Graphical Representation of Multivariate Dependencies

David Lee

Graphical Models Reading Group

August 7, 2015

Visualizing dependence structures with graphs

- The presentation is based on the book by Cox and Wermuth (1996)[1].
- This book mainly deals with dependence structures for variables commonly seen in observational studies in the social sciences.
- The use of graphs is emphasized, showing the relationships among purely explanatory variables, intermediate variables and responses.
- Statistical methods mentioned are fairly general.

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

Types of graphs

Several types of graphs are used in this book:

1. *Concentration graph*:
   - Undirected full edges with no boxes, i.e. no distinction among explanatory/response variables.
   - Absence of edge between nodes i and j means that Y_i is conditionally independent of Y_j given all other variables.
   - In multivariate normal distribution, this means a zero in the concentration (precision) matrix.

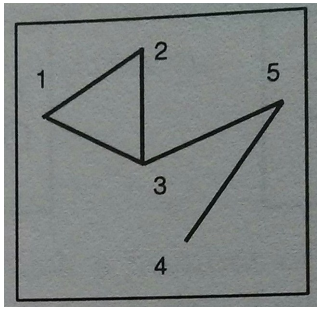


Figure : A concentration graph among 5 variables

Types of graphs (cont.)

2. *Covariance graph*:
   - Undirected dashed edges with no boxes.
   - Absence of edge between nodes i and j means that Y_i is marginally independent of Y_j.
   - In multivariate normal distribution, this means a zero in the covariance matrix.

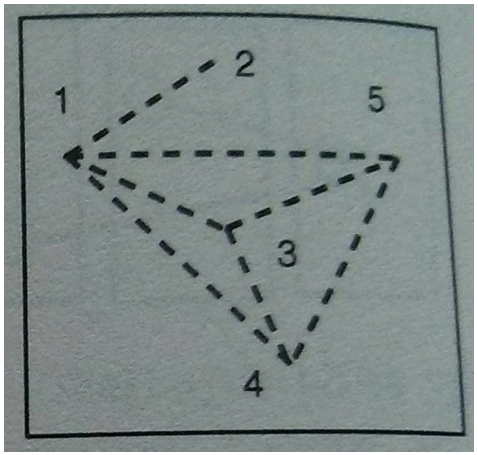


Figure : A covariance graph among 5 variables

Types of graphs (cont.)

3. *Directed acyclic graph*:
   - Directed full edges with no boxes.
   - No cycles allowed.
   - Mainly used to show predictor/response relationships.
   - A given variable is conditionally independent of all ancestors given the immediate parent(s).

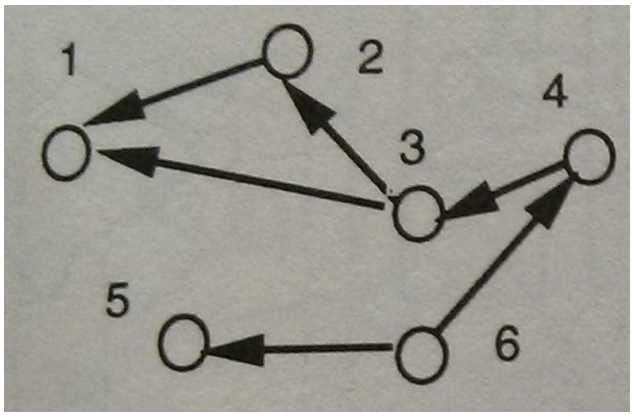


Figure : A directed acyclic graph. Here $Y_2 \perp\!\!\!\perp (Y_4, Y_5, Y_6) | Y_3$.

Types of graphs (cont.)

4. *Univariate recursive regression graph*:
   - Similar to directed acyclic graphs, but with boxes added.
   - Each box contains one variable.
   - Boxes (variables) are arranged from left to right so that a variable is an immediate response (child) to the one to the right.
   - For variables to be treated equally (i.e. neither response nor explanatory to each other), their boxes can stack vertically.
   - Sometimes a double-lined box is used for purely explanatory variables, whose values are regarded as fixed.

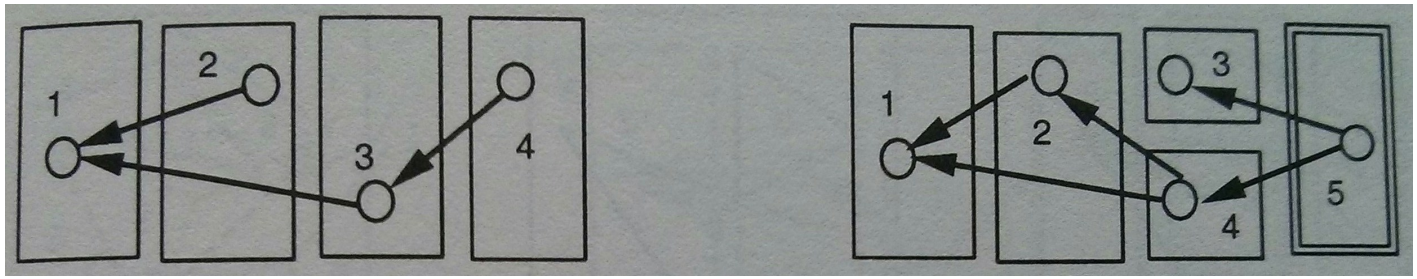


Figure : Examples of univariate recursive regression graphs

Types of graphs (cont.)

- Interpretation: If $i < j$, absence of edge between i and j means that Y_i is conditionally independent of Y_j given all $Y_k, i < k \neq j$. This follows from the regression setting where Y_i is regressed on all variables "to its right", and absence of a linkage means that particular regression coefficient is zero.

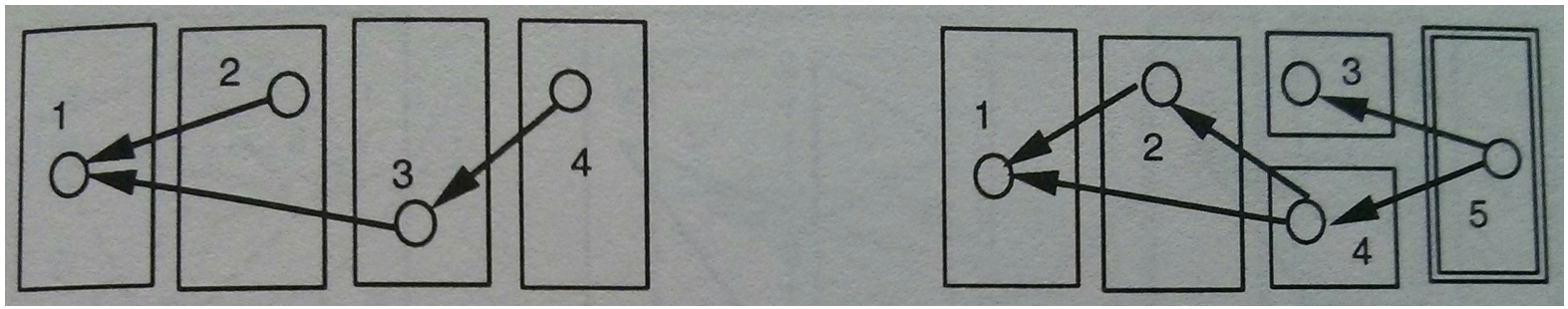


Figure : Same graphs as previous slide. In the left panel, $Y_1 \perp\!\!\!\perp Y_4 | (Y_2, Y_3)$ and $Y_2 \perp\!\!\!\perp (Y_3, Y_4)$.

Types of graphs (cont.)

5. *Joint-response chain graph*:
   - Similar to univariate recursive regression graph, but each box can contain several nodes.
   - Within a box, variables are linked by either undirected full or dashed lines that are interpreted like concentration/covariance graphs, conditional on all boxes to the right.
   - Arrows from one box to other are either all full or dashed:
     - Full arrows imply a variable is regressed on all other variables to the right and those in the same box.
     - Dashed arrows imply a variable is regressed on all other variables to the right only.

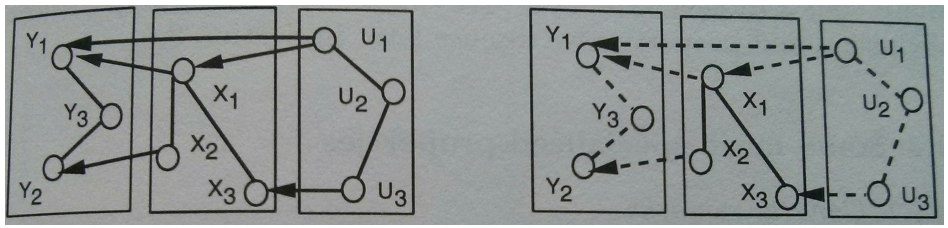


Figure : Two joint-response chain graphs

Types of graphs (cont.)

- In these two examples, U's are purely explanatory, X's are intermediate and Y's are purely response variables.
- Some independence relationships:
  - (Y_1, Y_2) : [Left] $Y_1 \perp\!\!\!\perp Y_2 | (Y_3, X_1, \ldots, U_3)$;
    [Right] $Y_1 \perp\!\!\!\perp Y_2 | (X_1, \ldots, U_3)$;
  - (X_1, U_3): [Left] $X_1 \perp\!\!\!\perp U_3 | (X_2, X_3, U_1, U_2)$; [Right] $X_1 \perp\!\!\!\perp U_3 | (U_1, U_2)$;
  - (X_2, X_3): $X_2 \perp\!\!\!\perp X_3 | (X_1, U_1, U_2, U_3)$ in both.

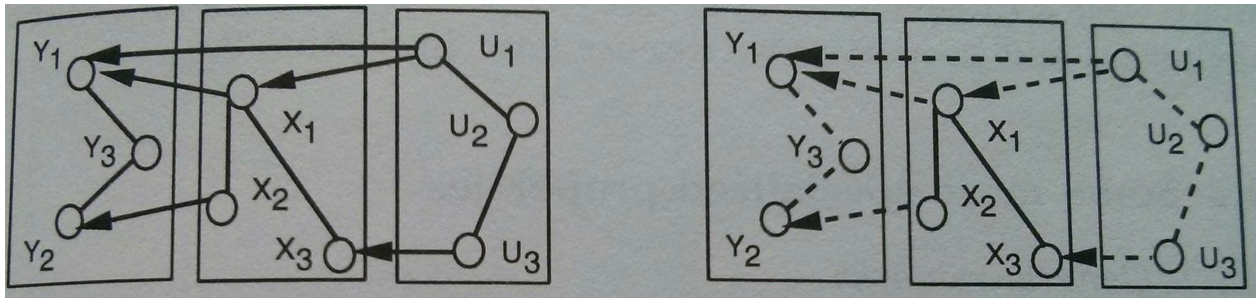


Figure : Same graphs as previous slide

Collision nodes, V- and U- configurations

- A node being response to two or more variables is also called a collision/sink node.
- The figure below shows some sink nodes. Panel (a) is called a sink-oriented V-configuration, while the others are all sink-oriented U-configuration. Sink-oriented configurations have implications on edges to be added when we transform graphs with directed arrows to concentration or covariance graphs.

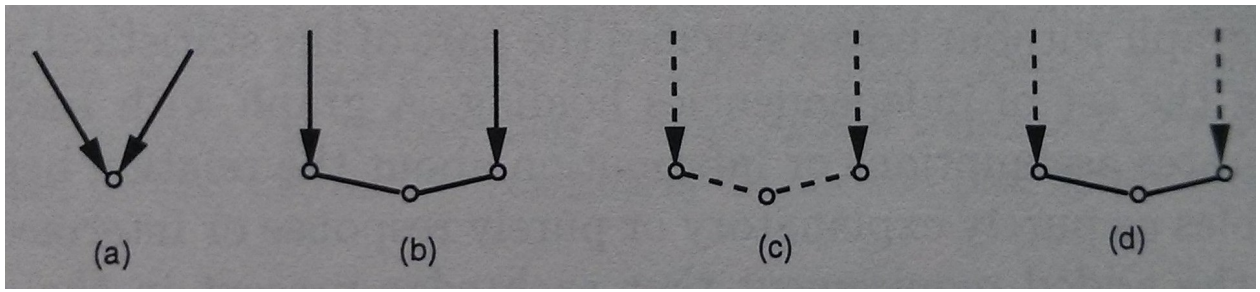


Figure : Examples of V- and U- configurations

Collision nodes, V- and U- configurations (cont.)

- If Y_i and Y_j are marginally or conditionally independent explanatory variables with common response (sink node) Y_t, then the corresponding concentration graph will have a full edge joining nodes i and j.
- If every path between nodes i and j in the directed acyclic graph has a collision node, then $Y_i \perp\!\!\!\perp Y_j$ and there will not be a dashed edge in the corresponding covariance graph.

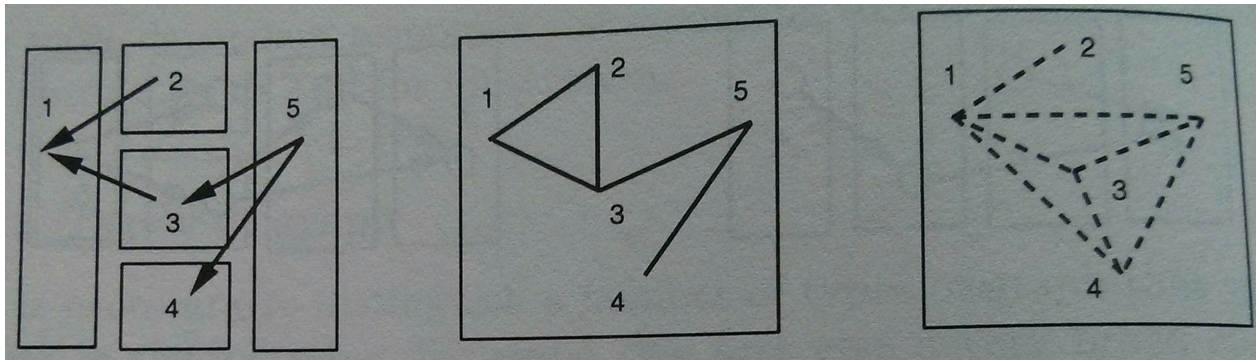


Figure : A directed graph and its corresponding concentration and covariance graphs

Markov properties

- There is a very convenient Markov property for concentration graphs. If a set of nodes C separates two other sets A and B, i.e. going from any node in A to any node in B requires passing through some node in C, then $Y_A \perp\!\!\!\perp Y_B | Y_C$ where Y_S is the collection of random variables represented by the set of nodes S.
- Some conditional independencies in the figure below:
  - (a): $(Y_1, Y_2) \perp\!\!\!\perp (Y_4, Y_5) | Y_3$ and $Y_1 \perp\!\!\!\perp Y_4 | (Y_2, Y_3, Y_5)$;
  - (b): $(Y_1, Y_2) \perp\!\!\!\perp (Y_4, Y_5) | (Y_3, Y_6)$ and $(Y_1, Y_2, Y_3, Y_6) \perp\!\!\!\perp Y_4 | Y_5$.

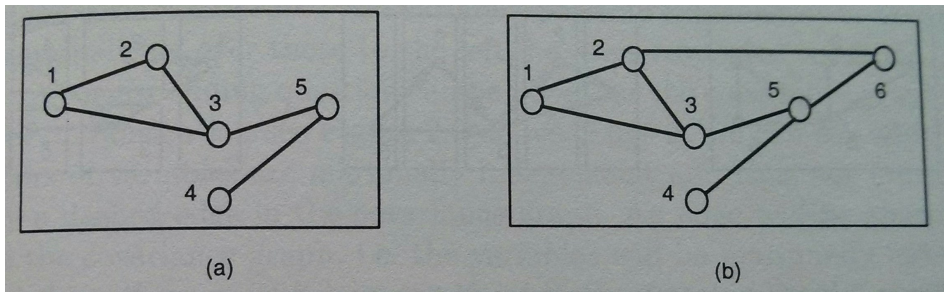


Figure : Two concentration graphs

Decomposability of independence hypotheses

- An independence hypothesis is decomposable if there is an equivalent sequence of independence hypotheses in a univariate recursive regression process, i.e. can be represented as the vanishing of regression coefficients in those multiple regressions.
- For instance, if (X, Y, Z) follows a trivariate normal distribution, the set of independence relationships

$$Y \perp\!\!\!\perp X|Z \quad \text{and} \quad X \perp\!\!\!\perp Z|Y$$

is equivalent to the set

$$Y \perp\!\!\!\perp X|Z \quad \text{and} \quad X \perp\!\!\!\perp Z \tag{1}$$

or to the single independence relationship

$$X \perp\!\!\!\perp (Y, Z). \tag{2}$$

Decomposability of independence hypotheses (cont.)

Relationship (1) corresponds to the regressions $Y \sim X + Z$ and $X \sim Z$, in which the regression coefficient for X in the former and for Z in the latter vanishes. Relationship (2) corresponds to the regression $X \sim Y + Z$ in which both regression coefficients are zero.

- Implication: A decomposable undirected (concentration) graph or partially directed graph, perhaps derived from a joint distribution, can be represented by a univariate recursive regression graph that preserves independencies, i.e. one that is independence-equivalent to the structure imposed by that distribution.

Decomposability of independence hypotheses (cont.)

- A concentration graph can be oriented to be independence-equivalent to a directed acyclic graph if and only if it contains no chordless n-cycle, $n \geq 4$.

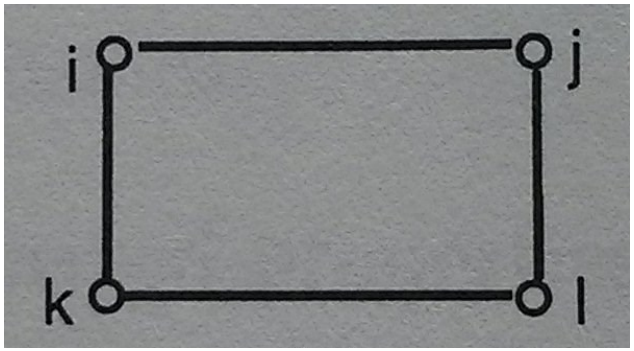


Figure : A chordless 4-cycle

Example 1: Four psychological variables

- In a study of anger and anxiety, test scores were obtained for trait (long-term features) anger/anxiety and state (short-term features) anger/anxiety. The estimated correlations and partial correlations are shown in the following table:

	Y	X	V	U
Y, state anxiety	1	0.45	0.47	-0.04
X, state anger	0.61	1	0.03	0.32
V, trait anxiety	0.62	0.47	1	0.32
U, trait anger	0.39	0.50	0.49	1

Table : Correlations (lower triangle) and partial correlations (upper triangle) for the two traits and two states

Example 1: Four psychological variables (cont.)

- Panel (a) below shows the concentration graph for these variables, and panel (b) is one possible independence-equivalent joint-response chain graph. It can be interpreted as Y regressed on X, V, U and X on Y, V, U, such that $Y \perp\!\!\!\perp U | (X, V)$ and $X \perp\!\!\!\perp V | (Y, U)$.

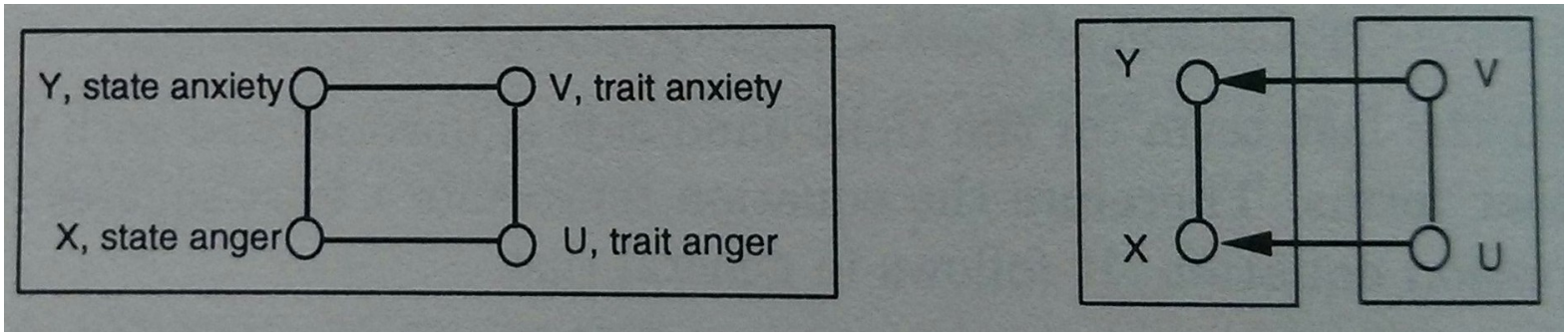


Figure : Concentration graph and a possible joint-response chain graph for the four variables

Example 2: Glucose control

- In this example, we are interested in identifying potential psychological and socio-economic variables that may impact glucose control for diabetic patients.
- Data contain the following for 68 patients in Germany with fewer than 25 years of diabetes:
  - Y [cont.]: Glucose level, measured by the concentration of glycosylated haemoglobin; the lower the better.
  - X [cont.]: Score of patient's knowledge on the illness
  - Scores on three attitudes towards the illness:
    - Z [cont.]: Fatalistic externality (chance determines what occurs)
    - U [cont.]: Social externality (powerful others are responsible)
    - V [cont.]: Internality (the patient him/herself is responsible)

Example 2: Glucose control (cont.)

- Three background/demographic variables considered as purely explanatory:
  - W [cont.]: Duration of illness in years
  - A [binary*]: Duration of formal schooling (1 — at least 13 years)
  - B [binary*]: Gender (1 — female)

  *Binary variables are coded 1 and -1.

- The following figure shows the categorization of these variables; Y the response variable of primary interest, X is a response of secondary interest and intermediate variable. The psychological (attitude) variables are potential explanatory variables to X and responses to intrinsic patient characteristics W, A, B.

Example 2: Glucose control (cont.)

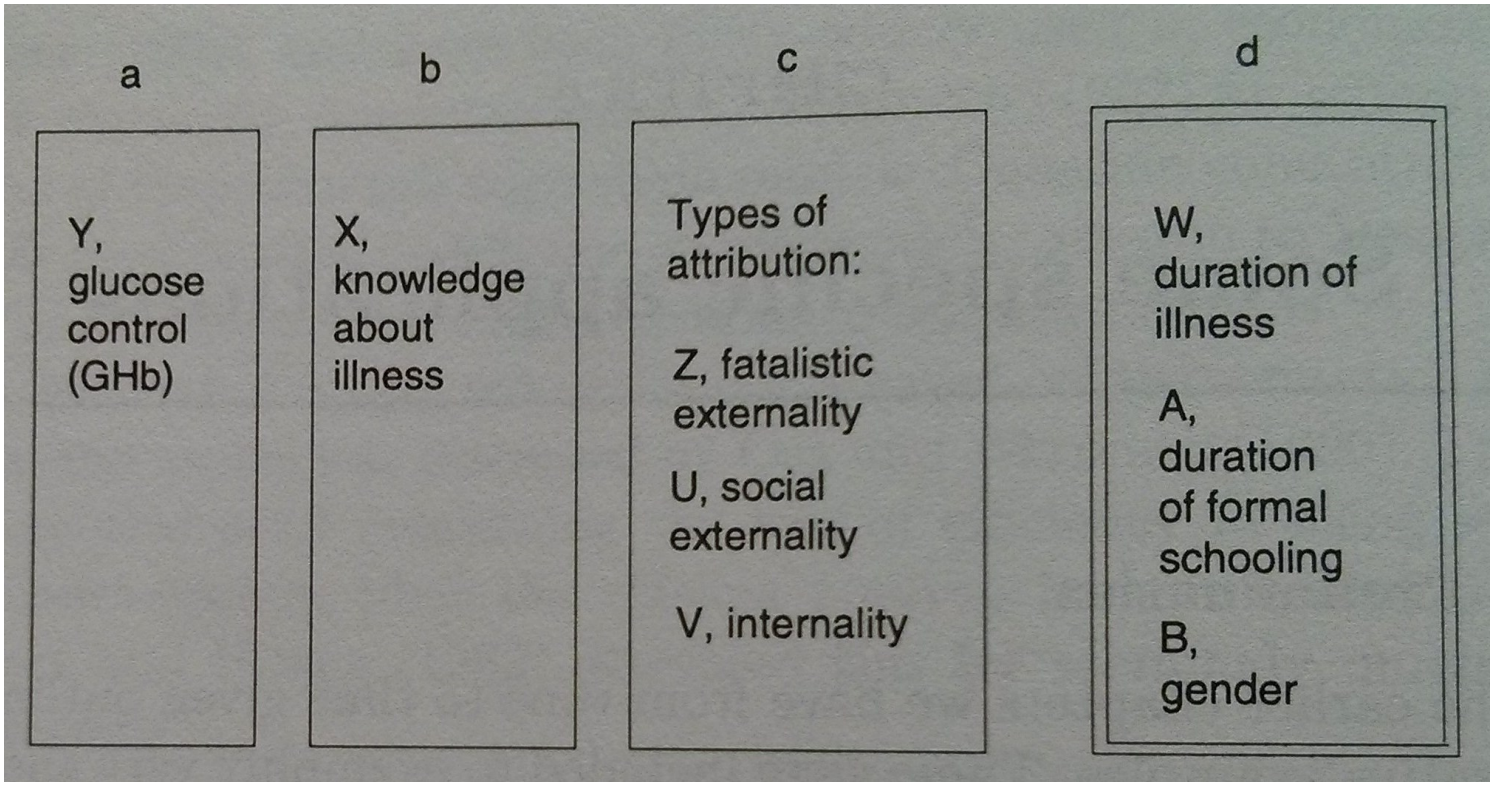


Figure : Grouping of the variables in the glucose control example

Example 2: Glucose control (cont.)

- The following figure is a summary of marginal and bivariate distributions for all variables.

Variable	Y	X	Z	U	V	W	A	B
Y, gl. con.	1	−.24	.09	−.13	.09	−.25	−.29	−.05
X, knowl.	−.34	1	−.34	−.08	.04	.00	.17	.14
Z, fat. ext.	.15	−.49	1	.41	−.26	.28	−.00	.09
U, soc. ext.	.03	−.32	.52	1	−.09	−.08	−.12	−.14
V, intern.	.04	.14	−.33	−.23	1	.16	−.07	−.23
W, dur. ill.	−.12	−.11	.28	.10	.05	1	−.25	.04
A, school.	−.32	.33	−.26	−.20	−.01	−.25	1	−.14
B, gender	−.07	.09	.08	−.06	−.22	.07	−.09	1
min.	5.4	11.0	8.0	8.0	27.0	0	−1	−1
max.	14.1	46.0	33.0	42.0	48.0	24	1	1
mean	9.3	35.4	19.0	24.2	41.3	10.4	-	-
st. dev.	2.0	7.3	5.4	7.0	4.7	7.0	-	-

Figure : Marginal summaries, correlations (lower triangle) and partial correlations (upper triangle) among the variables

Example 2: Glucose control (cont.)

- Regression of Y on all other variables finds only the parameters for X (knowledge), W (disease duration) and A (schooling) "significant". The added W-A interaction is guided by the examination of the cross-product terms.

Response	Explanatory variable							
Y	X	Z	U	V	W	A	B	$W.A$
(a) on all explanatory variables and one interaction term								
est. coeff.	−.06	.02	−.02	.05	−.04	−.53	−.08	.11
st. error	−.03	.05	.04	.05	.03	.23	.22	.03
ratio, t	−1.89	.35	−.71	.98	−1.12	−2.33	−.35	3.48
(b) on X and $A * W$								
est. coeff.	−.061	-	-	-	−.034	−.524	-	.111
st. error	.030	-	-	-	.031	.221	-	.031
ratio, t	−2.02	-	-	-	−1.10	−2.37	-	3.56

$W.A$ is the interaction represented by the product $(W - \bar{W}) \times A$;
$A * W$ consists of main effect terms A, W and the interaction $W.A$

Figure : Summaries of the regressions $Y \sim . + W.A$ and $Y \sim X + W*A$

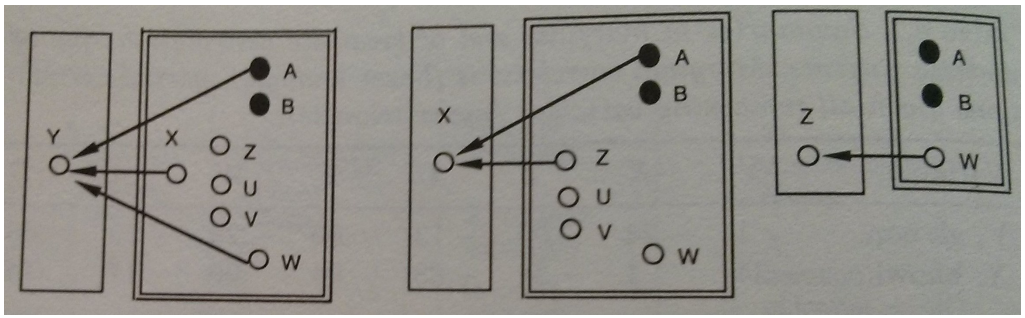
Example 2: Glucose control (cont.)

- Regression of X, knowledge (and Z, fatalistic ext.) on its ancestors results in the reduced models $X \sim Z + A$ and $Z \sim W$:

Response	Expl. variable			Response	Expl. variable	
X	Z	A	const.	Z	W	const.
est. coeff.	−.578	1.610	46.592	est. coeff.	.214	16.805
st. error	.144	.786	-	st. error	.092	-
ratio, t	−4.01	2.05	-	ratio, t	2.33	-

Figure : Summaries of the regressions $X \sim Z + A$ and $Z \sim W$

- These three regressions can be summarized by the following graphs. A filled (empty) circle represents binary (continuous) variable.



Example 2: Glucose control (cont.)

- And these graphs can be combined into the following independence (or univariate recursive regression) graph:

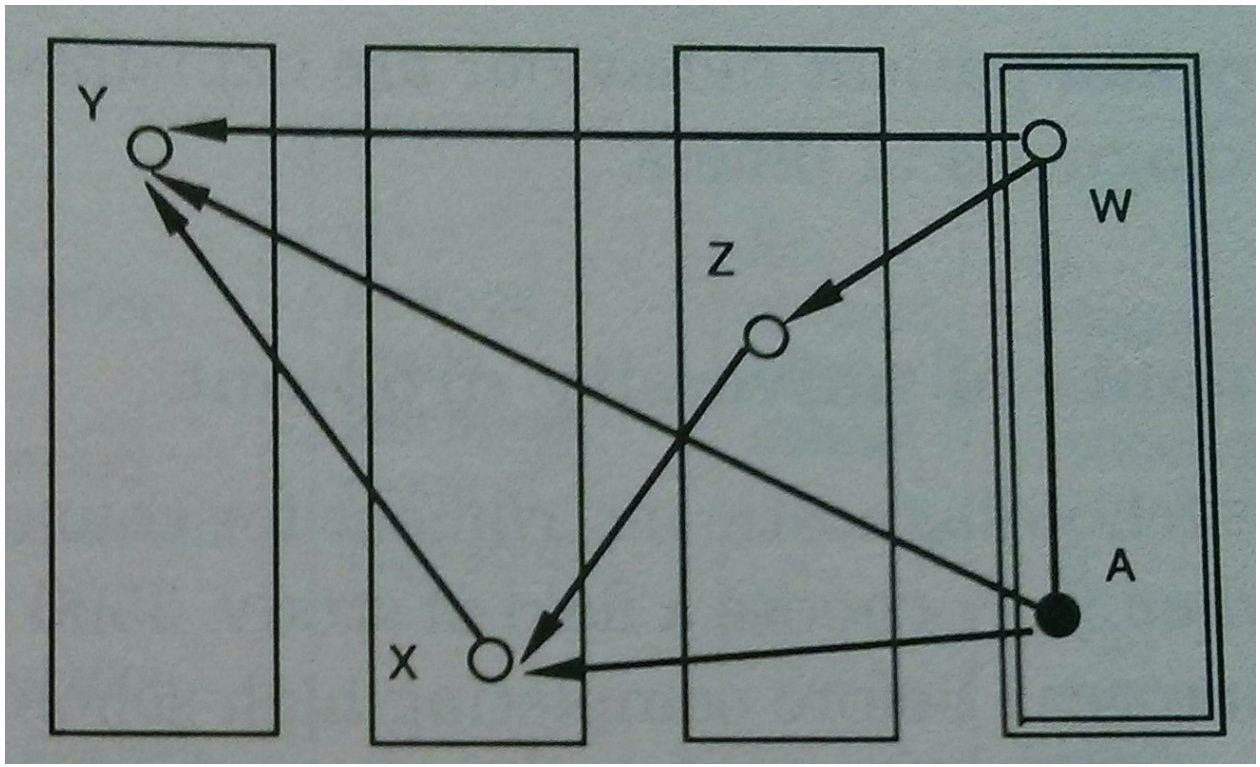


Figure : Independence graph for variables related to Y and X: $Y \sim X + W*A$; $X \sim Z + A$; $Z \sim W$.

Example 2: Glucose control (cont.)

- Some interpretations:
  1. Glucose control is better the more a patient knows about the illness (*from the regression of Y*), and for patients with less formal education, the longer the illness lasted (*from the interaction term*).
  2. Knowledge about diabetes is better for patients with more formal education and those who don't think that they get the disease by chance (*from the regression of X*).
  3. Fatalistic externality is higher with a longer duration of illness (*from the regression of Z*). One characteristic about the data set is that there are more patients with shorter formal education who have a longer duration of the disease (*from the correlation between W and A and also the distribution of W for each category of A*).

Example 2: Glucose control (cont.)

4. U (social externality), V (internality) and B (gender) appear to play little role in this example.

Further studies in Mainz and Düsseldorf later confirmed the importance of social and psychological factors in achieving desirable metabolic adjustment which were helpful in controlling glucose levels.

The End

Thank you!