Traceable regressions

Nanny Wermuth
Chalmers Technical University, Gothenburg, and
International Agency of Research on Cancer, Lyon

Fields Institute, Toronto, April 2012
Set-up for sequences of regressions in vector variables $Y_a Y_b \ldots$

\[
\begin{array}{cccc}
  & a & b & c & d & e \\
  Y_a & Y_b & Y_c & Y_d & Y_e \\
  \text{primary responses} & \text{intermediate variables} & \text{intermediate variables} & \text{treatment variables} & \text{context variables}
\end{array}
\]

Main goal: understanding development with data from

– cohort studies, multi-wave panel data
– studies with randomized, sequential interventions
– cross-sectional and even retrospective studies
General motivation

• Trying to understand short- and long-term effects of risks or of interventions is motivating empirical research in many fields of science

• For this, the main purpose of statistical planning, analysis and interpretation is to capture and use potential data generating processes and to trace pathways of dependence

• Sequences of multivariate or univariate regressions, simplified by independences, provide a flexible framework; joint responses may be discrete, continuous or be mixed of both types
A regression graph, $G_{\text{reg}}^N$, is traditionally the focus of interest.

$G_{\text{reg}}^N$ is a chain graph defined by node set $N$ and three types of edge sets $E\leftarrow$, $E\rightarrow$, and $E\rightarrow$

It has
- a split of $N = (u, v)$ with sequences of
- response nodes coupled as $\circ\cdots\circ$ in $u$ and
- context nodes coupled as $\circ\longrightarrow\circ$ in $v$
- a unique set of the concurrent nodes in $g_j$ for $j = 1, \ldots, J$
- in each compatible ordering of $g_j$, arrows, $\circ\leftarrow\circ$, never point to

$$g>_j = g_{j+1} \cup g_{j+2}, \ldots, \cup g_J$$
Example for a refined sets of concurrent nodes in $g_j$ obtained by statistical analyses after a first ordering into five blocks within a set of concurrent nodes, $g_j$, each node can be reached via at least one undirected path, no order is implied by stacked boxes
Example continued: deleting all arrows gives uniquely the sets of concurrent responses and concurrent context variables, the chain components \( g_j \)
A joint density $f_N$ is said to be generated over $G_{\text{reg}}^N$ if it has the basic factorizations with regressions $f_{g_j|g>_j}$ as

$$f_N = f_{u|v}f_v$$

with $f_{u|v} = \prod_{j \in u} f_{g_j|g>_j}$ and $f_v = \prod_{j \in v} f_{g_j}$

and satisfies the independences implied for each missing edge

For $i, k$ a node pair and $c \subset N \setminus \{i, k\}$, we have in general

$$i \perp k | c \iff (f_{i|kc} = f_{i|c}) \iff f_{ik|c} = (f_{i|c}f_{k|c})$$
For tracing pathways of dependence, the variable pairs needed to generate $f_N$ are instead the focus of interest and the substantive context determines which variable pairs are modeled by a conditional independence and which variable pairs are taken to be dependent.

Suppose one regressor is a risk factor for a response, then the prevention of the risk is generally judged to be of quite different importance if, for instance, the response is
– the occurrence of a common cold
– the infection with an HIV virus or
– an accident in a nuclear plant.
We write $i \nabla k \mid c$ for $Y_i, Y_k$ conditionally dependent given $Y_c$ for some $c \subset N \setminus \{i, k\}$

A graph is **edge-minimal** for a distribution generated over it, if every missing edge in the graph corresponds to a conditional independence statement and every edge present to a dependence statement.

A dependent variable pair $Y_i, Y_k$ is one needed in the generating process of $f_N$ and a family of densities $f_N$ generated over an edge-minimal graph changes if any one edge is removed from the graph.
Defining dependences and independences for an edge-minimal $G_{\text{reg}}^N$

**Definition 1**

An edge-minimal regression graph with $N = (u, v)$ and $g_1 < \cdots < g_J$ specifies a generating process for $f_N$, where

1. **edges present** in $G_{\text{reg}}^N$
   - $i \rightarrow k : i \triangleright k \mid g_j$ for $i, k$ concurrent response nodes in $g_j$ of $u$
   - $i \leftarrow k : i \triangleright k \mid g_j \setminus \{k\}$ for response $i$ in $g_j$ of $u$ and explanatory $k$ in $g_j$
   - $i \rightarrow k : i \triangleright k \mid v \setminus \{i, k\}$ for $i, k$ concurrent context nodes in $g_j$ of $v$

2. **edges missing** in $G_{\text{reg}}^N$ when the dependence sign $\triangleright$ is replaced by $\perp$
Thus, for an edge-minimal $G_{\text{reg}}^N$

– one fixed ordering of $g_j$ is assumed, so that the density of variables in $g_J$ is generated first,
the one of $g_{J-1}$ given $g_J$ next,
up to the density of $g_1$ given $g_{>1}$

– the graph implies for each variable pair either conditional dependence or independence given the same type of conditioning set

– for each node $i$ of $g_j$ in $u$, nodes in
$g_{>j} = g_{j+1} \cup g_{j+2}, \ldots, g_{J-1} \cup g_J$ are in the past of $g_j$
Requirements for two results on the independence structure of $G_N^{\text{reg}}$

Let $a, b, c, d$ denote disjoint subsets of $N$ where only $d$ may be empty and let any joint independence $b \perp\!\!\!\!\!\!\!\perp ac|d$ have three equivalent decompositions as

(i) $(b \perp a|cd$ and $b \perp c|d)$
(ii) $(b \perp a|d$ and $b \perp c|d)$
(iii) $(b \perp a|cd$ and $b \perp c|ad)$

then (i) named contraction, holds for all probability distributions (ii) combines decomposition and composition, holds in a regression when there is also a main-effect for every higher-order interactive or nonlinear dependence (iii) combines weak union and intersection, holds for all positive distributions.
Given the three equivalent decompositions of any joint dependence, active paths in $G^N_{\text{reg}}$ can be expressed in terms of anterior paths.

An **anterior $i{k}$-path** is a descendant-ancestor $i{q}$-path with a context-nodes $q{k}$-path attached to it (or any subpath)

$$i \leftarrow \underset{\text{ancestors of } i}{\circ \leftarrow \circ, \ldots, \circ \leftarrow q \leftarrow \circ, \ldots, \circ \leftarrow k} \underset{\text{antiters of } i}{\circ, \ldots, \circ \leftarrow k}$$

Let $\{a, b, c, m\}$ partition $N$, where $c$ denotes a conditioning set of interest for $a, b$ and $m$ the set of nodes to be ignored.

A **path in $G^N_{\text{reg}}$ is active given $c$** if of its inner nodes, every collision node is in $c \cup \text{ant}_c$ and every transmitting node is in $m$.
Lemma 1
Global Markov property of $G_{\text{reg}}^N$ (Sadeghi, 2009) $G_{\text{reg}}^N$ implies $a \perp \!\!\! \perp b \mid c$ if and only if there is no active path in $G_{\text{reg}}^N$ between $a$ and $b$ given $c$

Lemma 2
Equivalence of the pairwise and the global Markov property
(Sadeghi and Lauritzen, 2012) The independence structure of $G_{\text{reg}}^N$ is equivalently defined by its lists of the three types of missing edges and by its global Markov property.
Two-edge subgraphs induced by three nodes in $G_{\text{reg}}^N$, named Vs

There are just two basic types of Vs in $G_{\text{reg}}^N$

**collision Vs:**

\[ i \overset{\circ}{\leftarrow} k, \ i \rightarrow \overset{\circ}{\leftarrow} k, \ i \overset{\circ}{\leftarrow} \overset{\circ}{\leftarrow} k, \]

and **transmitting Vs:**

\[ i \leftarrow \overset{\circ}{\leftarrow} k, \ i \leftarrow \overset{\circ}{\leftarrow} k, \ i \rightarrow \overset{\circ}{\leftarrow} k, \ i \leftarrow \overset{\circ}{\rightarrow} k, \ i \leftarrow \overset{\circ}{\rightarrow} \rightarrow k \]
Lemma 3

Markov equivalence (Wermuth and Sadeghi, 2012) Two regression graphs with the same skeleton are Markov equivalent if and only if their sets of collision Vs are identical.

Lemma 4

The conditioning set of any independence statement implied by $G^N_{reg}$ for the endpoints of any of its Vs, includes the inner node if it is a transmitting V and excludes the inner node if it is collision V.
To make \( V \) dependence-inducing, we take an edge-minimal regression graph for \( f_N \), assume the three equivalent decompositions of a joint dependence and require in addition singleton transitivity.

**Singleton transitivity.** For \( i, h, k \) distinct nodes and \( d \subseteq N \setminus \{i, h, k\} \)

\[
(i \perp k|d \text{ and } i \perp k|hd) \implies (i \perp h|d \text{ or } k \perp h|d)
\]

Thus, for a conditional independence of \( Y_i, Y_k \) given \( Y_d \) and given \( Y_h, Y_d \) to hold both, there has to be at least one additional independence given \( Y_c \) involving \( Y_h \).

An edge-minimal \( G_{\text{reg}}^N \) forms a **dependence base** for \( f_N \), generated over it, if singleton transitivity holds (always for \( f_{g_j|g_{>j}}, f_{g_{>j}} \) a cut for all \( j \)).
Proposition 1

Dependence inducing $Vs$. For $(i, o, k)$ any $V$ of a dependence base $G^N_{\text{reg}}$ and each $c \subseteq N \setminus \{i, k, o\}$ such that this regression graph implies one of $i \perp k|c$ or $i \perp k|oc$, the following two equivalent statements hold:

- $(i, o, k)$ forms a collision $V \iff (i \perp k|c \implies i \bowtie k|oc)$
- $(i, o, k)$ forms a transmitting $V \iff (i \perp k|oc \implies i \bowtie k|c)$

Thus, in a dependence base $G^N_{\text{reg}}$, conditioning on the inner node of a collision $V$ and marginalizing over the inner node of transmitting $V$ is dependence-inducing for the endpoints of the $V$ given any appropriate $c$. 
Definition 2

**Traceable regressions.** For \( \{a, b, c, d\} \) partitioning \( N \), we say \( f_N \) results from traceable regressions if

1. it could have been generated over a dependence base regression graph, \( G_{\text{reg}}^N \),

2. it has the three equivalent decompositions of the joint independence \( b \perp \perp ac \!
\! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
\! \! \! \!
Next goal:

Obtaining a matrix criterion to decide whether a dependence base $G^N_{\text{reg}}$ implies $\alpha \perp \beta | c$ or $\alpha \perp \beta | c$ for partitioning

We use edge matrix representation of $G^N_{\text{reg}}$: adjacency matrices with ones added along the diagonal so that sums of products of submatrices become well-defined

First task:

Given $N = (u, v)$ and the edge matrices of $G^N_{\text{reg}}$ for $f_N = f_{u|v}f_v$

find the implied edge-matrices for another split $N = (a, b)$ with $a = \alpha \cup m$, $b = \beta \cup c$ and $G^N_{\text{reg}} - a|b$ for $f_N = f_{a|b}f_b$ having multivariate regression of $Y_a$ on $Y_b$ and a concentration graph for $Y_b$
Regression graphs have three types of edge sets, $E_\prec$, $E_-$, and $E_{\prec}$

The edge matrix components of $G^N_{\text{reg}}$ are a $d_N \times d_N$ upper block-triangular matrix $\mathcal{H}_{NN} = (\mathcal{H}_{ik})$ such that

$$\mathcal{H}_{ik} = \begin{cases} 1 & \text{if and only if } i \prec k \text{ or } i \rightarrow k \text{ in } G^N_{\text{reg}} \text{ or } i = k, \\ 0 & \text{otherwise}, \end{cases}$$

and a $d_u \times d_u$ symmetric matrix $\mathcal{W}_{uu} = (\mathcal{W}_{ik})$ such that

$$\mathcal{W}_{ik} = \begin{cases} 1 & \text{if and only if } i \rightarrow k \text{ in } G^N_{\text{reg}} \text{ or } i = k, \\ 0 & \text{otherwise}, \end{cases}$$

where, $E_-$ corresponds to $\mathcal{W}_{uu}$, $E_{\prec}$ to $\mathcal{H}_{vv}$, and $E_\prec$ to $\mathcal{H}_{uN}$ ($\mathcal{W}_{uv} = 0, \mathcal{W}_{vu} = 0, \mathcal{W}_{vv} = \mathcal{H}_{vv}$)
Example

For a Gaussian family in a mean-centered $Y_N$ generated over $G^N_{\text{reg}}$ with just two concurrent response sets $a, b$, the parameter matrices are for

$$H_{NN}Y_N = \varepsilon_N, \quad \text{cov}(\varepsilon_N) = W_{NN}$$

\[
H_{NN} = \begin{pmatrix}
I_{aa} & -\Pi_{a|b.v} & -\Pi_{a|v.b} \\
0_{ba} & I_{bb} & -\Pi_{b|v} \\
0_{va} & 0_{vb} & \Sigma_{vv.ab}
\end{pmatrix}
\]

\[
W_{NN} = \begin{pmatrix}
\Sigma_{aa|bv} & 0_{ab} & 0_{av} \\
0_{ba} & \Sigma_{bb|v} & 0_{bv} \\
0_{va} & 0_{vb} & \Sigma_{vv.ab}
\end{pmatrix}
\]

where the Yule-Cochran notation is used: $\Pi_{a|bv} = (\Pi_{a|b.v} \Pi_{a|v.b})$; edge matrices $\mathcal{H}_{NN}, \mathcal{W}_{NN}$ implicitly define such Gaussian families.
Partial closure

The edge matrix calculus of Wermuth, Wiedenbeck and Cox (2006) uses partial closure, denoted by $\text{zer}_a(\mathcal{F})$, which operates on all nodes $i$ in $a \subseteq N$ of a symmetric edge matrix $\mathcal{F}$.

After a reordering to have node $i$ in position $(1,1)$ of $\tilde{\mathcal{F}}$ and $b = N \setminus i$

$$\text{zer}_i \tilde{\mathcal{F}} = \text{In}[\begin{pmatrix} 1 & \mathcal{F}_{ib} \\ \mathcal{F}_{bi} & \mathcal{F}_{bb} + \mathcal{F}_{bi}\mathcal{F}_{ib} \end{pmatrix}]$$

is seen to close, by an edge, every $V$ with inner node $i$
Basic properties of partial closure

Partial closure is

(i) commutative

(ii) cannot be undone and

(iii) is exchangeable with selecting a submatrix

Lemma 5

Partial closure applied to $G^N_{\text{reg}}$. For $N = (a, b)$, the transformation $K_{NN} = \text{zer}_a(H_{NN})$ closes each $a$-line anterior path and $Q_{uu} = \text{zer}_b(W_{uu})$ each dashed, $b$-line collision path
Examples of three dependence base, 3-node graphs

Active path (1, 2, 3) induces in a) $1 \rhd 3$, in b) $1 \rhd 3|2$, and in c) $1 \rhd 3$

By letting the edge induced by the three $V$'s ‘remember the type of edge at the path endpoints’ the induced edges become in

a) $1 \leftarrow 3$, b) $1 \leftarrow 3$, c) $1 \leftarrow 3$
For $N = (a, b)$, $o_a$ nodes in $a$, $o_b$ nodes in $b$ and $i, k$ the endpoints of paths that are active for $G_{\text{reg}}^{N-a|b}$, there remain three types of active $ik$-path given $b$ in the graph having edge matrices $K_{NN}$ and $Q_{uu}$:

$$i \leftarrow o_a \rightarrow o_b \leftarrow k, \; i \leftarrow o_a \rightarrow o_a \rightarrow k, \; i \rightarrow o_b \rightarrow o_b \leftarrow k$$

**Proposition 2**

The active path remaining in $K_{NN} = \text{zer}_a(H_{NN})$, $Q_{uu} = \text{zer}_b(W_{uu})$ for $G_{\text{reg}}^{N-a|b}$ are closed with the induced edge matrices $P_{a|b}$, $S_{aa|b}$, $S_{bb}$

$$P_{a|b} = \text{In}[K_{ab} + K_{aa}Q_{ab}K_{bb}]$$

$$S_{aa|b} = \text{In}[K_{aa}Q_{aa}K_{aa}^T], \; S_{bb,a} = \text{In}[H_{bb}^TQ_{bb}H_{bb}]$$
After remembering the types of edge at the path endpoints, we have with $P_{a\mid b}$ an induced bipartite graph of arrows pointing from $b$ to $a$

$S_{aa\mid b}$ an induced covariance graph

$S^{bb.a}$ an induced concentration graph

**Lemma 6**

**Edge matrices induced by** $G_{\text{reg}}^N$ **for** $f_{\alpha\beta\mid c}$. The subgraph induced by nodes $\alpha \cup \beta$ in $G_{\text{reg}}^{N-a\mid b}$ captures the independence implications of $G_{\text{reg}}^N$ for $f_{\alpha\mid \beta c}f_{\beta\mid c}$ with multivariate regression of $Y_\alpha$ on $Y_\beta, Y_c$ and conditional concentration graph for $Y_\beta$ given $Y_c$

This subgraph has induced edge matrices

$P_{\alpha\mid \beta . c} = [P_{a\mid b}]_{\alpha, \beta}$

$S_{\alpha a\mid b} = [S_{aa\mid b}]_{\alpha, \alpha}$

$S^{\beta \beta . a} = [S^{bb.a}]_{\beta \beta}$
Proposition 3

Edge criteria for implied independences and dependences

A dependence base $G_{\text{reg}}^N$ implies $\alpha \perp \beta | c$ if $\mathcal{P}_{\alpha|\beta.c} = 0$ and it implies $\alpha \mid\!\mid \beta | c$ if $\mathcal{P}_{\alpha|\beta.c} \neq 0$

Corollary

The transformations of $G_{\text{reg}}^N$ to get $\mathcal{P}_{\alpha|\beta.c}$ use implicitly set transitivity since edges may be introduced but never removed.

For $a, b, c, d$ disjoint subsets of index set $N$, set transitivity means

$$ (a \perp b|d \text{ and } a \perp b|cd) \implies (a \perp c|d \text{ or } b \perp c|d) $$

Thus, the implications of the graph for a generated family ignores path cancellations, that are possible for a member
Most recent relevant work
Wermuth (2011) Bernoulli
Wermuth and Sadeghi (2012), to appear as invited discussion paper in TEST
A regular Gaussian family violating set transitivity. For $\mathcal{N} = (u, v)$, let $Y_u$ and $Y_v$ be mean-centered vector variables with a joint Gaussian distribution. Let them have equal dimension, $d_u$, the components of $Y_u$ and of $Y_u$ be mutually independent and all elements in the covariance matrix $\text{cov}(Y_u, Y_v) = \Sigma_{uv}$ be nonzero, then

$$\text{cov}(Y_u) = \Sigma_{uu} \text{ diagonal}, \quad \text{cov}(Y_v) = \Sigma_{vv} \text{ diagonal}$$

Let further the components of $Y_v$ have equal variances $\omega > 1$ and the equal variances of the components $Y_u$ be $\kappa > \omega + 1$. Whenever in the described situation $\Sigma_{uv}$ is orthogonal, then also

$$\text{cov}(Y_u|Y_v) = \Sigma_{uu|v} \text{ diagonal}, \quad \text{cov}(Y_v|Y_u) = \Sigma_{vv|u} \text{ diagonal}$$