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, where leaves represent the observable variables and where 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. Our analysis is based on combinatorial results generalizing the notion of cumulants and introduce a novel use of Möbius functions on partially ordered sets. The geometric structure we obtain links to the notion of a tree metric considered in phylogenetic analysis and to some interesting determinantal formulas involving hyperdeterminants of $2 \times 2 \times 2$ tables as defined in [19].

1. INTRODUCTION

Discrete graphical models have become a very popular tool in the statistical analysis of multivariate problems (see e.g. [22][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, when variables are discrete the model is described by polynomial equations in the ambient model space, maximum likelihood estimates are simple sample proportions and a 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 consequent additional inequality constraints tend to destroy the simple forms of maximum likelihood estimates 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]). One reason that these constraints are of practical importance is that when the sample size is not too large or if the sampling scheme is even slightly contaminated then the likelihood is often maximized on the boundaries of the parameter space. The constraints are therefore active and this can strongly affect input on the ensuing likelihood based inference, whether this is classical or Bayesian.

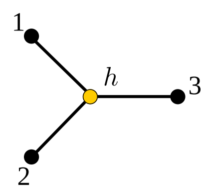
The motivation of this paper is to study the semi-algebraic geometry of underlying phylogenetic tree models over a collection of binary random variables. Phylogenetic analysis is based on Markov processes on trees which have the property

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

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 that all probabilities are strictly positive 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 [21], and Cavender and Felsenstein [6]. These invariant algebraic relations, expressed as zeros of a set of polynomial equations, over the observed probability tables must hold for a given phylogenetic model to be valid. The study of these algebraic structures has been recently embraced by computational algebraic geometers [1][13][39] and the consequent advances in understanding of these invariants now begins to clarify the statistical analyses of such models [5].

The main technical problem related to phylogenetic invariants is that they do not give a full geometric description of the statistical model. There are some nontrivial polynomial inequalities which also have to be satisfied.

Example 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 the 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 it is not a saturated model and not all the probability distributions lie in the model. Lazarsfeld [23, 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 the phylogenetic invariants do not give a complete specification of even a simple probability model like this one. Therefore inference based solely on the invariants is incomplete and derived estimates can be infeasible within the model class.

This example motivated the closer investigation of the geometry of these models. 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 [36]. The most important applications in biology involve variables that can take four values. Recently Matsen [24] 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 these constraints can be more easily expressed as relations among the central moments as in the example above.

Our analysis of moment structures induced by the models under consideration led us to an interesting application of the theory of partially ordered sets and its Möbius functions. Similar methods were used in the theory of graphical models (see for example the proof of Theorem 3.9 in [22]) for a poset of all subsets of a finite set and in the combinatorial theory of cumulants [29] for a poset of all partitions of a finite set. To our knowledge this paper is the first approach to use more general posets in the context of statistical analysis. It allows us to construct a useful reparametrization for our models. When expressed in this new coordinate system the underlying geometry of the models becomes transparent. 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 [8][32]. Combining these with 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 formulas we obtain involve determinants of the marginal 2×2 probability tables and the $2\times 2\times 2$ hyperdeterminant of three dimensional marginal probability tables as defined in [19] suggesting that this might be a general phenomenon for conditional independence models with hidden variables.

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 use central moments to describe general Markov models in a more efficient way. At the end of the section we also state the main theorem of the paper. In Section 4 we construct another, more intrinsic, coordinate system for the class of binary tree models. In the new coordinate system the parametrization of the model has a quasi-monomial form. This gives a better insight into the underlying geometry. In Section 5 we use the parametrization developed earlier to find an alternative form of equations given by Allman and Rhodes [1]. These are slightly simpler from the algebraic point of view and have a more transparent statistical interpretation. We prove our main theorem in Section 6. It gives the geometric description of general Markov models in the case of trivalent trees in terms of equations and inequalities involving moments between observable variables. The paper is concluded with a short discussion.

2. Independence models on trees

In this section we introduce models defined by global Markov properties on trees and models for rooted trees. We show the relations between them introducing the general Markov model on a tree which is the main subject of our study.

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. For a directed edge $(u, v)\in E$ we say that u is a *parent* of v. The set of all parents of $v\in V$ is denoted by $\mathrm{pa}(v)$. For a directed graph G a *moral graph* of G is an undirected graph obtained by joining all parents of a node by edges between each pair for all nodes in the graph and then changing all directed edges for undirected ones. 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*. 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 one is called a *leaf*. A vertex of T that is not a leaf is called an *inner node*. An edge of T is *inner* if both of its ends are inner 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 such that $v \in V \setminus r$ the set $\mathrm{pa}(v)$ is a singleton.

2.2. **Models defined by global Markov properties.** In this paper we always assume that random variables are binary taking one of the values in $\{0, 1\}$. 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 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. [22, 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. $Y_A = (Y_i)_{i\in A}$.

Statistical hypotheses often translate into algebraic relations defining a model. For example if Y_1, Y_2 are two independent binary random variables then with $P = [p_\alpha]$ denoting the joint probability table we have $\det P = p_{00}p_{11} - p_{01}p_{10} = 0$. Since $p_{00} + p_{01} + p_{10} + p_{11} = 1$ we can equivalently write the first equation as $p_{11} - (p_{10} + p_{11})(p_{01} + p_{11}) = 0$. However this is just a condition for the covariance between the two variables to be zero. So a probability distribution satisfies $X_1 \perp\!\!\!\perp X_2$ if and only if $\det P = 0$ if and only if $\mathrm{Cov}(X_1, X_2) = 0$.

More generally conditional independence statements imply that a set of polynomial equations in $(p_\alpha)_{\alpha\in\{0,1\}^{|V|}}$ holds 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 [9]) this collection of polynomials forms an ideal which we denote by I_{global} (see [15][38, Section 8.1]).

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 and we consider marginal distributions over the leaves 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 $(\alpha_1, \ldots, \alpha_{|V|}) \in \{0, 1\}^{|V|}$

$$p_\alpha(\theta) = \prod_{v\in V} \theta^{(v)}_{\alpha_v|\alpha_{\mathrm{pa}(v)}}, \tag{2}$$

where $\mathrm{pa}(r)$ is the empty set, $\theta = (\theta^{(v)}_{\alpha_v|\alpha_{\mathrm{pa}(v)}})$ and $\theta^{(v)}_{\alpha_v|\alpha_{\mathrm{pa}(v)}} = \mathbb{P}(Y_v = \alpha_v | Y_{\mathrm{pa}(v)} = \alpha_{\mathrm{pa}(v)})$. Since $\theta^{(r)}_0 + \theta^{(r)}_1 = 1$ and $\theta^{(v)}_{0|i} + \theta^{(v)}_{1|i} = 1$ for all $v \in V \setminus \{r\}$ and $i = 0, 1$ then the set of parameters consists of exactly $2|E| + 1$ free parameters: we have two parameters: $\theta^{(v)}_{1|0}$, $\theta^{(v)}_{1|1}$ for each edge $(u, v) \in E$ and one parameter $\theta^{(r)}_1$ for the root. We denote the parameter space by $\Theta_T = [0, 1]^{2|E|+1}$.

Standard results in the theory of graphical models tell us that if al probabilities in the 2×2 tables over adjacent variables are strictly positive the Markov process on T is equal to $\widehat{\mathcal{M}}_T$. Indeed, by [22, Theorem 3.27] the parametrization defined by (2) is equivalent to its directed global Markov properties on T^r. Moreover, since T^r has a uniquely defined root the moral graph of T^r (c.f. Section 2.1) 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 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 restrict ourselves to the interior of Δ_{2^n-1}. Equation (2) induces a polynomial map $f_T : \Theta_T \to \Delta_{2^n-1}$ obtained by marginalization over all the inner nodes of T

$$p_{\alpha_{[n]}}(\theta) = \sum_{\mathcal{H}} \prod_{v \in V} \theta^{(v)}_{\alpha_v|\alpha_{\mathrm{pa}(v)}}, \tag{3}$$

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 (c.f. [31, Section 8.3]). By definition these are models for the rooted tree T^r defined by (3). Note that with the positivity assumption the general Markov model is equivalent to $\mathcal{M}_T$. Moreover, since Θ_T is a *semi-algebraic set* (defined by polynomial equations and inequalities) then by the Tarski-Seidenberg theorem [3, Section 2.5.2] $\mathcal{M}_T$ is a semi-algebraic set as well.

3. Central moments and tree models

3.1. Moments and conditional independence. In this section we start by introducing a set of coordinates which will be useful to understand the geometry of $\mathcal{M}_T$. Let $X = (X_1, \ldots, X_n)$ be a random vector. Then we can obtain formulas relating the moments of these variables to the probability distribution of X. For each $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$ denote $X^\beta = \prod_i X_i^{\beta_i}$ and define $\lambda_\beta = \mathbb{E}X^\beta$ and $\mu_\beta = \mathbb{E}U^\beta$, where $U_i = X_i - \mathbb{E}X_i$. Below we give formulas for maps giving the reparametrization from the raw probabilities to central moments.

First we perform the reparametrization from the raw probabilities $p = [p_\alpha]$ to the non-central moments $\lambda = [\lambda_\alpha]$ for $\alpha = (\alpha_1, \ldots, \alpha_n) \in \{0, 1\}^n$. This is a linear map $f_{p\lambda} : \mathbb{R}^{2^n} \to \mathbb{R}^{2^n}$, where $\lambda = f_{p\lambda}(p)$ is defined as follows

$$\lambda_\alpha = \sum_{\alpha \leq \beta \leq \mathbf{1}} p_\beta \qquad \text{for any } \alpha \in \{0, 1\}^n, \tag{4}$$

where $\mathbf{1}$ denotes here the vector of ones and the sum is over all binary vectors β such that $\alpha \leq \beta \leq \mathbf{1}$ in the sense that $\alpha_i \leq \beta_i \leq 1$ for all $i = 1, \ldots, n$. In particular $\lambda_{\mathbf{0}} = 1$ for all probability distributions and hence the image $f_{p\lambda}(\Delta_{2^n-1})$ is contained in the hyperplane defined by $\lambda_{\mathbf{0}} = 1$.

The linearity of the expectation implies that the central moments can be expressed in terms of non-central moments. Thus

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

where $|\beta| = \sum_i \beta_i$. Using these equations we can transform variables from the non-central moments $[\lambda_\alpha]$ to another set of variables given by all the means $\lambda_{e_1}, \ldots, \lambda_{e_n}$, where $e_1, \ldots, e_n$ are unit vectors in $\mathbb{R}^n$, and central moments $[\mu_\alpha]$ for $\alpha \in \{0, 1\}$. The polynomial change of variables $f_{\lambda\mu} : \mathbb{R}^{2^n} \to \mathbb{R}^n \times \mathbb{R}^{2^n}$ is defined by (5) and on the first n coordinates it takes the values of $\lambda_{e_1}, \ldots, \lambda_{e_n}$. Denote $\mathcal{C}_n = (f_{\lambda\mu} \circ f_{p\lambda})(\Delta_{2^n-1})$ which is contained in a subspace of $\mathbb{R}^n \times \mathbb{R}^{2^n}$ given by

$$\mu_{\mathbf{0}} = 1 \quad \text{and} \quad \mu_{e_1} = \cdots = \mu_{e_n} = 0.$$

We can also easily define the inverse maps of $f_{p\lambda}$ and $f_{\lambda\mu}$. The map $f_{\lambda p} = f_{p\lambda}^{-1} : \mathbb{R}^{2^n} \to \mathbb{R}^{2^n}$ is given by

$$p_\alpha = \sum_{\alpha\leq\beta\leq\mathbf{1}} (-1)^{|\beta-\alpha|}\lambda_\beta \quad \text{for } \alpha = (\alpha_1, \ldots, \alpha_n) \in \{0,1\}^n \tag{6}$$

whilst $f_{\mu\lambda} = f_{\lambda\mu}^{-1} : \mathbb{R}^n \times \mathbb{R}^{2^n} \to \mathbb{R}^{2^n}$ is given by

$$\lambda_\alpha = \sum_{\mathbf{0}\leq\beta\leq\alpha} \mu_{\alpha-\beta}\prod_{i=1}^{n}\lambda_{e_i}^{\beta_i} \quad \text{for } \alpha \in \{0,1\}^n. \tag{7}$$

Moreover, $f_{\lambda\mu} \circ f_{p\lambda}$ and $(f_{\lambda p} \circ f_{\mu\lambda})^{-1}$ are equal as polynomial functions on Δ_{2^n-1}. Since all the maps defined above are regular polynomial maps with regular polynomial inverses they constitute an isomorphism between Δ_{2^n-1} and $\mathcal{C}_n$.

To simplify notation henceforth we will index moments not with $\{0,1\}^n$ but with the set of subsets of $[n]$. Here the set $A \subseteq [n]$ is identified with $\alpha \in \{0,1\}^n$ such that $\alpha_i = 1$ for all $i \in A$ and it is zero elsewhere. In particular for each $i \in [n]$ we write λ_i for λ_{e_i}. Let $[n]_{\geq k}$ denote the set of all subset of $[n]$ with at least k elements. 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 $\mu_{IJ} = \mu_I\mu_J$ for all nonempty $I \subseteq A$, $J \subseteq B$. But since the variables are binary we also have a converse result. If for all nonempty $I \subseteq A$, $J \subseteq B$ we have that $\mu_{IJ} = \mu_I\mu_J$ 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 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 $\beta \in \{0,1\}^{|B|}$ or equivalently $\mathrm{Cov}(U_A^\alpha, U_B^\beta) = 0$ for each non-zero $\alpha \in \{0,1\}^{|A|}$ and $\beta \in \{0,1\}^{|B|}$ which can be written as $\mu_{IJ} = \mu_I\mu_J$ for each nonempty $I \subseteq A$, $J \subseteq B$.

We can generalize the result above. For a random variable H_a let $\lambda_a = \mathbb{E}H_a$ and $U_a = H_a - \lambda_a$. For each $I \subseteq [n]$ let $U_I = \prod_{i\in I} U_i$ and $\eta_{a,I} = \mathbb{E}(U_I U_a)/\mathrm{Var}(H_a)$ where $\mathrm{Var}(H_a) = \lambda_a(1-\lambda_a)$. We have $X_A \perp\!\!\!\perp X_B | H_a$ if and only if for all nonempty

$I \subseteq A$, $J \subseteq B$

$$\begin{aligned} \mu_{IJ} &= \mu_I \mu_J + \lambda_a(1-\lambda_a)\eta_{a,I}\eta_{a,J}, \\ \eta_{a,IJ} &= \mu_I \eta_{a,J} + \eta_{a,I}\mu_J + (1-2\lambda_a)\eta_{a,I}\eta_{a,J}. \end{aligned} \tag{8}$$

Indeed, an equivalent condition for $X_A \perp\!\!\!\perp X_B | H_a$ can be written in terms of conditional covariances. Thus for each $I \subseteq A$, $J \subseteq B$ we have

$$\begin{aligned} \mathrm{Cov}(U_I, U_J | H_a = 0) = 0, \\ \mathrm{Cov}(U_I, U_J | H_a = 1) = 0. \end{aligned} \tag{9}$$

This set of equations is equivalent to

$$\begin{aligned} &\lambda_a \mathrm{Cov}(U_I, U_J | H_a = 1) + (1-\lambda_a)\mathrm{Cov}(U_I, U_J | H_a = 0) = 0, \\ &\mathrm{Cov}(U_I, U_J | H_a = 0) - \mathrm{Cov}(U_I, U_J | H_a = 1) = 0. \end{aligned} \tag{10}$$

Because any function of H_a is necessarily a linear function it follows that for any $I \subseteq [n]$

$$\mathbb{E}(U_I | H_a) = \mu_I + \eta_{a,I} U_a, \tag{11}$$

and hence

$$\mathrm{Cov}(U_I, U_J | H_a) = \mu_{IJ} - \mu_I \mu_J + (\eta_{a,IJ} - \eta_{a,I}\mu_J - \mu_I \eta_{a,J}) U_a - \eta_{a,I}\eta_{a,J} U_a^2.$$

Using this formula it is now straightforward to check that (10) is equivalent to (8).

3.2. **Reparametrization for general Markov models.** Let $f_{p\mu}$ denote the restriction of $f_{\lambda\mu} \circ f_{p\lambda}$ to Δ_{2^n-1}. We showed earlier that $f_{\lambda\mu} : \Delta_{2^n-1} \to \mathcal{C}_n$ is an isomorphism. Hence we can investigate the geometry of $\mathcal{M}_T$ in the new coordinate system. A similar approach is presented for example in [17][32]. We denote $\mathcal{M}_T^\mu = f_{p\mu}(\mathcal{M}_T) \subseteq \mathcal{C}_n$.

The next step is to reparametrize the parameter space of a tree model. Let $T = (V, E)$ be a rooted tree with n leaves and root r. Note that for a tree $1 + 2|E| = |V| + |E|$ so the number of free parameters in (2) and (3) is $|V| + |E|$. We define a polynomial map $f_{\theta\omega} : \mathbb{R}^{|V|+|E|} \to \mathbb{R}^{|V|+|E|}$ from the original set of parameters of Θ_T given by the root distribution and the conditional probabilities for each of the edges to a set of parameters given as follows: for every directed edge $(u, v) \in E$

$$\eta_{uv} = \theta^{(v)}_{1|1} - \theta^{(v)}_{1|0} \in [-1, 1] \quad \text{and} \tag{12}$$

$$\bar{\mu}_v = 1 - 2\lambda_v \in [-1, 1] \quad \text{for each inner node } v \in V,$$

where $\lambda_v = \mathbb{E} Y_v$ is a polynomial in the original parameters θ of degree depending on the distance of v from the root r. Indeed, let $r, v_1, \ldots, v_k, v$ be a directed path in T. Then

$$\lambda_v = \sum_{\alpha \in \{0,1\}^{k+1}} \theta^{(v)}_{1|\alpha_k} \theta^{(v_k)}_{\alpha_k|\alpha_{k-1}} \cdots \theta^{(r)}_{\alpha_r}.$$

The inverse map $f_{\omega\theta} : \mathbb{R}^{|V|+|E|} \to \mathbb{R}^{|V|+|E|}$ has an even simpler form. For each edge $(u, v) \in E$ we have

$$\begin{aligned} \theta^{(v)}_{1|0} &= \frac{1 + \bar{\mu}_v}{2} - \eta_{uv} \frac{1 + \bar{\mu}_u}{2}, \\ \theta^{(v)}_{1|1} &= \frac{1 + \bar{\mu}_v}{2} + \eta_{uv} \frac{1 - \bar{\mu}_u}{2} \end{aligned}$$

and $\theta^{(r)}_1 = \frac{1+\bar{\mu}_r}{2}$. Writing $\Omega_T = f_{\theta\omega}(\Theta_T)$ it follows that $f_{\theta\omega}$ is an isomorphism between Θ_T and Ω_T.

It is simple to check that $\eta_{uv} = \mathbb{E}(U_u U_v)/\mathrm{Var Y_u}$ where $\mathrm{Var}(Y_u) = \frac{1}{4}(1-\bar{\mu}_u^2)$. It follows that η_{uv} defined above coincides with the definition of $\eta_{u,v}$ just preceding equation (8), where in one notation we omit the coma. By definition $\eta_{u,v}$ is a linear regression coefficient of Y_v against Y_u, i.e. $\mathbb{E}(Y_v - \mathbb{E}Y_v|Y_u) = \eta_{uv}(Y_u - \mathbb{E}Y_u)$. When $(u,v) \in E$ then $\eta_{u,v}$ is a parameter of $\mathcal{M}_T$ (c.f. (12)) and hence we simply write η_{uv}. Unless otherwise indicated $\eta_{u,v}$ will denote $\mathbb{E}(U_u U_v)/\mathrm{Var Y_u}$ for any two variables Y_u, Y_v. In general for any T with n leaves we obtain the following sequence of polynomial maps

$$\Omega_T \xrightarrow{f_{\omega\theta}} \Theta_T \xrightarrow{f_T} \Delta_{2^n-1} \xrightarrow{f_{p\mu}} \mathcal{C}_n, \tag{13}$$

where $f_{\omega\theta}$ and $f_{p\mu}$ are regular polynomial maps.

The geometric description of the new parameter space Ω_T is quite complicated. For any choice of values for $(\bar{\mu}_v)_{v\in V}$, where $\bar{\mu}_v \in [-1,1]$, the constraints on the remaining parameters can be deduced from the following set of inequalities

$$\begin{aligned} -\min\{(1+\bar{\mu}_u)(1+\bar{\mu}_v),(1-\bar{\mu}_u)(1-\bar{\mu}_v)\} \le (1-\bar{\mu}_u^2)\eta_{u,v} \le \\ \le \min\{(1+\bar{\mu}_u)(1-\bar{\mu}_v),(1-\bar{\mu}_u)(1+\bar{\mu}_v)\} \end{aligned} \tag{14}$$

where $\eta_{uv} = \eta_{u,v}$ if $(u,v) \in E$.

To show (14) we check for which values of $\eta_{u,v}$ we can reconstruct a probability distribution $[p_{ij}]$ of (Y_u, Y_v) with given margins $p_{1+} = \lambda_u = \frac{1}{2}(1-\bar{\mu}_u)$, $p_{+1} = \lambda_v = \frac{1}{2}(1-\bar{\mu}_v)$. A sufficient condition for a table with margins summing to one to form a probability distribution is that $p_{01}, p_{10}, p_{11} \ge 0$ and $p_{01} + p_{10} + p_{11} \le 1$. We have $p_{01} = p_{+1} - p_{11} \ge 0$ if and only if $p_{11} \le \lambda_v$ equivalently $\mu_{uv} = p_{11} - \lambda_u\lambda_v \le \lambda_v(1-\lambda_u)$. In a similar way we show that $p_{10} \ge 0$ if and only if $\mu_{uv} \le (1-\lambda_v)\lambda_u$. The condition $p_{11} \ge 0$ is equivalent to $\mu_{uv} \ge -\lambda_v\lambda_u$ and the last condition is equivalent to $\mu_{uv} \ge -(1-\lambda_v)(1-\lambda_u)$. To obtain (14) write $\mu_{uv} = (1-\bar{\mu}_u^2)\eta_{u,v}$ and replace λ_u, λ_v with $\bar{\mu}_u, \bar{\mu}_v$ using (12).

In the remaining part of this subsection we present the semi-algebraic description of the tripod tree model.

Definition 2. Let P be a $2\times2\times2$ table then the hyperdeterminant of P as defined by Gelfand, Kapranov, Zelevinsky [19, Chapter 14] is given by

$$\begin{aligned} \mathrm{Det}\, P &= (p_{000}^2 p_{111}^2 + p_{001}^2 p_{110}^2 + p_{010}^2 p_{101}^2 + p_{011}^2 p_{100}^2) \\ &- 2(p_{000}p_{001}p_{110}p_{111} - p_{000}p_{010}p_{101}p_{111} + p_{000}p_{011}p_{100}p_{111} \\ &+ p_{001}p_{010}p_{101}p_{110} + p_{001}p_{011}p_{110}p_{100} + p_{010}p_{011}p_{101}p_{100}) \\ &+ 4(p_{000}p_{011}p_{101}p_{110} + p_{001}p_{010}p_{100}p_{111}). \end{aligned}$$

If $\sum p_{ijk} = 1$ then we can simplify this formula using the change of coordinates to central moments

$$\mathrm{Det}\, P = \mu_{123}^2 + 4\mu_{12}\mu_{13}\mu_{23}. \tag{15}$$

Lemma 3 (A semi-algebraic description of the tripod model)**.** *Let $\mathcal{M}_T$ be the general Markov model on a tripod tree T rooted in any node of T. Let P be a $2\times2\times2$ probability table for three binary random variables (X_1, X_2, X_3). Then $P \in \mathcal{M}_T$ if and only if*

$$\mu_{12}\mu_{13}\mu_{23} \ge 0, \tag{16}$$

$$\mu_{12}^2\mu_{13}^2 + \mu_{12}^2\mu_{23}^2 + \mu_{13}^2\mu_{23}^2 \le \mathrm{Det}\, P \le \min_{i,j} \mu_{ij}^2 \tag{17}$$

and

$$\mathrm{Det} P \leq ((1 \pm \bar{\mu}_i)\mu_{jk} \mp \mu_{123})^2 \quad \textit{for all } i = 1,2,3, \tag{18}$$

where for any $i = 1,2,3$ *by* j,k *we denote elements of* $\{1,2,3\} \setminus i$ *with a convention that* $j < k$.

Proof. The model is defined by $X_1 \perp\!\!\!\perp (X_2, X_3)|H$ and $X_2 \perp\!\!\!\perp X_3|H$. Then, since $\lambda_h(1-\lambda_h) = \frac{1}{4}(1-\bar{\mu}_h^2)$, (8) gives that

$$\begin{aligned} &\mu_{ij} = \tfrac{1}{4}(1-\bar{\mu}_h^2)\eta_{h,i}\eta_{h,j} \text{ for all } i \neq j \in \{1,2,3\} \text{ and} \\ &\mu_{123} = \tfrac{1}{4}(1-\bar{\mu}_h^2)\bar{\mu}_h\eta_{h,1}\eta_{h,2}\eta_{h,3}, \end{aligned} \tag{19}$$

where $\bar{\mu}_h \in [-1,1]$ and $(1-\bar{\mu}_h^2)\eta_{h,i}$ for all $i = 1,2,3$ satisfy the inequality (14). We now show that this implies the constraints on the moments given in the lemma. Using (19) we obtain

$$\mu_{12}\mu_{13}\mu_{23} = \left(\frac{1}{4}(1-\bar{\mu}_h^2)\right)^3 (\eta_{h,1}\eta_{h,2}\eta_{h,3})^2 \tag{20}$$

which in particular implies the inequality in (16) and it does not depend on a rooting of T. Moreover,

$$\mathrm{Det}\, P = \mu_{123}^2 + 4\mu_{12}\mu_{13}\mu_{23} = \frac{1}{16}(1-\bar{\mu}_h^2)^2(\eta_{h,1}\eta_{h,2}\eta_{h,3})^2. \tag{21}$$

and by the second equation in (19)

$$\bar{\mu}_h^2\, \mathrm{Det}\, P = \mu_{123}^2, \quad (1-\bar{\mu}_h^2)\, \mathrm{Det}\, P = 4\mu_{12}\mu_{13}\mu_{23} \tag{22}$$

and again this does not depend on the rooting of T. E s (19) and (21) imply that

$$\eta_{h,i}^2\, \mu_{jk}^2 = \mathrm{Det}\, P \quad \text{for all } i = 1,2,3. \tag{23}$$

Similarly one can show that

$$\mu_{12}^2\mu_{13}^2 + \mu_{12}^2\mu_{23}^2 + \mu_{13}^2\mu_{23}^2 = \frac{1}{16}(1-\bar{\mu}_h^2)^2(\eta_{h,1}^2 + \eta_{h,2}^2 + \eta_{h,3}^2)\mathrm{Det} P. \tag{24}$$

Since necessarily $\eta_{h,i}^2, \bar{\mu}_h^2 \in [0,1]$ then s (22), (23) and (24) imply that

$$\mu_{12}^2\mu_{13}^2 + \mu_{12}^2\mu_{23}^2 + \mu_{13}^2\mu_{23}^2 \leq \mathrm{Det}\, P \leq \min_{i,j} \mu_{ij}^2,$$

which is exactly (17).

Note that in the case when $\mathrm{Det}\, P = 0$ the inequalities in (18) are trivially satisfied. Hence all the constraints in the lemma for points such that $\mathrm{Det}\, P = 0$ hold. Now we can assume $\mathrm{Det}\, P > 0$. In this case $\bar{\mu}_h^2 \neq 1$, $\eta_{h,i} \neq 0$ for all $i = 1,2,3$. Let $\sigma_i = \mathrm{sgn}(\eta_{h,i})$, $\sigma_h = \mathrm{sgn}(\bar{\mu}_h)$, $\sigma_{ij} = \mathrm{sgn}(\mu_{ij})$ and $\sigma_{123} = \mathrm{sgn}(\mu_{123})$. By (19) we have $\sigma_{ij} = \sigma_i\sigma_j$ for all $i,j = 1,2,3$ and $\sigma_{123} = \sigma_h\sigma_1\sigma_2\sigma_3$ and hence $\sigma_h = \sigma_1\sigma_2\sigma_3\sigma_{123}$. Thus taking the square root of (22) we obtain $\bar{\mu}_h\sqrt{\mathrm{Det}\, P} = \sigma_1\sigma_2\sigma_3\mu_{123}$. We can now discover which constraints on the observed moments are induced by the further constraints on the parameter space given by (14). For each $i = 1,2,3$, if $\eta_{h,i} > 0$ ($\sigma_i = 1$) then from (23) $\eta_{h,i}|\mu_{jk}| = \sqrt{\mathrm{Det}\, P}$. It follows that $|\mu_{jk}|\sqrt{\mathrm{Det}\, P}\bar{\mu}_h = \sigma_{jk}|\mu_{jk}|\mu_{123} = \mu_{jk}\mu_{123}$. By (14) taken for each $i = 1,2,3$ after multiplying both sides by $|\mu_{jk}|\sqrt{\mathrm{Det}\, P}$ we obtain

$$(1-\bar{\mu}_h^2)\eta_{h,i}|\mu_{jk}|\sqrt{\mathrm{Det}\, P} = 4\mu_{12}\mu_{13}\mu_{23} \leq (1 \pm \bar{\mu}_i)(\sqrt{\mu_{jk}^2 \mathrm{Det}\, P} \mp \mu_{jk}\mu_{123}) \tag{25}$$

for all $i = 1,2,3$.

If $\eta_{h,i} < 0$ then $-\eta_{h,i}|\mu_{jk}| = \sqrt{\mathrm{Det}\, P}$ and $|\mu_{jk}|\sqrt{\mathrm{Det}\, P}\bar{\mu}_h = -\sigma_{jk}|\mu_{jk}|\mu_{123} = \mu_{jk}\mu_{123}$. Again by (14)

$$-(1-\bar{\mu}_h^2)\eta_{h,i}|\mu_{jk}|\sqrt{\mathrm{Det}\, P} = 4\mu_{12}\mu_{13}\mu_{23} \leq (1 \pm \bar{\mu}_i)(\sqrt{\mu_{jk}^2 \mathrm{Det}\, P} \mp \mu_{jk}\mu_{123})$$

and hence we obtain exactly the same inequalities.

To show that, given (16) and (17), these inequalities are equivalent to (18). Note that (16) implies that $\sqrt{\mu_{jk}^2 \mathrm{Det}\, P} \pm \mu_{jk}\mu_{123} > 0$. We can multiply both sides of (25) by this expression to obtain

$$4\mu_{12}\mu_{13}\mu_{23}(\sqrt{\mu_{jk}^2 \mathrm{Det}\, P} \pm \mu_{jk}\mu_{123}) \leq (1 \pm \bar{\mu}_i)(\mu_{jk}^2\, 4\mu_{12}\mu_{13}\mu_{23})$$

or

$$\sqrt{\mu_{jk}^2 \mathrm{Det}\, P} \leq (1 \pm \bar{\mu}_i)\mu_{jk}^2 \mp \mu_{jk}\mu_{123}. \tag{26}$$

Since Det$P > 0$ this is equivalent to

$$\begin{aligned} &0 \leq (1 \pm \bar{\mu}_i)\mu_{jk}^2 \mp \mu_{jk}\mu_{123}, \\ &\mathrm{Det} P \leq ((1 \pm \bar{\mu}_i)\mu_{jk} \mp \mu_{123})^2 \end{aligned} \tag{27}$$

The second inequality is exactly (18). The remaining task is to show that (18) already implies the first inequality in (27). To see this rewrite (18) as

$$\begin{aligned} &(1-\bar{\mu}_i)\mu_{jk}^2 \geq -2\mu_{123}\mu_{jk} \\ &(1+\bar{\mu}_i)\mu_{jk}^2 \geq 2\mu_{123}\mu_{jk}. \end{aligned} \tag{28}$$

noting that the left-hand sides are nonnegative. For each of the two inequalities if the right-hand side is negative then the inequality is trivially satisfied. If the right-hand side is nonnegative then in the first case $2\mu_{123}\mu_{jk} \geq \mu_{123}\mu_{jk}$ and in the second case $-2\mu_{123}\mu_{jk} \geq -\mu_{123}\mu_{jk}$. Hence the following set of inequalities is implied by (28)

$$(1-\bar{\mu}_i)\mu_{jk}^2 \geq -\mu_{123}\mu_{jk}, \quad (1+\bar{\mu}_i)\mu_{jk}^2 \geq \mu_{123}\mu_{jk}.$$

This is exactly the first inequality in (27). This shows that if $P \in \mathcal{M}_T$ then (16)–(18) must hold.

Now assume that P satisfies the inequalities in (16)–(18). We will show that a choice of parameters in (19) exists which satisfies constraints defining Ω_T. From (16) we know that $\mathrm{Det}\, P \geq 0$ so consider separately the two situations: first when $\mathrm{Det}\, P = 0$ and second when $\mathrm{Det}\, P > 0$. In the first case necessarily $\mu_{123} = 0$ and the inequality in (17) implies that at least two covariances are zero. If all the covariances are zero then setting all edge parameters to zero and $\bar{\mu}_h^2 = 1$ gives a valid choice of parameters satisfying (19). When one covariance, say $\mu_{12} \neq 0$, is non-zero then if a parametrization exists it has to satisfy $\bar{\mu}_h^2 \neq 1$, $\eta_{h,1}, \eta_{h,2} \neq 0$ and $\eta_{h,3} = 0$. Such a choice of parameters will exist if we can ensure that $\mu_{12} = (1-\bar{\mu}_h^2)\eta_{h,1}\eta_{h,2}$. This follows from Corollary 2 in [20] which states that if only $\mu_{12} \neq 0$ then there always exists a choice of parameters for model $X_1 \perp\!\!\!\perp X_2 | H$, where H is hidden.

Assume now that $\mathrm{Det}\, P > 0$ which by (17) implies that $\mu_{ij} \neq 0$ for each $i < j = 1, 2, 3$. Define $\bar{\mu}_h := \frac{\mu_{123}}{\sqrt{\mathrm{Det}\, P}}$ and $\eta_{h,i} := \frac{\mathrm{Det}\, P}{\mu_{jk}}$ for $i = 1, 2, 3$. One can easily check that $\mu_{ij} = (1-\bar{\mu}_h^2)\eta_{h,i}\eta_{h,j}$ for $i, j = 1, 2, 3$ and $(1-\bar{\mu}_h^2)\bar{\mu}_h\eta_{h,1}\eta_{h,2}\eta_{h,3} = \mu_{123}$. It remains to show that parameters defined in this way satisfy the constraints defining Ω_T. First note that by the inequalities we have $0 \leq 4\mu_{12}\mu_{13}\mu_{23} \leq \mathrm{Det}\, P$ and hence

$\bar{\mu}_h^2 \in [0,1]$ as required. Moreover since $\sigma_i = \sigma_{jk}$ then the inequalities in (25), holding by the above arguments, demand that $(1-\bar{\mu}_h^2)\eta_{h,i}$ satisfies the constraints in (14) for each $i = 1, 2, 3$.

□

Remark 4. In a phylogenetic analysis it is often assumed that $\eta_e > 0$ for all $e \in E$ and $\bar{\mu}_v^2 \neq 1$ for all $v \in V$ (c.f. assumptions (M1)-(M3) in Section 8.2 and Section 8.4 in [31]). The proof above is then much more straightforward since it restricts us to the case $\operatorname{Det} P > 0$ and $\mu_{ij} > 0$ for all $i < j = 1, 2, 3$. For a general tree this assumption also greatly simplifies computations.

3.3. **A relation to tree metrics.** Now let T be a general tree with n leaves. Before stating the main theorem of the paper we show how to obtain an elegant set of necessary constraints on $\mathcal{M}_T$. Assume that $\bar{\mu}_v^2 \neq 1$ for all $v \in V$ (c.f. Remark 4). The correlation between X_u and X_v is defined as $\rho_{uv} = \frac{\mu_{uv}}{\sqrt{(1-\bar{\mu}_u^2)(1-\bar{\mu}_v^2)}}$ which gives

$$\rho_{uv}^2 = \eta_{u,v}^2 \frac{1-\bar{\mu}_u^2}{1-\bar{\mu}_v^2}.$$

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 path joining them in T with the set of edges denoted by E_{kl}. Then using the first equation in (8) implies that $\rho_{ij} = \rho_{ih}\rho_{hj}$ for any h separating i and j. Using this argument repeatedly for the node adjacent to k, then for the next node in the path and so on, it can be seen that

$$\rho_{kl} = \prod_{e \in E_{kl}} \rho_e \tag{29}$$

for each probability distribution in $\widehat{\mathcal{M}}_T$ such that all the correlations are well defined.

The above equation allows us to demonstrate an interesting reformulation of our problem in term of tree metrics (c.f. [31, Section 7]) which we explain below (see also Cavender [7]).

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

$$d(k,l) = \begin{cases} -\log(\rho_{kl}^2), & \text{for all } k, l \in V \text{ such that } \rho_{kl} \neq 0, \\ +\infty, & \text{otherwise} \end{cases}$$

then $d(k,l) \geq 0$ because $\rho_{kl}^2 \leq 1$ and $d(k,k) = 0$ for all $k \in V$ since $\rho_{kk} = 1$.

If $R \in \mathcal{M}_T^{\mu}$ then by (29) we can define map $d_{(T;R)} : V \times V \to \mathbb{R}$

$$d_{(T;R)}(k,l) = \begin{cases} \sum_{(u,v)\in\mathcal{P}_T(k,l)} d(u,v), & \text{if } k \neq l, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

This map, restricted 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 (30) is satisfied? Or equivalently: What are the conditions on the absolute values of the second order correlations in order that $\rho_{ij}^2 = \prod_{e \in E_{ij}} \rho_e^2$ (for some edge correlations) is satisfied? We have the following theorem.

Theorem 5 (Tree-Metric Theorem, Buneman [4])**.** *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 $i, j, k, l \in [n]$,*

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

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

$$(\rho_{ij}\rho_{kl})^2 \geq \min\left\{(\rho_{ik}\rho_{jl})^2, (\rho_{il}\rho_{jk})^2\right\}. \tag{31}$$

As a corollary we can state the following well known result (c.f. [31, Section 8.4]).

Corollary 6. *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 which 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.

Lemma 7. *Let T be a tree such that each inner node has degree at least three. If the set of correlations ρ_{ij} for all pairs of leaves in the tree satisfies $\rho_{ij}\rho_{ik}\rho_{jk} \geq 0$ for all triples $\{i, j, k\} \subset [n]$ then there exists a choice of signs for the edge correlations consistent with the signs of the correlations between the leaves.*

Given that $\rho_{ij}\rho_{ik}\rho_{jk} \geq 0$ for all distinct triples $i, j, k \in [n]$ (31) can be rearranged as

$$\min\left\{\frac{\rho_{ik}\rho_{jl}}{\rho_{ij}\rho_{kl}}, \frac{\rho_{il}\rho_{jk}}{\rho_{ij}\rho_{kl}}\right\} \leq 1.$$

Using the fact that $\frac{\rho_{ik}\rho_{jl}}{\rho_{ij}\rho_{kl}} = \frac{\mu_{ik}\mu_{jl}}{\mu_{ij}\mu_{kl}}$ and $\frac{\rho_{il}\rho_{jk}}{\rho_{ij}\rho_{kl}} = \frac{\mu_{il}\mu_{jk}}{\mu_{ij}\mu_{kl}}$ these imply the set of inequalities

$$0 \leq \min\left\{\frac{\mu_{ik}\mu_{jl}}{\mu_{ij}\mu_{kl}}, \frac{\mu_{il}\mu_{jk}}{\mu_{ij}\mu_{kl}}\right\} \leq 1. \tag{32}$$

for all (not necessarily distinct) $i, j, k, l \in [n]$. It is important to note however from Lemma 3 that the inequalities given above cannot be sufficient even for the tripod tree model because any particular choice of means for the nodes of the tree constrains the space of possible edge parameters. The means of the leaf variables will be estimated and, except in the unlikely event of all being $1/2$, will actively constrain the tree metric space in the way described above. It follows that to be an effective tool, the tree metric space needs to be further truncated to respect the inequalities implied by Lemma 3 in the tripod case. Furthermore in the general case the observable higher order moments of a tree model further constrain this space and should not be ignored. The next section derives these constraints explicitly.

3.4. **The main theorem.** Since $\mathcal{M}_T$ is a semi-algebraic set, to describe it we need to provide the complete list of defining polynomial equations and inequalities. In this subsection we present the known results concerning the equality descriptions of the model. We state then the main theorem of the paper which gives the full semi-algebraic description.

Any conditional independence model has a defining set of equations (c.f. Section 2.2). Allman and Rhodes [1] identified equations defining the general Markov model for binary data when the defining tree is trivalent which means that all its inner nodes have valency three. To obtain this identification they studied the phylogenetic ideal, i.e. the set of all polynomials vanishing on $f_T(\mathbb{C}^{2|E|+1}) \subset \mathbb{C}^{2^n}$. Note that we do not restrict the parameters in (3) to lie in Θ_T but allow them to be any complex numbers. Extending the domain of defining equations to the complex field is a common approach in algebraic geometry and in this context this is done to make the analysis easier. To introduce the set of defining equations provided by Allman and Rhodes we need the following definition.

Definition 8. 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. In particular, for each $e \in E$, removing edge e from E induces a partition of the set of leaves into two subsets corresponding to the two 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 its 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 9 (Allman, Rhodes [1])**.** *Let T be a trivalent tree and $\mathcal{M}_T$ be the general Markov model on T for binary variables. Then the ideal defining the general Markov model is generated by all 3×3-minors of all the edge flattenings of T plus the trivial invariant $\sum_\alpha p_\alpha = 1$.*

This therefore identifies the set of equations defining the general Markov model for trivalent trees. Note that the result is true for the tripod tree model since in this case each edge flattening of the joint probability table is a 2×4 table so there are no 3×3 minors and hence there are no equations vanishing on the model.

The following result is well known (see e.g. [26]).

Lemma 10. *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).*

In what follows we often restrict ourselves to trivalent trees or by Lemma 10 equivalently to trees such that each inner node has degree at most three. This restriction is natural since for every T the model $\mathcal{M}_T$ can be realized as a submodel of $\mathcal{M}_{T^*}$ for some trivalent tree T^* with certain restrictions on the parameters in Ω_{T^*} (c.f. Remark 14).

In Section 6 we prove the following theorem which gives the exact geometric description of $\mathcal{M}_T$ in the case when T is a trivalent tree. However Proposition 18 shows that a similar result can be constructed for any tree topology if only we knew equations defining the image of $f_T : \Theta_T \to \Delta_{2^n-1}$.

Theorem 11. *Let $T = (V, E)$ be a trivalent 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. Then $P \in \mathcal{M}_T$ if and only if the following four conditions hold:*

(C1): *all 3×3-minors of all the edge flattenings of P vanish,*

(C2): *for all distinct triples $i, j, k \in [n]$ $\mu_{ij}\mu_{ik}\mu_{jk} \geq 0$ and*

$$(\mu_{ij}^2\mu_{ik}^2 + \mu_{ij}^2\mu_{jk}^2 + \mu_{ik}^2\mu_{jk}^2) \leq \operatorname{Det} P^{ijk} \leq \min\{\mu_{ij}^2, \mu_{ik}^2, \mu_{jk}^2\},$$

(C3): *for all distinct triples $i, j, k \in [n]$*

$$\operatorname{Det} P^{ijk} \leq \min_\sigma \left\{ \left((1 \pm \bar{\mu}_{\sigma(i)})\mu_{\sigma(j)\sigma(k)} \mp \mu_{ijk} \right)^2 \right\}.$$

for all three permutations σ of $\{i, j, k\}$ such that $\sigma(j) > \sigma(k)$.

(C4): *for any four distinct leaves i, j, k, l such that there exists $e \in E$ inducing a split $(A)(B)$ such that $i, j \in A$ and $k, l \in B$ (c.f. Figure 1) we have*

$$(2\mu_{ik}\mu_{jl})^2 \leq (\sqrt{\mu_{jl}^2 \operatorname{Det} P^{ijk}} + \mu_{jl}\mu_{ijk})(\sqrt{\operatorname{Det} P^{ikl}} - \mu_{ikl})$$

$$(2\mu_{ik}\mu_{jl})^2 \leq (\sqrt{\mu_{jl}^2 \operatorname{Det} P^{ijk}} - \mu_{jl}\mu_{ijk})(\sqrt{\operatorname{Det} P^{ikl}} + \mu_{ikl}).$$

The proof of the theorem is given in Section 6.

4. Tree cumulants

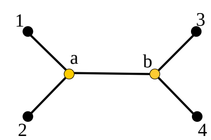
In previous sections we defined a reparametrization map $f_{p\mu} : \Delta_{2^n-1} \to \mathcal{C}_n$ of the model space. In this section we perform a further reparametrization of the system of central moments for another system of coordinates which is intrinsically linked to the given tree. The model in this coordinate system admits a quasi-monomial parametrization in the new parameters. Our approach in this section is more combinatorial and is based on the theory of Möbius functions (see [35]). This links to the concept of cumulants which are essentially model-free (see e.g. [25, Section 2] [27]). Our idea here is to develop some "tree cumulants" to obtain as simple parametric form of the model as possible.

4.1. **The poset of tree partitions.** Let $T = (V, E)$ be a tree with n leaves. A *split induced by* $e \in E$ is a partition of $[n]$ into two non-empty sets induced by removing e from E and restricting $[n]$ to the connected components of the resulting graph. Let $(A)(B)$ be a split induced by $e \in E$ and for $W \subset V$ let $T(W) = (V(W), E(W))$ denote the minimal subtree of T induced by W (c.f. Section 2.1). Then any split of A induced by $e' \in E(A)$ or a split of B induced by $e' \in E(B)$ 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 the partition.

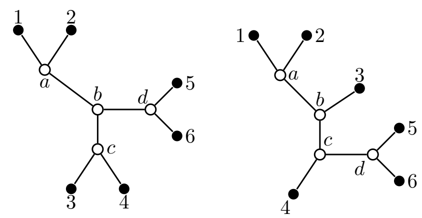
By Π_T we denote the partially ordered set (poset) of all multisplits of the set of leaves induced by inner edges of T. We can define the poset on the level of edges of T. Let $E^0 \subset E$ denote the set of inner edges of T. We define the following equivalence class on E^0. For $E_x \subseteq E^0$, $E_y \subseteq E^0$ we say $E_x \sim E_y$ if and only if removing E_x induces the same partition of $[n]$ as removing E_y. By $\overline{E}_x$ we denote the maximal with respect to inclusion element of the equivalence class of E_x. Let $x, y \in \Pi_T$ be defined by removing edges in $E_x \subseteq E$ and edges in $E_y \subseteq E$ respectively. We write $x \leq y$ if and only if $\overline{E}_x \subseteq \overline{E}_y$ and we say that y is a *subsplit* of x.

A *segment* $[x, y]$, for x and y in Π_T, is the set of all elements z such that $x \leq z \leq y$. The poset Π_T forms a lattice. To show this we define $x \vee y \in \Pi_T$ ($x \wedge y \in \Pi_T$) as an element in Π_T obtained induced by removing $E_x \cup E_y$ ($E_x \cap E_y$). Note that this definition does not depend on the choice of the representatives of the equivalence class of E_x and E_y. We have $x \vee y \geq x$, $x \vee y \geq y$ ($x \wedge y \leq x$, $x \wedge y \leq y$) and if there exists another $z \in \Pi_T$ with this property then $z \geq x \vee y$ ($z \leq x \wedge y$). The element $x \vee y$ ($x \wedge y$) is called a *join* (a *meet*) of x and y. It has a unique maximal element induced by removing E^0 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.

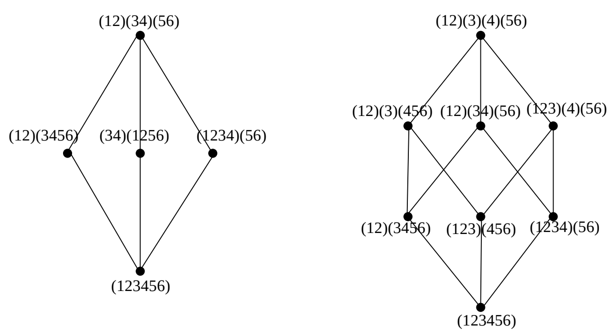
To illustrate these definitions let T be the quartet tree below.



This 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



Their respective posets Π_T^L (for the tree on the left) and Π_T^R (for the tree on the right) are given below



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 a *Möbius function* $m_\Pi : \Pi \times \Pi \to \mathbb{R}$ is defined in such a way that $m_\Pi(x, x) = 1$ for every $x \in \Pi$ and $m_\Pi(x, y) = -\sum_{x \leq z < y} m_\Pi(x, z)$ for $x < y$ in Π (c.f. [35, Section 3.7]). Let $W \subset V$ and 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 $\prod_{B\in x}\Pi_{T(B)}$ where the product is over all blocks B of x. By Proposition 3.8.2 in [35] the Möbius function on the product of posets $\prod_{B\in x}\Pi_{T(B)}$ can be written as a product of Möbius functions for each of the posets $\Pi_{T(B)}$. Thus

$$m(x,y)=\prod_{B\in x} m_B(0_B,y_B) \quad \text{for } y_B\in\Pi_{T(B)}, \tag{33}$$

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 $x\leq y$) and $x_B=0_B$ for each B.

4.2. **An induced reparametrization.** In this section unless otherwise stated we restrict ourselves to trees such that all the inner nodes have degree at most three. We use the combinatorial machinery developed in the previous subsection to define new coordinates $(\kappa_I)_{I\in[n]_{\geq 2}}$ using change of coordinates $f_{\mu\kappa}:\mathbb{R}^n\times\mathbb{R}^{2^n}\to\mathbb{R}^n\times\mathbb{R}^{2^n}$ defined by the Möbius function on $\Pi_{T(I)}$ for $I\in[n]_{\geq 2}$ in the following way.

Let T be a tree with n leaves then tree cumulants for T are obtained according to

$$\kappa_I=\sum_{\pi\in\Pi_{T(I)}} m_I(0_I,\pi)\prod_{B\in\pi}\mu_B \quad \text{for all } I\in[n]_{\geq 2}, \tag{34}$$

where by definition $\mu_i=0$ for all $i\in[n]$. By definition $f_{\mu\kappa}$ is an identity on the first n variables corresponding to the means.

The Jacobian of $f_{\mu\kappa}$ is equal to one. 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 μ_I, μ_J. Then it can be checked that the Jacobian matrix of (35) is lower triangular with all its diagonal terms taking the value one. The map $f_{\mu\kappa}$ is a regular polynomial map with a regular polynomial inverse. The exact form of the inverse map is given by the following lemma.

Lemma 12. *Let $T=(V,E)$ be a tree with n leaves. Then*

$$\mu_I=\sum_{x\in\Pi_{T(I)}}\prod_{B\in x}\kappa_B \quad \textit{for all } I\in[n]_{\geq 2}. \tag{35}$$

Proof. Define two functions on $\Pi_{T(I)}$

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

For each $y\in\Pi_{T(I)}$ by (34) we obtain that $\beta(y)$ is equal to

$$\prod_{B\in y}\kappa_B=\prod_{B\in y}\left(\sum_{x_B\in\Pi_{T(B)}} m_B(0_B,x_B)\prod_{C\in x_B}\mu_C\right)=\sum_{x\geq y}\prod_{B\in x} m_B(0_B,x_B)\prod_{C\in x}\mu_C,$$

where x is an element of $\Pi_{T(I)}$ obtained by concatenating x_B for $B\in y$. By the product formula in (33) we have $\prod_{B\in x} m_B(0_B,x_B)=m(y,x)$ which gives that $\beta(y)=\sum_{x\geq y} m_I(y,x)\alpha(x)$ for all $y\in\Pi_{T(I)}$. The proof now follows from the dual Möbius inversion (see Proposition 3.7.2 in [35]). □

By definition $\mathcal{K}_T=f_{\mu\kappa}(\mathcal{C}_n)$, where $\mathcal{K}_T$ denotes the space of tree cumulants, and $\mathcal{M}_T^{\kappa}=f_{\mu\kappa}(\mathcal{M}_T^{\mu})\subseteq\mathcal{K}_T$ and for any $I\in[n]_{\leq 3}$ we have $\kappa_I=\mu_I$, where $[n]_{\leq 3}$ denotes all the subsets of $[n]$ with at most three elements. Equation (34) justifies the name for the tree 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

$$\mathrm{Cum}((X_i)_{i\in I}) = \sum_{\pi\in\mathcal{P}(I)} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{B\in\pi} \mathbb{E}(X_B) \tag{36}$$

where the product is over all blocks of π and $|\pi|$ denotes the number of blocks in π. This is essentially the same as (34) but with a different defining poset (see [37][29]).

The following proposition motivates the whole section and demonstrates why our new coordinate system is particularly useful.

Proposition 13. *Let $T = (V, E)$ be a rooted tree with n leaves. Then $\mathcal{M}_T^\kappa$ is parametrized by a map defined coordinatewise as follows:*

$$\kappa_I = \frac{1}{4}\left(1-\bar{\mu}_{r(I)}^2\right) \prod_{v\in \mathrm{int}(V(I))} \bar{\mu}_v^{\deg(v)-2} \prod_{(u,v)\in E(I)} \eta_{uv} \quad \textit{for each } I\in[n]_{\geq 2}, \tag{37}$$

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

Let $\psi_T : \mathbb{R}^{|V|+|E|} \to \mathbb{R}^n \times \mathbb{R}^{2^n}$ be a map defined as follows. On the first n coordinates it is given by $\lambda_i = \frac{1-\bar{\mu}_i}{2}$ for $i = 1,\ldots,n$. Then we have $\kappa_\emptyset = 1$ and $\kappa_i = 0$ for $i = 1,\ldots,n$ and on the remaining coordinates the map is defined by (37). We have $\mathcal{M}_T^\kappa = \psi_T(\Omega_T) \subseteq \mathcal{K}_T$.

Proof of Proposition 13. 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.

First we show that for any $k, l \in V$ one has

$$\mu_{kl} = \frac{1}{4}(1-\bar{\mu}_k^2)\eta_{k,l} = \frac{1}{4}(1-\bar{\mu}_r^2)\prod_{e\in E_{kl}} \eta_e, \tag{38}$$

where r denotes the root of $\mathcal{P}_T(k,l)$. To show (38) first note that from (8) by taking $I = k$, $J = l$ we have

$$\mu_{kl} = \frac{1}{4}(1-\bar{\mu}_r^2)\eta_{r,k}\eta_{r,l}$$

so it suffices to show that $(1-\bar{\mu}_r^2)\eta_{r,k} = (1-\bar{\mu}_r^2)\prod_{e\in E_{rk}}\eta_e$ and $(1-\bar{\mu}_r^2)\eta_{r,l} = (1-\bar{\mu}_r^2)\prod_{e\in E_{rl}}\eta_e$. By symmetry it suffices to show the first. Let h_1 denote the child of r in $\mathcal{P}_T(r,k)$. Then $(r, h_1) \in E_{rk}$ and again by (8) we have

$$\frac{1}{4}(1-\bar{\mu}_r^2)\eta_{r,k} = \mu_{rk} = \frac{1}{4}(1-\bar{\mu}_{h_1}^2)\eta_{h_1,r}\eta_{h_1,k}.$$

Now we perform the same manipulations for $\eta_{h_1,k}$ denoting the child of h_1 in $\mathcal{P}_T(r,k)$ by h_2. We have $\frac{1}{4}(1-\bar{\mu}_{h_1}^2)\eta_{h_1,h_2} = \mu_{h_1h_2} = \frac{1}{4}(1-\bar{\mu}_{h_2}^2)\eta_{h_2,h_1}\eta_{h_2,k}$. Using a recursive argument we can write $(1-\bar{\mu}_r^2)\eta_{r,k} = \eta_{h_1,r}\eta_{h_2,h_1}\cdots\eta_{h_m,h_{m-1}}\eta_{h_m,k}(1-\bar{\mu}_{h_m}^2)$, where h_m is the parent of k in $\mathcal{P}_T(r,k)$. Note that we can write η_{h_mk} instead of $\eta_{h_m,k}$ because (h_m,k) is an edge of $\mathcal{P}_T(k,l)$ (c.f. Section 3.2). Moreover, $(1-\bar{\mu}_{h_i}^2)\eta_{h_i,h_{i-1}} = (1-\bar{\mu}_{h_{i-1}}^2)\eta_{h_{i-1}h_i}$ for each $i = 2,\ldots,m$. Then this can be rewritten as

$$(1-\bar{\mu}_r^2)\eta_{r,k} = (1-\bar{\mu}_r^2)\eta_{rh_1}\eta_{h_1h_2}\cdots\eta_{h_{m-1}h_m}\eta_{h_mk}.$$

To obtain (38) we perform the same manipulations on $(1-\bar{\mu}_r^2)\eta_{r,l}$.

Since by definition $\kappa_{ij} = \mu_{ij}$ for all $i, j \in [n]$, applying (38) to μ_{ij}, the proposition is clearly true for $n = 2$. In addition (19) implies that for every triple $i, j, k \in [n]$

$$\mu_{ijk} = \frac{1}{4}(1 - \bar{\mu}_h^2)\bar{\mu}_h \eta_{h,i}\eta_{h,j}\eta_{h,k}$$

where h is the inner node separating i, j, k in T. If h is the root of $T(ijk)$ then we are done. If not then the root r is on one of the three paths $\mathcal{P}_T(h,i)$, $\mathcal{P}_T(h,j)$, $\mathcal{P}_T(h,k)$ and h is a root for the remaining two paths. Hence using an exactly analogous argument to the one used to derive (38) we obtain the identity

$$\mu_{ijk} = \frac{1}{4}(1 - \bar{\mu}_h^2)\bar{\mu}_h \eta_{h,i}\eta_{h,j}\eta_{h,k} = \frac{1}{4}(1 - \bar{\mu}_r^2)\bar{\mu}_h \prod_{e \in E_{ijk}} \eta_e,$$

where E_{ijk} denotes the set of edges of $T(ijk)$. We have therefore shown that the lemma is true also for $n = 3$.

Now let us assume the proposition 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. We shall call such a pair an extended cherry (This differs slightly from the definition of a cherry (c.f. [31], p. 8) since the inner node does not have to be adjacent to the leaves.). 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 (8) we have

$$\mu_{12C} = \mu_{12}\mu_C + \frac{1}{4}(1 - \bar{\mu}_a^2)\eta_{a,12}\eta_{a,C} = \mu_{12}\mu_C + \eta_{a,12}\,\mu_{aC}. \tag{39}$$

Let $e \in E^0$ be the inner edge incident with a separating 1 and 2 from all other leaves. We define a closure relation in Π_T induced by e in the following way. If $x \in \Pi_T$ is induced by removing $E_x \subset E$ then $\overline{x}$ is induced by removing $E_x \cup e$. It is easily checked that this satisfies the following three conditions: (1) $\bar{x} \geq x$ (2) $\bar{\bar{x}} = \bar{x}$ and (3) $x \geq y$ implies $\bar{x} \geq \bar{y}$. Hence it defines a closure relation as defined by Rota [28]. An element $x \in \Pi$ is closed if $x = \overline{x}$. In our case 0 is never closed and 1 always is and hence $\overline{0} > 0$ and $\overline{1} = 1$.

Let $w = (12)(1_A) \in \Pi_T$. Since $\{1, 2\}$ form an extended 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 of $T(aA)$. A *trimming map* with respect to $\{1, 2\}$ 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 the trimming map constitutes an isomorphism of posets between $[0, w]$ and $\Pi_{T(aA)}$.

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 extended cherry $\{1, 2\}$. Let $w = (12)(1_A)$ then by (34)

$$\kappa_{[n]} = \sum_{x \in [0,w]} m(0, x) \prod_{B \in x} \mu_B + \sum_{x \notin [0,w]} m(0, x) \prod_{B \in x} \mu_B$$

and applying (39) to each μ_{12C} for each $x \in [0, w]$ we obtain

$$\prod_{B \in x} \mu_B = \prod_{B \in \overline{x}} \mu_B + \eta_{a,12} \prod_{B \in \widetilde{x}} \mu_B$$

and hence

$$\kappa_{[n]} = \sum_{x\in[0,w]} m(0,x)\prod_{B\in\overline{x}}\mu_B + \eta_{a,12}\sum_{x\in[0,w]} m(0,x)\prod_{B\in\widetilde{x}}\mu_B + \sum_{x\notin[0,w]} m(0,x)\prod_{B\in x}\mu_B. \tag{40}$$

Let $w' = (12)(0_A) \in \Pi_T$. Then $[w', w]$ is a subset of the closed elements in $[0, w]$. The first summand in (40) can be rewritten as

$$\sum_{y\in[w',w]}\left[\left(\sum_{\{x:\,\bar{x}=y\}} m(0,x)\right)\prod_{B\in y}\mu_B\right].$$

By [28, Proposition 4] if $\overline{x\vee y} = \overline{x}\vee\overline{y}$, which is clearly satisfied in our case since $(E_x \cup E_y) \cup e = (E_x \cup e) \cup (E_y \cup e)$, for all $y \in [w', w]$ the summands above are zero and hence the whole sum is zero. The third summand in (40) is zero as well because if $x \notin [0, w]$ then x contains (1) or (2) as one of the blocks and $\mu_1 = \mu_2 = 0$ by definition of central moments.

Since 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

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

Using (11) it can be checked that $\mu_{12a} = (1-\bar{\mu}_a^2)\bar{\mu}_a\eta_{a1}\eta_{a2}$ or equivalently that $\eta_{a,12} = \bar{\mu}_a\eta_{a1}\eta_{a2}$, where $\eta_{a1} = \prod_{e\in E(a1)}$ and $\eta_{a2} = \prod_{e\in E(a2)}$. Also since $|aA| = n-1$ by using the induction assumption

$$\kappa_{aA} = \left(1-\bar{\mu}^2_{r(aA)}\right)\prod_{(u,v)\in E(aA)}\eta_{uv}\prod_{v\in\mathrm{int}(V(aA))}\bar{\mu}_v^{\deg(v)-2},$$

where the degree is taken in $T(aA)$. But $E = E(aA) \cup E(1a) \cup E(2a)$ and

$$\mathrm{int}(V) = \mathrm{int}(V(aA)) \cup \{a\} \cup \mathrm{int}(V(1a)) \cup \mathrm{int}(V(2a))$$

and $r(aA) = r$. The degrees of a in T is three and the degree of all the inner nodes of $T(1a)$ and $T(2a)$ are two. Hence one can check that $\bar{\mu}_a\eta_{a1}\eta_{a2}\kappa_{aA}$ satisfies (37). This finishes the proof. □

Remark 14. The parameterization in (37) remains valid for general trees. If T is a tree with inner nodes of the degree higher than three then denote by T^* any tree such that all the inner nodes have degree at most three and such that T can be obtained from T^* by contracting some edges. Then $\mathcal{M}_T \subset \mathcal{M}_{T^*}$ given as the image of $\Omega_T \subset \Omega_{T^*}$ under ψ_{T^*}. The constraints on Ω_T are such that if we identify two inner nodes a, b (contracting (a, b)) then we set $\eta_{a,b} = 1$ and $\bar{\mu}_a = \bar{\mu}_b$.

Corollary 15. *Let $T = (V, E)$ be a tree and let $\mathcal{M}_T$ be the general Markov model on T. Then $\dim(\mathcal{M}_T) = |E| + |V|$ by which we mean that there exists a dense open subset of $\mathcal{M}_T$ diffeomorphic with a d-dimensional manifold.*

Proof. The parametrization in (37) is injective. Its image is diffeomorphic to $\mathcal{M}_T$. Since $\dim\mathcal{M}_T^{\kappa} = \dim\Omega_T = |V| + |E|$ so the dimension of $\mathcal{M}_T$ must be $|V| + |E|$. □

5. Phylogenetic invariants

Our new coordinates allow us to prove several useful results related to the structure of phylogenetic ideals defining $\mathcal{M}_T$ in the case when T is trivalent. In an analogous way to the edge flattenings of tables representing probability distributions we can define edge flattenings of $(\kappa_I)_{I\subseteq[n]}$, where by definition $\kappa_\emptyset = 1$ and $\kappa_i = 0$ for all $i \in [n]$. Let e be an edge of T inducing a split $(A)(B) \in \Pi_T$ such that $|A| = r$, $|B| = n - r$. Then $\widehat{N}_e$ is a $2^r \times 2^{n-r}$ matrix such that for any two subsets $I \subseteq A$, $J \subseteq B$ the element of $\widehat{N}_e$ corresponding to the I-th row and the J-th column is κ_{IJ}. Denote by N_e its submatrix given by removing the column and the row corresponding to empty subsets of A and B. Here the labeling for the rows and columns is induced by the ordering of the rows and columns for P_e (c.f. Definition 8), 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 9 in terms of the new coordinates.

Proposition 16. *Let $T = (V, E)$ be a tree and 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 split then* $\mathrm{rank}(P_e) = 2$ *if and only if* $\mathrm{rank}(N_e) = 1$.

Proof. By using elementary operations that do not change the determinant we will show that we can obtain, from the flattening matrix $P_e = [p_{\alpha\beta}]$ induced by a split $(A_1)(A_2)$, a block diagonal matrix $D_e = [d_{IJ}]$ with one as the first scalar block ($d_{\emptyset\emptyset} = 1$, $d_{\emptyset J} = 0$, $d_{I\emptyset} = 1$ for all $I \subseteq A$, $J \subseteq B$) and a matrix N_e as the second block. It will then follows that rank $(P_e) = 2$ if and only if 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 (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 D_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 the number of blocks in x by more than one which is not possible. Denote this block by $(I'J')$ where $I' \subseteq I$, $J' \subseteq J$. Note that by construction we have either both I', J' are empty sets if $x \geq (A_1)(A_2)$ or both $I', J' \neq \emptyset$ otherwise. We can rewrite (35) splitting the blocks

$$\mu_{IJ} = \sum_{x\in\Pi_{\overline{T}(IJ)}} \left(d_{I'J'} \prod_{I\supseteq B\in x} \kappa_B \prod_{J\supseteq B\in x} \kappa_B \right) = \sum_{I'\subseteq I} \sum_{J'\subseteq J} u_{II'} d_{I'J'} v_{J'J} \tag{41}$$

for some $u_{II'}$, $v_{J'J}$ and $\overline{T}$ is a trivalent tree covering T. Setting $u_{II'} = 0$ for $I \nsubseteq I'$, $v_{J'J} = 0$ for $J \nsubseteq J'$ we can write these coefficients in terms of a lower triangular matrix U and an upper triangular matrix V. 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 D_e and V records the column operations on D_e. In this way we have shown that using elementary operations one may obtain M_e from D_e. Because all the operations used in the transformation above are invertible we can go equivalently from M_e to D_e which finishes the proof. □

The proposition shows that the vanishing of all 3×3 minors of all the edge flattenings of P and the trivial invariant $\sum p_\alpha = 1$ are together equivalent to the vanishing all 2×2 minors of all edge flattenings of $\kappa = (\kappa_I)_{I \in [n]_{\geq 2}}$. An immediate corollary follows.

Corollary 17. *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 $I_1, I_2 \subset A$, $J_1, J_2 \subset B$. The set of equations of the form $\kappa_{I_1 J_1}\kappa_{I_2 J_2} = \kappa_{I_1 J_2}\kappa_{I_2 J_1}$ generate the phylogenetic ideal defining $\mathcal{M}_T$.*

In [12] 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 this might be so. Here, the invariants related to the four-point condition are only those involving second order covariances (c.f. Section 3). Moreover, 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 17. 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 central moments involved the higher variability of the invariant (see [25, Section 4.5]). This shows that invariants involving lower-order moments should be of a greater value in practice. The results of this paper further suggest that third-order moments are also helpful because they give us all the inequalities implicit in a phylogenetic model as well as some other invariants.

6. Proof of the main theorem

Note that the map $f_{\mu\kappa} : \mathcal{C}_n \to \mathcal{K}_T$ defined by (34) gives an isomorphism between $\mathcal{M}_T^\mu$ and $\mathcal{M}_T^\kappa$ and consequently also $\mathcal{M}_T$ by $f_{p\mu} : \Delta_{2^n-1} \to \mathcal{C}_n$. The following proposition gives the inequalities defining $\mathcal{M}_T^\kappa$. The result is formulated in terms of the complex numbers, i.e. we consider the domain of ψ_T without any constraints as $\mathbb{C}^{|V|+|E|}$. This allows us to make a link to the algebraic theory of phylogenetic invariants.

Proposition 18. *Let $T = (V, E)$ be a tree with n leaves. Let $\mathcal{M}_T$ be a general Markov model on T. If $\kappa \in \mathcal{K}_T$ is such that $\kappa \in \psi_T(\mathbb{C}^{|V|+|E|})$ then $\kappa \in \mathcal{M}_T^\kappa = \psi_T(\Omega_T)$ if and only if conditions (C2)-(C4) in Theorem 11 are satisfied.*

Constraints in Theorem 11 are formulated in terms of the second and the third order central moments. However, since $\mu_I = \kappa_I$ for all $I \in [n]_{\leq 3}$ they can be trivially translated to the constraints on κ. We prefer to keep the formulation in terms of moments but extended to the complex domain.

Proof. Note that by Lemma 10 we can assume that each of the inner nodes of T has degree greater than two. Let $\kappa \in \psi_T(\mathbb{C}^{|V|+|E|})$ satisfy (C2)-(C4). We will show that $(\psi_T)^{-1}(\kappa) \cap \Omega_T \neq \emptyset$ and consequently that $\kappa \in \mathcal{M}_T^\kappa$. Our analysis does not depend on a rooting of T. Indeed, without loss fix two different rootings r and r'. Let T be a tree rooted in r and by T' denote its copy rooted in r'. Then $\mathcal{M}_T = \mathcal{M}_{T'}$ and the parameters $(\eta_e), (\bar{\mu}_v)$ and (η'_e), $(\bar{\mu}'_v)$ are related as follows. We have $\bar{\mu}_v = \bar{\mu}'_v$

for all $v \in V$. Moreover, if $(u,v) \in E \cap E'$ then $\eta_{uv} = \eta'_{uv}$ and if $(u,v) \in E \setminus E'$ then $(1-\bar{\mu}_u^2)\eta_{uv} = (1-\bar{\mu}_v^2)\eta'_{vu}$. From the form of inequalities in (14) it suffices to check constraints on $(\bar{\mu}_h)$ and $\eta_{u,v}$ whenever $(u,v) \in E$ or $(v,u) \in E$.

The idea of the proof is to identify the preimage of κ and show that there is a point in the preimage which lies in Ω_T. First we identify $\bar{\mu}_h$ for each inner node $h \in V$. Fix h and let $i,j,k \in [n]$ be any three leaves separated in T by h. Then as in the proof of Lemma 3 (c.f. (22)) we have

$$\bar{\mu}_h^2 \mathrm{Det}\, P^{ijk} = \mu_{ijk}^2, \quad (1-\bar{\mu}_h^2)\mathrm{Det}\, P^{ijk} = 4\mu_{ij}\mu_{ik}\mu_{jk}, \tag{42}$$

where $\mathrm{Det}\, P^{ijk} = \mu_{ijk}^2 + 4\mu_{ij}\mu_{ik}\mu_{jk}$ denotes the hyperdeterminant of the $2\times2\times2$ table representing the marginal distribution of (X_i, X_j, X_k) as given by Definition 2. In particular because $\kappa \in \mathcal{K}_T$ it follows that all μ_{ij} and μ_{ijk} for $i,j,k \in [n]$ are real. Moreover, because κ satisfies condition (C2) then $\mathrm{Det}\, P^{ijk} \geq 0$. If there exists a triple $i,j,k \in [n]$ separated by h such that $\mathrm{Det}\, P^{ijk} > 0$ then we can divide by $\mathrm{Det}\, P^{ijk}$ in (42) obtaining a formula for $\bar{\mu}_h^2 \in [0,1]$ and so identify it up to a sign after taking the square root. Otherwise the value of this parameter is not identified but zero is always one of the possible values.

If h is adjacent to one of the leaves i then as in (23) we have

$$\eta_{h,i}^2\, \mu_{jk}^2 = \mathrm{Det}\, P^{ijk}, \tag{43}$$

where $j,k \in [n]$ are two other leaves such that i,j,k are separated by h. If we can find j,k such that $\mathrm{Det}\, P^{ijk} > 0$ then in particular $\mu_{jk}^2 > 0$ and (43) gives a formula for $\eta_{h,i}$ up to the choice of sign. By an identical argument to the one in proof of Lemma 3 this parameter takes a value within constraints defining Ω_T if the inequalities in condition (C3) hold for the given triple $i,j,k \in [n]$. If for all j,k the hyperdeterminant is zero then $\eta_{h,i}$ can be set to zero as in the proof of Lemma 3. Indeed, if there exists j,k such that $\mu_{jk} \neq 0$ then necessarily $\eta_{h,i} = 0$. Otherwise this parameter is not identified and zero is one of the possible values. Note that $\eta_{h,i} = 0$ is always allowed in Ω_T for all values of $\bar{\mu}_i$, $\bar{\mu}_h$ (c.f. (14)).

To compute the inner edges parameters $\eta_{a,b}$ 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

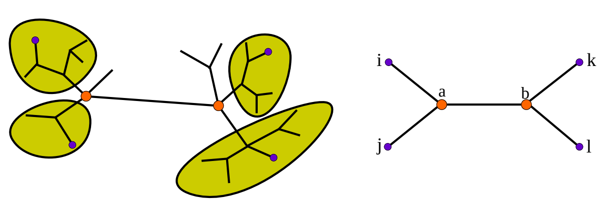


FIGURE 1. On the left we pick two adjacent inner nodes and four leaves (one from each of the shaded areas). On the right we have the marginal subtree for those four variables.

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

$$\mu_{ik}\mu_{jl}(1-\bar{\mu}_b^2) = \mu_{ij}\mu_{kl}(1-\bar{\mu}_a^2)\eta_{a,b}^2 \tag{44}$$

and $\mu_{ik}\mu_{jl} = \mu_{il}\mu_{jk}$. Take $\{i,j,k\}$ and $\{i,k,l\}$ as triples separated by a and b. Then we have the following possible situations. If for all choices of i,j,k,l we have

$\mathrm{Det}\, P^{ijk} = 0$ and $\mathrm{Det}\, P^{ikl} = 0$ then $\mu_{ij}\mu_{ik}\mu_{jk} = 0$ and $\mu_{ik}\mu_{il}\mu_{jl} = 0$. In this case we can set $\eta_{a,b} = 0$ which is a valid parameter for all choices of $\bar{\mu}_a$ and $\bar{\mu}_a$. Indeed, in this case by condition (C2) for both triples at most one covariance is non-zero. If all the six covariances are zero then the value of the parameter is not identified and it can be set to zero. If one of the covariances is non-zero then since $\mu_{ik}\mu_{jl} = \mu_{il}\mu_{jk}$ in $\psi_T(\mathbb{C}^{|V|+|E|})$ this can be only either μ_{ij} or μ_{kl} and in this case η_{ab} is not identified either and we can set $\eta_{a,b} = 0$. If both $\mu_{ij} \neq 0$ and $\mu_{kl} \neq 0$ then η_{ab} is necessarily zero.

Now consider one of the two cases: either for each i,j,k,l $\mathrm{Det}\, P^{ijk} = 0$ and there exists i,j,k,l such that $\mathrm{Det}\, P^{ikl} > 0$ or for each i,j,k,l $\mathrm{Det}\, P^{ikl} = 0$ and there exists i,j,k,l such that $\mathrm{Det}\, P^{ijk} > 0$. By symmetry without loss assume the first holds. In this case we consider the marginal model for $T(ikl)$. By Lemma 3, $\eta_{i,b} = \eta_{i,a}\eta_{a,b}$ can be identified and is a valid parameter for the marginal model for $T(ikl)$ when conditions (C3) for i,k,l are satisfied. Moreover, by Corollary 2 in [20] if $\eta_{i,b}$ takes value in the feasible region defined by (14) and it is non-zero then we can always find values of $\eta_{i,a}, \eta_{a,b}$ and $\bar{\mu}_a$ such that $(1-\bar{\mu}_a^2)\eta_{a,b}$ satisfies the inequalities in (14).

Finally, assume that there exists i,j,k,l such that $\mathrm{Det}\, P^{ijk} > 0$ and $\mathrm{Det}\, P^{ikl} > 0$. Using (42) write

$$(1-\bar{\mu}_a^2)\mathrm{Det}\, P^{ijk} = 4\mu_{ij}\mu_{ik}\mu_{jk}, \quad (1-\bar{\mu}_b^2)\mathrm{Det}\, P^{ikl} = 4\mu_{ik}\mu_{il}\mu_{kl}.$$

Multiply both sides of (44) by $\mathrm{Det} P^{ijk}\,\mathrm{Det} P^{ikl}$ and then use the above formula together with the fact that $\mu_{ik}\mu_{jl} = \mu_{il}\mu_{jk}$ to obtain

$$\mu_{ij}^2 \mathrm{Det}\, P^{ikl}\eta_{a,b}^2 = \mu_{il}^2 \mathrm{Det}\, P^{ijk}. \tag{45}$$

The above formulas identify values of the parameters $\bar{\mu}_a$, $\bar{\mu}_b$ and $\eta_{a,b}$ up to the choice of signs. Now we show that they give values within constraints defining Ω_T as long as the constraints in Theorem 11 are satisfied. Assume $\eta_{a,b} > 0$, i.e.

$$\eta_{a,b} = \left|\frac{\mu_{il}}{\mu_{ij}}\right| \sqrt{\frac{\mathrm{Det}\, P^{ijk}}{\mathrm{Det}\, P^{ikl}}}$$

then

$$\eta_{a,b}(1-\bar{\mu}_a^2) = \left|\frac{\mu_{il}}{\mu_{ij}}\right| \sqrt{\frac{\mathrm{Det}\, P^{ijk}}{\mathrm{Det}\, P^{ikl}}} \frac{4\mu_{ij}\mu_{ik}\mu_{jk}}{\mathrm{Det} P^{ijk}} = \frac{4\mu_{ik}^2|\mu_{jl}|}{\sqrt{\mathrm{Det}\, P^{ijk}\mathrm{Det}\, P^{ikl}}}, \tag{46}$$

where the second equation follows from the equation $\mu_{ik}\mu_{jl} = \mu_{il}\mu_{jk}$ and the fact that $\mathrm{sgn}(\mu_{ij}\mu_{il}) = \mathrm{sgn}(\mu_{jl})$. Since $\eta_{a,b} > 0$ (14) implies that the following constraint must be satisfied by (46)

$$\frac{4\mu_{ik}^2|\mu_{jl}|}{\sqrt{\mathrm{Det}\, P^{ijk}\mathrm{Det}\, P^{ikl}}} \le \min\{(1\pm\bar{\mu}_a)(1\mp\bar{\mu}_b)\}. \tag{47}$$

Adopting the notation from the proof of Lemma 3 we have that $\sigma_a = \sigma_{ijk}\sigma_{i,a}\sigma_{j,a}\sigma_{a,b}\sigma_{b,k}$ and $\sigma_b = \sigma_{ikl}\sigma_{i,a}\sigma_{a,b}\sigma_{b,k}\sigma_{b,l}$, where by construction $\sigma_{a,b} = 1$. So we can write

$$\bar{\mu}_a = \sigma_{i,a}\sigma_{j,a}\sigma_{b,k}\frac{\mu_{ijk}}{\sqrt{\mathrm{Det}\, P^{ijk}}}, \quad \bar{\mu}_b = \sigma_{i,a}\sigma_{b,k}\sigma_{b,l}\frac{\mu_{ikl}}{\sqrt{\mathrm{Det}\, P^{ikl}}}.$$

Multiply both sides of (47) by $|\mu_{jl}|\sqrt{\mathrm{Det}\, P^{ijk}\mathrm{Det}\, P^{ikl}}$. Since

$$|\mu_{jl}| = \sigma_{jl}\mu_{jl} = \sigma_{a,j}\sigma_{b,l}\mu_{jl}$$

we obtain

$$(2\mu_{ik}\mu_{jl})^2 \leq \min\left\{(\sqrt{\mu_{jl}^2 \mathrm{Det}\, P^{ijk}} \mp \mu_{jl}\mu_{ijk})(\sqrt{\mathrm{Det}\, P^{ikl}} \pm \mu_{ikl})\right\}$$

which holds by (C4). Consequently, the preimage of κ under ψ_T has a non-trivial intersection with Ω_T.

The converse implication is straightforward. This follows by checking that for each parameter the induced constraints are the same for all sign choices for the parameters. □

Proof of Theorem 11. To use Proposition 18 we show that up to the inequality constraints in (C2)-(C4) the image of ψ_T is described by all the 3×3 minors of all the edge flattenings of the joint probability table P of leaves of T. Denote by $\mathcal{J}$ the ideal in $\mathbb{C}[p_\alpha : \alpha \in \{0,1\}^n]$ generated by all 3×3-minors of all the edge flattenings supplemented with the trivial invariant $\sum p_\alpha = 1$ (c.f. [1]). Let $\mathcal{I}$ be an ideal defined as follows. For all inner edges of T consider a split of the set of leaves $(A)(B)$ and for all nonempty subsets $I_1, I_2 \subseteq A$, $J_1, J_2 \subseteq B$ take

$$\kappa_{I_1J_1}\kappa_{I_2J_2} - \kappa_{I_1J_2}\kappa_{I_2J_1} = 0,$$

where the κ's are the tree cumulants which are polynomials in the raw probabilities. By Proposition 16 the ideal is isomorphic to $\mathcal{J}$. Moreover, the zeros of $\mathcal{I}$ define the smallest algebraic variety containing $\psi_T(\mathbb{C}^{|V|+|E|})$.

Let $U_1 \subset \mathbb{C}^{|V|+|E|}$ be an open subset such that $\eta_e \neq 0$ for all $e \in E$ and $\bar{\mu}_v^2 \neq 1$ for all $v \in V$. Then in particular for any triple $i, j, k \in [n]$ $\mathrm{Det}\, P^{ijk} \neq 0$ (see (21)) and from the proof of Lemma 18 the map ψ_T restricted to U_1 is a proper and quasi-finite map (the preimage is always a finite set) and hence it is closed. Consequently if P is such that $\mathrm{Det}\, P^{ijk} \neq 0$ for all $i, j, k \in [n]$ and it satisfies all the equalities then it is in the image of the parametrization of ψ_T restricted to U_1.

If P for all $i, j, k \in [n]$ $\mathrm{Det}\, P^{ijk} \geq 0$ but for some of the triples the inequality is not strict then P is a limit of points (P_n) for which the inequalities are all strict. Moreover if P satisfies the equalities then we can assume that all P_n satisfy the equalities as well. From the paragraph above each of the P_n lies in the image of the parametrization. However, since ψ_T is a continuous map it also follows that P lies in the image. Again by Lemma 18 it lies in $\mathcal{M}_T^\kappa$ if and only if P satisfies the inequalities. □

7. Discussion

The new coordinate system proposed in this paper provides a better insight into the geometry of phylogenetic tree models with binary observations. The elegant form of the parametrization is useful and has already enabled us to find 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 derived coordinate system is based on a novel use of Möbius function in statistics, mimicking 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.

Since $\widehat{\mathcal{M}}_T$ forms a quadratic exponential family (see [22]) 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 [40]. It may be an interesting problem to understand in a better way the relationship between those two model classes.

Acknowledgements

Several people contributed to this paper in a substantial way. We are especially grateful to Diane Maclagan, John Rhodes and an anonymous referee for numerous helpful comments. We would also like to thank Bernd Sturmfels for a stimulating discussion at the early stage of 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] S. Basu, R. Pollack, and M. Roy, *Algorithms in Real Algebraic Geometry*, Springer, 2003.
[4] P. Buneman, *A note on the metric properties of trees*, J. Combinatorial Theory Ser. B, 17 (1974), pp. 48–50.
[5] 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.
[6] J. Cavender and J. Felsenstein, *Invariants of phylogenies in a simple case with discrete states*, Journal of Classification, 4 (1987), pp. 57–71.
[7] J. A. Cavender, *Letter to the editor*, Molecular Phylogenetics and Evolution, 8 (1997), pp. 443 – 444.
[8] J. Chang, *Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency*, Mathematical Biosciences, 137 (1996), pp. 51–73.
[9] D. A. Cox, J. B. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms*, Springer-Verlag, NY, 3rd ed., 2007.
[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] I. Gelfand, M. Kapranov, and A. Zelevinsky, *Discriminants, Resultants, and Multidimensional Determinants*, Birkhäuser, 1994.
[20] Z. Gilula, *Singular value decomposition of probability matrices: Probabilistic aspects of latent dichotomous variables*, Biometrika, 66 (1979), pp. 339–344.
[21] J. Lake, *A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony*, 1987.

[22] S. L. Lauritzen, *Graphical models*, vol. 17 of Oxford Statistical Science Series, The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.
[23] P. Lazarsfeld and N. Henry, *Latent structure analysis*, Houghton, Mifflin, New York, 1968.
[24] F. Matsen, *Fourier transform inequalities for phylogenetic trees*, Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 6 (2009), pp. 89–95.
[25] P. McCullagh, *Tensor methods in statistics*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1987.
[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.
[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] R. P. Stanley, *Enumerative combinatorics. Volume I*, no. 49 in Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2002.
[36] M. Steel and B. Faller, *Markovian log-supermodularity, and its applications in phylogenetics*, Applied Mathematics Letters, (2009).
[37] B. Streitberg, *Lancaster interactions revisited*, Ann. Statist., 18 (1990), pp. 1878–1885.
[38] 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.
[39] B. Sturmfels and S. Sullivant, *Toric Ideals of Phylogenetic Invariants*, Journal of Computational Biology, 12 (2005), pp. 204–228.
[40] 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.