THE GEOMETRY OF INDEPENDENCE TREE MODELS WITH HIDDEN VARIABLES

PIOTR ZWIERNIK AND JIM Q. SMITH

ABSTRACT. In this paper we investigate the geometry of undirected discrete graphical models of trees when all the variables in the system are binary and the inner nodes are unobserved . We obtain a full geometric description of these models which is given by polynomial equations and inequalities. We also give exact formulas for their parameters in terms of the marginal probability over the observed variables. A new system of coordinates is given that is intrinsically related to the defining trees and gives better insight into the underlying geometry.

1. INTRODUCTION

Discrete graphical models have become a very popular tool in the statistical analysis of multivariate problems (see e.g. [21][34]). When all the variables in the system are observed they exhibit a useful modularity. In particular it is possible to estimate all the conditional probabilities that parametrize such models. In addition, for discrete data the model is described by polynomial equations in the ambient model space, maximum likelihood estimates are simple sample proportions and conjugate Bayesian analysis is straightforward. However, if the values of some of the variables are unobserved then the resulting marginal distribution over the observed variables is usually more complicated both from the geometric and the inferential point of view [16][32]. The models usually impart additional inequality constraints, maximum likelihood estimates are much more complicated and no conjugate analysis is possible. The main problem with the geometric analysis of these models is that in general it is hard to obtain the inequality constraints defining a model even for very simple examples (see [11, Section 4.3][15, Section 7]).

The original motivation of this paper was to study the semi-algebraic geometry of underlying phylogenetic tree models for binary data. Phylogenetic analysis is based on Markov processes on trees which have the property that all the inner nodes in the tree represent hidden variables. The same family of models is considered in other contexts - e.g. Bayesian networks on rooted trees. Under a restriction to positive probabilities the undirected graphical models for trees in the case when all the inner nodes are hidden also represent the same family of distributions. A geometric understanding of these models led to the method of phylogenetic invariants introduced by Lake [20], and Cavender and Felsenstein [5]. The idea behind that is that when a given phylogenetic model holds then the probabilities of the observed tables of the observable variables must satisfy certain algebraic relations. These relations are given as zeros of a set of polynomial equations. Since the problem is

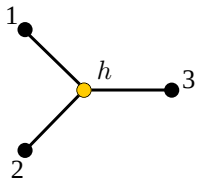
Key words and phrases. conditional independence, graphical models on trees, general Markov models, hidden data, binary data, higher-order correlations, cumulants, Möbius function, semi-algebraic statistical models, inequality constraints, phylogenetic invariants, tree metrics.

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

algebraic in nature it has been recently studied by algebraic geometers [1][13][38]. The substantial progress in the understanding of these invariants may result in their greater applicability in the statistical analysis [4].

There are still two main technical problems related to phylogenetic invariants. First, in general it is hard to obtain them. Second, the phylogenetic invariants do not give a full geometric description of the statistical model. There are some nontrivial polynomial inequalities which have to be satisfied.

Example 1.1. Let T be the tripod tree below



The inner node represents a binary hidden variable H and the leaves represent binary observable variables X_1, X_2, X_3. The tree represents conditional independence statements $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3 | H$. The model has full dimension and consequently there are no equations defining it. However not all the probability distributions lie in the model. Lazarsfeld [22, Section 3.1] showed that they must for example satisfy

$$\mathrm{Cov}(X_1, X_2)\mathrm{Cov}(X_1, X_3)\mathrm{Cov}(X_2, X_3) \geq 0.$$

Hence doing inference based only on the phylogenetic invariants introduces a systematic error.

This example motivated the closer investigation of the geometry of these models. So far the inequality constraints for tree models with hidden variables have been largely neglected. However, some results can be found in the literature. A solution in the case of a binary naive Bayes model was given by Auvray et al. [2]. In the binary case there are also some partial results for general tree structures given by Pearl and Tarsi [26] and Steel and Faller [35]. The most important applications in biology involve variables that can take four values. Recently Matsen [23] gave a set of inequalities in this case for group-based phylogenetic models (additional symmetries are assumed) using the Fourier transformation of the raw probabilities. Here we use ideas based on Settimi and Smith [32] who show that those constraints can be more easily expressed as relations among the central moments like in the example above. We develop a new parametrization based on higher order correlations. When expressed in this new parametrization the underlying geometry of the models becomes transparent. The inequalities defining the model are just simple restrictions on the second-order correlations between observable variables. We also obtain the exact formulas for the parameters of the models in terms of the marginal distribution of the observed variables extending results proved in [7][32]. Combining some earlier results we provide the exact semi-algebraic description of binary phylogenetic tree models in the case of the trivalent trees. However, the inequalities we develop also hold for general tree topologies.

The paper is organized as follows. In Section 2 we briefly introduce conditional independence models on trees stating the result of Allman and Rhodes [1] on equations defining the models. In Section 3 we introduce higher order correlations. Next in Section 4 we show a close link between the second order correlations induced by

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

the model and tree metrics in phylogenetic analysis. At the end of the section we also state the main theorem of the paper. In Section 5 we construct another, more intrinsic, coordinate system for the class of binary tree models. In the new coordinate system the parametrization of the model becomes monomial. This gives a better insight into the underlying geometry. We prove our main theorem in Section 6. In Section 7 we use the parametrization developed earlier to find an alternative form of equations given by Allman and Rhodes [1]. These are slightly simpler from algebraic point of view and have a more transparent statistical interpretation. Finally, in Section 8 we provide two simple examples: the tripod tree model and the hidden Markov model.

2. Independence models on trees

2.1. **Preliminaries on trees.** A *graph* G is an ordered pair (V, E) consisting of a non-empty set V of *vertices* and a set E of *edges* each of which is an element of $V \times V$. An edge $(u, v) \in E$ is *undirected* if $(v, u) \in E$ as well, otherwise it is *directed*. Graphs with only (un)directed edges are called (un)directed. If $e = (u, v)$ is an edge of a graph G, then u and v are called *adjacent* or *neighbours* and e is said to be *incident with* u and v. Let $v \in V$, the *degree* of v is denoted by $\deg(v)$, and is the number of edges incident with v. The graph obtained from G by identifying the ends of e and then deleting e, denoted G/e, is said to be obtained from G by *contracting* e. A *clique* is a graph in which each pair of distinct vertices is joined by an edge. A *path* in a graph G is a sequence of distinct vertices $v_1, v_2, \ldots, v_k$ such that, for all $i = 1, \ldots, k$, v_i and v_{i+1} are adjacent. If, in addition, v_1 and v_k are adjacent then the path is called a *cycle*. If G is directed and a cycle $v_1, v_2, \ldots, v_k$ is such that there is a directed edge from v_i to v_{i+1} for all $i = 1, \ldots, k-1$ then the cycle is called *directed*. A graph is *connected* if each pair of vertices in G can be joined by a path.

A (*directed*) *tree* $T = (V, E)$ is a connected (*directed*) graph with no cycles. A vertex of T of degree at most one is called a *leaf*. A vertex of T that is not a leaf is called an *interior node*. An edge of T is *interior* if both of its ends are interior vertices. Trees in this paper will always have n leaves. We denote the set of leaves by its labeling set $[n] = \{1, \ldots, n\}$. A connected subgraph of T is a *subtree* of T. For a subset V' of V, we let $T(V')$ denote the minimal connected subgraph of T that contains the vertices in V' and we say $T(V')$ is the subtree of T induced by V'. A *rooted tree* is a directed tree that has one distinguished vertex called the *root*, denoted by the letter r, and all the edges are directed away from r. For every vertex v of a rooted tree T^r we denote by $\mathrm{pa}(v)$ the vertex directly preceding v and call $\mathrm{pa}(v)$ the parent of v.

2.2. **Models defined by global Markov properties.** In this paper we always assume that random variables are binary. We consider models with hidden variables, i.e. variables whose values are never directly observed. The vector Y has as its components all variables in the graphical model, both those that are observed and those that are hidden. The subvector of Y of manifest variables, i.e. variables whose values are always observed, is denoted by X and the subvector of hidden variables by H.

Let $T = (V, E)$ be an undirected tree. For any three disjoint subsets A, B, C of the set of nodes we say that C *separates* A and B in T, denoted by $A \perp_T B|C$, if each path from a vertex in A to a vertex B passes through a vertex in C. We are

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

interested in statistical models for $Y = (Y_v)_{v\in V}$ defined by global Markov properties (GMP) on T, i.e. the set of conditional independence statements of the form (see e.g. [21, Section 3.2.1]):

$$\{Y_A \perp\!\!\!\perp Y_B | Y_C : \text{ for all } A, B, C \subset V \text{ s.t. } A \perp_T B|C\}, \tag{1}$$

where for any $A \subset V$ vector Y_A is the subvector of Y with elements indexed by A, i.e. $Y_A = (Y_i)_{i\in A}$.

Conditional independence statements give some polynomial equations in $(p_\alpha)_{\alpha\in\{0,1\}^{|V|}}$, where $p_\alpha := \mathbb{P}(Y_1 = \alpha_1, \ldots, Y_{|V|} = \alpha_{|V|})$ define the joint probability mass function of Y. From an algebraic view point (for basic definitions see [8]) this collection of polynomials forms an ideal which we denote by I_{global} (see [15][37, Section 8.1]).

When we are interested in the marginal model of the vector X of manifest variables corresponding to the leaves of T the algebraic analysis is much harder and this is the main concern of the paper. If all the variables in the system are observable then the model is denoted by $\widehat{\mathcal{M}}_T$. If all the variables related to the inner nodes of T are hidden then the model is denoted by $\mathcal{M}_T$. We call $\mathcal{M}_T$ a *general Markov model on* T.

2.3. **Models for rooted trees.** A Markov process on a rooted tree T^r is a sequence $\{Y_v : v \in V\}$ of random variables such that for each $(\alpha_1, \ldots, \alpha_{|V|}) \in \{0,1\}^{|V|}$

$$p_\alpha = \prod_{v\in V} \mathbb{P}(Y_v = \alpha_v | Y_{\text{pa}(v)} = \alpha_{\text{pa}(v)}), \tag{2}$$

where $\text{pa}(r)$ is an empty set.

Using standard results in the theory of graphical models we get that after the restriction to positive probabilities (we call it the positivity assumption) the Markov process on T is equal to $\widehat{\mathcal{M}}_T$. By [21, Theorem 3.27] the parametrization defined by Equation (2) is equivalent to directed global Markov properties on T^r. Moreover, since T^r has a uniquely defined root the moral graph of T^r is equal to its undirected version T. Hence, the directed global Markov properties on T^r are implied by the global Markov properties on T and they are equivalent under the positivity assumption. Note in passing that by Theorem 6 in [15] the underlying ideals of both models are the same as well.

Let $\Delta_{2^n-1} = \{p \in \mathbb{R}^{2^n} : \sum_\beta p_\beta = 1, p_\beta \geq 0\}$ with indices β ranging over $\{0,1\}^n$ be the probability simplex of all possible distributions on $X = (X_1, \ldots, X_n)$ represented by the leaves of T. In this paper, by the positivity assumption, we always restrict to the interior of Δ_{2^n-1}.

The model defined by (2) has exactly $2|E|+1$ free parameters - one for the root distribution and two for each edge. Equation (2) induces a map $\phi_T : [0,1]^{2|E|+1} \to \Delta_{2^n-1}$ obtained by marginalization over all the inner nodes of T. The parametrization of $\mathcal{M}_T$ is given by

$$p_{\alpha_{[n]}} = \sum_{\mathcal{H}} \mathbb{P}(Y_r = \alpha_r) \prod_{v\in V\setminus r} \mathbb{P}(Y_v = \alpha_v | Y_{\text{pa}(v)} = \alpha_{\text{pa}(v)}), \tag{3}$$

where $\mathcal{H}$ are all possible states of the vector of hidden variables, i.e. the sum is over $\alpha_{V\setminus[n]} \in \{0,1\}^{|V|-n}$ and for any $A \subseteq V$ $\alpha_A = (\alpha_i)_{i\in A}$. The name "general Markov model" for $\mathcal{M}_T$ comes from the theory of phylogenetic tree models. By definition these are models for the rooted tree T^r defined by (3). Since this model is equivalent to $\mathcal{M}_T$ the term is justifiable. Note that in particular by using $\mathcal{M}_T$

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

we do not distinguish the root in any sense, which agrees with the fact that the choice of the root is irrelevant to the analysis of general Markov models (e.g. [31, Section 8.2]).

2.4. **Equations of the model.** If X_1, X_2 are two independent binary random variables then with definition of p_α as above we have $p_{00}p_{11} - p_{01}p_{10} = 0$. Moreover, one can check that the equation is satisfied if and only if the joint distribution of X_1, X_2 satisfies this equation. The algebraic approach mentioned in the beginning of this section led Allman and Rhodes [1] to an answer to a more general question about equations defining the general Markov model for binary data. Equivalently we want to study the phylogenetic ideal, i.e. the set of all polynomials vanishing on $\phi_T(\mathbb{C}^{2|E|+1})$. Considering the complex numbers in this context is a common approach in algebraic geometry and in this context this is done to make the analysis easier.

To introduce the result we need the following definition.

Definition 2.1. Let $X = (X_1, \ldots, X_n)$ be a vector of binary random variables and let $P = (p_\gamma)_{\gamma \in \{0,1\}^n}$ be a $2 \times \ldots \times 2$ table of the joint distribution of X. Let $(A)(B)$ form a partition of $[n]$. Then the *flattening* of P induced by the partition is a matrix

$$P_{(A)(B)} = [p_{\alpha\beta}], \quad \alpha \in \{0,1\}^{|A|}, \beta = \{0,1\}^{|B|},$$

where $p_{\alpha\beta} = \mathbb{P}(X_A = \alpha, X_B = \beta)$. Let $T = (V, E)$ be a tree. Then, for each $e \in E$, removing edge e from E induces a partition of the set of leaves into two subsets according to the connected components of the resulting forest. The obtained flattening is called an *edge flattening* and we denote it by P_e.

Note that whenever we implicitly use some order on coordinates indexed by $\{0,1\}$-sequences we always mean the order induced by the lexicographic order on $\{0,1\}$-sequences such that $0\cdots00 > 0\cdots01 > \ldots > 1\cdots11$.

If P is the joint distribution of $X = (X_1, \ldots, X_n)$ then each of the flattenings is just a matrix representation of the joint distribution P and contains essentially the same probabilistic information. However, these different representations contain important geometric information about the model.

Theorem 2.1 (Allman, Rhodes [1])**.** *The ideal defining the general Markov model for a trivalent tree T (all the inner nodes have degree three) is generated by all 3×3-minors of all the edge flattenings of T plus the trivial invariant $\sum_\alpha p_\alpha = 1$.*

In a less algebraic language the result just provides equations defining the general Markov model.

3. Higher order correlations

In this section we introduce a useful set of coordinates for the model space. Let $X = (X_1, \ldots, X_n)$ be a vector of binary random variables. Then we can obtain formulas relating the moments of these variables to the probability distribution of X. For each $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$ $X^\beta = \prod_i X_i^{\beta_i}$ and define $\lambda_\beta := \mathbb{E}X^\beta$ and $\mu_\beta = \mathbb{E}[\prod_{i=1}^n (X_i - \mathbb{E}X_i)^{\beta_i}]$. For $\alpha = (\alpha_1, \ldots, \alpha_n) \in \{0,1\}^n$ we have

$$\lambda_\alpha := \mathbb{E}X^\alpha = \sum_{\alpha \leq \beta \leq \mathbf{1}} p_\beta, \tag{4}$$

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

where $\mathbf{1}$ denotes here the vector of ones and the sum is over all binary vectors β such that $\alpha \leq \beta \leq \mathbf{1}$ which means $\alpha_i \leq \beta_i \leq 1$ for all $i = 1, \ldots, n$. It is also straightforward to check that

$$\mu_\alpha = \sum_{0 \leq \beta \leq \alpha} (-1)^{|\beta|} \lambda_{\alpha-\beta} \prod_{i=1}^{n} \lambda_{e_i}^{\beta_i} \quad \text{for } \alpha \in \{0,1\}^n \text{ such that } |\alpha| > 1, \tag{5}$$

where $|\beta| = \sum_i \beta_i$ and $e_1, \ldots, e_n$ are unit vectors in $\mathbb{R}^n$ so $\lambda_{e_i} = \mathbb{E}X_i$. The inverse formula is

$$\lambda_\alpha = \sum_{0 \leq \beta \leq \alpha} \mu_{\alpha-\beta} \prod_{i=1}^{n} \lambda_{e_i}^{\beta_i} \quad \text{for } \alpha \in \{0,1\}^n. \tag{6}$$

For any $I \subseteq [n]$ the I-correlation (also called a higher order correlation, e.g. [9]) is defined as

$$\rho_I = \rho(X_I) := \mathbb{E}\left(\prod_{i \in I} U_i\right),$$

where $U_i = (X_i - \mathbb{E}X_i)/\sqrt{\mathrm{Var}(X_i)}$. Let $[n]_k$ denote the set of subsets of $[n]$ with exactly k-elements. If $I \in [n]_k$ then we sometimes call ρ_I a k-th order correlation. A formula for an I-correlation in terms of the non-central moments is obtained after normalization of Equation (5) for $\alpha = \sum a_i e_i$, where $a_i = 1$ if $i \in I$ and is zero otherwise. Note that under the positivity assumption the higher order correlations are always well-defined.

For elegance of forthcoming expressions we use skewness $\rho_{iii} = \mathbb{E}U_i^3$ instead of mean $\mathbb{E}X_i$. Provided that $\mathbb{E}X_i \neq 0, 1$ we have a one-to-one correspondence between the two. Indeed, by simple algebra calculations we can show that

$$\rho_{iii} = \frac{1 - 2\lambda_{e_i}}{\sqrt{\lambda_{e_i}(1 - \lambda_{e_i})}}, \quad \lambda_{e_i} = \frac{1}{2}\left(1 - \frac{\rho_{iii}}{\sqrt{4 + \rho_{iii}^2}}\right). \tag{7}$$

Using $\rho_{iii} \in \mathbb{R}$ instead of $\mathbb{E}X_i \in (0,1)$ enables us not only to obtain invariance with respect to any particular choice of the values for the variables in the system (changing the state space from $\{0,1\}$ to $\{a,b\}$ always affects the means but not the skewness) but also simplifies the parametric formulation of the model – see Section 5.

The models in the previous section are defined in terms of probabilities. We investigate the models under the suitable change of coordinates. A similar approach is presented for example in [32][17]. Let $\mathcal{R}_n$ be a space of all possible I-correlations for all $I \in [n]_{\geq 2}$, where $[n]_{\geq k}$ denotes the set of all subset of $[n]$ with at least k elements. Let $\mathbb{R}^n$ be the space of all possible skewnesses for elements of vector X. The reparametrization map $\phi_1 : \Delta_{2^n-1} \to \mathbb{R}^n \times \mathcal{R}_n$ induced by (5) is well defined on a dense open subset of Δ_{2^n-1} corresponding to its interior and is a diffeomorphism after constrained to this subset. The inverse of ϕ_1 may be obtained by combination of (4), (6) and (7) which results in the corrected version of the Bahadur's expansion [36].

For any two sets A, B let AB denote $A \cup B$. The basic condition on independence (see e.g. Feller [14], p 136) implies that if $X_A \perp\!\!\!\perp X_B$ then $\rho_{AB} = \rho_A \rho_B$. But since the variables are binary we also have a converse result. It is well know that if for each $I \subseteq A$, $J \subseteq B$ we have that $\rho_{IJ} = \rho_I \rho_J$ then $X_A \perp\!\!\!\perp X_B$. Indeed, the definition of the independence states that $X_A \perp\!\!\!\perp X_B$ if and only if $\mathrm{Cov}(f(X_A), g(X_B)) = 0$ for

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

any L^2 functions f and g. Since our variables are binary all the functions of X_A and X_B are just polynomials with square-free monomials. Settimi and Smith [32] concluded that the independence holds if and only if $\mathrm{Cov}(X_A^\alpha, X_B^\beta) = 0$ for each non-zero $\alpha \in \{0,1\}^{|A|}$ and $\beta \in \{0,1\}^{|B|}$. Equivalently we have $\mathrm{Cov}(U_A^\alpha, U_B^\beta) = 0$ which can be written as $\rho_{IJ} = \rho_I \rho_J$ for each nonempty $I \subseteq A$, $J \subseteq B$.

We also define a conditional I-correlation

$$\rho_{I|\mathcal{E}} = \rho(X_I|\mathcal{E}) := \mathbb{E}\left(\prod_{i\in I} U_i \Big| \mathcal{E}\right)$$

conditionally on the event $\mathcal{E}$. Of course $\rho(X_i)$ for $i = 1, \ldots, n$ is always zero but $\rho(X_i|\mathcal{E})$ is usually not. Let A, B, C be three disjoint subsets of $[n]$, then $\rho(X_B|X_C)$ will be denoted by $\rho_{B|C}$. If $X_A \perp\!\!\!\perp X_B | X_C$ then $\rho_{AB|C} = \rho_{A|C}\rho_{B|C}$. Note that this property generalizes the concept of the partial correlation (see e.g. [19, Chapter 27]).

Let Y be a binary variable and let $f : \mathbb{R} \to \mathbb{R}$ be any function. Then $f(Y) = c + dY$ for some $c, d \in \mathbb{R}$. Using this fact and writing a conditional expectation given a binary variable Y as a function of Y we can show that for each $I \subseteq [n]$

$$\rho_{I|Y} = \mathbb{E}(\prod_{i\in I} U_i | Y) = \rho_I + \rho_{I\,Y} U_Y.$$

In particular if X, Y are two random variables then $\rho_{X|Y} = \rho_{XY} U_Y$.

Let X_A, X_B be two vectors of binary variables. We have $X_A \perp\!\!\!\perp X_B | Y$ if and only if for all nonempty $I \subseteq A$, $J \subseteq B$

$$\begin{aligned} &\rho_{IJ} = \rho_I \rho_J + \rho_{IY}\rho_{JY}, \\ &\rho_{IJY} = \rho_I \rho_{JY} + \rho_{IY}\rho_J + \rho_{IY}\rho_{JY}\rho_{YYY}. \end{aligned} \tag{8}$$

Indeed, the equivalent condition for $X_A \perp\!\!\!\perp X_B | Y$ is that for each $I \subseteq A$, $J \subseteq B$ we have

$$\begin{cases} \mathrm{Cov}(X_I, X_J | Y = 0) = 0, \\ \mathrm{Cov}(X_I, X_J | Y = 1) = 0. \end{cases}$$

or after simple operations

$$\begin{cases} \mathbb{E}Y \mathrm{Cov}(X_I, X_J | Y = 1) + (1 - \mathbb{E}Y)\mathrm{Cov}(X_I, X_J | Y = 0) = 0, \\ \mathrm{Cov}(X_I, X_J | Y = 0) - \mathrm{Cov}(X_I, X_J | Y = 1) = 0. \end{cases}$$

which can be checked to be equivalent to (8).

4. Correlations and tree models

In the next two sections we focus on reparametrizations for graphical models on trees. The first step is to reparametrize the parameter space for the tree models using ideas from the previous section. Let $T = (V, E)$ be a tree with n leaves. Note that for a tree $1 + 2|E| = |V| + |E|$ is always true so the number of free parameters in (2) and (3) is $|V| + |E|$. Let $\phi_2 : [0,1]^{|V|+|E|} \to \mathbb{R}^n \times \mathbb{R}^{|V|-n} \times [-1,1]^{|E|}$ be a transformation from the original set of parameters to the set of parameters given by ρ_{uv} for all $(u, v) \in E$ (edge correlations) and ρ_{iii} for all $i \in V$ (skewnesses). We will denote $\Omega_T = \mathbb{R}^{|V|-n} \times [-1,1]^{|E|}$. So the new parameter space is $\mathbb{R}^n \times \Omega_T$. We split $\mathbb{R}^{|V|} = \mathbb{R}^n \times \mathbb{R}^{|V|-n}$ for a reason which will become clear later.

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

The most convenient way of defining the map ϕ_2 is as follows. Fix a rooting of T. For each directed edge $u \to v$ in E we have

$$\rho_{uv} = [\mathbb{P}(Y_v = 1|Y_u = 1) - \mathbb{P}(Y_v = 1|Y_u = 0)] \sqrt{\frac{\mathbb{P}(Y_u = 0)\mathbb{P}(Y_u = 1)}{\mathbb{P}(Y_v = 0)\mathbb{P}(Y_v = 1)}}. \tag{9}$$

Moreover for any node $i \in V$ the marginal distribution $\mathbb{P}(Y_i)$ and hence the skewness ρ_{iii} can be obtained using the root distribution and the conditional distributions. The inverse formula can be obtained in a similar manner. The map is well defined on an open dense subset corresponding to all the variables in the tree being non-degenerate. From now on we implicitly transform the parameter space using ϕ_2 assuming non-degeneracy.

By $\mathcal{M}_T^\rho$ we denote the image of a projection of $\phi_1(\mathcal{M}_T) \subset \mathbb{R}^n \times \mathcal{R}_n$ on $\mathcal{R}_n$. $\widehat{\mathcal{M}}_T^\rho$ is defined in a similar way.

Example 4.1. Recall the tripod tree model in Example 1.1. Since the model is defined by $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3|H$ using Equation (8) it can be shown that for $R \in \mathcal{R}_3$ we have $R \in \mathcal{M}_T^\rho$ if and only if:

$$\rho_{ij} = \rho_{ih}\rho_{jh} \text{ for all } i \neq j \in \{1, 2, 3\} \text{ and } \rho_{123} = \rho_{hhh}\rho_{1h}\rho_{2h}\rho_{3h} \tag{10}$$

for some $(\rho_{hhh}, \rho_{1h}, \rho_{2h}, \rho_{3h}) \in \mathbb{R} \times [-1, 1]^3$.

Let $k, l \in V$ be any two nodes representing variables Y_k, Y_l and let $\mathcal{P}_T(k, l)$ be the unique simple path (with no repeated vertices) joining them in T. Then using (8) recursively one can show that

$$\rho_{kl} = \prod_{(u,v)\in\mathcal{P}_T(k,l)} \rho_{uv} \tag{11}$$

for each probability distribution in $\widehat{\mathcal{M}}_T^\rho$.

There is an interesting reformulation of our problem in term of tree metrics (see [31, Section 7] or a definition below) which we explain below. The relation between the distance based methods used in phylogenetics and the correlation between variables in the binary case has been already pointed out for example by Cavender [6].

Let $d : V \times V \to \mathbb{R}$ be a map defined as

$$d(k, l) = \begin{cases} -\log(|\rho_{kl}|), & \text{for all } k, l \in V \text{ such that } \rho_{kl} \neq 0, \\ +\infty, & \text{otherwise} \end{cases}$$

then $d(k, l) \geq 0$ because $|\rho_{kl}| \leq 1$ and $d(k, k) = 0$ for all $k \in V$ since $\rho_{kk} = 1$.

If $R \in \mathcal{M}_T^\rho$ then by Equation (11) we can define map $d_{(T;R)} : V \times V \to \mathbb{R}$

$$d_{(T;R)}(k, l) = \begin{cases} \sum_{(u,v)\in\mathcal{P}_T(k,l)} d(u, v), & \text{if } k \neq l, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

This map after restriction to the product of the set of leaves $[n] \times [n] \subset V \times V$ is called a tree metric. In our case we have a point in the model space defining all the second order correlations and $d_{(T;R)}(i, j)$ for $i, j \in [n]$. The question is: What are the conditions for the "distances" between leaves so that there exists a tree T and edge lengths $d(u, v)$ for all $(u, v) \in E$ such that (12) is satisfied. Or equivalently: What are the conditions on the absolute values of the second order correlations in order that $|\rho_{ij}| = \prod_{(u,v)\in\mathcal{P}_T(i,j)} |\rho_{uv}|$ (for some edge correlations) is satisfied.

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

The answer to that question is well known in phylogenetic analysis. We have the following theorem.

Theorem 4.1 (Tree-Metric Theorem, Buneman [3]). *A function $\delta : [n] \times [n] \to \mathbb{R}$ is a tree metric on $[n]$ if and only if for every four (not necessarily distinct) elements $i, j, k, l \in [n]$,*

$$\delta(i,j) + \delta(k,l) \leq \max\left\{\delta(i,k) + \delta(j,l), \delta(i,l) + \delta(j,k)\right\}.$$

Note that since the elements $i, j, k, l \in [n]$ need not be distinct, every map satisfying the four-point condition defines a metric on $[n]$. Moreover, from the general theory we know that a tree metric defines the tree uniquely.

The four-point condition in terms of correlations translates to $|\rho_{ij}\rho_{kl}| \geq \min\{|\rho_{ik}\rho_{jl}|, |\rho_{il}\rho_{jk}|\}$. We can state the following well known corollary (c.f. [31, Section 8.4]).

Corollary. *If $P \in \mathcal{M}_T$ for some T then we can reconstruct T from the second order correlations between the leaves.*

Now we need an additional constraint on the second order correlations. This ensures that there exists a choice of signs for the correlations of all the edges consistent with the signs of the correlations between the leaves. It is easy to check that this condition is that $\rho_{ij}\rho_{ik}\rho_{jk} \geq 0$ for all triples $\{i, j, k\} \subset [n]$.

A surprising fact is that when the equations in Theorem 2.1 are supplemented with the above conditions then this provides the complete semi-algebraic description of the general Markov model. In Section 6 we prove the following theorem.

Theorem 4.2. *Let $T = (V, E)$ be a tree with n leaves such that all the inner nodes have degree at most three. Let $\mathcal{M}_T$ be a general Markov model on T. Suppose P is a joint probability distribution of n binary variables in the interior of Δ_{2^n-1}. Then $P \in \mathcal{M}_T$ if and only if P is such that all the 3×3-minors of all the edge flattenings of P vanish and in addition*

(I): *for all (not necessarily distinct) leaves $i, j, k, l \in [n]$*

$$|\rho_{ij}\rho_{kl}| \geq \min\left\{|\rho_{ik}\rho_{jl}|, |\rho_{il}\rho_{jk}|\right\}.$$

(II): *for all distinct triples $i, j, k \in [n]$*

$$\rho_{ij}\rho_{ik}\rho_{jk} \geq 0.$$

5. Tree-based cumulants

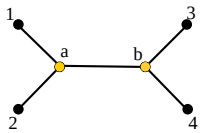
In Section 3 we defined a reparametrization map $\phi_1 : \Delta_{2^n-1} \to \mathbb{R}^n \times \mathcal{R}_n$ of the model space. In this section we perform a further reparametrization of the system of I-correlations for $I \in [n]_{\geq 2}$ for another system of coordinates which in some sense is intrinsically linked to the given tree. The model in this coordinate system admits a monomial parametrization in edge correlations and the skewnesses. The coordinates link to the concept of cumulants which are essentially model-free (see e.g. [24, Section 2] [27]). Our idea here is to develop some "tree-based cumulants" to obtain as simple parametric form of the model as possible. Our approach in this section is more combinatorial and is based on the theory of Möbius functions (see [28]).

Let $T = (V, E)$ be a tree with n leaves. A *split induced by* T is a partition of $[n]$ into two non-empty sets induced by removing an edge from T and restricting $[n]$ to the connected component of the resulting graph. Let $(A)(B)$ be a split induced

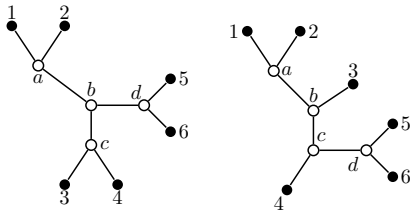
by T and for $W \subset V$ let $T(W)$ denote the minimal subtree of T induced by W (c.f. Section 2.1). Then any split of A induced by $T(A)$ or a split of B induced by $T(B)$ induces a partition of $[n]$ into three sets. We can iterate the procedure. By a *multisplit* we mean any partition $(A_1)\cdots(A_k)$ of the set of leaves induced by removing a subset of the set of edges of T. Each A_i is called a *block* of a partition. If x is an multisplit then $\prod_{B\in x}$ always denotes a product over all the blocks of x.

By Π_T we denote the partially ordered set (poset) of all multisplits induced by inner edges of T. For $x, y \in \Pi_T$ we write $x \leq y$ if and only if y is a *subsplit* of x which means that y can be obtained from multisplit x by further splits of its blocks. For two elements $x, y \in \Pi_T$ a *join* (a *meet*) of x and y denoted by $x \vee y$ ($x \wedge y$) is an element $z \in \Pi_T$ such that $z \geq x$, $z \geq y$ ($z \leq x$, $z \leq y$) and if there exists another $z' \in \Pi_T$ with this property then $z' \geq z$ ($z' \leq z$). In our case z always exists and it can be shown that if x is obtained by removing some subset of inner edges $E_x \subset E$ and y is induced by removing $E_y \subset E$ then $x \vee y$ is induced by $E_x \cup E_y$ and $x \vee y$ by $E_x \cap E_y$. Hence by definition Π_T forms a lattice. It has a unique maximal element induced by removing all inner edges and the minimal one with no edges removed which is equal to a single block $[n]$. The maximal element of a lattice is denoted by 1 and the minimal one is denoted by 0. A *segment* $[x, y]$, for x and y in Π_T, is the set of all elements z such that $x \leq z \leq y$.

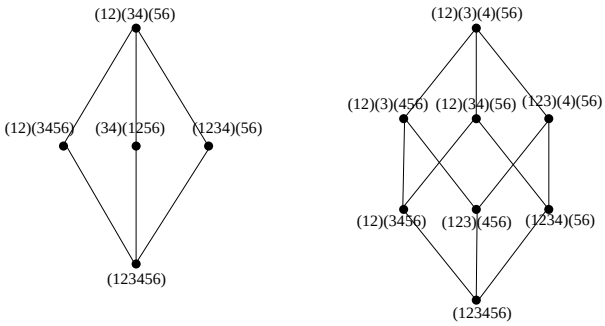
For an example let T be the quartet tree below



It has only one inner edge and hence the partially ordered set Π_T has exactly two elements $0 = (1234)$ and $1 = (12)(34)$. Now consider two different trivalent trees with six leaves



The posets Π_T^L (left) and Π_T^R (right) look as follows.



CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

So for example (12)(34)(56) is a multisplit in Π_T^L and it is a subsplit of any other multisplit $y \in \Pi_T^L$. Since for $x = (12)(34)(56)$ there are no subsplits of x apart from x itself then x is a maximal element in Π_T^L. However, it is not maximal in Π_T^R.

For any poset Π one defines a *Möbius function* $m_\Pi : \Pi \times \Pi \to \mathbb{R}$ such that $m_\Pi(x,x) = 1$ for every $x \in \Pi$ and $m_\Pi(x,y) = -\sum_{x \le z < y} m_\Pi(x,z)$ for $x < y$ in Π. Let $W \subset V$ then we denote $m_{\Pi_{T(W)}} := m_W$ and $m_{\Pi_T} := m$. We write 0_W and 1_W to denote the minimal and the maximal element of $\Pi_{T(W)}$ respectively. For any multisplit $x \in \Pi_T$ the interval $[x,1]$ has a natural structure of a product of posets for blocks of x namely $\prod_{B \in x} \Pi_{T(B)}$. By Proposition 4 in [28] the Möbius function on $\prod_{B \in x} \Pi_{T(B)}$ can be written as a product of Möbius functions for each of the posets $\Pi_{T(B)}$. Thus

$$m(x,y) = \prod_{B \in x} m_B(0_B, y_B) \quad \text{for } y_B \in \Pi_{T(B)}, \tag{13}$$

where y_B means the restriction of $y \in \Pi_T$ to the block containing only elements from $B \subset [n]$ (it is well definied since $x \le y$) and $x_B = 0_B$ for each B.

We use this combinatorial machinery to define new coordinates $(\kappa_I)_{I \in [n]_{\ge 2}}$ using reparametrization $f : \mathcal{R}_n \to \mathbb{R}^{2^n - n - 1}$ defined by the Möbius function on $\Pi_{T(I)}$ for $I \in [n]_{\ge 2}$ in the following way. Let

$$\kappa_I = \sum_{\pi \in \Pi_T(I)} m_I(0_I, \pi) \prod_{B \in \pi} \rho_B \quad \text{for all } I \in [n]_{\ge 2}, \tag{14}$$

where $\rho_i = 0$ for all $i \in [n]$. We define $\mathcal{C}_n = f(\mathcal{R}_n)$ and $\mathcal{M}_T^\kappa = f(\mathcal{M}_T^\rho) \subseteq \mathcal{C}_n$. By definition for any $I \in [n]_{\le 3}$ we can identify $\kappa_I = \rho_I$, where $[n]_{\le 3}$ denotes all the subsets of $[n]$ with at most three elements.

Equation (14) justifies the name for the tree-based cumulants. Indeed, one of the alternative definitions of cumulants is as follows. Let $\mathcal{P}(I)$ denote the set of all partitions of I. Then

$$K_I = \sum_{\pi \in \mathcal{P}(I)} (-1)^{|\pi|-1} (|\pi| - 1)! \prod_{B \in \pi} \mathbb{E}(X_B) \tag{15}$$

where the product is over all blocks of π and $|\pi|$ denotes the number of blocks in π. This is essentially the same as Equation (14) but with a different defining poset (see [36][29]).

The map in (14) is invertible. We can obtain an inverse map using the Möbius inversion or more specifically the dual Möbius inversion. We state the result for reader's convenience and to set the notation.

Lemma 5.1 (Dual Möbius Inversion)**.** *Let Π be a finite poset with 1 and $\alpha, \beta : \Pi \to \mathbb{R}$ be any two functions. Then*

$$\alpha(y) = \sum_{x \ge y} \beta(x), \qquad \textit{for all } y \in \Pi$$

if and only if

$$\beta(y) = \sum_{x \ge y} m_\Pi(y,x)\alpha(x), \qquad \textit{for all } y \in \Pi,$$

where $m_\Pi(\cdot,\cdot)$ is a Möbius function of Π.

To apply the lemma in our case we need the following result.

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

Lemma 5.2. *Let T be a tree and let $I \in [n]_{\geq 2}$. Define two functions on $\Pi_{T(I)}$*

$$\alpha(y) = \prod_{B \in y} \rho_B, \qquad \beta(y) = \prod_{B \in y} \kappa_B,$$

then $\beta(y) = \sum_{x \geq y} m_I(y,x)\alpha(x)$ for all $y \in \Pi_{T(I)}$.

Proof. For each $y \in \Pi_{T(I)}$ by definition $\beta(y) = \prod \kappa_B$ and using Equation (14) we obtain

$$\prod_{B \in y} \kappa_B = \prod_{B \in y} \left(\sum_{x_B \in \Pi_T(B)} m_B(0_B, x_B) \prod_{C \in x_B} \rho_C \right) = \sum_{x \geq y} \prod_{B \in x} m_B(0_B, x_B) \prod_{C \in x} \rho_C.$$

By the product formula in Equation (13) we have $\prod_{B \in x} m_B(0_B, x_B) = m(y,x)$ which gives the required result. □

It now follows by Lemma 5.1 and Lemma 5.2 that the inverse map for (14) is defined coordinatewise as follows - for each $I \in [n]_{\geq 2}$

$$\rho_I = \sum_{x \in \Pi_{T(I)}} \prod_{B \in x} \kappa_B, \tag{16}$$

where the product is over all blocks of the partition.

The above reparametrization is a diffeomorphism. To see this, order the variables in such a way that κ_I precedes κ_J as long as $I \subset J$ and do the same for ρ_I, ρ_J. Then it can be checked that the Jacobian matrix of (16) is lower triangular with all its diagonal terms taking the value one.

The following proposition motivates the whole section. It shows that our new coordinate system is particularly useful.

Proposition 5.3. *Let $T = (V, E)$ be a tree with n leaves. Then $\mathcal{M}_T^{\kappa}$ is parametrized as follows:*

$$\kappa_I = \prod_{(u,v) \in E(I)} \rho_{uv} \prod_{v \in \mathrm{int}(V(I))} \rho_{vvv}^{\deg(v)-2} \quad \text{for each } I \in [n]_{\geq 2}, \tag{17}$$

where the degree is taken in $T(I) = (V(I), E(I))$ and $\mathrm{int}(V(I))$ denotes the set of inner nodes of $T(I)$.

In order to prove the proposition we need some more general theory of partially ordered sets.

Definition 5.1. A closure relation in a partially ordered set Π (see [28]) is a function $x \mapsto \bar{x}$ of Π into itself with the properties (1) $\bar{x} \geq x$ (2) $\bar{\bar{x}} = \bar{x}$ and (3) $x \geq y$ implies $\bar{x} \geq \bar{y}$. An element $x \in \Pi$ is closed if $x = \overline{x}$.

Let $T = (V, E)$ be a tree with n leaves and let $i, j \in [n]$ such that $\{i, j\}$ forms a *cherry*, i.e. by definition a pair of leaves such that there exists a single inner node separating i and j from all the other leaves. We define a closure relation in Π_T induced by $\{i, j\}$ in the following way. If $x \in \Pi_T$ contains block (ijC) for some $C \subset [n]$ then we split the block into two blocks: $(ij)(C)$ otherwise $\overline{x} = x$. One can easily check that this satisfies conditions of Definition 5.1. Moreover, 0 is never closed and 1 always is and hence $\overline{0} > 0$ and $\overline{1} = 1$.

We have the following proposition.

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

Proposition 5.4 (Proposition 4 in [28]). *Let $x \to \bar{x}$ be a closure relation on a finite lattice Π with the property that $\overline{x \vee y} = \overline{x} \vee \overline{y}$ and that $\overline{0} > 0$. Then for all $y \in \Pi$,*

$$\sum_{\{x:\, \bar{x}=y\}} m(0,x) = 0.$$

Definition 5.2. Let $\{i,j\}$ be a cherry and $A = [n] \setminus \{i,j\}$. Let $w = (ij)(1_A) \in \Pi_T$ and $a \in V$ be the node of T separating i and j from A. A trimming map with respect to $\{i,j\}$ is a map $[0,w] \to \Pi_{T(aA)}$ such that $x \mapsto \widetilde{x}$ is defined by changing the block (ijC) in $x \in [0,w]$ for (aC).

Note that if a is a leaf in $T(aA)$ then the trimming map constitutes an isomorphism between $[0,w]$ and $\Pi_{T(aA)}$. Now we are ready to prove Proposition 5.3.

Proof of Proposition 5.3. First assume that all the inner nodes of T have degree at most three. In this case it suffices to prove the lemma for $I = [n]$. The general result obviously follows by restriction to the subtree $T(I)$ since each inner node of $T(I)$ has degree at most three. By Equation (10) the lemma is true for $n \leq 3$ since $\rho_{12} = \prod_{(u,v)\in\mathcal{P}_T(1,2)} \rho_{uv} = \kappa_{12}$, $\rho_{123} = \rho_{hhh}\rho_{1h}\rho_{2h}\rho_{3h} = \kappa_{123}$, where h denotes an inner node separating the leaves and ρ_{ih} is a shorthand for $\prod_{(u,v)\in\mathcal{P}_T(i,h)} \rho_{uv}$.

Now let us assume the theorem is true for all $k \leq n-1$ and let T be a tree with n leaves. We can always find two leaves separated from all the other leaves by an inner node (a cherry). Denote the leaves by $1,2$ and the inner node by a. Denote $A = \{3,\ldots,n\}$ and let $T(aA)$ be the minimal subtree of T induced by a and A. Note that the global Markov properties give that for each $C \subseteq A$ we have $(X_1, X_2) \perp\!\!\!\perp X_C | H_a$ so using Equation (8) we have

$$\rho_{12C} = \rho_{12}\rho_C + \rho_{12a}\rho_{aC}. \tag{18}$$

For each $x \in \Pi_T$ let $\overline{x}$ and $\widetilde{x}$ denote the image of x under the closure relation and the trimming map induced by the cherry $\{1,2\}$ respectively. By Equation (14) $\kappa_{[n]} = \sum_{x\in\Pi_T} m(0,x) \prod_{B\in x} \rho_B$. Let $w = (12)(1_A)$ then applying Equation (18) to each block of a form $(12C)$ we obtain

$$\begin{aligned}\kappa_{[n]} = \sum_{x\in\Pi_T} m(0,x) \prod_{B\in\overline{x}} \rho_B + \sum_{x\notin[0,w]} m(0,x) \prod_{B\in x} \rho_B + \\ + \rho_{12a} \sum_{\widetilde{x}\in[0,w]} m(0,x) \prod_{B\in\widetilde{x}} \rho_B.\end{aligned} \tag{19}$$

If we denote by $\overline{\Pi}_T$ the poset of all elements closed under the closure relation then the first summand can be rewritten as $\sum_{y\in\overline{\Pi}_T} \left[\left(\sum_{\{x:\, \bar{x}=y\}} m(0,x)\right) \prod_{B\in y} \rho_B \right]$, which is zero by Proposition 5.4 (the condition $\overline{x \vee y} = \overline{x} \vee \overline{y}$ can be easily checked if one thinks in terms of edge deletions inducing partitions). The second summand is zero as well because if $x \notin [0,w]$ then x contains (1) or (2) as one of the blocks and $\rho_1 = \rho_2 = 0$. Since $\{1,2\}$ form a cherry and all the inner nodes of T have degree at most three then a necessarily has degree three in T and it is a leaf in $T(aA)$ and the trimming map constitutes an isomorphism between $[0,w]$ and $\Pi_T(aA)$. By Proposition 4 in [28] the Möbius function of $[0,w]$ is equal to the restriction of the Möbius function on Π_T to the interval $[0,w]$. Since this is equal to the Möbius

function on $\Pi_{T(aA)}$ we have

$$\rho_{12a}\sum_{\widetilde{x}\in[0,w]} m(0,x)\prod_{B\in\widetilde{x}}\rho_B = \rho_{12a}\sum_{x\in\Pi_{T(aA)}} m_{aA}(0_{aA},x)\prod_{B\in x}\rho_B = \rho_{12a}\kappa_{aA}.$$

It can be checked that $\rho_{12a}=\rho_{1a}\rho_{2a}\rho_{aaa}$. Also since $|aA|=n-1$ by using the induction assumption

$$\kappa_{aA} = \prod_{(u,v)\in E(aA)}\rho_{uv}\prod_{v\in\mathrm{int}(V(aA))}\rho_{vvv}^{\deg(v)-2},$$

where the degree is taken in $T(aA)$. But $E = E(aA)\cup\mathcal{P}_T(1,a)\cup\mathcal{P}_T(2,a)$ and $\mathrm{int}(V)=\mathrm{int}(V(aA))\cup\{a\}\cup\mathrm{int}(V(1a))\cup\mathrm{int}(V(2a))$. Degree of a in T is three and the degree of all the inner nodes of $T(1a)$ and $T(2a)$ is two. Hence one can check that $\rho_{12a}\kappa_{aA}$ satisfies Equation (17). This finishes the proof in the case when all the inner nodes of T have degree at most three.

To obtain a general form of the formula for any tree T note that T can be obtained from a tree with all the inner nodes of degree at most three by identifying adjacent inner nodes (contracting some edges). If we identify two inner nodes a,b then we set $\rho_{ab}=1$ and $\rho_{aaa}=\rho_{bbb}$. □

We will denote the map defined by Equation (17) by ψ_T and hence $\mathcal{M}_T^\kappa = \psi_T(\Omega_T)\subseteq\mathcal{C}_n$, where Ω_T was defined in Section 4.

6. Proof of the main theorem

By the results of previous sections we can show that $\mathbb{R}^n\times\mathcal{M}_T^\kappa\subset\mathbb{R}^n\times\mathcal{C}_n$ is diffeomorphic to a dense open subset of $\mathcal{M}_T\subset\Delta_{2^n-1}$. Indeed, it follows from Proposition 5.3 that the parametrization of $\mathcal{M}_T^\kappa$ does not depend of ρ_{iii} for all $i\in[n]$. Hence

$$\Delta_{2^n-1}\xrightarrow{\phi_1}\mathbb{R}^n\times\mathcal{R}_n\xrightarrow{\mathrm{id}\times f}\mathbb{R}^n\times\mathcal{C}_n$$

constitutes the diffeomorphism. This means that for any tree T with n leaves both $\widehat{\mathcal{M}}_T$ and $\mathcal{M}_T$ restricted to the dense open subset defining non-degenerate random variables will always have a trivial component $\mathbb{R}^n$ being a projection on the n coordinates ρ_{iii} for $i=1,\ldots,n$. So we can focus on the non-trivial component and hence we call ψ_T the non-trivial part of the parametrization map $(\mathrm{id}\times\psi_T):\mathbb{R}^n\times\Omega_T\to\mathbb{R}^n\times\mathcal{C}_n$.

Corollary. *Let $T=(V,E)$ be a tree and let $\mathcal{M}_T$ be the general Markov model on T. Then* $\dim(\mathcal{M}_T)=|E|+|V|$.

Proof. The parametrization in Equation (17) is injective. Its image is diffeomorphic to a dense open subset of the non-trivial part of $\mathcal{M}_T$. Since $\dim\mathcal{M}_T^\kappa=\dim\Omega_T=|V|-n+|E|$ then the dimension of $\mathcal{M}_T$ has to be $|V|+|E|$. □

The following well known lemma (see e.g. [26]) shows that we can assume all the inner nodes in the given tree have degree at least three.

Lemma 6.1. *Let T be a tree. Let r be a vertex of degree two and let $e_1=(u,r)$, $e_2=(r,v)$ be the edges incident with r. Then $P\in\mathcal{M}_T$ if and only if $P\in\mathcal{M}_{T/e_1}=\mathcal{M}_{T/e_2}$, where T/e denotes a tree obtained from T by contracting edge e (c.f. Section 2.1).*

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

The following lemma gives the inequalities defining $\mathcal{M}_T^\kappa$. The lemma is formulated in terms of the complex numbers. This allows us to make a link to the algebraic theory of phylogenetic invariants.

Lemma 6.2. *Let $T = (V, E)$ be a tree with n leaves. Let $\mathcal{M}_T$ be a general Markov model on T. If $\kappa \in \mathcal{C}_n$ is such that $\kappa \in \psi_T(\mathbb{C}^{|V|-n+|E|})$ then $\kappa \in \mathcal{M}_T^\kappa$ if and only if inequalities in (I) and (II) in Theorem 4.2 are satisfied .*

Constraints (I) and (II) are formulated in terms of second order correlations but since $\rho_{ij} = \kappa_{ij}$ for all $i, j \subset [n]$ they are constraints on κ. We prefer to keep the formulation in terms of correlations.

Proof. Note that by Lemma 6.1 we can assume that each of the inner nodes of T has degree greater than two. Let $\kappa \in \psi_T(\mathbb{C}^{|V|-n+|E|})$ satisfy (I) and (II). We will show that $(\psi_T)^{-1}(\kappa) \in \Omega_T$. We assume that all the edge correlations are non-zero which defines an open dense subset in Ω_T. The image of this subset is an open dense subset of $\mathcal{M}_T^\kappa$ corresponding to all ρ_{ij} for $i, j \in [n]$ being non-zero.

We begin with the correlations $\rho_{i\mathrm{pa}(i)}$ for $i = 1, \ldots, n$. Since the degree of $\mathrm{pa}(i)$ is at least three, we can find two other leaves j, k in the tree such that $\mathrm{pa}(i)$ separates i, j and k in T. Definition of ψ_T implies that

$$\rho^2_{i\mathrm{pa}(i)} = \frac{\rho_{ij}\rho_{ik}}{\rho_{jk}} \tag{20}$$

and this does not depend on the choice of j and k as long as $\rho_{jk} \neq 0$. If (I) and (II) hold then $\rho_{i\mathrm{pa}(i)}$ is a valid correlation. Indeed, by (II) the right hand side is positive so $\rho_{i\mathrm{pa}(i)}$ is a real number and by (I) we have $|\rho_{i\mathrm{pa}(i)}| \leq 1$ (take two of the four leaves as equal).

To compute the inner edges correlations ρ_{ab} for each inner edge $(a, b) \in E$ note that we get at least four subsets of the set of leaves such that for any four leaves each from a different subset we have a quartet subtree (see Figure 1). Denote the

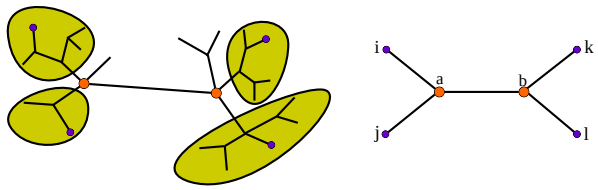


FIGURE 1. On the left we pick two adjacent inner nodes and four leaves (one from each of the shaded areas). On the right we have the marginal subtree for those four variables.

four chosen leaves as i, j, k, l. Again by the definiton of ψ_T we obtain

$$\frac{\rho_{ik}\rho_{jl}}{\rho_{ij}\rho_{kl}} = \frac{\rho_{il}\rho_{jk}}{\rho_{ij}\rho_{kl}} = \rho^2_{ab}. \tag{21}$$

As in the case of the pendant edges by (I) and (II) we have $|\rho_{ab}| \leq 1$.

For every inner node h representing a hidden variable H we can find $i, j, k \in [n]$ such that h is the only trivalent node of the subtree $T(ijk)$ and then we have

$$\kappa_{ijk} = \rho_{ijk} = \rho_{hhh}\rho_{ih}\rho_{jh}\rho_{kh}.$$

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

However, by Equation (7) ρ_{hhh} is an unbounded and one-to-one function of $\mathbb{E}H$. Moreover since $\kappa \in \mathcal{C}_n$ is a probability distribution then ρ_{ijk} is a real number and hence the equation

$$\rho_{hhh} = \frac{\rho_{ijk}}{\rho_{ih}\rho_{jh}\rho_{kh}} \tag{22}$$

does not impute any further inequality constraints. Consequently, the preimage of κ under ψ_T is in Ω_T.

Now let us assume that $\kappa \in \mathcal{M}_T^\kappa$. For any three leaves i, j, k we can find a unique inner node h such that $X_i \perp\!\!\!\perp X_j \perp\!\!\!\perp X_k | H$, so by Equation (20) we get $\rho_{ij}\rho_{jk}\rho_{ik} > 0$. For any four leaves consider a marginal quartet tree model. By Equation (21) we know that additionally

$$\frac{\rho_{ik}\rho_{jl}}{\rho_{ij}\rho_{kl}} = \frac{\rho_{il}\rho_{jk}}{\rho_{ij}\rho_{kl}} \leq 1,$$

which is equivalent to

$$|\rho_{ik}\rho_{jl}| = |\rho_{il}\rho_{jk}| \leq |\rho_{ij}\rho_{kl}|,$$

which of course implies that

$$|\rho_{ij}\rho_{kl}| \geq \min\{|\rho_{ik}\rho_{jl}|, |\rho_{il}\rho_{jk}|\}.$$

If we allow some of the second order correlations between leaves to be zero then we take the closure. This would change the strict inequality $\rho_{ij}\rho_{jk}\rho_{ik} > 0$ into an inequality.

□

Proof of Theorem 4.2. Parametrization $\mathrm{id} \times \psi_T$ defines a variety which under $(\mathrm{id} \times f) \circ \phi_1$ coincides with $\mathcal{M}_T$ on an open dense subset. Hence the equations defining $\mathcal{M}_T$ will also define (after the change of coordinates) $\mathbb{R}^n \times \mathcal{M}_T^\kappa$. By Theorem 2.1 the equations are given by all 3×3 minors of all the edge flattenings of P. On the other hand Lemma 6.2 gives the inequalities defining $\mathcal{M}_T^\kappa$. Again, since $\mathbb{R}^n \times \mathcal{M}_T^\kappa$ is diffeomorphic to a dense open subset of $\mathcal{M}_T$ then the inequalities hold also for $\mathcal{M}_T$. □

Lemma 6.2 shows that if all the inner nodes of T have degree at least three then $\mathcal{M}_T$ is identifiable (up to the switching of labels of the hidden variables). The identifiability of $\mathcal{M}_T$ was proved in [32] and in a more general case in [7]. Note however that the proof of Lemma 6.2 gives the explicit formulas for the parameters.

Remark. If T has nodes of degree greater than three than $\mathcal{M}_T \subset \mathcal{M}_{T'}$ where all inner nodes of T' have degree at most three and one can obtain T from T' by edge contractions. Hence the inequalities (I) and (II) hold also for $\mathcal{M}_T$.

Remark. In the phylogenetic analysis one usually assumes that $\rho_{uv} > 0$ for all $(u, v) \in E$ which reduces the set of the inequalities in Theorem 4.2.

7. Phylogenetic invariants

Our new coordinates allow us to prove several useful results related to the structure of phylogenetic ideals defining $\mathcal{M}_T$. In this section we will consider only a non-normalized version of κ's, i.e. based on central moments and not on the higher order correlations, which we denote by ν. We have that $\nu_I = \kappa_I \prod_{i \in I} \sqrt{\mathrm{Var}(X_i)}$

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

for every $I \in [n]_{\geq 2}$. In particular since both (14) and (16) can be multiplied on the both sides by $\sqrt{\prod_{i\in I} \mathrm{Var}(X_i)}$ we obtain

$$\nu_I = \sum_{x\in\Pi_T(I)} m_I(0_I, x) \prod_{B\in x} \mu_B \quad \text{for every } I \in [n]_{\geq 2}, \tag{23}$$

$$\mu_I = \sum_{x\in\Pi_T(I)} \prod_{B\in x} \nu_B \quad \text{for every } I \in [n]_{\geq 2}. \tag{24}$$

Moreover since the central moments and the raw probabilities are related by a polynomial map as well, there exists a polynomial change of coordinates from the raw probabilities to the ν's.

It is convenient to use another notational convention. In this section we sometimes exchange the usual notation ν_I for $I \in [n]_{\geq 2}$ and the notation ν_α for $\alpha \in \{0,1\}^n$ where $\alpha_i = 1$ if $i \in I$ and it is zero otherwise.

In an analogous way to the edge flattenings of probability distributions we can define edge flattenings of $(\nu_I)_{I\in[n]_{\geq 2}}$. Let e be an edge of T inducing a split $(A)(B) \in \Pi_T$ such that $|A| = r$, $|B| = n - r$. Then N_e is a $(2^r - 1) \times (2^{n-r} - 1)$ matrix such that for any two nonempty subsets $I \subseteq A$, $J \subseteq B$ the element of N_e corresponding to the I-th row and the J-th column is ν_{IJ}. Here the labeling for the rows and columns is induced by the ordering of the rows and columns for P_e (c.f. Definition 2.1), i.e. all the subsets of A and B are coded as $\{0,1\}$-vectors and we introduce the lexicographic order on the vectors with the vector of ones being the last one.

The following result allows us to rephrase equations from Theorem 2.1 in terms of the new coordinates.

Proposition 7.1. *Let $T = (V, E)$ be a tree and let let P be a probability distribution of a vector $X = (X_1, \ldots, X_n)$ of binary variables represented by the leaves of T. If $e \in E$ is an edge of T inducing a non-trivial split then* $\mathrm{rank}(P_e) = 2$ *if and only if* $\mathrm{rank}(N_e) = 1$.

Proof. We will show that by using elementary operations that do not change the determinant we can obtain from the flattening matrix $P_e = [p_{\alpha\beta}]$ induced by a split $(A_1)(A_2)$ a block diagonal matrix $\widehat{N}_e$ with 1 as the first scalar block and matrix N_e as the second block. It will then follows that $\mathrm{rank}\,(P_e) = 2$ if and only if $\mathrm{rank}\,(N_e) = 1$.

First note that the flattening matrix P_e can be transformed to the flattening of the non-central moments just by adding rows and columns according to Equation (4) and then to the flattening of the central moments $M_e = [\mu_{IJ}]$ such that $I \subseteq A_1$, $J \subseteq A_2$. It therefore suffices to show that we can obtain $\widehat{N}_e$ from M_e using elementary operations.

Let $I \subseteq A_1$, $J \subseteq A_2$ then for each $x \in \Pi_{T(IJ)}$ there is at most one block containing elements from both I and J, for otherwise removing e would increase number of blocks in x by more than one which is not possible. Denote the block by $(I'J')$ where $I' \subseteq I$, $J' \subseteq J$. Note that by construction we have either $I', J' = \emptyset$ if $x \geq (A_1)(A_2)$ or $I', J' \neq \emptyset$ otherwise. Set $\widehat{\nu}_{\emptyset\emptyset} = 1$, $\widehat{\nu}_{I\emptyset} = \widehat{\nu}_{\emptyset J} = 0$ and $\widehat{\nu}_{IJ} = \nu_{IJ}$ for all nonempty $I \subseteq A_1$, $J \subseteq A_2$. The elements $\widehat{\nu}_{IJ}$ form matrix $\widehat{N}_e$. We can rewrite

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

Equation (24) splitting the blocks

$$\mu_{IJ} = \sum_{x \in \Pi_T(IJ)} \left(\widehat{\nu}_{I'J'} \prod_{I \supseteq B \in x} \nu_B \prod_{J \supseteq B \in x} \nu_B \right) = \sum_{I' \subseteq I} \sum_{J' \subseteq J} u_{II'} \widehat{\nu}_{I'J'} v_{J'J} \tag{25}$$

for some $u_{II'}$, $v_{J'J}$. Setting $u_{II'} = 0$ for $I \not\subseteq I'$, $v_{J'J} = 0$ for $J \not\subseteq J'$ we can write the coefficients in matrices U, V. By construction the matrices U and V are lower and uppertriangular respectively. Since $u_{II} = 1$ for all $I \subseteq A_1$ and $v_{JJ} = 1$ for all $J \subseteq A_2$ we have $\det U = \det V = 1$. Matrix U records the row operations on $\widehat{N}_e$ and V records the column operations.

□

The proposition shows that the vanishing of all 3×3 minors of all the edge flattenings of P plus the trivial invariant $\sum p_\alpha = 1$ is equivalent to the vanishing all 2×2 minors of all edge flattenings of $\nu = (\nu_I)_{I \in [n]_{\geq 2}}$. As an immediate corollary we get the following theorem.

Theorem 7.2. *Let $T = (V, E)$ be a trivalent tree. Then the general Markov model $\mathcal{M}_T$ is defined by the following set of equations: for each split $(A)(B)$ induced by an edge consider any four nonempty sets $I_1, I_2 \subset A$, $J_1, J_2 \subset B$ then equations of the form $\nu_{I_1 J_1} \nu_{I_2 J_2} = \nu_{I_1 J_2} \nu_{I_2 J_1}$ generate the phylogenetic ideal defining $\mathcal{M}_T$.*

In [12] Nicholas Eriksson noted that some of invariants usually prove to be better in discriminating between different tree topologies than the others. His simulations showed that the invariants related to the four-point condition were especially powerful. The binary case we consider in this paper can give some partial understanding of why it might be so. Here, the invariants related to the four-point condition are only those involving second order correlations. By Corollary 4 we could essentially constrain our analysis to second order correlations. The projection of $\mathcal{R}_n$ or $\mathcal{C}_n$ to this space is constrained by the condition that the correlation matrix $[\rho_{ij}]$ is positive semi-definite. Moreover, the projection of $\mathcal{M}_T$ is described by inequalities in Theorem 4.2 plus some equations of the form

$$\mu_{ij}\mu_{kl} = \mu_{il}\mu_{kj} \quad \text{for all } i, k \in A, j, l \in B,$$

for all edge flattenings $(A)(B)$. And these are exactly the invariants related to the four-point condition.

There is also another argument for constraining to lower-order correlations or equivalently the lower order tree-cumulants. In general, the estimates of the higher-order moments (or cumulants) are sensitive to outliers and their variance generally grows with the order of the moment. Let $\hat{\mu}$ be a sample estimator of the central moments μ and let f be one of the equations in Theorem 7.2. Then using the delta method we have

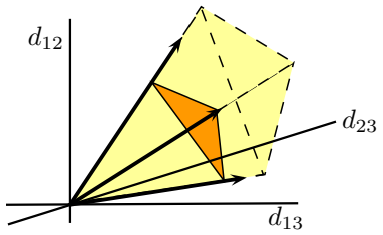
$$\mathrm{Var}(f(\hat{\mu})) \simeq \nabla f(\mu)^t \mathrm{Var}(\hat{\mu}) \nabla f(\mu).$$

Consequently, the higher order of the correlations involved the higher variability of the invariant (see [24, Section 4.5]). This shows that invariants involving lower-order moments should be of a greater value in practice.

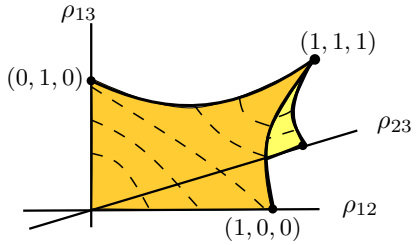
CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

8. Examples

8.1. **The tripod tree.** The new parametrization allows us to get an in-depth understanding of the geometry of the tripod model (c.f. Example 1.1). The trivial component corresponding to all possible values of $(\rho_{111}, \rho_{222}, \rho_{333})$ is $\mathbb{R}^3$. The parametrization of $\mathcal{M}_T^\kappa$ is $\psi_T : \mathbb{R} \times [-1,1]^3 \to \mathbb{R} \times [-1,1]^3$ given by equations in (10). It can be checked that ρ_{123} can be any real number as well as long as all edge correlations are non-zero. So for a moment we can restrict ourselves to the second-order correlations. By considerations of Section 4, modulo the signs we can equivalently deal with the cone of all possible tree metrics on three leaves (c.f. Theorem 4.1). The four-point condition for a triple $(d_{12}, d_{13}, d_{23}) \in \mathbb{R}^3_{\geq 0}$ translates to $d_{ij} \leq d_{ik} + d_{jk}$ for all $i, j, k \in \{1, 2, 3\}$ and hence the cone of all possible metrics is a polyhedral cone generated by $[1, 1, 0]$, $[1, 0, 1]$ and $[0, 1, 1]$.



To get the set in terms of the absolute values of the correlations we transform it back using $[x, y, z] \mapsto [\exp(-x), \exp(-y), \exp(-z)]$ and take the closure. This gives a set which is sketched below.

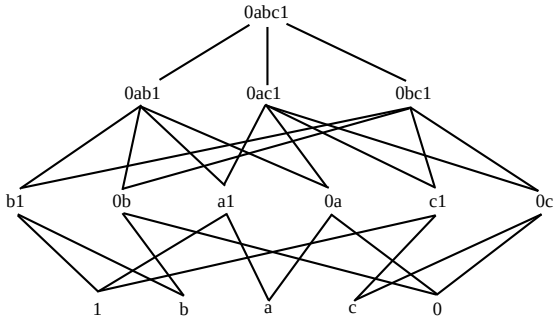


By Remark 6 the model considered in phylogenetic analysis is just the Cartesian product of the interior of this set with $\mathbb{R}^4$ (the trivial part plus a choice of ρ_{123}). However in general we do not assume that second-order correlations are strictly positive. Denote by $\mathcal{M}$ the four copies of the above set related to the four possible sign patterns of $(\rho_{12}, \rho_{13}, \rho_{23})$ (we have $\rho_{12}\rho_{13}\rho_{23} \geq 0$). Apart from the boundary points, the tripod model is just a Cartesian product of $\mathbb{R}^4$ with $\mathcal{M}$. However, the model has more complicated structure and is not just a Cartesian product of two simple spaces. Bad behavior can be spotted on the intersection of $\mathcal{M}$ with hyperplanes of the form $\rho_{12} = 0$, $\rho_{13} = 0$ and $\rho_{23} = 0$. Here we get disconnected graph models. Indeed $\rho_{ij} = 0$ if and only if $X_i \perp\!\!\!\perp X_j$. If one of ρ_{ij} is zero then necessarily $\rho_{123} = 0$ and hence we loose our product structure. Proceeding carefully we obtain a good understanding of other boundary points of $\mathcal{M}$.

Label the vertices of $\mathcal{M}$ as follows $0 : (0,0,0)$, $a : (1,0,0)$, $b : (0,1,0)$, $c : (0,0,1)$ and $1 : (1,1,1)$. And for example by $0ab1$ we denote the two dimensional face of $\mathcal{M}$ with vertices given by 0, a, b and 1. Three segments $a1$, $b1$, $c1$ parametrize a

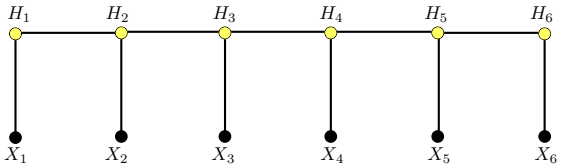
CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

situation of two variables being functionally linked. In this case the model degenerates and the defining graph is a clique with two manifest nodes. In general all the faces of $\mathcal{M}$ correspond to different graphical submodels of the tripod model. The correspondence between the faces and the graphs links to considerations in [25] linking the space of phylogenetic oranges and the Tuffley poset. The poset of faces of $\mathcal{M}$ is given by



Note that it corresponds exactly to the Tuffley poset depicted on Figure 2 in [25] (c.f. Theorem 3.3. therein). Moreover, all the faces giving that some of the observable variables are functionally linked correspond to degenerate distributions and hence lie on the boundary of the probability simplex.

8.2. **The Hidden Markov model.** Let $(X_i, H_i)_{i=1}^n$ be a sequence of binary random variables. The hidden Markov model is given by a caterpillar tree for which all the inner nodes have degree at most three.



In this case it is easy to give explicit formulas for κ's in terms of ρ's given by Equation (14). One can check that in the case of a caterpillar tree $\Pi_{T(I)}$ for each $I \in [n]_{\geq 2}$ is isomorphic to the set of all subsets of a set with $|I|-1$ elements. It follows that the Möbius function is given by $\mu_I(0,\pi) = (-1)^{|\pi|-1}$ where $|\pi|$ denotes the number of blocks in π. We can use it in Equation (14).

In general the parametrization defining the model is such that, for any $2 \leq k \leq n$ and any k indices $1 \leq i_1 < \ldots < i_k \leq n$, we have

$$\kappa_{i_1 \ldots i_k} = \prod_{(u,v) \in \mathcal{P}_T(h_{i_1}, h_{i_k})} \rho^h_{uv} \cdot \prod_{j=1}^{k} \rho^x_{i_j} \prod_{j=2}^{k-1} \rho_{i_j i_j i_j},$$

where ρ^h corresponds to correlations between adjacent inner nodes and ρ^x to the correlations between a leaf and an adjacent inner edge.

If the process relating the hidden variables is a homogeneous Markov chain starting from the stationary distribution then it can be checked that $\rho_{h_{i-1}h_i} = p(1|1) - p(1|0) = \rho_h \in [-1,1]$ for each $i = 1, \ldots, n$, where $p(\cdot|\cdot)$ refers to an element of the transition matrix $\mathbb{P}(H_i|H_{i-1})$. Moreover, if the conditional probabilities of the leaves given the parents are the same for each time we also have $\rho_{h_i x_i} = q(1|1) - q(1|0) = \rho_x \in [-1,1]$ for each $i = 1, \ldots, n$, where $q(\cdot|\cdot)$ is the

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

transition matrix coding $\mathbb{P}(X_i|H_i)$. Since we started in a stationary distribution we have in addition that $\mathbb{E}H_i = \lambda$ for all $i = 1, \ldots, n$ or equivalently $\rho_{h_i h_i h_i} = s$ for some $s \in \mathbb{R}$ and each $i = 1, \ldots, n$. In this simple case $\kappa_{[n]} = \rho_h^{n-1}\rho_x^n s^{n-2}$, $\kappa_{i,i+k} = \rho_{i,i+k} = \rho_h^k \rho_x^2$. In general for any $2 \leq k \leq n$ and any k indices $1 \leq i_1 < \ldots < i_k \leq n$ we have

$$\kappa_{i_1 \cdots i_k} = \rho_h^{i_k - i_1} \rho_x^k s^{k-2}.$$

This shows that this model is relatively simple with only three free parameters and a nice monomial parametrization.

9. Discussion

The proposed new coordinate system has allowed us to get a better insight into geometry of phylogenetic tree models with binary observations. The elegant form of the parametrization we obtain can be used to obtain some generalizations of the formulas for Bayesian information criteria in [30]. We will report these results in a later paper. We also believe that it can be used to derive asymptotic distributions of certain likelihood ratio statistics.

The coordinate system also rises a question about applicability of Möbius function in statistics. The way we define the coordinates mirrors the combinatorial definition of cumulants. A similar idea is exploited in the theory of free probabilities (see e.g. [33]). We believe that our approach can be extended to more general families of graphical models.

The link with tree metrics may possibly extend tools for analyzing this type of models. An extension for more general models is also possible. In a straightforward way we can consider models of trees such that only some of the inner nodes are hidden. Since the crucial assumption was that the hidden variables are binary probably we could also obtain similar results in the case when the observable variables are not binary.

Since $\widehat{\mathcal{M}}_T$ forms a quadratic exponential family (see [21]) its geometry is relatively simple [16] [18] and in some sense similar to tree models for Gaussian variables (see [10]). This partly explains why some of our results mirror the results obtained in [39]. It may be an interesting problem to understand in a better way the relationship between those two situations.

Acknowledgements

We are especially grateful to Diane Maclagan and an anonymous referee for an uncountable number of helpful comments. We would also like to thank Bernd Sturmfels for a stimulating discussion and for giving the idea of a link between our parametrization and the tree-metric. The first author would like to thank John Rhodes for giving an exciting introduction to the phylogenetic analysis and for providing a deeper motivation for our work.

References

[1] E. S. Allman and J. A. Rhodes, *Phylogenetic ideals and varieties for the general Markov model*, Adv. in Appl. Math., 40 (2008), pp. 127–148.

[2] V. Auvray, P. Geurts, and L. Wehenkel, *A Semi-Algebraic Description of Discrete Naive Bayes Models with Two Hidden Classes*, in Proc. Ninth International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, Florida, Jan 2006.

[3] P. Buneman, *A note on the metric properties of trees*, J. Combinatorial Theory Ser. B, 17 (1974), pp. 48–50.

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

[4] M. Casanellas and J. Fernandez-Sanchez, *Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees*, Molecular Biology and Evolution, 24 (2007), p. 288.
[5] J. Cavender and J. Felsenstein, *Invariants of phylogenies in a simple case with discrete states*, Journal of Classification, 4 (1987), pp. 57–71.
[6] J. A. Cavender, *Letter to the editor*, Molecular Phylogenetics and Evolution, 8 (1997), pp. 443 – 444.
[7] J. Chang, *Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency*, Mathematical Biosciences, 137 (1996), pp. 51–73.
[8] D. A. Cox, J. B. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms*, Springer-Verlag, NY, 3rd ed., 2007.
[9] D. R. Cox, *The analysis of multivariate binary data*, Applied Statistics, 21 (1972), pp. 113–120.
[10] D. R. Cox and N. Wermuth, *A note on the quadratic exponential binary distribution*, Biometrika, 81 (1994), pp. 403–408.
[11] M. Drton and S. Sullivant, *Algebraic Statistical Models*, Statistica Sinica, 17 (2007), pp. 1273–1297.
[12] N. Eriksson, *Using invariants for phylogenetic tree construction*, vol. 149 of The IMA Volumes in Mathematics and its Applications, Springer, 2007, pp. 89–108.
[13] N. Eriksson, K. Ranestad, B. Sturmfels, and S. Sullivant, *Phylogenetic algebraic geometry*, in Projective varieties with unexpected properties, Walter de Gruyter GmbH & Co. KG, Berlin, 2005, pp. 237–255.
[14] W. Feller, *An Introduction to Probability Theory and Applications*, vol. 2, John Wiley & Sons, New York, second ed., 1971.
[15] L. Garcia, M. Stillman, and B. Sturmfels, *Algebraic geometry of Bayesian networks*, J. Symbolic Comput, 39 (2005), pp. 331–355.
[16] D. Geiger, D. Heckerman, H. King, and C. Meek, *Stratified exponential families: graphical models and model selection*, Ann. Statist., 29 (2001), pp. 505–529.
[17] D. Geiger and C. Meek, *Graphical models and exponential families*, in Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, Madison, WI, August 1998, pp. 156–165.
[18] D. Geiger, C. Meek, and B. Sturmfels, *On the toric algebra of graphical models*, Annals of Statistics, 34 (2006), pp. 1463–1492.
[19] M. G. Kendall and A. Stuart, *The advanced theory of statistics. Vol. 2*, Hafner Publishing Co., New York, third ed., 1973. Inference and relationship.
[20] J. Lake, *A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony*, 1987.
[21] S. L. Lauritzen, *Graphical models*, vol. 17 of Oxford Statistical Science Series, The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.
[22] P. Lazarsfeld and N. Henry, *Latent structure analysis*, Houghton, Mifflin, New York, 1968.
[23] F. Matsen, *Fourier transform inequalities for phylogenetic trees*, Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 6 (2009), pp. 89–95.
[24] P. McCullagh, *Tensor methods in statistics*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1987.
[25] V. Moulton and M. Steel, *Peeling phylogenetic oranges*, Advances in Applied Mathematics, 33 (2004), pp. 710–727.
[26] J. Pearl and M. Tarsi, *Structuring causal trees*, J. Complexity, 2 (1986), pp. 60–77. Complexity of approximately solved problems (Morningside Heights, N.Y., 1985).
[27] G. Pistone and H. P. Wynn, *Cumulant varieties*, Journal of Symbolic Computation, 41 (2006), pp. 210–221.
[28] G. Rota, *On the foundations of combinatorial theory I. Theory of Möbius Functions*, Probability Theory and Related Fields, 2 (1964), pp. 340–368.
[29] G.-C. Rota and J. Shen, *On the combinatorics of cumulants*, J. Combin. Theory Ser. A, 91 (2000), pp. 283–304. In memory of Gian-Carlo Rota.
[30] D. Rusakov and D. Geiger, *Asymptotic model selection for naive Bayesian networks*, J. Mach. Learn. Res., 6 (2005), pp. 1–35 (electronic).
[31] C. Semple and M. Steel, *Phylogenetics*, vol. 24 of Oxford Lecture Series in Mathematics and its Applications, Oxford University Press, Oxford, 2003.

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism

[32] R. Settimi and J. Q. Smith, *Geometry, moments and conditional independence trees with hidden variables*, Ann. Statist., 28 (2000), pp. 1179–1205.
[33] R. Speicher, *Free probability theory and non-crossing partitions*, Sém. Lothar. Combin., 39 (1997), pp. Art. B39c, 38 pp. (electronic).
[34] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell, *Bayesian analysis in expert systems*, Statist. Sci., 8 (1993), pp. 219–283. With comments and a rejoinder by the authors.
[35] M. Steel and B. Faller, *Markovian log-supermodularity, and its applications in phylogenetics*, Applied Mathematics Letters, (2009).
[36] B. Streitberg, *Lancaster interactions revisited*, Ann. Statist., 18 (1990), pp. 1878–1885.
[37] B. Sturmfels, *Solving systems of polynomial equations*, vol. 97 of CBMS Regional Conference Series in Mathematics, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2002.
[38] B. Sturmfels and S. Sullivant, *Toric Ideals of Phylogenetic Invariants*, Journal of Computational Biology, 12 (2005), pp. 204–228.
[39] S. Sullivant, *Algebraic geometry of Gaussian Bayesian networks*, Advances in Applied Mathematics, 40 (2008), pp. 482–513.

Piotr Zwiernik, University of Warwick, Department of Statistics, CV7AL, Coventry, UK.
E-mail address: p.w.zwiernik@warwick.ac.uk

James Q. Smith, University of Warwick, Department of Statistics, CV7AL, Coventry, UK.

CRiSM Paper No. 09-32, www.warwick.ac.uk/go/crism