PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 25, Fasc. 2 (2005), pp. 231–239

PERFECT TREE-LIKE MARKOVIAN DISTRIBUTIONS

BY

ANN BECKER (HAIFA), DAN GEIGER* (HAIFA)
AND CHRISTOPHER MEEK (REDMOND, WA)

Abstract. We show that if a strictly positive joint probability distribution for a set of binary variables factors according to a tree, then vertex separation represents all and only the independence relations encoded in the distribution. The same result is shown to hold also for multivariate nondegenerate normal distributions. Our proof uses a new property of conditional independence that holds for these two classes of probability distributions.

AMS Mathematics Subject Classification: 60E05.

Key words and phrases: Conditional independence, graphical models, Markov models.

1. INTRODUCTION

A useful approach to multivariate statistical modeling is first to define the conditional independence constraints that are likely to hold in a domain, and then to restrict the analysis to probability distributions that satisfy these constraints. Directed and undirected graphical models where independence constraints are encoded through the topological properties of the corresponding graphs are an increasingly popular way of specifying independence constraints (Lauritzen (1989), Lauritzen and Spiegelhalter (1988), Pearl (1988), Whittaker (1900)).

The key idea behind these specification schemes is to utilize the correspondence between *vertex separation* in graphs and *conditional independence* in probability; each vertex represents a variable and if a set of vertices Z blocks all the paths between two vertices, then the corresponding two variables are asserted to be conditionally independent given the variables corresponding to Z. The success of graphical models stems in part from the fact that vertex separation and conditional independence share key properties which render graphs an effective language for specifying independence constraints.

---

* Part of this work was done while the author was on sabbatical at Microsoft Research.

In this paper we show that when graphical models are trees and distributions are from specific classes, then the relationship between vertex separation and conditional independence is much more pronounced. More specifically, we show that if a strictly positive joint probability distribution for a set of binary variables factors according to a tree, then vertex separation represents all and only the independence relations encoded in the distribution. The same result is shown to hold also for multivariate nondegenerate normal distributions.

The class of Markov trees has been studied in several contexts. Practical algorithms for learning Markov trees from data have been used for pattern recognition (Chow and Liu (1968)). Geometrical properties of families of tree-like distributions have been studied in Settimi and Smith (1999). Finally, the property of perfectness, when a graphical model represents all and only the conditional independence facts encoded in a distribution, is a key assumption in learning causal relationships from observational data (Glymour and Cooper (1999)).

2. PRELIMINARIES

Throughout this article we use lowercase letters for single variables (e.g., x, y, z) and boldfaced lowercase letters (e.g., $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$) for specific values for these variables. Sets of variables are denoted by capital letters (e.g., X, Y, Z), and their values are denoted by boldfaced capital letters (e.g., $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$). For example, if $Z = \{x, y\}$, then $\mathbf{Z}$ stands for $\{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x}$ is a value of x and $\mathbf{y}$ is a value of y. Let P stand for a density function or a probability distribution function as appropriate. We use $P(\mathbf{X})$ as a shorthand notation for $P(X = \mathbf{X})$. We say that $P(X)$ is *strictly positive* if $\forall \mathbf{X}\ P(\mathbf{X}) > 0$. We use Xy as a shorthand notation for $X \cup \{y\}$.

Let X, Y and Z be three disjoint sets of variables having a joint distribution $P(X, Y, Z)$. Then X and Y are conditionally independent given Z, denoted by $X \perp_P Y \mid Z$, if and only if

$$\forall \mathbf{X} \forall \mathbf{Y} \forall \mathbf{Z}\ P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) P(\mathbf{Z}) = P(\mathbf{X}, \mathbf{Z}) P(\mathbf{Y}, \mathbf{Z}).$$

The ternary relation $X \perp_P Y \mid Z$ was studied by Dawid (1979) and further investigated (e.g., in Spohn (1980), Pearl and Paz (1985), Pearl (1988), Geiger and Pearl (1993), Matúš (1992), Studeny (1992)). The ternary relation $X \perp_P Y \mid Z$ satisfies the following five properties which are called the *graphoid axioms* (Pearl and Paz (1985)):

- Symmetry:

(1) $$X \perp_P Y \mid Z \Rightarrow Y \perp_P X \mid Z.$$

- DECOMPOSITION:

(2) $$X \perp_P YW \mid Z \Rightarrow X \perp_P Y \mid Z.$$

- WEAK UNION:

(3) $$X \perp_P YW \mid Z \Rightarrow X \perp_P Y \mid ZW.$$

- CONTRACTION:

(4) $$X \perp_P Y \mid Z \wedge X \perp_P W \mid ZY \Rightarrow X \perp_P YW \mid Z.$$

If P is strictly positive, then

- INTERSECTION:

(5) $$X \perp_P Y \mid ZW \wedge X \perp_P W \mid ZY \Rightarrow X \perp_P YW \mid Z.$$

The following property holds for multivariate normal distributions $P(X, Y, Z, c)$ (Pearl (1988)). It also holds for discrete variables if $Z = \emptyset$ and c is a binary variable (Lauritzen and Spiegelhalter (1988), p. 219).

- WEAK TRANSITIVITY:

(6) $$X \perp_P Y \mid Z \wedge X \perp_P Y \mid Zc \Rightarrow X \perp_P c \mid Z \vee c \perp_P Y \mid Z.$$

A *Markov graph* of a probability distribution $P(x_1, \ldots, x_n)$ is an undirected graph $G = (V, E)$, where $V = \{x_1, \ldots, x_n\}$ is a set of vertices, one for each variable x_i, and E is a set of edges each represented as (x_i, x_j) such that

$$(x_i, x_j) \in E \text{ if and only if } \neg\, x_i \perp_P x_j \mid \{x_1, \ldots, x_n\} \backslash \{x_i, x_j\}.$$

A *Markov tree* is a Markov graph where G has no cycles; namely, every pair of vertices in G is connected with at most one path. This definition of a Markov tree permits G to be disconnected (often called a *forest*).

A key property of Markov graphs is the following. Let $A \perp_G B \mid C$ stand for the assertion that every path in G between a vertex in A and a vertex in B passes through a vertex in C, where A, B, and C are mutually disjoint sets of vertices. Note that whenever $A \perp_G B \mid C$ holds in G, A and B are separated by C. The ternary relation $A \perp_G B \mid C$ satisfies all the properties we listed for $A \perp_P B \mid C$ and some additional properties that do not hold for $A \perp_P B \mid C$ (Pearl (1988)).

THEOREM 1 (Pearl and Paz (1985)). *Let G be a Markov graph of $P(x_1, \ldots, x_n)$, and suppose Intersection holds for P. Then*

(7) $$A \perp_G B \mid C \Rightarrow A \perp_P B \mid C$$

for every disjoint set of vertices A, B, and C of G and their corresponding variables in $\{x_1, \ldots, x_n\}$.

The main result in this paper is a converse to the relation (7) under suitable conditions. When the converse holds, we say that G is a *perfect* *representation* of P. To facilitate our argument we must first introduce a new property for conditional independence.

- DECOMPOSABLE TRANSITIVITY:

(8) $$aB \perp_P De \mid c \wedge a \perp_P e \mid BD \Rightarrow a \perp_P c \mid B \vee c \perp_P e \mid D.$$

3. NEW PROPERTY OF CONDITIONAL INDEPENDENCE

We now prove that Decomposable transitivity holds for joint probability distributions of binary variables and for multivariate nondegenerate normal distributions. We then show that Decomposable transitivity holds also for vertex separation in undirected graphs.

THEOREM 2. *Let a, c, e be binary variables, B and D be (possibly empty) sets of variables, and $P(a, c, e, B, D)$ be a joint probability distribution for these variables. Then*

$$aB \perp_P De \mid c \wedge a \perp_P e \mid BD \Rightarrow a \perp_P c \mid B \vee c \perp_P e \mid D$$

holds for P.

Proof. We use $\mathbf{a}$ to denote a value for a, $\mathbf{B}$ to denote a value for a set of variables B, and a^0 and a^1 to denote the two values of a binary variable a. If $P(\mathbf{c}) = 0$ for any of the values of c, then the conclusion of the theorem is immediate.

From $aB \perp_P De \mid c$ it follows that

(9) $$P(\mathbf{a}, \mathbf{B}, \mathbf{c}, \mathbf{D}, \mathbf{e}) \cdot P(\mathbf{c}) = P(\mathbf{a}, \mathbf{B}, \mathbf{c}) \cdot P(\mathbf{c}, \mathbf{D}, \mathbf{e})$$

for every value $\mathbf{a}$, $\mathbf{c}$, $\mathbf{e}$, $\mathbf{B}$, $\mathbf{D}$ of the corresponding variables. From $a \perp_P e \mid BD$ we obtain

(10) $$P(\mathbf{a}^0, \mathbf{B}, \mathbf{D}, \mathbf{e}^0) \cdot P(\mathbf{a}^1, \mathbf{B}, \mathbf{D}, \mathbf{e}^1) = P(\mathbf{a}^1, \mathbf{B}, \mathbf{D}, \mathbf{e}^0) \cdot P(\mathbf{a}^0, \mathbf{B}, \mathbf{D}, \mathbf{e}^1)$$

for every value $\mathbf{B}$, $\mathbf{D}$ of B, D. Since c is a binary variable

(11) $$P(\mathbf{a}, \mathbf{B}, \mathbf{D}, \mathbf{e}) = P(\mathbf{a}, \mathbf{B}, \mathbf{c}^0, \mathbf{D}, \mathbf{e}) + P(\mathbf{a}, \mathbf{B}, \mathbf{c}^1, \mathbf{D}, \mathbf{e}).$$

Now, substituting (9) into (11), then substituting the result into (10), we get

$$\alpha(\mathbf{B})\beta(\mathbf{D}) = 0,$$

where

$$\alpha(\mathbf{B}) = P(\mathbf{a}^1, \mathbf{B}, \mathbf{c}^0) \cdot P(\mathbf{a}^0, \mathbf{B}, \mathbf{c}^1) - P(\mathbf{a}^0, \mathbf{B}, \mathbf{c}^0) \cdot P(\mathbf{a}^1, \mathbf{B}, \mathbf{c}^1)$$

and

$$\beta(\mathbf{D}) = P(\mathbf{c}^1, \mathbf{D}, \mathbf{e}^0) \cdot P(\mathbf{c}^0, \mathbf{D}, \mathbf{e}^1) - P(\mathbf{c}^0, \mathbf{D}, \mathbf{e}^0) \cdot P(\mathbf{c}^1, \mathbf{D}, \mathbf{e}^1).$$

Consequently, either $\alpha(\mathbf{B}) = 0$ or $\beta(\mathbf{D}) = 0$. Furthermore, since $\mathbf{B}$ and $\mathbf{D}$ are arbitrary values of B and D, respectively, we have $\forall \mathbf{B} \forall \mathbf{D} [\alpha(\mathbf{B}) =$

$0 \vee \beta(D) = 0]$, which is equivalent to $[\forall B\ \alpha(B) = 0] \vee [\forall D\ \beta(D) = 0]$, which is equivalent to

$$a \perp_P c \mid B \vee c \perp_P e \mid D. \blacksquare$$

THEOREM 3. *Let a, c, and e be continuous variables, B and D be (possibly empty) sets of continuous variables, and let $P(a, c, e, B, D)$ be a multivariate nondegenerate normal distribution for these variables. Then*

$$aB \perp_P De \mid c \wedge a \perp_P e \mid BD \Rightarrow a \perp_P c \mid B \vee c \perp_P e \mid D$$

holds for P.

Proof. We use a formal logical deduction style to emphasize that the only properties of normal distributions being used are the ones encoded in Symmetry, Decomposition, Intersection, Weak union, and Weak transitivity. Recall that Weak transitivity holds for every normal distribution and that Intersection holds for nondegenerate normal distributions. The other properties hold for every probability distribution.

We now derive the conclusion of the theorem from its antecedents.

1. $aB \perp_P De \mid c$ (given).
2. $a \perp_P e \mid BD$ (given).
3. $a \perp_P D \mid cB$ (Weak union, Decomposition, and Symmetry on (1)).
4. $B \perp_P e \mid cD$ (Weak union, Decomposition, and Symmetry on (1)).
5. $a \perp_P e \mid BDc$ (Weak union and Symmetry on (1)).
6. $a \perp_P c \mid BD \vee c \perp_P e \mid BD$ (Weak transitivity on (2) and (5)).
7. $a \perp_P cD \mid B \vee Bc \perp_P e \mid D$ (Intersection and Symmetry on (3), (4) and (6)).
8. $a \perp_P c \mid B \vee c \perp_P e \mid D$ (Symmetry and Decomposition on (7)). $\blacksquare$

THEOREM 4. *Let a, c, and e be distinct vertices of an undirected graph G, and let B and D be two (possibly empty) disjoint sets of vertices of G that do not include a, c or d. Then*

$$aB \perp_G De \mid c \wedge a \perp_G e \mid BD \Rightarrow a \perp_G c \mid B \vee c \perp_G e \mid D$$

holds for G.

Proof. Assume the assertion of the theorem does not hold in G but its antecedents hold. Then there exists a path γ_1 in G between a and c such that no vertices from B reside on γ_1, and there exists a path γ_2 in G between c and e such that no vertices from D reside on γ_2. If B and D are empty, then the concatenated path $\gamma_1 \gamma_2$ contradicts $a \perp_G e \mid BD$, which is assumed to hold in G. Thus, we can assume that either B or D are not empty. The concatenated path $\gamma_1 \gamma_2$ contains a vertex from B or D (or both) because $a \perp_G e \mid BD$ is assumed to hold in G. Assume that a vertex $d \in D$ resides on the path γ_1 between a and c or that a vertex $b \in B$ resides on the path γ_2 between c and e. In the first case vertices a and d are connected and the path that connects them does not include c, and in the second case vertices b and e are connected and the path that connects them does not include c. Thus, in both cases, $aB \perp_G De \mid c$ does not hold in G, contradicting our assumption. ∎

Note that the proof of Theorem 3 is also a valid proof of Theorem 4 because it merely uses properties that hold for graph separation, namely, Symmetry, Decomposition, Intersection, Weak union, and Weak transitivity. However, the proof of Theorem 3 is not a valid proof for Theorem 2 because Weak transitivity does not hold for binary variables (Meek (1997)).

4. PERFECT MARKOVIAN TREES

We are ready to prove the main result.

THEOREM 5. *Let G be a Markov tree for a probability distribution $P(x_1, \ldots, x_n)$, where P satisfies Intersection and Decomposable transitivity, then*

(12) $$A \perp_G B \mid C \text{ if and only if } A \perp_P B \mid C$$

for every disjoint set of vertices A, B, and C of G and their corresponding variables in $\{x_1, \ldots, x_n\}$.

Proof. Theorem 1 proves one direction of (12), and so it remains to show that

(13) $$A \perp_P B \mid C \text{ implies } A \perp_G B \mid C.$$

To prove (13) it is sufficient to show that $a \perp_P b \mid C$ implies $a \perp_G b \mid C$ for every pair of vertices $a \in A$ and $b \in B$. This is sufficient because $A \perp_P B \mid C$ implies $\forall a \forall b \; a \perp_P b \mid C$ (due to Decomposition and Symmetry) and $\forall a \forall b \; a \perp_G b \mid C$ is equivalent by definition to $A \perp_G B \mid C$. An extended version of this reduction from sets A, B to singletons a, b has been studied by Matúš (1992), Lemma 3.

We proceed by contradiction. Let x and y be a pair of vertices for which there exists a set of vertices Z satisfying

(14) $$x \perp_P y \mid Z \wedge \neg x \perp_G y \mid Z$$

and such that x and y are connected with the shortest path among all pairs x', y' for which there exists a set Z' satisfying

$$x' \perp_P y' \mid Z' \wedge \neg x' \perp_G y' \mid Z'.$$

Note that the unique path between x and y does not include vertices from Z.

Suppose first that the path between x and y is merely an edge connecting the two vertices. We will now reach a contradiction by showing that G cannot be a Markov tree of P because P would have to satisfy $x \perp_P y \mid U_{xy}$, where U_{xy} are all variables except x and y.

Let U_x be all the vertices on the x side of the edge (x, y) and U_y be the rest of the vertices. (Namely, U_x are the vertices in the component of x after removing the edge (x, y) from the tree G.) Let $B = U_x \cap Z$ and $D = U_y \cap Z$. By these definitions and because G is a Markov tree of P we have:

1. $x \perp_P y \mid BD$.
2. $U_x \perp_P yU_y \mid x$.
3. $xU_x \perp_P U_y \mid y$.

We proceed by a formal deduction using Symmetry, Decomposition, Contraction, Weak union, and Intersection, to show that (1), (2), and (3) imply $x \perp_P y \mid U_x U_y$.

4. $B \perp_P y \mid xD$ (Weak union, Symmetry, and Decomposition on (2)).
5. $xB \perp_P y \mid D$ (Intersection and Symmetry on (1) and (4)).
6. $x \perp_P y \mid D$ (Decomposition and Symmetry on (5)).
7. $x \perp_P D \mid y$ (Decomposition and Symmetry on (3)).
8. $x \perp_P yD \mid \emptyset$ (Intersection on (6) and (7)).
9. $x \perp_P y \mid \emptyset$ (Decomposition on (8)).
10. $x \perp_P U_y \mid y$ (Decomposition and Symmetry on (3)).
11. $x \perp_P yU_y \mid \emptyset$ (Contraction on (9) and (10)).
12. $xU_x \perp_P yU_y \mid \emptyset$ (Contraction and Symmetry on (2) and (11)).
13. $x \perp_P y \mid U_x U_y$ (Weak union and Symmetry on (12)).

Now suppose that the path between x and y has more than one edge and that c is a vertex on this path. We reach a contradiction by showing that the pair x, y is not the closest pair of vertices that satisfy (14) for some set Z', contrary to our selection of these vertices. Let BD be a partition of Z such that B is the set of the vertices in Z on the x side of c and $D = Z \backslash B$. The rest of the derivation is a formal deduction using properties of conditional independence.

1. $xB \perp_G Dy \mid c$ (by definition of B and D in G).
2. $xB \perp_P Dy \mid c$ (from (1) and since G is a Markov graph of P).
3. $x \perp_P y \mid BD$ ($Z = BD$ and $x \perp_P y \mid Z$ is assumed).
4. $x \perp_P c \mid B \vee c \perp_P y \mid D$ (Decomposable transitivity on (2) and (3)).
5. $\neg x \perp_G c \mid B \wedge \neg c \perp_G y \mid D$ (by definition of B and D in G).
6. $[x \perp_P c \mid B \wedge \neg x \perp_G c \mid B] \vee [c \perp_P y \mid D \wedge \neg c \perp_G y \mid D]$ (by (4) and (5)).

Each disjunct in Step 6 exhibits a pair of vertices that are closer to each other than x and y and yet satisfy (14) for some set Z', contradicting our choice of x and y. ■

Note that the theorem applies in two interesting cases: when $x_1, \ldots, x_n$ are binary variables and $P(x_1, \ldots, x_n)$ is strictly positive or when $x_1, \ldots, x_n$ are continuous variables and P is a multivariate nondegenerate normal distribution, because in these cases Intersection and Decomposable transitivity hold.

5. REMARKS

Our axiomatic approach to reasoning about conditional independence and its relationship to separation in graphs follows the approach taken by Pearl and Paz (1985). Our proof uses a new property of conditional independence that holds for the two classes of probability distributions we have focused on.

The algorithmic consequence of Theorem 5 is that in order to check whether a Markov tree G of P represents all the conditional independence statements that hold in P, assuming P satisfies Intersection and Decomposable transitivity, requires merely to test whether for each edge (x, y) in G, $x \perp_P y \mid \emptyset$ holds rather than testing the harder condition of whether $x \perp_P y \mid U_{xy}$ holds in P as required by the definition of a Markov graph. An open question remains as to what is the minimal computation needed to ensure that a Markov graph G (other than a tree) of a probability distribution P represents all the conditional independence statements that hold in P and what properties P or G need to satisfy to accommodate these computations.

A straightforward attempt to extend our results without changing the tests or the assumptions on P is quite limited because we have counterexamples to Theorem 5 when G is a polytree (a directed graph with no underlying undirected cycles) and when P does not satisfy Intersection or Decomposable transitivity.

REFERENCES

[1] C. Chow and C. Liu, *Approximating discrete probability distributions with dependence trees*, IEEE Trans. Inform. Theory 14 (1968), pp. 462–467.

[2] A. P. Dawid, *Conditional independence in statistical theory (with discussion)*, J. Roy. Statist. Soc. Ser. B 41 (1979), pp. 1–31.

[3] D. Geiger and J. Pearl, *Logical and algorithmic properties of conditional independence and graphical models*, Ann. Statist. 21 (1993), pp. 2001–2021.

[4] C. Glymour and G. Cooper, *Computation, Causation, Discovery*, AAAI press/The MIT press, Menlo Park 1999.

[5] S. Lauritzen, *Lectures on Contingency Tables*, 3rd edition, Aalborg University, 1989.

[6] S. Lauritzen and D. Spiegelhalter, *Local computations with probabilities on graphical structures and their application to expert systems (with discussion)*, J. Roy. Statist. Soc. Ser. B 50 (1988), pp. 157–224.

[7] F. Matúš, *On equivalence of Markov properties over undirected graphs*, J. Appl. Probab. 29 (1992), pp. 745–749.

[8] C. Meek, *Graphical Models: Selecting Causal and Statistical Models*, Ph. D. Thesis, Carnegie Mellon University, 1997.

[9] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, California, 1988.

[10] J. Pearl and A. Paz, *Graphoids: a graph based logic for reasoning about relevancy relations*, in: *Advances in Artificial Intelligence. II*, B. Boulay, D. Hogg and L. Steel (Eds.), North-Holland, Amsterdam 1985, pp. 357–363. See also Pearl (1988), p. 139.

[11] R. Settimi and J. Smith, *Geometry, moments and bayesian networks with hidden variables* in: *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, Morgan Kaufmann, San Francisco 1999, pp. 293–298.

[12] W. Spohn, *Stochastic independence, causal independence, and shieldability*, J. Philos. Logic 9 (1980), pp. 73–99.

[13] M. Studeny, *Conditional independences have no finite complete characterization*, in: *Transactions of the 11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Academia, Prague 1992, pp. 3–16.
[14] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester 1990.

Ann Becker and Dan Geiger
Computer Science Department
Technion, Haifa, 32000 Israel
E-mail: anyuta@cs.technion.ac.il
dang@cs.technion.ac.il

Christopher Meek
Microsoft Research
Redmond, WA, 98052, U.S.A.
E-mail: meek@microsoft.com

Received on 11.10.2001;
re-submitted version on 8.9.2005