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.

We can, of course, test the entire model as a unit and obtain an overall measure of fit between the model and the data [Chou and Bentler, 1995]. Such a global measure of fit, however, would be of little help in model construction, for it does not provide the investigator with information as to parameter is responsible for a bad fit (if any) and which substantive assumption requires reconsideration [Pearl, 1998]. An alternative would be to actually perform the regressions, using least-square methods, and test if each member of $C_0$ is indeed zero. This test provides the desired information on individual members of $C_0$, but may require high order regressions, hence loss of power, especially for large values of $i$. 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 to $i$,

3. test the hypothesis $r_{ij:Z_{ij}} = 0$ for each $i > j$, where $r_{ij:Z_{ij}}$ is the coefficient of $Y_j$ in the regression of $Y_i$ on $\{Y_j \cup Z_{ij}\}$.

We show in this paper that if the regression coefficients $r_{ij:Z_{ij}}$, 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, $Z_{ij} = P_i$, which yields the standard test used in validating Bayesian networks. However, when the size of $P_i$ is large, it might be advantageous to use non-parental separators, as shown below.

2 AN EXAMPLE

Consider the set $M$ of regression equations

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

(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:

$$
C_0 = \{\rho_{32-1} = 0, \rho_{41-23} = 0, \rho_{42-13} = 0, \rho_{51-234} = 0, \rho_{52-134} = 0\}
$$

(2)

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:

$$
B = \{\rho_{32-1} = 0, \rho_{41-3} = 0, \rho_{42-1} = 0, \rho_{51-3} = 0, \rho_{52-1} = 0\}
$$

(3)

We see that $B$ and $C_0$ contain the same number of elements, yet all elements of $B$ have at most three
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 $r_{ij}Z$ of $B$. This follows from the $d$-separation 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 $B_{pa} = \{\rho_{ij|pa_i} = 0|i > j\}$, 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: $B_{pa} = \{\rho_{321} = 0, \rho_{413} = 0, \rho_{423} = 0, \rho_{5143} = 0, \rho_{534} = 0\}$

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 to $i$. The set of zero partial correlations

$$B_{sep} = \{\rho_{ij|Z_{ij}} = 0|i > j\};$$

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].

An economical basis satisfying the condition of Theorem 1 can easily be found using graphical techniques. For example, if $Z_{ij}$ is any subset of $pa_i$ that separates $i$ from $j$, it must satisfy condition (ii) and constitutes, therefore, a legitimate basis for $M$. If we further wish to find a basis that minimizes the number of regressors, or some other cost function, the separator-minimization methods of [Acid and de Campos, 1996] and [Tian et al., 1998] can be used, since the restrictions imposed by condition (ii) merely exclude a set of nodes from entering the separator $Z_{ij}$.
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, called weak union and contraction in [Pearl, 1988].

weak union: 
\[ \rho_{ij \cdot Z} = 0 & \& \rho_{ik \cdot Z} = 0 \Rightarrow \rho_{ij \cdot k \cdot Z} = 0 \] (4)

contraction: 
\[ \rho_{ij \cdot k \cdot Z} = 0 & \& \rho_{ik \cdot Z} = 0 \Rightarrow \rho_{ij \cdot Z} = 0 \] (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 \( S_1 \) 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:

\[ (S_1, S_2, S_3)_D \Rightarrow (S_1, S_2, S_3)_P \] (6)

whenever \( (Y_i, p_{ai}, \{Y_1, \ldots, Y_{i-1}\})_P \) holds for \( i = 2, 3, \ldots, n \). 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 \( \rho_{ij \cdot S_r} = 0 \), 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

weak union: 
\[ (S_1, S_2, S_3)_P & (S_1, S_2, S_4)_P \Rightarrow (S_1, S_2, S_4, S_3)_P \] (7)

contraction: 
\[ (S_1, S_2, S_4, S_3)_P & (S_1, S_2, S_4)_P \Rightarrow (S_1, S_2, S_3)_P \] (8)

we now use a third property, called composition:

\[ (S_1, S_2, S_3)_P & (S_1, S_2, S_4)_P \Rightarrow (S_1, S_2, S_3, S_4)_P \]

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 \( (S_1 \setminus S_2, S_2, S_4) \).

4 PROOF OF THEOREM 1

Lemma 1

The set of independencies \( B_{pa} = \{(i, pa_i, j)_P \mid i > j \} \) 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 \( B_{pa} = \{(i, Z_{ij}, j)_P \mid i > j \} \) 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,

\[ (i, Z_{ij}, j)_D \Rightarrow (pa_i, Z_{ij}, j)_D \] (9)

Indeed, if the r.h.s of Eq. (9) is false, then for some node \( t \) in \( pa_i \setminus Z_{ij} \), there must be a path \( t \ldots \ldots j \) that is not blocked by \( Z_{ij} \). This implies that the path \( i \leftarrow t \ldots \ldots j \) 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 \( t \in pa_i \setminus Z_{ij} \), 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 \( i \), \( i', i = 1, 2, \ldots, i - 1 \) 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 \( (pa_i, Z_{ij}, j)_D \), since \( D_{i-1} \) is an \( I \)-map of \( P \). We also have \( (i, Z_{ij}, j)_D \) by the assumption of Lemma 2. Thus, by composition and weak union,

\[ (pa_i, Z_{ij}, j)_P \& (i, Z_{ij}, j)_P \Rightarrow (i, pa_i, Z_{ij}, j)_P \] (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

\[ \{(i, pa_i, j)_P \mid j < i, (i, j) \text{ nonadjacent} \} \] (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

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

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

\[
(i, pa_i, Z_{ij}, j)_{p} \Rightarrow (i, pa_i, j)_{p} 
\]

(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

\[
(i, pa_i, k)_{D} \Rightarrow (i, pa_i, k)_{p} 
\]

(14)

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

\[
(i, pa_i, Z_{ij})_{p} 
\]

(15)

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

\[
(i, pa_i Z_{ij}, j)_{p} \quad \text{and} \quad (i, pa_i, Z_{ij})_{p} \Rightarrow (i, pa_i, j)_{p} 
\]

which proves the theorem.

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, \( \rho_{31:3} \) is justified because \( \{X_3\} \subseteq pa_5 \). The term \( \rho_{32:1} \), 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 \( \{\rho_{41:23} = 0, \rho_{42:13} = 0, \rho_{43:12} = 0\} \) 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:

\[
(4,23,1), (4,13,2) \Rightarrow (4,3,12) \\
(4,3,12), (4,12,3) \Rightarrow (4,0,123). 
\]

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 \( \rho_{ij-Z_{ij}} = 0 \) on the covariance matrix. If these constraints lead to a unique solution \( \rho_{ij-pa_i} = 0 \) 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 \( \rho_{ij-pa_i} \), then \( Z_{ij} \) does not represent a basis. This algebraic approach to extending Theorem 1 could be 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 \( \rho_{42} \) and \( \rho_{43} \) and still satisfy the three triplets. These three triplets impose the following constraints on \( \rho_{12}, \rho_{13} \) and \( \rho_{22} \):

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

and

\[
\frac{\rho_{42}}{\rho_{43}} = \rho_{32} 
\]

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

\[
R = \begin{bmatrix}
1 & 3 & \frac{3}{4} & \frac{3}{5} & 0 \\
\frac{3}{4} & 1 & \frac{3}{4} & \frac{3}{5} & \frac{3}{4} \\
\frac{3}{5} & \frac{3}{4} & 1 & \frac{3}{5} & \frac{3}{4} \\
0 & \frac{3}{5} & \frac{3}{4} & 1 & \frac{3}{5} \\
\end{bmatrix} 
\]

(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.

Azaria Paz (personal communication, 1998) has suggested a promising approach to finding economic bases for undirected graphs. It rests on the conjecture that any given graph \( G_1 \) can be triangulated into two chordal graphs, \( G_2 \) and \( G_3 \), such that the edges added in creating \( G_2 \) are disjoint of those added in creating \( G_3 \). The disjointness of the two sets of edges ensures that the graphoid closure of the independencies entailed by \( G_2 \) and \( G_3 \) coincides with the set of independencies entailed by \( G_1 \). Therefore, the union of the bases of \( G_2 \) and \( G_3 \) is a basis of \( G_1 \). But \( G_2 \) and \( G_3 \) are DAG equivalent. Thus, we can find economical bases for the DAG equivalents of \( G_2 \) and \( G_3 \) using Theorem 1, and the union of those two bases is an economical base for \( G_1 \).

This and other extensions of Theorem 1 will be developed in a future paper.

References


