A comparison of a Gaussian and a binary ancestral graph model

MATHIAS DRTON
University of California, Berkeley

LUIS DAVID GARCIA
Mathematical Sciences Research Institute, Berkeley

Abstract

For an ancestral graph with four vertices, we show that the associated binary model exhibits different properties than the previously studied Gaussian analog. The Gaussian likelihood function may be multimodal, whereas the likelihood function of the binary model is guaranteed to be unimodal. Canonical hidden variable models associated with ancestral graph models impose additional polynomial constraints of non-independence type in the binary case, but no additional constraints in the Gaussian case.

Running title: Gaussian and binary ancestral graph models
Key words: ancestral graph, Bayesian network, hidden variables, graphical model, latent variables

1 Introduction

In graphical modelling, random variables are identified with the vertices of a graph and, via so-called Markov properties, the edges in the graph encode a pattern of probabilistic

independences used to define a statistical model (Lauritzen, 1996). Graphical models and Bayesian networks in particular have found wide-spread application (e.g. Friedman, 2004). In many applications, interpretation of the directed graph found by Bayesian network model selection is desired but complicated by the fact that effects, which appear to be of causal nature, may be induced by hidden (unobserved) variables (Pearl, 2000; Spirtes *et al.*, 2000). Ancestral graphs (Richardson & Spirtes, 2002) generalize the directed acyclic graphs (DAGs) that underlie Bayesian networks and may feature directed, undirected and bi-directed edges. A richer class of graphs, ancestral graphs can encode any conditional independence structure that may arise from a Bayesian network with hidden variables. Hence, they provide means to guard against misinterpretation induced by assuming absence of influential hidden variables (Richardson & Spirtes, 2003). We note that the classes of summary graphs (Cox & Wermuth, 1996) and MC-graphs (Koster, 2002, 1999) achieve the same goal, however contain ancestral graphs as a strict subclass (Richardson & Spirtes, 2002, §9).

For Gaussian ancestral graph models, i.e. all variables follow a joint multivariate normal distribution, Richardson & Spirtes (2002, §8) provide a parameterization and maximum likelihood (ML) estimates can be computed using iterative conditional fitting (Drton & Richardson, 2003, 2004b). For ancestral graph models for discrete random variables, however, statistical methodology has not yet been developed; in particular, there does not yet exist a parameterization of discrete ancestral graph models. As a first step, we study in this paper a particular binary ancestral graph model whose underlying graph G is shown in Figure 1(a). The graph G encodes an independence structure that can be summarized non-redundantly by the marginal independences

$$X_1 \perp\!\!\!\perp (X_3, X_4), \quad (X_1, X_2) \perp\!\!\!\perp X_4. \tag{1}$$

The graphs in Figure 1(b)-(d) are the three other ancestral graphs that are Markov equivalent to G. i.e. encode the exact same independences. The Gaussian model based

on G receives special attention in Wermuth *et al.* (2004); after appropriate conditioning it gives rise to the model studied in Drton & Richardson (2004c).

Figure 1 about here

Our work consists of a detailed comparison of the Gaussian ancestral graph model based on G and its binary analog. We first explore the relationships to canonical hidden variable models (Section 2), which in the Gaussian case are equal to the ancestral graph model, i.e. they impose only the independences (1). In the binary case, the models differ and we derive cubic polynomial constraints of non-independence type that are imposed by the hidden variable models but not by the ancestral graph model. We then study the likelihood function of the two ancestral graph models (Section 3). We provide a simple parameterization of the binary model, from which we obtain that the likelihood function of the binary model is guaranteed to be unimodal. This is surprising because the likelihood function of the Gaussian model may be multimodal (Drton, 2004; Drton & Richardson, 2004c). Our findings, summarized in Section 4, suggest that binary ancestral graph models may behave very differently from their Gaussian analogs. We remark that tools from computational algebra used in this paper are likely to be helpful in the study of discrete ancestral graph models beyond the particular example considered here (see also Geiger *et al.*, 2005; Pistone *et al.*, 2001; Sturmfels, 2002).

2 Non-independence constraints for canonical Bayesian networks with hidden variables

If an ancestral graph contains solely directed and bi-directed edges, as is the case for the graphs in Figure 1, then the conditional independences it encodes can be induced by marginalizing over hidden variables in a Bayesian network. Richardson & Spirtes (2002, §6) show this using so-called canonical DAGs, which, for the ancestral graphs in Figure

1(a)-(d), are shown in Figure 2(a)-(d). Here X_1, X_2, X_3 and X_4 are observed, whereas H_1, H_2 and H_3 are hidden.

Figure 2 about here

Assume now that the joint distribution of observed and hidden variables satisfies the global Markov property for one of the DAGs in Figure 2 (Lauritzen, 1996, §3.2.2). In particular, the marginal distribution of the observed variables (X_1, X_2, X_3, X_4) must satisfy the marginal independences in (1). By Richardson & Spirtes (2002, Thm. 6.3), no other conditional independences than the consequences of (1) will generally hold among (X_1, X_2, X_3, X_4). However, the hidden variable models may impose additional constraints on the observed margin that are not of independence type. Such constraints may be equality (e.g. Richardson & Spirtes, 2002, §7.3.1) or inequality constraints (e.g. Richardson & Spirtes, 2002, §8.6).

2.1 Gaussian canonical Bayesian networks

Let the joint distribution of observed and hidden variables be a multivariate normal distribution $\mathcal{N}(0, \Gamma)$, which is assumed to be centered merely for notational convenience. The covariance matrix Γ is assumed to be positive definite, denoted by $\Gamma > 0$. Let $\Sigma = (\sigma_{ij}) > 0$ be the 4×4 covariance matrix of the observed variables $(X_1, X_2, X_3, X_4)^t$. As argued above, Σ must be such that the marginal independences (1) hold, i.e.

$$\sigma_{13} = \sigma_{14} = \sigma_{24} = 0. \tag{2}$$

By lengthy elementary algebra using the parameterization of Gaussian Bayesian networks (e.g. Richardson & Spirtes, 2002, §8; Andersson & Perlman, 1998), one can show that for all the canonical DAGs D in Figure 2, the restrictions (2) are the only restrictions imposed by the Bayesian network with hidden variables based on D. This is stated more precisely in the following proposition.

Proposition 1. *Let D be any one of the four DAGs in Figure 2, and let $\Sigma > 0$ be a 4×4 matrix satisfying (2). Then one can find $\Gamma > 0$, where $\Gamma \in \mathbb{R}^{5\times5}$, $\Gamma \in \mathbb{R}^{6\times6}$ or $\Gamma \in \mathbb{R}^{7\times7}$ depending on which DAG D is considered, such that*

1. *the distribution $\mathcal{N}(0, \Gamma)$, the joint multivariate normal distribution of observed and hidden variables, is globally Markov with respect to D, and*

2. *the 4×4 submatrix of Γ that gives the covariance matrix of the observed variables $(X_1, X_2, X_3, X_4)^t$ is equal to Σ.*

2.2 Binary canonical Bayesian networks

Now assume that X_i, $i = 1, \ldots, 4$, and H_i, $i = 1, 2, 3$, are all binary with state space represented by $\{0, 1\}$. Are the marginal independences in (1) still the only restrictions imposed by the hidden variable models based on the canonical DAGs in Figure 2?

Let D be the DAG in Figure 2(a) featuring only one hidden variable. Let P be a probability distribution that is globally Markov with respect to D. Let

$$p_{ijk\ell h} = P(X_1 = i, X_2 = j, X_3 = k, X_4 = \ell, H_2 = h). \tag{3}$$

Furthermore, let

$$p_{ijk\ell} \;=\; p_{ijk\ell+} \;=\; \sum_{h=0}^{1} p_{ijk\ell\, h} \tag{4}$$

denote the *observable probabilities*. Every globally Markov distribution P induces a marginal distribution for the observed variables $(X_1, X_2, X_3, X_4)^t$, and we denote the family of all these marginal distributions by $\mathbb{D}$.

In order to find restrictions which must hold for a vector of observable probabilities to be in $\mathbb{D}$, we take an algebraic approach (cf. Garcia, 2004; Garcia *et al.*, 2004), in which polynomial (equality) constraints can be found (see Cox *et al.* (1997) for an introduction to the algebra involved). In general, additional inequality constraints will

have to hold among the observable probabilities to ensure that the marginal distribution of observed variables is in $\mathbb{D}$. In the algebraic approach, it is advantageous to consider the global Markov property rather than the local Markov property of the DAG D. The independences implied by the global Markov property for the DAG D can be summarized by the four independence statements

$$\begin{aligned} X_1 &\perp\!\!\!\perp (X_3, X_4, H_2), & X_4 &\perp\!\!\!\perp (X_1, X_2, H_2), \\ (X_1, X_2) &\perp\!\!\!\perp (X_3, X_4) \mid H_2, & (X_1, X_4) &\perp\!\!\!\perp H_2. \end{aligned}$$

These four independence statements define an ideal $I_{\mathcal{M}}$ in the polynomial ring $\mathbb{R}[p_{ijk\ell h}]$ generated by the 2^5 indeterminates $p_{ijk\ell h}$; compare Sturmfels (2002, §8). Loosely said, conditional independences for binary variables require some (conditional) odds ratios to be equal to one, which after clearing denominators yields polynomial relations. We write $\mathbb{R}[p_{ijk\ell}]$ for the polynomial subring of $\mathbb{R}[p_{ijk\ell h}]$ generated by the 2^4 observable probabilities $p_{ijk\ell}$ from (4). Let $P_{\mathcal{M}}$ be the distinguished prime ideal of $I_{\mathcal{M}}$ in the polynomial ring $\mathbb{R}[p_{ijk\ell h}]$, see Garcia *et al.* (2004, §4). The ideal $P_{\mathcal{M}}$ can be characterized as the set of all homogeneous polynomial functions on $\mathbb{R}^{2^5}$ which vanish on all probability distributions that factor according to the DAG D.

Proposition 2 (Garcia et al., 2004, Prop. 19). *The set of all polynomial functions which vanish on the set $\mathbb{D}$ of observable probability distributions is equal to the sum of the prime ideal*

$$P'_{\mathcal{M}} \quad = \quad P_{\mathcal{M}} \cap \mathbb{R}[p_{ijk\ell}] \tag{5}$$

and the ideal

$$\Big\langle \sum_{i,j,k,\ell=0}^{1} p_{ijk\ell} - 1 \Big\rangle \tag{6}$$

generated by the constraint that the observable probabilities sum to one.

For the considered model we find that $P_{\mathcal{M}} = I_{\mathcal{M}}$ and that $P'_{\mathcal{M}}$ is generated by three

groups of polynomials: (i) the six 2×2 subdeterminants of the matrix

$$\begin{pmatrix} p_{0000} + p_{0100} & p_{0001} + p_{0101} & p_{0010} + p_{0110} & p_{0011} + p_{0111} \\ p_{1000} + p_{1100} & p_{1001} + p_{1101} & p_{1010} + p_{1110} & p_{1011} + p_{1111} \end{pmatrix}, \tag{7}$$

which correspond to $X_1 \perp\!\!\!\perp (X_3, X_4)$; (ii) the six 2×2 subdeterminants of the matrix

$$\begin{pmatrix} p_{0000} + p_{0010} & p_{0100} + p_{0110} & p_{1000} + p_{1010} & p_{1100} + p_{1110} \\ p_{0001} + p_{0011} & p_{0101} + p_{0111} & p_{1001} + p_{1011} & p_{1101} + p_{1111} \end{pmatrix}, \tag{8}$$

which correspond to $X_4 \perp\!\!\!\perp (X_1, X_2)$; and (iii) the sixteen 3×3 subdeterminants of the 4×4 matrix

$$\begin{pmatrix} p_{0000} & p_{0001} & p_{0010} & p_{0011} \\ p_{0100} & p_{0101} & p_{0110} & p_{0111} \\ p_{1000} & p_{1001} & p_{1010} & p_{1011} \\ p_{1100} & p_{1101} & p_{1110} & p_{1111} \end{pmatrix}. \tag{9}$$

The sixteen cubics in (iii) are constraints of non-independence type that are imposed by the hidden variable model but are not imposed in the ancestral graph model.

For the DAGs in Figure 2(b)-(d), it is also the case that the polynomials (i), (ii), and (iii) generate the ideal $P'_{\mathcal{M}}$. In fact, it can be shown that the set $\mathbb{D}$ of vectors of observable probabilities $p_{ijk\ell}$, $(i, j, k, \ell) \in \{0, 1\}^4$, is the same regardless of which one of the four DAGs in Figure 2(a)-(d) induces $\mathbb{D}$.

Remark. The hidden variable model $\mathbb{D}$ is non-identifiable regardless of which one of the DAGs in Figure 2(a)-(d) is employed to parameterize the model. If the model $\mathbb{D}$ is parameterized using the conditional parameters associated with the smallest DAG, the one in Figure 2(a), then the parameterization comprises 11 parameters: one parameter for each one of the marginal distributions of H_2, X_1 and X_4, and eight more for the conditional distributions $(X_2 \mid X_1, H_2)$ and $(X_3 \mid X_4, H_2)$. The dimension of the hidden variable model $\mathbb{D}$, however, is equal to 9.

3 Likelihood inference in ancestral graph models

We now leave hidden variables models (and non-identifiability issues) behind and return to the ancestral graph model based on the equivalent graphs in Figure 1. For a study of the likelihood function, it is convenient to work with the graph G in Figure 1(a). We remark that the independences (1) can also be represented by the AMP chain graph (Andersson *et al.*, 2001)

$$X_1 \longrightarrow X_2 \text{ — } X_3 \longleftarrow X_4.$$

Thus the below results also apply to AMP chain graphs.

3.1 Gaussian model

The ancestral graph model based on G is the family of (centered) normal distributions $\mathcal{N}(0, \Sigma)$ for $(X_1, X_2, X_3, X_4)^t$ with covariance matrix $\Sigma = (\sigma_{ij}) > 0$ satisfying (2). ML estimation in this model is well understood. If S is the sample covariance matrix, then the ML estimators (MLE) of σ_{11} and σ_{44} are $\hat{\sigma}_{11} = S_{11}$ and $\hat{\sigma}_{44} = S_{44}$ (cf. Drton & Richardson, 2004a). The remaining parameters can be estimated in the model of conditional distributions $(X_3, X_4 \mid X_1, X_2)$ that constitutes a seemingly unrelated regressions model. This yields the following fact (see also Drton, 2004).

Proposition 3 (Drton & Richardson, 2004c, Thm. 2). *The Gaussian ancestral graph model based on the Markov equivalent graphs in Figure 1 has up to five solutions to the likelihood equations, three of which may be local maxima.*

3.2 Parameterization of the binary model

Assume now that X_i, $i = 1, \ldots, 4$, are binary with state space $\{0, 1\}$ and joint distribution P. Let

$$p_{ijk\ell} = P(X_1 = i, X_2 = j, X_3 = k, X_4 = \ell),$$

and

$$p = (p_{0000}, p_{0001}, \ldots, p_{1111}) \in \Delta,$$

where Δ is the probability simplex in $\mathbb{R}^{16}$. The binary ancestral graph model associated with the graph G is defined implicitly as the family of distributions

$$\mathbb{G} = \{p \in \Delta \mid p \text{ satisfies (1)}\}. \tag{10}$$

In order to find a parameterization of $\mathbb{G}$, we use property (C4) of conditional independence (Lauritzen, 1996, Ch. 3) to rewrite the independences (1) equivalently as

$$X_1 \perp\!\!\!\perp X_4, \qquad X_1 \perp\!\!\!\perp X_3 \mid X_4, \qquad X_2 \perp\!\!\!\perp X_4 \mid X_1. \tag{11}$$

This motivates the following parameterization consisting of the marginal parameters

$$s_1 = P(X_1 = 0), \qquad s_4 = P(X_4 = 0),$$

and the conditional parameters

$$t_{2i} = P(X_2 = 0 \mid X_1 = i), \quad i = 0, 1, \qquad t_{3\ell} = P(X_3 = 0 \mid X_4 = \ell), \quad \ell = 0, 1,$$
$$u_{i\ell} = P(X_2 = 0, X_3 = 0 \mid X_1 = i, X_4 = \ell), \quad i, \ell = 0, 1.$$

Overall we have specified 10 parameters that can be mapped to a vector $p \in \mathbb{G}$ as follows. First, we define the marginal distribution of (X_1, X_4) by setting

$$P(X_1 = 0, X_4 = 0) = s_1 s_4, \qquad P(X_1 = 0, X_4 = 1) = s_1(1 - s_4),$$
$$P(X_1 = 1, X_4 = 0) = (1 - s_1)s_4, \qquad P(X_1 = 1, X_4 = 1) = (1 - s_1)(1 - s_4).$$

Define the conditional distribution $(X_2, X_3 \mid X_1, X_4)$ by setting for all $(i, \ell) \in \{0, 1\}^2$:

$$\begin{aligned}
&P(X_2 = 0, X_3 = 0 \mid X_1 = i, X_4 = \ell) = u_{i\ell}, \\
&P(X_2 = 0, X_3 = 1 \mid X_1 = i, X_4 = \ell) = t_{2i} - u_{i\ell}, \\
&P(X_2 = 1, X_3 = 0 \mid X_1 = i, X_4 = \ell) = t_{3\ell} - u_{i\ell}, \\
&P(X_2 = 1, X_3 = 1 \mid X_1 = i, X_4 = \ell) = 1 - t_{2i} - t_{3\ell} + u_{i\ell}.
\end{aligned} \tag{12}$$

Now the parameterization is the injective map $\mathbb{R}^{10} \to \mathbb{R}^{16}$ with the 16 coordinates

$$p_{ijk\ell} = P(X_1 = i, X_4 = \ell)P(X_2 = j, X_3 = k \mid X_1 = i, X_4 = \ell). \tag{13}$$

It is not hard to verify via (11) that if the parameters are chosen such that $p \in \Delta$, then in fact $p \in \mathbb{G}$, and vice versa. Since the components of p defined in (13) sum to one, $p \in \Delta$ if and only if the 16 inequalities $p_{ijk\ell} \geq 0$ hold. Let $\Theta \subset \mathbb{R}^8$ be the set of vectors $\theta = (t_{20}, t_{21}, \ldots, u_{11})^t$ in $[0,1]^8$ that satisfy

$$t_{2i} + t_{3\ell} - 1 \leq u_{i\ell} \leq \min\{t_{2i}, t_{3\ell}\} \tag{14}$$

for all $i, \ell = 0, 1$. Then $p \in \Delta$ if and only if $(s_1, s_4, t_{20}, t_{21}, \ldots, u_{11}) \in [0,1]^2 \times \Theta$, which is the parameter space of $\mathbb{G}$.

3.3 Maximum likelihood estimation in the binary model

Given data $n_{ijk\ell} \in \mathbb{N}$ indicating how often the events $X_1 = i$, $X_2 = j$, $X_3 = k$, and $X_4 = \ell$ occurred and assuming multinomial sampling, the log-likelihood function of $\mathbb{G}$ is

$$\log L(p) = \sum_{i,j,k,\ell=0}^{1} n_{ijk\ell} \log(p_{ijk\ell}),$$

where we ignored a parameter independent additive constant. An MLE $\hat{p}$ of p satisfies

$$\hat{p} \in \arg\max\{\log L(p) \mid p \in \mathbb{G}\}. \tag{15}$$

Using the parameterization presented in Section 3.2, we write $\log L(p)$ as the sum

$$\log L(p) = \log L_1(s_1) + \log L_4(s_4) + \log L_{23|14}(t_{20}, \ldots, u_{11}), \tag{16}$$

where

$$\begin{aligned} \log L_1(s_1) =& n_{0+++} \log(s_1) + n_{1+++} \log(1 - s_1), \\ \log L_4(s_4) =& n_{+++0} \log(s_4) + n_{+++1} \log(1 - s_4), \end{aligned}$$

and

$$\log L_{23|14}(t_{20},\ldots,u_{11}) = n_{0000}\log(u_{00}) + n_{0001}\log(u_{01}) + \\ n_{0010}\log(t_{20}-u_{00}) + \cdots + n_{1111}\log(1-t_{21}-t_{31}+u_{11}).$$

Recall that if an index of a vector is replaced by plus-signs then this indicates that a sum was taken over the index; compare (4).

The graphical models literature does not contain any results on the structure of the maximization problem in (15). In particular, it is unknown if the likelihood function may have several local maxima over the model—a question we can now answer.

Theorem 4. *Let all counts in the data vector n be positive. Then the MLE $\hat{p}$ of $p \in \mathbb{G}$ is the unique local maximum of the likelihood function L over $\mathbb{G}$. Two of the components of the parametric representation of $\hat{p}$ can be computed explicitly as rational functions of the data, namely $\hat{s}_1 = n_{0+++}/n_{++++}$ and $\hat{s}_4 = n_{+++0}/n_{++++}$. The remaining components can be computed by solving a system of polynomial equations having 145 solutions (all real), but only one solution leading to an estimate in Δ.*

Proof. Since the parameter space $[0,1]^2 \times \Theta$ is a Cartesian product, the decomposition (16) permits to find the MLE of s_1 by maximizing $\log L_1$ individually. Thus, the MLE of s_1 is $\hat{s}_1 = n_{0+++}/n_{++++}$. Similarly the MLE of s_4 is $\hat{s}_4 = n_{+++0}/n_{++++}$. Therefore, after substituting $\hat{s}_1$ for s_1 and $\hat{s}_4$ for s_4, the parameterization (13) is a linear parameterization in 8 parameters; compare (12).

Catanese *et al.* (2004) introduce the concept of ML degree, which is the number of complex solutions to the likelihood equations of a statistical model, or equivalently the degree of the algebraic function that maps the data to the MLE. What we are claiming is that ML degree of the binary model $\mathbb{G}$ equals 145. In Catanese *et al.* (2004, Thm. 13), the authors show that the ML degree of a linearly parameterized model is equal to the number of bounded regions of the hyperplane arrangement defined by the linear

parametric equations, i.e. the equations

$$u_{00} = 0,\ u_{01} = 0,\ t_{20} - u_{00} = 0,\ \ldots, 1 - t_{21} - t_{31} + u_{11} = 0.$$

The number of bounded regions of a hyperplane arrangement $\mathcal{A}$ can be obtained by evaluating the Poincaré polynomial of $\mathcal{A}$ at -1, see Zaslavsky (1975). The Poincaré polynomial of the hyperplane arrangement (13) equals

$$P(t) = 1880\,t^8 + 4536\,t^7 + 5160\,t^6 + 3616\,t^5 + 1704\,t^4 + 552\,t^3 + 120\,t^2 + 16\,t + 1.$$

Thus, the ML degree equals $P(-1) = 145$. This computation was carried using D. Serenevy's *GAP Arrangements* package [1]. The same answer was independently obtained in *Singular* (Greuel *et al.*, 2001) using Algorithm 18 in Hosten *et al.* (2004).

Despite the fact that the likelihood equations generally have 145 solutions (all of which are in fact guaranteed to be real), only one solution will be in the probability simplex. This follows from the fact that, since $p_{ijk\ell}$ are products of linear forms in the parameters, the log-likelihood function $\log L$ is strictly concave and has exactly one local = global maximum over the model $\mathbb{G}$. □

Theorem 4 shows that the likelihood equations of the model $\mathbb{G}$ can be solved via a high-degree system of polynomial equations. In practice, however, it will be computationally more efficient to apply a hill-climbing method to compute the unique maximum of the likelihood function.

3.4 Zero counts in the binary model

In Theorem 4 the counts $n_{ijk\ell}$ were assumed to be all positive. This condition is sufficient but not necessary for strict concavity of the log-likelihood function, and can be relaxed as follows. Let

$$\mathcal{I} = \big\{(i,j,k,\ell) \in \{0,1\}^4 \mid n_{ijk\ell} \geq 1\big\} \tag{17}$$

[1] `http://dean.serenevy.net//?page=arrangement`

be the set indexing the positive counts. The mapping $\mathbb{R}^8 \to \mathbb{R}^{16}$ from the parameters $\theta = (t_{20}, t_{21}, \ldots, u_{11})^t$ to the 16 conditional probabilities $P(X_2 = j, X_3 = k \mid X_1 = i, X_4 = \ell)$ in (12) is linear. Thus the vector of conditional probabilities can be written as $C\theta + b$, where b is a vector whose components are equal to one if $j = k = 1$ and equal to zero otherwise. The matrix C equals

$$C = \begin{array}{c} \begin{array}{cccccccc} t_{20} & t_{21} & t_{30} & t_{31} & u_{00} & u_{01} & u_{10} & u_{11} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & -1 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

with the rows being ordered according to

$$(i, j, k, \ell) = (0,0,0,0), (0,0,0,1), (0,0,1,0), \ldots, (1,1,1,1).$$

Theorem 5. *Let $C_{\mathcal{I}}$ be the $|\mathcal{I}| \times 8$ submatrix of C obtained by selecting the rows of C corresponding to the indices in $\mathcal{I}$. Then the log-likelihood function $\log L$ of the model $\mathbb{G}$ is strictly concave if and only if the matrix $C_{\mathcal{I}}$ is of full rank 8.*

Proof. In the decomposition (16) of the log-likelihood function, the pieces $\log L_1$ and $\log L_4$ are strictly concave if and only if $n_{++++} \geq 1$, i.e. if there are any data at all. Therefore, $\log L$ is strictly concave if and only if the conditional log-likelihood function $\log L_{23|14} : \Theta \to \mathbb{R}$ is strictly concave.

Let $\theta \neq \bar{\theta}$ be two vectors in $\Theta \subseteq \mathbb{R}^8$. Let $c = C\theta + b$ and $\bar{c} = C\bar{\theta} + b$ be the two corresponding vectors of conditional probabilities. If $C_{\mathcal{I}}$ is of full rank then there exists $(i_0, j_0, k_0, \ell_0) \in \mathcal{I}$ such that $c_{i_0 j_0 k_0 \ell_0} \neq \bar{c}_{i_0 j_0 k_0 \ell_0}$. By strict concavity of the logarithm it follows that for any $\lambda \in (0, 1)$,

$$\log[\lambda c_{ijk\ell} + (1-\lambda)\bar{c}_{ijk\ell}] \geq \lambda \log(c_{ijk\ell}) + (1-\lambda)\log(\bar{c}_{ijk\ell}) \qquad \forall (i,j,k,\ell) \in \{0,1\}^4,$$

and

$$\log[\lambda c_{i_0 j_0 k_0 \ell_0} + (1-\lambda)\bar{c}_{i_0 j_0 k_0 \ell_0}] > \lambda \log(c_{i_0 j_0 k_0 \ell_0}) + (1-\lambda)\log(\bar{c}_{i_0 j_0 k_0 \ell_0}).$$

Thus $\log L_{23|14}$ is strictly concave.

If the rank of $C_{\mathcal{I}}$ is not full then there exists a vector $\theta_0 \neq 0$ in the kernel of $C_{\mathcal{I}}$. Choosing $\theta \in \Theta$ and $\nu \in \mathbb{R}$ such that $\bar{\theta} = \theta + \nu\theta_0 \in \Theta$, we find that $\log L_{23|14}$ is constant over the line segment between θ and θ_0, hence not strictly concave. □

If the log-likelihood function $\log L$ of the model $\mathbb{G}$ is strictly concave then it has a unique local = global maximum, and we saw that strict concavity depends on which counts $n_{ijk\ell}$ are positive. Regardless of the data, however, the log-likelihood function is concave, which yields that any local maximum is already a global maximum. The following two examples illustrate two points about the situation when the log-likelihood function is not strictly concave. On one hand the counts may be such that there are infinitely many local = global maxima. On the other hand they may be such that there is a unique local maximum even if the log-likelihood function is not strictly concave.

Example 6. Assume that the only positive counts are n_{0000} and n_{0001}. Then

$$\log L_{23|14}(\theta) = n_{0000}\log(u_{00}) + n_{0001}\log(u_{01}).$$

Clearly, this function is maximized by any $\hat{\theta} \in \Theta$ such that $\hat{u}_{00} = \hat{u}_{01} = 1$. By (14), it has to hold that $\hat{t}_{20} = \hat{t}_{30} = \hat{t}_{31} = 1$. The remaining components $(\hat{t}_{21}, \hat{u}_{10}, \hat{u}_{11})$ may take on any feasible values. □

Example 7. Assume now that the only positive counts are n_{0000} and n_{1111}. Then

$$\log L_{23|14}(\theta) = n_{0000} \log(u_{00}) + n_{1111} \log(1 - t_{21} - t_{31} + u_{11}).$$

If $\hat{\theta} \in \Theta$ maximizes this function (locally or globally), then $\hat{u}_{00} = 1$ and $1-\hat{t}_{21}-\hat{t}_{31}+\hat{u}_{11} = 1$. By (14), $\hat{u}_{00} = 1$ implies that $\hat{t}_{20} = \hat{t}_{30}$. From $\hat{u}_{11} = \hat{t}_{21} + \hat{t}_{31}$, it follows via (14) that $\hat{t}_{21} = \hat{t}_{31} = \hat{u}_{01} = \hat{u}_{10} = \hat{u}_{11} = 0$. Hence, the maximizer $\hat{\theta}$ is unique. □

3.5 Comment on non-binary discrete models

If the variables X_i, $i = 1, \ldots, 4$, have more than two levels, then a linear parameterization of conditional probabilities as in (12) can still be found. Suppose $X_i \in [m_i] := \{1, \ldots, m_i\}$ may take on one of m_i many states. Then the discrete ancestral graph model defined by (1) can be parameterized by the marginal probabilities

$$\begin{aligned} s_1(i) &= P(X_1 = i), \qquad i \in [m_1 - 1], \\ s_4(\ell) &= P(X_4 = \ell), \qquad \ell \in [m_4 - 1], \end{aligned}$$

and conditional probabilities

$$\begin{aligned} t_{2i}(j) &= P(X_2 = j \mid X_1 = i), \qquad (i, j) \in [m_1] \times [m_2 - 1], \\ t_{3\ell}(k) &= P(X_3 = k \mid X_4 = \ell), \qquad (k, \ell) \in [m_3 - 1] \times [m_4], \end{aligned}$$

and

$$\begin{aligned} u_{i\ell}(j, k) = P(X_2 = j, X_3 = k \mid X_1 = i, X_4 = \ell), \\ (i, j, k, \ell) \in [m_1] \times [m_2 - 1] \times [m_3 - 1] \times [m_4]. \end{aligned}$$

All conditional probabilities $P(X_2 = j, X_3 = k \mid X_1 = i, X_4 = \ell)$ can be written as linear expressions in these parameters similarly as in (12). Thus the results on unique local maxima of the likelihood function given in Theorems 4 and 5 carry over to discrete models for non-binary variables defined via (1).

4 Conclusion

In a detailed study, we showed that a binary ancestral graph model can behave very differently from its Gaussian analog in terms of uniqueness of local maxima of the likelihood function and relation to hidden variable models. An interesting observation to be made from our work is that in both the Gaussian and the binary case all canonical Bayesian networks with hidden variables associated with Markov equivalent ancestral graphs induce the same constraints on the observed margin. It is an open problem to find conditions on the ancestral graph which lead to canonical Bayesian networks imposing the same constraints. Another question that remains open is whether binary ancestral graph models may have a multimodal likelihood function. We believe that when studying this question algebraic tools such as the ML degree mentioned in the proof of Theorem 4 will be helpful. However, the degree 145 appearing in Theorem 4 suggests that necessary symbolic computations will be very demanding.

Acknowledgments

We would like to thank Thomas Richardson, Bernd Sturmfels and Seth Sullivant for helpful comments. The first author acknowledges support by NIH grant R01-HG2362-3. The second author thanks MSRI for hosting him during the Fall of 2004 and providing many resources that made this collaboration possible.

References

Andersson, S. A., Madigan, D. & Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.* **28**, 33–85.

Andersson, S. A. & Perlman, M. D. (1998). Normal linear regression models with recursive graphical Markov structure. *J. Multivariate Anal.* **66**, 133–187.

Catanese, F., Hosten, S., Khetan, A. & Sturmfels, B. (2004). The maximum likelihood degree. Manuscript, `math.AG/0406533`.

Cox, D., Little, J. & O'Shea, D. (1997). *Ideals, varieties, and algorithms* (Second ed.). Springer-Verlag, New York.

Cox, D. R. & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation.* Chapman and Hall, London.

Drton, M. (2004). Computing all roots of the likelihood equations of seemingly unrelated regressions. *J. Symbolic Comput.*, accepted.

Drton, M. & Richardson, T. S. (2003). A new algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence. In U. Kjærulff & C. Meek (Eds.), *Proceedings of the* 19^{th} *Conference on Uncertainty in Artificial Intelligence*, pp. 184–191. Morgan Kaufmann, San Francisco.

Drton, M. & Richardson, T. S. (2004a). Graphical answers to questions about likelihood inference for Gaussian covariance models. Technical Report 467, Department of Statistics, University of Washington. Submitted to *Ann. Statist.*

Drton, M. & Richardson, T. S. (2004b). Iterative conditional fitting for Gaussian ancestral graph models. In M. Chickering & J. Halpern (Eds.), *Proceedings of the* 20^{th} *Conference on Uncertainty in Artificial Intelligence*, pp. 130–137. AUAI Press, Arlington, VA.

Drton, M. & Richardson, T. S. (2004c). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika* **91**, 383–392.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799–805.

Garcia, L. D. (2004). Algebraic statistics in model selection. In M. Chickering & J. Halpern (Eds.), *Proceedings of the* 20^{th} *Conference on Uncertainty in Artificial Intelligence*, pp. 177–184. AUAI Press, Arlington, VA.

Garcia, L. D., Stillman, M. & Sturmfels, B. (2004). Algebraic geometry of Bayesian networks. *J. Symbolic Comput.*. Special Issue Méthodes Effectives en Géometrie Algébrique (MEGA).

Geiger, D., Meek, C. & Sturmfels, B. (2005). On the toric algebra of graphical models. *Ann. Statist.*, to appear.

Greuel, G.-M., Pfister, G. & Schönemann, H. (2001). Singular 2.0: A computer algebra system for polynomial computations. University of Kaiserslautern.

Hosten, S., Khetan, A. & Sturmfels, B. (2004). Solving the likelihood equations. Manuscript, `math.ST/0408270`.

Koster, J. (2002). Marginalizing and conditioning in graphical models. *Bernoulli* **8**, 817–840.

Koster, J. T. A. (1999). Linear structural equations and graphical models. Lecture notes, The Fields Institute, Toronto.

Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press, Oxford, UK.

Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge, UK.

Pistone, G., Riccomagno, E. & Wynn, H. P. (2001). *Algebraic statistics. computational commutative algebra in statistics*. Chapman & Hall/ CRC, Boca Raton, FL.

Richardson, T. S. & Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30**, 962–1030.

Richardson, T. S. & Spirtes, P. (2003). Causal inference via ancestral graph models (with discussion). In P. Green, N. Hjort, & S. Richardson (Eds.), *Highly Structured Stochastic Systems*, Chapter 3, pp. 83–105. Oxford University Press, Oxford, UK.

Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, prediction, and search* (Second ed.). MIT Press, Cambridge, MA.

Sturmfels, B. (2002). *Solving systems of polynomial equations*, Volume 97 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC.

Wermuth, N., Cox, D. R. & Marchetti, G. (2004). Consequences of a chain of covariances. Manuscript.

Zaslavsky, T. (1975). Facing up to arrangements: Face–count formulas for partitions of space by hyperplanes. *Memoirs Amer. Math. Soc.* **154**.

Mathias Drton, Department of Mathematics, University of California, Berkeley CA, 94720-3840, U.S.A.

E-mail: drton@math.berkeley.edu

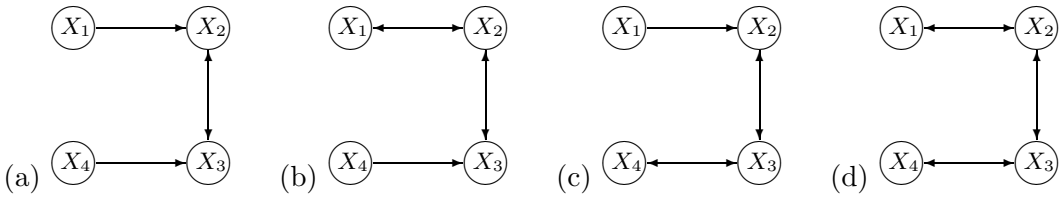

Figure 1: (a) The ancestral graph G, (b)-(d) the ancestral graphs that are Markov equivalent to G.

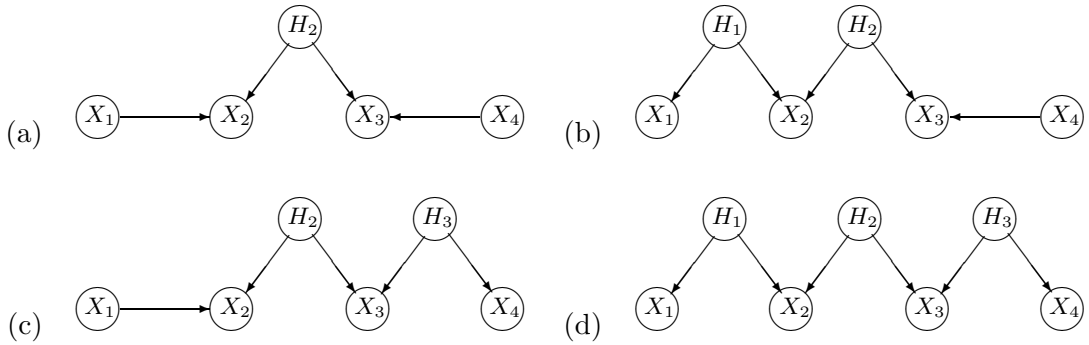

Figure 2: Canonical DAGs with hidden variables H_1, H_2 and H_3.