Test
DOI 10.1007/s11749-012-0290-6

INVITED PAPER

Sequences of regressions and their independences

Nanny Wermuth · Kayvan Sadeghi

Received: 2 March 2011 / Accepted: 9 March 2012

Abstract Ordered sequences of univariate or multivariate regressions provide statistical models for analysing data from randomized, possibly sequential interventions, from cohort or multi-wave panel studies, but also from cross-sectional or retrospective studies. Conditional independences are captured by what we name regression graphs, provided the generated distribution shares some properties with a joint Gaussian distribution. Regression graphs extend purely directed, acyclic graphs by two types of undirected graph, one type for components of joint responses and the other for components of the context vector variable. We review the special features and the history of regression graphs, prove criteria for Markov equivalence and discuss the notion of a simpler statistical covering model. Knowledge of Markov equivalence provides alternative interpretations of a given sequence of regressions, is essential for machine learning strategies and permits to use the simple graphical criteria of regression graphs on graphs for which the corresponding criteria are in general more complex. Under the known conditions that a Markov equivalent directed acyclic graph exists for any given regression graph, we give a polynomial time algorithm to find one such graph.

Communicated by Domingo Morales.

This invited paper is discussed in the comments available at doi:10.1007/s11749-012-0288-0, doi:10.1007/s11749-012-0287-1, doi:10.1007/s11749-012-0286-2, doi:10.1007/s11749-012-0285-3, doi:10.1007/s11749-012-0284-4.

N. Wermuth (✉)
Department of Mathematics, Chalmers Technical University, Gothenburg, Sweden
e-mail: wermuth@chalmers.se

N. Wermuth
International Agency of Research on Cancer, Lyon, France

K. Sadeghi
Department of Statistics, University of Oxford, Oxford, UK
e-mail: kayvan.sadeghi@jesus.ox.ac.uk

Published online: 02 May 2012

Keywords Chain graphs · Concentration graphs · Covariance graphs · Graphical Markov models · Independence graphs · Intervention models · Labelled trees · Lattice conditional independence models · Structural equation models

Mathematics Subject Classification Primary 62H99 · Secondary 62A99

1 Introduction

A common framework to model, analyse and interpret data for several, partially ordered joint or single responses is a sequence of multivariate or univariate regressions where the responses may be continuous or discrete or of both types. Each response is to be generated by a set of its regressors, called its *directly explanatory variables*. Based on prior knowledge or on statistical analysis, one is to decide which of the variables in a set of potentially explanatory ones are needed for the generating process. Thus, for each response, a first ordering determines what is potentially explanatory, named the past of the response, and what can never be directly explanatory, named the future. Furthermore, no variable is taken to be explanatory for itself.

Corresponding *regression graphs* consist of *nodes* and of *edges coupling distinct nodes*. The *nodes represent the variables* and the *edges stand for conditional dependences*, directed or undirected. The directly explanatory variables for an individual response variable Y_i show in the graph as the set of nodes from which arrows start and point to node i. These nodes are commonly named the *parents of node i*.

Every missing edge corresponds to a conditional independence statement. Edges are *arrows for directed dependences* and *lines for undirected dependences* among *variables on equal standing*, that is, among components of joint responses or of context variables. Undirected dependences are often also called associations. A given regression graph reflects a particular type of study which may be a simple experiment, a more complex sequence of interventions or an observational study.

One of the common features of pure experiments and of sequences of interventions with randomized, proportional allocation of individuals to treatments, is that, by study design, some variables can be regarded to act just like independent random variables. For instance, in an experiment with proportional numbers of individuals assigned randomly to each level combination of several experimental conditions, the set of explanatory variables contains no edge in the corresponding regression graph, reflecting a situation like mutual independence. Similarly, with fully randomized interventions, each treatment variable has exclusively arrows starting from its node but no incoming arrow. After statistical analysis, some conditional independences may be appropriate additional simplifications which show as further missing edges.

Sequences of interventions give a time-ordering for some of the variables. A time order is also present in cohort or multi-wave panel studies and in retrospective studies which focus on investigating effects of variables at one fixed time point in the past, without the chance of intervening. By contrast, in a strictly cross-sectional study, in which observations for all variables are obtained at the same time, any particular variable ordering is only assumed rather than implied by actual time.

The node set is at the planning stage of empirical studies partitioned into ordered sequences of single or joint responses, $Y_a, Y_b, Y_c, \ldots$, that we call *blocks of variables*

on equal standing and draw them in figures as *boxes*. This determines for the following statistical analyses that within each block there are undirected edges and between blocks there are directed edges, the arrows. The first block on the left contains the *primary responses* of Y_a and the last block on the right contains *context variables, also named the background variables*. After statistical analyses, arrows may start from nodes within any block but always end at a node in one of the blocks in the future. Thus, there are no arrows pointing to context variables and all arrows point in the same direction, from right to left. An *intermediate variable* is a response to some variables and also explanatory for other variables so that it has both incoming and outgoing arrows in the regression graph.

As an example, we take data from a retrospective study with 283 adult females answering questions about their childhood when visiting their general practitioner, mostly for some minor health problems; see Hardt et al. (2008). A well-fitting graph is shown in Fig. 1. It contains two binary variables, A and B, and six quantitative variables. Except for the directly recorded feature, age in years, all other variables are derived from answers to questionnaires, coded so that high values correspond to high scores.

The three blocks, a, b and c, reflect here a time-ordering of vector variables, Y_a, Y_b and Y_c, with Y_a representing the joint response of primary interest, Y_b an intermediate vector variable and Y_c a context vector variable. The three individual components of the primary response Y_a are variables capturing how the respondent recollects aspects of her relationship to the mother. The intermediate variable Y_b has two components that reflect severe distress during childhood. The three components of the context variable Y_c capture background information about the respondent and about her family.

The graph in Fig. 1, derived after statistical analyses, shows, among other independences, that Y_a is conditionally independent of Y_c given Y_b, written compactly in terms of sets of nodes as $a \perp\!\!\!\perp c|b$. None of the components of Y_c has an arrow pointing directly to a component of Y_a, but sequences of arrows lead indirectly from c to a via b.

This says, for instance, that prediction of Y_a is not improved by knowing the context variable Y_c if information on the more recent intermediate variable Y_b is available. More interpretations of the independences are given later. When some edges are missing and each edge present corresponds to a substantial dependence, the graph

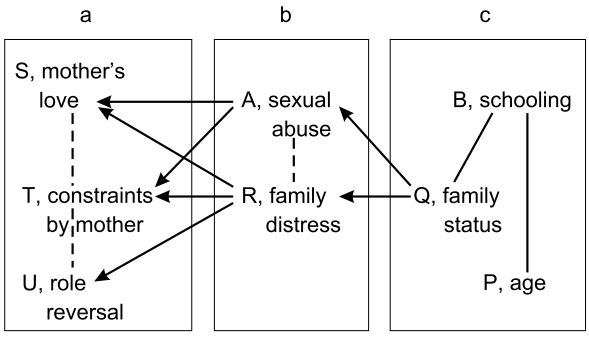


Fig. 1 A well-fitting regression graph for data on $n = 283$ adult females; within boxes are Y_a, Y_b, Y_c; corresponding ordered partitioning of the node set on top of the boxes

may also be viewed as a research hypothesis on which variables are needed to generate the joint distribution; see Wermuth and Lauritzen (1990). The goodness-of-fit of such a hypothesis can be tested in future studies.

Two models are *Markov equivalent* whenever their associated graphs capture the same *independence structure*, that is, the graphs lead to the same set of implied independence statements. Markov equivalent models cannot be distinguished on the basis of statistical goodness-of-fit tests for any given set of data. This may pose a problem in machine learning contexts. More precisely, knowledge about Markov equivalent models is essential for designing search procedures that converge with an increasing sample size to a true generating graph; see Castelo and Kocka (2003) for searches within the class of *directed acyclic graphs*, which consist exclusively of arrows and capture independences of ordered sequences in single response regressions.

More importantly though, Markov equivalent models may offer alternative interpretations of a given well-fitting model or open the possibility of using different types of fitting algorithms.

As we shall see in Sect. 7, the graph for nodes A, R, B, P, Q in blocks b and c of Fig. 1 is Markov equivalent to both graphs of Fig. 2. From knowing the Markov equivalence to the graph in Fig. 2(a), the joint response model for Y_b given Y_a may also be fitted in terms of univariate regressions and from the Markov equivalence to the graph in Fig. 2(b); one knows for instance directly, using Proposition 1 below, that sexual abuse is independent of age and schooling given knowledge about family distress and family status.

Regression graphs are a subclass of *the maximal ancestral graphs* of Richardson and Spirtes (2002) and both are subclasses of the *summary graphs* of Wermuth (2011). The two types are called *corresponding graphs* if they result after marginalizing over a node set m and conditioning on a disjoint node set c from a given directed acyclic graph. Both are *independence-preserving graphs* in the sense that they give the independence structure implied by the generating graph for all the remaining nodes and further conditioning or marginalizing can be carried out just as if the possibly much larger generating graph were used. The summary graph permits, in addition, to trace possible distortions of generating dependences as they arise in conditional dependences among the remaining variables, for instance in parameters of the maximal ancestral graph models.

In Sect. 2, we introduce further concepts and the notation needed to state at the end of Sect. 2 some of the main results of the paper and related results in the literature. In Sect. 3, a well-fitting regression graph is derived for data of chronic pain patients. Sections 4, 5 and 6 may be skipped if one wants to turn directly to formal definitions,

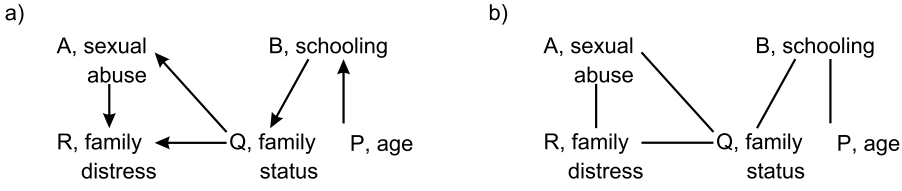


Fig. 2 Two Markov equivalent graphs to the one of Y_b, Y_c of Fig. 1

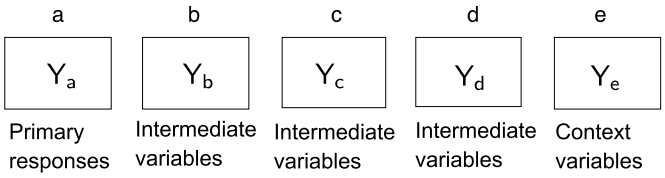


Fig. 3 A typical first ordering: here of five vector variables, $Y_a, \ldots, Y_e$; primary response Y_a listed *on the left*, context variable Y_e *on the right*, intermediate variables *in-between*

new results and proofs in Sect. 7. Section 4 reviews linear recursion relations that are mimicked by graphs and lead to the standard and to special ways of combining probability statements, summarized here in Sect. 5. In Sect. 6, some of the previous results in the literature for graphs and for Markov equivalences are highlighted. The Appendix contains details of the regressions analyses in Sect. 3.

2 Some further concepts and notation

Figure 3 shows five ordered blocks, to introduce the notion of connected components of the graph to represent conditionally independent responses given their common past.

In the example of a regression graph in Fig. 4 corresponding to Fig. 3, Y_a is a single response, Y_b has two component variables, both of Y_c and Y_e have four and Y_d has three. Each of the blocks b to e shows two *stacked boxes*, that is, subsets of nodes that are without any edge joining them. This is to indicate that disconnected components of a given response are conditionally independent given their past and that disconnected components of the context variables are completely independent.

Graphs with dashed lines are *covariance graphs denoted by* G^N_{cov}, those with full lines are *concentration graphs denoted by* G^N_{con}; see Wermuth and Cox (1998). The names are to remind one of their parameterization in *regular joint Gaussian distributions*, for which the covariance matrix is invertible and gives the *concentration matrix*. A zero ik-element in G^N_{cov} means $i \perp\!\!\!\perp k$ and a zero ik-element in G^N_{con} means $i \perp\!\!\!\perp k | \{1, \ldots, d\} \setminus \{i, k\}$; see Wermuth (1976a) or Cox and Wermuth (1996), Sect. 3.4.

The regression graph of Fig. 4 is consistent with the first ordering in Fig. 3 since no additional ordering is introduced, as it would have been by arrows within blocks a to e. After statistical analysis, blocks of the first ordering are often subdivided into the connected components of the graph, g_j, shown here in Fig. 4 with the help of the stacked boxes. For several nodes in g_j, each pair of nodes (i, k) is connected by at least one undirected ik-path within g_j. An *ik-path* connects its endpoint nodes i, k via a sequence of edges coupling distinct other nodes along the path, named *the path's inner nodes*.

For a regression graph, G^N_{reg}, the node set N has an ordered partitioning into two subsets, $N = (u, v)$, distinguishing response nodes within u from context nodes within v. The *connected components* g_j, for $j = 1, \ldots, J$, are the disconnected, undirected graphs that remain after removing all arrows from the graph. Thus, the displayed, stacked boxes in Fig. 4 are just a visual aid. We say that there is *an edge*

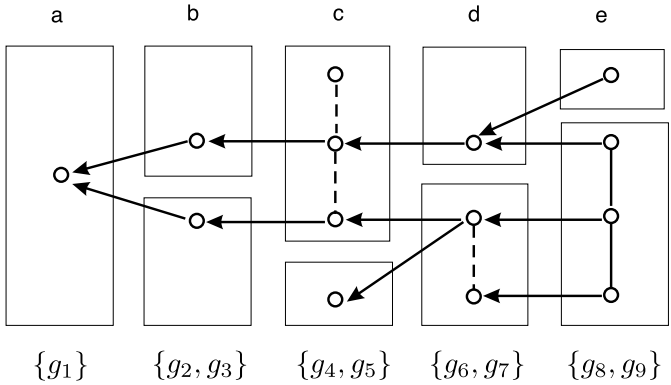


Fig. 4 A regression graph for 14 variables corresponding to blocks a to e of Fig. 3

between subsets a and b of N if there is an edge with one node in a and the other node in b. Then, the subgraph induced by nodes $a \cup b$ *is said to be connected in a and b*.

For any one block of stacked boxes, different orderings are possible. We speak of a *compatible ordering* if each *arrow* starting at a node in any g_j points to a node in $g_{<j} = g_1 \cup \cdots \cup g_{j-1}$, but never to a node in $g_{>j} = g_{j+1} \cup \cdots \cup g_J$, the *past of* g_j.

Full lines are edges coupling context variables within v. *Dashed lines* couple joint responses within u. The regression graph is *complete* if every node pair is coupled. In this case, the statistical model is *saturated* as it is unconstrained for some given family of distributions.

Let $g_1, \ldots, g_J$ denote any compatible ordering of the connected components of G^N_{reg}, then a corresponding joint density factorizes as

$$f_N = \prod_{j=1}^{J} f_{g_j|g_{>j}} \tag{1}$$

into sequences regressions for the joint responses g_j within u and for separate concentration graph models in disconnected g_j within v.

In a *generating process of* f_N *over a regression graph*, one starts with the density of g_J, and continues with the one of g_{J-1} given g_J up to the density of g_1 given $g_{>1}$ so that (1) is used for one given compatible ordering of the node set N. Every ik-edge present denotes a non-vanishing conditional dependence of Y_i and Y_k given some vector variable Y_c, written as $i \not\perp\!\!\!\perp k|c$, so that the graph is said to be edge-minimal or to capture a dependence structure. The generating process attaches the following meaning to each ik-edge present in G^N_{reg}:

$$\begin{aligned}
&\text{(i)} \quad i \not\perp\!\!\!\perp k|g_{>j} \quad \text{for } i,k \text{ both in a response component } g_j \text{ of } u;\\
&\text{(ii)} \quad i \not\perp\!\!\!\perp k|g_{>j} \setminus \{k\} \quad \text{for } i \text{ in } g_j \text{ of } u \text{ and } k \text{ in } g_{>j};\\
&\text{(iii)} \quad i \not\perp\!\!\!\perp k|v \setminus \{i,k\} \quad \text{for } i,k \text{ both in a context component } g_j \text{ of } v.
\end{aligned} \tag{2}$$

Notice that only for context variables, conditioning is on all other context variables while for responses, conditioning is exclusively on variables in their past. When the dependence sign $\pitchfork$ is replaced by the independence sign $\perp\!\!\!\perp$, equations (2) give with missing edges for node pairs i, k the *pairwise independence statements defining the independence structure of* G^N_{reg}, given the composition and the intersection property, discussed below, are applied.

An equivalent, more compact description of the set of defining pairwise independences and a proof of equivalence of this *pairwise Markov property* to the global Markov property has been given for the class of mixed loopless graphs, which contain regression graphs as a subclass; see Sadeghi and Lauritzen (2012); see also Kang and Tian (2009), Pearl and Paz (1987), Marchetti and Lupparelli (2011) for relevant, previous results. A *global Markov property* permits to read off the graph all independence statements implied by the graph.

Equation (2)(i) holds for the conditional covariance graphs of joint responses g_j within u having dashed lines as edges, (2)(ii) for the dependences of the single responses within g_j on variables in the past of g_j having arrows as edges, and Eq. (2)(iii) for the concentration graph of the context variables within v having full lines as edges. For instance, from the definition of the missing edges corresponding to (2), one can derive for Fig. 1, $S \perp\!\!\!\perp U|bc$ by (2)(ii), $P \perp\!\!\!\perp Q|B$ by (2)(iii), and both $A \perp\!\!\!\perp B|PQ$ and $A \perp\!\!\!\perp P|BQ$ by (2)(i) using first principles and the two special properties of the generated distributions named composition and intersection.

Notice that each missing edge of a regression graph corresponds to an independence statement for the uncoupled node pair; see also Lemmas 2 and 3 below. Therefore, regression graphs represent one special class of the so-called *independence graphs*. Whenever a regression graph G^N_{reg} consists of *two disconnected graphs*, for Y_a and Y_b say, since no path leads from a node in a to a node in b, and $a \cup b = N$, then $a \perp\!\!\!\perp b$ or $f_N = f_a f_b$, and the two vector variables may be analysed separately. Therefore, we treat in Sect. 7 of this paper only connected regression graphs.

All *graphs discussed* in this paper *have no loops*, that is, no edge connects a node to itself and they have *at most one edge between two different nodes*. Recall that an ik-path in such a graph can be described by a sequence of its nodes. By convention, an ik-path without inner nodes is an edge. For every ik-edge, the endpoints differ, $i \neq k$. An ik-path with $i = k$ has at least three nodes and is called a *cycle*.

A three-node path of arrows may contain only one of the three types of inner nodes shown in Fig. 5, called *transition, source and sink node*, respectively.

A *path is directed* if all its inner nodes are transition nodes. In a *directed cycle*, all edges are arrows pointing in the same direction and one returns to a starting node following the direction of the arrows. A regression graph contains no directed cycle and no *semi-directed cycles* which have at least one undirected edge in an otherwise directed cycle. If an arrow starts on a directed ik-path at k and points to i, then node k has been named an *ancestor* of node i and node i a *descendant* of node k.

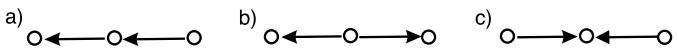


Fig. 5 The three types of three-node paths in directed acyclic graphs with inner nodes named (**a**) transition, (**b**) source, (**c**) sink node (or in directed acyclic graphs: collision node)

The *subgraph induced by a subset a* of the node set N consists of the nodes within a and of the edges present in the graph within a. A special type of induced subgraph, needed here, consisting of three nodes and two edges, is named a V-*configuration* or just a V. Thus, a three-node path forms a V if the induced subgraph has two edges.

An ik-*path is chordless* if for each of its three consecutive nodes (h, j, k), coupled by an hj-edge and jk-edge, there is no additional hk-edge present in the graph. In a *chordless cycle* of four or more nodes, the subgraph induced by every consecutive three nodes forms a V in the graph. An *undirected graph is chordal* if it contains no chordless cycle in four or more nodes.

In regression graphs, there may occur the three types of *collision* Vs of Fig. 6.

Notice that in a directed acyclic graph, the only possible collision V is directed and coincides with the sink V of Fig. 5(c).

An important common feature of the three Vs of Fig. 6 is that the inner node is excluded from every independence statements for its endpoints; see (2) and Lemma 2. In all other five possible types of V-configurations of a regression graph, named *transmitting* V*s*, the inner node is instead included in the independence statement for the endpoints; see (2) and Lemma 3 below. Notice that for uncoupled endpoints, both paths (a) and (b) of Fig. 5 are transmitting Vs. Similarly, the definition of transmitting and collision nodes remains unchanged if the Vs in Fig. 6 are interpreted as ik-paths for which there may be an additional ik-edge present in the graph.

A *collision path* has as inner nodes exclusively collision nodes, while a *transmitting path* has as inner nodes exclusively transmitting nodes. A chordless collision path in four nodes contains at least one dashed line. In particular, it is impossible to replace all the edges in such a four-node path by arrows and not generate at least one transmitting V. Thereby, the meaning of this missing edge would be changed and hence contradict its unique definition given from the generating process. The *skeleton* of a graph results by replacing each edge present by a full line. Now, two of the main new results of this paper can be stated.

Theorem 1 *Two regression graphs are Markov equivalent if and only if they have the same skeleton and the same sets of collision* V*s, irrespective of the type of edge.*

Theorem 2 *A regression graph with a chordal graph for the context variables can be oriented to be Markov equivalent to a directed acyclic graph in the same skeleton, if and only if it does not contain any chordless collision path in four nodes.*

Sequences of regressions were introduced and studied, without specifying a concentration graph model for the context variables, in Cox and Wermuth (1993) and Wermuth and Cox (2004), under the name of multivariate regression chains, reminding one of the sequences of unconstrained models that the class contains for Gaussian

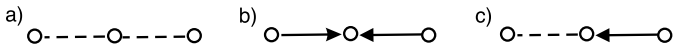


Fig. 6 The three types of collision Vs in regression graphs: (**a**) undirected, (**b**) directed or sink-oriented, (**c**) semi-directed; for uncoupled path endpoints, the inner node is excluded from every independence statement that the graph implies for these endpoints

joint responses. An extension to graphs including a concentration graph had already been proposed for directed acyclic graphs by Kiiveri et al. (1984). By this type of extension, the global Markov property of the graph remains unchanged.

A criterion for Markov equivalence of summary graphs has been derived by Sadeghi (2009) who also shows that two different criteria for maximal ancestral graphs are equivalent, those due to Zhao et al. (2005) and to Ali et al. (2009). These available Markov equivalence results and the associated proofs increase considerably in complexity, the larger the model class. On the other hand, the Markov equivalence criterion of Theorem 1 is simple and includes as special cases all available equivalence results for directed acyclic graphs, for covariance graphs and for concentration graphs, as set out in detail in Sects. 6 and 7 here.

For context variables taken as given, Gaussian regression graph models coincide with a large subclass of structural equation models (SEMs), those permitting local modelling due to the factorization property (1), and are without any *endogenous responses*. Such responses have residuals that are correlated with some of its regressors so that the so-called endogeneity problem is generated, by which, for joint Gaussian distributions, a zero equation parameter need not correspond to any conditional independence statement and a nonzero equation parameter is not a measure of conditional dependence. The consequence is that ordinary least-squares estimates of such equation parameters are typically strongly distorted. This was recognized by Haavelmo (1943) who received a Nobel Prize in Economics for this insight in 1989.

For traditional uses of SEMs see, for instance, Jöreskog (1981), Bollen (1989), Kline (2006), while Pearl (2009) advocates SEMs as a framework for causal inquiries. In the econometric literature forty years ago, independences were always regarded as 'overidentifying' constraints.

For discrete variables, more attractive features of regression graph models were derived by Drton (2009), who speaks of chain graph models of type IV for multivariate regression chains in the case all variables on equal standing have covariance graphs. He proves that each member in this class belongs to a curved exponential family; for a discussion of this notion see, for instance, Cox (2006), Sect. 6.8. Discrete type IV models form also a subclass of marginal models; see Rudas et al. (2010), Bergsma and Rudas (2002). Local independence statements that involve only variables in the past are equivalent to more complex local independences used by Drton (2009); see Marchetti and Lupparelli (2011). These local definitions imply the pairwise independence formulation for missing edges corresponding to Eq. (2) for any regression graph, G^N_{reg}.

Two other types of chain graph have been studied as joint response models in statistics, the so-called *AMP chain graphs* of Andersson et al. (2001), and the *LWF chain graphs* of Lauritzen and Wermuth (1989) and Frydenberg (1990). They use the same factorization as in Eq. (1), but they are suitable for modelling data from intervention studies only when they are Markov equivalent to a regression graph. The reason is that the conditioning set for pairwise independences of responses includes in general other nodes within the same connected component. For AMP graphs, the independence form of Eq. (2)(i) is replaced by

$$(\text{i}') \quad i \perp\!\!\!\perp k | g_{>j-1} \setminus \{i,k\} \quad \text{for } i,k \text{ both within a response component } g_j,$$

while (2)(ii) and (2)(iii) remain unchanged. For LWF graphs, (i) is also replaced by (i′) and the independence form of (ii) by

$$(\text{ii}') \quad i \perp\!\!\!\perp k | g_{>j-1} \setminus \{i, k\} \quad \text{for } i \text{ within a } g_j \text{ and } k \text{ in } g_{>j}.$$

As a consequence, each undirected subgraph in an AMP chain graph is a concentration graph, and an LWF chain graph consists of sequences of concentration graphs. For the corresponding different types of parameterizations of joint Gaussian distributions, see Wermuth et al. (2006b).

Not yet systematically approached is the search for *covering models that capture most but not all independences* in a more complex graph but which may be easier to fit than the reduced model; see Cox and Wermuth (1990). For regression graphs, details are explained here for a small example in Sect. 4, and in Sect. 7, first results are given in Propositions 8 to 10 and discussed using Figs. 16 and 17.

Before we turn to the different types of missing edges in more detail, we derive a well-fitting regression graph for data given by Kappesser (1997).

3 Deriving and interpreting a regression graph

For 201 chronic pain patients, the role of the site of pain during a three-week stay in a chronic pain clinic was to be examined. In this study, it was of main interest to investigate the changes in two main symptoms before and after stationary treatment and to understand determinants of the overall treatment success as rated by the patients, three months after they had left the clinic. Figure 7 shows a first ordering of the variables derived in discussions between psychologists, physicians and statisticians.

The first ordering of the variables gives, for each single or joint response, a list of its possible explanatory variables, shown in boxes to the right, but in Fig. 7 only

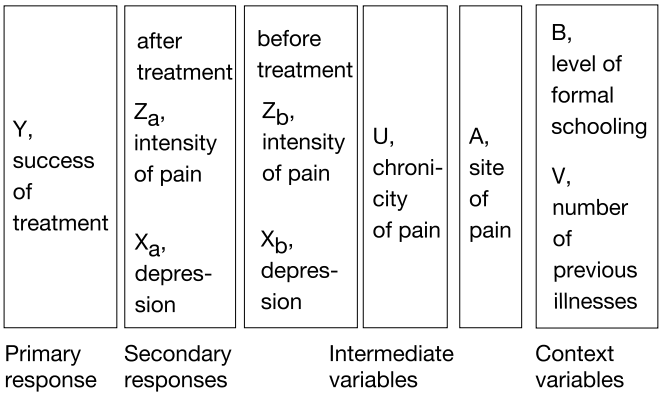


Fig. 7 First ordering of variables in the chronic pain study. There are two joint responses, intensity of pain and depression. They are the main symptoms of chronic pain, measured here before and after treatment. The components of each response are to be modelled conditionally given the variables listed in boxes to their right

those variables are displayed that remained, after statistical analyses, relevant for the responses of main interest.

Selecting for each response all its directly explanatory variables from this list and checking for remaining dependences among components of joint responses, provides enough insight to derive a well-fitting regression graph model. With this type of local modelling, the reasons for the model choice are made transparent.

Of the available background variables, age, gender, marital status and others, only the binary variables, level of formal schooling ($1 := $ less than ten years, $2 := $ ten or more years) and the number of previous illnesses in years ($\min := 0$, $\max := 16$) are displayed in the far right box as the relevant context variables. The response of primary interest, self-reported success of treatment, is listed in the box to the far left. It is a score that ranges between 0 and 35, combining a patient's answers to a specific questionnaire.

There are a number of intermediate variables. These are both explanatory for some variables and responses to others. Of these, two are regarded as joint responses since they represent two symptoms of a patient, intensity of pain and depression. Both are measured before treatment and directly after the three-week stationary stay. Questionnaire scores are available of depression ($\min := 0$, $\max := 46$) and of the self-reported intensity of pain ($\min := 0$, $\max := 10$). Chronicity of pain is a score ($\min := 0$, $\max := 8$) that incorporates different aspects, such as the frequency and duration of pain attacks, the spreading of pain and the use of pain relievers. In this study, the patients have one of two main sites of pain, the pain is either on their upper body, 'head, face, or neck' or on their 'back'.

A well-fitting regression graph is shown in Fig. 8. The graph summarizes some important aspects of the results of the statistical analyses for which details are given in the Appendix. In particular, it tells which of the variables are directly explanatory, that is, which are important for generating and predicting a response, by showing arrows that start from each of these directly explanatory variables and point to the response.

Variables listed to the right of a response but without an arrow ending at this response do not substantially improve the prediction of the response when used in addition to the directly explanatory variables. For instance, for treatment success, only

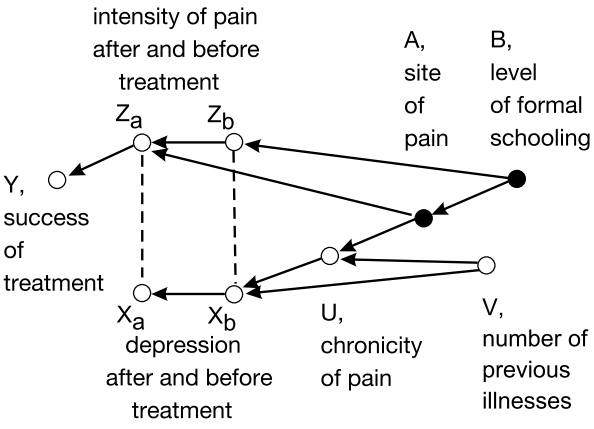


Fig. 8 Regression graph, well compatible with the data, that results from the reported statistical analyses. Discrete variables are drawn *as dots*, continuous ones *as circles*

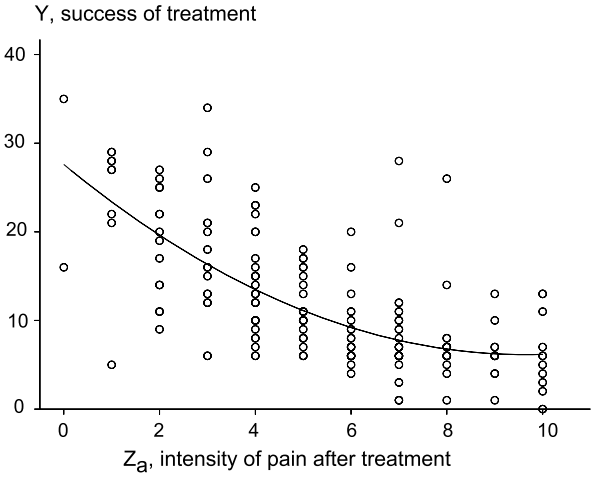


Fig. 9 Form of dependence of primary response Y on Z_a

the pain intensity after the clinic stay is directly explanatory and this pain intensity is an important mediator (intermediate variable) between treatment success and site of pain.

Scores of self-reported treatment success are low for almost all patients with high pain scores after treatment, that is, for scores higher than 6; see Fig. 9. Otherwise, treatment success is typically judged to be higher the lower the intensity of pain after treatment. This explains the nonlinear dependence of Y on Z_a.

As mentioned before, for back pain patients, the chronicity scores are on average higher than for headache patients and connected with a higher chronicity of the pain are higher scores of depression. These patients may possibly have tried too late, after the acute pain had started, to get well focused help. Both before and after treatment, highly depressed patients tend to report higher intensities of pain than others.

The study provides no information on which variables may explain these dependences between the symptoms that remain after having taken the available explanatory variables into account. However, hidden common explanatory variables may exist in both cases since these remaining dependences between the symptoms do not depend systematically on any other observed variable.

Some variables are *indirectly explanatory*. An arrow starts from an indirectly explanatory variable, and points via a sequence of arrows and intermediate variables to the response variable. For instance, the level of formal schooling and the site of pain are both indirectly explanatory for each of the symptoms after treatment and for the overall treatment success.

Once the types and directions of the direct dependence are taken into account, the regression graph helps to trace the development of chronic pain, starting from the context information on the level of schooling and the number of previous illnesses of a patient. Thus, patients with more years of formal schooling are more likely to be chronic headache patients. Patients with a lower level of formal schooling are more likely to be backache patients, possibly because more of them have jobs involving hard physical work. Backache patients reach higher stages of the chronicity of pain

and report higher intensity of pain still after treatment and are therefore typically less satisfied with the treatment they had received.

Graphical screening for nonlinear relations and interactive effects (Cox and Wermuth 1994) pointed to the nonlinear dependence of treatment success on intensity of pain after treatment but to no other such relations. The regression graph model is said to fit the data well because for each single response separately, there is no indication that adding a further variable would substantially change the generated conditional dependences. The seemingly unrelated dependences of the symptoms after treatment on those before treatment agree so well with the observations that they differ also little from regressions computed separately, see the appropriate tables in the Appendix.

Had there been no nonlinear relation and no categorical variables as responses, the overall model fit could also have been tested within the framework of structural equation models once the regression graph is available. This graph is derived here with the local modelling steps that use the first ordering of the variables, just in terms of univariate, multivariate and seemingly unrelated regressions. The regression graph provides a hypothesis that may be tested locally and/or globally in future studies that include the same set of nine variables. In this case, no variable selection strategy would be used or needed.

The available results for changes of the regression graph (Wermuth 2011) that result after marginalizing and conditioning provide a solid basis for comparing the results of any sequence of regressions with studies that contain the same set of core variables but which have some of the variables omitted or which consider subpopulations, defined by levels or level combinations of other variables. For instance, for comparisons with the current study, the same chronicity score may not be recorded in another pain clinic or data may be available only for patients with pain in the upper body.

The main substantive results of this empirical study are that site of pain needs to be taken into account also in future studies since it is an important mediator between the intrinsic characteristics of a patient, measured here by the given context variables, for both the overall treatment success and for the symptoms after treatment. For backache patients, the chronicity of pain and the depression score is higher than for the headache patients and the treatment is less successful since the intensity of pain remains high after the treatment in the clinic.

In the following section, we give three-variable examples of a Gaussian joint response regression and of the three subclasses of regression graphs that have only one type of edge, of the covariance, the concentration and the directed acyclic graph to discuss the different types of conditional dependences and the possible types of independence constraints associated with the corresponding regression graphs.

4 Regressions, dependences and recursive relations

For a quantitative response with linear dependences, the simple regression model dates back at least several centuries. The fitting of a least-squares regression line had been developed separately by Carl Friedrich Gauss (1777–1855), Adrien-Marie Legendre (1752–1833) and Robert Adrain (1775–1843). The method extends directly to models with several explanatory variables.

The most studied regression models are for joint Gaussian distributions. Regression graphs mimic important features of these linear models but represent also relations in other distributions of continuous and discrete variables, which permit in particular nonlinear and interactive dependences. In a regular joint Gaussian distribution, let the mean-centred vector variable Y have dimension three, then we write the covariance matrix, Σ, and the concentration matrix, Σ^{-1}, with graphs shown in Fig. 10, as

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

where the dot-notation indicates entries in a symmetric matrix.

With the edge of node pair $(1, 2)$ removed, both graphs turn into a V but have different interpretations. The resulting independence constraints are for parts (a) and (b) of Fig. 10, respectively,

$$1 \perp\!\!\!\perp 2 \iff (\sigma_{12} = 0) \quad \text{and} \quad 1 \perp\!\!\!\perp 2|3 \iff \big(\sigma^{12} = 0\big),$$

where the latter derives as an important property of concentration matrices; for proofs see Cox and Wermuth (1996), Sect. 3.4 or Wermuth et al. (2006a), Sect. 2.3. For other distributions, the independence interpretation of these two types of undirected graph remains unchanged, but not the parameterization. A similar statement holds for directed acyclic graphs and, more generally, for regression graphs.

For the linear equations that lead to a complete directed acyclic graph for a trivariate Gaussian distribution with mean zero, one starts with three mutually independent Gaussian residuals ε_i and takes the following system of equations, in which for instance $\beta_{1|3.2}$ is a regression coefficient for the dependence of response Y_1 on Y_3 when Y_2 is an additional regressor. Because of the form of the equations, one speaks of triangular systems also when the distribution of the residuals is not Gaussian, but the residuals are just uncorrelated, or expressed equivalently, if each residual is uncorrelated with the regressors in its equation:

$$\begin{aligned} Y_1 &= \beta_{1|2.3} Y_2 + \beta_{1|3.2} Y_3 + \varepsilon_1, \\ Y_2 &= \beta_{2|3} Y_3 + \varepsilon_2, \\ Y_3 &= \varepsilon_3. \end{aligned} \tag{3}$$

When the residuals do not follow Gaussian distributions, the probabilistic independence interpretation is lost, but the lack of a linear relation can be inferred with any vanishing regression coefficient.

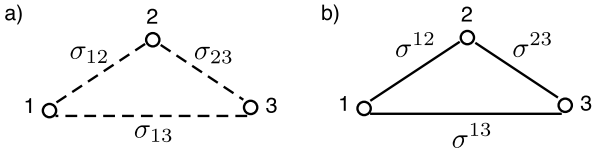


Fig. 10 For unconstrained trivariate Gaussian distributions, the parameters attached to the edges are those corresponding to (**a**) a covariance graph, (**b**) a concentration graph

In econometrics, Hermann Wold (1908–1992) introduced such systems as linear recursive equations with uncorrelated residuals. Harald Cramér (1893–1985) used the term linear least-squares equations for residuals in a population being uncorrelated with the regressors and the notation for the regression coefficients is an adaption of the one introduced by Udny Yule (1871–1951) and William Cochran (1909–1980).

In joint Gaussian distributions, independence constraints on triangular systems mean vanishing equation parameters and missing edges in directed acyclic graphs, such as

$$1 \perp\!\!\!\perp 2|3 \iff (\beta_{1|2.3} = 0) \quad \text{and} \quad 2 \perp\!\!\!\perp 3 \iff (\beta_{2|3} = 0).$$

The complete directed acyclic graph defined implicitly with equations (3) is displayed in Fig. 11(a).

For the smallest joint response model with the complete graph shown in Fig. 11(b), we take both Gaussian variables Y_1 and Y_2 to depend on a Gaussian variable Y_3, to get equations (4) with residuals having zero means and being uncorrelated with Y_3:

$$Y_1 = \beta_{1|3} Y_3 + u_1, \qquad Y_2 = \beta_{2|3} Y_3 + u_2, \qquad Y_3 = u_3. \tag{4}$$

Here, $\sigma_{12|3} = E(u_1 u_2)$. The generating processes and hence the interpretation differs for the two models in Eqs. (3) and (4). In the corresponding graphs of Fig. 11(a) and (b), the vanishing of the edges for pairs (1, 2) and (2, 3) mean the same independence constraints since

$$1 \perp\!\!\!\perp 2|3 \iff (\sigma_{12|3} = 0) \iff (\beta_{1|2.3} = 0) \quad \text{and} \quad 2 \perp\!\!\!\perp 3 \iff (\beta_{2|3} = 0),$$

but the edges for pair (1, 3) capture different dependences, $1 \perp\!\!\!\cap 3$ and $1 \perp\!\!\!\cap 3|2$, respectively. Again, taking away any edge generates a V. Taking away any two edges means to combine two independence statements. This is discussed further in the next section.

One of the special important features of the linear least-squares regressions is that the residuals are uncorrelated with the regressors. The effect is that the model part coincides with a conditional linear expectation as illustrated here with a model for response Y_1 and regressors Y_2, Y_3, which we take, as mentioned before, as measured in deviations from their means. For instance, one gets for

$$\begin{aligned} Y_1 &= \beta_{1|2.3} Y_2 + \beta_{1|3.2} Y_3 + \varepsilon_1, \\ E_{\text{lin}}(Y_1|Y_2, Y_3) &= \beta_{1|2.3} Y_2 + \beta_{1|3.2} Y_3. \end{aligned} \tag{5}$$

There is a recursive relation for least-squares regression coefficients; see Cochran (1938), Cox and Wermuth (2003), Ma et al. (2006). It shows for instance, with

$$\beta_{1|3} = \beta_{1|3.2} + \beta_{1|2.3}\beta_{2|3} \tag{6}$$

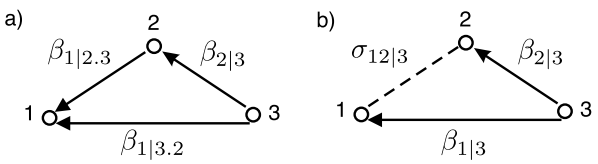


Fig. 11 Parameters of a Gaussian distribution in: (**a**) a complete G^N_{dag}, (**b**) a complete G^N_{reg}

that $\beta_{1|3.2}$, the partial coefficient of Y_3 given also Y_2 as a regressor for Y_1, coincides with the marginal coefficient, $\beta_{1|3}$, if and only if $\beta_{1|2.3} = 0$ or $\beta_{2|3} = 0$.

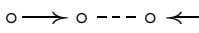
The method of maximizing the likelihood was recommended by Sir Ronald Fisher (1890–1962) as a general estimation technique that applies also to regressions with categorical or quantitative responses. One of the most attractive features of the method concerns properties of the estimates. Given two models with parameters that are in one-to-one correspondence, the same one-to-one transformation leads from the maximum-likelihood estimates under one model to those of the other.

Different single response regressions, such as logistic, probit, or linear regressions, were described as special cases of the generalized linear model by Nelder and Wedderburn (1972); see also McCullagh and Nelder (1989). In all of these regressions, the vanishing of the coefficient(s) of a regressor indicates conditional independence of the response given all directly explanatory variables for this response.

The general linear model with a vector response, also called multivariate linear regression, has identical sets of regressors for each component variable of a response vector variable. Maximum-likelihood estimation of regression coefficients for a joint Gaussian distribution reduces to linear least-squares fitting for each component separately; see Anderson (1958), Chap. 8.

With different sets of regressors for the components of a vector response, seemingly unrelated regressions (SUR) result and iterative methods are needed for estimation; see Zellner (1962). For small sample sizes, a given solution of the likelihood equations of a Gaussian SUR model may not be unique (see Drton and Richardson 2004; Sundberg 2010), while for exclusively discrete variables this will never happen (see Drton 2009). For mixed variables, no corresponding results are available yet.

In general, there often exists *a covering model with nice estimation properties*. For instance, one of the above described Gaussian SUR models that requires iterative fitting has regression graph



A generating process starts with independent explanatory variables, each of which relates only to one of the two response components, but these are correlated given both regressors. There is a simple covering model, in which two missing arrows are added to the graph to obtain a general linear model. In that case, in the new graph not every edge corresponds to a dependence, but closed-form maximum-likelihood estimates are available.

For a vector variable of categorical responses only, the multivariate logistic regression of Glonek and McCullagh (1995) reduces to separate main effect logistic regressions for each component of the response vector provided that certain higher-order interactions vanish; see Marchetti and Lupparelli (2011). In the context of structural equation models (SEMs), dependences of binary categorical variables are modelled in terms of probit regressions. These do not differ substantially from logistic regressions whenever the smallest and largest events occur at least with probability 0.1; see Cox (1966).

Multivariate linear regressions as well as SUR models belong to the framework of SEMs even though this general class had been developed in econometrics to deal

appropriately with endogenous responses. Estimation methods for SEMs were discussed in the Berkeley symposia on mathematical statistics and probability from 1945 to 1965, but some identification issues have been settled only recently; see Foygel et al. (2011) and for relevant previous results Brito and Pearl (2002), Stanghellini and Wermuth (2005).

In statistical models that treat all variables on equal standing, the variables are not assigned roles of responses or regressors and undirected measures of dependence are used instead of coefficients of directed dependence. In the concentration graph models, the undirected dependences are conditional given all remaining variables on equal standing.

For instance, for categorical variables, these models are better known as graphical log-linear models; see Birch (1963), Caussinus (1966), Goodman (1970), Bishop et al. (1975), Wermuth (1976a), Darroch et al. (1980). For Gaussian random variables, these had been introduced as covariance selection models (see Dempster 1972; Wermuth 1976b; Speed and Kiiveri 1986; Drton and Perlman 2004), and for mixed variables as graphical models for conditional Gaussian (CG) distributions (see Lauritzen and Wermuth 1989; Edwards 2000).

For a mean-centred vector variable Y, the elements of the covariance matrix Σ are $\sigma_{ij} = E(Y_i Y_j)$. If Σ is invertible, the covariances σ_{ij} are in a one-to-one relation with the concentrations σ^{ij}, the elements of the concentration matrix Σ^{-1}. There is a recursive relation for concentrations; see Dempster (1969). For a trivariate distribution,

$$\sigma^{23.1} = \sigma^{23} - \sigma^{12}\sigma^{13}/\sigma^{11}, \tag{7}$$

where $\sigma^{23.1}$ denotes the concentration of Y_2, Y_3 in their bivariate marginal distribution. Thus, the overall concentration σ^{23} coincides with $\sigma^{23.1}$ if and only if $\sigma^{12} = 0$ or $\sigma^{13} = 0$.

Alternatively, in covariance graph models, the undirected measures for variables on equal standing are pairwise marginal dependences. For Gaussian variables, these models had been introduced as hypotheses linear in covariances; see Anderson (1973), Kauermann (1996), Kiiveri (1987), Wermuth et al. (2006a), Chaudhuri et al. (2007). For categorical variables, covariance graph models have been studied only more recently; see Drton and Richardson (2008a), Lupparelli et al. (2009). Again, no similar estimation results are available for general mixed variables yet.

There is also a recursive relation for covariances; see Anderson (1958), Sect. 2.5. It shows for instance, for just three components of Y having a Gaussian distribution, with

$$\sigma_{12|3} = \sigma_{12} - \sigma_{13}\sigma_{23}/\sigma_{33}, \tag{8}$$

where $\sigma_{12|3}$ denotes the covariance of Y_1, Y_2 given Y_3. Therefore, $\sigma_{12|3}$ coincides with σ_{12} if and only if $\sigma_{13} = 0$ or $\sigma_{23} = 0$. By Eqs. (6), (7), (8), a unique independence statement is associated with the endpoints of any V in a trivariate Gaussian distribution.

In the context of multivariate exponential families of distributions, concentrations are special canonical parameters and covariances are special moment parameters with estimates of canonical and moment parameters being asymptotically independent; see

Barndorff-Nielsen (1978), p. 122. Regression graphs capture independence structures for more general types of distribution, where operators for transforming graphs mimic operators for transforming different parameterizations of joint Gaussian distributions; see Wermuth et al. (2006b), Wiedenbeck and Wermuth (2010), Wermuth (2011).

In particular, by removing an edge from any V of a regression graph, one introduces an additional independence constraint just as in a regular joint Gaussian distribution. For this, the generated distributions have to satisfy the composition and intersection property in addition to the general properties, as discussed in the next section.

5 Using graphs to combine independence statements

We now state the four standard properties of independences of any multivariate distribution; see e.g. Dawid (1979), Studený (2005), as well as two special properties of joint Gaussian distributions. The six, taken together, describe the combination and decomposition of independences in regression graphs, for instance those resulting by removing edges. We discuss when these six properties apply also to regression graph models.

Let X, Y, Z be random (vector) variables, continuous, discrete or mixed. By using the same compact notation, f_{XYZ} for a given joint density, a probability distribution or a mixture and by denoting the union of say X and Y by XY, one has

$$X \perp\!\!\!\perp Y|Z \iff (f_{XYZ} = f_{XZ} f_{YZ}/f_Z), \tag{9}$$

where for instance f_Z denotes the marginal density or probability distribution of Z. Since the order of listing variables for a given density is irrelevant, *symmetry of conditional independence* is one of the standard properties, that is,

$$\text{(i)} \quad X \perp\!\!\!\perp Y|Z \iff Y \perp\!\!\!\perp X|Z.$$

Equation (9) restated for instance for the conditional distribution of X given Y and Z, $f_{X|YZ} = f_{XYZ}/f_{YZ}$, is

$$X \perp\!\!\!\perp Y|Z \iff (f_{X|YZ} = f_{X|Z}). \tag{10}$$

When two edges are removed from a graph in Figs. 10 and 11, just one coupled pair remains, suggesting that the single node is independent of the pair.

For instance, in Fig. 11(a) with nodes 1, 2, 3 corresponding in this order to X, Y, Z, removing the arrows for (1, 2) and (2, 3), leaves (1, 3) disconnected from node 2. For any joint density, implicitly generated as $f_{XYZ} = f_{X|YZ} f_{Y|Z} f_Z$, one has equivalently,

$$(X \perp\!\!\!\perp Y|Z \text{ and } Y \perp\!\!\!\perp Z) \iff XZ \perp\!\!\!\perp Y.$$

In general, the *contraction property* is for a, b, c, d disjoint subsets of N:

$$\text{(ii)} \quad (a \perp\!\!\!\perp b|cd \text{ and } b \perp\!\!\!\perp c|d) \iff ac \perp\!\!\!\perp b|d.$$

It has become common to say that a *distribution is generated over a given* G^N_{dag} if the distribution factorizes as specified by the graph for any compatible ordering. For instance, for a trivariate distribution generated over the collision V of Fig. 11(b) obtained by removing the edge for $(2, 3)$, both orders $(1, 2, 3)$ and $(1, 3, 2)$ are compatible with the graph and $f_{XYZ} = f_{X|YZ} f_Y f_Z$.

Conversely, suppose that $XZ \perp\!\!\!\perp Y$ holds, then this implies $X \perp\!\!\!\perp Y$ and $Z \perp\!\!\!\perp Y$ so that for instance the same two edges as in Fig. 11(b) are missing in the corresponding covariance graph of Fig. 10(a). In general, the *decomposition property* is for a, b, c, d disjoint subsets of N:

$$\text{(iii)} \quad a \perp\!\!\!\perp bc|d \implies (a \perp\!\!\!\perp b|d \text{ and } a \perp\!\!\!\perp c|d).$$

In addition, $XZ \perp\!\!\!\perp Y$ implies $X \perp\!\!\!\perp Y|Z$ and $Z \perp\!\!\!\perp Y|X$ so that for instance the same two edges as in Fig. 11(a) are missing in the corresponding concentration graph of Fig. 10(b). In general, the *weak union property* is for a, b, c, d disjoint subsets of N:

$$\text{(iv)} \quad a \perp\!\!\!\perp bc|d \implies (a \perp\!\!\!\perp b|cd \text{ and } a \perp\!\!\!\perp c|bd).$$

Under some regularity conditions, all joint distributions share the four properties (i) to (iv).

Joint distributions, for which the reverse implication of the decomposition property (iii) and of the weak union property (iv) hold such as a regular joint Gaussian distribution, are said to have, respectively, the *composition property* (v) and the *intersection property* (vi), that is, for a, b, c, d disjoint subsets of N:

$$\text{(v)} \quad (a \perp\!\!\!\perp b|d \text{ and } a \perp\!\!\!\perp c|d) \implies a \perp\!\!\!\perp bc|d,$$

$$\text{(vi)} \quad (a \perp\!\!\!\perp b|cd \text{ and } a \perp\!\!\!\perp c|bd) \implies a \perp\!\!\!\perp bc|d.$$

The standard graph theoretical separation criterion has different consequences for the two types of undirected graph corresponding for Gaussian distributions to concentration and to covariance matrices. We say *a path intersects subset set* c of node set N if it has an inner node in c and let $\{a, b, c, m\}$ partition N to formulate known Markov properties. The notation is to remind one that with any independence statement $a \perp\!\!\!\perp b|c$, one implicitly has marginalized over the remaining nodes in $m = V \setminus \{a \cup b \cup c\}$, i.e. one considers the marginal joint distribution of Y_a, Y_b, Y_c.

Proposition 1 (Lauritzen 1996) *A concentration graph,* G^N_{con}, *implies* $a \perp\!\!\!\perp b|c$ *if and only if every path from* a *to* b *intersects* c.

Proposition 2 (Kauermann 1996) *A covariance graph,* G^N_{cov}, *implies* $a \perp\!\!\!\perp b|c$ *if and only if every path from* a *to* b *intersects* m.

Notice that Proposition 1 requires the intersection property, otherwise one could not conclude for three distinct nodes h, i, k e.g. that $(h \perp\!\!\!\perp i|k$ and $h \perp\!\!\!\perp k|i)$ implies $h \perp\!\!\!\perp ik$, while Proposition 2 requires the composition property, otherwise one could conclude e.g. that $(h \perp\!\!\!\perp i$ and $h \perp\!\!\!\perp k)$ implies $h \perp\!\!\!\perp ik$.

Corollary 1 *A covariance graph,* G^N_{cov}*, or a concentration graph,* G^N_{con}*, implies* $a \perp\!\!\!\perp b$ *if and only if in the subgraph induced by* $a \cup b$*, there is no edge between* a *and* b*.*

Corollary 2 *A regression graph,* G^N_{reg}*, captures an independence structure for a distribution with density* f_N *factorizing as* (1) *if the composition and intersection property hold for* f_N*, in addition to the standard properties of each density.*

Proof Given the intersection property (vi), any node i with missing edges to nodes k, l in a concentration graph of node set N implies $i \perp\!\!\!\perp \{k, l\}|N \setminus \{i, k, l\}$, and given the composition property (v), any node i with missing edges to nodes k, l in a covariance graph given Y_c implies $i \perp\!\!\!\perp \{k, l\}|c$. □

For purely discrete and for Gaussian distributions, necessary and sufficient conditions for the intersection property (vi) to hold are known; see San Martin et al. (2005). Too strong sufficient conditions are for joint Gaussian distributions that they are regular and for discrete variables, that the probabilities are strictly positive.

The composition property (v) is satisfied in Gaussian distributions and for triangular binary distributions with at most main effects in symmetric $(-1, 1)$ variables; see Wermuth et al. (2009). Both properties (v) and (vi) hold, whenever a distribution may have been generated over a possibly larger parent graph; see Wermuth (2011), Marchetti and Wermuth (2009), Wermuth et al. (2006b). *Parent graphs* are directed acyclic graphs that do not only capture an independence structure but are also edge-minimal with a unique independence statement assigned to each V of the graph. *A distribution generated over a parent graph* mimics these properties of the parent graph. It is known that every regression graph can be generated by a larger directed acyclic graph but not necessarily every statistical regression graph model can be generated in this way; see Richardson and Spirtes (2002), Sects. 6 and 8.6. One needs similar properties for distributions generated over a regression graph.

A graph is edge-minimal for the generated distribution if the distribution has a pairwise independence for each edge missing and a non-vanishing dependence for each edge present in the graph. For the generated distribution to have a unique independence statement assigned to each missing edge, it has to be *singleton transitive*, that is, for h, i, k, l distinct nodes of N,

$$(i \perp\!\!\!\perp k|l \text{ and } i \perp\!\!\!\perp k|lh) \implies (i \perp\!\!\!\perp h|l \text{ or } k \perp\!\!\!\perp h|l).$$

This says, that in order to have both a conditional independence of Y_i, Y_k given Y_l and given Y_l, Y_h, there has to be at least one additional independence involving the variable Y_h, the additional variable in the conditioning set. Graphs that are edge-minimal form *a dependence base* if they also satisfy singleton transitivity, expressed as

$$(i \not\perp\!\!\!\perp h|l \text{ and } k \not\perp\!\!\!\perp h|l \text{ and } i \perp\!\!\!\perp k|l) \implies i \not\perp\!\!\!\perp k|\{l, h\}$$

and

$$\big(i \not\perp\!\!\!\perp h|l \text{ and } k \not\perp\!\!\!\perp h|l \text{ and } i \perp\!\!\!\perp k|\{l, h\}\big) \implies i \not\perp\!\!\!\perp k|l,$$

which says that in the distribution there is a unique independence statement that corresponds to each V in the graph. For a $2 \times 2 \times 3$ contingency table, an example violating singleton-transitivity has been given with Eq. (5.4) by Birch (1963).

There exist these peculiar types of incomplete families of distributions (see Lehmann and Scheffé 1955; Brown 1986; Mandelbaum and Rüschendorf 1987), in which independence statements connected with a V may have the inner node both within and outside the conditioning set (see Wermuth and Cox 2004, Sect. 7; Darroch 1962). Such independences have also been characterized as being not representable in joint Gaussian distributions; see Lněnička and Matúš (2007). These distributions and those that are faithful to graphs are of limited interest in application in which one wants to interpret sequences of regressions.

Distribution is said to be faithful to a graph if every of its independence constraints is captured by a given independence graph; see Spirtes et al. (1993). As is proven in a forthcoming paper, this requires for regression graphs that (1) the graph represents both an independence and a dependence structure, and that (2) the distribution satisfies the composition and the intersection property and is set *transitive*, a property that is the following extension of singleton transitivity for node h replaced by a subset d of $N \setminus \{i, k, l\}$ that may contain several nodes:

$$\big(i \perp\!\!\!\perp k|l \text{ and } i \perp\!\!\!\perp k|\{l, d\}\big) \implies (i \perp\!\!\!\perp d|l \text{ or } k \perp\!\!\!\perp d|l).$$

This faithfulness property imposes strange constraints on parameters whenever more than two nodes induce a complete subgraph in the graph; see for instance Fig. 1 in Wermuth et al. (2009) for three binary variables. An early example of a regular Gaussian distribution that does not satisfy weak transitivity is due to Cox and Wermuth (1993), Eq. (8).

Notice that in general, the extension of singleton transitivity to weak transitivity excludes parametric cancellations that result from several paths connecting the same node pair. This is the only type of a possible parametric cancellation in regular Gaussian distributions; see Wermuth and Cox (1998).

However, the constraints are mild for distributions corresponding to regression graphs that are edge-minimal and that are forests. *Forests* are the union of disjoint trees and a *tree* is a connected undirected graph with one unique path joining every node pair.

Lemma 1 *A positive distribution is faithful to a forest representing both an independence and a dependence structure if it is singleton transitive.*

Proof Positive distributions satisfy the intersection property and for concentration graphs, the composition property is irrelevant. Given the above characterizations of faithfulness and of set transitivity, there are in a forest no cancellations due to several paths connecting the same node pair. Hence, set transitivity will be violated only if the singleton transitivity fails. □

Corollary 3 *A regular Gaussian distribution is faithful to a forest representing both an independence and a dependence structure.*

Notice that forests include trees and Markov chains as special cases. If they are edge-minimal, they are Markov equivalent to very special types of parent graphs which are are rarely of interest in statistics when studying sequences of regressions.

6 Some early results on graphs and Markov equivalence

In the past, results concerning graphs and Markov equivalence have been obtained quite independently in the mathematical literature on characterizing different types of graph, in the statistical literature on specifying types of multivariate statistical models, and in the computer science literature on deciding on special properties of a given graph or on designing fast algorithms for transforming graphs.

For instance, following the simple enumeration result for *labelled trees* in d nodes, d^{d-2}, by Karl-Wilhelm Borchardt (1817–1880), it could be shown that these trees are in one-to-one correspondence to distinct strings of size $d-2$; see Cayley (1889). Much later, labelled trees were recognized to form the subclass of directed acyclic graphs with exclusively source Vs and therefore to be also Markov equivalent to chordal concentration graphs that are without chordless paths in four nodes; see Castelo and Siebes (2003).

In the literature on graphical Markov models, a number of different names have been in use for a sink V, for instance 'two arrows meeting head-on' by Pearl (1988), 'unshielded collider' by Richardson and Spirtes (2002), and 'Wermuth-configuration' by Whittaker (1990), after it had been recognized that, for Gaussian distributions, the parameters of a directed acyclic graph model without sink Vs are in one-to-one correspondence to the parameters in its skeleton concentration graph model.

Proposition 3 (Wermuth 1980; Wermuth and Lauritzen 1983; Frydenberg 1990) *A directed acyclic graph is Markov equivalent to a concentration graph of the same skeleton if and only if it has no collision* V.

Efficient algorithms to decide whether an undirected graph can be oriented into a directed acyclic graph, became available in the computer science literature under the name of perfect elimination schemes; see Tarjan and Yannakakis (1984). When algorithms were designed later to decide which arrows may be flipped in a given G^N_{dag}, keeping the same skeleton and the same set of sink Vs, to get to a list of all Markov equivalent G^N_{dag}'s, these early results by Tarjan and Yannakakis are not referred to directly; see Chickering (1995).

The number of equivalent characterizations of concentration graphs that have perfect elimination schemes has increased steadily, since they were introduced as rigid circuit graphs by Dirac (1961). These graphs are not only named 'chordal graphs', but also 'triangulated graphs', 'graphs with the running intersection property' or 'graphs with only complete prime graph separators'; see Cox and Wermuth (1999).

By contrast, for a covariance graph that can be oriented to be Markov equivalent to a G^N_{dag} of the same skeleton, chordless paths are relevant.

Proposition 4 (Pearl and Wermuth 1994) *A covariance graph with a chordless path in four nodes is not Markov equivalent to a directed acyclic graph in the same node set.*

For distributions generated over directed acyclic graphs, sink Vs are needed again.

Proposition 5 (Frydenberg 1990; Verma and Pearl 1990) *Directed acyclic graphs of the same skeleton are Markov equivalent if and only if they have the same sink* V*s*.

Markov equivalence of a concentration graph and a covariance graph model is for regular joint Gaussian distributions equivalent to *parameter equivalence*, which means that there is a one-to-one relation between the two sets parameters. Therefore, an early result on parameter equivalence for joint Gaussian distributions implies the following Markov equivalence result for distributions satisfying both the composition and the intersection property.

Proposition 6 (Jensen 1988; Drton and Richardson 2008b) *A covariance graph is Markov equivalent to a concentration graph if and only if both consist of the same complete, disconnected subgraphs*.

Fast ways of inserting an edge for every transition V, of deciding on connectivity and on blocking flows have been available in the corresponding Russian literature since 1970 (see Dinitz 2006), but these results appear to have not been exploited for the so-called lattice conditional independence models, recognized as distributions generated over G^N_{dag}'s without any transition Vs by Andersson et al. (1997).

Markov equivalence of other than multivariate regression chain graphs has been given by Roverato (2005), Andersson and Perlman (2006) and Roverato and Studený (2006).

With the so-called global Markov property of a graph in node set N and any disjoint subsets a, b, c of N, one can decide whether the graph implies $a \perp\!\!\!\perp b|c$. To give this property for a regression graph, we use special types of path that have been called active; see Wermuth (2011). For this, let again $\{a, b, c, m\}$ partition the node set N of G^N_{reg}.

Definition 1 *A path from a to b in* G^N_{reg} *is active given c* if its inner collision nodes are in c or have a descendant in c and its inner transmitting nodes are in $m = N \setminus (a \cup b \cup c)$. Otherwise, *the path is said to break given c* or, equivalently, to break with m.

Thus, a path breaks when c includes an inner transmitting node or when m includes an inner collision node and all its descendants; see also Fig. 4 of Marchetti and Wermuth (2009).

For directed acyclic graphs, an active path of Definition 1 reduces to the d-connecting path of Geiger et al. (1990). Similarly, the following proposition coincides in that special case with a statement concerning their d-separation. Let node set N of G^N_{reg} be partitioned as above by $\{a, b, c, m\}$.

Proposition 7 (Cox and Wermuth 1996; Sadeghi 2009) *A regression graph,* G^N_{reg}, *implies* $a \perp\!\!\!\perp b|c$ *if and only if every path between a and b breaks given c*.

Thus, whenever G^N_{reg} implies $a \perp\!\!\!\perp b|c$, this independence statement holds in the corresponding sequence of regressions for which the density f_N factorizes as (1),

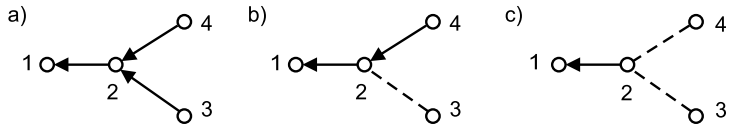


Fig. 12 Three regression graphs, which imply $3 \perp\!\!\!\perp 4$ but not $3 \perp\!\!\!\perp 4|1$

provided that f_N satisfies the same properties of independences, (i) to (vi) of Sect. 5, just like a regular Gaussian joint density. For example, in the graphs of Fig. 12, node 2 is an ancestor of node 1 so that G^N_{reg} does not imply $3 \perp\!\!\!\perp 4|2$.

Since covariance and concentration graphs consist only of one type of edge, the restricted versions in Propositions 1 and 2 of the defined path can be used for their global Markov property.

7 The main new results and proofs

We now treat connected regression graphs in node set N and corresponding distributions defined by sequences of regressions with joint discrete or continuous responses, ordered in connected components $g_1, \ldots, g_r$ of the graph, and with context variables in connected components, $g_{r+1}, \ldots, g_J$, which factorize as in (1), satisfy the pairwise independences of (2) as well as properties of independence statements, given as (i) to (vi) in Sect. 5.

For the main result of Markov equivalence for regression graphs, we consider distinct nodes i and k, node subsets c of $N \setminus \{i, k\}$ and the notion of shortest active paths.

Definition 2 An ik-path in G^N_{reg} is a shortest active path π with respect to c if every ik-path of G^N_{reg} with fewer inner nodes breaks given c.

Every chordless π is such a shortest path. If the consecutive nodes (k_{n-1}, k_n, k_{n+1}) on $\pi = (i = k_0, k_1, \ldots, k_m = k)$ induce a complete subgraph in G^N_{reg}, we say that there is *a triangle on the path*. In Fig. 13(a) nodes 2, 3, 4 form a triangle on the path $(1, 2, 4, 3, 5)$.

If this path is an active path connecting the uncoupled node pair $(1, 5)$, then nodes 2 and 4 are inner transmitting nodes outside c and the inner collision node 3 is in c. This path is then also the shortest active path connecting $(1, 5)$. The shorter path $(1, 2, 3, 5)$ has nodes 2 and 3 as inner transmitting nodes, but is inactive since node 3 is in c.

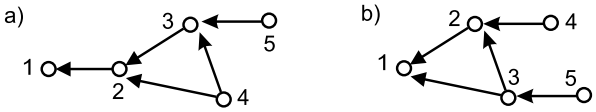


Fig. 13 Graphs of active five-node paths (**a**) with path $(1, 2, 4, 3, 5)$ the shortest active path, where 3 is in c, (**b**) active path $(4, 2, 1, 3, 5)$, where 1 is in c, and a shorter active path $(4, 2, 3, 5)$

By contrast in Fig. 13(b), when path $(4, 2, 1, 3, 5)$ is an active path connecting the uncoupled node pair $(4, 5)$, then path $(4, 2, 3, 5)$ is a shorter active path. To see this, notice that on an active $(4, 2, 1, 3, 5)$ path, the inner collision node 1 is in c and the inner transmitting nodes 2 and 3 are outside c. In this case, the inner collision node 2 on the path $(4, 2, 3, 5)$ has node 1 as a descendant in c, so that this shorter path is also active.

We also use the following results for proving Theorem 1. The first two are direct consequences of Proposition 7 and imply the pairwise independences of Eq. (2). Lemma 4 results with the independence form of (2). Let h, i, k be distinct nodes of N.

Lemma 2 *For (h, i, k) a collision V in G^N_{reg}, the inner node i is excluded from c in every independence statement for h, k implied by G^N_{reg}.*

Lemma 3 *For (h, i, k) a transmitting V in G^N_{reg}, the inner node i is included in c in every independence statement for h, k implied by G^N_{reg}.*

Lemma 4 *A missing ik-edge in G^N_{reg} implies at least one independence statement $i \perp\!\!\!\perp k|c$ for c a subset of $N \setminus \{i, k\}$.*

We can now derive the first of the main new results in this paper.

Theorem 1 *Two regression graphs are Markov equivalent if and only if they have the same skeleton and the same sets of collision* Vs*, irrespective of the type of edge.*

Proof Regression graphs G^N_{reg1} and G^N_{reg2} are Markov equivalent if and only if for every disjoint subset a, b, and c of the node set of N, where only c can be empty,

$$\left(G^N_{\text{reg1}} \implies a \perp\!\!\!\perp b|c\right) \iff \left(G^N_{\text{reg2}} \implies a \perp\!\!\!\perp b|c\right). \tag{11}$$

Suppose first that (11) holds. By Lemma 4, G^N_{reg1} and G^N_{reg2} have the same skeleton, and by Lemmas 2 and 3, G^N_{reg1} and G^N_{reg2} have the same collision Vs.

Suppose next that G^N_{reg1} and G^N_{reg2} have the same skeleton and the same collision Vs and consider two arbitrary distinct nodes i and k and any node subset c of $N \setminus \{i, k\}$. By Proposition 7, (11) is equivalent to stating that for every uncoupled node pair i, k, there is an active path with respect to c in G^N_{reg1} if and only if there is an active ik-path with respect to c in G^N_{reg2}.

Suppose further that path π is in G^N_{reg1} a shortest active ik-path with respect to c. Since G^N_{reg1} and G^N_{reg2} have the same skeleton, the path π exists in G^N_{reg2}. We need to show that it is active. If all consecutive two-edge subpaths of π are Vs then π is active in G^N_{reg2}. Therefore, suppose that nodes (k_{n-1}, k_n, k_{n+1}) on π form a triangle instead of a V. It may be checked first, that in all other possible triangles in regression graphs that can appear on π other than the two of Fig. 14, there is as in Fig. 13(b) a shorter active path. To complete the proof, we show that for the two types of triangles

Fig. 14 The two types of triangles in regression graphs without a shorter active path whenever the path with inner nodes (k_{n+1}, k_n, k_{n-1}) is active

shown in parts (a) and (b) of Fig. 14, path π is also in G^N_{reg2} an active ik-path with respect to c.

In G^N_{reg1} containing the triangle of Fig. 14(a) on a shortest active path π, node k_n is a transmitting node, which is by Lemma 3 outside c. By Lemma 2, node k_{n-1} is a collision node inside c. If instead k_{n-1} were a transmitting node on π in G^N_{reg1}, it would also be a transmitting node on $(k_{n-2}, k_{n-1}, k_{n+1})$ and give a shorter active path via the $k_{n-1}k_{n+1}$-edge, contradicting the assumption of π being a shortest path. Similarly, if collision node k_{n-1} on π were only an ancestor of c, then there were a shorter active path via the $k_{n-1}k_{n+1}$-edge.

In addition, node pair k_n, k_{n-2} is uncoupled in G^N_{reg1} since by inserting any such edge that is permissible in a regression graph, another shortest path via the $k_{n-2}k_n$-edge would result. Therefore, since G^N_{reg1} and G^N_{reg2} have the same collision Vs, the subpath (k_{n-2}, k_{n-1}, k_n) forms also a collision V in G^N_{reg2}. Similarly, $(k_{n-2}, k_{n-1}, k_{n+1})$ is a transmitting V and (k_{n+2}, k_{n+1}, k_n) is a V of either type. Hence k_{n-1} is a parent of k_{n+1} in G^N_{reg2} and the only permissible edge between k_n and k_{n+1} is an arrow pointing to k_{n+1}. Therefore, π forms an active path also in G^N_{reg2}.

The proof for Fig. 14(b) is the same as for Fig. 14(a) since the type of nodes along π, i.e. as collision or transmitting nodes, is unchanged. □

In the example of Fig. 15, all three regression graphs have the same skeleton. In G^N_{reg1} there are three collision Vs: $(3, 4, 5)$, $(1, 2, 5)$, and $(2, 1, 3)$. In G^N_{reg2} there are the same collision Vs. Therefore, these two graphs are Markov equivalent. However, there are only two collision Vs in G^N_{reg3}, and these are $(3, 4, 5)$ and $(2, 1, 3)$. Hence this graph is not Markov equivalent to G^N_{reg1} and G^N_{reg2}. The Markov equivalence of the graphs in Fig. 2 to the subgraph induced by $\{b, c\}$ in Fig. 1 is a further application of Theorem 1. Notice that Propositions 3 to 8 of Sect. 6 result as special cases of Theorem 1.

The following algorithm generates a directed acyclic graph from a given G^N_{reg} that fulfils its known necessary conditions for Markov equivalence to a directed acyclic

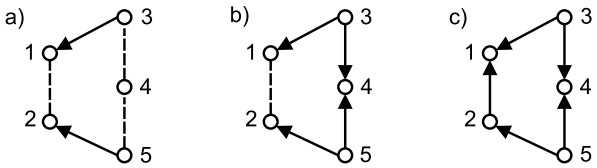


Fig. 15 (**a**) Regression graph G^N_{reg1}, (**b**) a Markov equivalent regression graph G^N_{reg2} to G^N_{reg1}, (**c**) a regression graph G^N_{reg3} that is directed acyclic and not Markov equivalent to G^N_{reg1}

graph; see Proposition 2 of Wermuth (2011). We refer to these connected components as the blocks of G^N_{reg}.

Algorithm 1 (Obtaining a Markov equivalent directed acyclic graph from a regression graph) Start from any given G^N_{reg} that has a chordal concentration graph and no chordless collision path in four nodes.

1. Apply the maximum cardinality search algorithm on the block consisting of full lines to order the nodes of the block.
2. Orient the edges of the block from a higher number to a lower one.
3. Replace collision Vs by sink Vs, i.e. replace $i - - - \circ - - - k$ and $i - - - \circ \prec\!\!- k$ by $i \longrightarrow \circ \prec\!\!- k$ when i and k are uncoupled. When a dashed line in a block is replaced by an arrow, label the endpoints such that the arrow is from a higher number to a lower one if the labels do not already exist.
4. Replace dashed lines $i - - - \circ - - - k$ of triangles by a sink path $i \longrightarrow \circ \prec\!\!- k$. When a dashed line in a block is replaced by an arrow, label the endpoints such that the arrow is from a higher number to a lower one if the labels do not already exist.
5. Replace dashed lines by arrows from a higher number to a lower one.

Continually apply each step until it is not possible to continue applying it further. Then move to the next step.

Lemma 5 *For a regression graph with a chordal concentration graph and without chordless collision paths in four nodes, Algorithm 1 generates a directed acyclic graph that is Markov equivalent to* G^N_{reg}.

Proof The generated graph is directed since, by Algorithm 1, all edges are turned into arrows. Since the block containing full lines is chordal, the graph generated by the perfect elimination order of the maximal cardinality search does not have a directed cycle; see Blair and Peyton (1993), Sect. 2.4 and Tarjan and Yannakakis (1984).

In addition, the arrows present in the graph do not change by the algorithm. Thus, to generate a cycle containing an arrow of the original graph, there should have been a cycle in the directed graph generated by replacing blocks by nodes. But, this is impossible in a regression graph. Therefore in the generated graph, there is no cycle containing arrows that have been between the blocks of the original graph.

Within a block, all arrows point from nodes with higher numbers to nodes with lower ones. Otherwise, there would have been at step 3 of the algorithm a chordless collision path with four nodes in the graph. Hence no directed cycle can be generated.

Theorem 1 gives Markov equivalence to G^N_{reg} since Algorithm 1 preserves the skeleton of G^N_{reg} and no additional collision V is generated because sink oriented Vs remain, only dashed lines are turned into arrows and no arrows are changed to dashed lines. □

Notice that this algorithm does not generate a unique directed acyclic graph, but every generated directed acyclic graph is Markov equivalent to the given regression graph. To obtain the overall complexity of Algorithm 1, we denote by n the number of nodes in the graph and by e the number of edges in the graph.

Corollary 4 *The overall complexity of Algorithm 1 is* $O(e^3)$.

Proof Suppose that the input of Algorithm 1 is a sequence of triples, each of which consists of the two endpoints of an edge and of the type of edge. The length of this sequence is equal to e and the highest number appearing in the sequence is n. For example, the sequence to the graph of Fig. 15(a) is $((1, 2, d), (3, 1, a), (5, 2, a), (4, 3, d), (4, 5, d))$, where '$d$' corresponds to a dashed line and 'a' corresponds to an arrow pointing from the first entry to the second one. Notice that this labelling is in general not the same as the ordering of nodes given by Algorithm 1.

The first two steps of Algorithm 1 can be performed in $O(e + n)$ time; see Blair and Peyton (1993). Step 3 of Algorithm 1 may be performed in $e(e + 1)(e - 2)/2$ steps since for each edge, one can go through the edge set to find the edges that give a three-node path with an inner collision node. This needs $e(e + 1)/2$ steps. For each collision node, one goes again through the edge set, excluding the two edges involved in the collision path, to check if the collision is a V. Other actions can be done in constant time.

Step 4 may require $ne(e + 1)/2$ steps since paths considered $\circ\text{---}\circ\text{---}\circ$ which do not form a V. Therefore, there is no reason to go through the edge set for the third time, but one might need to go through the node ordering to decide on the direction of the generated arrow. The last step may be performed with ne steps by going through the edge set changing 'd's to 'a's appropriately by looking at the node ordering. Therefore, the overall complexity of Algorithm 1 is $O(e^3)$. □

Corollary 2 and Propositions 4 to 8 can now be derived as special cases of Theorem 1 and Lemma 4. In addition by using Lemmas 1, 2 and pairwise independences, subclasses of regression graphs can be identified, which intersect with directed acyclic graphs, with other types of chain graphs, with concentration graphs or with covariance graphs.

Theorem 2 *A regression graph with a chordal graph for the context variables can be oriented to be Markov equivalent to a directed acyclic graph in the same skeleton, if and only if it does not contain any chordless collision path in four nodes.*

Proof Every chordal concentration graph can be oriented to be equivalent to a directed acyclic graph; see Tarjan and Yannakakis (1984). A missing edge for node pair $i < k$ in a directed acyclic graph means $i \perp\!\!\!\perp k | > i \setminus k$, which would contradict (2)(iii) if the graph contained a semi-directed chordless collision path in four nodes. No undirected chordless collision path in four nodes can be fully oriented without changing a collision V into a transmitting V, but G^N_{reg} can be oriented using Algorithm 1 if it contains no such path. □

Notice that for joint Gaussian distributions, Theorem 2 excludes Zellner's seemingly unrelated regressions and it excludes covariance graphs that cannot be made Markov equivalent to fully directed acyclic graphs; see Proposition 4.

Proposition 8 *A multivariate regression graph with connected components $g_1, \ldots, g_J$ is an AMP chain graph in the same connected components if and only if the covariance graph of every connected component of responses is complete.*

Proof The conditional relations of the joint response nodes in an AMP chain graph coincide with those of the regression graph with the same connected components. Furthermore, the subgraph induced by each connected component g_j of an AMP chain graph is a concentration graph given $g_{>j}$ while in G^N_{reg} it is a covariance graph given $g_{>j}$. By Proposition 6, these have to be complete for Markov equivalence. □

Proposition 9 *A multivariate regression graph with connected components $g_1, \ldots, g_J$ is an LWF chain graph in the same connected components if and only if it contains no semi-directed chordless collision path in four nodes and the covariance graph of every connected component of responses is complete.*

Proof The proof for the connected components of an LWF chain graph is the same as for an AMP chain graph since they both have concentration graphs for g_j given $g_{>j}$. The dependences of joint responses g_j on $g_{>j}$ coincide in an LWF chain graph with the bipartite part of the concentration graph in $g_j \cup g_{>j}$ so that Markov equivalent independence statements can only hold with these bipartite graphs being complete. □

Figure 16 illustrates Theorem 2 and Proposition 8, 9 with modified graphs of Fig. 4. The graphs in Fig. 16 are Markov equivalent to (a) a directed acyclic graph with the same skeleton obtainable by Algorithm 1, (b) an AMP chain graph in the same connected components, and (c) an LWF chain graph in the same connected components.

In general, by inserting some edges, a regression graph model can be turned into a model in one of the intersecting classes used in Propositions 2 to 9, just as a non-chordal graph may be turned into a chordal one by adding edges. When the independence structure of interest is captured by an edge-minimal regression graph, then the resulting graph after adding edges will no longer be an edge-minimal graph and hence will not give the most compact graphical description possible.

However, the graph with some added edges may define a covering model that is easier to fit than the reduced model corresponding to the edge-minimal graph, just as an unconstrained Gaussian bivariate response regression on two regressors may be fitted in closed form, while the maximum-likelihood fitting in the reduced model of Zellner's seemingly unrelated regression requires iterative fitting algorithms. Any well-fitting covering model in the three intersecting classes will show week dependences for the edges that are to be removed to obtain an edge-minimal graph.

Notice that sequences of regressions in the intersecting class with LWF chain graphs correspond for Gaussian distributions to sequences of the general linear models of Anderson (1958), Chap. 8, that is, to models in which each joint response has the same set of regressor variables. This shows in G^N_{reg} by identical sets of nodes from which arrows point to each node within a connected component.

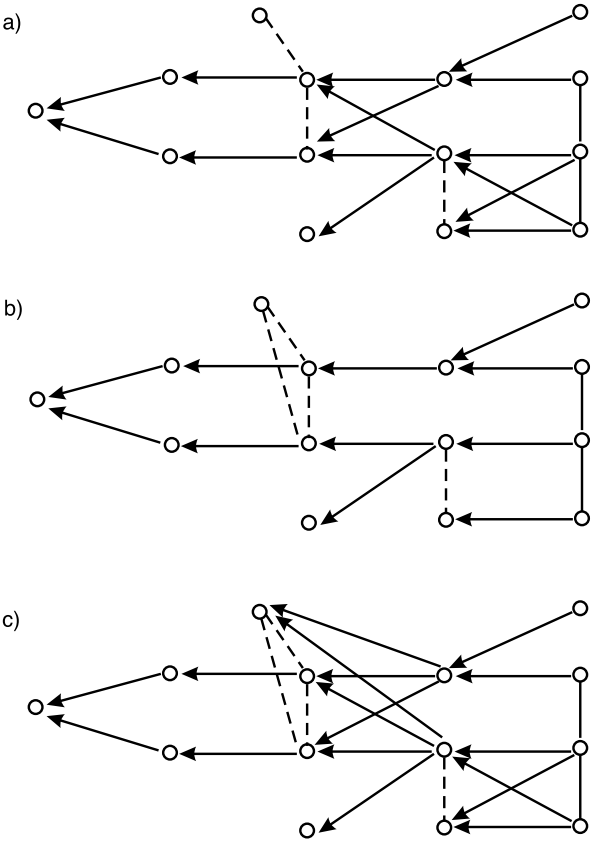


Fig. 16 The graph of Fig. 4 modified by adding edges to obtain a graph that is Markov equivalent to (**a**) a directed acyclic graph, (**b**) an AMP chain graph in the same connected components, and (**c**) an LWF chain graph in the same connected components

In contrast, the models in the intersecting classes with the two types of undirected graph may be quite complex in the sense of including many merely generated chordless cycles of size four or larger.

Proposition 10 *A multivariate regression graph has the skeleton concentration graph if and only if it contains no collision* V *and it has the skeleton covariance graph if and only if it contains no transmitting* V.

Proof Every V is a collision V in a covariance graph and a transmitting V in a concentration graph; see Lemmas 1 and 2. The first includes, the second excludes the inner node from the defining independence statement. Thus, in the presence of a V, one would contradict the uniqueness of the defining pairwise independences. □

Lastly, Fig. 17 shows the overall concentration graph induced by G^N_{reg} of Fig. 4. It may be obtained from the given G^N_{reg} by finding first the smallest covering LWF chain graph in the same connected components, then closing every sink V by an edge, i.e. adding an edge between its endpoints, and finally changing all edges to full lines.

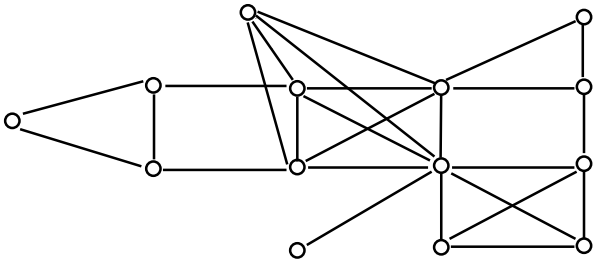


Fig. 17 The overall concentration graph induced by the regression graph in Fig. 4

In such a graph, several chordless cycles in four or more nodes may be induced and the connected components of G^N_{reg} may no longer show. In such a case, much of the important structure of the generating regression graph is lost. In addition, merely induced chordless cycles require iterative algorithms for maximum-likelihood estimation, even for Gaussian distributions. Thus, in the case of connected joint responses, it may be unwise to use a model search within the class of concentration graph models.

This contrasts with LWF chain graphs that coincide with regression graphs, such as in Fig. 16(c). These preserve the available prior knowledge about the connected components and give Markov equivalence to directed acyclic graphs so that model fitting is possible in terms of single response regressions, that is, by using just univariate conditional densities. In addition, the simplified criteria for Markov equivalence of directed acyclic graphs apply.

On the other hand, sequences of regressions that coincide with LWF chains, permit us to model simultaneous intervention on a set of variables since the corresponding independence graphs are directed and acyclic in nodes representing vector variables. This represents a conceptually much needed extension of distributions generated over directed acyclic graphs in nodes representing single variables, but excludes the more specialized seemingly unrelated regressions and incomplete covariance graphs.

Acknowledgements The work of the first author has been supported in part by the Swedish Research Society via the Gothenburg Stochastic Centre and by the Swedish Strategic Fund via the Gothenburg Mathematical Modelling Centre. We thank R. Castelo, D.R. Cox, G. Marchetti and the referees for their most helpful comments.

Appendix: Details of regressions for the chronic pain data

Tables 1–8 show the results of linear least-squares regressions or logistic regressions, one at a time, for each of the response variables and for each component of a joint response separately. At first, each response is regressed on all its potentially explanatory variables given by their first ordering. The tables give the estimated constant term and for each variable in the regression, its estimated coefficient (coeff), the estimated standard deviation of the coefficient (s_{coeff}), as well as the ratio $z_{\text{obs}} = \text{coeff}/s_{\text{coeff}}$. These ratios are compared with 2.58, the 0.995 quantile of a random variable Z having a standard Gaussian distribution, for which $\Pr(|Z| > 2.58) = 0.01$. In backward selection steps, the variable with the smallest observed value $|z_{\text{obs}}|$ is deleted from a regression equation, one at a time, until the threshold is reached.

Table 1 Response: Y, success of treatment; linear regression including a quadratic term

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
Constant	23.40	–	–	20.50	–	–	–
Z_a, pain intensity after	−1.73	0.15	–11.19	−1.89	0.15	−12.77	–
X_a, depression after	−0.16	0.05	−3.04	–	–	–	−1.86
Z_b, pain intensity before	0.04	0.16	0.26	–	–	–	0.65
X_b, depression before	0.10	0.05	1.82	–	–	–	0.33
U, pain chronicity	−0.15	0.30	−0.51	–	–	–	−0.99
A, site of pain	−2.27	0.91	−2.48	–	–	–	−2.33
V, previous illnesses	0.19	0.11	1.76	–	–	–	1.24
B, level of schooling	−0.50	0.78	−0.64	–	–	–	−0.22
$(Z_a - \text{mean}(Z_a))^2$	0.18	0.23	3.41	0.23	0.05	4.28	–

$R^2_{\text{full}} = 0.54$; Selected model $Y : Z_a + Z_a^2$; $R^2_{\text{sel}} = 0.49$

Table 2 Response: Z_a, intensity of pain after treatment; linear regression

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
Constant	2.74	–	–	2.98	–	–	–
Z_b, pain intensity before	0.12	0.08	1.60	0.16	0.07	2.16*	–
X_b, depression before	0.03	0.02	1.28	–	–	–	1.76
U, pain chronicity	0.11	0.14	0.75	–	–	–	1.43
A, site of pain	1.07	0.42	2.51	1.27	0.39	3.26	–
V, previous illnesses	0.00	0.05	0.03	–	–	–	0.83
B, level of schooling	−0.19	0.37	−0.52	–	–	–	−0.70

$R^2_{\text{full}} = 0.09$; Selected model $Z_a : Z_b + A$; $R^2_{\text{sel}} = 0.07$

*: Depression before treatment needed because of the repeated measurement design; the low correlation for Z_a, Z_b is due to a change in measuring, before and after treatment

The procedure defines a selected model, unless one of the excluded variables has a contribution of $|z'_{\text{obs}}| > 2.58$ when added alone to the selected directly explanatory variables; then such a variable needs also to be included as an important directly explanatory variable. This did not happen in the given data set.

The tables show for linear models also R^2, the coefficient of determination, both for the full and for the selected model. Multiplied by 100, it gives the percentage of the variation in the response explained by the model.

In the linear regression of Z_a on X_a and on the directly explanatory variables of both Z_a and X_a, that is, on Z_b, X_b, A, the contribution of X_a leads to $z_{\text{obs}} = 3.51$, which coincides—by definition—with z_{obs} computed for the contribution of Z_a in the linear regression of X_a on Z_a and on Z_b, X_b, A. Hence the two responses are

Table 3 Response: X_a, depression after treatment; linear regression

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
Constant	2.54	–	–	4.55	–	–	–
Z_b, pain intensity before	−0.05	0.22	−0.23	–	–	–	−0.21
X_b, depression before	0.62	0.06	10.43	0.68	0.05	12.68	–
U, pain chronicity	0.96	0.42	2.28	–	–	–	2.31
A, site of pain	−1.19	1.25	−0.95	–	–	–	−0.10
V, previous illnesses	0.05	0.15	0.35	–	–	–	1.08
B, level of schooling	0.15	1.09	0.14	–	–	–	−0.01

$R^2_{\text{full}} = 0.46$; Selected model $X_a : X_b$; $R^2_{\text{sel}} = 0.45$

Table 4 Response: Z_b, intensity of pain before; linear regression

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
Constant	7.60	–	–	7.38	–	–	–
U, pain chronicity	0.10	0.13	0.77	–	–	–	0.59
A, site of pain	−0.58	0.40	−1.44	–	–	–	−1.20
V, previous illnesses	0.02	0.05	0.46	–	–	–	0.72
B, level of schooling	−0.94	0.35	−2.70	−0.89	0.33	−2.65	–

$R^2_{\text{full}} = 0.05$; Selected model $Z_a : B$; $R^2_{\text{sel}} = 0.03$

Table 5 Response: X_b, depression before; linear regression

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
Constant	10.96	–	–	7.31	–	–	–
U, pain chronicity	1.97	0.49	4.02	1.78	0.46	3.87	–
A, site of pain	−2.33	1.50	−1.55	–	–	–	−1.42
V, previous illnesses	0.54	0.18	2.99	0.55	0.18	3.06	–
B, level of schooling	−1.10	1.31	−0.84	–	–	–	−0.57

$R^2_{\text{full}} = 0.18$; Selected model $X_b : U + V$; $R^2_{\text{sel}} = 0.17$

correlated even after considering the directly explanatory variables and a dashed line joining Z_a and Z_b is added to the well-fitting regression graph in Fig. 8.

In the linear regression of Z_b on X_b and on the directly explanatory variables of both Z_b and X_b, that is, on U, A, V, B, the contribution of X_b leads to $z_{\text{obs}} = 2.64$. Hence the two responses are associated after considering their directly explanatory variables and there is a dashed line joining Z_b and X_b in the regression graph of Fig. 8.

Table 6 Response: U, chronicity of pain; linear regression

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
Constant	2.93	–	–	2.47	–	–	–
A, site of pain	0.95	0.21	4.58	1.02	0.20	5.02	–
V, previous illnesses	0.14	0.02	5.83	0.14	0.02	5.92	–
B, level of schooling	−0.27	0.19	−1.43	–	–	–	−1.43

$R^2_{\text{full}} = 0.26$; Selected model $X_b : A + V$; $R^2_{\text{sel}} = 0.25$

Table 7 Response: A, site of pain; logistic regression

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
Constant	0.26	–	–	0.60	–	–	–
V, previous illnesses	0.05	0.04	1.22	–	–	–	1.22
B, level of schooling	−1.25	0.40	−3.11	−1.28	0.40	−3.18	–

Selected model $A : B$; response recoded to $(0, 1)$ instead of $(1, 2)$

Table 8 Response: V, previous illnesses; linear regression

Explanatory variables	Starting model			Selected			Excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
Constant	6.41	–	–	5.53	–	–	–
B, level of schooling	−0.65	0.54	−1.20	–	–	–	–

Selected model V : –

The relatively strict criterion, for excluding variables, assures that all edges in the derived regression graph correspond to dependences that are considered to be substantive in the given context. Had instead a 0.975 quantile been chosen as threshold, then one arrow from A to Y and another from U to X_a would have been added to the regression graph. Although this would correspond to a better goodness-of-fit, such weak dependences are less likely to become confirmed as being important in follow-up studies.

The subgraph induced by Z_a, Z_b, X_a, X_b of the regression graph in Fig. 8 corresponds to two seemingly unrelated regressions. Even though separate least-squares estimates can in principle be severely distorted, for the present data, the structure is so well-fitting in the unconstrained multivariate regression of Z_a and X_a on Z_b, X_b, U, V, A, B, that is, in a simple covering model, that none of these potential problems are relevant.

With $C = \{U, V, A, B\}$, this is evident from the observed covariance matrix of Z_a, X_a given Z_b, X_b, C, denoted here by $\tilde{\Sigma}_{aa|bC}$ and the observed regression coeffi-

cient matrix $\tilde{\Pi}_{a|b.C}$ being almost identical to the corresponding maximum likelihood estimators $\hat{\Sigma}_{aa|bC}$ and $\hat{\Pi}_{a|b.C}$.

The former can be obtained by sweeping or partially inverting the observed covariance matrix of the eight variables with respect to Z_b, X_b, C and the latter by using an adaption of the EM-algorithm, due to Kiiveri (1987), on the observed covariance matrix of the four symptoms, corrected for linear regression on C. In this way, one gets

$$\tilde{\Sigma}_{aa|bC} = \begin{pmatrix} 5.61 & 3.91 \\ 3.91 & 48.37 \end{pmatrix}, \qquad \hat{\Sigma}_{aa|bC} = \begin{pmatrix} 5.66 & 3.94 \\ 3.94 & 48.41 \end{pmatrix},$$

$$\tilde{\Pi}_{a|b.C} = \begin{pmatrix} 0.12 & 0.03 \\ -0.05 & 0.62 \end{pmatrix}, \qquad \hat{\Pi}_{a|bC} = \begin{pmatrix} 0.14 & 0.00 \\ 0.00 & 0.60 \end{pmatrix}.$$

The assumed definition of the joint distribution in terms of univariate and multivariate regressions assures that the overall fit of the model can be judged locally in two steps. First, one compares each unconstrained, full regression of a single response with regressions constrained by some independences, that is, by selecting a subset of directly explanatory variables from the list of the potentially explanatory variables. Next, one decides for each component pair of a joint response whether this pair is conditionally independent given their directly explanatory variables considered jointly. This can again be achieved by single univariate regressions, as illustrated above for the joint responses Z_a and X_a.

References

Ali RA, Richardson TS, Spirtes P (2009) Markov equivalence for ancestral graphs. Ann Stat 37:2808–2837

Anderson TW (1958) An introduction to multivariate statistical analysis. Wiley, New York (3rd edn, 2003)

Anderson TW (1973) Asymptotically efficient estimation of covariance matrices with linear structure. Ann Stat 1:135–141

Andersson SA, Perlman MD (2006) Characterizing Markov equivalence classes for AMP chain graph. Ann Stat 34:939–972

Andersson SA, Madigan D, Perlman MD, Triggs CM (1997) A graphical characterization of lattice conditional independence models. Ann Math Artif Intell 21:27–50

Andersson SA, Madigan D, Perlman MD (2001) Alternative Markov properties for chain graphs. Scand J Stat 28:33–86

Barndorff-Nielsen OE (1978) Information and exponential families in statistical theory. Wiley, Chichester

Bergsma W, Rudas T (2002) Marginal models for categorical data. Ann Stat 30:140–159

Birch MW (1963) Maximum likelihood in three-way contingency tables. J R Stat Soc B 25:220–233

Bishop YMM, Fienberg SF, Holland PW (1975) Discrete multivariate analysis. MIT Press, Cambridge

Blair JRS, Peyton BW (1993) An introduction to chordal graphs and clique trees. In: George JA, Gilbert JR, Liu JWH (eds) Graph theory and sparse matrix computations. IMA volumes in mathematics and its applications, vol 56. Springer, New York, pp 1–30

Brito C, Pearl J (2002) A new identification condition for recursive models with correlated errors. Struct Equ Model 9:459–474

Bollen KA (1989) Structural equations with latent variables. Wiley, New York

Brown LD (1986) Fundamentals of statistical exponential families with applications in statistical decision theory. LNMS, vol 9. Inst Math Stat, Beachwood

Castelo R, Kocka T (2003) On inclusion-driven learning of Bayesian networks. J Mach Learn Res 4:527–574

Castelo R, Siebes A (2003) A characterization of moral transitive acyclic directed graph Markov models as labeled trees. J Stat Plan Inference 115:235–259
Caussinus H (1966) Contribution á l'analyse statistique des tableaux de corrélation. Ann Fac Sci Univ Toulouse 29:77–183
Cayley A (1889) A theorem on trees. Q J Math 23:376–378
Chaudhuri S, Drton M, Richardson TS (2007) Estimation of a covariance matrix with zeros. Biometrika 94:199–216
Cochran WG (1938) The omission or addition of an independent variate in multiple linear regression. Suppl J R Stat Soc 5:171–176
Chickering DM (1995) A transformational characterization of equivalent Bayesian networks. In: Besnard P, Hanks S (eds) Proc 10th UAI conf. Kaufman, San Mateo, pp 87–98
Cox DR (1966) Some procedures associated with the logistic qualitative response curve. In: David FN (ed) Research papers in statistics: Festschrift for J Neyman. Wiley, New York, pp 55–71
Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge
Cox DR, Wermuth N (1990) An approximation to maximum-likelihood estimates in reduced models. Biometrika 77:747–761
Cox DR, Wermuth N (1993) Linear dependencies represented by chain graphs (with discussion). Stat Sci 8:204–218; 247–277
Cox DR, Wermuth N (1994) Tests of linearity, multivariate normality and adequacy of linear scores. J R Stat Soc C 43:347–355
Cox DR, Wermuth N (1996) Multivariate dependencies: models, analysis, and interpretation. Chapman and Hall/CRC Press, London
Cox DR, Wermuth N (1999) Likelihood factorizations for mixed discrete and continuous variables. Scand J Stat 26:209–220
Cox DR, Wermuth N (2003) A general condition for avoiding effect reversal after marginalization. J R Stat Soc B 65:937–941
Darroch JN (1962) Interactions in multi-factor contingency tables. J R Stat Soc B 24:251–263
Darroch JN, Lauritzen SL, Speed TP (1980) Markov fields and log-linear models for contingency tables. Ann Stat 8:522–539
Dawid AP (1979) Conditional independence in statistical theory (with discussion). J R Stat Soc B 41:1–31
Dempster AP (1969) Elements of continuous multivariate analysis. Addison-Wesley, Reading
Dempster AP (1972) Covariance selection. Biometrics 28:157–175
Dinitz Y (2006) Dinitz' algorithm: the original version and even's version. In: Even S, Goldreich O, Rosenberg AL, Selman AL (eds) Essays in memory of Shimon Even. Springer, New York, pp 218–240
Dirac GA (1961) On rigid circuit graphs. Abh Math Semin Univ Hamb 25:71–76
Drton M (2009) Discrete chain graph models. Bernoulli 15:736–753
Drton M, Perlman MD (2004) Model selection for Gaussian concentration graphs. Biometrika 91:591–602
Drton M, Richardson TS (2004) Multimodality of the likelihood in the bivariate seemingly unrelated regression model. Biometrika 91:383–392
Drton M, Richardson TS (2008a) Binary models for marginal independence. J R Stat Soc B 70:287–309
Drton M, Richardson TS (2008b) Graphical methods for efficient likelihood inference in Gaussian covariance models, J. J Mach Learn Res 9:893–914
Edwards D (2000) Introduction to graphical modelling, 2nd edn. Springer, New York
Foygel R, Draisma J, Drton M (2011) Half-trek criterion for generic identifiability of linear structural equation models (submitted). Available under http://arxiv.org/abs/1107.5552
Frydenberg M (1990) The chain graph Markov property. Scand J Stat 17:333–353
Geiger D, Verma TS, Pearl J (1990) Identifying independence in Bayesian networks. Networks 20:507–534
Glonek GFV, McCullagh P (1995) Multivariate logistic models. J R Stat Soc B 53:533–546
Goodman LA (1970) The multivariate analysis of qualitative data: interaction among multiple classifications. J Am Stat Assoc 65:226–256
Haavelmo T (1943) The statistical implications of a system of simultaneous equations. Econometrica 11:1–12
Hardt J, Sidor A, Nickel R, Kappis B, Petrak F, Egle UT (2008) Childhood adversities and suicide attempts: a retrospective study. J Fam Violence 23:713–718
Jensen ST (1988) Covariance hypotheses which are linear in both the covariance and the inverse covariance. Ann Stat 16:302–322
Jöreskog KG (1981) Analysis of covariance structures. Scand J Stat 8:65–92

Kang C, Tian J (2009) Markov properties for linear causal models with correlated errors. J Mach Learn Res 10:41–70
Kappesser J (1997) Bedeutung der Lokalisation für die Entwicklung und Behandlung chronischer Schmerzen. Thesis, Department of Psychology, University of Mainz
Kauermann G (1996) On a dualization of graphical Gaussian models. Scand J Stat 23:115–116
Kiiveri HT (1987) An incomplete data approach to the analysis of covariance structures. Psychometrika 52:539–554
Kiiveri HT, Speed TP, Carlin JB (1984) Recursive causal models. J Aust Math Soc A 36:30–52
Kline RB (2006) Principles and practice of structural equation modeling, 3rd edn. Guilford Press, New York
Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford
Lauritzen SL, Wermuth N (1989) Graphical models for association between variables, some of which are qualitative and some quantitative. Ann Stat 17:31–57
Lehmann EL, Scheffé H (1955) Completeness, similar regions and unbiased estimation. Sankhya 14:219–236
Lněnička R, Matúš F (2007) On Gaussian conditional independence structures. Kybernetika 43:323–342
Lupparelli M, Marchetti GM, Bergsma WP (2009) Parameterization and fitting of discrete bi-directed graph models. Scand J Stat 36:559–576
Ma ZM, Xie XC, Geng Z (2006) Collapsibility of distribution dependence. J R Stat Soc B 68:127–133
Mandelbaum A, Rüschendorf L (1987) Complete and symmetrically complete families of distributions. Ann Stat 15:1229–1244
Marchetti GM, Lupparelli M (2011) Chain graph models of multivariate regression type for categorical data. Bernoulli 17:845–879
Marchetti GM, Wermuth N (2009) Matrix representations and independencies in directed acyclic graphs. Ann Stat 47:961–978
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall/CRC Press, London
Nelder JA, Wedderburn R (1972) Generalized linear models. J R Stat Soc, A 135:37–384
Pearl J (1988) Probabilistic reasoning in intelligent systems. Kaufmann, San Mateo
Pearl J (2009) Causality: models, reasoning, and inference, 2nd edn. Cambridge University Press, New York
Pearl J, Paz A (1987) Graphoids: a graph based logic for reasoning about relevancy revelations. In: Boulay BD, Hogg D, Steel L (eds) Advances in artificial intelligence II. North Holland, Amsterdam, pp 357–363
Pearl J, Wermuth N (1994) When can association graphs admit a causal interpretation? In: Cheeseman P, Oldford W (eds) Models and data, artificial intelligence and statistics IV. Springer, New York, pp 205–214
Richardson TS, Spirtes P (2002) Ancestral Markov graphical models. Ann Stat 30:962–1030
Roverato A (2005) A unified approach to the characterisation of Markov equivalence classes of directed acyclic graphs, chain graphs with no flags and chain graphs. Scand J Stat 32:295–312
Roverato A, Studený M (2006) A graphical representation of equivalence classes of AMP chain graphs. J Mach Learn Res 7:1045–1078
Rudas T, Bergsma WP, Nemeth R (2010) Marginal log-linear parameterization of conditional independence models. Biometrika 97:1006–1012
Sadeghi K (2009) Representing modified independence structures. Transfer thesis, Oxford University
Sadeghi K, Lauritzen SL (2012) Markov properties of mixed graphs (submitted). Also available on http://arxiv.org/abs/1109.5909
San Martin E, Mouchart M, Rolin JM (2005) Ignorable common information, null sets and Basu's first theorem. Sankhya 67:674–698
Speed TP, Kiiveri HT (1986) Gaussian Markov distributions over finite graphs. Ann Stat 14:138–150
Spirtes P, Glymour C, Scheines R (1993) Causation, prediction and search. Springer, New York
Stanghellini E, Wermuth N (2005) On the identification of path analysis models with one hidden variable. Biometrika 92:337–350
Studený M (2005) Probabilistic conditional independence structures. Springer, London
Sundberg R (2010) Flat and multimodal likelihoods and model lack of fit in curved exponential families. Scand J Stat 37:632–643
Tarjan RE, Yannakakis M (1984) Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. SIAM J Comput 13:566–579

Verma T, Pearl J (1990) Equivalence and synthesis of causal models. In: Bonissone PP, Henrion M, Kanal LN, Lemmer JF (eds) Proc 6th UAI conf. Elsevier, Amsterdam, pp 220–227
Wermuth N (1976a) Analogies between multiplicative models for contingency tables and covariance selection. Biometrics 32:95–108
Wermuth N (1976b) Model search among multiplicative models. Biometrics 32:253–263
Wermuth N (1980) Linear recursive equations, covariance selection, and path analysis. J Am Stat Assoc 75:963–997
Wermuth N (2011) Probability models with summary graph structure. Bernoulli 17:845–879
Wermuth N, Cox DR (1998) On association models defined over independence graphs. Bernoulli 4:477–495
Wermuth N, Cox DR (2004) Joint response graphs and separation induced by triangular systems. J R Stat Soc B 66:687–717
Wermuth N, Lauritzen SL (1983) Graphical and recursive models for contingency tables. Biometrika 70:537–552
Wermuth N, Lauritzen SL (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). J R Stat Soc B 52:21–75
Wermuth N, Cox DR, Marchetti GM (2006a) Covariance chains. Bernoulli 12:841–862
Wermuth N, Wiedenbeck M, Cox DR (2006b) Partial inversion for linear systems and partial closure of independence graphs. BIT Numer Math 46:883–901
Wermuth N, Marchetti GM, Cox DR (2009) Triangular systems for symmetric binary variables. Electron J Stat 3:932–955
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester
Wiedenbeck M, Wermuth N (2010) Changing parameters by partial mappings. Stat Sin 20:823–836
Zellner A (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J Am Stat Assoc 57:348–368
Zhao H, Zheng Z, Liu B (2005) On the Markov equivalence of maximal ancestral graphs. Sci China Ser A 48:548–562

Test
DOI 10.1007/s11749-012-0284-4

DISCUSSION

Comments on: Sequences of regressions and their independences

Robert Castelo

I would like first to congratulate Nanny Wermuth and Kayvan Sadeghi for their comprehensive and insightful review of regression graphs, which also takes the reader through some of the milestones in the theory of graphical Markov models. Among the many aspects involved in the thorough description of the concepts and properties of regression graphs provided by the authors, I would like to draw the attention of this commentary to their relationship with respect to other types of graphical Markov models.

Markov properties, such as the global Markov property on purely undirected graphs, provide the connection between graphs and probability distributions, which allows one to derive intuitive interpretations from complex statistical models. The extent of these interpretations depends on the type of graph employed to define the graphical Markov model, and therefore, its topological properties tell us something about the class of statistical models represented by the graph. A canonical example are concentration graphs Markov equivalent to acyclic digraphs (DAGs), which correspond to chordal graphs (Wermuth 1980; Kiiveri et al. 1984). As pointed out in the paper, chordal graphs permit one to employ efficient fitting procedures to estimate structure and parameters from data, while chordless cycles on more than three vertices occurring in concentration graphs impose the requirement of iterative fitting algorithms for that purpose. Analogous findings on computational advantages conferred by chordal graphs have been also exploited in the field of databases (Beeri et al. 1983) in computer science.

Communicated by Domingo Morales.

This comment refers to the invited paper available at doi:10.1007/s11749-012-0290-6.

R. Castelo (✉)
Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain
e-mail: robert.castelo@upf.edu

Published online: 03 April 2012

This picture may suggest that different types of graph lead to a mosaic of distinct isolated classes of graphical Markov models. However, Wermuth and Sadeghi in this paper rightly show that by inserting edges in a regression graph it can be converted into another graphical Markov model of one of the intersecting classes, suggesting in fact that all the distinct classes of graphical Markov models are interlaced. This observation was already made by Andersson et al. (1995, p. 38) in the context of lattice conditional independence (LCI) models, which coincide with the class of transitive DAG models (DAGs without transition Vs), where they show that every LCI model includes and it is included in at least one DAG model. Those authors also concluded that every conditional independence restriction is equivalent to a simple LCI model and thus any graphical Markov model could be described by the intersection of all LCI models that contain it. This result was later expanded with the characterization of the class of DAGs with exclusively source Vs, also known as tree conditional independence (TCI) models (Castelo and Siebes 2003), whose Markov equivalence classes are represented by P4-free chordal graphs, as recalled also in this paper by Wermuth and Sadeghi. In this latter restricted subclass of chordal graphs the intersection of all cliques is always non-empty in each connected component (Castelo and Wormald 2003) and one such connected graph constitutes the simplest and most constraint graphical representation of one single conditional independence restriction.

The interlaced structure of graphical Markov models can provide a solution to the problem of selecting a regression graph from data within the vast search space of such models by searching first within a model class with fast data fitting procedures and then refine that model to select the final regression graph. While the authors claim that LWF chain graphs may constitute one such class of models because model fitting is possible in terms of single response regressions and Markov equivalence can be handled in the same way as in DAGs, the concept of vertex separator is much more complex. This contrasts with undirected graphs where this concept is much simpler and thus more amenable for working with marginal distributions of size $(q + 2) < n$ whenever the number of random variables p is larger than the number of observations n (Castelo and Roverato 2006).

References

Andersson S, Madigan D, Perlman M, Triggs C (1995) On the relation between conditional independence models determined by finite distributive lattices and by directed acyclic graphs. J Stat Plan Inference 48:25–46

Beeri C, Fagin R, Maier D, Yannakakis M (1983) On the desirability of acyclic database schemes. J ACM 30(3):479–513

Castelo R, Roverato A (2006) A robust procedure for Gaussian graphical model search from microarray data with p larger than n. J Mach Learn Res 7:2621–2650

Castelo R, Siebes A (2003) A characterization of moral transitive acyclic directed graph Markov models as labeled trees. J Stat Plan Inference 115:235–259

Castelo R, Wormald N (2003) Enumeration of P4-free chordal graphs. Graphs Comb 19:467–474

Kiiveri H, Speed T, Carlin J (1984) Recursive causal models. J Aust Math Soc A 36:30–52

Wermuth N (1980) Linear recursive equations, covariance selection, and path analysis. J Am Stat Assoc 75:963–972

Test
DOI 10.1007/s11749-012-0285-3

DISCUSSION

Comments on: Sequences of regressions and their independencies

Mathias Drton · Chris Fox · Andreas Käufl

1 Introduction

A graphical model is a statistical model that is associated with a graph whose nodes correspond to random variables. The model is defined by requiring distributions to obey a factorization property determined by the graph's edges or, alternatively, to exhibit a collection of conditional independencies associated with the pattern of edges absent from the graph. This latter point of view is the one stressed in the paper by Wermuth and Sadeghi who treat models associated with graphs that they term 'regression graphs'. We would like to take the opportunity to briefly comment on the motivation of regression graphs, their relationship with other mixed graphs, and on constraints that are not of conditional independence type. While there has been recent progress on models for categorical data (see, for instance, Evans and Richardson 2011, and references therein), our discussion will focus on multivariate normal distributions.

When variables are related through acyclic cause and effect relationships, the dependence structure they exhibit can be represented by directed acyclic graphs; see

Communicated by: Domingo Morales

This comment refers to the invited paper available at doi:10.1007/s11749-012-0290-6.

M. Drton (✉) · C. Fox
Department of Statistics, The University of Chicago, Chicago, IL, USA
e-mail: drton@uchicago.edu

C. Fox
e-mail: chrisfox@uchicago.edu

A. Käufl
Institute for Mathematics, University of Augsburg, Augsburg, Germany
e-mail: andreas.kaeufl@googlemail.com

Published online: 28 April 2012

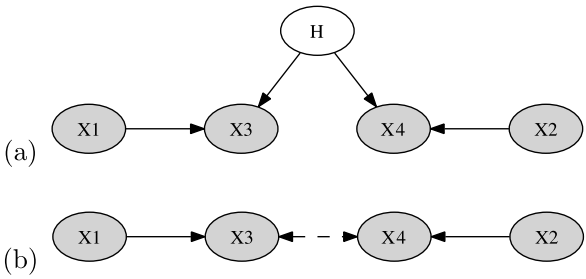


Fig. 1 (**a**) Seemingly unrelated regressions graph with hidden variable H. (**b**) Representation of the hidden variable model via a regression graph

e.g. Pearl (2009) and Spirtes et al. (2000). However, when selection effects or dependencies due to hidden variables are to be represented it is useful to consider graphs with more than one type of edge. Wermuth and Sadeghi allude to this point in their introduction, where they point the reader to work by Richardson and Spirtes (2002) and Wermuth (2011). Regression graphs form a special class of graphs that are of interest in this context. Their full lines can be thought of as arising through selection, and their dashed lines allow one to represent correlations due to hidden variables. In our opinion, this provides the strongest motivation for the use of regression graph models. As we will not treat selection effects in this commentary, our examples will involve graphs without full lines. Our pictures of graphs will have dashed lines drawn as bidirected edges with two arrowheads as is customary in the literature on structural equation models.

2 Regression graphs and hidden variables

For a concrete example of how regression graphs can be used to model the effects of hidden variables, consider the directed graph in Fig. 1(a), and assume that node H represents a hidden variable. In the statistical model associated with this graph, every marginal distribution for (X_1, X_2, X_3, X_4) exhibits the independencies

$$X_1 \perp\!\!\!\perp (X_2, X_4) \quad \text{and} \quad X_2 \perp\!\!\!\perp (X_1, X_3), \tag{1}$$

and every other conditional independence holding in all the marginal distributions for (X_1, X_2, X_3, X_4) is a consequence of standard conditional independence implications. The independencies in (1) are represented faithfully by the regression graph in Fig. 1(b); compare also the discussion of seemingly unrelated regressions in Wermuth and Sadeghi's Sect. 4. However, more is true. For instance, every four-variate normal distribution that satisfies the independencies in (1) does in fact arise as a marginal distribution of (X_1, X_2, X_3, X_4) under some (normal) joint distribution associated with the directed graph that includes the hidden variable H.

3 Non-independence constraints

While regression graphs are appropriate for conditional independence constraints, they cannot represent all conditional independence patterns that may arise from di-

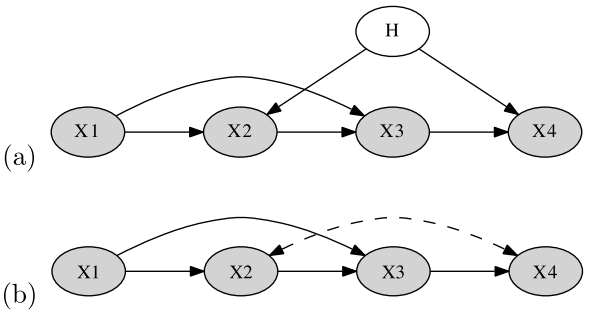


Fig. 2 (**a**) Verma graph with hidden variable H. (**b**) Mixed graph representation of the hidden variable model

rected graphs whose nodes include hidden variables. Moreover, conditional independence is not the only type of constraint of interest. As a concrete example, consider the directed graph in Fig. 2(a), sometimes referred to as the 'Verma graph'. Treating node H as a hidden variable, it is natural to ask whether the directed graphical model leads to constraints on the marginal distribution of (X_1, X_2, X_3, X_4). The answer is 'yes', but the constraints are not of conditional independence type. When restricting to normal distributions the marginal for (X_1, X_2, X_3, X_4) has a covariance matrix $\Sigma = (\sigma_{ij})$ that satisfies

$$\begin{aligned}\sigma_{12}\sigma_{13}\sigma_{14}\sigma_{23} - \sigma_{11}\sigma_{14}\sigma_{23}^2 - \sigma_{12}\sigma_{13}^2\sigma_{24} + \sigma_{11}\sigma_{13}\sigma_{23}\sigma_{24} \\ - \sigma_{12}^2\sigma_{14}\sigma_{33} + \sigma_{11}\sigma_{14}\sigma_{22}\sigma_{33} + \sigma_{12}^2\sigma_{13}\sigma_{34} - \sigma_{11}\sigma_{13}\sigma_{22}\sigma_{34} = 0, \qquad (2)\end{aligned}$$

and every other relation among the entries of the covariance matrix is a polynomial multiple of the given relation. The problem of finding relations in covariance matrices, or in tables of probabilities when discrete random variables are considered, can be solved using methods from computational algebra; compare Drton et al. (2009b) where the Verma graph is discussed in Sect. 3.3. It is also clear that the constraint given in (2) is not a conditional independence. The polynomial in (2) has degree four, but conditional independence constraints in a 4×4 covariance matrix are of degree no more than 3; compare Sect. 3.1 in Drton et al. (2009b).

While computer algebra instantly produces the polynomial in (2), it is not immediately clear how this polynomial arises. One possible explanation uses a result of Tian (2005). Write Σ for the parametrized 4×4 covariance matrix obtained from the Verma graph in Fig. 2(a); readers unfamiliar with the parametrization we have in mind may simply refer to the parametrization of the graph in Fig. 2(b) discussed below. Create another parametrized 4×4 matrix Σ' from the same parameters with the exception that the two coefficients associated with the edges $X_1 \to X_3$ and $X_2 \to X_3$ are set to zero and the variance parameter for X_3 is set such that X_3 has variance 1. This matrix Σ' is thus associated with the subgraph with $X_1 \to X_3$ and $X_2 \to X_3$ removed. Applying d-separation to this subgraph reveals that in the submodel X_1 and X_4 are marginally independent. We may deduce from the work of Tian (2005) that there is a rational map g, defined on the entire positive definite cone, that takes the covariance matrix Σ and returns the covariance matrix $\Sigma' = g(\Sigma)$. Hence, we know that an unconstrained symmetric matrix $\Sigma = (\sigma_{ij})$ can only be a covariance matrix associated with the Verma graph in Fig. 2(a) if $g(\Sigma)_{14} = 0$. Algebraically, to

compute $g(\Sigma)$, we calculate the Cholesky decomposition $\Sigma^{-1} = LL^T$ with L lower-triangular, replace the third row of L by $(0, 0, 1, 0)$ to obtain the lower-triangular matrix $\tilde{L}$, and form $g(\Sigma) = (\tilde{L}\tilde{L}^T)^{-1}$. We see that $g(\Sigma)_{14}$ is the ratio that has the Verma polynomial from (2) in the numerator and the product of σ_{11} and the determinant $\det(\Sigma_{123\times 123})$ in the denominator.

We remark that the calculation we just outlined is closely related to a nonparametric version in Sect. 7.3.1 in Richardson and Spirtes (2002). There it is shown that the observed margin in the model given by the Verma graph is still constrained even without parametric assumptions; see also Verma and Pearl (1991). Nonparametrically, we may think of the modification of the Cholesky factor just described as factoring the joint density of (X_1, X_2, X_3, X_4) into conditionals as

$$f(x_1, x_2, x_3, x_4) = f_4(x_4 \,|\, x_1, x_2, x_3) f_3(x_3 \,|\, x_1, x_2) f_2(x_2 \,|\, x_1) f_1(x_1)$$

and replacing the conditional density $f_3(x_3 \,|\, x_1, x_2)$ by the density of a standard normal distribution with argument x_3. Then, Σ is the covariance matrix associated with the density f, and $g(\Sigma) = \Sigma'$ is the covariance matrix after replacing f_3.

4 Mixed graphs and structural equation models

As the Verma example makes clear, it can be interesting to go beyond graphs that encode solely conditional independencies and instead adopt a more general framework. Mixed graphs that may feature both arrows (directed edges) and dashed lines without any further constraints on their structure can provide such a framework; see Wermuth (2011) for a general treatment of how mixed graphs can be associated with directed graphs that have hidden variables among their nodes. The Verma graph leads to the mixed graph in Fig. 2(b), which is not a regression graph.

The classical approach of structural equations allows one to give statistical meaning to any mixed graph. Sticking with the Verma example and considering the linear case with zero means, the graph in Fig. 2(b) is translated into the equation system

$$\begin{aligned} X_1 &= \epsilon_1, & X_3 &= \lambda_{13} X_1 + \lambda_{23} X_2 + \epsilon_3, \\ X_2 &= \lambda_{12} X_1 + \epsilon_2, & X_4 &= \lambda_{34} X_3 + \epsilon_4, \end{aligned}$$

where $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ are jointly normal with ϵ_1, ϵ_3 and (ϵ_2, ϵ_4) mutually independent but possible correlation between ϵ_2 and ϵ_4. Writing $\Omega = (\omega_{ij})$ for the covariance matrix of $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ it is clear that the equations determine the covariance matrix for (X_1, X_2, X_3, X_4) to be of the form

$$(I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}, \tag{3}$$

where I denotes the identity matrix,

$$\Lambda = \begin{pmatrix} 0 & \lambda_{12} & \lambda_{13} & 0 \\ 0 & 0 & \lambda_{23} & 0 \\ 0 & 0 & 0 & \lambda_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \Omega = \begin{pmatrix} \omega_{11} & 0 & 0 & 0 \\ 0 & \omega_{22} & 0 & \omega_{24} \\ 0 & 0 & \omega_{33} & 0 \\ 0 & \omega_{24} & 0 & \omega_{44} \end{pmatrix}.$$

The display of the two matrices makes explicit that the support of Λ is given by the arrows in the mixed graph whereas the dashed lines determine the support of Ω. Despite the fact that there are no conditional independencies involved, the set of covariance matrices associated with the mixed graph in Fig. 2(b) is equal to the set of marginal covariance matrices arising in the hidden variable model given by Fig. 2(a). Hence, the two graphs encode the same set of four-variate normal distributions. We would like to remark that the Verma model and related linear models can be fitted by maximum likelihood using the algorithm presented in Drton et al. (2009a), which is based on iterative least squares computations that converge reliably. By the main theorems of Drton et al. (2011), the Verma graph gives a globally (or everywhere) identifiable model in which maximum likelihood estimators are asymptotically normal given large samples.

5 Model equivalence

At this point, the reader may wonder which mixed graphs determine linear structural equation models that are 'cut out' by conditional independencies. In other words, which mixed graphs yield a set of covariance matrices that is equal to all positive definite matrices obeying some set of conditional independence constraints? One class of mixed graphs for which this is true are the regression graphs considered by Wermuth and Sadeghi. A regression graph is a mixed graph without semi-directed cycles, that is, the graph does not contain cycles with at least one arrow and all arrows pointing in the same direction. Graphs without semi-directed cycles have also been called chain graphs in the literature. To our knowledge, the most general class of graphs known to define linear models cut out by conditional independencies are the maximal ancestral graphs of Richardson and Spirtes (2002). A mixed graph is ancestral if all its semi-directed cycles involve at least two dashed lines. Clearly, the Verma graph is not ancestral because of the cycle $X_2 \to X_3 \to X_4 \leftrightarrow X_2$.

Two mixed graphs can be equivalent in the sense of having the same set of associated covariance matrices. For answering questions about such model equivalence, it is useful to have implicit representations of models in terms of conditional independence, or possibly other types of constraint. When only conditional independencies are of concern, model equivalence is typically referred to as Markov equivalence. To our knowledge, the most general results about model/Markov equivalence were given by Ali et al. (2009) who develop a polynomial-time criterion for ancestral graphs. However, as clarified by Wermuth and Sadeghi, easier and faster to check conditions can be given for regression graphs. A recent result that holds promise to help resolve further model equivalence questions is the trek-separation criterion due to Sullivant et al. (2010), which allows one to characterize the set of all determinants that vanish for covariance matrices in linear mixed graph models. This new criterion extends an earlier result on 2×2 determinants that is known as the tetrad representation theorem; see Spirtes et al. (2000).

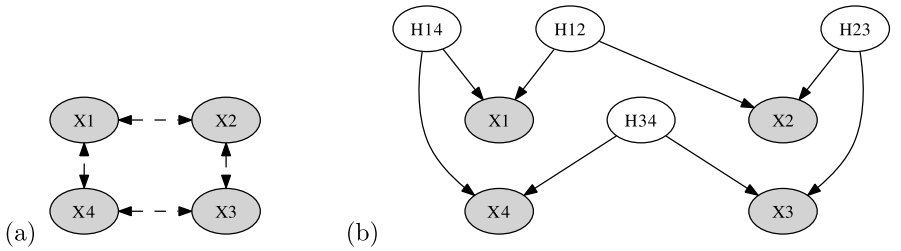


Fig. 3 (**a**) Four-cycle as covariance graph. (**b**) The 'canonical' hidden variable model inducing the four–cycle

6 Inequality constraints

We would like to conclude this commentary by pointing out that mixed graph models need not always be precisely equal to hidden variable models. The four-cycle in Fig. 3(a) has the independence interpretation

$$X_1 \perp\!\!\!\perp X_3 \quad \text{and} \quad X_2 \perp\!\!\!\perp X_4.$$

Hence, the multivariate normal distributions associated with the graph simply correspond to all positive definite matrices $\Sigma = (\sigma_{ij})$ of the form

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & 0 \\ 0 & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & 0 & \sigma_{34} & \sigma_{44} \end{pmatrix}. \tag{4}$$

The 'canonical' hidden variable model leading to the four-cycle is depicted in Fig. 3(b). Let us consider joint multivariate normal distributions for the eight nodes in this directed graph. The results in Drton and Yu (2010) then imply that a positive definite matrix with zeros as in (4) is the covariance matrix of some marginal distribution of (X_1, X_2, X_3, X_4) in the directed hidden variable model if and only if we have

$$\begin{aligned} &\sigma_{11}\sigma_{22}\sigma_{33}\sigma_{44} - \sigma_{11}\sigma_{22}\sigma_{34}^2 - \sigma_{11}\sigma_{23}^2\sigma_{44} - \sigma_{12}^2\sigma_{33}\sigma_{44} \\ &\quad + \sigma_{12}^2\sigma_{34}^2 + 2\sigma_{12}\sigma_{23}\sigma_{34}\sigma_{14} - \sigma_{22}\sigma_{33}\sigma_{14}^2 + \sigma_{23}^2\sigma_{14}^2 \geq 0. \end{aligned}$$

The left hand side of this inequality is the determinant of the matrix obtained by negating σ_{12}, or any other non-zero off-diagonal entry of the matrix Σ in (4). We conjecture that there does not exist a normal directed graphical model with hidden variables that gives as observed covariance matrices precisely the positive definite matrices with $\sigma_{13} = \sigma_{24} = 0$. We would be interested in hearing the thoughts of Wermuth and Sadeghi on this point as well as the potential statistical use of non-independence constraints in graphical modelling.

References

Ali RA, Richardson TS, Spirtes P (2009) Markov equivalence for ancestral graphs. Ann Stat 37(5B):2808–2837

Drton M, Yu J (2010) On a parametrization of positive semidefinite matrices with zeros. SIAM J Matrix Anal Appl 31(5):2665–2680

Drton M, Eichler M, Richardson TS (2009a) Computing maximum likelihood estimates in recursive linear models with correlated errors. J Mach Learn Res 10:2329–2348

Drton M, Sturmfels B, Sullivant S (2009b) Lectures on algebraic statistics. Oberwolfach seminars, vol 39. Birkhäuser, Basel

Drton M, Foygel R, Sullivant S (2011) Global identifiability of linear structural equation models. Ann Stat 39(2):865–886

Evans RJ, Richardson TS (2011) Marginal log-linear parameters for graphical Markov models. arXiv:1105.6075

Pearl J (2009) Causality, 2nd edn. Cambridge University Press, Cambridge

Richardson T, Spirtes P (2002) Ancestral graph Markov models. Ann Stat 30(4):962–1030

Spirtes P, Glymour C, Scheines R (2000) Causation, prediction, and search, 2nd edn. Adaptive computation and machine learning. MIT Press, Cambridge

Sullivant S, Talaska K, Draisma J (2010) Trek separation for Gaussian graphical models. Ann Stat 38(3):1665–1685

Tian J (2005) Identifying direct causal effects in linear models. In: Veloso MM, Kambhampati S (eds) Proceedings of the twentieth national conference on artificial intelligence (AAAI). AAAI Press/MIT Press, Menlo Park/Cambridge, pp 346–353

Verma T, Pearl J (1991). Equivalence and synthesis of causal models. Technical report R-150, UCLA

Wermuth N (2011) Probability distributions with summary graph structure. Bernoulli 17(3):845–879

Test
DOI 10.1007/s11749-012-0286-2

DISCUSSION

Comments on: Sequences of regressions and their independences

Monia Lupparelli · Alberto Roverato

Applied researchers can now rely on several families of graphical models for data analysis. The availability of a wide range of tools is clearly an advantage but, on the other hand, the initial assumption that the data generating process belongs to a specific class of graphical models is a crucial step with non-negligible consequences for the results of the analysis.

Background knowledge on the problem at hand sometimes suggests that the variables involved in the analysis, Y_V with $V = \{1, \ldots, p\}$, can be partitioned into groups $Y_{B_1}, \ldots, Y_{B_K}$ called *blocks* such that: (i) variables within each block are considered *on equal standing*; (ii) blocks can be partially ordered on the basis of time or subject-matter considerations in such a way that the first block Y_{B_1} contains primary responses, all background or context variables are collected in the last block Y_{B_K} and blocks $Y_{B_2}, \ldots, Y_{B_{K-1}}$ contain intermediate responses. In this case, it seems natural to restrict attention to chain graph (CG) models where blocks of variables are joined by arrows pointing from blocks with higher position to blocks with lower position. The statement that variables within blocks are on equal standing is traditionally implemented by imposing that blocks induce undirected graphs corresponding to either concentration or covariance graph models.

The paper by Wermuth and Sadeghi is devoted to the class of CG models known as regression graph (RG) models, but sometimes also referred to as CG models of

Communicated by Domingo Morales.

This comment refers to the invited paper available at doi:10.1007/s11749-012-0290-6.

M. Lupparelli · A. Roverato (✉)
Department of Statistical Sciences, University of Bologna, Bologna, Italy
e-mail: alberto.roverato@unibo.it

M. Lupparelli
e-mail: monia.lupparelli@unibo.it

Published online: 28 April 2012

type IV (see Drton 2009; Sadeghi and Lauritzen 2011). It is an illuminating paper that addresses a number of different issues. It contains a comprehensive review of the historical genesis of these models clarifying the connections with methodological tools developed in other and affine research fields. It also includes a worthwhile discussion concerning the interpretation of both dependencies and independencies represented by these models, thereby making it clear which are the most appropriate contexts of application. To this aim, the paper describes some applications illustrating the use of these models and suggesting effective model selection procedures. Finally, novel methodological contributions concerning Markov equivalence are given. We wish to congratulate the authors for their excellent work.

An interesting feature of RGs is that they belong to the family of loopless mixed graphs (Sadeghi and Lauritzen 2011) and are also a special case of both summary graphs (Wermuth 2011) and ancestral graphs (Richardson and Spirtes 2002), as made clear by Sadeghi and Lauritzen (2011) who provided a hierarchical classification of several classes of graphical models. We notice that the definition of RG models given by Wermuth and Sadeghi differs from the definition of CG models of type IV given in Drton (2009) and subsequently used in the classification of Sadeghi and Lauritzen (2011). The former assumes a concentration graph model for the variables in the last block Y_{B_K} and covariance graph models in the remaining blocks, whereas in the latter there is no distinction between blocks which all induce covariance graph models. This is a minor difference that only involves the marginal distribution of the context variables, but we take it as a starting point for a critical discussion of the role played by blocks.

Every block of either primary or intermediate responses is preceded by at least one block of explanatory variables and the assumption of a covariance graph for these blocks is motivated by the fact that these graphs encode the correlation structure of residuals in joint regressions. From this perspective, a concentration graph model seems a natural choice for the variables in the last block, which are purely explanatory.

The partition of variables into blocks is the first step of the analysis, and the model selection procedure will return a RG model where variables are partitioned into *chain components*. Every block contains one or more chain components which are the connected components of the block. The chain components have a compatible ordering that is used to obtain a recursive factorization of the joint density of Y_V. Consider the case where, in the selected model, some response variables form a chain component with no incoming arrows so that the corresponding term in the generating process is a marginal density rather than a conditional density. Since these variables are labelled as responses, their marginal distribution is assumed to belong to a covariance graph model, but one may wonder if a concentration graph model would be more appropriate in this case. More generally, we consider the partition of variables into blocks to be used to study a subclass of RG models that is sometimes unduly restrictive.

Let $\mathcal{M}_V$ denote the family of RG models for Y_V. The assumed block structure $Y_{B_1}, \ldots, Y_{B_K}$ implies that only a subset of the models, which we denote by $\mathcal{M}_V^B$, is considered. The process that identifies $\mathcal{M}_V^B$ as a subset of $\mathcal{M}_V$ can be split into two steps. The first step implements the background information on the block ordering by restricting the attention to the subclass of RG models for Y_V such that edges between

blocks are arrows whose direction is compatible with the block ordering, and we denote by $\mathcal{M}_V^{\leftarrow} \subseteq \mathcal{M}_V$ such subclass of models. RGs in $\mathcal{M}_V^{\leftarrow}$ allow variables within blocks to be coupled by either arrows or undirected edges. The fact that variables within blocks are on equal standing motivates the second step, which identifies $\mathcal{M}_V^B$ as the subclass of $\mathcal{M}_V^{\leftarrow}$ made up of all models inducing a concentration graph for Y_{B_K} and covariance graphs for the remaining blocks.

The researcher may be uncertain about the appropriate block structure, but the class $\mathcal{M}_V^B$ is very sensitive to the block specification. Uncertainty may concern, for instance, whether a subset of variables should form two adjacent blocks B_j and B_{j+1} or a single block $B_j \cup B_{j+1}$. This decision may heavily affect the results of the analysis because constraining variables to belong to a common block implies, for instance, that an asymmetric relationship between them is precluded. This drawback is, to some extent, mitigated by the results on Markov equivalence. Indeed, an RG that is Markov equivalent to the selected graph in $\mathcal{M}_V^B$ may include arrows inside the blocks, as highlighted in the example given in Figs. 1 as 2 of the paper by Wermuth and Sadeghi.

The main point is that an unambiguous meaning should be associated with the statement "variables are on equal standing". When it is used to mean that nothing is known about the independence structure of those variables, then assuming either a concentration or a covariance graph model may not be appropriate because these models specify a very special kind of equal standing between variables. In this case, it might be more appropriate to perform a model search within the class $\mathcal{M}_V^{\leftarrow}$ that confers more flexibility to the analysis. Note that, following this approach, merging blocks leads one to consider wider classes of models and they can therefore be motivated by uncertainty about the exact block structure.

Hence, considering the wider class $\mathcal{M}_V^{\leftarrow}$ makes the analysis more robust with respect to misspecification of the block structure, at the cost of an increased complexity in the exploration of the search space.

References

Drton M (2009) Discrete chain graph models. Bernoulli 15(3):736–753

Richardson TS, Spirtes P (2002) Ancestral graph Markov models. Ann Stat 30(4):962–1030

Sadeghi K, Lauritzen S (2011) Markov properties for loopless mixed graphs. Technical report. arXiv:1109.5901v1

Wermuth N (2011) Probability distributions with summary graph structure. Bernoulli 17(3):845–879

Test
DOI 10.1007/s11749-012-0288-0

DISCUSSION

Comment on: Sequences of regressions and their independences

Bala Rajaratnam

1 Introduction

The class of regression graph models studied by Wermuth and Sadeghi (2012) are adept at modelling data from intervention studies, and thus they are a useful addition to the different classes of graphical models that have been introduced in the literature. A part of their appeal comes from their simplicity and the natural way in which different types of graphical models (covariance, concentration, and directed acyclic graph or DAG) are combined to yield a graphical model with connected components that "represent conditional independent responses given their common past" (Wermuth and Sadeghi 2012). As Wermuth and Sadeghi (2012) already note, regression graph models can be regarded as one of three types of chain graphs that have been studied as joint response models. The other two are the AMP chain graph models of Andersson and Wojnar (2004) and LWF chain graphs of Lauritzen and Wermuth (1989) and Frydenberg (1990). Wermuth and Sadeghi (2012) also mention that AMP and LWF chain graph models are useful for intervention studies only if they correspond to regression graph models in the sense of Markov equivalence. This is an important point as the conditioning set of variables for AMP and LWF chain graph models can contain nodes from the same connected component.

The probabilistic properties of regression graph models have also been established by Wermuth and Sadeghi (2012). These include generating processes over a regression graph, Markov properties, Markov equivalence, and a discussion of faithfulness.

Communicated by Domingo Morales.

This comment refers to the invited paper available at doi:10.1007/s11749-012-0290-6.

B. Rajaratnam (✉)
Stanford University, Stanford, CA, USA
e-mail: brajarat@stanford.edu

Published online: 28 April 2012

These probabilistic properties, though illuminating in their own right, are also useful for deriving (both maximum likelihood and Bayesian) statistical inference procedures. In particular, conditions for Markov equivalence between regression graph models and DAG models can be exploited in order to specify inferential procedures for classes of regression models.

To this end, Wermuth and Sadeghi (2012) also note that covering models, those that manage to retain most of the independences in a complex graph but not all, can be highly useful in this regard. In particular, they comment that a systematic approach to search for covering models is not available in the literature.

The authors of Wermuth and Sadeghi (2012) are to be commended for a thought provoking paper that highlights the use of regression graph models. In this short discussion note of Wermuth and Sadeghi (2012), we highlight some of the aspects of estimation and model selection that can be potentially useful while working with regression graph models. In particular, we explore properties pertaining to estimation and model selection when the variables are jointly Gaussian distributed. We shall consider the distribution of the maximum likelihood estimator and proceed to specify Bayes procedures for inference in regression graph models. These endeavours naturally lead to generalization and amalgamation of the various classes of Wishart distributions that have been recently introduced in the mathematical statistics literature. Extensions to continuous distributions other than Gaussian and discrete variables which permit nonlinear and interactive dependences are also briefly discussed.

This short discussion note is organized as follows. Section 2 introduces some preliminaries on various classes of Wishart distributions that have been introduced in the literature within the context of maximum likelihood and Bayesian inference for graphical models. Section 3 gives methodology and results that illustrate the task of estimation and model selection in regression graph models. For the sake of brevity, details are provided elsewhere. Section 4 concludes by summarizing and outlining future research directions.

2 Wisharts for DAG, concentration, and covariance graphical models

We now briefly study four different graphical Wishart distributions that have appeared in the mathematical statistics literature and special cases of these. The reader is referred to Letac and Massam (2007) and Ben-David and Rajaratnam (2011) for the relevant notation (including definitions of the spaces $P_{\mathcal{G}}$, $Q_{\mathcal{G}}$, $R_{\mathcal{G}}$, $S_{\mathcal{G}}$, etc.).

Concentration Wishart distributions The W_{P_G} and W_{Q_G} "concentration Wisharts" of Letac and Massam (2007) can be used for Bayesian Analysis for graphical Gaussian models, as in Rajaratnam et al. (2008), or to describe the distribution of the maximum likelihood estimator in a decomposable graphical Gaussian model. The W_{P_G} density and the corresponding inverse Wishart are specified below. The density of W_{Q_G} is also given below.

$$W_{P_G}(\alpha, \beta, \theta; dy) = e^{-\langle \theta, y \rangle} \frac{H_G(\alpha, \beta; \phi(y)}{\Gamma_{II}(\alpha, \beta) H_G(\alpha, \beta; \theta)} \nu_G(dy), \tag{1}$$

$$\Gamma_{II}(\alpha,\beta) = \pi^{((c_1-s_2)s_2+\sum_{j=2}^{k}(c_j-s_j)s_j)/2} \times \Gamma_{s_2}\left[-\alpha_1 - \frac{c_1-c_2}{2} - \gamma_2\right]\Gamma_{c_1-s_2}(-\alpha_1)\prod_{j=2}^{k}\Gamma_{c_j-s_j}(-\alpha_j), \tag{2}$$

$$IW_{P_G}(\alpha,\beta,\theta;dx) = \frac{e^{-\langle\theta,\hat{x}^{-1}\rangle}H_G(\alpha,\beta;x)}{\Gamma_{II}(\alpha,\beta)H_G(\alpha,\beta;\theta)}\mu_G(dx), \tag{3}$$

$$W_{Q_G}(\alpha,\beta,\sigma;dx) = e^{-\langle\theta,\hat{x}^{-1}\rangle}\frac{\Gamma_I(\alpha,\beta)H_G(\alpha,\beta;x)}{H_G\alpha,\beta;\sigma}. \tag{4}$$

Letac and Massam (2007) demonstrate that the highly useful hyper-inverse Wishart and hyper Wishart of Dawid and Lauritzen (1993) belong to the class of enriched LM concentration graph Wishart distributions. These Wishart distributions appear naturally in both Bayesian and frequentist inferences for graphical Gaussian models. In particular, Rajaratnam et al. (2008) use the LM priors for flexible covariance estimation in high-dimensional settings. Roverato (2002) also studies the problem of inference in the nondecomposable concentration graph setting.

Covariance Wishart distributions A recent addition to the class of Wishart distributions is the "covariance Wishart" distribution of Khare and Rajaratnam (2011) specified below. These have been introduced in the context of Bayesian inference for Gaussian covariance graph models.

$$\tilde{\pi}_{U,\alpha}^{P_G}(\Sigma) \propto e^{-(\mathrm{tr}(\hat{x}U)+\sum_{i=1}^{m}(2n_i+\alpha_i)\log D_{ii}((\hat{x})^{-1}))/2}, \quad \frac{\prod_{S\in\mathcal{S}}|x_S|^{(|S|+1)\nu(S)}}{\prod_{C\in\mathcal{C}}|x_C|^{|C|+1}}, \quad x \in Q_G. \tag{5}$$

We note that Gibbs sampling can be used in order to sample from the KR covariance Wishart distributions (see Khare and Rajaratnam 2011). The reader is also referred to the work by Silva and Ghahramani (2009) and Andersson and Wojnar (2004) for related work.

DAG Wishart distributions Yet an even more recent addition to extensions of the Wishart distribution on the cone is the DAG Wishart distributions of Ben-David and Rajaratnam (2011).

$$\pi_{U,\alpha}^{R_\mathcal{G}}(\Upsilon) = z_\mathcal{G}(U,\alpha)^{-1}\exp\left\{-\frac{1}{2}\mathrm{tr}(\widehat{\Upsilon}U)\right\}\prod_{i=1}^{m}D_{ii}^{-\frac{1}{2}\alpha_i+pa_i+2}.$$

Standard conjugate priors in the Cholesky parameterization of the DAG model have been known in the literature for some time (see, for example, Geiger and Heckerman 1994 and other texts/papers on the topic). In the above work, Ben-David and Rajaratnam (2011) derive (for any arbitrary DAG) the DAG Wishart distributions on covariance and concentration spaces that correspond to enriched standard conjugate priors in the Cholesky parameterization of the DAG model.

3 Generalized Wishart distributions for regression graphs

In this section we briefly examine how the Wishart distributions specified above for simple graphical models (those which are purely concentration graph, covariance graph, or DAG models) can be combined for Bayesian inference for regression graph models. We examine three possible approaches.

First, following the spirit of Wermuth and Sadeghi (2012), we first state a result that exploits the Markov equivalence result in Theorem 2.

Proposition 1 *Consider a regression graph model with a chordal graph for the context variables which does not contain any chordless collision path in four nodes. Then the class of enriched DAG Wishart distributions specified in Ben-David and Rajaratnam* (2011) *yields a class of hyper Markov conjugate priors for this regression graph model.*

Moreover, note that the priors are also strong directed hyper Markov. The proof of the above proposition follows simply by noting that in Theorem 2 of Wermuth and Sadeghi (2012) it is proved that the regression graph can be "oriented to be Markov equivalent to a directed acyclic graph in the same skeleton." This result can be exploited to construct a DAG, which is then sufficient to make the use of the enriched conjugate priors studied in Ben-David and Rajaratnam (2011). An advantage of this line of thinking is that the approach followed in Ben-David and Rajaratnam (2011) allows one to evaluate closed-form posterior covariance and concentration quantities which can be beneficial in higher-dimensional settings. We also note that if the quantities of interest are the Cholesky parameters, then the DAG Wisharts are not directly required.

An equally relevant question pertains to specifying procedures when the regression graph is not Markov equivalent to a model with only one type of edges. Wermuth and Sadeghi (2012) suggest a useful approach to dealing with this in the context of maximum likelihood estimation. In particular, they suggest using a covering model with additional edges that is easier to fit than the original (regression graph) model. The relationships to LWF model are specifically mentioned in this regard. The authors justifiably also caution that this approach may mean that important structure of the original generating regression graph is lost in the process (see p. 38 of Wermuth and Sadeghi 2012). The "covering" approach can also be useful when undertaking Bayesian inference for regression graph models. A covering model that corresponds to a model for which any of the above graphical Wishart distributions apply is ideal since Bayes procedures and their theoretical properties have already been studied at length. Details on what would be the "best" covering model and links to Markov equivalences and Bayesian inference require further analysis and will be discussed elsewhere.

Identifying a simple Markov equivalent model, or using a covering model, as discussed above, may not always be applicable in general due to the limitations outlined above. A third more general approach is now discussed below.

Consider the factorization given by Eq. (1) in Wermuth and Sadeghi (2012),

$$f_N = \prod_{j=1}^{J} f_{g_j|g_{>j}}, \tag{6}$$

and recall Eq. (2) from Wermuth and Sadeghi (2012), which gives the following meaning to each ik edge ($i \neq k$) present in the regression graph G^N_{reg}:

$$\begin{aligned}
&\text{(i)} \quad i \pitchfork k|g_{>j} \quad \text{for } i,k \text{ both in a response component } g_j \text{ of } u,\\
&\text{(ii)} \quad i \pitchfork k|g_{>j}\backslash\{k\} \quad \text{for } i \text{ in } g_j \text{ of } u \text{ and } k \text{ in } g_{>j},\\
&\text{(iii)} \quad i \pitchfork k|v\backslash\{i,k\} \quad \text{for } i,k \text{ both in a context component } g_j \text{ of } v.
\end{aligned} \tag{7}$$

Also note that proposition (7) of Wermuth and Sadeghi (2012) states that for disjoint states a, b, c, a given regression graph implies $a \perp\!\!\!\perp b \mid c$ if and only if every path between the sets a and b breaks given c.

A natural "system of priors" that combines the graphical Wishart priors that have already been introduced/studied in the literature can be constructed by recursively applying the standard conjugate prior for each block conditional on previous blocks. Recall that the different graphical Wishart priors (i.e., the covariance, concentration, and DAG Wishart priors) can serve as natural candidates in this process of building the overall prior. This process can be symbolically represented (with some abuse of notation) as follows:

$$\pi(\theta) \propto \prod_{j=1}^{J} f_{g_j|g_{>j}}(x_j|\theta_j)\pi_{g_j|g_{>j}}(\theta_j), \tag{8}$$

where $\pi_{g_j|g_{>j}}(\theta_j)$ denotes the prior for the jth block given the parameters for previous blocks, and where θ represents the full parameter of the regression graph model. We shall refer to the overall prior $\pi(\theta)$ as the generalized regression graph Wishart distribution.

Broadly speaking, the above approach allows us to sequentially specify conditional priors which are already available from the literature. Moreover posterior samples for a given block can be obtained directly by sampling conditionally on previous blocks. In this case, MCMC is used only within blocks and not between blocks (see, for instance, Khare and Rajaratnam 2011 for the covariance graph case). This topic is more involved, and the precise statement of the above assertion is omitted from this discussion note.

Note that in the Gaussian setup conjugacy is retained overall due to the properties of the graphical Wishart priors outlined in Sect. 2. The above approach specifies conditional graphical Wishart priors within each block given previous blocks. Bayesian modelling of parameters between blocks can be specified in a manner similar to that of a typical DAG model. The above scheme also allows us to define a new concept, the "block hyper Markov" property, to describe the deep structure in the prior. We shall not discuss this here and mention that details will be given elsewhere. Moreover, we remark that the same approach used above to specify priors above can also be used to build a system of regression graph model priors outside the Gaussian setting.

4 Closing remarks

We conclude this short note by once again commending the authors for an interesting and useful paper. We have broadly outlined procedures for Bayesian inference for regression graph models including prior specification and sampling from the posterior. The Bayesian approach has the distinct advantage of also directly giving variability estimates of Bayes procedures. We note that under certain conditions the graphical Wishart distributions studied above are useful in specifying the distributions of graphical maximum likelihood estimates. We also remark that the above discussion does not specifically address Bayesian model selection or choice of hyper parameters. These details will be elaborated on elsewhere.

Acknowledgements The author acknowledges funding from the Humboldt-Foundation which in part allowed him to attend a workshop in Gullmarsstrand, Sweden. The author thanks the organizers/funders for the kind invitation to attend this workshop. The author also acknowledges Sang Oh for LaTex assistance. The work was also supported in part by National Science Foundation under Grant Nos. DMS-CMG 1025465, AGS-1003823, DMS-1106642 and grants NSA H98230-11-1-0194, DARPA-YFA N66001-11-1-4131, and SUWIEVP10-SUFSC10-SMSCVISG0906.

References

Andersson SA, Wojnar GG (2004) Wishart distributions on homogeneous cones. J Theor Probab 17(4):781–818

Ben-David E, Rajaratnam B (2011) Generalized hyper Markov laws for directed acyclic graphs. Technical report, Department of Statistics, Stanford University, 45, September

Dawid AP, Lauritzen SL (1993) Hyper-Markov laws in the statistical analysis of decomposable graphical models. Ann Stat 21(3):1272–1317

Frydenberg M (1990) The chain graph Markov property. Scand J Stat 17(4):333–353

Geiger D, Heckerman D (1994) Learning Gaussian networks. Microsoft research. Technical report: Msr-tr-94-10, July

Khare K, Rajaratnam B (2011) Wishart distributions for decomposable covariance graph models. Ann Stat 39(1):514–555

Lauritzen SL, Wermuth N (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat 17(1):31–57

Letac G, Massam H (2007) Wishart distributions for decomposable graphs. Ann Stat 35(3):1278–1323

Rajaratnam B, Massam H, Carvalho CM (2008) Flexible covariance estimation in graphical Gaussian models. Ann Stat 36(6):2818–2849

Roverato A (2002) Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. Scand J Stat 29(3):391–411

Silva R, Ghahramani Z (2009) The hidden life of latent variables: Bayesian learning with mixed graph models. J Mach Learn Res 10:1187–1238

Wermuth N, Sadeghi K (2012) Sequences of regressions and their independences. TEST (to appear)

Test
DOI 10.1007/s11749-012-0287-1

DISCUSSION

Comments on: Sequence of regressions and their independences

Elena Stanghellini

First of all I wish to congratulate the authors for this interesting paper, which addresses different aspects of graphical models and provides a vast list of references that connect this research area with contributions in the field of structural equation and econometric models, analysis of binary data and contingency tables, etc. I believe that the novelty of the paper can be best summarized by noting that a symbol to represent conditional associations is introduced. It means that the focus has now been shifted from conditional independences to conditional associations. To do so, the notion of faithfulness has been replaced by a group of properties that a distribution has to satisfy for the corresponding graph to be faithful. This capacity to break down the faithfulness assumption into the intersection of weaker assumptions allows one to modulate it. As a consequence, the authors show that many properties can be derived under a milder assumption, which they call traceable regressions. These derivations lead to a simple and elegant formulation of Markov equivalence between regression graphs, which I also believe is a major interesting advance.

My reflexion originates from the notion of *covering model with nice estimation properties*. This happens to be particularly useful in studies for which the outcome of interest is measured only on a selected population. Let X be an explanatory variable of the outcome Y and Z be a variable that drives the selection mechanism. Two situations can occur, one in which Y is defined for all units in the population but observed only for Z taking on some particular values, and the other in which Y is

Communicated by Domingo Morales.

This comment refers to the invited paper available at doi:10.1007/s11749-012-0290-6.

E. Stanghellini (✉)
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli 1, 06100 Perugia, Italy
e-mail: elena.stanghellini@stat.unipg.it

Published online: 31 March 2012

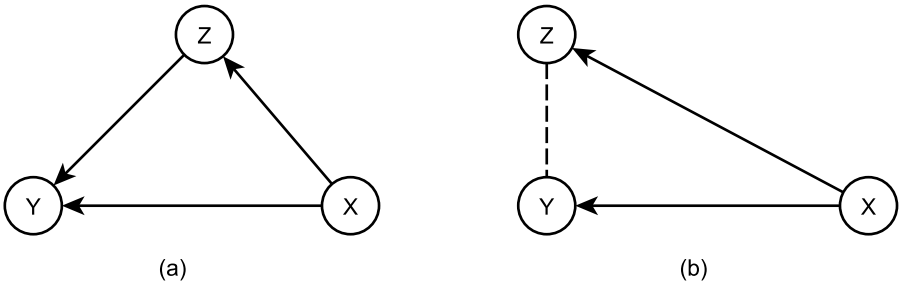


Fig. 1 Two hypothetical data generating processes with the selection variable Z, (**a**) an intermediate and (**b**) an outcome variable

defined only for those units with Z taking on particular values. I shall discuss the first one and build an example to summarize what I mean, and to clarify my comment. For a group of Italian freshman students we want to know how predictive is X, the mark they get at the end of their high school studies, of Y, the mark they get when the finish their undergraduate three year degree at a given university. Suppose now that in order to access this university they have to pass a test. Let Z be the score a student gets at the test. Only students with a score higher than a fixed threshold c, equal for all students, are allowed to enroll at the university. To keep the argument simple, assume no further selection is taking place, so that all students with $Z > c$ enroll at the university and get their final mark Y. Suppose further that the relationship between all variables can be captured by linear regressions.

In Fig. 1(a) and (b) are the two hypothesized regression graphs to describe the underlying data generating process. We can think of it as conditional to some context variables. In Fig. 1(a) the variable Z is a directly explanatory variable of Y, while in Fig. 1(b) they are correlated. In this second case, we can think of Z and Y as response variables of some hidden node U representing the level of motivation of the student. As discussed in Sect. 4 of the paper, both models are saturated but they are not Markov equivalent. In particular, we would like to stress the different meaning of the arrows from X to Y in Fig. 1(a) and (b), the former representing the coefficient of X in the linear regression of Y against X and Z (which we denote by $\beta_{Y|X.Z}$), the latter representing the coefficient of X in the linear regression of Y against X only (which we denote by $\beta_{Y|X}$).

The interest is in the linear regression coefficient $\beta_{Y|X}$ in the overall population. The linear regression of Y on X on the selected population is clearly distorted. We can, however, distinguish between a situation, which we denote Case A, where Z is observed from the situation, which we denote Case B, where only the indicator $I = I(Z > c)$ is observed. In Case A, when the postulated model is as in Fig. 1(a), the marginal regression coefficient of interest can be determined by making use of Cochran's formula:

$$\beta_{Y|X} = \beta_{Y|X.Z} + \beta_{Z|X}\beta_{Y|Z.X}.$$

Since Z and X are observed for all units, $\beta_{Z|X}$ can be consistently estimated from a random sample drawn from the model. Furthermore, since selection acts as truncation on an explanatory node for Y, $\beta_{Y|Z.X}$ and $\beta_{Y|X.Z}$ can also be consistently estimated

from a random sample drawn from the truncated distribution (perhaps with a loss of efficiency). Similarly, when the postulated model is as in Fig. 1(b), the regression coefficient of interest can also be estimated by making use of the above relationship, after noting, however, that the estimations are not made under the model, but under an auxiliary model with nice estimation properties. Notice that, despite the symmetry of Fig. 1(b), alternative estimations based on auxiliary models that make use of a regression with Z as a response variable are clearly distorted as truncation acts on a response variable.

The situation complicates under Case B, in which only $I = I(Z > c)$ is observed (and again Y is observed for units with $I = 1$ only). In this second case, the above formula is of no use. It may be useful to notice that the distortion may be bounded in modulo. For simplicity let x be a continuous random variable and let $\tilde{\beta}_{Y|X}(x) = \frac{dE(Y|x,I=1)}{dx}$ be the derivatives w.r.t. x of the nonlinear regression function expressing $E[Y \mid x, I = 1]$. Let $\text{dist}(x) = \beta_{Y|X} - \tilde{\beta}_{Y|X}(x)$ be the distortion of the regression coefficient of interest induced by selection. Notice that the distortion varies with x. It is possible to show that under both models $|\text{dist}(x)| \leq \beta_{Z|X}\beta_{Y|Z.X}$ (see Marchetti and Stanghellini 2008, and Hutton and Stanghellini 2011). Although the second term $\beta_{Y|Z.X}$ cannot be estimated from the observable data, this formula allows one to derive bounds for the coefficient of interest. Notice that distortion is null when, in both graphs of Fig. 1, at least one edge but the arrow from X to Y is missing, a situation described in the literature as a *missing (conditionally) at random* mechanism (see e.g. Little and Rubin 2002, Chap. 1).

To summarize, the notion of a covering model may vary from one context to the other. Regression graph models are useful tools to exemplify the differences between models and to point to one choice or the other.

References

Hutton JL, Stanghellini E (2011) Modelling bounded health scores with censored skew-normal distributions. Stat Med 30:368–376

Little RA, Rubin BR (2002) Statistical analysis with missing data, 2nd edn. Wiley, London

Marchetti GM, Stanghellini E (2008) A note on distortions induced by truncation, with application to linear regression systems. Stat Probab Lett 78:824–829

Test
DOI 10.1007/s11749-012-0291-5

DISCUSSION

Rejoinder on: Sequences of regressions and their independences

Nanny Wermuth · Kayvan Sadeghi

Received: 8 March 2012 / Accepted: 9 March 2012

We thank the discussants for their careful reading of the manuscript and their thoughtful comments. It is nice to see that almost each discussant stresses a different contribution of the paper, that the paper initiated already extensions of distributional results and that many other special aspects are emphasized in the discussions. We respond in detail to the comments, with the discussants ordered alphabetically.

Response to Dr. Robert Castelo We agree with Robert that the global Markov property of concentration graphs, also known as their separation criterion, is simple and computationally more attractive than those for directed acyclic graphs and regression graphs. However, concentration graphs have also only one type of edge so that they do not, for instance, permit to integrate a priori available knowledge about a time ordering among the variables into model building processes.

Furthermore, if such an ordering holds and leads to simplifying factorizations of the joint density, these important properties may no longer show in the concentration graph for this density, as illustrated in the paper with a generating graph in Fig. 4 and the corresponding induced concentration graph in Fig. 17. What we try to emphasize is that in the model class intersecting regression and LWF chain graphs, the simpler

Communicated by Domingo Morales.

This rejoinder refers to the comments available at doi:10.1007/s11749-012-0288-0, doi:10.1007/s11749-012-0287-1, doi:10.1007/s11749-012-0286-2, doi:10.1007/s11749-012-0285-3, doi:10.1007/s11749-012-0284-4.

N. Wermuth (✉)
Chalmers Technical University, Gotherburg, Sweden
e-mail: wermuth@chalmers.se

K. Sadeghi
Department of Statistics, University of Oxford, Oxford, UK

Published online: 28 April 2012

global Markov property of the former is available for the latter and that a known ordering of the variables remains unchanged if one moves from a regression graph to its covering LWF graph. Any search in this intersecting class may indeed be computationally more expensive. In such a case, preserving solid prior knowledge is to be weighted against computational gains.

For more variables (p) than observations (n), the prior knowledge about the relations among the variables under study may also be weak. In such situations, the question always arises whether the results of any analysis can be generalized to apply to an underlying real population even if it captures the dependences in a given set of data well. To us, the $p \gg n$ situation is like observing many detailed features of a handful of individuals in one region and trying to judge from this the well-being of the whole population in the given country.

We thank Robert especially for pointing to us further results for nice, intersecting subclasses of graphical Markov models. They are clearly conceptually important, possibly more so for applications in computer science and in artificial intelligence than in statistics.

Response to Dr. Mathias Drton and his colleagues We agree with Mathias, Chris and Andreas that the extension of directed acyclic graphs involves mainly the inclusion of dashed lines, or equivalently of bi-directed arrows, whereas adding a concentration graph for the background variables just leads to a natural parameterization of the variables taken as given, having dependences that are not further explained.

But in contrast to mixed graphs in which a dashed edge is added anywhere to represent a possible underlying hidden variable, they are added in regression graphs as at most one single edge for a node pair, within a set of responses on equal standing that we name joint responses. Such variables are permitted to change directly when there is an intervention on the variables of their past.

Thereby, the attractive recursive factorization property of directed acyclic graphs is preserved for joint responses, and the dashed lines represent undirected dependences that remain still unexplained when all variables of their past have been used to generate their joint conditional distribution. Thus for us, the strongest motivation for using regression graphs is the extension from a set of only single responses to joint responses such that arrows in the regression graphs capture simple research hypotheses based on what is known up to the time when a new joint response becomes available.

Most of the more complex types of mixed graphs, which may have dashed lines added anywhere, correspond to models that are unlikely to be of direct interest when one is formulating research hypotheses in a given substantive context. They are however important as consequences of transforming regression graphs. As already noted in the discussion, such edges result, for instance, after marginalizing over all nodes along two paths to a common ancestor. Then, a dashed line represents a direct confounder, distorting an existing dependence. The discussants illustrate nicely how complex such constraints can become in general.

If, on the other hand, a set of variables is mutually independent given a common parent node, then marginalizing over the parent induces a complete covariance graph as well as simple parametric constraints. This is known, for instance, from

linear factor analysis with a single unobserved factor. We think that searching for non-independence constraints in general is likely to be unfruitful, that is, when there is no prior knowledge, no well-understood stepwise data-generating process or no design to generate simple scores.

The quoted results by Ali et al. on Markov equivalence of maximal ancestral graphs have been simplified for regression graphs; see Sadeghi (2012). The conditions for Markov equivalence of maximal ancestral graphs as well as the simpler conditions for Markov equivalence of regression graphs have been implemented in the ggm-package in R; see Sadeghi and Marchetti (2011, 2012).

Mathias and his colleagues present also a chordless 4-cycle for a Gaussian 4×4 covariance matrix and say that it cannot be generated by an underlying hidden variable model when there is precisely one off-diagonal negative element. We have not checked this, but want to point out that there is a data-generating process in the given sets of two variables, each with zero means and equal variances.

For two marginally independent regressors, say $b = \{2, 4\}$, two conditionally independent responses $a = \{1, 3\}$ given b, and an orthogonal regression coefficient matrix $\Pi_{a|b}$, it follows that the responses are also marginally independent and the regressors are also conditionally independent given a. Thus, in such a regular Gaussian family of distributions with $E(Y_a|Y_b = y_b) = \Pi_{a|b} y_b$ and $E(Y_b) = 0$, there is one negative element in $\Pi_{a|b}$ and the covariance matrix as well as the concentration matrix has zeros in positions $(1, 3)$ and $(2, 4)$ but nowhere else. Here, the non-independence constraint is the orthogonality of $\Pi_{a|b}$.

Response to Dr. Bala Rajaratnam The discussion by Bala is focused on Gaussian regression graph models after pointing out that this class of models is now established as a useful framework in general. He builds mainly on the recursive factorization of regression graph models and on our notion of covering graph models. The latter may lead to simpler estimation procedures than a corresponding, more parsimonious, reduced model that contains more independence statements.

His contribution shows in particular, how much more additional work is needed if precise distributional results are to be derived for estimators based on specific distributional assumptions, especially in the case of multivariate Bayes estimation with flexible conjugate prior distributions. Bala is to be congratulated for describing a promising approach by which the available results for the three main subclasses of regression graphs (directed acyclic, concentration and covariance graphs) can be combined to obtain distributional results of Bayes estimators for sequences of regressions.

The model selection and fitting approach described in the paper for one given research question and one set of given data emphasizes instead how far one can get by using repeatedly standard single response regressions, linear and logistic regressions that include possibly nonlinear and interactive effects. One can use such a local modeling approach to obtain one or several hypotheses on which specific sequences of regressions may have generated the data.

In general, one has to pay a price for this. One does not obtain an estimator under a precisely specified distribution nor under the precise global independence structure specified by a graph. The gain is that an excellent or poor fit of each independence

statement defining the graph becomes evident in terms of single-response regressions even in these joint response models. The reasons are (1) that each independence statement defining a regression graph also shows in zero parameters of appropriately defined single response regressions, and (2) that for quantitative response regressions and for binary response regressions, the parameters are interpretable measures of conditional dependence, for which point and interval estimates can be obtained that are robust under a variety of sampling schemes.

Response to Dr. Monia Lupparelli and Dr. Alberto Roverato We agree with Monia and Alberto that the first ordering of the variables into a sequence of joint and single responses is a crucial step in the analysis of data. In fact, we see it as major advantages of regression graphs that corresponding solid knowledge can be incorporated into the model selection process, that it is possible not to condition on other components of the same joint response and that one still can rely on standard statistical tools.

The ordering has indeed non-negligible consequences for the results. Thus, in case the prior knowledge about an ordering is weak, the ordering is not agreed upon by all involved in the substantive research on the variables under study, or alternative orderings appear equally plausible, it is important to understand what a given generating process implies when the ordering, and with it the conditioning sets of responses, is changed. For this, the results on Markov equivalence and on independence-preserving graphs that result after marginalizing and conditioning in regression graphs are important, but the more direct task is to obtain graphs induced by just changing which variables are considered as possibly explanatory to which responses and to understand under which conditions on a given generating process, an induced edge in such a graph will also correspond to an induced dependence. This has been treated, in an admittedly not too reader-friendly way, for generating graphs that are directed and acyclic (see Wermuth and Cox 2004), and more recently for those that are regression graphs (see Wermuth 2012).

We also agree that responses should not be regarded as being on equal standing just because 'nothing is known about the independence structure of those variables.' We have been in the lucky situation of having had intensive discussions with psychologists and physicians in the case of the data for which analyses are reported in the paper, and with researcher in other substantive fields on different occasions, where a good agreement on the most likely ordering of joint and single responses was reached. Furthermore, the individual regression results, used to build a regression graph from data analyses, were in good agreement with what was known from previous studies or from theoretical considerations.

However, if one has to rely only on results of model search procedures, we agree that a wide class of models could be preferable to one that narrows down the options. In this connection, we have experienced that searches in model classes that do not permit to specify at least some nonlinear or interactive dependences may choose wrong sets of explanatory variables.

Response to Dr. Elena Stanghellini We thank Elena for appreciating our results on distributions that are faithful to regression graphs. It has indeed taken a long time to come up with its essential components. Only recently it was brought to our attention

though that the term 'weak transitivity' has been in use with different meanings in the literature. In the computer science papers that we quote it means marginalizing over a set of variables, whereas for instance Milan Studený uses it to mean marginalizing over a single variable which coincides with what we name singleton-transitivity. Thus one has to watch out for such confusing differences in definitions.

Requiring a distribution to be faithful to a graph imposes typically severe constraints on the parameter space (see, for instance, the discussion of Fig. 1 for just three variables in Wermuth et al. 2009), and it excludes even subclasses of regular Gaussian distributions such as the family described in this rejoinder when answering Mathias Drton and his colleagues.

It is now known that one can trace dependences in sequences of regressions, in a similar way as Sewall Wright did almost a century ago for his exclusively linear, directed acyclic graph models, provided the distribution has the properties of what has been named a compositional graphoid by Sadeghi and Lauritzen (2012), and it satisfies singleton-transitivity as well; see Wermuth (2012). Such models, are similar to faithful distributions but do not require weak transitivity with respect to sets of variables. Thereby, for instance the whole family of regular Gaussian distributions is again covered; see the proof by Studený (2005), Corollary 2.5 in Sect. 2.3.6.

So far, we have never considered the extension of a regression graph as described by Elena, that is, to include nodes in the graph that are not observable random variables but represent a decision or a fact on how individuals are selected into a study. It is enlightening to see how this can lead, in the case of a simple directed cyclic graph and a simple regression graph, to different results when there is an outcome-dependent selection. Also, we have now learned by her contribution that the missing-at-random assumption in such graphs can be represented by a missing edge that has the selection node as one of its endpoints.

Outcome-dependent sampling is carried to its extreme in case-control studies, where samples from two populations arise. Cases are those diagnosed to have a disease and controls are those not diagnosed with the disease under study. For rare diseases, the observed controls form essentially a sample from the general population. This special sampling design leads to cost-effective data and to logistic regression as the standard tool to identify those features that are risks and other important explanatory variables for the binary response having the two disease classifications as its levels.

It has now been shown how the sampling design of case-control data leads to concentration graphs capturing structure for cases and controls even though the goal is typically to understand which sequences of regressions are relevant. In such studies, the separate dependence structures of cases and controls, together with convincing data summaries to supplement model based estimates, are the key for gaining more insights than with logistic regressions alone; see Wermuth et al. (2012).

End of rejoinder To end, we want to summarize what we see as the most promising aspects of the discussion paper: (1) how simple the criteria for the global Markov property and for the Markov equivalence of regression graphs are, even though the class includes directed acyclic graphs and two types of undirected graphs, (2) how one can use standard single response regressions to get closer to an understanding of how

sequences of joint response regressions might have been generated and to using the full potential of a graph both regarding its dependence and independence structure, and (3) how one can get easily in terms of covering models to classes intersecting with other types of graphical Markov models. For instance, for the class of regression graphs intersecting with LWF chain graphs, the given ordering of the joint responses in the regression graphs is preserved and by ignoring relevant local independences of regression graphs, one just excludes seemingly unrelated regressions and models with incomplete covariance graphs of any connected response component.

References

Sadeghi K (2012) Markov equivalences for subclasses of loopless mixed graphs (submitted). http://arxiv.org/abs/1110.4539

Sadeghi K, Lauritzen SL (2012) Markov properties of mixed graphs. Ann Stat (submitted). http://arxiv.org/abs/1109.5909

Sadeghi K, Marchetti GM (2011). Subroutines available in the R-package ggm vignette

Sadeghi K, Marchetti GM (2012) Graphical Markov models with mixed graphs in R (submitted)

Studený M (2005) Probabilistic conditional independence structures. Springer, London

Wermuth N (2012) Sequences of regressions and their dependences (submitted). http://arxiv.org/abs/1110.1986

Wermuth N, Cox DR (2004) Joint response graphs and separation induced by triangular systems. J R Stat Soc B 66:687–717

Wermuth N, Marchetti GM, Cox DR (2009) Triangular systems for symmetric binary variables. Electron J Stat 3:932–955

Wermuth N, Marchetti GM, Byrnes G (2012) Case-control studies for rare diseases: improved estimation of several risks and of feature dependences (submitted). http://arxiv.org/abs/1203.1829