The Annals of Statistics
2000, Vol. 28, No. 4, 1179–1205

GEOMETRY, MOMENTS AND CONDITIONAL INDEPENDENCE TREES WITH HIDDEN VARIABLES[1]

BY RAFFAELLA SETTIMI AND JIM Q. SMITH

University of Chicago and University of Warwick

We study the geometry of the parameter space for Bayesian directed graphical models with hidden variables that have a tree structure and where all the nodes are binary. We show that the conditional independence statements implicit in such models can be expressed in terms of polynomial relationships among the central moments. This algebraic structure will enable us to identify the inequality constraints on the space of the manifest variables that are induced by the conditional independence assumptions as well as determine the degree of unidentifiability of the parameters associated with the hidden variables. By understanding the geometry of the sample space under this class of models we shall propose and discuss simple diagnostic methods.

1. Introduction. Graphical models have proved to be a powerful tool for building Bayesian models to analyze multivariate problems where all variables are observed [e.g., Spiegelhalter, Dawid, Lauritzen and Cowell (1993)]. In particular it is possible to estimate all the conditional probabilities that parameterize such models by using a conjugate analysis. However, when all the data on certain variables in an explanatory model are missing, conjugacy usually disappears, estimates of these conditional probabilities become highly dependent on one another and they often cannot be determined from data no matter how extensive that data is [see, e.g., Settimi and Smith (1998)].

In this paper we propose a geometrical approach to analyze such difficulties. We first observe that conditional independence assumptions implicit in directed graphical models induce some constraints on the model space, that can be expressed as polynomial equations among the central moments. We then exploit such an algebraic structure to explore the geometry and the irregularities of the parameter space for Bayesian directed graphical models with hidden variables, defined over a set of binary variables, and such that the conditional independence assumptions are represented via a directed tree. Understanding the geometry and the singularities of the parameter spaces will enable us to investigate practical statistical issues, such as parameter identifiability, model dimension and diagnostic methods.

In the statistical analyses of problems with missing data it has been common practice either to use various methods of approximation to calculate the posterior probabilities of the model parameters [see, e.g., Spiegelhalter and

Received January 1999; revised July 2000.
[1]Supported in part by EPSRC Grant GR/K 72254.
AMS 1991 *subject classifications*. Primary 62F15; secondary 62H17, 68R10.
Key words and phrases. Conditional independence, Bayesian networks, Bayesian multinomial models, model identifiability.

Cowell (1992), Cowell (1998), Ramoni and Sebastiani (1997)] or to resort to numerical algorithms like MCMC approaches [Madigan and York (1995)] or data augmentation methods [Tanner and Wong (1987)]. Although these methods are obviously useful tools in problems with missing data where all the variables are at least partially observed, when used in the context of hidden variables they are vulnerable to various difficulties.

A typical problem when estimating models from incomplete datasets, when data on some variables in the model are completely missing, is that different combinations of values of the conditional probabilities provide equally likely explanations for the observed data and the approximating methods cited above may only identify a subset of these. In a Bayesian setting this will mean that the typical posterior distribution of parameters will have many isolated modes. When the hidden variables have an interpretative value and help determine action, as is often the case, for example, in medical models, missing any possible good explanation may be disastrous. Furthermore there may exist a continuum of equally likely combinations of parameter values [see, e.g., Settimi and Smith (1998)]. This feature will make certain inferences within the model extremely sensitive to the prior density in a Bayesian analysis; some functions of the parameters being unidentifiable from the data.

For these reasons, when estimating probabilities in directed graphical models with hidden variables, it is useful to acquire a good understanding of the geometry of the likelihood of the data before embarking on any numerical search or approximation algorithm.

A complementary issue is how directed graphical models with manifest and hidden nodes constrain the joint distribution over the margin of the manifest variables. It is known that such conditional independence models induce either equality and inequality constraints over the marginal probabilities of the manifest variables [see, e.g., Spirtes, Richardson and Meek (1997)]. A study of the geometry of the feasible regions of the probability distributions of the manifest variables has already begun [see Geiger, Heckerman and Meek (1996), Geiger and Meek (1998) and Settimi and Smith (1998)].

For the sake of simplicity this paper will concentrate its study on the geometry of the parameter space for directed graphical models with hidden variables, when all the variables are binary and the conditional independence statements implied by the graphical models are represented by directed trees. In Section 2 we describe a systematic way of analyzing the geometry of these probability spaces by using polynomial equations of *noncentral moments* induced on the sample space [cf. Pistone, Riccomagno and Wynn (1999)] and polynomial equations on *central moments* which express the assumed conditional independence structure as discussed in Section 2.2.

Notice that because the geometry of the probability spaces over the binary trees is expressed in terms of polynomials on the central moments, the additive parameterization can be regarded as a more natural choice for these categorical models and in this paper is preferred to the commonly used multiplicative parameterization.

In Section 3 we give an explicit formula for the dimension of the space of the unidentifiable parameters of directed graphical models which are described by arbitrary directed (or equivalently undirected) trees. In Section 4 we present a few results on the geometry of the parameter space of the marginal distributions over the manifest variables for such tree models. We show that there is a very rich structure of inequalities on moment parameters which allow certain simple diagnostics to be constructed. In Section 5 we shall discuss how the geometrical study of the sample space can be used to implement effective diagnostic methods to test whether a given model with hidden variables is consistent with the observed data and make some initial steps towards constructing such diagnostics.

2. Moments of binary variables with dependence structure.

2.1. *Noncentral moments of a vector of binary variables.* Here we review some basic results about the noncentral moments of the joint distribution of n binary random variables $Y_1, \ldots, Y_n$ whose state space is $\{-1, 1\}$. Write $\mathbf{Y} := (Y_1, \ldots, Y_n)$ and let $p(y) := p(Y_1 = y_1, \ldots, Y_n = y_n)$ for $\mathbf{y} := (y_1, \ldots, y_n)$ be the probability that the random vector $\mathbf{Y}$ takes value $\mathbf{y}$ chosen in the set $\mathscr{Y}_n$.

Given a vector of nonnegative integers $\mathbf{a} := (a_1, \ldots, a_n)$, we shall define $\mathbf{Y^a} := \prod_{i=1}^n Y_i^{a_i}$. First, note that for any nonnegative integer $a_i, 1 \le i \le n$, then

$$\mathbf{Y}^{2\mathbf{a}} = \prod_{i=1}^{n} Y_i^{2a_i} = 1.$$

Thus it follows that

$$\mathbf{Y^a} = \mathbf{Y^{b(a)}} \text{ for } \mathbf{b}(\mathbf{a}) = (b_1, \ldots, b_n) \quad \text{with } b_i = a_i \bmod 2, \qquad 1 \le i \le n. \tag{2.1}$$

In particular considering the noncentral moment $m(\mathbf{a}) := \mathbb{E}(\mathbf{Y^a})$ we have that

$$m(\mathbf{a}) = m(\mathbf{b}(\mathbf{a})) \tag{2.2}$$

where $\mathbf{b}(\mathbf{a})$ is defined above.

By considering for example its characteristic function, it is clear that any distribution on $\mathbf{Y}$ is uniquely specified by its noncentral moments and hence, because of (2.2), by the set of moments $\mathscr{M} := \{m(\mathbf{b}) : \mathbf{b} \in B_n\}$, where B_n is the set of nonzero binary n-dimensional vectors. Clearly there are $2^n - 1$ such moments corresponding to the $2^n - 1$ probabilities associated with $\mathbf{Y}$.

Second, note that the moments set $\mathscr{M}$ is a proper subset of $[-1, 1]^n$ since the probabilities on $\mathbf{Y}$ are constrained to lie in a simplex. The simplex constraints are simply the redundant moment condition

$$m(\mathbf{0}) = \sum_{\mathbf{y} \in \mathscr{Y}_n} p(\mathbf{y}) = 1,$$

where $\mathbf{0}$ is a vector of all zeros, together with 2^n positivity constraints

$$p(\mathbf{y}) \ge 0 \quad \text{for all } \mathbf{y} \in \mathscr{Y}_n. \tag{2.3}$$

However because each $m(\mathbf{b})$ is a linear function of $p(\mathbf{y})$ for $\mathbf{b} \in B_n$, the 2^n linear inequalities (2.3) transform into 2^n inequality constraints on $\mathscr{M} := \{m(\mathbf{b})\colon \mathbf{b} \in B_n\}$. It is easy to check that this admissible region in the $\mathscr{M}$ moment space contains the zero vector, is convex (being the intersection of convex regions) and contains an open ball of dimension $2^n - 1$. An example of this space for $n = 4$ is given in Section 4.1.

2.2. *Conditional independence and central moments.* It is well known that two random vectors $\mathbf{Y} := (Y_1, \ldots, Y_{n_1})$ and $\mathbf{Z} := (Z_1, \ldots, Z_{n_2})$ are independent if and only if

$$\operatorname{Cov}(f_1(\mathbf{Y}), f_2(\mathbf{Z})) = 0$$

for all L_2 functions $f_1(\cdot), f_2(\cdot)$ [Feller (1971), page 136]. The general forms and dimensions of such relationships are studied and characterized in some detail, for example, in Streitberg (1990).

Here we consider only random vectors whose components can each take at most N integer values (usually $N = 2$). In this case all functions are equal to polynomials of order at most $(N-1)n_1$ and $(N-1)n_2$ respectively. It follows by the linearity of the expectation operator that $\mathbf{Y}$ and $\mathbf{Z}$ will be independent if and only if

$$\operatorname{Cov}(\mathbf{Y}^{\boldsymbol{\alpha}}, \mathbf{Z}^{\boldsymbol{\beta}}) = 0$$

for all monomials

$$\begin{aligned} \mathbf{Y}^{\boldsymbol{\alpha}} &:= \prod_{i=1}^{n_1} Y_i^{\alpha_i} \quad \text{for } \alpha_i = 0, 1, 2, \ldots, \qquad \boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_{n_1}), \\ \mathbf{Z}^{\boldsymbol{\beta}} &:= \prod_{i=1}^{n_2} Z_i^{\beta_i} \quad \text{for } \beta_i = 0, 1, 2, \ldots, \qquad \boldsymbol{\beta} := (\beta_1, \ldots, \beta_{n_2}). \end{aligned} \tag{2.4}$$

Note that in the particular case when the state space of each component of the random vector is $\{-1, 1\}$, the identities of the last sections imply that $\mathbf{Y}$ and $\mathbf{Z}$ are independent if and only if equations (2.4) hold for binary vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$, that is whenever α_i and β_i are equal to 0 or 1 for $1 \le i \le n_j$, $j = 1, 2$.

Consider now a binary random variable W. Analogous arguments to the unconditioned case above, provided that W is nondegenerate, that is, $\operatorname{Var}(W) > 0$, show that $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | W$ if and only if

$$\mathbb{E}(\mathbf{Y}^{\boldsymbol{\alpha}} - \mathbb{E}(\mathbf{Y}^{\boldsymbol{\alpha}}|W))(\mathbf{Z}^{\boldsymbol{\beta}} - \mathbb{E}(\mathbf{Z}^{\boldsymbol{\beta}}|W)) = 0 \tag{2.5}$$

for all binary vectors of $\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Furthermore, when W is binary, the conditional expected values $\mathbb{E}(\mathbf{Y}^{\boldsymbol{\alpha}}|W)$ and $\mathbb{E}(\mathbf{Z}^{\boldsymbol{\beta}}|W)$ must be linear in W. Therefore these expectations can be written as

$$\begin{aligned} \mathbb{E}(\mathbf{Y}^{\boldsymbol{\alpha}}|W) &= \mathbb{E}(\mathbf{Y}^{\boldsymbol{\alpha}}) + A[W - \mathbb{E}(W)] \quad \text{for } A = \operatorname{Cov}(\mathbf{Y}^{\boldsymbol{\alpha}}, W)\operatorname{Var}(W)^{-1}, \\ \mathbb{E}(\mathbf{Z}^{\boldsymbol{\beta}}|W) &= \mathbb{E}(\mathbf{Z}^{\boldsymbol{\beta}}) + B[W - \mathbb{E}(W)] \quad \text{for } B = \operatorname{Cov}(\mathbf{Z}^{\boldsymbol{\beta}}, W)\operatorname{Var}(W)^{-1}. \end{aligned}$$

By substituting these into (2.5) and simplifying, we can write

$$\begin{aligned}(2.6)\qquad &\mathbb{E}(Y^{\boldsymbol{\alpha}}Z^{\boldsymbol{\beta}}) - \mathbb{E}(Y^{\boldsymbol{\alpha}})\mathbb{E}(Z^{\boldsymbol{\beta}}) - A(\mathbb{E}(Z^{\boldsymbol{\beta}}W) - \mathbb{E}(Z^{\boldsymbol{\beta}})\mathbb{E}(W)) \\ &\quad -B(\mathbb{E}(Y^{\boldsymbol{\alpha}}W) - \mathbb{E}(Y^{\boldsymbol{\alpha}})\mathbb{E}(W)) + AB\mathbb{E}(W - \mathbb{E}(W))^2 = 0.\end{aligned}$$

By using the definition of the covariance function in (2.6), we have that $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|W$ if and only if

$$(2.7)\qquad \operatorname{Var}(W)\operatorname{Cov}(\mathbf{Y}^{\boldsymbol{\alpha}}, \mathbf{Z}^{\boldsymbol{\beta}}) = \operatorname{Cov}(\mathbf{Y}^{\boldsymbol{\alpha}}, W)\operatorname{Cov}(W, \mathbf{Z}^{\boldsymbol{\beta}})$$

for all binary vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$. Note that these sets of equations are typically not independent of each other.

The equations (2.7) characterize the family of distribution on $(\mathbf{Y}, \mathbf{Z}, W)$ for which $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|W$ by determining explicitly the constraints imposed on the probability space of $\mathbf{Y}, \mathbf{Z}, W$ by the conditional independence statement. Even when $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are binary vectors, the dimension of the feasible probability space is not always immediate, but can be determined by solving the set of central moments equations (2.7).

3. A result for the estimation of hidden variables on directed trees. In this section we shall examine those moment relationships that describe the parameter space of directed graphical models whose conditional independence assumptions are represented via directed trees and when some variables are hidden. In order to do this we need first to introduce some terminology and results for directed graphical models that will be used throughout the paper.

3.1. *Graphical models and the curved exponential family.* A directed acyclic graph (DAG) $\mathcal{G}(V, E)$ consists of a set of nodes V and a set of directed edges or arrows E, that link ordered pairs of distinct nodes in V; thus if $v_i \to v_j$ for $v_i, v_j \in V$, then the edge $e(v_i, v_j)$ is in E. The graph does not contain any *directed cycle*, that is there is no sequence of nodes $v_1, \ldots, v_k$ in V such that $e(v_i, v_{i+1})$ for $i = 1, \ldots, k-1$ and $e(v_k, v_1)$ are in E. Two nodes v_i, v_j are *neighbors* or *adjacent* if they are linked by an edge, that is if $e(v_i, v_j)$ or $e(v_j, v_i)$ are in E.

A graph $\widetilde{\mathcal{G}}(V, E)$ is called *undirected* if the edges in E are undirected or lines, so for any edge $e(v_i, v_j)$ in E, also $e(v_j, v_i)$ is in E. The *undirected version* $\widetilde{\mathcal{G}}(V, \widetilde{E})$ of a DAG $\mathcal{G}(V, E)$ is the undirected graph obtained from $\mathcal{G}(V, E)$, by replacing the arrows with undirected edges.

Let $\{Y_1, \ldots, Y_n\}$ be a set of random variables with joint distribution $p(Y_1, \ldots, Y_n)$. A DAG $\mathcal{G}(V, E)$ whose nodes are the random variables $Y_1, \ldots, Y_n$ can be assumed to represent the interdependencies among $Y_1, \ldots, Y_n$. For convenience, in the rest of the paper, the terms "nodes" and "variables" will be used interchangeably. The node Y_i is said a *parent* of Y_j if $e(Y_i, Y_j) \in E$ and we let $Pa(Y_j)$ denote the *parent set* of Y_j, that is, the set of all its parents. A *path* between two nodes Y_i and Y_j in V is a sequence of distinct vertices $\{v_1, \ldots, v_r\}$ in V such that $v_1 = Y_i$ and $v_r = Y_j$ and $e(v_k, v_{k+1}) \in E$ for all

$k = 1, \ldots, r-1$. The *ancestral set* $An(Y_i)$ of a node Y_i in V is defined as the set of all the nodes Y_j, $j \neq i$ in V such that there exists a path from Y_j to Y_i.

Thus a directed acyclic graphical (DAG) model with graph $\mathcal{G}(V, E)$ defines the set of probability distributions over the sample space of $\{Y_1, \ldots, Y_n\}$ that obey the *directed Markov property* expressed by

$$Y_i \perp\!\!\!\perp An(Y_i) | Pa(Y_i)$$

for each variable Y_i in V with respect to the DAG $\mathcal{G}(V, E)$ [Lauritzen (1996), page 50]. Equivalently, the probability distributions on the sample space of $\{Y_1, \ldots, Y_n\}$ specified by a DAG model with graph $\mathcal{G}(V, E)$ factor as

$$p(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} p(Y_i | Pa(Y_i))$$

accordingly to the graph $\mathcal{G}(V, E)$. A probability distribution defined as above is said to be Markov with respect to the graph $\mathcal{G}(V, E)$.

For the purpose of this paper it is useful to formulate the following definition.

DEFINITION 3.1. The distribution of $\{Y_1, \ldots, Y_n\}$ is connected, if no variable $Y_i, i = 1, \ldots, n$ is independent of the rest.

Obviously the results of this section can be applied in a straightforward way to each connected subvector of $\{Y_1, \ldots, Y_n\}$, if the distribution of $\{Y_1, \ldots, Y_n\}$ is not connected.

In this paper the attention is focused on a subclass of DAG models that is defined as follows.

DEFINITION 3.2. A directed tree $\mathcal{T}(V, E)$ is a DAG with edges E and nodes $V = \{Y_1, \ldots, Y_n\}$ such that each node in V has exactly one parent, except one node, called *root*, which has none. The nodes with no children are called *terminal*. A tree model with respect to $\mathcal{T}(V, E)$ is the set of probabilities that factor accordingly to the tree $\mathcal{T}(V, E)$.

We shall call a tree model with graph $\mathcal{T}(V, E)$ *binary* if all the variables $Y_i, i = 1, \ldots, n$ in V are binary; that is, Y_i can only take two values, coded by -1 or 1.

Given three distinct subsets $A, B, C \subset V, C$ is said to *separate* A from B in an undirected graph if for any node $v_A \in A$ and $v_B \in B$ there exists a path between v_A and v_B intersecting C. Given a DAG $\mathcal{G}(V, E)$, if A, B are subsets of vertices in V, such that $V = A \cup B, C = A \cap B$ separates B from A in the undirected version of $\mathcal{G}(V, E)$, and $e(v, w) \in E$ for any two nodes $v, w \in C$, then the subgraphs $\mathcal{G}(A, E_A)$ and $\mathcal{G}(B, E_B)$ form a *decomposition* of $\mathcal{G}(V, E)$.

The separation property is very useful and enables us to read easily off conditional independence assumptions which are coded in directed trees, as stated in the following theorem.

THEOREM 3.1 (Separation for directed trees). *Let the probability distribution* $p(Y_1, \ldots, Y_n)$ *be Markov with respect to the directed tree* $\mathcal{T}(V, E)$ *with nodes* $V = \{Y_1, \ldots, Y_n\}$ *and edges* E. *If* $A, B, C \subset V$ *are subsets of nodes in* $\mathcal{T}(V, E)$, *then*

$$A \perp\!\!\!\perp B | C$$

whenever C *separates* A *from* B *in the undirected version of* $\mathcal{T}(V, E)$.

For the proof, see, for instance, Lauritzen [(1996), page 47].

A directed tree model will be named a *nest* if, in the associated directed tree, the root node and all the terminal nodes correspond to observed variables, and the remaining nodes are hidden. An example of such a tree is displayed in Figure 1. Nests are particularly interesting directed tree models. Firstly by the separation theorem 3.1 we notice that no conditional independence statements between subsets of manifest variables can be deduced in general. Secondly it is simple to transform an arbitrary tree model with a given set of hidden and manifest variables into a Markov equivalent tree model which is formed by a set of nests. Thus in this sense a *nest* provides a building block for the analysis of any tree model with hidden variables. Call a tree model *triadic*, if it is a nest with respect to a graph in which each hidden variable has one parent and two children (see the example displayed in Figure 1). An important class of directed graphical models which usually have a triadic structure is used in phylogenetic applications [see, e.g., Swofford, Olsen, Waddell and Hillis (1996)].

Throughout the paper we shall denote with $\mathcal{H} = \{H_1, \ldots, H_m\}$ a set of *hidden* or *latent* variables, which are not observed and we assume $\mathrm{Var}(H_i) > 0$ for $1 \le i \le m$, that implies that no hidden variable has a degenerate distribution. The set of variables $\mathcal{X} = (X_1, \ldots, X_n)$ that are assumed to be observed will be called *manifest* or *observed* variables.

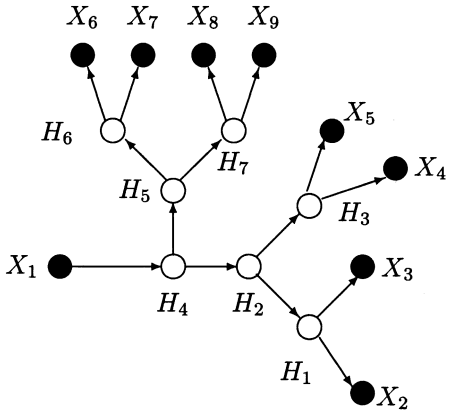

FIG. 1. *A triadic nest.*

It has recently been shown that multinomial directed graphical models define a family of probability distributions over the finite set of variables V that is curved exponential [Geiger and Meek (1998)]. In particular directed tree models being Markov equivalent to their undirected version define a linear exponential family of probability distributions [Lauritzen (1996), Chapter 4]. Recall that a linear or regular exponential family $\mathscr{F}$ is defined as the set of probability distributions of the form

$$f(y, \theta) = \exp\{\langle \theta, t(y)\rangle - \psi(\theta)\},$$

where t is the *canonical statistic* defined on a sample space $\mathscr{Y} \in \mathbb{R}^n$, taking values in a real Euclidean vector space $\mathscr{T} \in \mathbb{R}^k$ with inner product $\langle \cdot \cdot \rangle$. The natural parameter space is an open set given by

$$\Theta = \left\{ \theta \in \mathscr{T} | \int \exp(\langle \theta, t(y)\rangle)\, \nu(dy) < \infty \right\}.$$

The *dimension* of an exponential family, called the *order*, is the dimension of the natural parameter space Θ when the family is in its minimal form, that is, when $f(x, \theta)$ cannot be represented with a parameter vector θ' of dimension smaller than k. We say that a curved exponential family $\mathscr{F}_C$ of dimension h is an embedded subfamily of $\mathscr{F}$ if and only if its natural parameter space Θ_C is a smooth manifold of dimension h embedded in Θ.

When some variables are hidden, the sample space on the manifest variables $\mathscr{X}$ of a directed graphical model can be defined as the mapping of the parameter space Θ onto the marginal probability space Θ_X associated to the observed random variables and we say that there exists a map ν: $\Theta \to \Theta_X$. For most points $\theta_{x_0} \in \Theta_X$, the preimages $\mathscr{I}_{x_0} = \nu^{-1}(\theta_{x_0})$ are manifolds of the same dimension given by

$$\dim(\mathscr{I}_{x_0}) = \dim(\Theta) - \dim(\Theta_X), \tag{3.1}$$

[see, e.g., Hartshorne (1977), Chapter 10]. For a certain class of directed graphical models with one hidden variable, Geiger, Heckerman, King and Meek (1998) show that such a mapping is not a smooth manifold and therefore such models are not curved exponential models.

We shall say that a subspace Θ_H of the parameter space Θ associated to a directed graphical model is unidentifiable for any data $\mathbf{x}$ on the observed variables $\mathscr{X}$ if for all parameters θ and θ' in Θ_H, the joint probability distribution over the directed graphical model is such that $p(\mathbf{x}|\theta) = p(\mathbf{x}|\theta')$. In statistical terminology this is commonly known as global unidentifiability and is a stronger definition than the one used in latent variable models analysis where it is typically assumed that a parameter θ in Θ is *locally identifiable* if for any θ' in a neighborhood of θ $p(\mathbf{x}|\theta) \neq p(\mathbf{x}|\theta')$ for all values $\mathbf{x}$.

The dimension of the unidentifiable space is defined as $\dim(\mathscr{I}_{x_0})$ calculated in (3.1), that is, the dimension of the preimage at smooth points θ_{x_0} of the map $\nu(\cdot)$. Heuristically, this means that within the space defined by the natural parameterization of the exponential family for a directed tree model, there exists an open ball, whose dimension is the dimension of the unidentifiable

space, in which all combinations of parameter values give the same likelihood function for all datasets $\mathbf{x}$.

3.2. *Results on the identifiability of binary trees.* In this section we shall explore the identifiability issues for binary tree models with hidden variables by using the algebraic constraints on the probability distribution of the manifest variables induced by the conditional independence models as described in Section 2.

Because of the local nature of conditional independence in tree models, a triadic nest forms the building block to develop an understanding of the statistical properties of more complicated tree models with hidden variables. In particular it allows us to determine the dimension of the space of the unidentifiable parameters when (1) the random variables are binary, (2) the joint probability distribution obeys the Markov properties with respect to a directed tree and (3) the marginal distribution of its observed variables $\mathscr{X}$ is known. In practice, of course, the distribution on $\mathscr{X}$ will be estimated from the observed marginal counts. If $\mathscr{D}$ is a random sample of the vector $\mathscr{X}$, then clearly, in the notation of Section 3.1, by definition $\mathscr{D} \perp\!\!\!\perp \Theta | \Theta_X$. In particular this means that the observed data $\mathscr{D}$ are informative about the parameters of Θ only through what they tell us about the marginal parameter space Θ_X. One important consequence is that this will bound what we could expect to learn about the set of hidden variables $\mathscr{H} = \{H_1, \ldots, H_m\}$ from a random sample of the observed population $\{X_1, \ldots, X_n\}$. In particular for a Bayesian model it will tell us which features of the prior distribution over $(\mathscr{X}, \mathscr{H})$ will endure after extensive sampling on $\mathscr{X}$ or, in other words, to which features of the prior specification the model will be particularly sensitive.

LEMMA 3.2. *For any binary triadic model with tree $\mathscr{T}(V, E)$ and variables $V = \mathscr{H} \cup \mathscr{X}$, if the distribution of the manifest variables $\mathscr{X}$ is connected then we can calculate*:

(i) $|\mathrm{Cov}(H_i, H_j)|$ *for all hidden variables such that* $e(H_i, H_j) \in E,\ 1 \le i \ne j \le m$.

(ii) $\mathrm{Var}(H_j),\ 1 \le j \le m$.

(iii) $|\mathrm{Cov}(X_i, H_j)|$ *for all adjacent variables in* V *such that* $e(H_i, H_j) \in E,\ 1 \le i \le n,\ 1 \le j \le m$.

For the proof, see the Appendix.

The result in Lemma 3.2 can be generalized to directed graphical models that assume conditional independence statements represented by more complicated tree structures as described in the following theorem.

THEOREM 3.3. *For any binary directed nest with tree $\mathscr{T}(V, E)$ and variables $V = \mathscr{H} \cup \mathscr{X}$, if the distribution of the manifest variables $\mathscr{X}$ is connected*

and all hidden variables have at least three neighbors, then we can calculate:

(i) $|\mathrm{Cov}(H_i, H_j)|$ *for all hidden variables such that* $e(H_i, H_j) \in E$, $1 \leq i \neq j \leq m$.

(ii) $\mathrm{Var}(H_j)$, $1 \leq j \leq m$.

(iii) $|\mathrm{Cov}(X_i, H_j)|$ *for all adjacent variables in* $\mathcal{T}(V, E)$ *such that* $e(X_i, H_j) \in E$, $1 \leq i \leq n$, $1 \leq j \leq m$.

PROOF. For any directed edges connecting a pair of neighbors $e(H_i, H_j)$ or $e(H_j, X_l)$ in $\mathcal{T}(V, E)$, where H_i, H_j are hidden and X_l is manifest, there is a subtree of $\mathcal{T}(V, E)$, say $\mathcal{T}[H_i, H_j]$ or $\mathcal{T}(H_j, H_l)$, which is triadic. To construct such a tree first find the unique path from the root node to either X_l or H_i and delete all but two children of hidden nodes sequentially together with their ancestors, always ensuring no edge on this path are deleted.

From the separation theorem 3.1, the family of probability distributions, which are Markov with respect of this subtree, specifies a triadic model over its variables. The result now follows from Lemma 3.2. □

It follows that the unidentifiability will depend solely on the aliasing and the unidentifiable space is zero-dimensional as proved in the Appendix. Because different solutions correspond to sign changes on the hidden variables there are only finitely many of them and so they are topologically separated. So in common terminology [see Goodman (1974a)] we can assert that such models are *locally identifiable* given sufficient data. We are then in a relatively favorable situation where we can overcome the identifiability problem in our model by eliciting a prior distribution for the parameters that assigns probability 1 to a particular ordering on the probabilities of the hidden variables, for example by demanding that $P(H_i = 1) > 1/2$. Note however that most of the standard directed graphical models do not assume this. Indeed, for reasons of interpretation, such arbitrary restrictions may well be unrealistic. When this is the case our data gives us no additional information with respect to the prior settings on where the best explanation between these aliasing alternatives lies.

The following result provides a method to calculate the dimension of the unidentifiable space for triadic models which are nests. We shall assume that the data arises from a probability distribution on the observed variables $\mathcal{X}$, that is a smooth point in Θ_X. In general the singular points in the parameter space Θ_X correspond either to zeros in the probability tables on $\mathcal{X}$ or to degenerate distributions on $\mathcal{H} \cup \mathcal{X}$, such that $X_i = H_j$ for some i, j, with $1 \leq i \leq n$ and $1 \leq j \leq m$. An example of this is shown, for instance, in the geometrical analysis of the triadic model in Section 4.1. Hence the following results shall demand that the variables $\mathcal{X}$ are connected and that the joint probability table associated to $\mathcal{H} \cup \mathcal{X}$ has nonzero entries.

THEOREM 3.4. *For any binary nest with tree* $\mathcal{T}(V, E)$ *and variables* $V = \mathcal{H} \cup \mathcal{X}$, *the number of the unidentifiable parameters in the joint parameter space over* $(\mathcal{X}, \mathcal{H})$ *is* $\delta(\mathcal{T}) = 2k$, *where* k *is the number of hidden variables in* $\mathcal{T}(V, E)$ *with exactly two neighbors*.

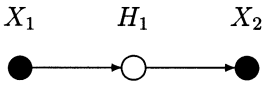

FIG. 2. *A simple nest in three variables.*

For the proof see the Appendix.

Theorem 3.4 embeds the known case of the simplest nest with two manifest variables and one hidden node, displayed in Figure 2, which has a two-dimensional unidentifiable space as proved in Gilula (1979). The unidentifiable space corresponds to all the possible binary hidden variables H_1 that are consistent with the observed margins (X_1, X_2); see, for instance, Settimi and Smith (1998) for a discussion on the geometry of the parameter space of this particular model.

We can extend the result in Theorem 3.4 to directed graphical models with conditional independence assumptions corresponding to more general tree structures as stated in the theorem below.

DEFINITION 3.3. A hidden variable in a tree $\mathscr{T}(V, E)$ is called *bounded* if it lies on a path between two manifest variables in $\mathscr{T}(V, E)$, and *unbounded* otherwise.

Let $\mathscr{H}_{B,2}$ denote the set of bounded hidden variables with two neighbors and $\mathscr{H}_U$ be the set of unbounded hidden variables in V. We write $N(\mathscr{H}_{B,2})$ and $N(\mathscr{H}_U)$ to denote the number of variables in $\mathscr{H}_{B,2}$ and $\mathscr{H}_U$, respectively. For example, in Figure 3 the node H_1 is bounded, H_2 is unbounded and H_3 is bounded with two neighbors.

THEOREM 3.5. *For a binary tree model with tree $\mathscr{T}(V, E)$ and variables $V = \mathscr{H} \cup \mathscr{X}$, if the distribution of the manifest variables $\mathscr{X}$ is connected, then the number of unidentified parameters in the joint parameter space over $(\mathscr{X}, \mathscr{H})$ is $\delta(\mathscr{T})$ where*

$$\delta(\mathscr{T}) = 2[N(\mathscr{H}_{B,2}) + N(\mathscr{H}_U)].$$

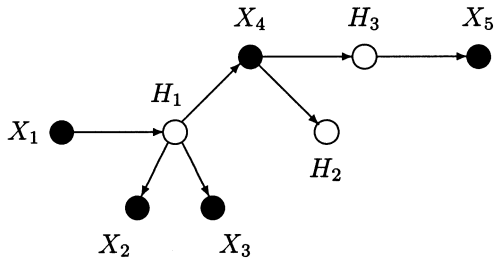

FIG. 3. *A directed tree with bounded and unbounded hidden variables.*

For the proof, see the Appendix.

Despite the fact that we can calculate the dimension of the unidentified parameter space $p(\mathscr{H}|\mathscr{X})$ in general, unless all hidden variables are unbounded, the space itself can be very complicated. Even when this space has zero dimension it will be subject to "aliasing" because alternative admissible solutions can be obtained by switching the sign of any bounded H_j giving 2^K, $K = N(\mathscr{H}_{B,2})$, separate equally likely solutions to the problem.

4. Some results on the geometry of the observed margins. Recently there has been an increasing interest in the nature and in particular in the dimension of the observed space of the manifest variables in a directed graphical model with hidden variables. The feasible region specified by the marginal distribution of directed graphical models with hidden variables is generally very complicated whenever that graphical model contains a nest.

The regularity conditions concerning asymptotic results we might want to use are typically broken, see Geiger, Heckerman, King and Meek (1998) for a discussion of related issues. Instead of studying the local geometry of the space, as these authors do, we try to obtain an insight into the nature of the more global features of its geometry, focusing our attention on binary tree models with *isolated* hidden variables, which correspond to directed trees in which each hidden variable has its parent and its children that are observed. In particular notice that the submodels relative to the subgraphs containing a hidden variable are nests. Hence any probability distribution which is Markov with respect to a tree with isolated hidden variables can be factored into the product of probability distributions, that are defined by submodels with at most one hidden variable. This is shown in the following example.

EXAMPLE (Tree model with isolated hidden variables). The tree model in Figure 4 has isolated hidden nodes H_1, H_2, H_3. The corresponding tree can be decomposed into the subgraphs $\mathscr{G}_i(V_i, E_i)$, $i = 1, \ldots, 4$ with set of nodes $V_1 = \{X_1, X_2, X_3, X_4, H_1\}$, $V_2 = \{X_4, H_2, X_8\}$, $V_3 = \{X_5, H_3, X_6, X_7\}$ and $V_4 = \{X_4, X_5\}$. Notice that $G_1(V_1, E_1), G_2(V_2, E_2), G_3(V_3, E_3)$ correspond to nest models. The joint probability distribution defined by this tree model can be factored as

$$p(X_1, \ldots, X_8, H_1, H_2, H_3) = g_1(X_1, X_2, X_3, X_4, H_1)\, g_2(X_4, X_8, H_2)\, g_3(X_5, X_6, X_7, H_2)\, g_4(X_4, X_5),$$

where each function $g_i(\cdot)$, for $1 \leq i \leq 4$, belongs to the family of probability distributions defined by the submodels associated to the subgraphs $G_i(V_i, E_i)$, $i = 1, \ldots, 4$.

Thus, the parameter space associated to the manifest variables of a binary tree model with isolated hidden nodes can be analyzed by first decomposing the tree, representing the conditional independence assumptions of the model, into the set of subgraphs containing at most one hidden variables and then

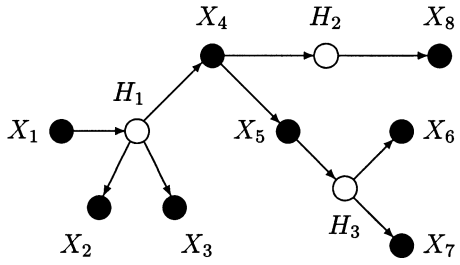

FIG. 4. *A directed tree with isolated hidden variables.*

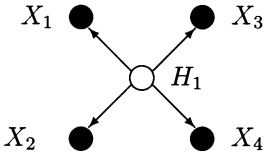

FIG. 5. *A directed tree with one hidden variable.*

exploring separately the marginal subspaces associated to the marginal distributions of the manifest variables in the corresponding nest models.

The geometry of the parameter space for a nest submodel with two observed variables is studied in detail in Settimi and Smith (1998) and the case of a triadic nest with three observed variables is analyzed in Settimi and Smith (1999). The following section presents the geometry of the projection onto the marginal space Θ_X of nest model with $n = 4$ manifest variables.

4.1. *Nest with four observed variables.* Let us suppose that four variables $X_1,\ X_2,\ X_3,\ X_4$ are observed and that the interrelationships among them are represented by a hidden variable model with directed tree structure displayed in Figure 5. Such a graphical model implies the conditional independence assumption $\perp\!\!\!\perp_{i=1}^{4} X_i | H_1$, that is X_1, X_2, X_3, X_4 are all conditional independent given H_1. The projection of such a model onto the marginal probability space of the observed variables $X_1, \ldots, X_4$ is analyzed by examining the algebraic constraints characterizing the family of marginal distributions associated to $X_1, \ldots, X_4$ which are consistent with the model in Figure 5.

The additive model of the probability distribution of $X_1, \ldots, X_4$ can be expressed by

$$
\begin{aligned}
&p(X_1 = i_1,\ X_2 = i_2,\ X_3 = i_3,\ X_4 = i_4) \\
&\quad = \tfrac{1}{16}\Bigg[\prod_{j=1}^{4}(1 + i_j\mu_j) + \sum_{1\le j<k\le 4} i_j i_k \mu_{jk} \\
&\qquad + \sum_{1\le j<k<h\le 4} i_j i_k \lambda_{jkh} + i_1 i_2 i_3 i_4 \lambda_{1234} \Bigg]
\end{aligned} \tag{4.1}
$$

for $i_l \in \{-1, 1\}$ with $l = 1, \ldots, 4$ where $\mu_{ijk} := \mathbb{E}((X_j - \mu_j)(X_k - \mu_k)(X_h - \mu_h))$ and $\mu_{1234} := \mathbb{E}(\prod_{i=1}^{4}(X_i - \mu_i))$. The usual constraints are imposed on the moments defining the reparametrization in (4.1) because of the coherence requirements that probabilities must lie in the simplex. Well-known algorithms in linear programming provide the boundaries on these moments for fixed means [Chvátal (1983), page 240].

Note that we can express $\lambda_{jkh} := \mathbb{E}(X_j X_k X_h)$ for $1 \le j < k < h \le 4$ and $\lambda_{1234} := \mathbb{E}(X_1 X_2 X_3 X_4)$ in terms of the central moments, by using the Bahadur expansion for high-order probability distributions [Streitberg (1990)], that is,

$$\begin{aligned}
\lambda_{jkh} &= \mu_{jkh} + \mu_j \mu_k \mu_h + \sum_{j \neq k < h} \mu_j \mu_{kh}, \\
\lambda_{1234} &= \mu_{1234} + \mu_1 \mu_2 \mu_3 \mu_4 + \sum_{j \neq k < h < s} \mu_j \mu_{khs} \\
&\quad + \sum_{j < k \neq h < s} \mu_{jk} \mu_{hs} + \sum_{j \neq k \neq h < s} \mu_j \mu_k \mu_{hs}
\end{aligned} \tag{4.2}$$

The conditional independence model can be expressed in terms of the central moments as

$$\begin{aligned}
&\mu_{jk} = (1 - \mu_H^2)\eta_j \eta_k, \qquad 1 \le j < k \le 4, \\
&\mu_{jkh} = -2\mu_H(1 - \mu_H^2)\eta_j \eta_k \eta_h, \qquad 1 \le j < k < h \le 4, \\
&\mu_{1234} = 2(1 - \mu_H^2)(3\mu_H^2 - 1)\eta_1 \eta_2 \eta_3 \eta_4
\end{aligned} \tag{4.3}$$

for $\eta_j = \mathrm{Cov}(X_j, H_1)/\mathrm{Var}(H_1)$.

It can be shown that projecting the parameter space Θ of the tree model onto the space of the sample distributions defines a feasible region within the sample space that can be expressed as

$$\begin{aligned}
&\mu_{jkh}^2(1 - \mu_H^2) = 4\mu_H^2 \mu_{jk} \mu_{jh} \mu_{kh} \qquad \text{for } 1 \le j < k < h \le 4, \\
&\mu_{1234}^3(1 - \mu_H^2) = 8(3\mu_H^2 - 1)^3 \textstyle\prod_{j \neq k = 1}^{4} \mu_{jk}.
\end{aligned} \tag{4.4}$$

These equations give the condition

$$2\mu_{1234}^3 \mu_H^4 (1 - \mu_H^2) = (3\mu_H^2 - 1)^3 |\mu_{123} \mu_{124} \mu_{134} \mu_{234}|$$

Notice that the solutions of the set of equations (4.3) must satisfy the following inequalities:

$$\begin{aligned}
&0 \le \tau_{jk} = \sqrt{4\mu_{jk}\mu_{jh}\mu_{kh} + \mu_{jkh^2}} \Big/ (2|\mu_{jk}|) \le 1, \\
&0 \le \tau_{jh} = \sqrt{4\mu_{jk}\mu_{jh}\mu_{kh} + \mu_{jkh^2}} \Big/ (2|\mu_{jh}|) \le 1 \quad \text{for } 1 \le j < k < h \le 4, \\
&0 \le \tau_{kh} = \sqrt{4\mu_{jk}\mu_{jh}\mu_{kh} + \mu_{jkh^2}} \Big/ (2|\mu_{kh}|) \le 1.
\end{aligned} \tag{4.5}$$

The quantities τ_{jk} for $1 \le j < k \le 4$ provide useful and simple test statistics to check whether the data are compatible with the conditional independence

assumptions implied by the model. In Section 5 we show an application of these statistics through an example.

Hence from the system of equations (4.4) we can deduce the set of restrictions which are imposed by the conditional independence tree model of Figure 5 on the parameters of the marginal distribution

$$
\begin{aligned}
\mu_{12}\mu_{34} &= \mu_{14}\mu_{23},\\
\mu_{13}\mu_{24} &= \mu_{14}\mu_{23},\\
\mu_{123}\mu_{14} &= \mu_{124}\mu_{13},\\
\mu_{123}\mu_{14} &= \mu_{134}\mu_{12},\\
\mu_{124}\mu_{23} &= \mu_{234}\mu_{12},\\
\mu_{124}^2\mu_{134}^2 &= (\mu_{1234}\mu_{124}\mu_{134} + 2\mu_{123}\mu_{234}\mu_{14}^2)\mu_{14}.
\end{aligned} \tag{4.6}
$$

Incidentally we notice that these relationships suggest a reparametrization of the sample space in terms of the sample means μ_i, $i = 1, \ldots, 4$ and a subset of the central moments, for example considering the parameters μ_{jk}, μ_{hs} for $j \neq k \neq h \neq s$ and μ_{jhk} for $2 \leq h < k \leq 4$.

From the first equation of (4.4) the admissible region in the parameter space Θ_X is defined by the second-order moments that satisfy the sign conditions $\mu_{jk}\mu_{jh}\mu_{kh} > 0$ for each j, k, h.

We can also write down boundary constraints on the sample distributions of all triplets X_j, X_k, X_h given by

$$
2|\mu_{jk}| \geq |\mu_{jh}||\mu_{hk}| + \sqrt{\mu_{jh}^2\mu_{hk}^2 + \mu_{jhk}^2} \qquad \text{for } |\mu_{jk}| \leq |\mu_{jh}||\mu_{hk}|, \tag{4.7}
$$

for $1 \leq j < k \neq h \leq 4$.

Because of the sign constraints above, the region in the second-order moments given in (4.7) is defined as the union of four symmetrical nonintersecting regions contained in the four quadrants in which those sign conditions hold. Figure 6 displays one of these regions in the positive quadrant relatively to the triplet X_1, X_2, X_3 for $\mu_1 = 0.3,\ \mu_2 = 0.2,\ \mu_3 = 0.1,\ \mu_{123} = 0.1$.

Further constraints on the higher order moments can be derived. It can be easily seen that the third-order moments, satisfying the system of equations (4.4), must lie within the region defined by

$$
\frac{4}{3\sqrt{3}} \geq |\mu_{jkh}| \geq \frac{4(|\mu_{jhs}\mu_{jks} - \mu_{js}\mu_{1234}|)\mu_{hks}^2}{(|\mu_{jhs}\mu_{jks}|)^{1/2}\,(3\mu_{jhs}\mu_{jks} - 2\mu_{js}\mu_{1234}|)^{3/2}}, \tag{4.8}
$$

for $1 \leq j < h < k < s \leq 4$.

The marginal distributions which are consistent with the model in Figure 5 have moments which lie in the nonlinear subspace defined by (4.6), (4.7) and (4.8). Even in this simple case we have that the constraints imposed on the sample space by the conditional independence trees become quite severe, defining restrictions also on the high-order moments. The boundary points of the feasible sample space correspond to zeros in the marginal table of the distributions of X_i, H_1 for $i = 1, \ldots, 4$. Therefore they define degenerate distribution on H_1, X_i for $i = 1, \ldots, 4$. The three-way interaction terms

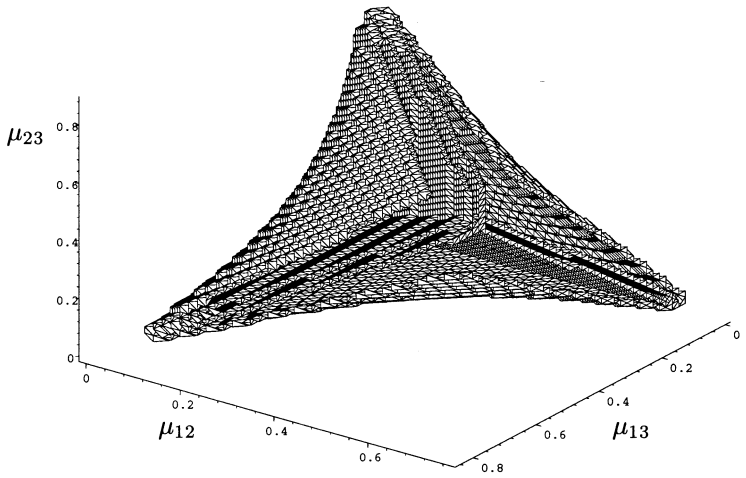

FIG. 6. *The feasible region in the positive quadrant relative to the triplet* X_1, X_2, X_3 *for* $\mu_{123} = 0.3$ *and the mean values* $\mu_1 = 0.3$, $\mu_2 = 0.2$, $\mu_3 = 0.1$.

μ_{jkh} can take values less or equal to $4\sqrt{3}/3$. Notice that the maximum value of μ_{jkh} corresponds to cusp in the feasible region, defined by the boundary point $\mu_{jh} = \mu_{jk} = \mu_{hk} = 2/3, \mu 1234 = 0$, that is, the degenerate distribution $X_j = X_k = X_h = H_1$. Thus the extrema of the feasible region all correspond to models which are in some way degenerate because they set some cell probabilities to zero.

In the statistical literature for latent variable models these degeneracies have been studied for a long time. In particular when data are not consistent with the conditional independence assumptions, a most likely model will have some structural zeroes in the table of the conditional probabilities, implying that certain functional relationships among the variables hold [see, e.g., Goodman (1974a), De Leeuw, van der Heijden and Verboon (1990)]. As a practical consequence, the likelihood will often take its maximum value on a boundary of the parameter space. In discrete directed graphical models with hidden variables, this feature has only recently attracted attention. This focus is timely because many of the logistic or informative Dirichlet priors currently used in practice have densities which tend to zero at their extremes and so obscure these boundary solutions.

4.2. *Extension to nests with n observed variables.* In this section we consider a nest with n observed variables $X_1, \ldots, X_n$ such that $\coprod_{i=1}^{n} X_i | H_1$; that is, all variables are conditionally independent on each other given a binary hidden variable H_1. Such a conditional independence model can be expressed in terms of a set of equations in the central moments over the variables $X_1, \ldots, X_n, H_1$ that will describe the geometrical structure of the parameter space of a nest.

To find such a set of equations, the expressions of the central moments are simplified by exploiting some algebraic results on the classification of complementary set partitions and on the Moebius inversion function that are described in McCullagh [(1987, pages 65 and 251)] and Streitberg (1990). Such results show how to write the kth central moment $\mu_{i_1 \cdots i_k} := \mathbb{E}(\prod_{j=1}^{k}(X_{i_j} - \mu_{i_j}))$, for $1 \le i_j \le n$ and $k = 2, \ldots, n$, as a polynomial function in terms of the kth noncentral moments $\lambda_{i_1 \cdots i_k} := \mathbb{E}(\prod_{j=1}^{k} X_{i_j})$, and of the central moments of order lower than k.

For instance, the expressions of the central moments for $k = 3$ and $k = 4$ can be derived from equations (4.2). The fifth order moment can be expressed as

$$
\begin{aligned}
\mathbb{E}\Big(\prod_{j=1}^{5}(X_j - \mu_j)\Big) = \lambda_{12345} - \mu_1\mu_2\mu_3\mu_4\mu_5 - \sum_{\substack{i=1 \\ i\neq j<k<r<s}}^{5} \mu_i\mu_{jkrs} \\
- \sum_{\substack{1\le i<j\le 5 \\ i,\, j\neq k<r<s}}^{5} \mu_{ij}\mu_{krs} - \sum_{\substack{1\le i<j\le 5 \\ i,\, j\neq k<r<s}}^{5} \mu_i\mu_j\mu_{krs} \\
- \sum_{\substack{1\le i\le 5 \\ i\neq j<k\neq r<s}}^{5} \mu_i\mu_{jk}\mu_{rs} - \sum_{\substack{1\le i<j<k\le 5 \\ i,\, j,\, k\neq r<s}}^{5} \mu_i\mu_j\mu_k\mu_{rs}.
\end{aligned} \tag{4.9}
$$

From the results discussed in Section 2.2 each variable X_i, $1 \le i \le n$ is a linear function of H_1 and can be written as

$$
X_i = \mu_i + \eta_i(H - \mu_H) + \varepsilon_i \qquad \text{for } \mathbb{E}(\varepsilon_i) = \mathbb{E}(\varepsilon_i H) = 0,\ i = 1, \ldots, n, \tag{4.10}
$$

where $\eta_i = \mathrm{Cov}(X_i, H_1)/\mathrm{Var}(H_1)$ for each $i = 1, \ldots, n$.

Hence the noncentral moments of the joint distribution over $X_1, \ldots, X_n$ can be easily found from (4.10) as functions of the central moments of the distribution H_1. Substituting these expressions into the polynomial equations of the central moments, we obtain the set of simultaneous equations describing the parameter space of the given nest.

For instance, the noncentral moments for $n = 5$ are given by

$$
\begin{aligned}
\lambda_{ij} &= \mu_i\mu_j + \eta_i\eta_j(1-\mu_H^2) \quad \text{for } 1\le i<j\le 5, \\
\lambda_{ijk} &= \mu_i\mu_j\mu_k + (\mu_i\eta_j\eta_k + \mu_j\eta_i\eta_k + \mu_k\eta_i\eta_j)(1-\mu_H^2) \\
&\quad + \eta_i\eta_j\eta_k\mathbb{E}(H_1-\mu_H)^3 \quad \text{for } 1\le i<j<k\le 5, \\
&\ \ \vdots \\
\lambda_{12345} &= \mu_1\mu_2\mu_3\mu_4\mu_5 + \textstyle\sum \mu_i\mu_j\mu_k\eta_r\eta_s(1-\mu_H^2) \\
&\quad + \textstyle\sum \mu_i\mu_j\eta_k\eta_r\eta_s(-2\mu_H(1-\mu_H^2)) \\
&\quad + \textstyle\sum \mu_i\eta_j\eta_k\eta_r\eta_s(1+2\mu_H^2-3\mu_H^4)
\end{aligned} \tag{4.11}
$$

$$+4\mu_H(\mu_H^4-1)\eta_1\eta_2\eta_3\eta_4\eta_5$$

sums are for $1\le i\ne j\ne k\ne r\ne s\le 5$.

The first fourth-order central moments are derived from (4.3) and the equation relative to the fifth central moment μ_{12345} is obtained by substituting the equations (4.11) in (4.9). After some algebraic simplifications we can write

$$\begin{aligned}
\mu_{jk} &= (1-\mu_H^2)\eta_j\eta_k, \qquad 1\le j<k\le 5,\\
\mu_{jkr} &= -2\mu_H(1-\mu_H^2)\eta_j\eta_k\eta_r, \qquad 1\le j<k<r\le 5,\\
\mu_{ijkr} &= 2(1-\mu_H^2)(3\mu_H^2-1)\eta_i\eta_j\eta_k\eta_r, \qquad 1\le i<j<k<r\le 5,\\
\mu_{12345} &= 8\mu_H(\mu_H^2-2)(3\mu_H^2-4)\eta_1\eta_2\eta_3\eta_4\eta_5.
\end{aligned}$$

Hence the computation of these polynomial equations for a conditional independence model with n observed variables such that $\coprod_{i=1}^n X_i|H_1$ is straightforward. Notice that computer algebra packages, such as Maple [Char, Geddes, Gonnet, Leong and Monogan (1995)], can be used to calculate automatically such sets of simultaneous equations.

The explicit algebraic characterization of these models allows us to analyze the geometrical structure of the region of the probability distributions over $\mathscr{X}$ that belong to any binary tree model with isolated hidden variables. For instance, the geometrical analyses of nest models with two, three or four observed variables, discussed above, can be applied in an analogous way to nest models with $n\ge 4$ manifest variables.

4.3. *Probability spaces over the manifest variables.* We conclude this section with a result that arises from a generalization of the TETRAD condition [Cox and Wermuth (1996), page 72] and is valid for general trees. This result forms the basis for a simple check on the compatibility of data with directed graphical models that are nests, whenever we have extensive data on the manifest variables $\mathscr{X}$ so that we can reasonably assume that the estimates of the marginal probabilities over $\mathscr{X}$ can be calculated up to a negligible sample error. In fact for large datasets we assume that these marginal probabilities can be substituted by their sample proportions without loss.

For an arbitrary random vector $\mathbf{Y}=(Y_1,\dots,Y_n)$, define the matrix $S(\mathbf{Y})=\{s_{i,j}\}_{i,j}$ where

$$s_{i,j}=\begin{cases} 1, & \text{if } \operatorname{Cov}(Y_i,Y_j)>0,\\ 0, & \text{if } \operatorname{Cov}(Y_i,Y_j)=0 \text{ for } 1\le i,\ j\le n,\\ -1, & \text{if } \operatorname{Cov}(Y_i,Y_j)<0.\end{cases}$$

THEOREM 4.1. *For any nest with tree $\mathscr{T}(V,E)$ and variables $V=\mathscr{X}\cup\mathscr{H}$, if the distribution of the manifest variables $\mathscr{X}$ is connected, then the rank of $S(\mathbf{X})$ is* 1, *where $\mathbf{X}$ is the random vector $\mathbf{X}=(X_1,\dots,X_n)$.*

For the proof, see the Appendix.

REMARK. Although this is just one of the constraints on the covariance matrix of the terminal and root node of a conditional independence structure on binary variables with m hidden interior nodes, it does give a simple initial check to see whether a sample might be described in this way. It appears that we can get anything of rank up to n, $n \geq 3$. Thus if $n = 2$, then $S(X) = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ or $S(X) = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ both of which are of rank 1, so we have no restrictions. However for $n = 3$ we can have matrices that are not acceptable, that is of rank greater than 1, such as, for instance,

$$\begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

5. Toward simple diagnostics on trees. An important motivation for our study is that commonly used methods for model selection cannot be applied to directed graphical models with hidden variables. Indeed we have seen that the projection of the model space Θ onto the marginal space of the observed variables defines a region with singular points on the boundaries that correspond to degenerate distribution on the manifest variables. The existence of such singularities implies that the usual regularity conditions, that justify the use of asymptotic tests such as χ^2 test, likelihood ratio test and Laplace approximations yielding the Bayesian information criterion, are violated. Rather than develop general criterial which do not necessarily provide an explanation on how a possible model is inappropriate, we shall propose a small number of diagnostic statistics based on the values of sample moments. This sort of diagnostics has a long tradition in statistics; for instance, the well-known TETRAD condition on the second-order moments can be used to check the conditional independence assumption between two variables [Cox and Wermuth (1996), page 72].

The most obvious way for a Bayesian to perform a diagnostic check of a given model is to compare the value of the observation vector with its predictive probability [see, e.g., Geisser (1993)]. Although we throughly recommend such methods, it is also true that, because of the complexity of directed graphical models, in common practice the prior probabilities are typically chosen to be in common families with assumptions of independence made for no other reason than convenience, for example, local and global independence. Other safeguards are therefore helpful. From the analysis above we suggest that a possible additional diagnostic is to check whether the posterior probabilities of the hidden variables in a Bayesian tree network appear to be close to zero or one. This will happen if the likelihood takes maximum values on the boundary of the parameter space. If this is so then we might suspect that the given model does not explain adequately the observed data. Identifying where these extreme values occur helps to indicate which aspects of the embedded conditional independence assumptions might be suspect. Two other features of tree models are worth checking routinely. When we have reasonably extensive data

we should examine whether:

1. The sample distribution on the manifest variables satisfies the sign constraints and the rank condition above;
2. The algebraic constraints on the sample estimates of lower order moments are satisfied, in particular the sign conditions on the second-order moments given in the example of Section 4.1 for any three manifest variables that form a triadic subtree.

THEOREM 5.1. *For any nest with tree* $\mathscr{T}(V, E)$ *and variables* $V = \mathscr{H} \cup \mathscr{X}$, *all triples of manifest variables* $\{X_i, X_j, X_k\}$ *for* $1 \leq i \neq j \neq k \leq n$ *in* $\mathscr{T}(V, E)$ *have probability distributions which define a triadic model*, *with* H *one of the hidden variables in the nest*.

PROOF. From the separation theorem 3.1, the set of probability distributions which are Markov with respect to the directed tree $\mathscr{T}^*(V, E)$ with the same undirected tree as $\mathscr{T}(V, E)$ but with root node X_i belong to the nest with tree $\mathscr{T}(V, E)$. To verify the assertion of the theorem we now just choose a hidden variable H to be the unique hidden node lying on both the paths $\{X_i, X_j\}, \{X_i, X_k\}$ and which is closest to X_i. Hence the separation theorem on $\mathscr{T}^*(V, E)$ allows us to assert that $[\![X_i, X_j, X_k | H$. □

It follows that, with reasonably extensive data, we can expect that all $\binom{n}{3}$ triples of the n manifest variables lie within or close to their corresponding three dimensional regions of the type given in Figure 6 in Section 4.1.

EXAMPLE (Application to a simple dataset). We consider a dataset analyzed by Goodman (1974b) where the responses of 3,398 schoolboys to two interviews about their self-perceived membership in the "leading crowd" are cross-classified with respect to four dichotomous variables X_1, X_2, X_3 and X_4. The variables X_1 and X_2 correspond to questions on self-perceived membership and on the attitude with respect to it at the first interview, respectively, and X_3 and X_4 correspond to the same questions at the second interview. The sample proportions of the responses to the two interviews are displayed in Table 1.

Our geometrical approach is used to explore whether the nest model, such that the manifest variables X_i for $i = 1, \ldots, 4$ are independent of each other conditionally on a binary hidden variable H_1, is supported by such a dataset.

Table 2 shows the sample estimates of the moments for the probability distribution on X_1, X_2, X_3, X_4. Recalling the results in Section 4.1, the sample values of the statistics $\{\tau_{jk}, \tau_{jh}, \tau_{kh}\}$ defined in (4.5) for each triplet, X_j, X_k, X_h for $1 \leq j < k < h \leq 4$ are given by

$$\{\tau_{12}, \tau_{13}, \tau_{23}\} = \{0.8339, 0.1579, 0.6388\},$$
$$\{\tau_{12}, \tau_{14}, \tau_{24}\} = \{0.1569, 0.8364, 0.6475\},$$
$$\{\tau_{13}, \tau_{14}, \tau_{34}\} = \{0.5341, 0.5390, 0.1822\},$$
$$\{\tau_{23}, \tau_{24}, \tau_{34}\} = \{0.1994, 0.5341, 1.2234\}.$$

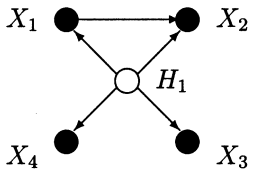

FIG. 7. *A simple directed acyclic graph with one hidden variable.*

Since these values lie outside $[0, 1]$ they suggest that the conditional independence assumption $[\![X_2, X_3, X_4 | H_1$ might be inappropriate to explain the data and confirm the analysis carried out by Goodman (1974b) in his paper, where this conditional independence model is rejected and the data are analyzed by using models with a more complicated hidden structure. More formal Bayesian data analyses for hidden variables models with tree structure will be presented in a later paper where we shall discuss more sophisticated diagnostic methods.

6. Discussion. The extension of our geometrical analysis to more complicated graphical models, defining set of probability distributions which are Markov with respect to directed acyclic graphs that are not trees, is not straightforward. However, there are obvious generalizations to certain models, whose conditional independence assumptions are not represented by directed trees, that can be derived directly from our results, such as, for instance, the simple DAG model represented in Figure 7 that has zero-dimensional unidentifiable space and the class of DAG models which assume a conditional independence structure over the hidden variables that is described via a directed tree.

Further extensions that will be explored in a future paper include nonbinary DAG models. The algorithm of identifying the noncentral moments equations determined on the sample spaces of the different variables and also the conditional independence statements of the graph in terms of central

TABLE 1

Sample proportions of the responses of 3,398 *schoolboys classified according to two interviews about self-perceived membership in "leading crowds"*

			Second interview		
	$\mathbf{X}_3$ membership	+	+	−	−
	$\mathbf{X}_4$ attitude	+	−	+	−
First interview					
X_1 membership	X_2 attitude				
+	+	0.135	0.041	0.032	0.014
+	−	0.050	0.054	0.016	0.026
−	+	0.055	0.022	0.156	0.083
−	−	0.025	0.029	0.099	0.163

TABLE 2
Sample estimates of moments associated to the observed variables X_1, X_2, X_3, X_4

		Covariances			
	Means	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	**Three-way interaction**
X_1	$\hat{\mu}_1 = -0.2625$	0.0976	0.5627	0.0606	$\hat{\mu}_{123} = 0.0295$
X_2	$\hat{\mu}_2 = 0.076$		0.1274	0.2861	$\hat{\mu}_{124} = 0.0082$
X_3	$\hat{\mu}_3 = -0.181$			0.1249	$\hat{\mu}_{134} = 0.0357$
X_4	$\hat{\mu}_4 = 0.1377$				$\hat{\mu}_{234} = -0.2742$
		$\hat{\mu}_{1234} = -0.0865$			

moments equations are still valid. However, there is no longer a binary tree structure on the hidden variables, the equations become much more complicated and it is often necessary to resort to computer algebra techniques, such as Gröbner bases [Cox, Little and O'Shea (1991)]. Interestingly it appears that similar techniques as given in this paper can be used when all variables are Gaussian. Here, however, the sample space restrictions need to be specified in terms of cumulants rather than in terms of noncentral moments.

APPENDIX

A. Proofs of lemmas and theorems.

PROOF OF LEMMA 3.2. Proceed by induction on m. There is only one binary triadic model when $m = 1$, namely the triadic model represented by the tree in Figure 8. It is easy to show that the theorem holds in this case. Such a triadic model implies the conditional independence assumptions $\coprod X_1,\ X_2,\ X_3|H$, which are equivalent to the statements

$$X_i \perp\!\!\!\perp X_j|H_1 \quad \text{for } 1 \le i \ne j \le 3,\ (X_1X_2) \perp\!\!\!\perp X_3|H.$$

By using (2.7) where we set $W = H_1$, $Y = X_i$, $Z = X_j$ for $1 \le i \ne j \le 3$ and $\alpha = 1$, $\beta = 1$ the conditional independence assumptions $X_i \perp\!\!\!\perp X_j|H_1$ can be written as

$$\text{Var}(H_1)\,\text{Cov}(X_i, X_j) = \text{Cov}(X_i, H_1)\,\text{Cov}(X_j, H_1). \tag{A.1}$$

The other relationship $(X_1X_2) \perp\!\!\!\perp X_3|H_1$ can be expressed by

$$\text{Var}(H_1)\,\text{Cov}(X_1X_2, X_3) = \text{Cov}(X_1X_2, H_1)\,\text{Cov}(X_3, H_1). \tag{A.2}$$

By substituting (A.1) into (A.2) and rearranging, equation (A.2) can be replaced by

$$\begin{aligned}&\mathbb{E}((X_1 - \mu_1)(X_2 - \mu_2)(X_3 - \mu_3))\\ &\quad = -2\mathbb{E}(H_1)\,\text{Cov}(X_1, H_1)\,\text{Cov}(X_2, H_1)\,\text{Cov}(X_3, H_1)/\text{Var}(H_1)^2.\end{aligned} \tag{A.3}$$

Hence (i), (ii) and (iii) hold as solutions of the system of equations (A.1) and (A.3) and the theorem is true for $m = 1$. Suppose now that the statement

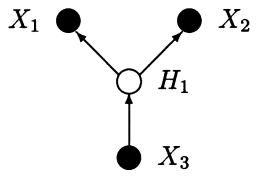

FIG. 8. *A triadic nest with one hidden variable.*

is true for all binary triadic trees with $m = k, k \geq 1$. Consider a binary triadic model with $k+1$ binary hidden variables, corresponding to the tree $\mathcal{T}(V, E)$. Then $\mathcal{T}(V, E)$ will have at least one hidden variable, say H_1, whose node has as its neighbors two manifest variables, say X_1, X_2, and one hidden variable, say H_2. It is easily checked, using the separation theorem 3.1, that the directed tree $\mathcal{T}^*(V, E)$, obtained from $\mathcal{T}(V, E)$ deleting the edges $e(H_2, H_1), e(H_1, X_1), e(H_1, X_2)$ and the nodes H_1 and X_1 and adding the edge $e(H_2, X_2)$ represents a set of conditional independence assumptions that define a triadic model with k hidden nodes. From the inductive hypothesis therefore, we can immediately state that (i), (ii) and (iii) are true if we can prove that $|\mathrm{Cov}(H_1, H_2)|, \mathrm{Var}(H_1), |\mathrm{Cov}(X_1, H_1)|$ and $|\mathrm{Cov}(X_2, H_1)|$ can be evaluated.

Also by the separation theorem 3.1, labelling the root manifest variable X_3 we have that $\coprod_{i=1}^{3} X_i | H_1$. From the example in Section 4.1 we can therefore assert that $\mathrm{Var}(H_1), |\mathrm{Cov}(X_1, H_1)|$ and $|\mathrm{Cov}(X_2, H_1)|$ are determined.

Finally, we can read from the tree using separation that $X_2 \perp\!\!\!\perp H_2 | H_1$ which implies

$$\mathrm{Var}(H_1)\,|\mathrm{Cov}(X_2, H_1)| = |\mathrm{Cov}(X_2, H_1)|\,|\mathrm{Cov}(H_1, H_2)|. \tag{A.4}$$

Note that $|\mathrm{Cov}(X_2, H_1)|$ is calculated above and also nonzero, for otherwise we would have $X_2 \perp\!\!\!\perp H_1$ because X_2 and H_1 are binary.

By separation on $\mathcal{T}(V, E)$ we would then have $X_2 \perp\!\!\!\perp (X_1, X_3, \ldots, X_n)$; this would mean that the distribution of $\mathcal{X}$ is not connected, which is not allowed in the hypotheses. The variance $\mathrm{Var}(H_1)$ is calculated from (A.4). Finally, $|\mathrm{Cov}(X_2, H_2)|$ is determined from the inductive hypothesis applied to $\mathcal{T}^*(V, E)$.

So we have shown that if the hypothesis is true for $m = k$ it is also true for $m = k+1$; thus by induction the theorem holds for all m. □

PROOF OF THEOREM 3.4. First note that if $\delta(\mathcal{T}) = 0$, then Lemma 3.3 proves this result since all second moments of the distribution are determined in modulus and so up to possible sign changes in the hidden variables.

So suppose H_2 is a hidden node in $\mathcal{T}(V, E)$ with exactly one parent, H_1, and one child, H_3.

Let V be the vertex set and E denote the edge set of the tree $\mathcal{T}(V, E)$ and construct a new directed tree $\mathcal{T}'(V', E')$ where

$$V' = V \backslash \{H_2\} \quad \text{and} \quad E' = (E \backslash \{e(H_1, H_2), e(H_2, H_3)\}) \cup \{e(H_1, H_3)\}.$$

It is easy to check that the separation theorem 3.1 implies that the set of conditional independence statements represented in $\mathcal{T}(V, E)$ are equivalent to those coded in the tree $\mathcal{T}'(V, E)$, none of which directly concern the joint distribution of functions of H_2, together with the two further statements

$$H_1 \perp\!\!\!\perp H_3 | H_2, \tag{A.5}$$

$$H_2 \perp\!\!\!\perp V \backslash \{H_1, H_2, H_3\} | H_1, H_3. \tag{A.6}$$

Because all nodes are binary, the second statement (A.6) tells us that, given $[\mathbb{E}(H_1), \mathbb{E}(H_3), \mathrm{Cov}(H_1, H_3)]$, the quantities $\mathbb{E}(H_2), \mathrm{Cov}(H_1, H_2)$ and $\mathrm{Cov}(H_2, H_3)$ are functionally independent of all the other parameters determining the conditional distributions of the hidden variables given the manifest variables. The first equation (A.5) is equivalent to demanding

$$\mathrm{Var}(H_2)\,\mathrm{Cov}(H_1, H_3) = \mathrm{Cov}(H_1, H_2)\,\mathrm{Cov}(H_2, H_3), \tag{A.7}$$

where $\mathrm{Var}(H_2) = 1 - \mathbb{E}(H_2)^2$. Note that $\mathrm{Cov}(H_1, H_3) \neq 0$, for otherwise $H_1 \perp\!\!\!\perp H_2$ and so the distribution of $\mathscr{X}$ would not be connected, contrary to the hypothesis. So (A.7) defines a solution space of dimension 2. By replacing $\mathcal{T}(V, E)$ by $\mathcal{T}'(V, E)$ in the above argument and repeating this argument successively a residual DAG having only hidden variables with at least three neighbors is constructed. Hence the directed graphical model defined by the probabilities that are Markov with respect to such a residual DAG has an unidentifiable space of dimension $2k$, where k is the number of hidden variables with two neighbors as required. □

PROOF OF THEOREM 3.5. A simple application of the separation theorem 3.1 shows that all DAG models which assume conditional independence statements coded via directed trees that have the same undirected version are Markov-equivalent; that is, they induce the same conditional independence restrictions on the joint distribution of $(\mathscr{X}, \mathscr{H})$. So without loss we can assume that the root node is a manifest variable.

Directly from the definition of a DAG model, we have that the implied conditional independence statements in $\mathcal{T}(V, E)$ concerning the conditional distribution $p(\mathscr{H}|\mathscr{X})$ are equivalent to those coded in the directed trees $\{\mathcal{T}_1(V_1, E_1), \ldots, \mathcal{T}_J(V_J, E_J)\}$ where $V = \bigcup_{j=1}^{J} V_j$ and $E = \bigcup_{j=1}^{J} E_j$. The set V_j is defined as $V_j = \mathscr{X}[j] \cup \mathscr{H}[j]$ where $\mathscr{X}[j]$ is a set of manifest variables and $\mathscr{H}[j]$ is a set of hidden variables, together with a set of conditional independence statements over V given below.

Thus we have for $1 \le j \ne j' \le J$,

$$\mathscr{X}[j] \cap \mathscr{X}[j'] = \begin{cases} \varnothing & \text{or,} \\ \{X_{j,j'}\}, \end{cases}$$

where $X_{j,j'}$ denotes a manifest variable which is terminal or root in both $\mathscr{T}_j$ and $\mathscr{T}_{j'}$. Besides,

$$\mathscr{H}[j] \cap \mathscr{H}[j'] = \varnothing \quad \text{and} \quad \bigcup_{j=1}^{J} \mathscr{H}[j] = \mathscr{H}, 1 \le j \ne j' \le J,$$

$$E_j \cap E_{j'} = \varnothing, \qquad 1 \le j \ne j' \le J,$$

where $\mathscr{T}_j, 1 \le j \le J$ is a nest or its vertex set V_j contains exactly one manifest variable, which is its root node, and the other nodes are hidden.

The additional conditional independence statements needed are

$$\mathscr{H}[j] \perp\!\!\!\perp V \backslash V_j | \mathscr{X}[j], \qquad 1 \le j \le J. \tag{A.8}$$

Because of (A.8), to find the dimension of the required unidentifiable space we need only to add the dimension of the unidentifiable space associated with the J models corresponding to these J trees $\mathscr{T}_j(V_j, E_j)$. If $\mathscr{T}_j(V_j, E_j)$ corresponds to a nest model then the dimension of the unidentifiable space associated to this model is given in Theorem 3.4. Otherwise it clearly contains $2N(\mathscr{H}[j])$ undetermined parameters, namely, for each variable $H_i \in \mathscr{H}[j]$ those are the probabilities $p(H_i = 1|\text{parent}(H_i) = 1)$ and $p(H_i = 1|\text{parent}(H_i) = 2)$. Adding over J now gives the result of the theorem. □

PROOF OF THEOREM 4.1. For convenience, define the hidden vector $\mathbf{H} = (H_1, \ldots, H_n)$. First note that the distribution of the manifest variable $\mathbf{X}$ is invariant to sign changes in the hidden variables $\mathscr{H}$. Therefore, by considering the vector form, if $\mathbf{H}' = B\mathbf{H}$ where B is an arbitrary diagonal matrix whose eigenvalues are either -1 or 1, then if $\mathbf{H}'$ is given the same distribution as another $\mathbf{H}$, the margin on $\mathbf{X}$ will be the same in two cases provided that $p(\mathbf{X}|\mathbf{H}) = p(\mathbf{X}|\mathbf{H}')$ under the transform B above.

Beginning with the parent node X_1, change the sign of its hiddden child, say H_1, if and only if $\text{Cov}(H_1, X_1) < 0$. Moving from parent X_i to child H_i of the hidden node, change the sign of H_i to $-H_i$ if and only if $\text{Cov}(X_i, H_i) < 0$. The root node and all the transformed hidden variables $\mathbf{H}'$ will now be positively correlated. Now consider the linear transformation $\mathbf{X}' = A\mathbf{X}$ where $A = \text{diag}\{a_i : 1 \le i \le n\}$ with entries $a_i = (-1)^d, d = 1 + \text{sign}(\text{Cov}(X_i, H_{j(i)})$ where $H_{j(i)}$ is the unique parent of X_i in $\mathscr{T}$. By the string rule and the assumption that $\text{Cov}(X_i, X_j) \ne 0$, we therefore have that $S(\mathbf{X}') = J_n$ where J_n is the $n \times n$ matrix of ones.

In particular it follows that

$$\text{rank}(S(\mathbf{X})) = \text{rank}(S(A^{-1}\mathbf{X}')) = \text{rank}(S(\mathbf{X}')) = \text{rank}(J) = 1. \qquad \square$$

Acknowledgments. We are extremely grateful a referee and an Associate Editor for their helpful comments and suggestions.

REFERENCES

CHAR, B., GEDDES, K., GONNET, G., LEONG, B. and MONOGAN, M. (1995). *MAPLE V Library Reference Manual*. Springer, New York.

CHVATÁL, V. (1983). *Linear Programming*. Freeman, New York.

COWELL, R. G. (1998). Mixture reduction via predictive scores. *Statist. Comput.* **8** 97–103.

COX, D., LITTLE, J. and O'SHEA, D. (1991). *Ideals, Varieties and Algorithms. An Introduction to Computational Algebraic Geometry and Commutative Algebra.* Springer, New York.

COX, D. R. and WERMUTH, N. (1996). *Multivariate Dependencies. Models Analysis and Interpretation*. Chapman and Hall, London.

DE LEEUW, J., VAN DER HEIJDEN, P. G. M. and VERBOON, P. (1990). A latent time-budget model. *Statist. Neerlandica* **44** 1–22.

FELLER, W. (1971). *An Introduction to Probability Theory and its Applications* **2**. Wiley, New York.

GEIGER, D., HECKERMAN, D., KING, H. and MEEK, C. (1998). Stratified exponential families: graphical models and model selection. Technical Report MSR-TR-98-31, Microsoft Research Center, WA.

GEIGER, D., HECKERMAN, D. and MEEK, C. (1996). Asymptotic model selection for directed networks with hidden variables. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence* 283–290. Morgan Kaufmann, San Mateo, CA.

GEIGER, D. and MEEK, C. (1998). Graphical models and exponential families. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* 156–165. Morgan Kaufmann, San Mateo, CA.

GEISSER, S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall, London.

GILULA, Z. (1979). Singular value decomposition of probability matrices: probabilistic aspects of latent dichotomous variables. *Biometrika* **66** 339–344.

GOODMAN, L. A. (1974a). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231.

GOODMAN, L. A. (1974b). The analysis of systems of qualitative variables when some of the variables are unobservable. A modified latent structure approach. *Amer. J. Sociology* **79** 1179–1259.

HARTSHORNE, R. (1977). *Algebraic Geometry*. Springer, New York.

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford Univ. Press.

MADIGAN, D. and York, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **63** 215–232.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, London.

PISTONE, G., RICCOMAGNO, E. and WYNN, H. P. (1999). Gröbner bases and factorisation in discrete probability and Bayes. *Statist. Comput.* (Special issue for the Workshop on Symbolic Computation, CRM, Montreal.)

RAMONI, M. and SEBASTIANI, P. (1997). Learning Bayesian networks from incomplete databases. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* 401–408. Morgan Kaufmann, San Mateo, CA.

SETTIMI, R. and SMITH, J. Q. (1998). On the geometry of Bayesian graphical models with hidden variables. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* 472–479. Morgan Kaufman, San Mateo, CA.

SETTIMI, R. and SMITH, J. Q. (1999). Geometry, moments and Bayesian networks with hidden variables. In *Proceedings of the Seventh International Workshop on Statistics and Artificial Intelligence* 293–298. Morgan Kaufmann, San Mateo, CA.

SPIEGELHALTER, D. J. and COWELL, R. G. (1992). Learning in probabilistic expert systems. In *Bayesian Statistics* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) **4** 447–466. Clarendon, Oxford.

SPIEGELHALTER, D. J., DAWID, A. P. LAURITZEN, S. L. and COWELL, R. G. (1993). Bayesian analysis of expert systems. *Statist. Sci.* **8** 219–282.

SPIRTES, P., RICHARDSON, T. and MEEK, C. (1997). The dimensionality of mixed ancestral graphs. Technical Report CMU-PHIL-83, Dept. Philosophy, Carnegie Mellon Univ.

STREITBERG, B. (1990). Lancaster interactions revisited. *Ann. Statist.* **18** 1878–1885.

SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. J. and HILLIS, D. M. (1996). Phylogenetic inference. In *Molecular Systematics*, 2nd ed. (Hillis, D. M., Moritz, C. and Mable, B. K., eds.) 407–514. Sinauer Assoc., Sunderland, MA.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distribution by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637
E-MAIL: settimi@galton.uchicago.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY, CV4 7AL
UNITED KINGDOM
E-MAIL: j.q.smith@warwick.ac.uk