Linearity Properties of Bayes Nets with Binary Variables

David Danks and Clark Glymour[1]

Abstract

It is "well known" that in linear models:
(1) testable constraints on the marginal distribution of observed variables distinguish certain cases in which an unobserved cause jointly influences several observed variables;
(2) the technique of "instrumental variables" sometimes permits an estimation of the influence of one variable on another even when the association between the variables may be confounded by unobserved common causes;
(3) the association (or conditional probability distribution of one variable given another) of two variables connected by a path or pair of paths with a single common vertex (a trek) can be computed directly from the parameter values associated with each edge in the trek;
(4) the association of two variables produced by multiple treks can be computed from the parameters associated with each trek; and
(5) the independence of two variables conditional on a third implies the corresponding independence of the sums of the variables over all units conditional on the sums over all units of each of the original conditioning variables.

These properties are exploited in search procedures. We show that (1) and (2) hold for all Bayes nets with binary variables. We further show that for Bayes nets parameterized as noisy *or* and noisy *and* gates, all of these properties save (4) hold.

1. Introduction.

Linear models have special advantages for model search and for the estimation of causal effects. Among them are those listed in the Abstract. Property (1) permits the detection of common causes via the "Tetrad Representation Theorem" and in combination with properties (3) and (4) is sufficient for the determination of latent structural relations from rather weak background assumptions (Spirtes, et al, 1993/2001; Shafer, et al., 1995). Property (2) provides a standard technique for estimating causal influence in econometrics, epidemiology and elsewhere. Property (5) is an essential assumption of many search methods that attempt to identify the causal structure of units from aggregated data, for example, several proposed methods of discovering genetic regulatory networks from measurements of mRNA concentrations.

In many models that are objects of automated search, for example networks for genetic regulation, it is assumed that the variables under study are binary. An important body of questions therefore concerns which of the properties of linear systems relevant to search hold for Bayes nets of binary variables, either in general or in an interesting class of special cases. Some results are known. For example the rules (3) and (4) for computing correlations in linear models are known to hold as well for singly trek-connected Bayes nets with binary variables, and

[1] Affiliations: Danks, Institute for Human and Machine Cognition, University of West Florida, and University of California, San Diego; Glymour, Carnegie Mellon University; Institute for Human and Machine Cognition, University of West Florida; and University of California, San Diego. Research for this paper was supported by grants to the IHMC and to Carnegie Mellon University by the National Aeronautics and Space Administration. We thank Richard Scheines and Peter Spirtes for valuable discussions concerning aggregation.

counterexamples are known for networks that have multiple treks between pairs of variables (Pearl, 1988). Techniques are known for using instrumental variables to bound causal effects in binary Bayes nets (Pearl, 2000). We supply a further result for Bayes nets of binary variables generally, and we discuss these properties for Bayes nets of binary variables parameterized as noisy *or* and noisy *and* gates, a parameterization of particular interest because of its use as a model of naïve human causal judgment (Cheng, 1997).

2. General Results

One technical notion and one Lemma will be used throughout this paper. A *trek* in a directed acyclic graph (DAG) is a directed path from one vertex to another, or a pair of directed paths terminating in two distinct vertices and intersecting in a single vertex. The unique vertex on any trek that has no edges in the trek directed into it is the *source* of the trek.

Lemma For any DAG with only binary variables, if A _||_ $C \mid B$, then $\rho(A, C) = \rho(A, B) * \rho(B, C)$.

Proof of Lemma:

We have the following general formula for the correlation of two binary variables:

$$\rho(X,Y) = \frac{P(X) * [P(Y \mid X) - P(Y)]}{\sigma_X \sigma_Y}$$

Therefore, we have the following two formulae:

$$\rho(A,C) = \frac{P(A) * [P(C \mid A) - P(C)]}{\sigma_A \sigma_C}$$

$$\rho(A,B) * \rho(B,C) = \frac{P(A) * [P(B \mid A) - P(B)] * P(B) * [P(C \mid B) - P(C)]}{\sigma_A \sigma_C \sigma_B{}^2}$$

Since $\sigma_B{}^2 = P(B) * [1 - P(B)]$, these two equations are equal iff:

$$[P(C \mid A) - P(C)] * [1 - P(B)] = [P(B \mid A) - P(B)] * [P(C \mid B) - P(C)] \quad (2.1)$$

Furthermore, we know that

$$[P(C \mid B) - P(C)] = [1 - P(B)] * [P(C \mid B) - P(C \mid \sim B)] \quad (2.2)$$

Plugging (2.2) into (2.1) and simplifying, the correlations are equal iff:

$$[P(C \mid A) - P(C)] = [P(B \mid A) - P(B)] * [P(C \mid B) - P(C \mid \sim B)] \quad (2.3)$$

Expanding the right-hand side, we have:

$$P(C \mid B)*P(B \mid A) - P(C \mid B)*P(B) - P(C \mid \sim B)*P(B \mid A) + P(C \mid \sim B)*P(B)$$
$$= P(C \mid B)*P(B \mid A) - P(C \mid \sim B) + P(C \mid \sim B)*P(\sim B \mid A) - P(C \,\&\, B) + P(C \mid \sim B) - P(C \,\&\, \sim B)$$

Since we have only binary variables, and since A _||_ $C \mid B$, the first and third terms combine to form $P(C \mid A)$. The second and fifth terms cancel. And the fourth and sixth combine to form $-P(C)$. Therefore, since the equality in (2.3) holds, $\rho(A, C) = \rho(A, B) * \rho(B, C)$.

2.1 A Tetrad Representation Theorem for Bayes Nets with Binary Variables

For systems of binary variables we formulate the Tetrad Representation Theorem (TRT) as follows:

> **Tetrad Representation Theorem for Binary Variables**: In a DAG G, there is a choke point between $\{I_1, I_2\}$ and $\{J_1, J_2\}$ if and only if $\rho_{11}\rho_{22} - \rho_{12}\rho_{21} = 0$ over a set of parameters of measure 1 (where ρ_{ij} is the correlation between I_i and J_j).

A variable C is a choke point between two sets **I** and **J** if and only if every trek between $I \in$ **I** and $J \in$ **J** includes C. The two proofs of the TRT for linear systems have both used the generalized trek rule, which holds for all linear systems. The generalized trek rule states that:

$$\rho(X_0, X_n) = \sum_{t \in T} \prod_{i=1}^{t_n} \rho(X_{i-1}, X_i)$$

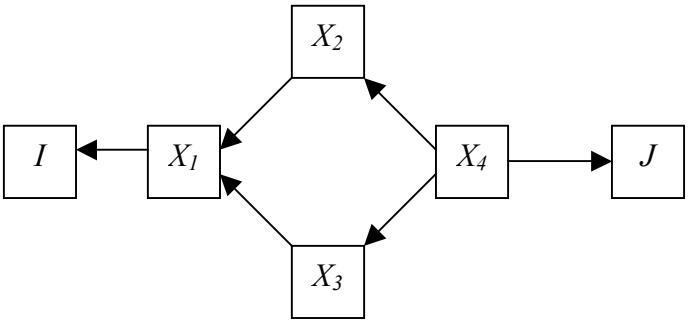
where T is the set of all and only the treks connecting X_0 and X_n, and t_n is the number of nodes on trek t. In fact, if the generalized trek rule holds for a system, then the TRT naturally follows, since the generalized trek rule is the only part of the TRT proof that depends on non-graphical properties. Unfortunately, the generalized trek rule does not hold in general for Bayes nets with binary variables, and we therefore adopt a modified strategy.

Let **T** be the set of treks from I to J, where X_i ranges over the variables on $T \in$ **T** (i.e., X_i ranges over every variable, including I and J, on all of the treks between I and J). We then define the following two sets:

U(**T**) = $\{<X_i, X_j> : \forall T \in \mathbf{T}(X_i \rightarrow X_j \in T)\}$

S(**T**) = $\{<X_k, X_l> : [<X_k, X_l> \notin \mathbf{U}(\mathbf{T})] \;\&\; [\forall T \in \mathbf{T}(X_k, X_l \in T)] \;\&\; \neg\exists X_i[[\forall T \in \mathbf{T}(X_i \in T)] \;\&\; [X_i$ is between X_k and $X_l]^2]\}$

These sets are actually quite easily described in English. **U**(**T**) consists of all of the pairs (i.e., directed edges) that appear in every trek from I to J. **S**(**T**) consists of the first and last vertex of each portion of the treks that do not overlap. Note that at least one of the two sets will be non-empty (if **T** is non-empty). Figure 2.1.1 provides **U**(**T**) and **S**(**T**) for a sample graph.



U(**T**) = $\{<X_1, I>, <X_4, J>\}$
S(**T**) = $\{<X_4, X_1>\}$

Figure 2.1.1: **U**(**T**) and **S**(**T**) for a sample graph

Note that we will omit the "(**T**)" when there is only one set of treks to consider. Given this notation, the following two theorems prove that a variant of the generalized trek rule holds for systems of binary variables.

Theorem 2.1.1:
Given the above notation, if **T** consists entirely of directed paths from I to J,

[2] Note that "between" is well-defined here, since X_k, X_i, and X_l are on every trek, and each trek must go through them in the same order.

$$\rho(I,J)=\left[\prod_{<X_i,X_j>\in \mathbf{U}}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathbf{S}}\rho(X_i,X_j)\right]$$

Theorem 2.1.2:
If **T** is the set of all treks between *I* and *J* (not necessarily all of which are directed paths), then

$$\rho(I,J)=\left[\prod_{<X_i,X_j>\in \mathbf{U}}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathbf{S}}\rho(X_i,X_j)\right]$$

Theorems 2.1.1 and 2.1.2 thus show that something like the generalized trek rule holds for systems of binary variables. It turns out that this variant is actually all we need for the TRT, as the following two theorems show.

Theorem 2.1.3:
If there is at least one choke point between $\{I_1, I_2\}$ and $\{J_1, J_2\}$, then:
$\rho(I_1, J_1) * \rho(I_2, J_2) - \rho(I_1, J_2) * \rho(I_2, J_1) = 0$

Theorem 2.1.4:
If there is no choke point between $\{I_1, I_2\}$ and $\{J_1, J_2\}$, then for a measure 1 set of parameters,
$\rho(I_1, J_1) * \rho(I_2, J_2) - \rho(I_1, J_2) * \rho(I_2, J_1) \neq 0$

Corollary 2.1.1: (Tetrad Representation Theorem for binary variables)
There is at least one choke point between $\{I_1, I_2\}$ and $\{J_1, J_2\}$ iff:
$\rho(I_1, J_1) * \rho(I_2, J_2) - \rho(I_1, J_2) * \rho(I_2, J_1) = 0$

Proof of Theorem 2.1.1:
We prove by induction on $|\mathbf{U}| + |\mathbf{S}| = n$.
Base case (n = 1): Since a set cannot have negative cardinality, the base case occurs when exactly one of the sets has exactly one element. If **U** has only one element, then it must be the element $<I, J>$, in which case the equation is trivially true. Similarly, if **S** has only one element, then it must be $<I, J>$, and so the equation is trivially true.
Induction step: For the induction step, we will assume that the sum of the cardinalities is $n - 1$. Then, we will show that the equation still holds when we add an element either to **U** or to **S**, where the element comes immediately before *J*. We can assume this without loss of generality, since we can "build" the set of directed paths iteratively.

Case 1: Assume that we add another element to **U**. That is, find the element $<X_j, J>$ in either **U** or **S**, replace that element by $<X_j, X_n>$ (in the same set), and add $<X_n, J>$ to **U**.[3] Then, since every path from *I* to *J* passes through X_n, $I \_||\_ J \mid X_n$. Therefore, by the Lemma, $\rho(I, J) = \rho(I, X_n) * \rho(X_n, J)$. Now consider the subgraph that excludes *J*. For this graph, $|\mathbf{U}| + |\mathbf{S}| = n - 1$, since the only element we remove is $<X_n, J>$ from **U**. Therefore, we know that the decomposition equation holds for $\rho(I, X_n)$, and so the decomposition equation holds for the full graph.

[3] And adjust the graph accordingly, by redirecting every edge into *J* into X_n, and adding $X_n \rightarrow J$ to the graph.

Case 2: Now consider adding an element to **S**. That is, we find the element $\langle X_j, J\rangle$ in either **U** or **S**, replace that element by $\langle X_j, X_n\rangle$ (in the same set), and add $\langle X_n, J\rangle$ to **U**.[4] As in case 1, we can use the Lemma to show that adding an element to **S** leads to $\rho(I, J) = \rho(I, X_n) * \rho(X_n, J)$, since $I \perp\!\!\!\perp J \mid X_n$. Therefore, since we assume the decomposition equation holds for the subgraph correlation (given by $\rho(I, X_n)$), the equation holds in this case.

Since the equation holds in both cases of the induction step, we have proven the theorem.

Proof of Theorem 2.1.2:

There are two different cases to consider, based on the number of distinct sources.

Case 1: Assume that there is exactly one source for all of the treks, call it W. Then we can use the Lemma to derive $\rho(I, W) * \rho(W, J) = \rho(I, J)$. The I sides of all of the treks form the set of directed paths from W to I, and the J sides form the set of directed paths from W to J. Using theorem 2.1.1, we can derive the above equation (since the total correlation is just the product of the sides, which are the products of the pieces).

Case 2: Assume that there are m distinct sources of the treks.

Claim: All of the sources must be between X_i and X_j, where $\langle X_i, X_j\rangle \in \mathbf{S}$.

Proof: Assume there are distinct $\langle X_i, X_j\rangle$, $\langle X_k, X_l\rangle \in \mathbf{S}$ such that some of the sources were between X_i and X_j, and some between X_k and X_l. Without loss of generality, assume that X_i is closest to I, and X_l is closest to J. Since there are sources between X_i and X_j, there must be edge(s) from X_j towards J. Similarly, there must be edge(s) from X_k towards I since there are sources between X_k and X_l. However, the edges out from X_j and the edges out from X_k must converge at some node X_r, or else we would have had $\langle X_i, X_l\rangle \in \mathbf{S}$. But this implies that X_r is a collider and so the paths do not form treks.

Since all of the sources fall between X_i and X_j, we need only show that $\rho(I, J) = \rho(I, X_i) * \rho(X_i, X_j) * \rho(X_j, J)$, since the trek pieces from X_i to I form the set of directed paths, and the pieces from X_j to J form the set of directed paths. However, this is straightforwardly shown by two applications of the Lemma (first to $\rho(I, J)$ in terms of X_j as the screener, and then to $\rho(I, X_j)$ with X_i as the screener).

Proof of Theorem 2.1.3:

Lemma 2.1 in Shafer, *et al*. (1995) tells us that, if there is more than one choke point, then all of the treks go through the choke points in the same order. Therefore, we can designate the choke point closest to $\{I_1, I_2\}$ as W, and without loss of generality, we can assume that W is between the source(s) and $\{I_1, I_2\}$ (and is possibly identical to one of these elements). W must occur in either **U** or **S** (or possibly both), since otherwise there would be a trek such that W did not occur in that trek, which would contradict W's status as a choke point. Therefore, if $\mathbf{T}_{i,j}$ is the set of treks between I_i and J_j, then we can divide $\mathbf{U}(\mathbf{T}_{i,j})$ into two disjoint sets: $\mathbf{U_I}(\mathbf{T}_{i,j})$ and $\mathbf{U_J}(\mathbf{T}_{i,j})$ corresponding to the elements of $\mathbf{U}(\mathbf{T}_{i,j})$ on the I_i side, and the J_j side respectively. We can similarly split $\mathbf{S}(\mathbf{T}_{i,j})$ into two disjoint sets. Now, consider the two terms in the theorem's equation (using theorem 2.1.2 to expand them):

[4] And adjust the graph accordingly, by redirecting every edge into J into X_n, and adding arbitrarily many children for X_n, with arbitrary connections among the children and into J.

$$\rho(I_1,J_1)*\rho(I_2,J_2)=\left[\prod_{<X_i,X_j>\in \mathrm{U_I}(\mathrm{T}_{1.1})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{U_J}(\mathrm{T}_{1.1})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{S_I}(\mathrm{T}_{1.1})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{S_{JI}}(\mathrm{T}_{1.1})}\rho(X_i,X_j)\right]$$
$$*\left[\prod_{<X_i,X_j>\in \mathrm{U_I}(\mathrm{T}_{2.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{U_J}(\mathrm{T}_{2.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{S_I}(\mathrm{T}_{2.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{S_J}(\mathrm{T}_{2.2})}\rho(X_i,X_j)\right]$$

and

$$\rho(I_1,J_2)*\rho(I_1,J_2)=\left[\prod_{<X_i,X_j>\in \mathrm{U_I}(\mathrm{T}_{1.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{U_J}(\mathrm{T}_{1.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{S_I}(\mathrm{T}_{1.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{S_{JI}}(\mathrm{T}_{1.2})}\rho(X_i,X_j)\right]$$
$$*\left[\prod_{<X_i,X_j>\in \mathrm{U_I}(\mathrm{T}_{1.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{U_J}(\mathrm{T}_{1.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{S_I}(\mathrm{T}_{1.2})}\rho(X_i,X_j)\right]*\left[\prod_{<X_i,X_j>\in \mathrm{S_J}(\mathrm{T}_{1.2})}\rho(X_i,X_j)\right]$$

However, since *W* is a choke point, every trek between $\{I_1, I_2\}$ and $\{J_1, J_2\}$must pass through *W*. Therefore, $\mathbf{U_I}(\mathbf{T}_{\mathrm{i.j}})$ and $\mathbf{S_I}(\mathbf{T}_{\mathrm{i.j}})$ are dependent only on the value of *i*. Similarly, $\mathbf{U_J}(\mathbf{T}_{\mathrm{i.j}})$ and $\mathbf{S_J}(\mathbf{T}_{\mathrm{i.j}})$ are dependent only on *j*. Therefore, we have the following equalities:

$\mathbf{U_I}(\mathbf{T}_{1.1}) = \mathbf{U_I}(\mathbf{T}_{1.2})$
$\mathbf{U_J}(\mathbf{T}_{1.1}) = \mathbf{U_J}(\mathbf{T}_{2.1})$
$\mathbf{S_I}(\mathbf{T}_{1.1}) = \mathbf{S_I}(\mathbf{T}_{1.2})$
$\mathbf{S_J}(\mathbf{T}_{1.1}) = \mathbf{S_J}(\mathbf{T}_{2.1})$

$\mathbf{U_I}(\mathbf{T}_{2.1}) = \mathbf{U_I}(\mathbf{T}_{2.2})$
$\mathbf{U_J}(\mathbf{T}_{1.2}) = \mathbf{U_J}(\mathbf{T}_{2.2})$
$\mathbf{S_I}(\mathbf{T}_{2.1}) = \mathbf{S_I}(\mathbf{T}_{2.2})$
$\mathbf{S_J}(\mathbf{T}_{1.2}) = \mathbf{S_J}(\mathbf{T}_{2.2})$

Therefore, every term in the first product has an equal term in the second product, and therefore the difference between the products must be zero.

Proof of Theorem 2.1.4:
If there is no choke point, then $\mathbf{U}(\mathbf{T}_{1.1}) \cup \mathbf{U}(\mathbf{T}_{2.2}) \cup \mathbf{S}(\mathbf{T}_{1.1}) \cup \mathbf{S}(\mathbf{T}_{2.2}) \neq \mathbf{U}(\mathbf{T}_{1.2}) \cup \mathbf{U}(\mathbf{T}_{2.1}) \cup \mathbf{S}(\mathbf{T}_{1.2}) \cup \mathbf{S}(\mathbf{T}_{2.1})$, since if the unions were equal, then we would have a choke point. Therefore, the tetrad difference (which is defined by these unions) is zero only on a set of measure zero in parameter space (where the measure is absolutely continuous with the uniform measure on [0,1]).

2.2 Aggregation
In the case of gene expression, we typically take data on several variables, but we actually receive data summed or averaged over many individuals at once. The aim of inquiry is a Bayes net representing the conditional independence and causal relations among the properties of individual units. So we pose the question:

> If $X \perp\!\!\!\perp Z \mid Y$ for each individual (in a large, i.i.d. sample of size *N*), is $\sum_{i=1}^{N} X_i \perp\!\!\!\perp \sum_{i=1}^{N} Z_i \mid \sum_{i=1}^{N} Y_i$, and conversely?

We abbreviate the conditional independence of the summed variables as ΣX _||_ ΣZ | ΣY. We argue informally that for large N almost certainly the conditional independence above holds for the summed variables if and only if it holds for the individual variables. For large N, the distribution of ΣX, ΣZ ,ΣY is approximately normal by the Central Limit Theorem, and, to that approximation, a conditional independence holds if and only if the corresponding conditional covariance or partial correlation vanishes. We have the following formula for the conditional covariance of ΣX and ΣZ:

$Cov(\Sigma X, \Sigma Z \mid \Sigma Y) = E(\Sigma X \,\&\, \Sigma Z \mid \Sigma Y) - E(\Sigma X \mid \Sigma Y) * E(\Sigma Z \mid \Sigma Y)$

The first term factors into: $E(\Sigma X \mid \Sigma Y) * E(\Sigma Z \mid \Sigma X \,\&\, \Sigma Y)$. Therefore, the covariation (and so also the correlation) is zero if and only if $E(\Sigma Z \mid \Sigma X \,\&\, \Sigma Y) = E(\Sigma Z \mid \Sigma Y)$.

We can express the expected value of ΣZ as functions of the probabilities of X and of Y as:

$$E(\Sigma Z) = N * P(Z = 1) = N * [P(Z = 1 \mid X = 1, Y = 1) * P(Y = 1 \mid X = 1) * P(X = 1) \qquad (2.2.1)$$

$$+ P(Z = 1 \mid X = 1, Y = 0) * P(Y = 0 \mid X = 1) * P(X = 1)$$

$$+ P(Z = 1 \mid X = 0, Y = 1) * P(Y = 1 \mid X = 0) * P(X = 0)$$

$$+ P(Z = 1 \mid X = 0, Y = 0) * P(Y = 0 \mid X = 0) * P(X = 0)]$$

and

$$E(\Sigma Z) = N * P(Z = 1) = N * [P(Z = 1 \mid Y = 1) * P(Y = 1) + P(Z = 1 \mid Y = 0) * P(Y = 0)]. \qquad (2.2.2)$$

Conditioning (2.2.2) on $\Sigma Y = N_Y$ results in

$$E(\Sigma Z \mid \Sigma Y = N_Y) = N [P(Z = 1 \mid Y = 1) * N_Y/N + P(Z = 1 \mid Y = 0) * (1- (N_Y/N))] \qquad (2.2.3)$$

Conditioning (2.2.1) on $\Sigma Y = N_Y$, $\Sigma X = N_X$ in the analogous way, rearranging and using the fact that $P(Z \mid X, Y) = P(Z \mid Y)$ also results in equation (2.2.3). Hence within the approximations noted, almost certainly X _||_ $Z \mid Y$ if and only if $\sum_{i=1}^{N} X_i$ _||_ $\sum_{i=1}^{N} Z_i \mid \sum_{i=1}^{N} Y_i$.

3. Bayes Nets of Noisy-Or/ Noisy-And Gates

Consider an arbitrary directed acyclic graph (DAG) whose vertices are binary variables taking values in {0,1}. We say a model is a Noisy-OR and -AND gate model, or more briefly a Cheng model if, for each variable X, the set of parents of X, P(X), can be partitioned into two sets, GEN(X) and PRE(X) such that:

$$X = [U_X + \Sigma_{K \in GEN(X)} q_{KX} K] [\Pi_{L \in PRE(X)} (1 - q_{LX} L)]$$

where all sums are Boolean, and U_X is distributed independently of all variables other than X and the descendants of X, and q_{KX} and q_{LX} are separate parameters for each variable K and L, respectively, and all such parameters are jointly independent of each other and of all variables in the network . Intuitively, the variables in GEN(X) and U are generative or positive causes of X, while the variables in PRE(X) prevent X (taking X = 1 as the occurrence of X or the marked case.) Again, intuitively, the probability that q_{KX} = 1 is the probability that, given that K = 1, K causes X = 1, and the probability that q_{LX} = 1 is the probability that, given that L = 1, L prevents X = 1 (Cheng, 1997). Sources of variation not represented in the network are required to be generative, since otherwise none of the parameters of the model can be estimated from

observational data (Glymour, 1998). Such models have been applied in electrical engineering and developed as models of human judgment of non-interactive causal relations. Our concern is to find the linear analogies valid in such models.

3.1 Instrumental Variable Calculations.
Instrumental variable models have the graphical structure shown in figure 3.1.1.

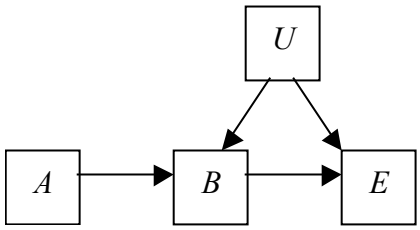


Figure 3.1.1: Instrumental variable graph

where U is unobserved. The object is to estimate the conditional probability distribution of E on values of B determined by an intervention that randomizes B. Suppose all causes are generative, so that

$E = q_{be}B \oplus q_{ue}U$ and $B = q_{ab}A \oplus q_{ub}U$ (3.1.1)

where $\oplus$ is Boolean addition. Following Spirtes, et al. (1993/2001), and Pearl (2000), what must be estimated is

$P_{B=0}(E = 1) = P(q_{ue}U = 1)$ and $P_{B=1}(E = 1) = P(q_{be} \oplus q_{ue}U = 1)$. (3.*)

It is easily verified that

$P(q_{ab} = 1) = [P(B = 1 \mid A = 1) - P(B = 1 \mid A = 0)] / [1 - P(B = 1 \mid A = 0)$. (3.1.2)

(The derivation is in Cheng, 1997). Substituting and factoring in (3.1.1):

$E = q_{be}\, q_{ab}A \oplus (q_{be}q_{ub} \oplus q_{ue})U$, (3.1.3)

It follows by an analogous argument to that for (3.1.2) that

$P(q_{be} = 1) * P(q_{ab} = 1) = [P(E = 1 \mid A = 1) - P(E = 1 \mid A = 0)] / [1 - P(E = 1 \mid A = 0)]$ (3.1.4)

The ratio of (3.1.3) to (3.1.2) gives $P(q_{be} = 1)$. The r.h.s. of the first equation in 3.* is obtained by

$P(q_{ue}U = 1) = P(q_{ue}U = 1 \mid B = 1) * P(B = 1) + P(q_{ue}U = 1 \mid B = 0) * P(B = 0)$ (3.1.5)

which after some algebra reduces to a formula in observed probabilities:

$P(q_{ue}U = 1) = [P(E = 1 \mid B = 1) - P(q_{be} = 1)] * P(B = 1) / [1 - P(q_{be} = 1)]$ (3.1.6)
$+ P(E = 1 \mid B = 0) * P(B = 0)$

Hence the r.h.s. of each equation in 3.* can be estimated. Analogous results are obtained with similar algebra when the influence of B is preventive and A is generative.

3.2 Trek Rules
In this section we assume the following typical structure:

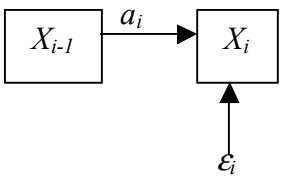


Figure 3.2.1: Typical graphical unit structure
where the response functions (and associated probabilities) are:
Noisy-OR gate:
$X_i = a_i X_{i-1} \oplus \varepsilon_i$
$P(X_i) = P(a_i) * P(X_{i-1}) + P(\varepsilon_i) - P(a_i) * P(X_{i-1}) * P(\varepsilon_i)$

Noisy-AND gate:
$X_i = \varepsilon_i \bullet (1 - a_i X_{i-1})$
$P(X_i) = P(\varepsilon_i) * [1 - P(a_i) * P(X_{i-1})]$

Theorem 3.2.1:
If a directed path of length $n \geq 1$ composed of noisy-OR and noisy-AND gates (in any combination and order) is the only trek between X_0 and X_n, then:

$$\rho(X_0, X_n) = \left[\prod_{i=1}^{n} P(a_i) * g(i)\right] * \frac{\sqrt{P(X_0) * [1 - P(X_0)]}}{\sqrt{P(X_n) * [1 - P(X_n)]}},$$

where $g(i) = \begin{cases} [1 - P(\varepsilon_i)], \text{ if the } i \text{ - th gate is noisy - OR; or} \\ -P(\varepsilon_i), \text{ if the } i \text{ - th gate is noisy - AND.} \end{cases}$

Theorem 3.2.2:
If a directed path of length $n \geq 1$ composed of noisy-OR and noisy-AND gates (in any combination and order) is the only trek between X_0 and X_n, then:

$$\rho(X_0, X_n) = \prod_{i=1}^{n} \rho(X_{i-1}, X_i)$$

Theorem 3.2.3:
If a trek of length $n \geq 1$ composed of noisy-OR and noisy-AND gates (in any combination and order) with X_k as the source of the trek ($n \geq k \geq 0$) is the only trek between X_0 and X_n, then:

$$\rho(X_0, X_n) = \prod_{i=1}^{n} \rho(X_{i-1}, X_i)$$

Corollary 3.2.1: (follows directly from theorems 3.2.3 and 3.2.1)
If a trek of length $n \geq 1$ composed of noisy-OR and noisy-AND gates (in any combination and order) with source X_k ($n \geq k \geq 0$) is the only trek between X_0 and X_n, then:

$$\rho(X_0, X_n) = \left[\prod_{i=0}^{k-1} P(a_i) * g(i)\right] * \left[\prod_{i=k+1}^{n} P(a_i) * g(i)\right] * \frac{P(X_k) * [1 - P(X_k)]}{\sqrt{P(X_0) * [1 - P(X_0)]} * \sqrt{P(X_n) * [1 - P(X_n)]}},$$

where $g(i) = \begin{cases} [1 - P(\varepsilon_i)], \text{ if the } i \text{ - th gate is noisy - OR; or} \\ -P(\varepsilon_i), \text{ if the } i \text{ - th gate is noisy - AND.} \end{cases}$

Proof of Theorem 3.2.1: (by induction)

Base case (n = 1):
Consider the case in which we have just one noisy-OR gate. Using the above formula for a noisy-OR gate, we can straightforwardly derive (after some algebraic manipulation) the following covariance and correlation:

$Cov(X_0, X_1) = P(a_1) * [1 - \varepsilon_1] * P(X_0) * [1 - P(X_0)]$

$$\rho(X_0, X_1) = P(a_1) * [1 - P(\varepsilon_1)] * \frac{\sqrt{P(X_0) * [1 - P(X_0)]}}{\sqrt{P(X_1) * [1 - P(X_1)]}}$$

Now consider the case in which we have just one noisy-AND gate. Using the above formula for a noisy-AND gate, we can derive (after some algebra) the following covariance and correlation:

$Cov(X_0, X_1) = -P(a_1) * P(\varepsilon_1) * P(X_0) * [1 - P(X_0)]$

$$\rho(X_0, X_1) = -P(a_1) * P(\varepsilon_1) * \frac{\sqrt{P(X_0) * [1 - P(X_0)]}}{\sqrt{P(X_1) * [1 - P(X_1)]}}$$

Induction step:
Assume the theorem holds for $n - 1$, and now we will show that it holds for n. Consider first the case in which we add a noisy-OR gate from X_{n-1} to X_n. Since there is only one trek between X_0 and X_n, and it passes through X_{n-1}, we know that $X_0 \perp\!\!\!\perp X_n \mid X_{n-1}$. Therefore,

$\rho(X_0, X_n) = \rho(X_0, X_{n-1}) = \rho(X_{n-1}, X_n)$. (3.2.1)

By the same reasoning as in the base step, we know that:

$$\rho(X_{n-1}, X_n) = P(a_n) * [1 - P(\varepsilon_n)] * \frac{\sqrt{P(X_{n-1}) * [1 - P(X_{n-1})]}}{\sqrt{P(X_n) * [1 - P(X_n)]}}. \quad (3.2.2)$$

Since, the theorem was assumed to hold for $n - 1$, we can substitute the equation for the theorem into (3.2.1) for $\rho(X_0, X_{n-1})$, and use (3.2.2) for $\rho(X_{n-1}, X_n)$. After we roll $P(a_n)$ and $[1 - P(\varepsilon_n)]$ into the Π term (from the theorem equation), we have:

$$\rho(X_0, X_n) = \left[\prod_{i=1}^{n} P(a_i) * g(i)\right] * \frac{\sqrt{P(X_0) * [1 - P(X_0)]}}{\sqrt{P(X_{n-1}) * [1 - P(X_{n-1})]}} * \frac{\sqrt{P(X_{n-1}) * [1 - P(X_{n-1})]}}{\sqrt{P(X_n) * [1 - P(X_n)]}}$$

Hence, after canceling, the formula still holds if we add a noisy-OR gate to the end of the path.
Now consider adding a noisy-AND gate from X_{n-1} to X_n. In that case, we can use exactly the same reasoning as in the noisy-OR case, except that we are incorporating different terms into the theorem equation for $n - 1$. Therefore, the formula holds if we add a noisy-AND gate to the end of the path. Since the theorem holds for both base case possibilities and both induction possibilities, the theorem holds for all Cheng models.

Proof of Theorem 3.2.2:
By Theorem 3.2.1, we know that:

$$\rho(X_{i-1}, X_i) = P(a_i) * g(i) * \frac{\sqrt{P(X_{i-1}) * [1 - P(X_{i-1})]}}{\sqrt{P(X_i) * [1 - P(X_i)]}}$$

Therefore, if we multiply together the correlations for each of the gates, we have

$$\prod_{i=1}^{n} \rho(X_{i-1}, X_i) = \left[\prod_{i=1}^{n} P(a_i) * g(i)\right] * \left[\prod_{i=1}^{n} \frac{\sqrt{P(X_{i-1}) * [1 - P(X_{i-1})]}}{\sqrt{P(X_i) * [1 - P(X_i)]}}\right].$$

Since all of the terms in the second Π term cancel out except for the initial numerator and final denominator, the right-hand side reduces to $\rho(X_0, X_n)$, as given in Theorem 3.2.1.

Proof of Theorem 3.2.3:
By Theorem 3.2.2, if suffices to show that $\rho(X_0, X_n) = \rho(X_0, X_k) * \rho(X_k, X_n)$, since the decomposition holds for each of the directed paths ($X_0 \leftarrow X_k$ and $X_k \rightarrow X_n$). Since $\rho(Y, Y) = 1$, this condition trivially holds for $k = 0$ or n. Therefore, we will assume that $0 < k < n$.
Now, consider the covariance between X_0 and X_n. After much algebra, we get the following formula (independently of the ordering of noisy-OR and noisy-AND gates on the trek):
$Cov(X_0, X_n) = P(X_k) * [1 - P(X_k)] * [P(X_0 \mid X_k) - P(X_0 \mid \sim X_k)] * [P(X_n \mid X_k) - P(X_n \mid X_k)]$
Therefore, we have the following formula for the correlation:

$$\rho(X_0, X_n) = P(X_k) * [1 - P(X_k)] * \frac{[P(X_0 \mid X_k) - P(X_0 \mid \sim X_k)] * [P(X_n \mid X_k) - P(X_n \mid \sim X_k)]}{\sqrt{P(X_0) * [1 - P(X_0)]} * \sqrt{P(X_n) * [1 - P(X_n)]}} \quad (3.2.3)$$

Now, we also have the following formulae for the correlations between X_0 and X_k, and X_k and X_n:

$$\rho(X_0, X_k) = P(X_k) * \frac{P(X_0 \mid X_k) - P(X_0)}{\sqrt{P(X_0) * [1 - P(X_0)]} * \sqrt{P(X_k) * [1 - P(X_k)]}}$$

$$\rho(X_k, X_n) = P(X_k) * \frac{P(X_n \mid X_k) - P(X_n)}{\sqrt{P(X_n) * [1 - P(X_n)]} * \sqrt{P(X_k) * [1 - P(X_k)]}}$$

If we multiply these two correlations together, we have:

$$\rho(X_0, X_k) * \rho(X_k, X_n) = P^2(X_k) * \frac{[P(X_0 \mid X_k) - P(X_0)] * [P(X_n \mid X_k) - P(X_n)]}{P(X_k) * [1 - P(X_k)] * \sqrt{P(X_0) * [1 - P(X_0)]} * \sqrt{P(X_n) * [1 - P(X_n)]}} \quad (3.2.4)$$

Consider the terms in the numerator of the fraction. We can perform the following transformation on the first term (and similarly for the second):
$P(X_0 \mid X_k) - P(X_0) = P(X_0 \mid X_k) - P(X_0 \,\&\, X_k) - P(X_0 \,\&\, \sim X_k) =$
$P(X_0 \mid X_k) - P(X_0 \mid X_k) * P(X_k) - P(X_0 \mid \sim X_k) * [1 - P(X_k)] =$
$[1 - P(X_k)] * [P(X_0 \mid X_k) - P(X_0 \mid \sim X_k)]$.
Substituting this transformation (and the analogous one for $[P(X_n \mid X_k) - P(X_n)]$) back into equation (3.2.4) and canceling terms, we have:

$$\rho(X_0, X_k) * \rho(X_k, X_n) = P(X_k) * [1 - P(X_k)] * \frac{[P(X_0 \mid X_k) - P(X_0 \mid \sim X_k)] * [P(X_n \mid X_k) - P(X_n \mid \sim X_k)]}{\sqrt{P(X_0) * [1 - P(X_0)]} * \sqrt{P(X_n) * [1 - P(X_n)]}} \quad (3.2.5)$$

These equations show that $\rho(X_0, X_n) = \rho(X_0, X_k) * \rho(X_k, X_n)$, which is what we needed to establish the theorem.

4. Counterexamples

The trek rules for singly connected Cheng models do not generalize. Further, Cheng models make it easy to show that the aggregation invariance that holds in all Bayes nets with binary variables when conditioning on a single variable does not hold when conditioning on multiple variables.

4.1 Failure of the Trek Rule

The Trek Rule does not generalize to multiply trek connected variables in noisy-AND/OR networks. That is: If T is the set of all and only the treks between X_0 and X_n, and $|T| > 1$, then it is not necessarily the case that: $\rho(X_0, X_n) = \sum_{t \in T} \rho_t(X_0, X_n)$ (where $\rho_t(X_0, X_n)$ is the correlation between X_0 and X_n if trek t were the only trek).

Consider the following graph composed solely of noisy-AND gates (the a_i and ε_I terms are left out for simplicity):

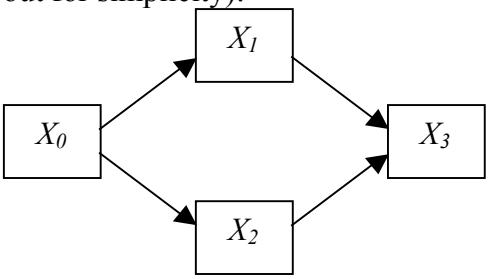


Figure 4.1.1: Counterexample to trek rule

So, the equations for the dependent variables are:

$X_1 = [1 - a_1X_0] \bullet \varepsilon_1$

$X_2 = [1 - a_2X_0] \bullet \varepsilon_2$

$X_3 = [1 - a_{31}X_1] \bullet [1 - a_{32}X_2] \bullet \varepsilon_3$

We only need to determine $\rho(X_0, X_3)$ directly, since we can use Theorem 3.2.1 to compute the correlations along each trek (since each is a directed path). When we substitute the equations for X_1 and X_2 into the equation for X_3, we get:

$P(X_3) = [1 - P(a_{31})*P(\varepsilon_1)*[1 - P(a_1)*P(X_0)]] * [1 - P(a_{32})*P(\varepsilon_2)*[1 - P(a_2)*P(X_0)]] * P(\varepsilon_3)$

After lots of algebra, we can then derive the following covariance and correlation:

$Cov(X_0, X_3) = P(\varepsilon_3) * P(X_0) * [1 - P(X_0)] * [P(\varepsilon_1)*P(a_1)*P(a_{31}) + P(\varepsilon_2)*P(a_2)*P(a_{32}) -$
$P(\varepsilon_1)*P(\varepsilon_2)*P(a_1)*P(a_{31})*P(a_{32}) - P(\varepsilon_1)*P(\varepsilon_2)*P(a_2)*P(a_{31})*P(a_{32}) +$
$P(\varepsilon_1)*P(\varepsilon_2)*P(a_1)*P(a_2)*P(a_{31})*P(a_{32})*P(X_0)]$

$$\rho(X_0, X_3) = P(\varepsilon_3) * Q * \frac{\sqrt{P(X_0) * [1 - P(X_0)]}}{\sqrt{P(X_3) * [1 - P(X_3)]}}, \quad (4.1.1)$$

where $Q = [P(\varepsilon_1)*P(a_1)*P(a_{31}) + P(\varepsilon_2)*P(a_2)*P(a_{32}) - P(\varepsilon_1)*P(\varepsilon_2)*P(a_1)*P(a_{31})*P(a_{32}) -$
$P(\varepsilon_1)*P(\varepsilon_2)*P(a_2)*P(a_{31})*P(a_{32}) + P(\varepsilon_1)*P(\varepsilon_2)*P(a_1)*P(a_2)*P(a_{31})*P(a_{32})*P(X_0)]$

Using the Lemma to compute the correlations along each individual trek, we have

$$\rho_1(X_0, X_3) + \rho_2(X_0, X_3) = P(\varepsilon_3) * W * \frac{\sqrt{P(X_0) * [1 - P(X_0)]}}{\sqrt{P(X_3) * [1 - P(X_3)]}}, \quad (4.1.2)$$

where $W = [P(\varepsilon_1)*P(a_1)*P(a_{31}) + P(\varepsilon_2)*P(a_2)*P(a_{32})]$.

Therefore, when we compare equations (4.1.1) and (4.1.2), we can see that the generalized trek rule will hold for this case if and only if $Q = W$, which is true if and only if:

$P(\varepsilon_1)*P(\varepsilon_2)*P(a_{31})*P(a_{32}) * [P(a_1) + P(a_2) - P(a_1)*P(a_2)*P(X_0)] = 0$

Since we assume that all of the probabilities are non-extremal, this equality cannot possibly be satisfied (since $P(a_1)*P(a_2)*P(X_0) < P(a_1)$ and $P(a_1)*P(a_2)*P(X_0) < P(a_2)$). Therefore, the generalized trek rule does not hold for all graphs composed of noisy-OR and noisy-AND gates.

4.2 Failure of aggregation

Aggregation does not generalize. That is: If X is an ancestor of Z, and $Y_1, \ldots, Y_n$ ($n > 1$) are the parents of Z, then it is not necessarily the case that $\rho(\Sigma X, \Sigma Z \mid \Sigma Y_1, \ldots, \Sigma Y_n) = 0$.

Proof by counterexample. Consider the following graph:

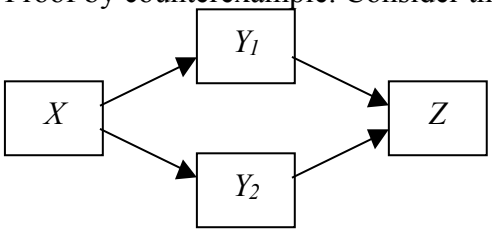


Figure 4.2.1: Counterexample to aggregation

We have the following formula:

$Cov(\Sigma X, \Sigma Z \mid \Sigma Y_1, \Sigma Y_2) = E(\Sigma X \,\&\, \Sigma Z \mid \Sigma Y_1, \Sigma Y_2) - E(\Sigma X \mid \Sigma Y_1, \Sigma Y_2) * E(\Sigma Z \mid \Sigma Y_1, \Sigma Y_2)$

The first term in the formula factors into: $E(\Sigma X \mid \Sigma Y_1, \Sigma Y_2) * E(\Sigma Z \mid \Sigma X, \Sigma Y_1, \Sigma Y_2)$. Therefore, the covariance (and hence the correlation) equals zero just in case:

$E(\Sigma Z \mid \Sigma X, \Sigma Y_1, \Sigma Y_2) = E(\Sigma Z \mid \Sigma Y_1, \Sigma Y_2)$.

Now consider the left-hand side of the equation. If we assume that there are N individuals in the summation, that the summations are given by N_X, N_{Y1}, and N_{Y2}, and that all of the connections are noisy-AND gates, then we have:

$E(\Sigma Z \mid \Sigma X, \Sigma Y_1, \Sigma Y_2) = N * [(1 - P(a_{Y1})*P(Y_1)) * (1 - P(a_{Y2})*P(Y_2)) * P(\varepsilon_Z)] =$

$P(\varepsilon_Z) * [N - P(a_{Y1})*N_{Y1} - P(a_{Y2})*N_{Y2} - P(a_{Y1})*P(a_{Y1})*N*P(Y_1 \,\&\, Y_2)]$.

Now, $P(Y_1 \,\&\, Y_2)$ is a function of X, and so we can reduce $E(\Sigma Z \mid \Sigma X, \Sigma Y_1, \Sigma Y_2)$ to a formula having only known values (including N_X, N_{Y1}, and N_{Y2}).

Consider a similar operation on $E(\Sigma Z \mid \Sigma Y_1, \Sigma Y_2)$. In this case, our simplification must stop with a $P(Y_1 \,\&\, Y_2)$ term still in the formula. That is, we cannot determine whether, in fact, these two equations are equal. It depends on the probability of the joint occurrence of Y_1 and Y_2, which we do not know.

5. Comments

The counterexample to aggregation invariance argues that, except in special cases, attempts to infer an underlying structure among binary variables from aggregated data ought to be suspect. On the positive side, the explicit characterization of trek rules and the applicability of instrumental variables to noisy-or/noisy-and gate models may be of use both in the design of psychological experiments and in data analysis where such parameterizations are plausible.

The most important positive result in this paper is surely the extension of the Tetrad Representation Theorem to systems of binary variables. Combined with the absence of conditional independence relations among the measured variables (as in Spirtes, et al, 1993/2001) it provides a necessary and sufficient condition (assuming "faithfulness" – see Spirtes, et al., 1993/2001) for four measured variables in a structure of binary variables to have a single unmeasured common cause. The applicability of the result bears comparison with recent statistical work (Junker and Ellis, 1997) that provides a sufficient condition (implicitly with the same faithfulness assumption) for a single common cause given an infinite sequence of measured variables. An interesting open question concerns whether results similar to the TRT can be

obtained for models now popular in psychometrics in which the probability distribution on measured binary variables is a function of a continuous latent variable.

References

Cheng, Patricia W. 1997. "From Covariation to Causation: A Causal Power Theory." *Psychological Review*, 104: 367-405.

Glymour, Clark. 1998. "Learning Causes: Psychological Explanations of Causal Explanation." *Minds and Machines*, 8: 39-60.

Junker, B.W. and J.L. Ellis. 1997. "A Characterization of Monotone Unidimensional Latent Variable Models." *Annals of Statistics*, 25, 1327-1343.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann Publishers, Inc.

Pearl, Judea.. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge Univeristy Press.

Shafer, Glenn, Alexander Kogan, and Peter Spirtes. 1995. "A Generalization of the Tetrad Representation Theorem." *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*. Fort Lauderdale, Fl. pp. 476-487.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993/2001. *Causation, Prediction, and Search*, Springer. 2nd edition, 2001, Cambridge, Mass.: AAAI Press & The MIT Press.