ —} \not\oslash \text{— } j, \qquad i \text{---} \boxed{\bigcirc} \text{---} j, \tag{12}$$

$$i \text{ —— } j, \qquad\qquad i \text{---} j\,.$$

The edge matrix of a complete generating graph of $G^N_{\rm cov}$ or $G^N_{\rm con}$ is a $d \times d$ matrix of ones. It has $d-1$ zero eigenvalues and one eigenvalue equal to d. Hence, it is not invertible, but by subtracting it from a $(d+1)$ multiple of an identity matrix, one obtains a well-posed inversion task, [90]. In the statistical literature, this type of ***Tikhonov regularization*** was introduced some fifteen years later in the form of ridge regression; a seemingly ill-posed problem is solved by increasing the diagonal elements of the matrix.

If we denote by $\boldsymbol{\mathcal{W}}$ any of the symmetric edge matrices of a generating $G^N_{\rm con}$ or $G^N_{\rm cov}$, then the corresponding edge matrices induced for the covariance or the concentration graph are of the type:

$$\boldsymbol{\mathcal{W}}^- = \mathrm{In}[\{(d+1)\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{W}}\}^{-1}]. \tag{13}$$

By definition, the matrix $(d+1)\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{W}}$ preserves the zero pattern of a given edge matrix $\boldsymbol{\mathcal{W}}$ and it is a M-matrix, so that its inverse is nonnegative. The concept of a M(inkowski)-matrix was introduced and studied by Ostrowski, [63] [64] without any applications concerning graphs or statistics; it is an invertible matrix with exclusively nonpositive off-diagonal elements. For undirected generating graphs, the M-matrix in equation (13) turns each connected component into a complete subgraph.

Figs. 2 and 3 above illustrate, in particular, that a generating, undirected graph is typically different from a corresponding induced graph. The latter summarizes all independences of a defined type implied by, say, a starting $G^N_{\rm par}$. It can, in general, not be used to derive further implied independences of another type; exceptions are discussed here later. The concentration graph and the covariance graph induced by the $G^N_{\rm par}$ in Figs. 2 and 3 are both incomplete, connected graphs. If they were also generating graphs, this would, by Corollary 2 or by equation (13), give a contradiction.

In spite of the similarities of the two types of undirected graph, estimation of covariance graph structures, [4], [107], [12], [51], [43], [118], is typically much more complex than estimation of concentration graph structures, [23], [98], [83], [10], [50]). The latter but not the former have, for instance, reduced sets of minimal sufficient statistics, [8], [15], in exponential families with independences constraints, and for Gaussian distributions, there is a unique maximum of the likelihood function whenever there are less variables than observations (for '$p < n$'), [23].

Regression graphs

Regression graphs are simple graphs with response nodes in a set u and context nodes in a set v such that for an '**ordered split**' of the node set as $N = (u, v)$, the density of joint responses, $\mathbf{X}_u$ is considered conditionally given the context variables in vector $\mathbf{X}_v$ and the joint density factorizes as

$$f_N = f_{u|v} f_v \,. \tag{14}$$

Furthermore, the response set u has an ordered partition into connected components as $u = (g_1, \ldots, g_k, \ldots, g_K)$ so that for all nodes in the subgraph of g_k, the nodes in their past are in $g_{>k} = \{g_{k+1}, \ldots, g_K, v\}$ and

$$f_{u|v} = \textstyle\prod_{k=1}^{K} f_{g_k|g_{>k}} \,. \tag{15}$$

Simplifying conditional independences are captured by the '**regression graph $\boldsymbol{G^N_{\rm reg}}$**'. This simple graph uses Definition 1 and consists of a concentration graph for the context nodes, a conditional covariance graph for each of the '**concurrent responses**', that is for $\boldsymbol{X}_k$ within each connected component g_k, and a directed acyclic graph in the vector variables $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K, \boldsymbol{X}_v)$.

The set-up for a regression graph model starts with the response-vector variable $\boldsymbol{X}_1$ of primary interest, possibly followed by one of secondary interest and ends with a context-vector variable $\boldsymbol{X}_v$; for an example see Fig. 4.

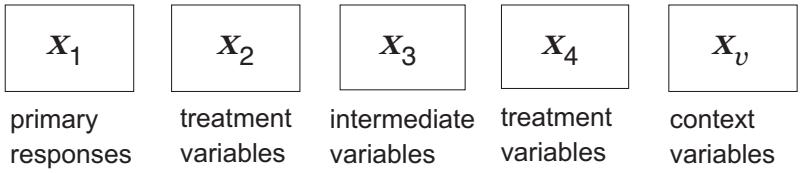


Figure 4: *A typical ordering of vector variables for a regression graph model.*

Intermediate variables form a sequence of variables between $\boldsymbol{X}_1$ and $\boldsymbol{X}_v$.

We let G^N_{reg} form a dependence base, so that edges present mean non-vanishing dependences; one compatible ordering of the nodes is fixed.

Definition 1 Wermuth and Sadeghi, [113]. *An ij-edge present in G^N_{reg} means*

i---j with i, j in g_k:	$i \Cap j \mid g_{>k}$,
$i \prec\!\!- j$ with i in g_k and j in $g_{>k}$:	$i \Cap j \mid g_{>k} \setminus \{j\}$,
i ——j with i, j in v:	$i \Cap j \mid v \setminus \{i, j\}$,

while for uncoupled pairs (i, j), the dependence sign $\Cap$ is replaced by the independence sign $\perp\!\!\!\perp$, but the conditioning sets remain unchanged.

There are equivalent pairwise properties of G^N_{reg}, [77], important for interpretation, such as $i \perp\!\!\!\perp j | \text{par}_i$ for uncoupled (i, j) with $i \in g_k$ and $j \in g_{>k}$.

A distribution is said to be '**generated over a regression graph**' when it satisfies the factorizations of equations (14), (15) while independences as well as dependences are specified by Definition 1 for a given node ordering $N = (1, \ldots, d)$. Note that with Definition 1, G^N_{reg} is unchanged for a reordering of the nodes within any response set g_k.

In a regression graph, three additional Vs may occur compared to those in a parent graph, see equation (2), and in the two types of undirected graph, see equation (12):

$$i \prec\!\!- \not\emptyset \text{---} j, \quad i \prec\!\!- \not\emptyset \text{——} j, \quad i \text{---} \boxed{\circ} \prec\!\!- j \tag{16}$$

$$i \text{---} j, \qquad i \prec\!\!- j, \qquad i \prec\!\!- j\,.$$

With Definition 1 and a fixed compatible ordering of the nodes, three edge sets of different types are given for G^N_{reg}, in a self-explanatory notation, as $E_{--}, E_{\prec\!\!-}, E_{--}$. Their union defines one edge set E. Three different Vs are edge-inducing by conditioning on the inner node, see the last V on the right-hand side of equations (16), (12), (2). These are the '**collision Vs**', the other five possible types of a G^N_{reg} are the '**transmitting Vs**'. Accordingly, the inner nodes are '**collision nodes**' or '**transmitting nodes**'.

One justification for the types of induced edge stems from the construction of summary graphs, one class of '**independence-preserving**' graphs, those which preserve all independences implied by a generating G^N_{par} or G^N_{reg} in a smaller graph obtained after marginalizing, conditioning and removing nodes as well as their edges, [99], [74], [76]. In particular, the above different types of Vs , plus two more that are used in constructing summary graphs, can be combined in any order in a consistent way, [99] Appendix.

There are other classes of independence-preserving graphs, not described here, by which additional implications of a generating graph may be derived from a smaller graph. These may have different types of edge, [44], [70], [74], serve different purposes but define the same independence structures.

To derive independences implied by G^N_{reg}, further concepts are useful. The notion of anteriors of a response i, [73], extends the one of ancestors in G^N_{par} to G^N_{reg}. For $N = (u, v)$ and $i \prec\!\!- j$, node i in g_k is within u, but the parent node j can be any node in $g_{>k}$, the past of i. '**Anterior paths**' join paths among context nodes in v with an arrow to descendant-ancestor paths in u:

$$i \prec\!\!- \overbrace{\underbrace{\circ \prec\!\!- \circ, \ldots, \circ \prec\!\!- d_u}_{\text{ancestors of } i} \text{——} 1 \text{——} \circ, \ldots, \circ \text{——} d_v}^{\text{anteriors of } i}.$$

Recall that an a-line path connects a node pair by a path with all inner nodes in a subset a of N. With this, the notion of an ancestor graph of G^N_{par} can be extended.

Definition 2 *An* ***a*-line anterior graph** *of G^N_{reg} has edge $i \prec\!\!-\!\!- j$ for every a-line anterior j of i in G^N_{reg} and a-line paths for context nodes in v are closed.*

This graph permits to express the effects of separation in G^N_{reg}, [73], [75], in a way comparable to those in undirected graphs, see Prop. 2 and 3. Again, let $N = (a, b)$, $a = \{\alpha, m\}$ and $b = \{\beta, c\}$, where only m or c may be empty.

Proposition 4 Wermuth, [101]. *A regression graph implies $\alpha \perp\!\!\!\perp \beta | c$ if along every path between α and β, in the a-line anterior graph of G^N_{reg}, a collision node intersects m or a transmitting node intersects c.*

The converse of Prop. 4 holds with Definition 1, and a fixed compatible ordering of the nodes. Prop. 4 specializes to the effects of separation in directed acyclic graphs; see [55], Criterion 1, also for a proof of equivalence to other path criteria for the independence implications of the G^N_{par}, [65], [33], [46].

Corollary 3 *A path between α and β in the a-line anterior graph of G^N_{reg} is edge-inducing if every collision node is in c and every other node is in m.*

Corollary 4 *A path between α and β in the a-line ancestor graph of G^N_{par} is edge-inducing if every sink node is in c and every other node is in m.*

In particular, G^N_{reg} and G^N_{par} induce a complete graph by marginalizing over N if, for node 1, the last node d is an anterior in G^N_{reg} or an ancestor in G^N_{par}.

One further important question is whether two regression graphs with different types of edge can define the same independence structure if they have the same node set N and an identical edge set $E = E_{--} \cup E_{\prec\!\!-} \cup E_{---}$.

Proposition 5 Wermuth and Sadeghi, [113]. *Two regression graphs, with different types of edge but an identical node set N and an identical edge set E, are Markov equivalent if and only if their sets of collision Vs coincide.*

Thus for instance, a given regression graph is Markov equivalent to its induced concentration graph if and only if G^N_{reg} does not contain a collision V, and to its induced covariance graph if and only if G^N_{reg} does not contain any transmitting V. A covariance and a concentration graph are Markov equivalent if and only if they consist of identical sets of complete subgraphs.

Before we derive graphs induced by G^N_{reg}, we introduce two basic types of Gaussian distributions that may get generated over a regression graph.

Two types of Gaussian regression graph model

We note first that for mean-centered variables and $N = (a, b)$, a '**linear regression of a joint response $\boldsymbol{X}_a$ on $\boldsymbol{X}_b$** gives, [95],

$$\boldsymbol{X}_a = \boldsymbol{\Pi}_{a|b}\boldsymbol{X}_b + \boldsymbol{\eta}_a, \;\; E(\boldsymbol{\eta}_a) = 0, \;\; \text{cov}(\boldsymbol{\eta}_a, \boldsymbol{X}_b) = 0, \;\; \text{cov}(\boldsymbol{\eta}_a) \text{ invertible.} \qquad (17)$$

The parameters are a matrix of population least-squares regression coefficients $\mathbf{\Pi}_{a|b}$ and a residual covariance matrix $\mathbf{\Sigma}_{aa|b} = E(\boldsymbol{\eta}_a \boldsymbol{\eta}_a^{\mathrm{T}})$. The interpretation of $\mathbf{\Pi}_{a|b}$ results by post multiplication with $\boldsymbol{X}_b^{\mathrm{T}}$ and taking expectations: $E(\boldsymbol{X}_a \boldsymbol{X}_b^{\mathrm{T}}) - \mathbf{\Pi}_{a|b} E(\boldsymbol{X}_b \boldsymbol{X}_b^{\mathrm{T}}) = \mathbf{0}$.

Joint Gaussian distributions generated over a corresponding regression graph are non-degenerate, have a concentration matrix $\mathbf{\Sigma}^{bb.a}$ for $\boldsymbol{X}_b$ and zeros in the defined parameter matrices are given by Definition 1 for $K = 1$.

The well-known relations of these parameter matrices, [22], [23] Appendix B, [55] Appendix 1, to $\mathbf{\Sigma} = \mathrm{cov}(\boldsymbol{X}_N)$ and to $\mathbf{\Sigma}^{-1}$ are for $N = (a, b)$:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ . & \mathbf{\Sigma}_{bb} \end{pmatrix}, \quad \mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{\Sigma}^{aa} & \mathbf{\Sigma}^{ab} \\ . & \mathbf{\Sigma}^{bb} \end{pmatrix},$$

$$\begin{aligned} \mathbf{\Sigma}_{aa|b} &= \mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab} \mathbf{\Sigma}_{bb}^{-1} \mathbf{\Sigma}_{ba} = (\mathbf{\Sigma}^{aa})^{-1}, \\ \Pi_{a|b} &= \mathbf{\Sigma}_{ab} \mathbf{\Sigma}_{bb}^{-1} = -(\mathbf{\Sigma}^{aa})^{-1} \mathbf{\Sigma}^{ab}, \\ \mathbf{\Sigma}^{bb.a} &= \mathbf{\Sigma}^{bb} - \mathbf{\Sigma}^{ba} (\mathbf{\Sigma}^{aa})^{-1} \mathbf{\Sigma}^{ab} = \mathbf{\Sigma}_{bb}^{-1}, \end{aligned} \tag{18}$$

where the expressions for $\mathbf{\Sigma}^{bb.a}$ and $\mathbf{\Sigma}_{aa|b}$ are the matrix forms of the recursion relations for concentrations and covariances in equation (6). As is explained later, these matrix results can all be obtained by applying the matrix operator named partial inversion. The result analogous to Corollary 1 is the following direct consequence of equation (18).

Corollary 5 *For any subset a of N, marginalizing in $\mathbf{\Sigma}^{-1}$ over a gives $\mathbf{\Sigma}_{aa|b}$ and $\mathbf{\Pi}_{a|b}$, while conditioning in $\mathbf{\Sigma}^{-1}$ on a leaves the submatrix $\mathbf{\Sigma}^{bb}$ unchanged. For b subset of N, conditioning in $\mathbf{\Sigma}$ on b gives $\mathbf{\Sigma}^{bb.a}$ and $\Pi_{a|b}$ while marginalizing in $\mathbf{\Sigma}$ over b leaves the submatrix $\mathbf{\Sigma}_{aa}$ unchanged.*

This applies, in similar form also, for $a = (\alpha, \gamma)$, to $\mathbf{\Sigma}_{\alpha\alpha|b}$ and with $b = (\beta, \delta)$ to $\mathbf{\Sigma}^{\beta\beta.a}$, that is marginalizing in any covariance matrix leads to a submatrix and conditioning in any concentration matrix leads to a submatrix, while more non-vanishing parameters may get induced, otherwise.

For $a = (\alpha, \gamma)$ and $b = (\beta, \delta)$, marginalizing over γ and conditioning on δ gives also a submatrix: $\mathbf{\Pi}_{\alpha|\beta.\delta}$, where α indicates the response, β the regressor and δ the remaining regressors conditioned on:

$$\mathbf{\Pi}_{a|b} = \begin{pmatrix} \mathbf{\Pi}_{\alpha|\beta.\delta} & \mathbf{\Pi}_{\alpha|\delta.\beta} \\ \mathbf{\Pi}_{\gamma|\beta.\delta} & \mathbf{\Pi}_{\gamma|\delta.\beta} \end{pmatrix}. \tag{19}$$

Thus, by Corollary 5 and the same partition as in equation (19), the parameters for $f_{\alpha|\beta\delta}$ and $f_{\beta|\delta}$ are simply submatrices of those for $f_{a|b}$ and f_b.

$$\mathbf{\Sigma}_{\alpha\alpha|b} = [\mathbf{\Sigma}_{aa|b}]_{\alpha,\alpha}, \quad \mathbf{\Pi}_{\alpha|\beta.\delta} = [\mathbf{\Pi}_{a|b}]_{\alpha,\beta}, \qquad \mathbf{\Sigma}_{\beta\beta.a} = [\mathbf{\Sigma}^{bb.a}]_{\beta\beta}. \tag{20}$$

The matrices $\mathbf{\Sigma}_{aa|b}$, $\mathbf{\Pi}_{a|b}$, $\mathbf{\Sigma}^{bb.a}$ arise also, with $\boldsymbol{I}$ denoting an identity matrix, in orthogonalized equations, [103], corresponding to equations (17):

$$\begin{pmatrix} \boldsymbol{I}_{aa} & -\mathbf{\Pi}_{a|b} \\ \mathbf{0} & \mathbf{\Sigma}_{bb}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_a \\ \boldsymbol{X}_b \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta}_a \\ \boldsymbol{\eta}_b \end{pmatrix}. \tag{21}$$

Note that $\mathrm{cov}(\mathbf{\Sigma}_{bb}^{-1} \boldsymbol{X}_b) = \mathbf{\Sigma}_{bb}^{-1}$ so that this concentration matrix plays several roles.

For a **Gaussian regression graph model**, recall that equations (14), (15) and Definition 1 apply. Covariance matrices after regressing $\boldsymbol{X}_k$ on $\boldsymbol{X}_{>k}$ and $\boldsymbol{\Sigma}_{vv}^{-1}$ are in a block-diagonal matrix $\boldsymbol{W}_{NN}$. The matrix of equation parameters, $\boldsymbol{H}_{NN}$, is upper block-triangular with identity matrices in the sizes of g_k along the diagonal, $\boldsymbol{\Sigma}_{vv}^{-1}$ in the last block and off-diagonally $-\boldsymbol{\Pi}_{g_k|g_{>k}}$:

$$\boldsymbol{H}_{NN}\boldsymbol{X}_N = \boldsymbol{\eta}_N \text{ with } \boldsymbol{W}_{NN} = \mathrm{cov}(\boldsymbol{\eta}_N). \tag{22}$$

As an example, we choose $K = 2$ and $u = (\alpha, \gamma)$:

$$\boldsymbol{H}_{NN} = \begin{pmatrix} \boldsymbol{I}_{\alpha\alpha} & -\boldsymbol{\Pi}_{\alpha|\gamma.v} & -\boldsymbol{\Pi}_{\alpha|v.\gamma} \\ & \boldsymbol{I}_{\gamma\gamma} & -\boldsymbol{\Pi}_{\gamma|v} \\ \boldsymbol{0} & & \boldsymbol{\Sigma}_{vv}^{-1} \end{pmatrix} \quad \boldsymbol{W}_{NN} = \begin{pmatrix} \boldsymbol{\Sigma}_{\alpha\alpha|\gamma v} & & \boldsymbol{0} \\ & \boldsymbol{\Sigma}_{\gamma\gamma|v} & \\ \boldsymbol{0} & & \boldsymbol{\Sigma}_{vv}^{-1} \end{pmatrix}.$$

Equation (22) implies for the single joint response regression of $\boldsymbol{X}_u$ on $\boldsymbol{X}_v$:

$$\boldsymbol{P}_{u|v} = -\boldsymbol{H}_{uu}^{-1}\boldsymbol{H}_{uv}, \quad \boldsymbol{\Sigma}_{uu|v} = \boldsymbol{H}_{uu}^{-1}\boldsymbol{W}_{uu}(\boldsymbol{H}_{uu}^{-1})^{\mathrm{T}}, \tag{23}$$

hence with equations (18) also simple matrix expressions for $\boldsymbol{\Sigma}^{uN}$ and $\boldsymbol{\Sigma}_{Nv}$.

The edge sets of G_{reg}^N are captured by edge matrices $\boldsymbol{\mathcal{H}}_{NN}$ and $\boldsymbol{\mathcal{W}}_{NN}$.

Definition 3 We denote the dimension of g_k by d_k, the one of v by d_v, so that $d = \sum_{k=1}^{k=K} d_k + d_v$ for the ordered node set $N = (1, \ldots, d)$. The edge matrix, $\boldsymbol{\mathcal{H}} = (\mathcal{H}_{ij})$, is upper block-triangular, with K identity matrices of size $d_k \times d_k$ along the diagonal and a symmetric edge matrix for the concentration graph of $\boldsymbol{X}_v$ alone in the last block. In the upper, off-diagonal parts are ones for arrows pointing in G_{reg}^N from $g_{>k}$ to g_k:

$$\mathcal{H}_{ij} = \begin{cases} 1 & \text{if and only if } i \prec\!\!\!- j \text{ or } i \text{ ——} j \text{ in } G_{\mathrm{reg}}^N \text{ or } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

The edge matrix, $\boldsymbol{\mathcal{W}}_{uu} = (\mathcal{W})_{ij}$, for dashed lines, is block-diagonal with K symmetric $d_k \times d_k$ edge matrices for covariance graphs of $\boldsymbol{X}_k$ given $\boldsymbol{X}_{>k}$:

$$\mathcal{W}_{ij} = \begin{cases} 1 & \text{if and only if } i\text{---}j \text{ in } G_{\mathrm{reg}}^N \text{ or } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

For $\boldsymbol{\mathcal{W}}_{NN}$, the last block is taken to be $\boldsymbol{\mathcal{W}}_{vv} = \boldsymbol{\mathcal{I}}_{vv}$ since the full line edges of the concentration graph of $\boldsymbol{X}_u$ are already captured by $\boldsymbol{\mathcal{H}}_{vv}$.

For nodes u as a single response node and v its regressors, one gets e.g.

$$\boldsymbol{\mathcal{P}}_{u|v} = \mathrm{In}[\boldsymbol{\mathcal{H}}_{uu}^{-}\boldsymbol{\mathcal{H}}_{uv}], \quad \boldsymbol{\mathcal{S}}_{uu|v} = \mathrm{In}[\boldsymbol{\mathcal{H}}_{uu}^{-}\boldsymbol{\mathcal{W}}_{uu}(\boldsymbol{\mathcal{H}}_{uu}^{-})^{\mathrm{T}}], \tag{26}$$

as the induced edge matrices for equation (23). Note that some induced edge matrices are denoted by using a close calligraphic equivalent to the parameters in the Gaussian case. These are then either edge matrices of a starting graph, or their submatrices, or they can be derived directly in terms of the matrix operator described next. All others get $\boldsymbol{\mathcal{N}}$ as notation.

Partial closure

We now let the ordered node set $N = \{1, \ldots, d\}$ denote also the rows and columns of an edge matrix $\mathcal{M}$ containing Vs of one type. Then, for an ordered partioning as $N = (a, b)$, the '**partial closure**' operator, denoted by $\mathrm{zer}_a\boldsymbol{\mathcal{M}}$, closes a-line paths in a corresponding graph with edge matrix $\mathcal{M}$ and finds structural zeros induced in a corresponding parameter matrix $\boldsymbol{M}$ of a Gaussian distribution.

Definition 4 The partial closure operator is, with $a = \{1\}$ and a square matrix $\mathbf{m}$ for

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} 1 & \mathbf{v}^{\mathrm{T}} \\ \mathbf{w} & \mathbf{m} \end{pmatrix} : \quad \mathrm{zer}_{\{1\}}\boldsymbol{\mathcal{M}} = \begin{pmatrix} 1 & \mathbf{v}^{\mathrm{T}} \\ \mathbf{w} & \mathrm{In}[\mathbf{m} + \mathbf{w}\mathbf{v}^{\mathrm{T}}] \end{pmatrix},$$

where $\mathrm{zer}_a\boldsymbol{\mathcal{M}}$, for t elements in a, may be thought of as applying the above operation t times, using repeatedly appropriate permutations of $\boldsymbol{\mathcal{M}}$.

An off-diagonal $i, j \neq k$ of $\mathrm{zer}_{\{k\}}\boldsymbol{\mathcal{M}}$ contains an additional one compared to $\boldsymbol{\mathcal{M}}$ if and only if $\mathcal{M}_{ij} = 0$ and $\mathcal{M}_{ik}\mathcal{M}_{kj} = 1$, hence indicating the presence of a V in the graph. The operator preserves all ones of $\boldsymbol{\mathcal{M}}$ and it closes paths with Vs which must be of the same type in the graph represented by $\boldsymbol{\mathcal{M}}$.

Proposition 6 Wermuth, Wiedenbeck and Cox, [114]. *Partial closure is commutative, cannot be undone and is exchangeable with taking submatrices.*

One may for instance get the edge matrix $\boldsymbol{\mathcal{A}}^{-}$, that is obtain the transitive closure of a directed acyclic graph, and $\boldsymbol{\mathcal{W}}^{-}$ of equation (13), that is complete all connected components of an undirected graph, with $N = \{a, b\}$ as

$$\mathrm{zer}_b\mathrm{zer}_a\boldsymbol{\mathcal{A}} = \mathrm{zer}_N\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}^{-}, \qquad \mathrm{zer}_b\mathrm{zer}_a\boldsymbol{\mathcal{W}} = \mathrm{zer}_N\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{W}}^{-}.$$

By Definition 1 and Prop. 6, $\mathrm{zer}_a\boldsymbol{\mathcal{A}}$ for G^N_{par} remains unit-upper triangular in the starting order, $\mathrm{zer}_a\boldsymbol{\mathcal{W}}$ for G^N_{con} remains symmetric and disconnected components, such as the graphs for conditional covariances of joint responses in G^N_{reg}, remain disconnected.

The edge matrices $(\boldsymbol{\mathcal{P}}_{u|v}, \boldsymbol{\mathcal{S}}_{uu|v})$ in equation (26), induced by G^N_{reg} for $f_{u|v}$ with a single joint response u, may be obtained with $[\mathrm{zer}_u\boldsymbol{\mathcal{H}}_{NN}]_{u,N}$. For $a = (\alpha, \gamma)$ and $b = (\beta, \delta)$, the edge matrix components for $f_{\alpha|\beta\delta}$ and $f_{\beta|\delta}$ as induced by G^N_{reg} with $f_{a|b}$ and f_b are given by the subgraph of $\alpha \cup \beta$, just as the Gaussian parameters in equation (20) are given by submatrices.

Algorithms for finding the transitive closure in directed graphs, possibly containing cycles, started to be developed independently in the Russian, French and American computer science literature; for a recent survey see [92]. Algorithms for finding connected components for general graphs, [88], are also still being developed, [69].

One advantage of partial closure is that its properties justify stepwise procedures using just the Vs in a G^N_{reg}. Another is that properties of this matrix operator prove some features of the regression graph transformations.

Edge matrices induced by G^N_{reg}

The edge matrix of the a-line anterior graph of G^N_{reg}, see Definition 2, arises for a any subset of N, $b = N \setminus a$ and a reordering as $N = (a, b)$ with:

$$\boldsymbol{K}_{NN} = \mathrm{zer}_a\boldsymbol{H}_{NN}. \tag{27}$$

This operation closes all full, a-line paths within v and for each i in u, it turns every a-line anterior j into a parent of i. Similarly,

$$\boldsymbol{\mathcal{V}}_{NN} = \mathrm{zer}_b \boldsymbol{\mathcal{W}}_{NN}, \tag{28}$$

closes all dashed, b-line paths in the conditional covariance graphs of the responses.

Thus, the two partial closure operations in equations (27), (28) close all of the following four types of Vs, where o_g denote nodes in a subset g of N:

$$i_u \prec\!\!-\mathrm{o}_a \prec\!\!- j_N, \quad i_u \prec\!\!-\mathrm{o}_a \!\!- j_v, \quad i_v \!\!-\mathrm{o}_a \!\!- j_v, \quad i_u \text{---} \mathrm{o}_b \text{---} j_u\,,$$

and the types of induced edge are as specified in equations (2), (12), (16) for o_a, a node to be marginalized over, and o_b, a node to be conditioned on. These induced edges preserve the ordered split of the nodes, $N = (u, v)$. The corresponding model may be interpreted as a covering model, one with fewer constraints than the reduced model specified by the generating G^N_{reg}, [16].

Four types of V remain to be closed for consequences of G^N_{reg} with $f_{u|v} f_v$ for $f_{a|b} f_b$:

$$i_a \prec\!\!-\mathrm{o}_a \text{---} j_a, \quad i_a \prec\!\!-\mathrm{o}_a \!\!\succ j_b \quad i_a \text{---} \mathrm{o}_b \prec\!\!- j_b, \quad i_b \!\!\succ \mathrm{o}_b \prec\!\!- j_b\,.$$

To achieve this, $\boldsymbol{\mathcal{V}}_{uu}$ is combined with $\boldsymbol{\mathcal{K}}_{vv}$ to give $\boldsymbol{\mathcal{Q}}_{NN}$:

$$\boldsymbol{\mathcal{Q}}_{uu} = \boldsymbol{\mathcal{V}}_{uu}, \quad \boldsymbol{\mathcal{Q}}_{vv} = \boldsymbol{\mathcal{K}}_{vv}, \quad \boldsymbol{\mathcal{Q}}_{uv} = \mathbf{0} \quad , \boldsymbol{\mathcal{Q}}_{vu} = \mathbf{0}. \tag{29}$$

Then, these remaining Vs are closed with the following edge matrix products:

$\mathrm{In}[\boldsymbol{\mathcal{K}}_{aa} \boldsymbol{\mathcal{Q}}_{aa} \boldsymbol{\mathcal{K}}^{\mathrm{T}}_{aa}]$ gives for $i_a \prec\!\!- k_a \text{---} l_a \!\!\succ j_a$ and $i_a \prec\!\!- k_a \!\!- l_a \!\!\succ j_a$ a complete covariance graph,

$\mathrm{In}[\boldsymbol{\mathcal{K}}_{aa} \boldsymbol{\mathcal{V}}_{ab} \boldsymbol{\mathcal{K}}_{bb}]$ gives for $i_a \prec\!\!- k_a \text{---} l_b \prec\!\!- j_b$ a complete graph of response nodes $\{i_a, k_a\}$ and regressor nodes $\{l_b, j_b\}$,

$\mathrm{In}[\boldsymbol{\mathcal{H}}^{\mathrm{T}}_{bb} \boldsymbol{\mathcal{V}}_{bb} \boldsymbol{\mathcal{H}}_{bb}]$ gives for $i_b \!\!\succ k_b \text{---} l_b \prec\!\!- j_b$ a complete concentration graph.

This leads to the edge matrix components, $\boldsymbol{\mathcal{N}}_{a|b}$, $\boldsymbol{\mathcal{N}}_{aa|b}$, of the graph for regressing $\boldsymbol{X}_a$ on $\boldsymbol{X}_b$ and, $\boldsymbol{\mathcal{N}}^{bb.a}$, for the concentration graph of $\boldsymbol{X}_b$, as induced by $\boldsymbol{G}^{\boldsymbol{N}}_{\mathbf{reg}}$; induced arrows point from regressor $\boldsymbol{X}_b$ to response $\boldsymbol{X}_a$.

Proposition 7 Wermuth, [100]. *Edge matrix components induced by G^N_{reg} for $N = (a, b)$, by marginalizing over any $a \subset N$, conditioning on $b = N \setminus a$, are*

$$\begin{aligned} \boldsymbol{\mathcal{N}}_{aa|b} &= \mathrm{In}[\boldsymbol{\mathcal{K}}_{aa} \boldsymbol{\mathcal{Q}}_{aa} \boldsymbol{\mathcal{K}}^{\mathrm{T}}_{aa}], \\ \boldsymbol{\mathcal{N}}_{a|b} &= \mathrm{In}[\boldsymbol{\mathcal{K}}_{ab} + \boldsymbol{\mathcal{K}}_{aa} \boldsymbol{\mathcal{V}}_{ab} \boldsymbol{\mathcal{K}}_{bb}], \\ \boldsymbol{\mathcal{N}}^{bb.a} &= \mathrm{In}[\boldsymbol{\mathcal{H}}^{\mathrm{T}}_{bb} \boldsymbol{\mathcal{V}}_{bb} \boldsymbol{\mathcal{H}}_{bb}]. \end{aligned}$$

The zeros in these induced edge matrices represent also the structural zeros in $\boldsymbol{\Sigma}_{aa|b}$, $\boldsymbol{\Pi}_{a|b}$, $\boldsymbol{\Sigma}^{bb.a}$. As in Definition 1 for $K = 1$, these ij-zeros mean:

$$i \perp\!\!\!\perp j | b \text{ in } \boldsymbol{\mathcal{N}}_{aa|b}, \quad i \perp\!\!\!\perp j | b \setminus j \text{ in } \boldsymbol{\mathcal{N}}_{a|b}, \quad i \perp\!\!\!\perp j | b \setminus \{i, j\} \text{ in } \boldsymbol{\mathcal{N}}^{bb.a}.$$

Example 1 For G^N_{par} of Fig. 5 with edge matrix $\boldsymbol{\mathcal{A}}$, marginalizing with an order-respecting split, $a = \{1,2,3\}$, and conditioning on $b = N \setminus a$ gives with $\boldsymbol{\mathcal{A}}^-_{aa} = \boldsymbol{\mathcal{K}}_{aa}$ and $\boldsymbol{\mathcal{K}}_{ab} = \text{In}[\boldsymbol{\mathcal{A}}^-_{aa}\boldsymbol{\mathcal{A}}_{ab}]$ a direct generalization of equation (9):

$$\boldsymbol{\mathcal{N}}_{aa|b} = \text{In}[\boldsymbol{\mathcal{A}}^-_{aa}(\boldsymbol{\mathcal{A}}^-_{aa})^{\text{T}}], \quad \boldsymbol{\mathcal{N}}_{a|b} = \boldsymbol{\mathcal{K}}_{ab}, \quad \boldsymbol{\mathcal{N}}^{bb.a} = \text{In}[\boldsymbol{\mathcal{A}}^{\text{T}}_{bb}\boldsymbol{\mathcal{A}}_{bb}].$$

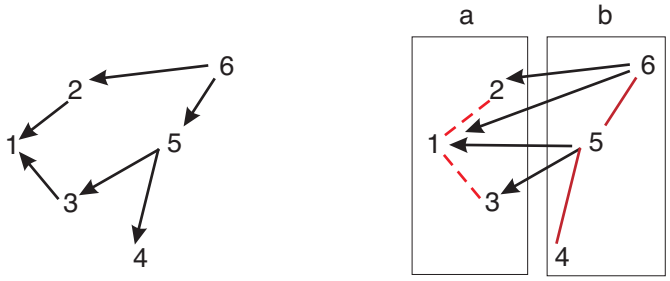


Figure 5: *Left: the generating parent graph, right: induced graph for one set of response nodes,* $a = \{1,2,3\}$*, and one set of regressor nodes* $b = \{4,5,6\}$*.*

An alternative to the edge matrix results is to use Corollary 4 to derive, separately for each missing edge in G^N_{par}, the consequences of the new conditioning sets specified by $N = (a,b)$ for the regression graph of Fig. 5, which has only one joint response.

Example 2 A generating G^N_{reg} for determinants of the well-being of diabetic patients having a lower level of formal schooling, Y, is given in Fig. 6 left. The following description of this graph attaches to it a plausible, substantive story. This uses statistical results, [18], not given here.

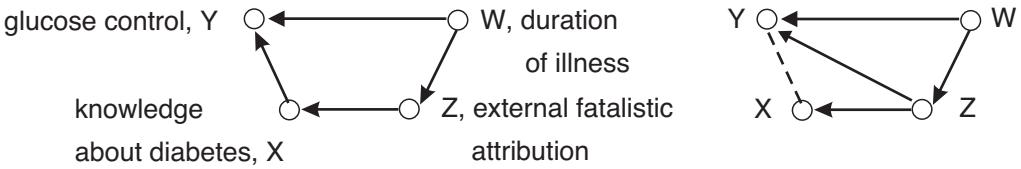


Figure 6: *Left: a generating* G^N_{par}*, right: the induced graph for regression of* (Y,X) *on* (W,Z)*;* $Y \prec\!\!-\!\!- Z$ *induced by marginalizing over* X *in the left* G^N_{reg}*.*

Glucose control improves, the more a patient knows about diabetes and the longer ago diabetes was diagnosed. Thus, glucose control depends directly on the knowledge about the illness, X, and on the time since the illness was diagnosed, W, hence $Y \not\perp\!\!\!\perp X|W$ and $Y \not\perp\!\!\!\perp W|X$. Knowledge, X, is better, the lower the external fatalistic attribution, Z, that is the less patients tend to think that their well-being depends mainly on their physicians, so that $X \not\perp\!\!\!\perp Z$. And, fatalistic attribution, Z, decreases with the time since diagnosis, W, so that $Z \not\perp\!\!\!\perp W$. This well-fitting graph contains the path (Y,X,Z,W). This path, together with the type of involved dependences, suggests that intervening on the variables along it may improve the well-being of diabetic patients.

By constructing induced graphs, one can answer queries like: which additional dependences result from a given generating process by using another type of process for the same variables? This may for instance arise in empirical studies when researchers

disagree on the ordering of the variables. In the example, $X \prec\!\!\!\leftarrow\!\!\!- W$ results with X as primary and Y as secondary response, due to conditioning on Y in the sink V , (X, Y, W), in the starting graph, while for Y, X as a joint response, the arrow $Y \prec\!\!\!\leftarrow\!\!\!- Z$ is added.

Edge criteria for effects of separation in $G^N_{\mathbf{reg}}$

By Corollary 1 and by $\boldsymbol{\mathcal{N}}_{a|b}$ representing the edge matrix induced by G^N_{reg} for the bipartite graph of arrows when $\boldsymbol{X}_a$ is regressed on $\boldsymbol{X}_b$, submatrices of the edge matrices in Prop. 7 give also the structural zeros induced by G^N_{reg} for the joint conditional distribution with density of $f_{\alpha\beta|c} = f_{\alpha|\beta c} f_{\alpha|c}$:

$$\boldsymbol{\mathcal{N}}_{\alpha|\beta.c} = [\boldsymbol{\mathcal{N}}_{a|b}]_{\alpha,\beta}, \qquad \boldsymbol{\mathcal{N}}_{\alpha\alpha|b} = [\boldsymbol{\mathcal{N}}_{aa|b}]_{\alpha,\alpha}, \qquad \boldsymbol{\mathcal{N}}^{\beta\beta.a} = [\boldsymbol{\mathcal{N}}^{bb.a}]_{\beta,\beta}.$$

Proposition 8 Wermuth, [100]. *A regression graph G^N_{reg} with edges given by Definition 1 implies $\alpha \perp\!\!\!\perp \beta|c$ if $\boldsymbol{\mathcal{N}}_{\alpha|\beta.c}=0$ and it implies $\alpha \not\perp\!\!\!\perp \beta|c$ if $\boldsymbol{\mathcal{N}}_{\alpha|\beta.c}\neq 0$.*

Thus, the absence of ones in a matrix indicates directly a queried independence and the presence of ones shows where dependencies occur. Instead, with any path criterion, one has to study the properties of paths before a decision can be reached. This may get cumbersome in large graphs when one has to check for each collision V whether its collision node is within the anterior set of c.

Properties of regression graphs

A regression graph, G^N_{reg}, shares the three properties in Prop. 1 of a joint Gaussian distribution generated over G^N_{reg}. It is dependence-inducing and independences combine downwards and upwards, that is it satisfied singleton transitivity, intersection and composition, in addition to the general properties of all probability distributions.

Its composition and intersection property have been proven in general, [75], and were discussed above for just three variables. Singleton transitivity requires an additional independence involving node h, say, if the conditioning set for independence of i, j includes and excludes h. For i, j, h distinct nodes of N and c a subset of $N \setminus \{i, j, h\}$

$$(i \perp\!\!\!\perp j|c \text{ and } i \perp\!\!\!\perp j|hc) \implies (i \perp\!\!\!\perp h|c \text{ or } j \perp\!\!\!\perp h|c). \tag{30}$$

The equivalent statement 'for ($i \not\perp\!\!\!\perp h|c$ and $j \not\perp\!\!\!\perp h|c$), either $i \perp\!\!\!\perp j|c$ can hold or $i \perp\!\!\!\perp j|hc$ but not both', was proven with equations (10) and (11) for two types of V in G^N_{par} .The same types of argument prove singleton transitivity of G^N_{reg}. Recall that traceable regressions satisfy these same three properties, hence '**mimic independence properties of a Gaussian distribution**' generated over G^N_{reg} .

For '**set transitivity**' as defined in the literature, the single node h in equation (30) is replaced by a subset of N; disjoint of $\{i, j, c\}$. Set transitivity is for instance violated when both of the independence structures hold which are defined with the concentration and with the covariance graph in Fig. 7.

This may happen in Gaussian distributions, [17], [100], but not for undirected graphs; see Corollary 2. More generally, since graphs induced by G^N_{reg} may be derived by partial closure and by adding products of binary matrices, contributions of several paths to a conditional dependence of any node pair i, j can never cancel out.

Figure 7: *Left: concentration graph with $i \perp\!\!\!\perp j|\{h,k\}$ and $h \perp\!\!\!\perp k|\{i,j\}$; right: covariance graph with $i \perp\!\!\!\perp j$ and $h \perp\!\!\!\perp k$; connector for i,j is $\{h,k\}$ in both.*

Proposition 9 Wermuth, [101]. *The structures captured and induced by $G^N_{\rm reg}$ are like traceable regressions with exclusively positive dependences.*

We show next how source, transition and sink Vs of $G^N_{\rm par}$ in Fig. 1 and equation (2) generalize to source, transition and sink Us in Fig. 8. By remembering the path ends for the four ij-paths in Fig. 8, induced are either a dashed ij-line, an ij-arrow or a full ij-line; see equations (2), (12), (16) for the involved, repeated closing of Vs:

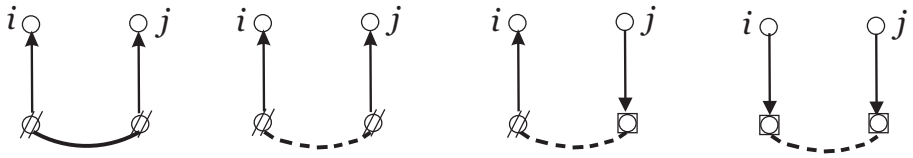


Figure 8: *Types of U with undirected edges and arrows to or from i,j. The first two on the left: source U; the third: transition U; the fourth on the right: sink U.*

We now let i and j be again an uncoupled node pair of N. Sets $\delta \neq \emptyset$ and c be disjoint subsets of $N \setminus \{i,j\}$. Then, δ is called a '**connector**' if $G^N_{\rm reg}$ implies ($i \perp\!\!\!\perp j|\delta c$ and $i \not\perp\!\!\!\perp j|c$) or ($i \perp\!\!\!\perp j|c$ and $i \not\perp\!\!\!\perp j|\delta c$) and the inner nodes of undirected ij-paths exhaust the nodes of δ. With this definition, a previous claim of set transitivity of $G^N_{\rm reg}$, [100], can be corrected as follows.

Proposition 10 Wermuth, [101]. *Regression graphs are connector-transitive, compositional graphoids.*

Equation (30) is changed into **connector transitivity** by replacing the single node h by a connector δ. Connector-transitivity extends singleton-transitivity. It concerns chordless cycles in undirected graphs; the simplest are in Fig. 7. Furthermore, it concerns Us with mixed edges inducing an undirected ij-edge. These have two incoming or two outgoing arrows at i,j and an undirected path via the nodes of δ. It is the exchangeability property of partial closure which permits one to argue by just using subgraphs.

Next, we describe the operator which corresponds closely to partial closure since it transforms parameter matrices for Gaussian distributions generated over $G^N_{\rm reg}$ in a similar way as partial closure modifies edge matrices.

Partial inversion

Let $N = (1,\ldots,d)$ denote the rows and columns of a real-valued matrix $\boldsymbol{M}$ having invertible leading principal submatrices, where $\boldsymbol{M}$ connects real-valued vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ as $\boldsymbol{M}\boldsymbol{x} = \boldsymbol{y}$. The '**partial inversion**' operator, denoted by $\mathrm{inv}_a\boldsymbol{M}$, for '$a$' any subset

of N and an ordering as $N = (a, b)$, exchanges argument and image relating to a, [114], [115], that is

$$\boldsymbol{M} \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_a \\ \boldsymbol{y}_b \end{pmatrix} \text{ is turned into: } \operatorname{inv}_a \boldsymbol{M} \begin{pmatrix} \boldsymbol{y}_a \\ \boldsymbol{x}_b \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_b \\ \boldsymbol{y}_b \end{pmatrix}. \tag{31}$$

Applied for instance to rows a of $\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = \boldsymbol{\zeta}$, two correlated sets of equations turn directly into two sets of orthogonalized equations; see equation (21).

Definition 5 With $a = \{1\}$, the partial inversion operator is for

$$\boldsymbol{M} = \begin{pmatrix} s & \boldsymbol{v}^{\mathrm{T}} \\ \boldsymbol{w} & \boldsymbol{m} \end{pmatrix} : \quad \operatorname{inv}_{\{1\}} \boldsymbol{M} = \begin{pmatrix} 1/s & -\boldsymbol{v}^{\mathrm{T}}/s \\ \boldsymbol{w}/s & \boldsymbol{m} - \boldsymbol{w}\boldsymbol{v}^{\mathrm{T}}/s \end{pmatrix},$$

where matrix $\boldsymbol{m} - \boldsymbol{w}\boldsymbol{v}^{\mathrm{T}}/s$ is a Schur complement, [79], and $\operatorname{inv}_a \boldsymbol{M}$ for t elements in a, may be thought of as applying the above operation t times, by using repeatedly appropriate permutations of $\boldsymbol{M}$.

Partial inversion, [114], generalizes the sweep operator, [22],[23], and other methods for Gaussian elimination, [36], to non-symmetric matrices. A small modification of the sweep operator leads to the '**symmetric difference**', so that an action on a, say, is undone by using this same operator again on a.

Proposition 11 Wermuth, Wiedenbeck and Cox, [114]. *Partial inversion is commutative, can be undone and is exchangeable with taking submatrices.*

In particular, the operator gives $\operatorname{inv}_b \boldsymbol{\Sigma} = -\operatorname{inv}_a \boldsymbol{\Sigma}^{-1}$ and the corresponding three Gaussian parameter matrices in equation (18). The Schur complements involved in the two operations, $\operatorname{inv}_b \boldsymbol{\Sigma}$ and $\operatorname{inv}_a \boldsymbol{\Sigma}^{-1}$, are matrix forms of the recursion relations in equation (6). A matrix form of the recursion relation for regression coefficients arises by partial inversion on v in the matrix example to equation (22).

By starting from a general regression graph model in equation (22), the parameter matrices $\boldsymbol{H}_{NN}$ and $\boldsymbol{W}_{NN}$ and $N = (u, v)$ are given. Parameter transformations that are analogous to those of the edge matrices of Prop. 7 have been derived using the partial inversion operator and sums of matrix products, [100]. In contrast to partial closure, partial inversion may lead to negative elements in the induced matrices and therefore permit path cancellations.

Some special aspects

For many regression graph models, the parameters in the regressions of $\boldsymbol{X}_k$ given the past $\boldsymbol{X}_{>k}$ can be estimated by using standard methods, [58], [95], [2], but some are based on special multivariate models, [35], [54], [72], [28] or on special features of the data, [27], [31], [10]. Possible shortcomings have been identified for some estimation methods, [57], and for some models, [71], [60]. New estimation results are needed for joint responses of both categorical and quantitative components; exceptions are CG-regressions, [47], [26].

Features of special models may give unexpected insights and often lead to simplified properties. For instance, parent graphs without any transition V, shown in Fig. 1, are lattice conditional independence models, [6]. For these models, G^N_{reg} coincides with the

ancestor graph. Hence, the separating paths of Prop. 4 apply directly to $G^N_{\rm reg}$. Parent graphs of exclusively source Vs are labelled trees, [11]. These have exactly one path connecting each node pair and $\alpha \perp\!\!\!\perp \beta | c$ if every path between α and β intersects c.

Parent graphs without any sink Vs are said to be decomposable. By Prop. 5, they are Markov equivalent to concentration graphs in the same node and edge set. Finding well-fitting models for them may often be based on small subsets of variables and, for judging their goodness of fit, re estimation of parameters may not be needed, [97], [85]. Complex properties of estimates, simplify for decomposable models as well, [21], [48]. Strong analogies to Gaussian models result for binary variables with special types of graph, [110], especially when their distributions are jointly symmetric, [111], [112].

For observational studies, it is of concern whether dependences can be well estimated when some variables are unobserved. As a first step, one needs to know, when the parameters of such models can be identified. Considerable progress has been made regarding this in the last years; see [80], [30], [86], [1].

Some regression graph models for symmetric binary variables

We now consider special models for symmetric binary variables, which compare most closely to Gaussian distributions generated over some regression graph for variables standardized to have mean zero and unit variance. The purpose is to illustrate for some binary distributions generated over simple Markov equivalent graphs that the corresponding models are also '**parameter equivalent**', that is there is a one-to-one relation between the parameters of two different models. This assures that the same transformation, which relates the parameters of the models, applies also to the maximum-likelihood estimates, [29], an important property that appears not to be shared by any of the more recently developed estimation methods.

The binary variables have levels $-1, 1$ and equal probabilities, $\frac{1}{2}$, allocated to each of its two levels. A consequence is that they have mean zero and unit variance by definition. Their covariance matrix $\mathbf{\Sigma}$ coincides therefore with their correlation matrix; it has elements $\sigma_{ij} = \rho_{ij}$ and $\sigma_{ii} = 1$.

Induced marginal and partial correlations are just as for Gaussian distributions generated over the same graph, but a zero partial correlation in an induced concentration graph need not correspond to an independence statement; for an example see [111] Appendix C, and see also [49].

For an ordered node set $N = (1, 2, 3, 4)$, we denote four symmetric binary variables by A, B, C, D, their respective levels by i, j, k, l, and abbreviate joint and conditional probabilities for instance as

$$f_{1234} = \pi^{ABCD}_{ijkl} = \Pr(A = i, B = j, C = k, D = l), \quad \pi^{ABC|D}_{ijk|l} = \pi_{ijkl} / \textstyle\sum_l \pi_{ijkl}.$$

Their joint distributions are generated over parent graphs with just main effects, for the complete parent graph as:

$$\begin{aligned} \pi^{A|BCD}_{i|jkl} &= \tfrac{1}{2}(1 + \eta_{12} ij + \eta_{13} ik + \eta_{14} il) \\ \pi^{B|CD}_{j|kl} &= \tfrac{1}{2}(1 + \eta_{23} jk + \eta_{24} jl) \\ \pi^{C|D}_{k|l} &= \tfrac{1}{2}(1 + \eta_{34} kl) \\ \pi^{D}_{l} &= \tfrac{1}{2}\,, \end{aligned} \tag{32}$$

with η's resulting from $\mathbf{\Sigma}$ just like linear least-squares regression coefficients:

$$\begin{array}{rcl} \eta_{34} & = & \rho_{34} \\ (\eta_{23} \quad \eta_{24}) & = & (\rho_{23} \quad \rho_{24})\mathbf{\Sigma}^{-1}_{\{>2\}\{>2\}} \\ (\eta_{12} \quad \eta_{13} \quad \eta_{14}) & = & (\rho_{12} \quad \rho_{13} \quad \rho_{14})\mathbf{\Sigma}^{-1}_{\{>1\}\{>1\}}\,, \end{array}$$

where the inverse of, say, a submatrix $\boldsymbol{M}_{aa}$ is written as $\boldsymbol{M}_{aa}^{-1}$.

This form of the η-parameters generalizes directly to $d > 4$ variables and stems from the close connection for binary variables between probabilities and expectations. For instance, by using equation (32)

$$E(B|C = k, D = l) = \eta_{23}k + \eta_{24}l,$$

and correlation coefficients are cross-sum differences in probabilities, [112], such as:

$$E(CD) = (\pi^{CD}_{11} + \pi^{CD}_{-1-1}) - (\pi^{CD}_{-11} + \pi^{CD}_{1-1}) = 2(\pi^{CD}_{11} - \pi^{CD}_{-11}) = \rho_{34}.$$

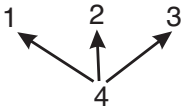
The second equality holds since in the generated distributions all odd order moments vanish so that there is also **joint symmetry**, [25], Appendix C. In this case, the probability of any level combination of these binary variables equals the probability of the level combination having each sign switched.

For binary variables in general, logit regressions, [2], are best suited to model conditional independence constraints. A logit regression is already close to a linear regression whenever the extreme events are not rare, but instead are more probable than say 0.1, [14]. It is in the special case of symmetric binary variables that the vanishing of linear regression coefficients in equation (32) coincides with the vanishing of logit regression coefficients in corresponding sequences of main-effect logit regressions. Thus, for instance,

$$1 \perp\!\!\!\perp 4|\{2,3\} \Leftrightarrow (\eta_{14} = 0), \qquad 2 \perp\!\!\!\perp 3|4 \Leftrightarrow (\eta_{23} = 0), \qquad 3 \perp\!\!\!\perp 4 \Leftrightarrow (\eta_{34} = 0)\,.$$

These binary distributions, generated over a given $G^N_{\rm par}$, have the edge matrix $\boldsymbol{A}$ of equation (8), the same triangular decompositions of $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$ as in equation (7), and the same induced covariance and concentration graphs as in equation (9), even though $\mathbf{\Delta}$ does not contain the conditional variances but, for $d > 2$, their expected values with respect to the past variables. We now turn to some Markov-equivalent regression graphs and models.

Example 3 The following graph captures **mutual conditional independence** of A, B, C given D for $(1, 2, 3, 4) = (A, B, C, D)$



For any type of distribution generated over this graph, the edge matrix $\boldsymbol{\mathcal{A}}$ is the binary matrix defined by equation (8) and the generated density is

$$f_{1234} = f_{1|4}f_{2|4}f_{3|4}f_4 \;\Leftrightarrow\; (1 \perp\!\!\!\perp 2 \perp\!\!\!\perp 3)|4\,.$$

Here, the four binary symmetric variables have the constraints $0 = \eta_{12} = \eta_{13} = \eta_{23}$ in equations (32). The triangular decomposition of $\mathbf{\Sigma}$, that is the matrix pair $(\boldsymbol{A}^{-1}, \mathbf{\Delta})$, leads to the special form of the correlation matrix with

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & -\rho_{14} \\ & 1 & 0 & -\rho_{24} \\ & & 1 & -\rho_{34} \\ \mathbf{0} & & & 1 \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{14}\rho_{24} & \rho_{14}\,\rho_{34} & \rho_{14} \\ . & 1 & \rho_{24}\rho_{34} & \rho_{24} \\ . & . & 1 & \rho_{34} \\ . & . & . & 1 \end{pmatrix}$$

and $\delta_{ss} = 1 - \rho_{s4}^2$ for $s = 1, 2, 3$, and $\delta_{44} = 1$. The induced correlations, corresponding to the three missing edges of the graph, are as specified for the outer nodes of a source V in equation (32). Here every ancestor is a parent, hence $\mathcal{A}^- = \mathcal{A}$, and equation (9) gives a complete induced covariance graph and an induced concentration graph with no additional edge.

Since the given $G^N_{\rm par}$ contains no collision V, it is Markov equivalent to $G^N_{\rm con}$ with the same node and edge set. The joint probabilities obtained from equations (32) show directly that the more important parameter equivalence holds in addition. Markov equivalence often implies parameter equivalence whenever a single parameter is attached to each edge present in $G^N_{\rm par}$.

Example 4 The example here is a **Markov chain** graph, which is a parent graph consisting of a single direction-preserving path of arrows, here:

$$1 \prec\!\!-\!\!- 2 \prec\!\!-\!\!- 3 \prec\!\!-\!\!- 4\,,$$

where each response node remembers from its past only the most recent node. For any type of distribution generated over this graph, the edge matrix $\boldsymbol{\mathcal{A}}$ is the binary matrix defined by equation (8) and the generated density is

$$f_{1234} = f_{1|2} f_{2|3} f_{3|4} f_4 \;\Leftrightarrow\; (1 \perp\!\!\!\perp \{3,4\}|2 \text{ and } 2 \perp\!\!\!\perp 4|3)\,.$$

For four binary symmetric variables and constraints $0 = \eta_{13} = \eta_{14} = \eta_{24}$ in equations (32), the triangular decomposition of $\mathbf{\Sigma}$, the matrix pair $(\boldsymbol{A}^{-1}, \mathbf{\Delta})$ gives the special form of the correlation matrix with

$$\boldsymbol{A} = \begin{pmatrix} 1 & -\rho_{12} & 0 & 0 \\ & 1 & -\rho_{23} & 0 \\ & & 1 & -\rho_{34} \\ \mathbf{0} & & & 1 \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{12} & \rho_{12}\,\rho_{23} & \rho_{12}\,\rho_{23}\,\rho_{34} \\ . & 1 & \rho_{23} & \rho_{23}\,\rho_{34} \\ . & . & 1 & \rho_{34} \\ . & . & . & 1 \end{pmatrix},$$

and $\delta_{ss} = 1 - \rho_{s,s+1}^2$ for $s = 1, 2, 3$ and $\delta_{44} = 1$. The correlation induced for each missing edge in $G^N_{\rm par}$ equals the product of the correlations along the path connecting the uncoupled node pair. Here, every node in the past of i is an ancestor of i, hence the ancestor graph with edge matrix $\boldsymbol{\mathcal{A}}^-$ is complete. Consequently, the induced covariance graph is also complete.

As in Example 3, equation (9) gives an induced concentration graph with no additional edge. Here, $G^N_{\rm par}$ is Markov equivalent to a $G^N_{\rm con}$ which is a **concentration chain** in nodes $(1, 2, 3, 4)$:

$$1 \text{ —— } 2 \text{ —— } 3 \text{ —— } 4,$$

where each edge present means $i \pitchfork j | N \setminus \{i,j\}$. Also as in Example 1, there is parameter equivalence obtained from equation (32) to the parameters in the joint distribution:

$$\pi_{ijkl}^{ABCD} = \tfrac{1}{16}\{(1+\rho_{12}ij)(1+\rho_{23}jk)(1+\rho_{34}kl)\,.$$

Example 5 This last example is quite different from the previous one. It is a **covariance chain** in nodes $(1,2,3,4)$:

$$1\text{---}2\text{---}3\text{---}4\,,$$

where each ij-edge present represents in general $i \pitchfork j$. For the symmetric binary variables, the dependence is captured by the marginal correlation coefficient, $\rho_{ij} \neq 0$. The simplifying independences are in $\boldsymbol{A}^{-1}$ and in $\boldsymbol{\Sigma}$ but there are none for the joint distribution generated over a parent graph with node ordering $(1,2,3,4)$. Accordingly, the factorizations and the independence structure are

$$(f_{124} = f_{12}f_4 \text{ and } f_{134} = f_1 f_{34}) \Leftrightarrow (\{1,2\} \perp\!\!\!\perp 4 \text{ and } 1 \perp\!\!\!\perp \{3,4\})$$

Since dashed-line Vs are edge-inducing by conditioning, induced regression coefficients appear in $\boldsymbol{A}$, where $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$,

$$\boldsymbol{A} = \begin{pmatrix} 1 & -\eta_{12} & \eta_{12}\,\eta_{23} & -\eta_{12}\,\eta_{23}\,\eta_{34} \\ & 1 & -\eta_{23} & \eta_{23}\,\eta_{34} \\ & & 1 & -\eta_{34} \\ \boldsymbol{0} & & & 1 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho_{12} & 0 & 0 \\ . & 1 & \rho_{23} & 0 \\ . & . & 1 & \rho_{34} \\ . & . & . & 1, \end{pmatrix} \quad . \tag{33}$$

Since there are no vanishing regression coefficients, there are also no independences of the type $i \perp\!\!\!\perp j | N \setminus \{i,j\}$: hence the induced concentration graph is complete.

There can be sign changes for induced coefficients, for instance $a_{24} = -a_{23}a_{34}$. Separate estimation of the parameters in the regressions for responses $i < 3$ is not feasible for this model, since some of the regression coefficients depend on coefficients in the past of node i.

However by Prop. 5, the given covariance chain is Markov equivalent to the following regression graph

$$1 \longrightarrow 2\text{---}3 \prec\!\!\!\!- 4\,,$$

which represents, for Gaussian distributions, the simplest type of Zellner's, [121], [24], seemingly unrelated regression. After reordering to $(2,3,1,4)$, the covariance matrix here becomes $\boldsymbol{\Sigma}'$ while partial inversion on the regressors $1,4$ gives the parameters for the joint response regression of $\boldsymbol{Y}_a$ on $\boldsymbol{Y}_b$, where $a = \{2,3\}$ on $b = \{1,4\}$

$$\boldsymbol{\Sigma}' = \begin{pmatrix} 1 & \rho_{23} & \rho_{12} & 0 \\ . & 1 & 0 & \rho_{34} \\ . & . & 1 & 0 \\ . & . & . & 1 \end{pmatrix} \quad \mathrm{inv}_{3,4}\,\boldsymbol{\Sigma}' = \begin{pmatrix} 1-\rho_{12}^2 & \rho_{23} & \rho_{12} & 0 \\ . & 1-\rho_{34}^2 & 0 & \rho_{34} \\ \sim & \sim & 1 & 0 \\ \sim & \sim & . & 1 \end{pmatrix}.$$

The $\sim$ notation denotes entries that are symmetric up to the sign. This parametrization is equivalent to requesting

$$2 \pitchfork 3 | \{1,4\}, \quad 2 \pitchfork 1 | 4, \quad 2 \perp\!\!\!\perp 4 | 1, \quad 3 \perp\!\!\!\perp 1 | 4, \quad 3 \pitchfork 4 | 1, \quad 1 \perp\!\!\!\perp 4.$$

It leads to joint probabilities which become, see [111] Appendix A:

$$\pi^{BCAD}_{jkil} = \tfrac{1}{16}(1 + \rho_{12}ij + \rho_{23}jk + \rho_{34}kl + \rho_{12}\rho_{34}ijkl)\,,$$

so that there is parameter equivalence in spite of a four-factor interaction.

For $N = (2, 3, 1, 4)$, the three types of edge sets, $E_{--}, E_{\leftarrow}, E_{—}$ are captured by

$$\boldsymbol{\mathcal{H}} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\mathcal{W}} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Here, $\boldsymbol{\mathcal{E}} = \mathrm{In}(\boldsymbol{\mathcal{W}} + \boldsymbol{\mathcal{H}} + \boldsymbol{\mathcal{H}}^{\mathrm{T}})$ is the generating edge matrix of $\boldsymbol{\Sigma}$, due to Markov equivalence. With it, the edge matrix of the induced concentration graph becomes, by equation (13), $\boldsymbol{S}^{NN} = \mathrm{In}(\boldsymbol{L}^{-1})$, where $\boldsymbol{L} = 5\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{E}}$. This is just one way to see here that the induced concentration graph is complete.

These last three examples show that independences may mean simple zero constraints on parameters of one model but may appear as complex constraints, even in parameter equivalent models. It is therefore in general rarely useful to restrict model search and data analysis to one particular class of models. Strong prior knowledge would be an exception. Even then, using Markov- and parameter-equivalence may aid in finding alternative interpretations and alternative fitting algorithms.

If the motivation for designing an empirical study are causal hypotheses, then undirected graphical models alone are typically of little interest. But similarly, directed acyclic graph models are of little help when one expects that an intervention will lead to changes in several connected responses at the same time. For instance, when effects of a drug to reduce blood pressure are to be studied, this intervention will affect systolic and diastolic blood pressure simultaneously and not one before the other.

Discussion

It took nearly 40 years of research until the present form of the regression graph, G^N_{reg}, was defined and its properties and consequences could be studied. The graph represents ordered sequences of joint response regressions. Responses may depend on all or on only some of the variables in their past. The graph contains three types of edge, one undirected type for dependences among responses, another undirected type for dependences among context variables and directed edges pointing to a response from nodes in its past. Conditional dependences show in edges present in G^N_{reg} .These dependences simplify the more conditional independence constraints there are, that is with more missing edges.

To make the graphs useful tools for tracing developmental pathways and for predicting structure in alternative models, the generated distributions have to mimic some properties of joint Gaussian distributions, see Prop. 1. If in G^N_{reg}, independences did not combine downwards and upwards, that is if the intersection and the composition properties were not satisfied, it would be impossible to infer mutual independence of disconnected subgraphs. Then, the graphical representations would be nearly useless.

If regression graphs were not, in addition, singleton-transitive, then they would not even well represent Gaussian distributions which have this property and are the simplest and most studied types of joint distribution. Also, edges present in induced graphs would

not point to non-vanishing conditional dependences in traceable regressions. But this is a prerequisite for useful tracings of pathways of development in the graphs.

Connector-set transitivity will illuminate the distinction between structural independences and those that may result due to special parametric constellations. The distinction between the two types reflects a long-standing practice in empirical research. Whenever a result has been replicated in several studies under essentially the same conditions, one typically still wants to establish it under modified conditions.

Even when all edges present in a graph correspond to positive dependences, negative linear dependences are induced by closing collision Vs; see equation (5) and equation (33). Some first results for preserving positive dependences have been obtained, [52], [40], others are expected for totally positive distributions generated over $G^N_{\rm con}$ and for decomposable regression graphs, those without any collision V.

Among the open theoretical questions are the following: Can necessary and sufficient conditions be derived for the properties of traceable regressions, such as those for the intersection property, [78]? For this, can methods of algebraic statistics also be helpful, such as those for binary tree models, [122]? How will independence structures and their properties change when graphs are no longer finite, [59]? When may models with fewer independence constraints, that is with more edges in the graph, be safely used as covering models, [16], for simpler estimation and useful interpretation?

At least equally important are further direct applications of traceable regressions; for a summary of the related tasks and links to detailed reports on finding well-fitting models in different research contexts see [106]. With traceable regressions, it has become feasible, for the first time, to derive the structural consequences of (1) ignoring some of the variables, of (2) selecting subpopulations via fixed levels of some other variables or of (3) changing the order in which the variables might get generated. With the currently used methods for combining results from empirical studies, called '**meta-analyses**', such effects are not taken care off. Therefore, the, most important future applications of these models will aim at the best possible integration of knowledge from related studies.

References

[1] Allman, E.S., Rhodes, J.A., Stanghellini, E. and Valtorta, M. (2015). Parameter identifiability of discrete Bayesian networks with hidden variables. *J. Causal Inference*, to appear.

[2] Andersen P.K. and Skovgaard L.T. (2010). *Regression with linear predictors.* Springer, New York.

[3] Anderson, T.W. (1958). *An introduction to multivariate statistical analysis.* (3rd ed., 2003) Wiley, New York.

[4] Anderson, T.W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1**, 135 – 141.

[5] Andersson, S. A., Madigan, D. and Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.* **28**, 33–85.

[6] Andersson, S.A., Madigan, D., Perlman, M.D. and Triggs, C.M. (1997). A graphical characterization of lattice conditional independence models. *Ann. Math. Artific. Intellig.* **21**, 27–50

[7] Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, Chichester.

[8] Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. B* **25**, 220 – 233.

[9] Bollen, K.A. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

[10] Castelo, R. and Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Research* **7**, 2621–2650.

[11] Castelo, R. and Siebes, A. (2003). A characterization of moral transitive acyclic directed graph Markov models as labeled trees. *J. Stat. Plan. Inf.* **115**, 235–259.

[12] Chaudhuri, S., Drton, M. and Richardson, T.S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* **94**, 199–216.

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

[14] Cox, D.R. (1966). Some procedures connected with the logistic qualitative response curve. In (F.N. David, ed.) *Research Papers in Statistics: Essays in Honour of J. Neyman's 70th Birthday.* Wiley, London, 55 – 71.

[15] Cox, D.R.(2006). *Principles of statistical inference.* Cambridge University Press, Cambridge.

[16] Cox, D.R. and Wermuth, N. (1990). An approximation to maximum-likelihood estimates in reduced models. *Biometrika* **77**, 747–761.

[17] Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Science* **8**, 204 – 218, 247 – 277.

[18] Cox, D.R. and N. Wermuth (1996). *Multivariate Dependencies: Models, Analysis, and Interpretation.* Chapman and Hall, London.

[19] Darroch, J.N., Lauritzen, S.L., and Speed, T.P. (1980). Markov fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522 – 539.

[20] Dawid, A.P. (1979). Conditional independence in statistical theory. *J. Roy. Statist. Soc. B* **41**, 1 – 31.

[21] Dawid, A.P. and Lauritzen, S.L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–1317.

[22] Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading, Mass.

[23] Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157 – 175.

[24] Drton, M. (2009). Discrete chain graph models. *Bernoulli* **5**, 736–753.

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

[26] Edwards, D. and Lauritzen, S.L. (2001). The TM algorithm for maximising a conditional likelihood function. *Biometrika* **88**, 961–972.

[27] Eichler, M, Dahlhaus R. and Sandkühler J. (2003). *Partial correlation analysis for the identification of synaptic connections.* Biological Cybernetics **89**, 289–302.

[28] Evans, R. and Forcina, A. (2013). Two algorithms for fitting constrained marginal models. *Comput Stat Data Anal.* **66**, 1–7.

[29] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222**, 309–368.

[30] Foygel, R., Draisma J. and Drton, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. Ann. Statist. **40**, 1682–1713.

[31] Fried R. and Didelez, V. (2003). Decomposability and selection of graphical models for time series. *Biometrika* **90**, 251–267.

[32] Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. *Ann. Statist.* **18**, 790–805.

[33] Geiger, D., Verma, T.S. and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks* **20**, 507–534.

[34] Gibbs, W. (1902). *Elementary Principles of Statistical Mechanics.* Yale University Press, New Haven.

[35] Glonek G.F.V. and McCullagh P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc. B* **53**, 533–546.

[36] Grcar, J. F. (2011). Mathematicians of Gaussian elimination. *Notices Amer. Math. Soc.* **58**, 782–792.

[37] Green, P.J., Hjort, N. and Richardson, S. (eds.) (2002) *Models for Highly Structured Stochastic Systems.* Oxford University Press, Oxford.

[38] Haberman, S. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5**, 815–841.

[39] Højsgaard, S.D. Edwards, D. and Lauritzen, S.L. (2012). *Graphical Models with R.* Springer, Berlin, Heidelberg, New York.

[40] Jiang, Z., Ding, P. and Geng, Z. (2015). Qualitative evaluation of associations by the transitivity of the association signs. *Statistica Sinica*, **25**. doi:10.5705/ss.2013.095, also under arXiv:1405.4258.

[41] Jöreskog, K.G. (1981). Analysis of covariance structures. *Scand. J. Statistics* **8**, 65–92.

[42] Kauermann, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.* **23**, 105 – 116.

[43] Khare, K. and Rajaratnam, B. (2011). Wishart distributions for covariance graph models. *Ann. Statist* **39**, 514–555.

[44] Koster, J. (2002). Marginalising and conditioning in graphical models. *Bernoulli* **8**, 817–840.

[45] Lauritzen, S.L. (1996). *Graphical Models.* Oxford University Press, Oxford.

[46] Lauritzen, S.L., Dawid, A.P., Larsen, B., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491 – 505.

[47] Lauritzen, S.L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31 – 54.

[48] Letac, G, and Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35**, 1278–1323.

[49] Loh, P. and Wainwright, M.J. (2013). Structure estimation for discrete graphical models: generalized covariance matrices and their inverses. *Ann. Statist.* **41**, 3022–3049.

[50] Lněnička, R. and Matúš, F. (2007). On Gaussian conditional independence structures. *Kybernetika* **43**, 323–342.

[51] Lupparelli M., Marchetti, G.M. and Bergsma, W.P. (2009). Parameterization and fitting of discrete bi-directed graph models. *Scand. J. Statist.* **36**, 559–576.

[52] Ma, Z.M., Xie, X.C. and Geng, Z. (2006). Collapsibility of distribution dependence. *J. Roy. Statist. Soc. B* **68**, 127–133.

[53] Mabry, R. (1999). Proof without words. (Thirds of a triangle) *Math. Magaz.* **72**, 63.

[54] Marchetti, G.M. and Lupparelli, M. (2011). Chain graph models of multivariate regression type for categorical data. *Bernoulli* **17**, 845–879.

[55] Marchetti, G.M. and Wermuth, N. (2009). Matrix representations and independencies in directed acyclic graphs. *Ann. Statist.* **47**, 961–978.

[56] Markov, A.A. (1912). *Wahrscheinlichkeitsrechnung.* Teubner, Leipzig. (German translation of 2nd Russian ed., 1908).

[57] McCullagh, P. (2008). Sampling bias and logistic models. (2008). *J. Roy. Statist. Soc. B* **70**, 643–677.

[58] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

[59] Montague, D. and Rajaratnam, B. (2015). Graphical Markov models for infinitely many variables; under arXiv:1501.07878.

[60] Nemeth, R. and Rudas, T. (2013). On the application of discrete marginal graphical models. *Sociol. Methodology* **43**, 70–100.

[61] Neumann, C.G. (1884). *Vorlesungen über Riemann'sche Theorie der Abel'schen Integrale*, 2nd ed., Teubner, Leipzig.

[62] Oliver, R.E. and Smith, J.Q. (1990). *Influence Diagrams, Belief Nets and Decision Analysis*. Wiley, London.

[63] Ostrowski, A. (1937). Über die Determinanten mit überwiegender Hauptdiagonale. *Commentarii Mathematici Helvetici* **10**, 69–96.

[64] Ostrowski, A. (1956). Determinanten mit überwiegender Hauptdiagonale und die absolute Konvergenz von linearen Iterationsprozessen. *Commentarii Mathematici Helvetici* **29** 175– 210.

[65] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California.

[66] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.* 2nd ed., Cambridge University Press, New York.

[67] Pearl J. and Paz, A. (1987). Graphoids: a graph based logic for reasoning about relevancy relations. In: (eds. B.D. Boulay D. Hogg, and L. Steel) *Advances in Artificial Intelligence II*, North Holland, Amsterdam, 357–363.

[68] Pearl, J. and Wermuth, N. (1994). When can association graphs admit a causal interpretation? In: (eds. P. Cheeseman and W. Oldford) *Models and data, artificial intelligence and statistics IV.* Springer, New York, 205–214.

[69] Reingold, O. (2008). Undirected connectivity in log-space. *J. ACM* **55**, Art 17.

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

[71] Robins, J.M., Scheines, R., Spirtes, P. and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika* **90**, 491–515.

[72] Roverato, A., Lupparelli, M. and La Rocca, L. (2013). Log-mean linear models for binary data. *Biometrika*, **100**, 485–494.

[73] Sadeghi, K. (2009). Representing modified independence structures. *Transfer thesis, Oxford University.*

[74] Sadeghi, K. (2013). Stable mixed graphs. *Bernoulli* **19**, 2330–2358.

[75] Sadeghi, K. and Lauritzen, S.L. (2014). Markov properties for mixed graphs. *Bernoulli* **20**, 676–696.

[76] Sadeghi K. and Marchetti, G.M. (2012). Graphical Markov models with mixed graphs in R. *The R Journal* **4**, 65–73.

[77] Sadeghi, K. and Wermuth, N. (2015). Pairwise Markov properties for regression graphs. In preparation.

[78] San Martin E., Mochart M. and Rolin, J.M. (2005). Ignorable common information, null sets and Basu's first theorem. *Sankhya* **67**, 674–698.

[79] Schur, J. (1917). Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *J. Reine Angew. Mathem.* **147**, 205–232.

[80] Shpitser, I. and Tian, J. (2010). On identifying causal effects. In: (eds. Dechter, R., Geffner, H. and Halpern, J. Y.) Heuristics, Probability, and Causality: A Tribute to Judea Pearl. College Publications, London.

[81] Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc., B* **13**, 238–241.

[82] Speed, T.P. (1979). A note on nearest-neighbour Gibbs and Markov distributions over graphs. *Sankhya A* **41**, 184 – 197.

[83] Speed, T.P. and Kiiveri, H.T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138 – 150.

[84] Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search.* Springer-Verlag, New York (2nd edition (2001), MIT Press, Cambridge).

[85] Sundberg, R. (1975). Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. *Scand. J. Statist.* **2**, 71 – 79.

[86] Stanghellini, E. and Vantaggi, B. (2013). Identification of discrete concentration graph models with one hidden binary variable. *Bernoulli* **19**, 1920–1937.

[87] Studený, M. (2005). *Probabilistic Conditional Independence Structures*. Springer, London.

[88] Tarjan. R.E. (1972). Depth-first search and linear graph algorithms. *SIAM J. Computing* **1**, 146–160.

[89] Tukey, J.W. (1954). Causation, regression, and path analysis. In: (O. Kempthorne et al., eds.) *Statistics and Mathematics in Biology*, Iowa State College Press, Ames, 35 – 66.

[90] Tikhonov, A.N., (1963). Resolution of ill-posed problems and the regularization method (in Russian), Doklady Akademii Nauk SSSR, 151, 501–504. Translated in Soviet Mathematics **4**, 1035–1038.

[91] Uhler, C., Raskutti, G., Bühlmann, P. and Yu, B. (2013). Geometry of faithfulness assumption in causal inference. *Ann. Statist.* **41**, 436–463.

[92] van Schaik, S.J. (2010). Answering reachability queries on large directed graphs. *Thesis, Utrecht University*.

[93] Wainer, H. (1983). Book review of 'Social Indicators III: selected data on social conditions and trends in the United States'. *J. Americ. Statist. Assoc.* **78**, 492–496.

[94] Wainwright, M.J. and Jordan, M.I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**, 1– 305.

[95] Weisberg, S. (2014). *Applied Linear Regression*, 4th edition, Wiley, Hoboken, New Jersey.

[96] Wermuth, N. (1976a). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95 – 108.

[97] Wermuth, N. (1976b). Model search among multiplicative models. *Biometrics* **32**, 253–263.

[98] Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Ass.* **75**, 963 – 997.

[99] Wermuth, N. (2011). Probability models with summary graph structure. *Bernoulli* **17**, 845-879.

[100] Wermuth, N. (2012). Traceable regressions. *Internat. Statist. Review* **80**, 415–438.

[101] Wermuth, N. (2015). On properties of regression graphs and models. *Observational Studies* **1**. In preparation.

[102] Wermuth, N. and Cox, D.R. (1998). On association models defined over independence graphs. *Bernoulli* **4**, 477–495.

[103] Wermuth, N. and Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J. Roy. Stat. Soc. B* **66**, 687–717.

[104] Wermuth, N. and Cox, D.R. (2008). Distortions of effects caused by indirect confounding. *Biometrika* **95**, 17–33.

[105] Wermuth, N. and Cox, D.R. (2013). Concepts and a case study for a flexible class of graphical Markov models. In: (Becker, C., Fried, R. and Kuhnt, S. eds.) *Robustness and Complex Data Structures. Festschrift in Honour of Ursula Gather*. Springer, Heidelberg, 327– 347; also under arXiv 1303.1436

[106] Wermuth, N. and Cox, D.R. (2015). Graphical Markov models: overview. In: (Wright, J. ed.) *Internat. Encyc. Social Behav. Sciences*, 2nd ed. Elesevier, Amsterdam, 341–350, also under arXiv 1407.7783.

[107] Wermuth, N., Cox, D.R. and Marchetti, G.M. (2006). Covariance chains. *Bernoulli* **12**, 841-862.

[108] Wermuth, N. and Lauritzen, S.L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537–552.

[109] Wermuth, N. and Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. B* **52**, 21 – 72.

[110] Wermuth, N. and Marchetti, G.M. (2014). Star graphs induce tetrad correlation for Gaussian as well as for binary distributions. *Electr. J. Statist.* **8**, 253–273, also on arXiv: 1307.5396.

[111] Wermuth, N., Marchetti, G.M. and Cox, D.R. (2009). Triangular systems for symmetric binary variables. *Electr. J. Statist.* **3**, 932–955.

[112] Wermuth, N., Marchetti, G.M. and Zwiernik, P. (2014). Binary distributions of concentric rings. *J. Multiv. Analysis* **130**, 252–260; also on arXiv: 1311.5655

[113] Wermuth N. and Sadeghi, K. (2012). Sequences of regressions and their independences (with discussion). *TEST* **21**, 215–279. Also under arXiv:1103.2523.

[114] Wermuth, N., Wiedenbeck, M. and Cox, D.R. (2006). Partial inversion for linear systems and partial closure of independence graphs. *BIT, Numerical Math.* **46**, 883–901.

[115] Wiedenbeck, M. and Wermuth, N. (2010). Changing parameters by partial mappings. *Statistica Sinica* **20**, 823–836.

[116] Whittaker, J. L. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

[117] Strotz, R. H. and Wold, H. O. A. (1960). Recursive vs. nonrecursive systems: an attempt at synthesis. *Econometrica* **28**, 417–427.

[118] Won, J., Kim, S.J., Lim, J. and Rajaratnam, B. (2013). Condition number-regularized covariance estimation. *J. Roy. Statist. Soc. B* **75**, 427–450.

[119] Wright, S. (1923). The theory of path coefficients: a reply to Niles' criticism. *Genetics*, **8**, 239 – 255.

[120] Wright, S. (1934). The method of path coefficients. *Ann. Math. Statist.*, **5**, 161 – 215.

[121] Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57**, 348–368.

[122] Zwiernik, P. and Smith, J. Q. (2011). Implicit inequality constraints in a binary tree model. *Electr. J. Statist.* **5**, 1276–1312.