Testing Regression Models With Fewer Regressors

Judea Pearl and Peyman Meshkat
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

1 INTRODUCTION

Let M be a recursive regression model, that is, a set of regression equations for ordered variables $Y_1, \ldots, Y_n$, where each variable Y_i is regressed on $Y_1, \ldots, Y_{i-1}$ and where a set C_0 of coefficients are set to zero in advance, while others remain "free". We wish to test if this model fits a given set of data, namely, if the data support the starting assumption of setting the coefficients in C_0 to zero.

A straightforward way of testing this assumption would be to actually perform the regressions and test if all members of C_0 are indeed zero. This, however, may require high order regressions, especially for large values of i, and the question arises whether we can run a different set of regressions, each with a smaller number of variables, and still test the original model M.

We show that the answer to this question is affirmative, and that the following procedure accomplishes the task:

Graphical Procedure (GP)

1. Construct the directed acyclic graph of M, in which nodes represent variables and arrows represent non-zero coefficients,

2. for each pair (i, j) of non adjacent variables, $i > j$, find a set Z_{ij} of nodes such that:

   2.1 Z_{ij} d-separates i from j in the graph, and

   2.2 Z_{ij} contains only nodes that are closer to i than j is,

3. test the hypothesis $r_{ij \cdot Z_{ij}} = 0$ for each $i > j$.

We show in this paper that if the regression coefficients $r_{ij \cdot Z_{ij}}$, chosen according to the procedure above, vanish, then all members of C_0 and only those members, must vanish as well.

A special, well known choice for Z_{ij} is the set of parent nodes of i, namely, $Z_{ij} = pa_i$, which yields the standard test used in validating Bayesian networks. However, when the size of pa_i is large, it might be advantageous to use non-parental separators, as shown below.

2 AN EXAMPLE

Consider the set M of regression equations

$$\begin{array}{rcl} X_2 & = & a_{21}X_1 + \epsilon_2 \\ X_3 & = & a_{31}X_1 + 0X_2 + \epsilon_3 \\ X_4 & = & 0X_1 + 0X_2 + a_{43}X_3 + \epsilon_4 \\ X_5 & = & 0X_1 + 0X_2 + a_{53}X_3 + a_{54}X_4 + \epsilon_5 \end{array} \tag{1}$$

The assumptions embedded in this regression model are represented by the zero coefficients, and correspond to the vanishing of the following set of partial correlations:

$$C_0 = \{\rho_{32\cdot1} = 0, \rho_{41\cdot23} = 0, \rho_{42\cdot13} = 0, \rho_{51\cdot234} = 0, \rho_{52\cdot134} = 0\} \tag{2}$$

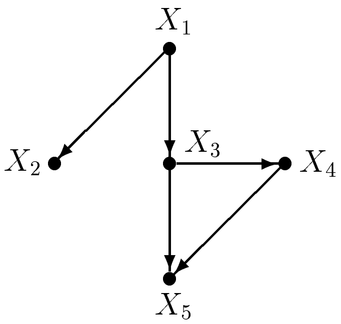

Figure 1:

The graph representing M is shown in Figure 1, from which the separating sets for each nonadjacent pair of nodes can easily be identified. One choice of separators leads to the following set of partial regression coefficients that need be tested:

$$B = \{\rho_{32\cdot1} = 0, \rho_{41\cdot3} = 0, \rho_{42\cdot1} = 0, \rho_{51\cdot3} = 0, \rho_{52\cdot1} = 0\} \tag{3}$$

We see that B and C_0 contain the same number of elements, yet all elements of B have at most three indices and will require, therefore, only two regressors in their corresponding tests. (The parental scheme would require three regressors for testing $\rho_{52\cdot34} = 0$.)

The paper proves that, in general, the elements of C_0 are zero if and only if the elements of B are zero, where B is any set that meets conditions 2.1 and 2.2. We further extend this result to nonrecursive regression and outline economical ways of testing both directed and undirected graph models with Gaussian variables.

3 THEORETICAL BACKGROUND

Definition 1 (basis) *Let S be a set of zero partial correlations. A basis B for S is a set of zero partial correlations that (1) implies (using the laws of probability) every element of S, and (2) no proper subset of B sustains such implication.*

The target of our investigation is a set S which corresponds to C_0, the zero elements in a recursive system of regression equations. Such a set is characterized by a distinct feature: the indices in every element of the i-th row are precisely the set of predecessors of i. It is well known, that such sets of zero partial correlations can be represented by separation in a directed acyclic graph $D(M)$, such as the one constructed by procedure GP in Section 1. This follows from the fact that in any DAG D, the parents of node i separate i from all its nondescendants in D [Pearl, 1988, pp. 119–120]. Moreover, the DAG $D(M)$ also enables us to identify a basis for C_0, choosing the set of parents of node i as the conditioning variables Z in each member $r_{ij \cdot Z}$ of B. This follows from the d-separation[1] theorem [Verma and Pearl, 1988], which states that every separation conditions in $G(M)$ corresponds to conditional independence relationship in the model M from which G was constructed. When dealing with regression models, conditional independence translates to zero partial correlation and, therefore, every partial correlations that corresponds to pair-wise separations in $D(M)$ is guaranteed to vanish in M. We denote the set of all these vanishing partial correlations by $R(D)$, and we say that each member of $R(D)$ is *entailed by D*.

Thus, an obvious choice of a basis for C_0, as well as for $R(D)$, is the set of equalities $B_{pa} = \{\rho_{ij \cdot pa_i} = 0 | i > j\}$, where i ranges over all nodes in D, j ranges over all predecessors of i in any order that agrees with the arrows of D, and pa_i stands for the set of parents of node i in D. For example, the parent basis for the model in Fig. 1 would consist of the elements: $B_{pa} = \{\rho_{32 \cdot 1} = 0, \rho_{41 \cdot 3} = 0, \rho_{42 \cdot 3} = 0, \rho_{51 \cdot 43} = 0, \rho_{52 \cdot 43} = 0\}$ Testing for these equalities is sufficient therefore for testing the vanishing of all elements of C_0. However, when the parent sets pa_i are large, it may be possible to select a more economical basis (see Eq. (3)), as stated in the next theorem.

Theorem 1 *Let (i, j) be a pair of nonadjacent nodes in a DAG D, and Z_{ij} any set of nodes such that:*

(i) Z_{ij} d-separates i from j in the graph, and

(ii) Z_{ij} contains only nodes that are closer to i than j is.

The set of zero partial correlations $B_{sep} = \{\rho_{ij \cdot Z_{ij}} = 0 | i > j\}$, consisting of one element per nonadjacent pair, constitutes a basis for the set $R(D)$ of all vanishing partial correlations entailed by D.

That no proper subset of B_{sep} implies the vanishing of C_0 follows from the observation that for every DAG D there exists a covariance matrix whose vanishing partial correlations coincide precisely with the separation conditions in D. Had any proper subset B' of B_{sep} been a basis for C_0, the missing inequalities would have to be implied by B', and this would mean that the diagram created by adding arrows to D for each element of $B_{sep} \backslash B$ would be inconsistent, contrary to the theorem of [Geiger and Pearl, 1990].

Section 4 establishes several lemmas which provide weak versions of Theorem 1 and lead the way toward the proof. These lemmas are based on two properties of partial correlations,

[1] The d in d-separation connotes "directional." In this paper, however, we will use the term "separation" to mean d-separation.

called weak union and contraction in [Pearl, 1988].

$$\text{weak union}: \quad \rho_{ij\cdot Z} = 0 \ \& \ \rho_{ik\cdot Z} = 0 \quad \Rightarrow \quad \rho_{ij\cdot Zk} = 0 \tag{4}$$

$$\text{contraction}: \quad \rho_{ij\cdot Zk} = 0 \ \& \ \rho_{ik\cdot Z} = 0 \quad \Rightarrow \quad \rho_{ij\cdot Z} = 0 \tag{5}$$

To facilitate the derivation, we introduce additional notation. For any three sets of variables, S_1, S_2 and S_3, we shall write $(S_1, S_2, S_3)_D$ if, in diagram D, the nodes associated with S are separated from those associated with S_3, by the nodes associated with set S_2. Correspondingly, we write $(S_1, S_2, S_3)_P$ if, in a probability function P, the set of variables S_1 is conditionally independent of the set S_3, given the variables in set S_2. Thus, the d-separation theorem mentioned above can be stated succinctly as:

$$(S_1, S_2, S_3)_D \Longrightarrow (S_1, S_2, S_3)_P \tag{6}$$

whenever $(Y_i, pa_i, \{Y_1, \dots, Y_{i-1}\} \backslash pa_i)_P$ holds for $i = 2, 3, \dots, n$. Whenever this implication holds, we will say that D is an I-map of P (see [Pearl, 1988, p. 96]).

In this paper, our concern lies not with general conditional independencies but rather with vanishing partial correlations. To this end, we will continue to use the notation $(S_1, S_2, S_3)_P$ to denote zero partial correlation $\rho_{ij\cdot S_3} = 0$, where i is any element of S_1 and j is any element of S_3. However, in addition to the properties of weak union and contraction, written

$$\text{weak union}: \quad (S_1, S_2, S_3)_P \& (S_1, S_2, S_4)_P \Rightarrow (S_1, S_2 S_4, S_3)_P \tag{7}$$

$$\text{contraction}: \quad (S_1, S_2 S_4, S_3)_P \& (S_1, S_2, S_4)_P \Rightarrow (S_1, S_2, S_3)_P \tag{8}$$

we now use a third property, called composition:

$$\text{composition} \ (S_1, S_2, S_3)_P \& (S_1, S_2, S_4)_P \Leftrightarrow (S_1, S_2, S_3 S_4)_P$$

which holds for partial correlations. We will permit the sets $S_1 S_3$ to intersect with S_2, with the understanding that $(S_1, S_2, S_4)_P$ stands for $(S_1 \backslash S_2, S_2, S_3 \backslash S_2)_P$

4 PROOF OF THEOREM 1

Lemma 1 *The set of independencies $B_{pa} = \{(i, pa_i, j)_P | i > j\}$ is a basis for $R(D)$.*

Proof: As mentioned in Section 2, Lemma 1 is a special case of the d-separation theorem, in the context of *compositional* independencies, that is, independence of individual elements in a set implies the independence of the entire set.

Lemma 2 *The set $B_{pa'} = \{(i, Z_{ij}, j)_P | i > j\}$ is a basis of $R(D)$ if Z_{ij} is any subset of pa_i that separates i from j in D.*

Proof:
This proof will consist of two parts. In part 1 we will show that any set Z_{ij} that separates i from j also separates pa_i from j, that is,

$$(i, Z_{ij}, j)_D \Rightarrow (pa_i, Z_{ij}, j)_D \tag{9}$$

Indeed, if the r.h.s of Eq. (9) is false, then for some node t in $pa_i \backslash Z_{ij}$ there must be a path $t \ldots\ldots k$ that is not blocked by Z_{ij}. This implies that the path $i \leftarrow t \ldots\ldots k$ is, likewise, not blocked by Z_{ij}, which contradicts the assumption $(i, Z_{ij}, j)_D$. Thus, $(t, Z_{ij}, j)_D$ holds for every $t \in pa_i \backslash Z_{ij}$, which establishes the first part of the proof.

For the second part, we proceed by induction. We assume that Lemma 2 is true for the predecessors of i, $i' = 1, 2, \ldots, i-1$ in some ordering of the nodes that agrees with the arrows in D. This means that, D_{i-1}, the subgraph induced by the nondescendants of i, is an I-map of P – all separations in D_{i-1} stand for valid independencies in P. We then set to prove that Lemma 2 holds in D_i, the graph induced by i together with its predecessors.

Noting that the separation condition on the r.h.s of Eq. (9), $(pa_i, Z_{ij}, j)_D$, involves only nondescendants of i, we have $(pa_i, Z_{ij}, j)_P$, since D_{i-1} is an I-map of P. We also have $(i, Z_{ij}, j)_P$ by the assumption of Lemma 2. Thus, by composition and weak union,

$$(pa_i, Z_{ij}, j)_P \ \& \ (i, Z_{ij}, j)_P \Rightarrow (i, pa_i Z_{ij}, j)_P \tag{10}$$

Clearly, for $Z_{ij} \subseteq pa_i$, Eq. (10) implies $(i, pa_i, j)_P$, and establishes Lemma 2, because the set of independencies

$$\{(i, pa_i, j)_P \mid j < i, (i, j) \text{ nonadjacent}\} \tag{11}$$

coincides with the basis B_{pa} of Lemma 1.

Proof of Theorem 1: Let $d(i, j)$ denote the shortest distance between nodes j and i, and $d(Z_{ij})$ the highest $d(i, k)$ of any member k of set Z_{ij}. We will prove Theorem 1 by double induction; first on i and, second, for any fixed i, on $d(i, j)$.

For a fixed i, Eq. (10) holds for all $j < i$ whenever a separating set Z_{ij} is found that satisfies $(i, Z_{ij}, j)_P$. We need to show that the set of independencies

$$\{(i, pa_i Z_{ij}, j)_P \mid j < i, (i, Z_{ij}, j)_D\} \tag{12}$$

implies Eq. (11), whenever Z_{ij} satisfies $d(Z_{ij}) < d(i, j)$ of all $i > j$.

The lemma is certainly true for $d(i, j) = 2$, namely for any node j that is adjacent to pa_i. For any such node, the separating set Z_{ij} that enters Eq. (12) must be a subset of pa_i, hence, we immediately obtain (from (10)):

$$(i, pa_i Z_{ij}, j)_P \Rightarrow (i, pa_i, j)_P \tag{13}$$

Now assume the relation $(i, pa_i, j)_P$ holds for any j such that $d(i, j) < d$, and consider an arbitrary node j such that $d(i, j) = d > 2$. Based on condition (ii) of the theorem, every member k of Z_{ij} must have $d(i, k) < d$ and, therefore, the induction hypothesis entails

$$(i, pa_i, k)_D \Rightarrow (i, pa_i, k)_P \tag{14}$$

and since every k in Z_{ij} satisfies the l.h.s. of (14), we have

$$(i, pa_i, Z_{ij})_P \tag{15}$$

Putting (12) and (15) together, and using contraction, we get

$$(i, pa_i Z_{ij}, j)_P \text{ and } (i, pa_i, Z_{ij})_P \Rightarrow (i, pa_i, j)_P$$

which proves the lemma.

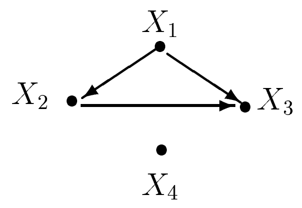

Figure 2:

5 REMARKS TOWARD EXTENDING THEOREM 1

Theorem 1 is sufficient for showing that Eq. (3) forms a basis for the model of Fig. 1. For example, $\rho_{51\cdot 3}$ is justified because $\{X_3\} \subseteq pa_5$. The term $\rho_{52\cdot 1}$, though it does not satisfy the conditions of Lemma 2, meets those of Theorem 1 and qualifies Eq. (3) as a basis.

To see why Theorem 1 may be extendable, consider the model in Fig. 2, the basis of which is $\{(\rho_{4j} = 0, j = 1, 2, 3)\}$. The set $\{\rho_{41\cdot 23} = 0, \rho_{42\cdot 13} = 0, \rho_{43\cdot 12} = 0\}$ forms a basis for $R(D)$, though it does not meet condition (ii) of the theorem. This follows by applying the axiom of intersection to the given three independencies:

$$\begin{aligned}(4, 23, 1), (4, 13, 2) &\rightarrow (4, 3, 12)\\ (4, 3, 12), (4, 12, 3) &\rightarrow (4, 0, 123).\end{aligned}$$

which yields the correct basis $(4, 0, 123)_P$

This example points to another method of testing bases. The statistics of regression models is completely specified by the covariance matrix $cov(i, j)$, which has $n(n-1)/2$ off-diagonal terms. Every nonadjacent pair (i, j) in the diagram is separated by some set Z_{ij} and imposes a constraint $\rho_{ij\cdot Z_{ij}} = 0$ on the covariance matrix. If these constraints lead to a unique solution $\rho_{ij\cdot pa_i} = 0$ for every pair $i > j$, then the sets Z_{ij} constitute a basis. Else, if the constraints are not sufficient for imposing a unique solution $(= 0)$ on each $\rho_{ij\cdot pa_i}$, then Z_{ij} does not represents a basis. This algebraic approach to extending Theorem 1 should be the used as a last resort, after exhausting the axiomatic approach.

To show that condition (ii) of Theorem 1 cannot be relaxed to allow just any separating set Z_{ij}, consider the model in Fig. 2, and assume we are given the following three independencies:

$$(4, 12, 3),\ (4, 3, 2),\ (4, 0, 1)$$

each is represented by a genuine separation in the graph. To show that this set is not a basis for Fig. 2, we note that none of the graphoid axioms is applicable to these triples, and therefore we can't prove $(4, 0, 23)$.

This is still not a proof, because the graphoid axioms are not complete relative to correlational independencies; a direct proof is feasible in this case. Using the recursion relation for partial correlations, we can obtain a non-zero solution for ρ_{42} and ρ_{43} and still satisfy the three triplets. These three triplets impose the following constraints on ρ_{12}, ρ_{13} and ρ_{32}:

$$\rho_{23}^2 + \rho_{12}^2 - \rho_{23}\rho_{13}\rho_{12} = 1.$$

and

$$\frac{\rho_{42}}{\rho_{43}} = \rho_{32}$$

Clearly, one can easily satisfy these constraints and obtain a nonzero values for ρ_{42} and ρ_{43}, as the following (positive definite) matrix shows:

$$R = \begin{matrix} 1 & \frac{3}{4} & \frac{2}{9} & 0 \\ \frac{3}{4} & 1 & \frac{3}{4} & \frac{3}{5} \\ \frac{2}{9} & \frac{3}{4} & 1 & \frac{4}{5} \\ 0 & \frac{3}{5} & \frac{4}{5} & 1 \end{matrix} \tag{16}$$

The last example illustrates some of considerations needed for extending Theorem 1 to nonrecursive regression, such as the one used in Markov fields over undirected graphs. The fundamental basis for Markov fields is given by the pair-wise Markov condition [Pearl 1988, Chapter 3; Lauritzen 1996], which consists of all nonadjacent pairs, each separated by all other nodes in the model. The local Markov condition (invoking the neighbors of each node in the graph) is not a basis, because some neighborhood-based separations can be derived from other such separations. If the graph is decomposable, it can be oriented into a DAG (preserving I-mapness) and we can choose a basis by Theorem 1 along any such orientation. The interesting question is how to deal with nondecomposable graphs when the pair-wise basis is too wasteful. One possibility is to make the graph decomposable by filling-in some edges, orient the graph and find a basis according to Theorem 1, and finally, to handle the filled-in edges using a pair-wise Markov condition on the corresponding clicks. This and other extensions to Theorem 1 will be discussed in a follow-up report.

References

[Geiger and Pearl, 1990] D. Geiger and J. Pearl. Logical and algorithmic properties of independence and their application to Bayesian networks. *Annals of Mathematics and AI*, 2(1–4):165–178, 1990.

[Lauritzen, 1996] S.L. Lauritzen. *Graphical Models.* Clarendon Press, Oxford, 1996.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligence Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

[Verma and Pearl, 1988] T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*, pages 352–359, Mountain View, CA, 1988. Also in R. Shachter, T.S. Levitt, and L.N. Kanal (Eds.), *Uncertainty in AI 4*, Elesevier Science Publishers, 69–76, 1990.