Electronic Journal of Statistics
Vol. 5 (2011) 1276–1312
ISSN: 1935-7524
DOI: 10.1214/11-EJS640

Implicit inequality constraints in a binary tree model

Piotr Zwiernik

Institute for Pure and Applied Mathematics,
460 Portola Plaza,
Box 957121,
Los Angeles, CA 90095-7121
e-mail: piotr.zwiernik@gmail.com

and

Jim Q. Smith

University of Warwick
Department of Statistics
CV7AL, Coventry, UK
e-mail: j.q.smith@warwick.ac.uk

Abstract: In this paper we investigate the geometry of a discrete Bayesian network whose graph is a tree all of whose variables are binary and the only observed variables are those labeling its leaves. We provide the full geometric description of these models which is given by a set of polynomial equations together with a set of complementary implied inequalities induced by the positivity of probabilities on hidden variables. The phylogenetic invariants given by the equations can be useful in the construction of simple diagnostic tests. However, in this paper we point out the importance of also incorporating the associated inequalities into any statistical analysis. The full characterization of these inequality constraints derived in this paper helps us determine how and why routine statistical methods can break down for this model class.

AMS 2000 subject classifications: Primary 62H05, 62E15; secondary 60K99, 62F99.
Keywords and phrases: Graphical models on trees, binary data, tree cumulants, semialgebraic statistical models, phylogenetic invariants, inequality constraints.

Received October 2010.

Contents

1. Introduction

A Bayesian network whose graph is a tree all of whose inner nodes represent variables which are not directly observed defines an important class of models containing both phylogenetic tree models and hidden Markov models. Inference for this model class tends to be challenging and often needs to employ fragile numerical algorithms. In [40] we established a useful new coordinate system to analyze such models when all of the variables are binary. This reparametrization enabled us not only to address various identifiability issues but also helped us to derive exact formulae for the maximum likelihood estimators given that the sample proportions were in this model class.

However, as well as making identifiability issues more transparent and open to systematic analysis, this new coordinate system can be also used to analyze the global structure of tree models. In particular, it enables us to obtain the full description of these models in terms of implicit polynomial equations and inequalities. Knowing this full semi-algebraic description is extremely useful when used in conjunction with the identifiability structure as discussed in [40]. We explain in Section 3 how this study impacts the stability of the maximum likelihood and Bayesian estimation procedures within the class of phylogenetic tree models. It is also helpful in the construction of tree diagnostics and model selection procedures within this class.

This paper builds on the results in [12] where some partial understanding of the analytic approach to the maximum likelihood estimation was presented. The problem here is that routinely fitted phylogenetic models often violate the inequality constraints defining the model. One effect of this phenomenon is then that the maximum likelihood estimators (MLEs) usually lie on to the boundaries of the parameter space (see Section 3 for an example). In a full Bayesian analysis it will make the ensuing inference about probabilities highly sensitive to the settings of prior distributions on the parameters (see [32, 33]). This, in turn, automatically interferes with the appropriate functioning of model selection algorithms. For example Bayes Factor scores will be highly influenced again by priors. On the other hand more classical methods like for example AIC or BIC algorithms, when used routinely, misbehave because many of the MLEs will lie on the boundary of the feasible region since usual dimension counting penalties are implicitly too large (see [38]). For these and other reasons explained in more detail in Section 3, the inequality conditions are of considerable practical importance.

This paper is part of an explosion of work which apply techniques in algebraic geometry to study and develop statistical methodologies. The particular geometric study of tree models was first introduced by Lake [21], and Cavender

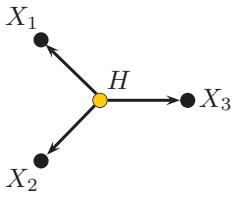

FIG 1. *The graphical representation of the tripod tree model.*

and Felsenstein [9]. This research was initially focused on so called phylogenetic invariants. These are algebraic relationships expressed as a set of polynomial equations over the observed probability tables which must hold for a given phylogenetic model to be valid. We note that these algebraic techniques have also been embraced by computational algebraic geometers [2, 17, 37] enhancing statistical and computational analysis of such models [7] (see also [1] and references therein).

The main technical deficiency of using phylogenetic invariants alone in this way is that they do not give a *full* geometric description of the statistical model. However, the additional inequalities obtained as the main result of this paper complete this description. Where and how these inequality constraints can helpfully supplement an analysis based on phylogenetic invariants is illustrated by the simple example given below.

Example 1.1. Let T be the tripod tree in Figure 1 where we use the convention that observed nodes are depicted by black nodes. The inner node represents a binary hidden variable H and the leaves represent binary observable variables X_1, X_2, X_3. The model is given by all probability distributions p_α for $\alpha \in \{0,1\}^3$ such that

$$p_\alpha = \theta_0^{(H)} \prod_{i=1}^{3} \theta^{(i)}_{\alpha_i|0} + \theta_1^{(H)} \prod_{i=1}^{3} \theta^{(i)}_{\alpha_i|1},$$

where $\theta_i^{(H)} = \mathbb{P}(H = i)$ for $i = 0, 1$ and $\theta^{(i)}_{j|k} = \mathbb{P}(X_i = j|H = k)$ for $i = 1, 2, 3$ and $j, k = 0, 1$. The model has full dimension over the space of observed marginal distributions (X_1, X_2, X_3) and consequently there are no non-trivial equalities defining it. However, it is not a saturated model since not all the marginal probability distributions over the observed vector (X_1, X_2, X_3) lie in the model class. For example Lazarsfeld and Henry [23, Section 3.1] showed that the second order moments of the observed distribution must satisfy

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

Together with many other constraints we derive later, this constraint, which clearly impacts the inferences we might want to make (see Section 3), is not acknowledged through the study of phylogenetic invariants. Therefore inference based solely on these invariants is incomplete. For example naive estimates de-

rived through these methods can be infeasible within the model class in a sense illustrated later in this paper.

This example and the discussion of some inferential issues discussed above motivated the closer investigation of the semi-algebraic features associated with the geometry of binary tree models with hidden inner nodes. The main problem with the geometric analysis of these models is that, in general, it is hard to obtain all the inequality constraints defining a model explicitly even for very simple examples (see [15, Section 4.3], [18, Section 7]). Despite this, some results can be found in the literature. A binary naive Bayes model was studied by Auvray et al. [3]. There are also some partial results for general tree structures on binary variables given by Pearl and Tarsi [27] 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 provide a simpler and more statistically transparent way to express the constrained space.

The semialgebraic description we obtain here also has an elegant mathematical structure. For example [8] gave an intriguing correspondence between, on the one hand, a correlation system on tree models and on the other distances induced by trees where the length between two nodes in a tree is given as a sum of the length of edges in the path joining them. The new coordinate system for tree models that we introduced in [40] enables us to explore in detail this relationship between probabilistic tree models (also called the tree decomposable distributions in [27]) and tree metrics and extend these results.

It has been known for some time that the constraints on possible distances between any two leaves in the tree imply some additional inequality constraints on the possible covariances between the binary variables represented by the leaves. These inequalities, given in (16), follow from the four-point condition ([29], Definition 7.1.5) together with some other simple non-negativity constraints. By using our new parametrization we are able to show in this paper that these two types of inequality constraints cannot be sufficient to describe the model class. Thus any probability distribution in the model class must satisfy many other additional constraints involving higher order moments. Using our methods we are able to provide the full set of the defining constraints in Theorem 4.7. This is given by a list of polynomial equations and inequalities which describe the set of all probability distributions in the model.

The paper is organized as follows. In Section 2 we briefly introduce general Markov models. We then proceed to describe a convenient new change of coordinates for these models given in [40]. In the new coordinate system the parametrization of the model has an elegant product form. We use this to obtain the full semi-algebraic description of a simple naive Bayes model. In Section 3 we discuss various ways in which an awareness of these implicit inequalities can enrich a statistical analysis of this model class. In Section 4 we state our main theorem and illustrate how it can be used. In Section 5 we discuss these results for a simple quartet tree model.

2. Tree models and tree cumulants

We begin by defining and reviewing a new coordinate system for tree models and demonstrate how it can be used to provide a better understanding of this model class. We list the main results from our previous paper [40] and link it to the results presented in the next sections.

Parametrizations based on moments are one way of providing a structured model a structure more amenable to an algebraic analysis (see [4, 14]). This approach has proved particularly effective in the presence of hidden data (see [31]) since then the analysis of a particular marginal distributions over a subset of the observed variables can be specified as a function of the joint moments containing that subset only. On the other hand when a model class is defined by a set of conditional independences further insight may be provided by reparametrizing to other *functions* of these moments to elegantly represent this additional underlying structure. These functions typically resemble cumulants.

One useful property of standard cumulants is that joint cumulants always vanish whenever the random vector under analysis can be split into two independent subvectors. Here we exploit analogous property using a reparametrization customized to the topology of a particular tree. These tree cumulants are introduced in [40]. They vanish only if some of the edges in the defining tree model are missing. This corresponds to the marginal independence of the leaves of two connected components of the induced forest. The property follows from a more general result in [39, Proposition 4.3] and partly explains the elegant product-like structure of the resulting parametrization in Proposition 2.3.

In this paper we assume that random variables are binary taking values either 0 or 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 observed variables is denoted by X and the subvector of hidden variables by H. A (*directed*) *tree* $T = (V, E)$, where V is the set of vertices (or nodes) and $E \subseteq V \times V$ is the set of edges of T, is a connected (*directed*) graph with no cycles. 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. A rooted tree is usually denoted by T^r. For each $v \in V$ by $\mathrm{pa}(v)$ we denote the node preceding v in T^r. In particular $\mathrm{pa}(r) = \emptyset$. 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*.

Let T denote an undirected tree with n leaves and let $T^r = (V, E)$ denote T rooted in $r \in V$. A Markov process on a rooted tree T^r is a sequence $\{Y_v : v \in V\}$ of random variables such that for each $\alpha = (\alpha_v)_{v \in V} \in \{0,1\}^V$ its joint distribution satisfies

$$p_\alpha(\theta) = \theta^{(r)}_{\alpha_r} \prod_{v \in V \setminus r} \theta^{(v)}_{\alpha_v | \alpha_{\mathrm{pa}(v)}}, \tag{1}$$

where $\theta^{(r)}_{\alpha_r} = \mathbb{P}(Y_r = \alpha_r)$ 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 $\Theta_T = [0,1]^{2|E|+1}$ and the Markov process on T^r by $\widetilde{\mathcal{M}}_T$.

Remark 2.1. The reason to omit the root r in the notation is that this model does not depend on the rooting and is equivalent to the undirected graphical model given by global Markov properties on T. To prove this note that T^r is a perfect directed graph and hence by [22, Proposition 3.28] parametrization in (1) is equivalent to factorization with respect to T. Since T is decomposable, by [22, Proposition 3.19], this factorization is equivalent to the global Markov properties.

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 of $X = (X_1, \ldots, X_n)$ represented by the leaves of T. We assume now that all the inner nodes represent hidden variables. Equation (1) induces a polynomial map $f_T : \Theta_T \to \Delta_{2^n-1}$ obtained by marginalization over all the inner nodes of T

$$p_\beta(\theta) = \sum_{\mathcal{H}} \theta^{(r)}_{\alpha_r} \prod_{v \in V \setminus r} \theta^{(v)}_{\alpha_v|\alpha_{\mathrm{pa}(v)}}, \tag{2}$$

where $\mathcal{H}$ is the set of all $\alpha \in \{0,1\}^V$ such that the restriction to the leaves of T is equal to β. We let $\mathcal{M}_T = f_T(\Theta_T)$ denote the *general Markov model* over the set of observable random variables (c.f. [29, Section 8.3]).

A *semialgebraic set* in $\mathbb{R}^d$ is a finite union of sets given by a finite number of polynomial equations and inequalities. Since Θ_T is a semialgebraic set and f_T is a polynomial map then by [5, Proposition 2.2.7] $\mathcal{M}_T$ is a semialgebraic set as well. Moreover, if f is a polynomial isomorphism from Δ_{2^n-1} to another space then $f(\mathcal{M}_T)$ is also a semialgebraic set. The semialgebraic description of $f(\mathcal{M}_T)$ in $f(\Delta_{2^n-1})$ gives the semialgebraic description of $\mathcal{M}_T$.

The idea behind tree cumulants was to define a polynomial isomorphism from Δ_{2^n-1} to the space of new coordinates $\mathcal{K}_T$. We defined a partially ordered set (poset) of all the partitions of the set of leaves induced by removing edges of the given tree T. Then tree cumulants are given as a function of probabilities induced by a Möbius function on the poset. The details of this change of coordinates are given in Appendix A and are illustrated below.

The tree cumulants are given by $2^n - 1$ coordinates: n means $\lambda_i = \mathbb{E}X_i$ for all $i = 1, \ldots, n$ and a set of real-valued parameters $\{\kappa_I : I \subseteq [n] \text{ where } |I| \geq 2\}$. Each formula for κ_I is expressed as a function of the higher order central moments of the observed variables. These formulae are given explicitly in equation (19) of Appendix A. Since the change of coordinates is a polynomial isomorphism then, by [5, Proposition 2.2.7], the image of $\mathcal{M}_T$ in the space of tree cumulants, denoted by $\mathcal{M}^\kappa_T$, is a semialgebraic set. In this paper we provide the full semialgebraic description of $\mathcal{M}^\kappa_T$, that is the complete set of polynomial equations and inequalities involving the tree cumulants which describes $\mathcal{M}^\kappa_T$ as the subset of $\mathcal{K}_T$, for subsequent use in a statistical analysis of the model class.

Example 2.2. Consider the quartet tree model, i.e. the general Markov model given by the graph in Figure 2. The tree cumulants are given by 15 coordinates:

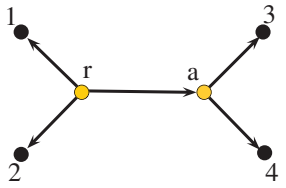

FIG 2. *A quartet tree*

λ_i for $i = 1, 2, 3, 4$ and κ_I for $I \subseteq [4]$ such that $|I| \geq 2$. Denoting $U_i = X_i - \mathbb{E}X_i$ we have $\kappa_{ij} = \mathbb{E}U_iU_j = \mathrm{Cov}(X_i, X_j)$ for $1 \leq i < j \leq 4$ and

$$\kappa_{ijk} = \mathbb{E}\left(U_iU_jU_k\right)$$

for all $1 \leq i < j < k \leq 4$ which we note is a third order central moment. However, in general tree cumulants of higher order cannot be equated with their corresponding central moments but only expressed as functions of them. These functions are obtained by performing an appropriate Möbius inversion. Thus for example from equation (19) in Appendix A we have that

$$\kappa_{1234} = \mathbb{E}\left(U_1U_2U_3U_4\right) - \mathbb{E}\left(U_1U_2\right)\mathbb{E}\left(U_3U_4\right).$$

Note that since the observed higher order central moments can be expressed as functions of probabilities, tree cumulants can also be expressed as functions of these probabilities.

Let $X_{\hat{i}} = (X_1, X_2, X_3, X_4) \setminus \{X_i\}$ for $i = 1, 2, 3, 4$. From [39, Proposition 4.3] it follows in particular that, like for the joint cumulant, $\kappa_{1234} = 0$ whenever $X_i \perp\!\!\!\perp X_{\hat{i}}$ for any $i = 1, 2, 3, 4$ or $(X_1, X_2) \perp\!\!\!\perp (X_3, X_4)$. However, in general, $\kappa_{1234} \neq 0$ for example if $(X_1, X_3) \perp\!\!\!\perp (X_2, X_4)$ and hence tree cumulants differ from classical cumulants. Vanishing of the tree cumulants corresponds to an edge being missing in the particular defining tree. This generalizes for other trees and gives a heuristic explanation for the nice product-like parametrization presented in Proposition 2.3 below. We explain this now formally.

Let $T^r = (V, E)$ and let Ω_T denote the set of parameters with coordinates given by $\bar{\mu}_v$ for $v \in V$ and $\eta_{u,v}$ for $(u, v) \in E$. Define a reparametrization map $f_{\theta\omega} : \Theta_T \to \Omega_T$ as follows:

$$\begin{array}{ll} \eta_{u,v} = \theta^{(v)}_{1|1} - \theta^{(v)}_{1|0} & \text{for every } (u, v) \in E \text{ and} \\ \bar{\mu}_v = 1 - 2\lambda_v & \text{for each } v \in V, \end{array} \tag{3}$$

where $\lambda_v = \mathbb{E}Y_v$ is a polynomial in the original parameters θ. To see this let $r, v_1, \ldots, v_k, v$ be a directed path in T. Then

$$\lambda_v = \mathbb{P}(Y_v = 1) = \sum_{\alpha \in \{0,1\}^{k+1}} \theta^{(v)}_{1|\alpha_k}\theta^{(v_k)}_{\alpha_k|\alpha_{k-1}} \cdots \theta^{(r)}_{\alpha_r}. \tag{4}$$

It can be easily checked that if $\mathrm{Var}(Y_u) > 0$ then $\eta_{u,v} = \mathrm{Cov}(Y_u, Y_v)/\mathrm{Var}(Y_u)$. Hence $\eta_{u,v}$ is just the regression coefficient of Y_v with respect to Y_u.

The parameter space Ω_T is given by the following constraints:

$$\begin{aligned} &-1 \le \bar{\mu}_r \le 1, \qquad \text{and for each } (u,v) \in E \\ &-(1+\bar{\mu}_v) \le (1-\bar{\mu}_u)\eta_{u,v} \le (1-\bar{\mu}_v) \\ &-(1-\bar{\mu}_v) \le (1+\bar{\mu}_u)\eta_{u,v} \le (1+\bar{\mu}_v). \end{aligned} \tag{5}$$

In Appendix A we show that there is a polynomial isomorphism between Δ_{2^n-1} and the space of tree cumulants $\mathcal{K}_T$ giving the following diagram, where the dashed arrow denotes the induced parametrization.

$$\begin{array}{ccc} \Theta_T & \xrightarrow{\quad f_T \quad} & \Delta_{2^n-1} \\ {\scriptstyle f_{\omega\theta}}\uparrow \;\downarrow{\scriptstyle f_{\theta\omega}} & & {\scriptstyle f_{\kappa p}}\uparrow \;\downarrow{\scriptstyle f_{p\kappa}} \\ \Omega_T & \overset{\psi_T}{- - - - \to} & \mathcal{K}_T \end{array} \tag{6}$$

One motivation behind this change of coordinates is that the induced parametrization $\psi_T : \Omega_T \to \mathcal{K}_T$ has a particularly elegant form.

Proposition 2.3 ([40], Proposition 4.1)**.** *Let T be an undirected tree with n leaves. Assume that T is trivalent which here means that all of its inner nodes have degree at most three. Let $T^r = (V, E)$ be T rooted in $r \in V$. Then $\mathcal{M}_T^\kappa$ is parametrized by the map $\psi_T : \Omega_T \to \mathcal{K}_T$ given as $\lambda_i = \frac{1}{2}(1-\bar{\mu}_i)$ for $i = 1, \ldots, n$ and*

$$\kappa_I = \frac{1}{4}\left(1 - \bar{\mu}_{r(I)}^2\right) \prod_{v \in \mathrm{int}(V(I))} \bar{\mu}_v^{\deg(v)-2} \prod_{(u,v)\in E(I)} \eta_{u,v} \quad \text{for } I \subseteq [n], |I| \ge 2 \tag{7}$$

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 $T^r(I)$.

Proposition 2.3 has been formulated for trivalent trees. However, it can be easily extended to the general case as explained in [40, Section 4].

This result enabled us to completely understand identifiability of tree models extending results in [10]. In particular [40, Theorem 5.4] identifies the cases when the model is identified up to label switching. This condition is rather technical and here we usually would recommend the use of the sufficient condition that all the covariances between the leaves are nonzero. Further results focus on the geometry of the unidentified space in the case when the identifiability fails. More importantly, [40, Corollary 5.5] gives us formulae for parameters given a probability distribution in the case when identifiability holds. This result gives us a closed-form formulae for MLEs in certain special cases (see Corollary 3.1).

To illustrate our technique we next obtain the full semialgebraic description of the tripod tree model. This result is not new (see e.g. [3, 30] and a special case given by [26, Theorem 3.1]). However, this allows us not only to unify notation but also to introduce the strategy we use to prove the general case. We begin with a definition.

Definition 2.4. Let A be a $2 \times 2 \times 2$ table. The hyperdeterminant of A as defined by Gelfand, Kapranov, Zelevinsky [19, Chapter 14] is given by

$$\begin{aligned}\operatorname{Det} A &= (a_{000}^2 a_{111}^2 + a_{001}^2 a_{110}^2 + a_{010}^2 a_{101}^2 + a_{011}^2 a_{100}^2)\\ &- 2(a_{000}a_{001}a_{110}a_{111} + a_{000}a_{010}a_{101}a_{111} + a_{000}a_{011}a_{100}a_{111}\\ &+ a_{001}a_{010}a_{101}a_{110} + a_{001}a_{011}a_{110}a_{100} + a_{010}a_{011}a_{101}a_{100})\\ &+ 4(a_{000}a_{011}a_{101}a_{110} + a_{001}a_{010}a_{100}a_{111}).\end{aligned}$$

If $\sum a_{ijk} = 1$ then treating all entries formally as joint cell probabilities (without positivity constraints) we can simplify this formula using the change of coordinates to central moments. The reparametrizations in Appendix A are well defined for this extended space of probabilities and we have that

$$\operatorname{Det} A = \mu_{123}^2 + 4\mu_{12}\mu_{13}\mu_{23}, \tag{8}$$

which can be verified by direct computations.

From the construction of tree cumulants (c.f. Appendix A) it follows that $\kappa_I = \mu_I$ for all $I \subseteq [n]$ such that $2 \leq |I| \leq 3$. Henceforth, for clarity, these lower order tree cumulants will be written as their more familiar corresponding central moments.

Proposition 2.5 (The semialgebraic description of the tripod model)**.** *Let $\mathcal{M}_3$ be the general Markov model on a tripod tree T rooted in any node of T. Let P be a $2 \times 2 \times 2$ probability table for three binary random variables (X_1, X_2, X_3) with central moments $\mu_{12}, \mu_{13}, \mu_{23}$, μ_{123} (equivalent to the corresponding tree cumulants) and means λ_i, for $i = 1, 2, 3$. Then $P \in \mathcal{M}_3$ if and only if one of the following two cases occurs:*

(i) $\mu_{123} = 0$ and at least two of the three covariances $\mu_{12}, \mu_{13}, \mu_{23}$ vanish.
(ii) $\mu_{12}\mu_{13}\mu_{23} > 0$ and

$$\begin{aligned}|\mu_{jk}|\sqrt{\operatorname{Det} P} - \mu_{123}\mu_{jk} &\leq (1 - \bar{\mu}_i)\mu_{jk}^2,\\ |\mu_{jk}|\sqrt{\operatorname{Det} P} + \mu_{123}\mu_{jk} &\leq (1 + \bar{\mu}_i)\mu_{jk}^2\end{aligned} \tag{9}$$

for all $i = 1, 2, 3$ where by j, k we denote elements of $\{1, 2, 3\} \setminus i$.

Sketch of the proof. The proof is given in Appendix B. Here, for convenience, we give its outline. Denote by $\mathcal{M} \subseteq \Delta_7$ the family of distributions described by (i) and (ii). We need to show that $\mathcal{M}_3 = \mathcal{M}$. To show that $\mathcal{M}_3 \subseteq \mathcal{M}$ we use the parametrization in Proposition 2.3 to prove that either (i) holds or it does not, and then, inequalities in (ii) are equivalent to (5). To show the opposite inclusion we propose formulae for the parameters in terms of the observed distribution given by [40, Corollary 5.5], and show that this formulae agree with the parametrization in Proposition 2.3 up to the sign. The inequality $\mu_{12}\mu_{13}\mu_{23} > 0$ assures that there is a choice of signs for the parameters such that the parametrization holds exactly. □

All the points satisfying (i) correspond to submodels of $\mathcal{M}$ where some of the observed variables are independent of each other.

3. Inferential issues related to the semialgebraic description

There are at least three reasons why the implicit inequality constraints of this model class can have a critical impact on a statistical analysis of this model class. First, used in conjunction with other geometric techniques these inequalities help us determine, whether or not the likelihood associated with a given tree model has multiple local maxima. Second, it gives us the basis for developing simple model diagnostics which complement those associated with implicit algebraic constraints. Finally, awareness of whether these constraints are active for given data set enables us to identify when standard numerical methods might fail both for estimation and model selection across different candidate trees. We consider and illustrate all these issues below.

Proposition 2.5 and Theorem 4.7 give explicit descriptions of tree models as subsets of the probability simplex and hence also as submodels of the multinomial model. The literature on constrained multinomial models (see [13] for a review) gives many examples of what may go wrong in this case. If the multiway marginal table of observed random variables is sampled at random then its likelihood will be given as the multinomial likelihood constrained to the model. The unconstrained multinomial likelihood is of course a very well-behaved function. In particular it is log-concave and its unique maximum is given by the sample proportions $\hat{p}$ as long as all the entries of $\hat{p}$ are nonzero. However, after constraining to the model this function may become much more complicated.

We know that unidentifiability of parameters causes estimation problems associated for example with multiple local maxima of the likelihood and the posterior density. However, because the constraints on the model do not define a convex region, the constrained likelihood will not necessarily have a unique maximum (see Figure 4). So even if we use ways of cleverly accounting for the aliasing caused by unidentifiability we can still be left with other multiple local solutions induced by the violations of the constraints. This, in turn, can make estimation schemes unstable. The discussion below complements results presented in [12].

If the unconstrained multinomial maximum likelihood estimator given by the sample proportions satisfies the equation but does not satisfy some of the inequalities then the MLE of the given tree model will *always* lie on the boundary of the parameter space Θ_T. Of course, if all the inequalities hold but some of the equalities do not then, in principle, it is not such a serious problem as the estimates will typically lie in the interior of the parameter space. However, if there are even the smallest perturbations of the model class we are likely to be drawn outside the feasible region. This is a phenomenon observed in many applied analyzes of these models (see, e.g. [12]). This occurs even in the simple tripod tree above where the feasible region accounts for only 8% of Δ_7. Of course simply sampling from the tree model itself will not identify this potential difficulty since such samples will automatically not violate the constraints in any significant way. But if the tree only approximately holds then we begin to encounter certain difficulties.

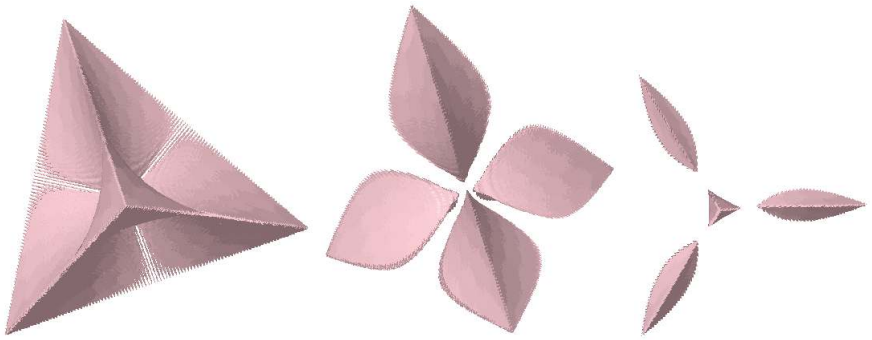

FIG 3. *The space of all possible covariances* $\mu_{12}, \mu_{13}, \mu_{23}$ *for the tripod tree model in the case when* $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{2}$ *and* μ_{123} *is equal to* 0, 0.005 *and* 0.02 *(from left to right).*

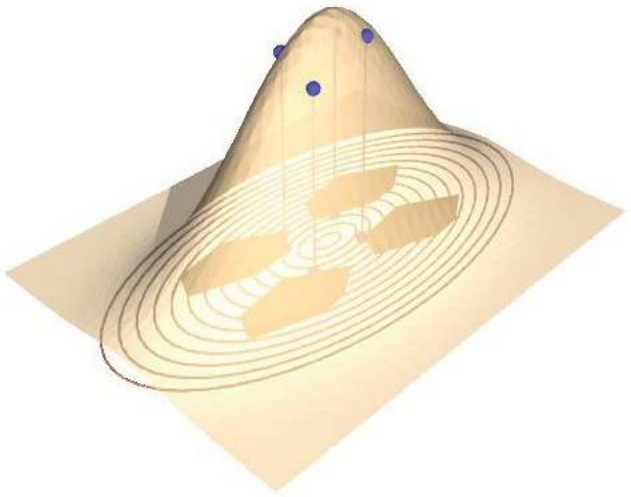

FIG 4. *The multinomial likelihood and a submodel of the saturated model given by four disjoint regions. The four local maxima are obtained on boundaries of these regions.*

Since the tripod tree model $\mathcal{M}_3$ is of full dimension there are no non-trivial phylogenetic invariants and so the feasible regions of the model class are purely associated with inequality constraints and so particularly straightforward. In Figure 3 we depict these constraints as they apply to the second order moments of the three observed variables given some typical values of the other coordinates. For example there are four components corresponding to four possible choices of signs for covariances satisfying $\mu_{12}\mu_{13}\mu_{23} \geq 0$.

We can now give an explicit illustration of the type of multimodality that can be induced in this context. The likelihood function $\ell : \Theta_T \to \mathbb{R}$ for the tripod tree model can be also treated as a function on Δ_7 by $\ell(\theta) = \ell(p(\theta))$ in

which case it will be denoted by $\ell(p)$. Since we understand the parametrization $p : \Theta_T \to \Delta_7$ of $\mathcal{M}_3$ then understanding $\ell(p)$ gives us automatically understanding of $\ell(\theta)$. The advantage is that in this setting $\ell(p)$ is just obtained as the multinomial likelihood function $\ell(p) = \ell(p; x) = \prod p_{ijk}^{x_{ijk}}$ constrained to the model as explained above. If $\hat{p}$ lies in the model class $\mathcal{M}_3$ then $\ell(p)$ has a unique maximum and the maxima of $\ell(\theta)$ can be obtained by mapping back $\hat{p}$ to the parameter space Θ_T by using [40, Equation (3)]. This result generalizes.

Corollary 3.1. *Let $T = (V, E)$ be a phylogenetic tree with n leaves and let $\mathcal{M}_T$ be the corresponding tree model. If $\hat{p} \in \mathcal{M}_T$ then [40, Corollary 5.5] gives the formulae for the maximum likelihood estimators. In the case when the number of MLEs is finite, there are always exactly $2^{|V|-n}$ MLEs which are equivalent up to switching labels of the hidden variables.*

We have however argued that usually $\hat{p} \notin \mathcal{M}_T$. In this case there is potentially more than one local maximum of the constrained multinomial likelihood function. Let $\hat{p}$ the sample proportions for some observed data on three binary random variables. We have three possible scenarios:

(i) $\hat{p} \in \mathcal{M}_3$ and then $\ell(p)$ is unimodal.
(ii) $\hat{p} \notin \mathcal{M}_3$ and $\ell(p)$ is multimodal but there exists only one global maximum.
(iii) $\hat{p} \notin \mathcal{M}_3$ and $\ell(p)$ has multiple global maxima.

The situation in (iii) raises an interesting question related to the model identifiability. For every data point satisfying (iii) we are not able to identify the parameters using the maximum likelihood estimation even if we take into account the label switching problem.

Of course from the numerical point of view the situation in (ii) and (iii) may describe equally bad scenarios since in both cases the algorithms become unstable even for arbitrary large sample sizes. Thus suppose that a sample of size 10000 has been observed

$$\left[\begin{array}{cc|cc} x_{000} & x_{001} & x_{100} & x_{101} \\ x_{010} & x_{011} & x_{110} & x_{111} \end{array}\right] = \left[\begin{array}{cc|cc} 2069 & 16 & 2242 & 331 \\ 2678 & 863 & 442 & 1359 \end{array}\right]. \tag{10}$$

By direct computations we check that all the constraint in Proposition 2.5 hold apart from $\mu_{12}\mu_{13}\mu_{23} \geq 0$ and hence $\hat{p}$ does not lie in $\mathcal{M}_3$. The corresponding parameters will lie on the boundary of the parameter space. We performed the following simulation. We sampled uniformly from $\Theta_T = [0, 1]^7$ the starting parameters for the EM algorithm and noted the results of the EM approximation. For 100 iterations the procedure found four different isolated maxima given in Table 1.

Up to label switching on the inner node these are two distinct maximizers of the log-likelihood function $\ell(\theta)$ corresponding to rows $1, 3$. The value of the log-likelihood function, computed as $\sum_{ijk} x_{ijk} \log p_{ijk}$, is equal to -18387 and -18917 respectively. Both points correspond to somewhat degenerate tripod tree models where one of the observed variables is functionally related to the hidden variable. For example the first point lies on the submodel given by $X_1 \perp\!\!\!\perp X_3 | X_2$. We performed a similar analysis for other data points for which

TABLE 1
Results of the EM algorithm

	$\theta_1^{(r)}$	$\theta_{1\|0}^{(1)}$	$\theta_{1\|1}^{(1)}$	$\theta_{1\|0}^{(2)}$	$\theta_{1\|1}^{(2)}$	$\theta_{1\|0}^{(3)}$	$\theta_{1\|1}^{(3)}$
1	0.4658	0.3371	0.5524	1.0000	0.0000	0.4159	0.0745
2	0.5342	0.5524	0.3371	0.0000	1.0000	0.0745	0.4159
3	0.4771	0.0000	0.9167	0.6369	0.4216	0.1468	0.3775
4	0.5229	0.9167	0.0000	0.4216	0.6369	0.3775	0.1468

only $\mu_{12}\mu_{13}\mu_{23} \geq 0$ fails and three different EM maximizers were often found. In every case the maximizers corresponded to degenerate submodels. In conjunction with [40, Theorem 5.4] we also have data for which the likelihood function $\ell(\theta)$ is maximized over an infinite number of points. This for example holds for any data such that the constrained multinomial likelihood is maximized over a point such that $p_{0ij} = \lambda p_{1ij}$ for some λ and each $i, j = 0, 1$. In this case $\mu_{12} = \mu_{13} = 0$ and the MLEs form a set of a positive dimension by [40, Theorem 5.4].

We note that the whole discussion above remains valid for more general tree models. The conditional independence properties of tree models imply that, since any three leaves are separated by an inner node, the corresponding marginal distributions form a tripod tree model. Demanding that tripod tree constraints must be satisfied by *all* triples of observed random variables cuts out all but a small proportion of the probability simplex. Furthermore by Theorem 4.7 we know that, in addition, many other constraints involving higher order moments will also apply. Therefore, the types of issues we illustrated above become increasingly critical for inference on trees, which in practical applications are of a much higher dimension. Thus real-world data will typically satisfy all the constraints defining the model very rarely. This, in turn, tends to result in multimodality of the likelihood function and MLEs lying on the boundary of the parameter space.

By acknowledging the existence of the inequality constraints we have already demonstrated how graphical methods can be used to identify why and where the fitted tree model might be flawed. Most naively, when samples are very large we could calculate the sample moments and notice which inequality constraints are active on the data set presented. When these lie outside these regions then we have strong information that the fitted tree model is inappropriate and we can expect there to be problems with both estimation - as illustrated above - and model selection. Slightly more sophisticatedly we could also compare the model MLE: constrained as it is by these inequalities, with the MLE in the saturated model. Likelihood ratio statistics can then be used to measure the extent of the model inaccuracy. Of course this comparison can be performed directly. However, then we lose the geometrical insight as to exactly why and how the model is failing. This insight will be helpful in guiding us in identifying alternative models that might better explain the data. We note that the likelihood ratio statistics for a constrained multinomial model against the saturated model in general will not asymptotically have the χ^2 distribution (see e.g. [11]). If the constrains are linear then the underlying distribution is called the chi-bar

squared distribution (see [13]). The situation is however much more complicated for tree models since here the constraints define a union of non-convex bodies. In the end of Section 4 we provide a short discussion on a description of $\mathcal{M}_T$ in terms of convex sets.

Inequalities are also relevant for the model choice. Suppose that the sufficient statistic does not satisfy some inequalities for each of the models under analysis. Then asymptotic model selection techniques like BIC can mislead. The effective parameter size will be miscounted because at least some of the MLEs will lie of the boundary of the space (see e.g. [28, 38]). Model selection based on Bayes factors will also tend to be unrobust. Since the estimates lie on the boundary the marginal likelihood for each of the models depends heavily on the tail behavior of the prior distribution on that boundary. See [32] and [33] for explanations of why this is so. For example a standard choice of a prior distribution for conditional distributions in tree models is the Dirichlet distribution. However, for different choices of its prior parameters the Bayes factors generated by the prior tails can be very different. Note that within the Bayesian paradigm the sampling of the tripod tree is straightforward once we recognize the constraint structure using a simple importance sampler generating samples from Δ_7 and rejecting if they do not satisfy the defining inequalities. Of course this is not the only way of specifying a prior density for selecting between the saturated model and the tree model. However, our suggestion is very simple to implement and its inferential consequences are more transparent than more conventional methods using default priors within the conventional probabilitistic parametrization, where the selection can be highly dependent on the tails of priors.

4. Explicit expression of implied inequality constraints

In this section we discuss the geometry of general tree models. First, we use some links to tree metrics to provide a simple set of algebraic constraints on the model space. Then, in Theorem 4.7, we provide the complete semialgebraic description for this model class.

Let $T = (V, E)$ be a general undirected tree with n leaves and T^r the tree T rooted in $r \in V$. Before stating the main theorem of the paper we first show how to obtain an elegant set of necessary constraints on $\mathcal{M}_T$. In this section we assume that $\bar{\mu}_r^2 \neq 1$ and $\eta_{u,v} \neq 0$ for all $(u, v) \in E$. By [40, Remark 4.3], this implies that $\bar{\mu}_v^2 \neq 1$ for all $v \in V$. Since $\mathrm{Var}(Y_u) = \frac{1}{4}(1 - \bar{\mu}_u^2)$ the correlation between Y_u and Y_v is defined as $\rho_{uv} = \frac{4\mu_{uv}}{\sqrt{(1-\bar{\mu}_u^2)(1-\bar{\mu}_v^2)}}$. This gives

$$\rho_{uv} = \eta_{u,v}\sqrt{\frac{1-\bar{\mu}_u^2}{1-\bar{\mu}_v^2}} = \eta_{v,u}\sqrt{\frac{1-\bar{\mu}_v^2}{1-\bar{\mu}_u^2}}. \tag{11}$$

Lemma 4.1. *For any $i, j \in [n]$ let $E(ij)$ be the set of edges on the unique path joining i and j in T. Then*

$$\rho_{ij} = \prod_{(u,v)\in E(ij)} \rho_{uv} \tag{12}$$

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

Proof. By (7) applied to $T(ij)$ we have $\mu_{ij} = \frac{1}{4}(1-\bar{\mu}_r^2)\prod_{(u,v)\in E(ij)}\eta_{u,v}$, where r is the root of the path between i and j and hence

$$\rho_{ij} = \sqrt{\frac{1-\bar{\mu}_r^2}{1-\bar{\mu}_i^2}}\sqrt{\frac{1-\bar{\mu}_r^2}{1-\bar{\mu}_j^2}}\prod_{(u,v)\in E(ij)}\eta_{u,v}.$$

Now apply (11) to each $\eta_{u,v}$ in the product above to show (12). □

The above equation allows us to demonstrate an interesting reformulation of our problem in term of tree metrics (c.f. [29, Section 7]) which we explain below (see also Cavender [8]).

Definition 4.2. A function $\delta : [n]\times[n]\to\mathbb{R}$ is called a *tree metric* if there exists a tree $T=(V,E)$ with the set of leaves given by $[n]$ and with a positive real-valued weighting $w : E\to\mathbb{R}_{>0}$ such that for all $i,j\in[n]$

$$\delta(i,j) = \begin{cases} \sum_{e\in E(ij)} w(e), & \text{if } i\neq j,\\ 0, & \text{otherwise.}\end{cases}$$

Let now $d : V\times V\to\mathbb{R}$ 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 $K\in\mathcal{M}_T^\kappa$ then by (12) $\rho_{ij}^2 = \prod_{e\in E(ij)}\rho_e^2$ and we can define map $d_{(T;K)} : [n]\times[n]\to\mathbb{R}$

$$-\log(\rho_{ij}^2) = d_{(T;K)}(i,j) = \begin{cases} \sum_{(u,v)\in E(ij)} d(u,v), & \text{if } i\neq j,\\ 0, & \text{otherwise.}\end{cases} \tag{13}$$

This map is a tree metric by Definition 4.2. In our case we have a point in the model space defining all the second order correlations and $d_{(T;K)}(i,j)$ 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 (13) 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 4.3 (Tree-Metric Theorem, Buneman [6])**.** *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 $i,j,k,l\in[n]$,*

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

Moreover, a tree metric defines the tree uniquely.

This theorem gives us a set of explicit constraints on the distributions in a tree model. Since $\delta(i,j) = \log(-\rho_{ij})$ the constraints in Theorem 4.3 translate in terms of correlations to

$$-\log(\rho_{ij}^2\rho_{kl}^2) \leq -\min\{\log(\rho_{ik}^2\rho_{jl}^2), \log(\rho_{il}^2\rho_{jk}^2)\}.$$

Since log is a monotone function we obtain

$$\min\left\{\frac{\rho_{ik}^2\rho_{jl}^2}{\rho_{ij}^2\rho_{kl}^2}, \frac{\rho_{il}^2\rho_{jk}^2}{\rho_{ij}^2\rho_{kl}^2}\right\} = \min\left\{\frac{\mu_{ik}^2\mu_{jl}^2}{\mu_{ij}^2\mu_{kl}^2}, \frac{\mu_{il}^2\mu_{jk}^2}{\mu_{ij}^2\mu_{kl}^2}\right\} \leq 1 \tag{14}$$

for all not necessarily distinct leaves $i,j,k,l \in [n]$. Hence, using the relation between correlations and tree metrics given in [8] we managed to provide a set of simple semialgebraic constraints on the model. Furthermore, later in Theorem 4.7 we show that these constraints are not the only active constraints on the model $\mathcal{M}_T$. Before we present this theorem it is helpful to make some simple observations about the relationship between correlations and probabilistic tree models.

Since ρ_{uv} can have different signs we define a signed tree metric as a tree metric with an additional sign assignment for each edge of T.

Lemma 4.4. *Let T be a tree with set of leaves $[n]$. Suppose that we have a map $\sigma : [n] \times [n] \to \{-1,1\}$. Then there exists a map $s_0 : E \to \{-1,1\}$ such that for all $i,j \in [n]$*

$$\sigma(i,j) = \prod_{(u,v)\in E(ij)} s_0(u,v) \tag{15}$$

if and only if for all triples $i,j,k \in [n]$ $\sigma(i,j)\sigma(i,k)\sigma(j,k) = 1$.

The proof is given in Appendix B.

The following proposition gives a set of simple constraints on probability distribution in tree models. This may be particularly useful in practice since it involves only computing pairwise margins of the data and it enables us to check if a data point may come from a phylogenetic tree model.

Proposition 4.5. *Let $P \in \Delta_{2^n-1}$ be a probability distribution. If $P \in \mathcal{M}_T$ for some tree T with n leaves then*

$$0 \leq \min\left\{\frac{\mu_{ik}\mu_{jl}}{\mu_{ij}\mu_{kl}}, \frac{\mu_{il}\mu_{jk}}{\mu_{ij}\mu_{kl}}\right\} \leq 1 \tag{16}$$

for all (not necessarily distinct) $i,j,k,l \in [n]$ whenever $\mu_{ij}, \mu_{kl} \neq 0$.

Proof. Lemma 4.4 implies that for all $i,j,k \in [n]$ necessarily $\mu_{ij}\mu_{ik}\mu_{jk} \geq 0$. This in particular implies that $\frac{\mu_{ik}\mu_{jl}}{\mu_{ij}\mu_{kl}} \geq 0$ for all $i,j,k,l \in [n]$. By taking the square root in (14) these constraints can be combined to give the inequalities in (16). □

In Theorem 4.7 we show that (16) provides the complete set of inequality constraints on $\mathcal{M}_T$ that involve only second order moments in their expression.

The fact that additional constraints involving higher order moments exist is illustrated in the following simple example.

Example 4.6. Consider the tripod tree model in Proposition 2.5. Let K be a point in $\mathcal{K}_T$ given by $\lambda_i = 0.15$ for $i = 1, 2, 3$, $\mu_{ij} = 0.0625$ (or equivalently $\rho_{ij} = 0.49$) for each $i < j$ and $\mu_{123} = 0.0526$. This point lies in the space of tree cumulants $\mathcal{K}_T$ which can be checked by mapping back the central moments to probabilities, since the resulting vector $[p_\alpha]$ lies in Δ_7.

Clearly K satisfies all the tree metric constraints in (16). The equation (12) is satisfied with $\rho_{hi} = 0.7$ for each $i = 1, 2, 3$. We now show that despite this $K \notin \mathcal{M}_T^\kappa$. For if $K \in \mathcal{M}_T^\kappa$ then we could find $\bar{\mu}_h$ and $\eta_{h,i}$ satisfying constraints in (5) so that (21) held. Using the formulae in [40, Corollary 5.5] it is easy to compute that $\bar{\mu}_h = 0.86$ and $\eta_{h,i} \approx 0.98$. However, K is not in the model since these parameters do not lie in Ω_T. Indeed,

$$(1 + \bar{\mu}_h)\eta_{h,i} \approx 1.8228 > (1 + \bar{\mu}_i) = 1.7$$

and hence (5) is not satisfied.

The consequence of the fact that the parameters do not lie in Ω_T is that this parametrization does not lead to a valid assignment of conditional probabilities to the edges of the tree. For example with the values given above we can calculate that the induced marginal distribution for (X_i, H) would have to satisfy $\mathbb{P}(X_i = 0, H = 1) = -0.0043$ which is obviously not a consistent assignment for a probability model. Thus, there must exist other constraints involving observed higher order moments that need to hold for a probability model to be valid. We note that for the tripod tree these were given by Proposition 2.5.

The following theorem gives the complete set of constraints which have to be satisfied by tree cumulants to lie in $\mathcal{M}_T$ in the case when T is a trivalent tree. Let $P \in \Delta_{2^n-1}$ be the probability distribution of the vector $(X_1, \ldots, X_n)$ then for any $i, j, k \in [n]$ let P^{ijk} denote the $2 \times 2 \times 2$ table of the marginal distribution of (X_i, X_j, X_k).

Theorem 4.7. *Let $T = (V, E)$ be a trivalent tree with n leaves and $\mathcal{M}_T \subseteq \Delta_{2^n-1}$ be the model defined as an image of the parametrization in (2). Suppose P is a joint probability distribution on n binary variables. Then $P \in \mathcal{M}_T$ if and only if the following conditions hold:*

(C1) *For each edge split $A|B$ (c.f. Definition A.1) of the set of leaves of T whenever we have four nonempty subsets (not necessarily disjoint) $I_1, I_2 \subseteq A$, $J_1, J_2 \subseteq B$ then*

$$\kappa_{I_1 J_1}\kappa_{I_2 J_2} - \kappa_{I_1 J_2}\kappa_{I_2 J_1} = 0.$$

(C2) *For all $1 \leq i < j < k \leq n$ the corresponding marginal distribution P^{ijk} lies in the tripod model.*

(C3) *for all $I \subseteq [n]$ if there exist $i, j \in I$ such that $\mu_{ij} = 0$ then $\kappa_I = 0$*

(C4) *for any $i,j,k,l \in [n]$ such that there exists $e \in E$ inducing a split $A|B$ such that $i,j \in A$ and $k,l \in B$ we have*

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

Moreover, if $\mu_{ij} \neq 0$ for all $i,j \in [n]$ then the constraints in Proposition 4.5 are the only constraints involving only second order moments.

Sketch of the proof. The proof is given in Appendix C. Here, for convenience, we give its outline. Denote by $\mathcal{M} \subseteq \Delta_{2^n-1}$ the family of distributions described by (C1)-(C4). We need to show that $\mathcal{M}_T = \mathcal{M}$. To show that $\mathcal{M}_T \subseteq \mathcal{M}$ we use the parametrization in Proposition 2.3 to show that (C1) and (C3) always hold, and that (C2) and (C4) are equivalent to (5). To show the opposite inclusion we propose formulae for the parameters in terms of the observed distribution given by [40, Corollary 5.5], and show that this formulae agree with the parametrization in Proposition 2.3 up to the sign. The last part is technical since we need to show that (C1)-(C4) also imply that there is a choice of signs for the parameters such that the parametrization in Proposition 2.3 holds exactly. □

Theorem 4.7 has been formulated for trivalent trees. However, any tree with degrees of some nodes higher than three can be realized as a submodel of a trivalent tree model as explained in [40, Section 4]. Also, including degree two nodes does not change anything in the induced marginal distribution. This result is well known (see e.g. [40, Lemma 2.1]).

A natural question arises for how large trees it is feasible to verify the constraints defining the model. The equality constraints in (C1) can be expressed directly in the raw probabilities and they are easy to check even for relatively large trees. This, by [2, Theorem 4],can be done using so called edge flattenings, which is explained in more details in Appendix D. Checking the other constraints requires only computing $\binom{n}{2}$ covariances between the observed variables and $\binom{n}{3}$ third order central moments. In particular, in practice there is no need of changing the coordinates from the raw probabilities to tree cumulants which can be quite complicated even for relatively small trees.

Another important practical aspect is whether there exist some efficient convex bounds for the model in the space of the raw probabilities. The answer to this question is negative, which follows from the fact that $\mathrm{conv}(\mathcal{M}_T) = \Delta_{2^n-1}$. This is easily seen from the fact that $\mathcal{M}_{\mathrm{ind}} \subseteq \mathcal{M}_T$, where $\mathcal{M}_{\mathrm{ind}}$ denotes the model of full independence $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n$, and that $\mathrm{conv}(\mathcal{M}_{\mathrm{ind}}) = \Delta_{2^n-1}$. To get some informative convex bounds one possibility is to generalize the tripod tree case. Here the model consists of four components depicted in Figure 4 corresponding to different sign patterns of the observed covariances. These components are equivalent up to rotation and symmetry. Instead of taking the convex hull of the whole model we suggest the analysis of the convex hull of each of the components separately. This is also well motivated by the fact that in phylogenetics it is usually assumed that $\eta_{u,v} > 0$ for all $(u,v) \in E$ which means restriction to one of the components with all the observed covariances positive. We will not discuss this issue here in more detail.

TABLE 2
Moments and tree cumulants for a probability assignment which lies in $\mathcal{M}_T$, where T is the quartet tree

α	I	p_α	λ_I	κ_I
0000	$\emptyset$	$\frac{163837}{1417176}$	1	0
0001	4	$\frac{100735}{1417176}$	$\frac{1}{2}$	0
0010	3	$\frac{48167}{708588}$	$\frac{1}{2}$	0
0011	34	$\frac{45955}{708588}$	$\frac{253}{972}$	$\frac{5}{486}$
0100	2	$\frac{85507}{1417176}$	$\frac{1}{2}$	0
0101	24	$\frac{76007}{1417176}$	$\frac{251}{972}$	$\frac{2}{243}$
0110	23	$\frac{36559}{708588}$	$\frac{85}{324}$	$\frac{1}{81}$
0111	234	$\frac{35531}{708588}$	$\frac{2489}{17496}$	$\frac{4}{2187}$
1000	1	$\frac{41255}{708588}$	$\frac{1}{2}$	0
1001	14	$\frac{37315}{708588}$	$\frac{253}{972}$	$\frac{5}{486}$
1010	13	$\frac{73199}{1417176}$	$\frac{43}{162}$	$\frac{5}{324}$
1011	134	$\frac{75355}{1417176}$	$\frac{1271}{8748}$	$\frac{5}{2187}$
1100	12	$\frac{43471}{708588}$	$\frac{829}{2916}$	$\frac{25}{729}$
1101	124	$\frac{44171}{708588}$	$\frac{8107}{52488}$	$\frac{20}{6561}$
1110	123	$\frac{97063}{1417176}$	$\frac{1405}{8748}$	$\frac{10}{2187}$
1111	1234	$\frac{130547}{1417176}$	$\frac{130547}{1417176}$	$\frac{40}{59049}$

5. Example: The quartet tree model

We can check that the point $K \in \mathcal{K}_T$ provided in Table 2 satisfies all the constraints in Theorem 4.7. It is convenient to provide the numbers as rationals so that the equalities can be checked exactly. To check (C1), note for example that

$$\kappa_{13}\kappa_{24} - \kappa_{14}\kappa_{23} = \frac{5}{324}\cdot\frac{2}{243} - \frac{5}{486}\cdot\frac{1}{81} = 0,$$

$$\kappa_{123}\kappa_{134} - \kappa_{1234}\kappa_{13} = \frac{10}{2187}\cdot\frac{5}{2187} - \frac{40}{59049}\cdot\frac{5}{324} = 0.$$

To check (C2) verify for example that $\mathrm{Det}P^{123} = \frac{25}{531441}$ and

$$\left((1 \pm \bar{\mu}_1)\mu_{23} \mp \mu_{123}\right)^2 = \left\{\frac{1369}{4782969}, \frac{289}{4782969}\right\}$$

$$\left((1 \pm \bar{\mu}_2)\mu_{13} \mp \mu_{123}\right)^2 = \left\{\frac{30625}{76527504}, \frac{9025}{76527504}\right\}$$

$$\left((1 \pm \bar{\mu}_3)\mu_{12} \mp \mu_{123}\right)^2 = \left\{\frac{7225}{4782969}, \frac{4225}{4782969}\right\}$$

and hence

$$\mathrm{Det}P^{123} \le \min\left\{\left((1 \pm \bar{\mu}_{\sigma(i)})\mu_{\sigma(j)\sigma(k)} \mp \mu_{ijk}\right)^2\right\} = \frac{289}{4782969}$$

is satisfied.

TABLE 3
Moments and tree cumulants of the given probability assignment which does not lie in $\mathcal{M}_T$

α	I	p_α	λ_I	κ_I
0000	$\emptyset$	$\frac{163837}{1417176}$	1	0
0001	4	$\frac{83213}{1417176}$	$\frac{1}{2}$	0
0010	3	$\frac{10999}{177147}$	$\frac{1}{2}$	0
0011	34	$\frac{11519}{177147}$	$\frac{1009}{2916}$	$\frac{70}{729}$
0100	2	$\frac{105785}{1417176}$	$\frac{1}{2}$	0
0101	24	$\frac{52489}{1417176}$	$\frac{97}{324}$	$\frac{4}{81}$
0110	23	$\frac{6875}{177147}$	$\frac{95}{324}$	$\frac{7}{162}$
0111	234	$\frac{8515}{177147}$	$\frac{4285}{17496}$	$\frac{56}{2187}$
1000	1	$\frac{13834}{177147}$	$\frac{1}{2}$	0
1001	14	$\frac{7226}{177147}$	$\frac{283}{972}$	$\frac{10}{243}$
1010	13	$\frac{61777}{1417176}$	$\frac{139}{486}$	$\frac{35}{972}$
1011	134	$\frac{51137}{1417176}$	$\frac{6113}{26244}$	$\frac{140}{6561}$
1100	12	$\frac{13760}{177147}$	$\frac{293}{972}$	$\frac{25}{486}$
1101	124	$\frac{3088}{177147}$	$\frac{3749}{17496}$	$\frac{40}{2187}$
1110	123	$\frac{13445}{1417176}$	$\frac{1805}{8748}$	$\frac{35}{2187}$
1111	1234	$\frac{278965}{1417176}$	$\frac{278965}{1417176}$	$\frac{560}{59049}$

From the point of view of the original motivation a different scenario is of interest. Imagine that we have $K \in \mathcal{K}_T$ such that all the equalities in (C1) are satisfied, i.e. all the phylogenetic invariants hold. If one of the constraints in (C2)-(C5) does not hold then $K \notin \mathcal{M}_T^\kappa$. This shows that the method of phylogenetic invariants as commonly used can lead to spurious results. For example consider sample proportions and the corresponding tree cumulants as in Table 3. It can be checked that for this point all the equations in (C1) are satisfied. However, this point does not lie in the model space. Using the formulae in [40, Corollary 5.5], which gives the inverse map for the parametrization, it is simple to confirm that the point mapping to K satisfies $\theta^{(4)}_{1|1} = \frac{67}{54} > 1$. This cannot therefore be a probability and so $\theta \notin \Theta_T$.

6. Discussion

The new coordinate system proposed in [40] provides a better insight into the geometry of phylogenetic tree models with binary observations. The product form of the parametrization is useful and has already enabled us to obtain the full geometric description of the model class.

Of course it is one thing formally being able to identify the constraints in the model and quite another to use this understanding for model selection and estimation in realistically large scale problems. The results in this paper only formally allow us to determine explicitly the extremely complex nature of the feasible solution space of a given tree model and determine whether a proposed

estimate is feasible. So they simply represent the first stage in constructing methodology which supports these insights with an inferential technology that can address statistical issues in large tree. In particular, there remains the much more challenging issue of designing samplers that use our results explicitly to efficiently estimate and explore the tree model space. We are currently investigating this issue and hope to report such algorithms in a later paper.

One of the interesting implications of our results for phylogenetic analysis is that it enables us to consider different, simpler model classes containing the original one in such a way that the whole evolutionary interpretation in terms of the tree topologies remains valid. If we were interested only in the tree we could consider the model defined only by a subsets of constraints in Theorem 4.7 involving only covariances. The cost of this reduction is that the conditional independencies induced by the original model no longer hold, which, in turn, affects the interpretation of the model. We note that this approach is in a similar spirit to that employed to motivate the MAG model class introduced in [34].

Acknowledgments

Diane Maclagan and John Rhodes contributed substantially to this paper. We would also like to thank Bernd Sturmfels for a stimulating discussion at the early stage of our work and Lior Pachter for pointing out reference [8].

Appendix A: Change of coordinates

In this section we index raw probabilities with subsets of $[n]$ instead of $\{0,1\}^n$. We identify $I \subseteq [n]$ with $\alpha \in \{0,1\}^n$ such that $\alpha_i = 1$ only if $i \in I$. We first change our coordinates from the raw probabilities $p = [p_I]_{I\subseteq[n]}$ to the non-central moments $\lambda = [\lambda_I]_{I\subseteq[n]}$, where $\lambda_I = \mathbb{E}(\prod_{i\in I} X_i)$. This is a linear map $f_{p\lambda} : \mathbb{R}^{2^n} \to \mathbb{R}^{2^n}$ with determinant equal to one, where the components λ_I of the vector $\lambda = f_{p\lambda}(p)$ are defined by

$$\lambda_I = \sum_{J\supseteq I} p_J \qquad \text{for any } I \subseteq [n]. \tag{17}$$

In particular $\lambda_\emptyset = 1$ for all probability distributions and the image $f_{p\lambda}(\Delta_{2^n-1})$ is contained in the hyperplane defined by $\lambda_\emptyset = 1$. Moreover, from (17), it follows that the λ's are just marginal probabilities. The linearity of the expectation implies that the central moments can be expressed in terms of non-central moments. Define $\mu_I = \mathbb{E}(\prod_{i\in I} U_i)$, where $U_i = X_i - \mathbb{E}X_i$. Then

$$\mu_I = \sum_{J\subseteq[n]} (-1)^{|J|}\lambda_{I\setminus J}\prod_{i\in J}\lambda_i \quad \text{for } I \subseteq [n]. \tag{18}$$

Using these equations we can transform coordinates from the non-central moments $\lambda = [\lambda_I]$ to another set of variables given by all the means $\lambda_1, \ldots, \lambda_n$ and

central moments $[\mu_I]$ for $I \subseteq [n]$. The polynomial map $f_{\lambda\mu} : \mathbb{R}^{2^n} \to \mathbb{R}^n \times \mathbb{R}^{2^n}$ is an identity on the first n coordinates corresponding to the means $\lambda_1, \ldots, \lambda_n$ and is defined on the remaining coordinates using the equations (18). Let $\mathcal{C}_n = (f_{\lambda\mu} \circ f_{p\lambda})(\Delta_{2^n-1})$. This is contained in a subspace of $\mathbb{R}^n \times \mathbb{R}^{2^n}$ given by

$$\mu_\emptyset = 1 \quad \text{and} \quad \mu_1 = \cdots = \mu_n = 0.$$

Since $f_{\lambda\mu}$ is invertible (see [40, Appendix A.1]) it provides a change of coordinates from the non-central moments to a coordinate system on $\mathcal{C}_n$ given by $\lambda_1, \ldots, \lambda_n$ together with μ_I for all $I \subseteq [n]$ such that $|I| \geq 2$. Note that the Jacobian of $f_{\lambda\mu} \circ f_{p\lambda} : \Delta_{2^n-1} \to \mathcal{C}_n$ is constant and equal to one.

The final change of coordinates requires some combinatorics.

Definition A.1. Let $T = (V, E)$ be a tree with n leaves. An *edge split* is a partition of $[n]$ into two non-empty sets induced by removing an edge $e \in E$ and restricting $[n]$ to the connected components of the resulting graph. By an *edge partition* we mean any partition $B_1|\cdots|B_k$ of the set of leaves of T induced by removing a subset of E. Each B_i is called a *block* of the partition.

Let Π_T denote the partially ordered set (poset) of all tree partitions of the set of leaves. The ordering in this poset is induced from the ordering in the lattice Π_n of all partitions of $[n]$ (see [35, Example 3.1.1.d]). Thus for $\pi = B_1|\cdots|B_r$ and $\nu = B'_1|\cdots|B'_s$ we have $\pi \leq \nu$ if every block of π is contained in one of the blocks of ν. The poset Π_T has a unique minimal element $1|2|\cdots|n$ induced by removing all edges in E and the maximal one with no edges removed which is equal to a single block $[n]$. The maximal element is denoted by $\hat{1}$ and the minimal one is denoted by $\hat{0}$.

For any poset Π a *Möbius function* $\mathfrak{m}_\Pi : \Pi \times \Pi \to \mathbb{R}$ can be defined in such a way that $\mathfrak{m}_\Pi(\pi, \pi) = 1$ for every $\pi \in \Pi$, $\mathfrak{m}_\Pi(\nu, \pi) = -\sum_{\nu \leq \delta < \pi} \mathfrak{m}_\Pi(\nu, \delta)$ for $\nu < \delta$ in Π and is zero otherwise (c.f. [35, Section 3.7]). Let $T(W)$, for $W \subset V$, denote the minimal subtree of T containing W in its set of vertices. Then $\Pi_{T(W)}$ is the poset of all multisplits of the set of leaves of $T(W)$ induced by edges of $T(W)$. The Möbius function on $\Pi_{T(W)}$ will be denoted by $\mathfrak{m}_W$ and the Möbius function on Π_T will be denoted by $\mathfrak{m}$. Let $\hat{0}_W$ and $\hat{1}_W$ denote the minimal and the maximal element of $\Pi_{T(W)}$ respectively.

Consider a map $f_{\mu\kappa} : \mathbb{R}^n \times \mathbb{R}^{2^n} \to \mathbb{R}^n \times \mathbb{R}^{2^n}$ where the coordinates in the domain are denoted by $\lambda_1, \ldots, \lambda_n$ and μ_I for $I \subseteq [n]$ and let the coordinates of the image space be denoted by $\lambda_1, \ldots, \lambda_n$ and κ_I for $I \subseteq [n]$. The map is defined as the identity on the first n coordinates corresponding to $\lambda_1, \ldots, \lambda_n$ and

$$\kappa_I = \sum_{\pi \in \Pi_{T(I)}} \mathfrak{m}_I(\pi, \hat{1}_I) \prod_{B \in \pi} \mu_B \quad \text{for all } I \subseteq [n], \tag{19}$$

where by convention $\kappa_\emptyset = \mu_\emptyset$. Let $\mathcal{K}_T = f_{\mu\kappa}(\mathcal{C}_n)$. Note that for any $I \subseteq [n]$ such that $|I| \leq 3$, $\kappa_I = \mu_I$. In particular $\mathcal{K}_T$ is contained in the subspace of $\mathbb{R}^n \times \mathbb{R}^{2^n}$ given by

$$\kappa_\emptyset = 1, \quad \kappa_1 = \cdots = \kappa_n = 0$$

The map $f_{\mu\kappa} : \mathcal{C}_n \to \mathcal{K}_T$ is a polynomial isomorphism with a polynomial inverse $f_{\kappa\mu}$. It therefore gives a change of coordinates to a coordinate system on $\mathcal{K}_T$ given by $\lambda_1, \ldots, \lambda_n$ and κ_I for $|I| \geq 2$. The exact form of the inverse map is given by the Möbius inversion formula (c.f. [40, Section 3.2])

$$\mu_I = \sum_{\pi \in \Pi_{T(I)}} \prod_{B \in \pi} \kappa_B \quad \text{for all } I \subseteq [n], |I| \geq 2. \tag{20}$$

Note that after restriction to Δ_{2^n-1}, $f_{p\lambda}(\Delta_{2^n-1})$ and $\mathcal{C}_n$ respectively all $f_{p\lambda}$, $f_{\lambda\mu}$ and $f_{\mu\kappa}$ are polynomial maps with polynomial inverses (c.f. [40, Appendix A]). This therefore implies that there is a polynomial isomorphism between Δ_{2^n-1} and $\mathcal{K}_T$.

Appendix B: Proofs

Proof of Proposition 2.5. By Remark 2.1 $\mathcal{M}_3$ does not depend on the rooting. Therefore, we can assume that T is rooted in h. In this case Proposition 2.3 implies that $\mathcal{M}_3^\kappa$ is given by $\lambda_i = \frac{1}{2}(1 - \bar{\mu}_i)$ for $i = 1, 2, 3$ and

$$\begin{aligned} \mu_{ij} &= \frac{1}{4}(1 - \bar{\mu}_h^2)\eta_{h,i}\eta_{h,j} \text{ for all } i \neq j \in \{1, 2, 3\} \text{ and} \\ \mu_{123} &= \frac{1}{4}(1 - \bar{\mu}_h^2)\bar{\mu}_h\eta_{h,1}\eta_{h,2}\eta_{h,3}, \end{aligned} \tag{21}$$

subject to constraints in (5).

Denote the subset of $\mathcal{K}_T$ given by constraints (i),(ii) by $\mathcal{M}$. We need to show that $\mathcal{M} = \mathcal{M}_3^\kappa$. First, we prove that $\mathcal{M}_3^\kappa \subseteq \mathcal{M}$. Let $K = \psi_T(\omega)$ for some $\omega \in \Omega_T$ with coordinates given by $\bar{\mu}_h$ and $\bar{\mu}_i$, $\eta_{h,i}$ for $i = 1, 2, 3$. We consider two cases. Either $(1 - \bar{\mu}_h^2)\eta_{h,1}\eta_{h,2}\eta_{h,3}$ is zero or not. In the first case $\mu_{123} = 0$ and at least two covariances vanish and hence (i) holds.

Now we show that if $(1 - \bar{\mu}_h^2)\eta_{h,1}\eta_{h,2}\eta_{h,3} \neq 0$ then (ii) holds. From (21)

$$\mu_{12}\mu_{13}\mu_{23} \quad = \quad \left(\frac{1}{4}(1 - \bar{\mu}_h^2)\right)^3 (\eta_{h,1}\eta_{h,2}\eta_{h,3})^2 \quad > \quad 0. \tag{22}$$

To show that K satisfies (9) we can simply substitute for the corresponding moments using (21). After trivial reductions we then obtain that

$$|\eta_{h,i}| \pm \bar{\mu}_h\eta_{h,i} \leq (1 \pm \bar{\mu}_i),$$

which is equivalent to (5). Therefore, since by hypothesis (5) holds, we also have that $\mathcal{M}_3^\kappa \subseteq \mathcal{M}$.

To show $\mathcal{M} \subseteq \mathcal{M}_3^\kappa$ we prove that for $K \in \mathcal{M}$ a parameter ω in (21) exists which satisfies the constraints defining Ω_T and $K = \psi_T(\omega)$. Let P be the probability distribution corresponding to K. First, consider the points satisfying (i). If all three covariances vanish for this point then taking $\eta_{h,1} = \eta_{h,2} = \eta_{h,3} = 0$ and $\bar{\mu}_h^2 = 1$ we obtain a valid choice of parameters in (21) and their values

satisfy (5). When one covariance is non-zero, say $\mu_{12} \neq 0$, then, if a choice of parameters exists it must 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 [20, Corollary 2] 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.

Consider now case (ii). Since $\mu_{12}\mu_{13}\mu_{23} > 0$ then in particular $\operatorname{Det} P > 0$. Set $\bar{\mu}_h^2 = \frac{\mu_{123}^2}{\operatorname{Det} P}$ and $\eta_{h,i}^2 = \frac{\operatorname{Det} P}{\mu_{jk}^2}$ for $i = 1, 2, 3$. It follows that $(\frac{1}{4}(1 - \bar{\mu}_h^2))^2\eta_{h,i}^2\eta_{h,j}^2 = \mu_{ij}^2$ for $i, j = 1, 2, 3$ and $(\frac{1}{4}(1 - \bar{\mu}_h^2))^2\bar{\mu}_h^2\eta_{h,1}^2\eta_{h,2}^2\eta_{h,3}^2 = \mu_{123}^2$. This coincides with (21) modulo the sign. It can be easily shown that $\mu_{12}\mu_{13}\mu_{23} > 0$ implies that there exist a choice of signs for $\eta_{h,i}$ for $i = 1, 2, 3$ such that

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

for all $1 \leq i < j \leq 3$ as in (21). For example set $\operatorname{sgn}(\eta_{h,i}) = \operatorname{sgn}(\mu_{jk})$ and use the fact that, by our assumption, $\operatorname{sgn}(\mu_{ij}) = \operatorname{sgn}(\mu_{ik})\operatorname{sgn}(\mu_{jk})$. This choice of signs already determines the sign of $\bar{\mu}_h$ so that

$$\frac{1}{4}(1 - \bar{\mu}_h^2)\bar{\mu}_h\eta_{h,1}\eta_{h,2}\eta_{h,3} = \mu_{123}$$

holds.

It remains to show that parameters set in this way satisfy the constraints defining Ω_T. First note that since $0 < 4\mu_{12}\mu_{13}\mu_{23} \leq \operatorname{Det} P$ then $\bar{\mu}_h^2 \in (0, 1)$ as required. From [40, Appendix D] we know that if $(\eta_{h,1}, \eta_{h,2}, \eta_{h,3}, \bar{\mu}_h)$ is one choice of parameters then there exists only one alternative choice and it is $(-\eta_{h,1}, -\eta_{h,2}, -\eta_{h,3}, -\bar{\mu}_h)$. For a fixed $i = 1, 2, 3$ it is easily checked that $(\eta_{h,i}, \bar{\mu}_h)$ satisfies (5) if and only if $(-\eta_{h,i}, -\bar{\mu}_h)$ does. Therefore, we can assume that $\eta_{h,i} = \frac{\sqrt{\operatorname{Det} P}}{|\mu_{jk}|} > 0$. In this case $\bar{\mu}_h = \operatorname{sgn}(\mu_{jk})\frac{\mu_{123}}{\sqrt{\operatorname{Det} P}}$. It follows that (5) is satisfied if and only if (9) holds. □

Proof of Lemma 4.4. First assume that the map $s_0 : E \to \{-1, 1\}$, given in the statement of the lemma, exists. This induces a map $s : V \times V \to \{-1, 1\}$ such that $s(k, l) = \prod_{(u,v)\in E(kl)} s_0(u, v)$. For any triple i, j, k there exists a unique inner node h which is the intersection of all three paths between i, j, k. By the above equation the choice of signs for all $(u, v) \in E$ gives $s(i, h), s(j, h)$ and $s(k, h)$. Since $s(i, j) = s(i, h)s(j, h)$ and the same for the two other pairs, we get that $s(i, j)s(i, k)s(j, k) = s^2(i, h)s^2(j, h)s^2(k, h) = 1$ and the result follows since by construction $\sigma(i, j) = s(i, j)$ for all $i, j \in [n]$.

Now we prove the converse implication. Whenever there is a path $E(uv)$ in T such that all its inner nodes have degree two then a sign assignment satisfying (15) exists if and only if there exists a sign assignment for the same tree but with $E(uv)$ contracted to a single edge (u, v). Hence we can assume that the degree of each inner node is at least three.

We use an inductive argument with respect to number of hidden nodes. First we will show that the theorem is true for trees with one inner node (star trees) denoted by h. In this case we will use induction with respect to number of leaves.

It can easily be checked directly that the theorem is true for the tripod tree. Assume it works for all star trees with $k \leq m-1$ leaves and let T be a star tree with m leaves. By assumption for any three leaves i,j,k: $\sigma(i,j)\sigma(i,k)\sigma(j,k)=1$. If we consider a subtree with $(1,h)$ deleted then by induction assumption we can find a consistent choice of signs for all remaining edges. A choice of a sign for $(1,h)$ consistent with (15) exists if for all $i \geq 2$ $\sigma(1,i) = s_0(1,h)s_0(i,h)$. This is true if either $\sigma(1,i)s_0(i,h)=1$ for all i or $\sigma(1,i)s_0(i,h)=-1$ for all i. Assume it is not true, i.e. there exist two leaves i,j such that $\sigma(1,i)s_0(i,h)=1$ and $\sigma(1,j)s_0(j,h)=-1$. Then in particular since $\sigma(i,j)=s_0(i,h)s_0(j,h)$ we would have that $\sigma(1,i)\sigma(1,j)\sigma(i,j)=-1$ which contradicts our assumption.

If the number of the inner nodes is greater than one then pick an inner node h adjacent to exactly one inner node. Let h' be the inner node adjacent to h and let I be a subset of leaves which are adjacent to h. Choose one $i \in I$ and consider a subtree T' obtained by removing all leaves in I and the incident edges apart from the node i and the edge (h,i). By the induction, since h has degree two in the resulting subtree, we can find signs for all edges of T'. Set $s_0(h,h')=1$ then $s_0(h,i)=s(h',i)$ which identifies $s_0(h,i)$. Similarly it can be showed that there exists a choice of signs for all remaining edges (i',h). The result follows since the choice of $i \in I$ was arbitrary. □

Appendix C: The proof of the main theorem

Let $K \in \mathcal{K}_T$ have coordinates given by λ_i for $i=1,\ldots,n$ and κ_I for $I \subseteq [n]$ such that $|I| \geq 2$. Let K^J, $J \subseteq [n]$, denote the projection onto the coordinates given by λ_i for $i \in J$ and κ_I, $I \subseteq J$, $|I| \geq 2$. Directly from the definition of $\mathcal{M}_T$ it follows that $K \in \mathcal{M}_T^\kappa$ if and only if $K^I \in \mathcal{M}_{T(I)}^\kappa$ for all $I \subseteq [n]$.

Let $\mathcal{M}$ denote the subset of $\mathcal{K}_T$ defined by constraints in (C1)-(C4). We need to show that $\mathcal{M} = \mathcal{M}_T^\kappa$. We divide the proof into series of lemmas.

Lemma C.1. *The inclusion $\mathcal{M}_T^\kappa \subseteq \mathcal{M}$ holds.*

Proof. Since the rooting is not relevant by Remark 2.1, we choose an arbitrary inner node as the root node. Let $K \in \mathcal{M}_T^\kappa$ and hence $K = \psi_T(\omega)$ for some $\omega \in \Omega_T$.

To show that the equations in (C1) hold let $A|B$ be an edge split and let $e=(w,w')$ be the edge inducing this split. By $T \setminus e$ we denote the graph obtained from T by removing the edge e. We assume that w lies in the same connected component of $T \setminus e$ as A and w' lies in the second component of $T \setminus e$. For every non-empty $I \subseteq A$ and $J \subseteq B$ from Proposition 2.3

$$\begin{aligned}\kappa_{IJ} \quad = \quad & \frac{1}{4}(1-\bar{\mu}_{r(IJ)}^2) \prod_{v \in \mathrm{int}(V(Iw'))} \bar{\mu}_v^{\deg(v)-2} \prod_{v \in \mathrm{int}(V(Jw))} \bar{\mu}_v^{\deg(v)-2} \\ & \cdot \eta_{w,w'} \prod_{(u,v)\in E(Iw)} \eta_{u,v} \prod_{(u,v)\in E(Jw')} \eta_{u,v}.\end{aligned}$$

From this it easily follows that for any non-empty $I_1, I_2 \subseteq A$ and $J_1, J_2 \subseteq B$,

$\kappa_{I_1J_1}\kappa_{I_2J_2} - \kappa_{I_1J_2}\kappa_{I_2J_1} = 0$ if and only if

$$(1-\mu^2_{r(I_1J_1)})(1-\mu^2_{r(I_2J_2)}) = (1-\mu^2_{r(I_1J_2)})(1-\mu^2_{r(I_2J_1)}). \tag{23}$$

To show that (23) is always true, we consider two cases: either $r(AB) \in V(Aw)$ or $r(AB) \in V(Bw')$. If $r(AB) \in V(Aw)$ then $r(I_1J_1) = r(I_1w)$, $r(I_1J_2) = r(I_1w)$, $r(I_2J_1) = r(I_2w)$ and $r(I_2J_2) = r(I_2w)$. Hence in this case (23) holds. The case $r(AB) \in V(Bw')$ follows by symmetry. Therefore the equations in (C1) always hold.

To show that K satisfies (C2) consider the projection K^{ijk} for each $i,j,k \in [n]$. By [40, Corollary 2.2] $\mathcal{M}^{\kappa}_{T(ijk)}$ is equal to the tripod tree model. Since $K^{ijk} \in \mathcal{M}^{\kappa}_{T(ijk)}$ then, by Proposition 2.5, (C2) must hold. To show that K satisfies (C3) let $i,j \in [n]$ be such that $\mu_{ij} = 0$. Let $I \subseteq [n]$ be such that $i,j \in I$ and assume that $\kappa_I(\omega) \neq 0$. Then by (7) in particular $\mu^2_{r(I)} \neq 1$ and $\eta_{u,v} \neq 0$ for all $(u,v) \in E(I)$. By [40, Remark 4.3] this implies in particular that $\bar{\mu}^2_{r(ij)} \neq 1$. From this, again by (7), it follows that $\mu_{ij} \neq 0$ and we get a contradiction. Hence if $\mu_{ij} = 0$ then $\kappa_I = 0$ for all I such that $i,j \in I$.

To show that K satisfies (C4) let $i,j,k,l \in [n]$ be the four leaves mentioned in the condition. Let u and v be two inner nodes such that u separates i from j, v separates k from l and $\{u,v\}$ separates $\{i,j\}$ from $\{k,l\}$. In other words u, v are the only inner nodes of degree three in $T(ijkl)$. By [40, Lemma 2.1], $T(ijkl)$ gives the same model as the quartet tree with four leaves i,j,k,l and two inner nodes u, v. Moreover, by Remark 2.1, $\mathcal{M}_{T(ijkl)}$ does not depend on the rooting so we can assume that the tree is rooted in u. Since $K^{ijkl} \in \mathcal{M}_{T(ijkl)}$ then for some parameter choices

$$\mu_{ik} = \frac{1}{4}(1-\bar{\mu}^2_u)\eta_{u,i}\eta_{u,v}\eta_{v,k}, \quad \mu_{jl} = \frac{1}{4}(1-\bar{\mu}^2_u)\eta_{u,j}\eta_{u,v}\eta_{v,l}$$

$$\mu_{ijk} = \frac{1}{4}(1-\bar{\mu}^2_u)\bar{\mu}_u\eta_{u,i}\eta_{u,j}\eta_{u,v}\eta_{v,k}, \quad \mu_{ikl} = \frac{1}{4}(1-\bar{\mu}^2_u)\bar{\mu}_v\eta_{u,i}\eta_{u,v}\eta_{v,k}\eta_{v,l}.$$

Substitute these equations into (C4). There are then two cases to consider: $\mu_{uv} \geq 0$, $\mu_{uv} < 0$. Laborious but elementary algebra shows that the condition in (C4) is equivalent to (5) applied to $(1-\bar{\mu}^2_u)\eta_{u,v}$ and hence (C4) holds by definition. Consequently $\mathcal{M}^{\kappa}_T \subseteq \mathcal{M}$. □

To show the opposite inclusion is a bit more complicated. We consider two separate cases. Let $K \in \mathcal{M}$. We construct a point $\omega_0 \in \mathbb{R}^{|V|+|E|}$ such that $\omega_0 \in \Omega_T$ and $\psi_T(\omega_0) = K$, i.e. ω_0 is such that, for all $I \subseteq [n]$ such that $|I| \geq 2$, κ_I can be written in terms of the parameters in ω_0 as in (7).

Lemma C.2. *Let K be such that $\mu_{ij} \neq 0$ for all $i,j \in [n]$. If $K \in \mathcal{M}$ then $K \in \mathcal{M}^{\kappa}_T$.*

Proof. We set squares of values of all the parameters in terms of the observed moments using [40, Corollary 5.5]. We will show that the equations in (7) must hold for their absolute values. We will then need to ensure there is at least one assignment of signs for a set of parameters such that all equations in (7) hold

exactly. Finally, we will show that the parameter vector ω_0 defined in this way lies in Ω_T.

For each inner node h of T let $i,j,k \in [n]$ be any three leaves separated by h in T. By (C2) we have that $\mu_{ij}\mu_{ik}\mu_{jk} > 0$ and hence also that $\mathrm{Det}P^{ijk} > 0$. Now set

$$(\bar{\mu}_h^0)^2 = \frac{\mu_{ijk}^2}{\mathrm{Det}P^{ijk}}. \tag{24}$$

We show that (C1), which K satisfies by assumption, implies that the value of $(\bar{\mu}_h^0)^2$ does not depend on the choice of i,j,k. It suffices to show that if k is replaced by another leaf k' such that i,j,k' are separated by h in T then $\frac{\mu_{ijk}^2}{\mathrm{Det}P^{ijk}} = \frac{\mu_{ijk'}^2}{\mathrm{Det}P^{ijk'}}$. Since h has degree three in T then there exists an edge $e \in E$ inducing a split $A|B$ such that $i,j \in A$ and $k,k' \in B$. From (C1) it follows that

$$\mu_{ik}\mu_{jk'} = \mu_{ik'}\mu_{jk}, \quad \mu_{ijk}\mu_{ik'} = \mu_{ijk'}\mu_{ik}, \quad \mu_{ijk}\mu_{jk'} = \mu_{ijk'}\mu_{jk} \tag{25}$$

and consequently

$$\mathrm{Det}P^{ijk}\mu_{ij}\mu_{ik'}\mu_{jk'} = \mathrm{Det}P^{ijk'}\mu_{ij}\mu_{ik}\mu_{jk} \tag{26}$$

which implies that

$$\frac{\mu_{ijk}^2}{\mathrm{Det}P^{ijk}} = \frac{\mu_{ijk}^2\mu_{ij}\mu_{ik'}\mu_{jk'}}{\mathrm{Det}P^{ijk}\mu_{ij}\mu_{ik'}\mu_{jk'}} = \frac{\mu_{ijk'}^2\mu_{ij}\mu_{ik}\mu_{jk}}{\mathrm{Det}P^{ijk'}\mu_{ij}\mu_{ik}\mu_{jk}} = \frac{\mu_{ijk'}^2}{\mathrm{Det}P^{ijk'}}$$

as required.

For terminal edges (v,i) of T such that $i \in [n]$, let $j,k \in [n]$ be any two leaves of T such that v separates i, j, k. Set

$$(\eta_{v,i}^0)^2 = \frac{\mathrm{Det}P^{ijk}}{\mu_{jk}^2}. \tag{27}$$

As in the previous case it is straightforward to check that, given (C1), this value does not depend on the choice of j,k. For example, if instead of k we have k' and v separates i,j,k' in T then there exists an edge split such that $\{i,j\}$ and $\{k,k'\}$ are in different blocks. By (25), we can show that

$$\frac{\mathrm{Det}P^{ijk}}{\mu_{jk}^2} = \frac{\mu_{ik}\mathrm{Det}P^{ijk}}{\mu_{ik'}\mu_{jk'}\mu_{jk}} = \frac{\mathrm{Det}P^{ijk'}}{\mu_{jk'}^2}.$$

For inner edges $(u,v) \in E$ let $i,j,k,l \in [n]$ be any four leaves such that u separates i from j, v separates k from l and $\{u,v\}$ separates $\{i,j\}$ from $\{k,l\}$. Set

$$(\eta_{u,v}^0)^2 = \frac{\mu_{il}^2}{\mu_{ij}^2}\frac{\mathrm{Det}P^{ijk}}{\mathrm{Det}\,P^{ikl}} \tag{28}$$

which is well-defined since μ_{ij}^2 and $\mathrm{Det}P^{ikl}$ are strictly positive. We now show that this value does not depend on the choice of i,j,k,l. By symmetry it suffices

to show that we obtain the same value if instead of l we took another leaf l' such that u, v are the only degree three nodes in $T(ijkl')$. Since v has degree three then there must exist an inner edge separating i, j, k from l, l'. From (C1) it follows that

$$\mu_{il'}\mu_{kl'}\mathrm{Det}P^{ikl} = \mu_{il}\mu_{kl}\mathrm{Det}P^{ikl'}, \quad \mu_{il}\mu_{kl'} = \mu_{il'}\mu_{kl}$$

and hence

$$\frac{\mu_{il}^2}{\mu_{ij}^2}\frac{\mathrm{Det}P^{ijk}}{\mathrm{Det}\,P^{ikl}} = \frac{\mu_{il'}\mu_{kl'}}{\mu_{il'}\mu_{kl'}}\frac{\mu_{il}^2}{\mu_{ij}^2}\frac{\mathrm{Det}P^{ijk}}{\mathrm{Det}\,P^{ikl}} = \frac{\mu_{il'}^2}{\mu_{ij}^2}\frac{\mathrm{Det}P^{ijk}}{\mathrm{Det}\,P^{ikl'}}$$

as required.

We now show that with the choice of parameters satisfying (24), (27) and (28) the modulus of equations in (7) hold. First consider the case $I = \{i, j\}$. Label the inner nodes of $E(ij)$ by $v_1, \ldots, v_k$ beginning from the node adjacent to i. For each $s = 1, \ldots, k$ let i_s denote a leaf such that v_s separates i, j, i_s in T. By Remark 2.1, we can choose any rooting. We assume that the root $r(ij)$ of this path is in v_1. We now proceed to check that

$$\begin{aligned}\mu_{ij}^2 &= \left(\frac{1}{4}(1-(\bar{\mu}^0_{r(ij)})^2)\right)^2 \prod_{(u,v)\in E(ij)}(\eta^0_{u,v})^2 \qquad (29)\\ &= \left(\frac{1}{4}(1-(\bar{\mu}^0_{r(ij)})^2)\right)^2 (\eta^0_{v_1,u})^2\left(\prod_{s=2}^{k}(\eta^0_{v_{s-1},v_s})^2\right)(\eta^0_{v_k,v})^2.\end{aligned}$$

Since v_1 separates i, j, i_1 by construction, from (24) we therefore have

$$\frac{1}{4}(1-(\bar{\mu}^0_{v_1})^2) = \frac{\mu_{ij}\mu_{ii_1}\mu_{ji_1}}{\mathrm{Det}(P^{iji_1})}.$$

Now substitute this equation and all the set values in (27), (28) into the right hand side of (29). Use the fact that v_k separates i, j, i_k in T and i_{s-1}, i_s are the only degree three nodes in $T(ii_{s-1}ji_s)$. Since (v_1, i) and (v_k, j) are the only terminal edges we obtain

$$\left(\frac{\mu_{ij}\mu_{ii_1}\mu_{ji_1}}{\mathrm{Det}(P^{iji_1})}\right)^2 \cdot \frac{\mathrm{Det}P^{iji_1}}{\mu_{ji_1}^2} \cdot \left(\prod_{s=2}^{k}\frac{\mu_{ii_s}^2}{\mu_{ii_{s-1}}^2}\frac{\mathrm{Det}P^{iji_{s-1}}}{\mathrm{Det}\,P^{iji_s}}\right) \cdot \frac{\mathrm{Det}P^{iji_k}}{\mu_{ji_k}^2} \qquad (30)$$

It can now be checked that all the expressions with hyperdeterminants cancel out and the formula reduces to μ_{ij}^2 as required.

Now we need to show that for every $I = \{i, j, k\}$

$$\mu_{ijk}^2 = \left(\frac{1}{4}(1-\bar{\mu}^0_{r(ijk)})^2)\right)^2 (\bar{\mu}^0_w)^2 \prod_{(u,v)\in E(ijk)}(\eta^0_{u,v})^2, \qquad (31)$$

where by w we denote the node separating i, j and k. Assume that $T(ijk)$ is rooted somewhere on the path between i and j. Using (29) the right hand side of (31) can be rewritten as

$$\mu_{ij}^2(\bar{\mu}_w^0)^2 \prod_{(u,v)\in E(wk)} (\eta_{u,v}^0)^2. \tag{32}$$

Number the degree three nodes in $E(wk)$ by $v_1,\dots,v_l$ and let i_s denote a leaf such that the inner nodes of $T(ijki_s)$ of degree three are exactly v_{s-1} and v_s, where $v_0 = w$. By an exactly analogous argument as in the case above we obtain

$$\begin{aligned}&\prod_{(u,v)\in E(wk)} (\eta_{u,v}^0)^2 \\ &\quad= \frac{\mu_{ii_1}^2}{\mu_{ij}^2}\frac{\mathrm{Det}P^{ijk}}{\mathrm{Det}P^{iki_1}} \cdot \left(\prod_{s=2}^{l} \frac{\mu_{i_{s-1}i_s}^2}{\mu_{i_{s-2}i_{s-1}}^2}\frac{\mathrm{Det}P^{i_{s-2}i_{s-1}k}}{\mathrm{Det}P^{i_{s-1}i_sk}}\right)\frac{\mathrm{Det}P^{i_{l-1}i_lk}}{\mu_{i_{l-1}i_l}^2},\end{aligned} \tag{33}$$

where $i_0 = i$. It can be easily checked that all the hyperdeterminants apart from the term $\mathrm{Det}P^{ijk}$ cancel out. Moreover, all the covariances apart from the term μ_{ij}^{-2} cancel out as well. Hence (33) is equal to $\frac{\mathrm{Det}P^{ijk}}{\mu_{ij}^2}$. Now, by using the definition of $(\bar{\mu}_w^0)^2$ in (24), it can be easily checked that (32) is equal to μ_{ijk}^2 as required.

So far we have confirmed only that the squares of parameters in ω_0 satisfy required equations at least for the tree cumulants up to the third order. Next, we show that there exists a consistent choice of signs for these parameters such that the equations are satisfied exactly. Let $\sigma(i,j) = \mathrm{sgn}(\mu_{ij})$. Since by assumption $\mu_{ij} \neq 0$ for all $i,j \in [n]$ then the conditions in (C2) imply that $\sigma(i,j)\sigma(i,k)\sigma(j,k) = 1$ for all triples $i,j,k \in [n]$. Hence by Lemma 4.4 there exists a choice $s_0(u,v) \in \{-1,+1\}$ for all $(u,v) \in E$ such that $\sigma(i,j) = \prod_{(u,v)\in E(ij)} s_0(u,v)$ for all $i,j \in [n]$. For any two nodes $k,l \in V$ we define $s(k,l) = \prod_{(u,v)\in E(kl)} s_0(u,v)$. A choice of signs for the parameters can be obtained as follows: For each edge $(u,v) \in E$ we set $\mathrm{sgn}(\eta_{u,v}^0) = s_0(u,v)$ and, for each inner node v, set $\mathrm{sgn}(\bar{\mu}_v^0) = \mathrm{sgn}(\mu_{ijk})s(v,i)s(v,j)s(v,k)$ where i,j,k are any three leaves of T separated by v.

Assume now that the choice of the signs of the parameters, induced by $s_0(u,v)$ for $(u,v) \in E$, has been made. This choice of signs gives

$$\bar{\mu}_v^0 = s(v,i)s(v,j)s(v,k)\frac{\mu_{ijk}}{\sqrt{\mathrm{Det}P^{ijk}}}, \tag{34}$$

$$\eta_{v,i}^0 = s(v,i)\frac{\sqrt{\mathrm{Det}P^{ijk}}}{|\mu_{jk}|}, \tag{35}$$

$$\eta_{u,v}^0 = s_0(u,v)\left|\frac{\mu_{il}}{\mu_{ij}}\right|\sqrt{\frac{\mathrm{Det}P^{ijk}}{\mathrm{Det}\,P^{ikl}}}. \tag{36}$$

Note that, in particular, with this choice of signs $\operatorname{sgn}(\eta^0_{u,v}) = s_0(u,v)$ for all $(u,v) \in E$ and $\operatorname{sgn}(\bar\mu^0_v) = \operatorname{sgn}(\mu_{ijk}) \prod_{(u,v)\in E(ijk)} s_0(u,v)$. Since (29) holds, it follows that

$$|\mu_{ij}| = \frac{1}{4}(1-(\bar\mu^0_{r(ij)})^2) \prod_{(u,v)\in E(ij)} |\eta^0_{u,v}|.$$

Now multiply both sides by $s(i,j) = \prod_{(u,v)\in E(ij)} s_0(u,v)$ to get

$$\begin{aligned}\mu_{ij} = s(i,j)|\mu_{ij}| &= \frac{1}{4}(1-(\bar\mu^0_{r(ij)})^2) \prod_{(u,v)\in E(ij)} s_0(u,v)|\eta^0_{u,v}| \\ &= \frac{1}{4}(1-(\bar\mu^0_{r(ij)})^2) \prod_{(u,v)\in E(ij)} \eta^0_{u,v}. \end{aligned} \tag{37}$$

Similarly, from (31), we have that

$$|\mu_{ijk}| = \frac{1}{4}(1-(\bar\mu^0_{r(ijk)})^2)|\bar\mu^0_w| \prod_{(u,v)\in E(ijk)} |\eta^0_{u,v}|.$$

Multiply both sides by $\operatorname{sgn}(\mu_{ijk})$ and use the fact that $(\prod_{(u,v)\in E(ijk)} s_0(u,v))^2 = 1$ to get

$$\begin{aligned}\mu_{ijk} &= \frac{1}{4}(1-(\bar\mu^0_{r(ijk)})^2) \left(|\bar\mu^0_w| \operatorname{sgn}(\mu_{ijk}) \prod_{(u,v)\in E(ijk)} s_0(u,v) \right) \\ &\quad \cdot \prod_{(u,v)\in E(ijk)} s_0(u,v)|\eta^0_{u,v}| \\ &= \frac{1}{4}(1-(\bar\mu^0_{r(ijk)})^2)\bar\mu^0_w \prod_{(u,v)\in E(ijk)} \eta^0_{u,v}\end{aligned}$$

as desired.

We now show (7) for $|I| \geq 4$ by induction. Let $(u,v) \in E$ be any edge splitting I into two subsets I_1 and I_2 such that $|I_1|, |I_2| \geq 2$ and u is the node closer to I_1. Let $i \in I_1$ and $j \in I_2$ then, by (C1),

$$\kappa_{I_1 I_2} = \frac{\kappa_{I_1 j}\kappa_{i I_2}}{\kappa_{ij}}.$$

By induction we can assume that $\kappa_{I_1 j}$, $\kappa_{i I_2}$ and κ_{ij} have form as in (7). Moreover,

$$\frac{\prod_{(u,v)\in E(iI_2)} \eta_{u,v} \prod_{(u,v)\in E(I_1 j)} \eta_{u,v}}{\prod_{(u,v)\in E(ij)} \eta_{u,v}} = \prod_{(u,v)\in E(I)} \eta_{u,v},$$

$$\prod_{h\in N(iI_2)} \bar\mu_h^{\deg h - 2} = \prod_{h\in N(vI_2)} \bar\mu_h^{\deg h - 2},$$

$$\prod_{h\in N(I_1 j)} \bar{\mu}_h^{\deg h-2} = \prod_{h\in N(I_1 u)} \bar{\mu}_h^{\deg h-2}.$$

Using this we can write

$$\kappa_{I_1 I_2} = \frac{1}{4}\frac{(1-\bar{\mu}^2_{r(iI_2)})(1-\bar{\mu}^2_{r(I_1 j)})}{(1-\bar{\mu}^2_{r(ij)})} \prod_{h\in N(I)} \bar{\mu}_h^{\deg h-2} \prod_{(u,v)\in E(I)} \eta_{u,v}. \tag{38}$$

The root of $T(I)$ is either in $T(I_1 u)$ or in $T(vI_2)$. In the first case $r(I_1 j) = r(I)$ and $r(iI_2) = r(ij)$. In the second case $r(I_1 j) = r(ij)$ and $r(iI_2) = r(I)$. Hence in both cases

$$\frac{(1-\bar{\mu}^2_{r(iI_2)})(1-\bar{\mu}^2_{r(I_1 j)})}{(1-\bar{\mu}^2_{r(ij)})} = (1-\bar{\mu}^2_{r(I)})$$

and (38) has the required form given by (20). It follows that $K = \psi_T(\omega_0)$.

It now remains to show that the parameters defined in (34), (35) and (36) define a parameter vector ω_0 which lies in Ω_T. Since, by (C2), $\mu^2_{ijk} \le \mathrm{Det}P^{ijk}$ for all $i,j,k \in [n]$ for all inner nodes h we have $\bar{\mu}^0_h \in [-1,1]$ as required. For a terminal edge (v,i) consider the marginal model induced by $T(ijk)$, where j,k are any two leaves such that v separates i,j,k in T. From Proposition 2.5 constraints (C2) and (C3) imply that $\eta_{v,i}$ is a valid parameter. To show that (36) satisfies (5) write

$$(1\pm\bar{\mu}^0_u)\eta^0_{u,v} = \left(1 \pm s(u,i)s(u,j)s(u,k)\frac{\mu_{ijk}}{\sqrt{\mathrm{Det}P^{ijk}}}\right) s(u,v)\left|\frac{\mu_{il}}{\mu_{ij}}\right| \sqrt{\frac{\mathrm{Det}P^{ijk}}{\mathrm{Det}P^{ikl}}}.$$

Now substitute this together with the expressions for $\bar{\mu}^0_u$ and $\bar{\mu}^0_v$, given by (34), into (5). First assume $s(u,v) = 1$. Then $s(u,k) = s(v,k)$, $s(v,i) = s(u,i)$ and (5) becomes

$$\left(\sqrt{\mathrm{Det}P^{ijk}} \pm s(u,i)\mu_{ijk}\right)\left|\frac{\mu_{il}}{\mu_{ij}}\right| \le \left(\sqrt{\mathrm{Det}P^{ikl}} \pm s(v,l)\mu_{ikl}\right).$$

By multiplying both sides by a positive expression $|\mu_{jl}|(\sqrt{\mathrm{Det}P^{ijk}} \mp s(u,i)\mu_{ijk})$ we obtain

$$4\mu^2_{ik}\mu^2_{jl} \le \left(\sqrt{\mathrm{Det}P^{ijk}} \pm s(u,l)\mu_{jl}\mu_{ijk}\right)\left(\sqrt{\mathrm{Det}P^{ikl}} \mp s(v,l)\mu_{ikl}\right).$$

However, $s(u,l) = s(v,l)$ hence this is satisfied by (C5). It is easily calculated that the case $s(u,v) = -1$ leads to the same constraint. This finishes the proof of Lemma C.2. □

Lemma C.3. *The inclusion $\mathcal{M} \subseteq \mathcal{M}^{\kappa}_T$ holds.*

Proof. Let $K \in \mathcal{M}$ be a tree cumulant and let $\Sigma = [\mu_{ij}] \in \mathbb{R}^{n\times n}$ be the matrix of all covariances between the leaves. We say that an edge $e \in E$ is *isolated relative to* K if $\mu_{ij} = 0$ for all $i,j \in [n]$ such that $e \in E(ij)$. By $\widehat{E} \subseteq E$ we denote the set of all edges of T which are isolated relative to K. By $\widehat{T} = (V, E\setminus\widehat{E})$ we

denote the forest obtained from T by removing edges in $\widehat{E}$ and we call it the K-forest. We define relations on $\widehat{E}$ and $E \setminus \widehat{E}$. For two edges e, e' with either $\{e, e'\} \subset \widehat{E}$ or $\{e, e'\} \subset E \setminus \widehat{E}$ write $e \sim e'$ if either $e = e'$ or e and e' are adjacent and all the edges that are incident with both e and e' are isolated relative to K. Let us now take the transitive closure of $\sim$ restricted to pairs of edges in $\widehat{E}$ to form an equivalence relation on $\widehat{E}$. This transitive closure is constructed as follows. Consider a graph with nodes representing elements of $\widehat{E}$ and put an edge between e, e' whenever $e \sim e'$. Then the equivalence classes correspond to connected components of this graph. Similarly, take the transitive closure of $\sim$ restricted to the pairs of edges in $E \setminus \widehat{E}$ to form an equivalence relation in $E \setminus \widehat{E}$. We will let $[\widehat{E}]$ and $[E \setminus \widehat{E}]$ denote the set of equivalence classes of $\widehat{E}$ and $E \setminus \widehat{E}$ respectively (for details see [40, Section 5]).

Again we show that there exists $\omega_0 \in \Omega_T$ such that $\psi_T(\omega_0) = K$. Set $\eta^0_{u,v} = 0$ for all $(u, v) \in \widehat{E}$ and $\bar{\mu}^0_v = 0$ for all inner nodes of T with degree zero in $\widehat{T}$. It then follows that $(1 \pm \bar{\mu}_u)\eta_{u,v} = 0$ satisfies (5) for all $(u, v) \in \widehat{E}$ and $\bar{\mu}^0_v \in [-1, 1]$ for all $v \in \widehat{V}$ and hence these parameters satisfy constraints defining Ω_T. If $I \subseteq [n]$ is such that $E(I) \cap \widehat{E} \neq \emptyset$ then $\kappa_I = 0$ by (C3). Hence in this case we can assert that

$$\kappa_I = \frac{1}{4}(1 - (\bar{\mu}^0_{r(I)})^2) \prod_{v \in N(I)} (\bar{\mu}^0_v)^{\deg(v)-2} \prod_{(u,v) \in E(I)} \eta^0_{u,v}$$

simply because both sides of this equation are zero. By [40, Remark 5.2 (iv)] every connected component of $\widehat{T}$ is a subtree which is either an inner node or a tree with the set of leaves contained in $[n]$. Denote the connected subtrees which are not inner nodes by $T_1, \ldots, T_k$ and their sets of leaves by $[n_l]$ for $l = 1, \ldots, k$. For every $l = 1, \ldots, k$ and all $i, j \in [n_l]$ we have that $\mu_{ij} \neq 0$. Hence for each T_l applying Lemma C.2 we have $K^{[n_l]} \in \mathcal{M}_{T_l}$. If $I \subseteq [n]$ is such that $E(I) \cap \widehat{E} = \emptyset$ then $I \subseteq [n_l]$ for some $l = 1, \ldots, k$. Since $K^{[n_l]} \in \mathcal{M}_{T_l}$ then there exists a choice of parameters such that κ_I can be written as (7). Therefore $K \in \mathcal{M}_T$ and we are done. □

The proof that $\mathcal{M} = \mathcal{M}^{\kappa}_T$ follows from Lemma C.1 and Lemma C.3. It suffices to show that, given that all covariances are non-zero, the only constraints of $\mathcal{M}$ involving only second order moments are (16). In the formulation of the main result the only such constraints are all the equations in (C1) involving only covariances and the positivity constraints in (C2). By the four-point condition (c.f. (14)) the inequalities

$$\min\left\{\left(\frac{\mu_{ik}\mu_{jl}}{\mu_{ij}\mu_{kl}}\right)^2, \left(\frac{\mu_{il}\mu_{jk}}{\mu_{ij}\mu_{kl}}\right)^2\right\} \leq 1$$

for all not necessarily distinct $i, j, k, l \in [n]$ uniquely define the underlying tree metric and hence they are equivalent to all the equations in (C1) involving only second order moments. The inequalities

$$\min\left\{\frac{\mu_{ik}\mu_{jl}}{\mu_{ij}\mu_{kl}}, \frac{\mu_{il}\mu_{jk}}{\mu_{ij}\mu_{kl}}\right\} \geq 0$$

are equivalent to $\mu_{ij}\mu_{ik}\mu_{jk} \geq 0$ for all $i,j,k \in [n]$. However, the two above sets of inequalities are exactly equivalent to (16). □

Appendix D: Phylogenetic invariants

In a seminal paper Allman and Rhodes [2] identified equations defining the general Markov $\mathcal{M}_T$ in the case when T is a trivalent tree. In this section we relate their results to ours. To introduce their main theorem we need the following definition.

Definition D.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 split of $[n]$. Then the *flattening* of P induced by the split 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 edge partitions the induced flattening is called an *edge flattening* and we denote it by P_e, where $e \in E$ is the edge inducing the split.

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 $0\cdots00 > 0\cdots01 > \ldots > 1\cdots11$. This gives in particular the ordering of rows and columns of flattenings.

Theorem D.2 (Allman, Rhodes [2])**.** *Let T^r be a trivalent tree rooted in r and $\mathcal{M}_T$ be the general Markov model on T^r as defined by (2). Then the smallest algebraic variety, i.e. a subset of a real space defined by a finite set of polynomial equations, containing the general Markov model, is defined by vanishing of all 3×3-minors of all the edge flattenings of T^r together with the trivial polynomial equation $\sum_\alpha p_\alpha = 1$.*

Note that the result includes the case of 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 non-trivial polynomials vanishing on the model.

Just as we defined edge flattenings of probability tables we can also define edge flattenings of $(\kappa_I)_{I\subseteq[n]}$ where $\kappa_\emptyset = 1$ and $\kappa_i = 0$ for all $i \in [n]$ (c.f. Appendix A). 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}. Let N_e denote 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 D.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 the equations in Theorem D.2 in terms of our new coordinates.

Proposition D.3. *Let $T = (V, E)$ be a tree and let P be a probability distribution of a vector $X = (X_1, \dots, X_n)$ of binary variables represented by the leaves of T. If $e \in E$ is an edge of T inducing a split $A_1|A_2$ then* $\operatorname{rank}(P_e) = 2$ *if and only if* $\operatorname{rank}(N_e) = 1$.

Proof. Let $P_e = [p_{\alpha\beta}]$ be the matrix induced by a split $A_1|A_2$. We will show that $\operatorname{rank}(P_e) = \operatorname{rank}(D_e)$ where $D_e = [d_{IJ}]$ is a block diagonal matrix with 1 as the first 1×1 block (i.e. $d_{\emptyset\emptyset} = 1$, $d_{\emptyset J} = 0$, $d_{I\emptyset} = 0$ for all $I \subseteq A_1$, $J \subseteq A_2$) and the matrix N_e as the second block. It will then follow that $\operatorname{rank}(P_e) = 2$ if and only if $\operatorname{rank}(N_e) = 1$.

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 (17) and then to the flattening of the central moments $M_e = [\mu_{IJ}]$ such that $I \subseteq A_1$, $J \subseteq A_2$ using (18). It therefore suffices to show that $\operatorname{rank}(M_e) = \operatorname{rank}(D_e)$.

Let $I \subseteq A_1$, $J \subseteq A_2$. Then for each $\pi \in \Pi_{T(IJ)}$ there is at most one block containing elements from both I and J. For if this were not so then removing e would increase the number of blocks in π 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 $\pi \geq A_1|A_2$ in $\Pi_{T(IJ)}$ or both $I', J' \neq \emptyset$ otherwise. We can rewrite (20) as

$$\mu_{IJ} = \sum_{\pi \in \Pi_{T(IJ)}} \left(\kappa_{I'J'} \prod_{I \supseteq B \in \pi} \kappa_B \prod_{J \supseteq B \in \pi} \kappa_B \right). \tag{39}$$

We have $d_{I'J'} = \kappa_{I'J'}$ and it can be further rewritten as

$$\mu_{IJ} = \sum_{I' \subseteq I} \sum_{J' \subseteq J} u_{II'} d_{I'J'} v_{J'J}$$

where $u_{II'} = \sum_{\pi \in \Pi_T(I \setminus I')} \prod_{B \in \pi} \kappa_B$ and $v_{J'J} = \sum_{\pi \in \Pi_T(J \setminus J')} \prod_{B \in \pi} \kappa_B$. Setting $u_{II'} = 0$ for $I' \not\subseteq I$, $v_{J'J} = 0$ for $J' \not\subseteq J$ we can write these coefficients in terms of a lower triangular matrix U and an upper triangular matrix V. Since by construction $u_{II} = 1$ for all $I \subseteq A_1$ and $v_{JJ} = 1$ for all $J \subseteq A_2$ we have $\det U = \det V = 1$. Therefore, M_e has the same rank as D_e. □

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 which gives the equations in (C1) in Theorem (4.7).

Corollary D.4. *Let $T = (V, E)$ be a trivalent tree. Then the smallest algebraic variety containing $\mathcal{M}_T^\kappa$ is defined by the following set of equations. For each split $A|B$ induced by an edge consider any four (not necessarily disjoint) nonempty sets $I_1, I_2 \subseteq A$, $J_1, J_2 \subseteq B$ and the induced equation $\kappa_{I_1J_1}\kappa_{I_2J_2} - \kappa_{I_1J_2}\kappa_{I_2J_1} = 0$.*

In [16] Eriksson noted that some of the 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 the only ones which involve second order moments (c.f. Section 4). 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 polynomials in Theorem D.4 but expressed in terms of the central moments. 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, in this loose sense at least, the higher the order of the central moments (or equivalently the higher the order of the tree cumulants) the higher the variability of we might expect the invariant to exhibit (see [25, Section 4.5]).

References

[1] Allman, E. S. and Rhodes, J. A. (2007). Phylogenetic invariants. In *Reconstructing evolution* 108–146. Oxford Univ. Press, Oxford. MR2359351

[2] Allman, E. S. and Rhodes, J. A. (2008). Phylogenetic ideals and varieties for the general Markov model. *Adv. in Appl. Math.* **40** 127–148. MR2388607 (2008m:60145)

[3] Auvray, V., Geurts, P. and Wehenkel, L. (2006). A Semi-Algebraic Description of Discrete Naive Bayes Models with Two Hidden Classes. In *Proc. Ninth International Symposium on Artificial Intelligence and Mathematics*.

[4] Beerenwinkel, N., Eriksson, N. and Sturmfels, B. (2007). Conjunctive Bayesian networks. *Bernoulli* **13** 893–909. MR2364218 (2009c:62013)

[5] Bochnak, J., Coste, M. and Roy, M.-F. (1998). *Real Algebraic Geometry*. Springer. MR1659509

[6] Buneman, P. (1974). A note on the metric properties of trees. *J. Combinatorial Theory Ser. B* **17** 48–50. MR0363963 (51 ##218)

[7] Casanellas, M. and Fernández-Sánchez, J. (2007). Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees. *Molecular Biology and Evolution* **24** 288.

[8] Cavender, J. A. (1997). Letter to the editor. *Molecular Phylogenetics and Evolution* **8** 443–444.

[9] Cavender, J. A. and Felsenstein, J. (1987). Invariants of phylogenies in a simple case with discrete states. *Journal of Classification* **4** 57–71.

[10] Chang, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences* **137** 51–73. MR1410044

[11] Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics* **25** 573–578. MR0065087

[12] Chor, B., Hendy, M. D., Holland, B. R. and Penny, D. (2000). Multiple Maxima of Likelihood in Phylogenetic Trees: An Analytic Approach. *Molecular Biology and Evolution* **17** 1529–1541.
[13] Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology* **53** 1–13. MR2500683
[14] Drton, M. and Richardson, T. S. (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 287–309. MR2424754
[15] Drton, M. and Sullivant, S. (2007). Algebraic Statistical Models. *Statistica Sinica* **17** 1273–1297. MR2398596
[16] Eriksson, N. (2007). *Using invariants for phylogenetic tree construction. The IMA Volumes in Mathematics and its Applications* **149** 89-108. Springer. MR2500465
[17] Eriksson, N., Ranestad, K., Sturmfels, B. and Sullivant, S. (2005). Phylogenetic algebraic geometry. In *Projective varieties with unexpected properties* 237–255. Walter de Gruyter GmbH & Co. KG, Berlin. MR2202256 (2006k:14119)
[18] Garcia, L. D., Stillman, M. and Sturmfels, B. (2005). Algebraic geometry of Bayesian networks. *J. Symbolic Comput* **39** 331–355. MR2168286
[19] Gelfand, I. M., Kapranov, M. M. and Zelevinsky, A. V. (1994). *Discriminants, Resultants, and Multidimensional Determinants*. Birkhäuser. MR1264417
[20] Gilula, Z. (1979). Singular value decomposition of probability matrices: Probabilistic aspects of latent dichotomous variables. *Biometrika* **66** 339–344. MR0548203
[21] Lake, J. A. (1987). A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution* **4** 167.
[22] Lauritzen, S. L. (1996). *Graphical models. Oxford Statistical Science Series* **17**. The Clarendon Press Oxford University Press, New York. Oxford Science Publications. MR1419991 (98g:62001)
[23] Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Houghton, Mifflin, New York.
[24] Matsen, F. A. (2009). Fourier Transform Inequalities for Phylogenetic Trees. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* **6** 89-95.
[25] McCullagh, P. (1987). *Tensor methods in statistics. Monographs on Statistics and Applied Probability*. Chapman & Hall, London. MR907286 (88k:62004)
[26] Pearl, J. (1986). Fusion, propagation, and structuring in belief networks* 1. *Artificial intelligence* **29** 241–288. MR0858200
[27] Pearl, J. and Tarsi, M. (1986). Structuring causal trees. *J. Complexity* **2** 60–77. Complexity of approximately solved problems (Morningside Heights, N.Y., 1985). MR925434 (89g:68056)
[28] Rusakov, D. and Geiger, D. (2005). Asymptotic model selection

for naive Bayesian networks. *J. Mach. Learn. Res.* **6** 1–35 (electronic). MR2249813

[29] Semple, C. and Steel, M. (2003). *Phylogenetics. Oxford Lecture Series in Mathematics and its Applications* **24**. Oxford University Press, Oxford. MR2060009 (2005g:92024)

[30] Settimi, R. and Smith, J. Q. (1998). On the Geometry of Bayesian Graphical Models with Hidden Variables. In *UAI* (G. F. Cooper and S. Moral, eds.) 472-479. Morgan Kaufmann.

[31] Settimi, R. and Smith, J. Q. (2000). Geometry, moments and conditional independence trees with hidden variables. *Ann. Statist.* **28** 1179–1205. MR1811324 (2002b:62068)

[32] Smith, J. and Daneshkhah, A. (2010). On the robustness of Bayesian networks to learning from non-conjugate sampling. *International Journal of Approximate Reasoning* **51** 558–572. MR2644597

[33] Smith, J. Q. and Rigat, F. (2008). Isoseparation and Robustness in Finitre Parameter Bayesian Inference. *CRiSM Res Rep* 07–22.

[34] Spirtes, P., Richardson, T. and Meek, C. Heuristic greedy search algorithms for latent variable models In *Proceedings of AI & STAT'97* 481–488. Citeseer.

[35] Stanley, R. P. (2002). *Enumerative combinatorics. Volume I. Cambridge Studies in Advanced Mathematics* 49. Cambridge University Press. MR1442260

[36] Steel, M. and Faller, B. (2009). Markovian log-supermodularity, and its applications in phylogenetics. *Applied Mathematics Letters*. MR2523016

[37] Sturmfels, B. and Sullivant, S. (2005). Toric Ideals of Phylogenetic Invariants. *Journal of Computational Biology* **12** 204–228.

[38] Zwiernik, P. An asymptotic approximation of the marginal likelihood for general Markov models. arXiv:1012.0753. submitted.

[39] Zwiernik, P. (2010). L-cumulants, L-cumulant embeddings and algebraic statistics. arXiv:1011.1722.

[40] Zwiernik, P. and Smith, J. Q. (2010). Tree-cumulants and the geometry of binary tree models. *to appear in Bernoulli*.