Bernoulli **27**(1), 2021, 135–154
https://doi.org/10.3150/20-BEJ1231

Discrete statistical models with rational maximum likelihood estimator

ELIANA DUARTE[1], ORLANDO MARIGLIANO[2,*] and BERND STURMFELS[2,†]

[1]*Fakultät für Mathematik, Otto-von-Guericke Universität Magdeburg, 39106 Magdeburg, Germany.*
E-mail: eliana.duarte@ovgu.de
[2]*Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstraße 22, 04103 Leipzig, Germany.*
E-mail: [*] orlando.marigliano@mis.mpg.de; [†] bernd@mis.mpg.de

A discrete statistical model is a subset of a probability simplex. Its maximum likelihood estimator (MLE) is a retraction from that simplex onto the model. We characterize all models for which this retraction is a rational function. This is a contribution via real algebraic geometry which rests on results on Horn uniformization due to Huh and Kapranov. We present an algorithm for constructing models with rational MLE, and we demonstrate it on a range of instances. Our focus lies on models familiar to statisticians, like Bayesian networks, decomposable graphical models and staged trees.

Keywords: algebraic statistics; discrete statistical models; graphical models; likelihood geometry; maximum likelihood estimator; real algebraic geometry

1. Introduction

A *discrete statistical model* is a subset $\mathcal{M}$ of the open probability simplex Δ_n. Each point p in Δ_n is a probability distribution on the finite state space $\{0, 1, \ldots, n\}$, that is, $p = (p_0, p_1, \ldots, p_n)$, where the p_i are positive real numbers that satisfy $p_0 + p_1 + \cdots + p_n = 1$. The model $\mathcal{M}$ is the set of all distributions $p \in \Delta_n$ that are relevant for an application.

In data analysis, we are given an empirical distribution $u = (u_0, u_1, \ldots, u_n)$. This is the point in Δ_n whose ith coordinate u_i is the fraction of samples in state i. The *maximum likelihood estimator* (MLE) of $\mathcal{M}$ is a function $\Phi\colon \Delta_n \to \mathcal{M}$ that takes the empirical distribution u to a distribution $\hat{p} = (\hat{p}_0, \hat{p}_1, \ldots, \hat{p}_n)$ that best explains the given observations. Here "best" is understood in the sense of likelihood inference, so that $\hat{p} = \Phi(u)$ is the point in $\mathcal{M}$ that maximizes the *log-likelihood function* $p \mapsto \sum_{i=0}^{n} u_i \cdot \log(p_i)$. For any vector u in $\mathbb{R}^{n+1}_{>0}$, we set $\Phi(u) := \Phi(u/|u|)$ where $|u| = u_0 + \cdots + u_n$.

Likelihood inference is consistent. This means that $\Phi(u) = u$ for $u \in \mathcal{M}$. This follows from the fact that the log-likelihood function is strictly concave on Δ_n and its unique maximizer is $p = u$. Hence, the MLE Φ is a retraction from the simplex onto the model.

This point is fundamental for two fields at the crossroads of mathematics and data science. *Information Geometry* [1] views the MLE as the nearest point map of a Riemannian metric on Δ_n, given by the Kullback–Leibler divergence of probability distributions. *Algebraic Statistics* [4,17] is concerned with models $\mathcal{M}$ whose MLE Φ is an algebraic function of u. This happens when the constraints that define $\mathcal{M}$ are given in terms of polynomials in p. In this article, we address a question that is fundamental for both fields:

For which models $\mathcal{M}$ is the MLE Φ a rational function in the empirical distribution u?

The most basic example where the MLE is rational is the independence model for two binary random variables ($n = 3$). Here, $\mathcal{M}$ is a surface in the tetrahedron Δ_3. That surface is a familiar picture that serves as a point of entry for both Information Geometry and Algebraic Statistics. Points in $\mathcal{M}$ are

 © 2021 ISI/BS

positive rank one 2×2 matrices $\left[\begin{smallmatrix} p_0 & p_1 \\ p_2 & p_3 \end{smallmatrix}\right]$ whose entries sum to one. The data takes the form of a nonnegative integer 2×2 matrix u of counts of observed frequencies. Hence $|u| = u_0+u_1+u_2+u_3$ is the sample size, and $u/|u|$ is the empirical distribution. The MLE $\hat{p} = \Phi(u)$ is evaluated by multiplying the row and column sums of u:

$$\hat{p}_0 = \frac{(u_0+u_1)(u_0+u_2)}{|u|^2}, \qquad \hat{p}_1 = \frac{(u_0+u_1)(u_1+u_3)}{|u|^2},$$
$$\hat{p}_2 = \frac{(u_2+u_3)(u_0+u_2)}{|u|^2}, \qquad \hat{p}_3 = \frac{(u_2+u_3)(u_1+u_3)}{|u|^2}.$$

These four expressions are rational, homogeneous of degree zero and their sum is equal to 1. See [10], Example 2, for a discussion of these formulas from our present perspective.

The surface $\mathcal{M}$ belongs to the class of graphical models [14]. Fix an undirected graph G whose nodes represent random variables with finitely many states. The undirected graphical model $\mathcal{M}_G$ is a subset of Δ_n, where $n+1$ is the number of states in the joint distribution. The graphical model $\mathcal{M}_G$ is *decomposable* if and only if the graph G is chordal. Each coordinate $\hat{p}_i$ of its MLE is an alternating product of linear forms given by maximal cliques and minimal separators of G. A similar formula exists for directed graphical models, which are also known as Bayesian networks.

In both cases, the coordinates of the MLE are not only rational functions, but even alternating products of linear forms in $u = (u_0, u_1, \ldots, u_n)$. This is no coincidence. Huh [10] proved that if Φ is a rational function then each of its coordinates is an alternating product of linear forms, with numerator and denominator of the same degree. Huh further showed that this alternating product must take a very specific shape. That shape was discovered by Kapranov [12], who named it the *Horn uniformization*. The results by Kapranov and Huh are valid for arbitrary complex algebraic varieties. They make no reference to a context where the coordinates are real, positive and add up to 1.

The present paper makes the leap from complex varieties back to statistical models. Building on the remarkable constructions by Kapranov and Huh, we here work in the setting of real algebraic geometry that is required for statistical applications. Our main result (Theorem 1) characterizes all models $\mathcal{M}$ in Δ_n whose MLE is a rational function. It is stated in Section 2 and all its ingredients are presented in a self-contained manner.

In Section 3, we examine models with rational MLE that are familiar to statisticians, such as decomposable graphical models and Bayesian networks. Our focus lies on *staged tree models*, a far-reaching generalization of discrete Bayesian networks, described in the book by Collazo, Görgen and Smith [3]. We explain how our main result applies to these models. The proof of Theorem 1 is presented in Section 4. This is the technical heart of our paper, building on the likelihood geometry of [11], §3. We also discuss the connection to toric geometry and geometric modeling developed by Clarke and Cox [2]. In Section 5, we present our algorithm for constructing models with rational MLE, and we discuss its implementation and some experiments. The input is an integer matrix representing a toric variety, and the output is a list of models derived from that matrix. Our results suggest that only a very small fraction of Huh's varieties in [10] are statistical models.

2. How to be rational

Let $\mathcal{M}$ be a discrete statistical model in the open simplex Δ_n that has a well-defined maximum likelihood estimator $\Phi : \Delta_n \to \mathcal{M}$. We also write $\Phi : \mathbb{R}^{n+1}_{>0} \to \mathcal{M}$ for the induced map $u \mapsto \Phi(u/|u|)$ on positive vectors. If the $n+1$ coordinates of Φ are rational functions in u, then we say that $\mathcal{M}$ *has rational MLE*. The following is our main result.

Theorem 1. *The following are equivalent for the statistical model $\mathcal{M}$ with MLE Φ:*

(1) *The model $\mathcal{M}$ has* rational MLE.
(2) *There exists a* Horn pair (H, λ) *such that $\mathcal{M}$ is the image of the Horn map*

$$\varphi_{(H,\lambda)}: \quad \mathbb{R}^{n+1}_{>0} \to \mathbb{R}^{n+1}_{>0}.$$

(3) *There exists a* discriminantal triple $(A, \Delta, \mathbf{m})$ *such that $\mathcal{M}$ is the image under the monomial map $\phi_{(\Delta,\mathbf{m})}$ of precisely one orthant* (9) *of the dual toric variety Y_A^*.*

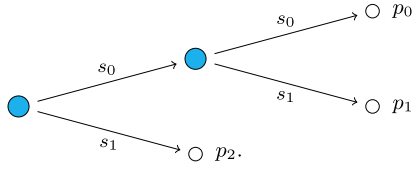
The MLE of the model satisfies the following relation on the open orthant $\mathbb{R}^{n+1}_{>0}$:

$$\Phi = \varphi_{(H,\lambda)} = \phi_{(\Delta,\mathbf{m})} \circ H. \tag{1}$$

This theorem matters for statistics because it reveals when a model has an MLE of the simplest possible closed form. Property (2) says that the polynomials appearing in the numerators and denominators of the rational formulas must factor into linear forms with positive coefficients. Property (3) offers a recipe, based on toric geometry, for explicitly constructing such models. The advance over [10] is that Theorem 1 deals with positive real numbers. It hence furnishes the definitive solution in the case of applied interest.

The goal of this section is to define all the terms seen in parts (2) and (3) of Theorem 1.

Example 2. We first discuss Theorem 1 for a simple experiment: *Flip a biased coin. If it shows heads, flip it again*. This is the model with $n = 2$ given by the tree diagram below. The model $\mathcal{M}$ is a curve in



the probability triangle Δ_2. The tree shows its parametrization

$$\Delta_1 \to \Delta_2, \qquad (s_0, s_1) \mapsto \left(s_0^2, s_0 s_1, s_1\right) \quad \text{where } s_0, s_1 > 0 \text{ and } s_0 + s_1 = 1.$$

The implicit representation of the curve $\mathcal{M}$ is the equation $p_0 p_2 - (p_0 + p_1) p_1 = 0$. Let (u_0, u_1, u_2) be the counts from repeated experiments. A total of $2u_0 + 2u_1 + u_2$ coin tosses were made. We estimate the parameters as the empirical frequency of heads, respectively, tails:

$$\hat{s}_0 = \frac{2u_0 + u_1}{2u_0 + 2u_1 + u_2} \quad \text{and} \quad \hat{s}_1 = \frac{u_1 + u_2}{2u_0 + 2u_1 + u_2}.$$

The MLE is the retraction from the triangle Δ_2 to the curve $\mathcal{M}$ given by the formula

$$\Phi(u_0, u_1, u_2) = \left(\hat{s}_0^2, \hat{s}_0 \hat{s}_1, \hat{s}_1\right) = \left(\frac{(2u_0 + u_1)^2}{(2u_0+2u_1+u_2)^2}, \frac{(2u_0+u_1)(u_1+u_2)}{(2u_0 + 2u_1 + u_2)^2}, \frac{u_1 + u_2}{2u_0+2u_1+u_2} \right).$$

Hence $\mathcal{M}$ has rational MLE. We see that the Horn pair from part (2) in Theorem 1 has

$$H = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 1 \\ -2 & -2 & -1 \end{pmatrix} \quad \text{and} \quad \lambda = (1, 1, -1).$$

We next exhibit the discriminantal triple $(A, \Delta, \mathbf{m})$ in part (3) of Theorem 1. The matrix $A = (1\ 1\ 1)$ gives a basis of the left kernel of H. The second entry is the polynomial

$$\Delta = x_3^2 - x_1^2 - x_1x_2 + x_2x_3 = (x_3 - x_1)(x_1 + x_2 + x_3). \tag{2}$$

The third entry marks the leading term $\mathbf{m} = x_3^2$. These data define the monomial map

$$\phi_{(\Delta,\mathbf{m})}: \quad (x_1, x_2, x_3) \mapsto \left(\frac{x_1^2}{x_3^2}, \frac{x_1x_2}{x_3^2}, -\frac{x_2}{x_3}\right).$$

The toric variety of the matrix A is the point $Y_A = \{(1:1:1)\}$ in $\mathbb{P}^2$. Our polynomial Δ vanishes on the line $Y_A^* = \{x_1 + x_2 + x_3 = 0\}$ that is dual to Y_A. The relevant orthant is the open line segment $Y_{A,\sigma}^* := \{(x_1 : x_2 : x_3) \in Y_A^* : x_1, x_2 > 0 \text{ and } x_3 < 0\}$. Part (3) in Theorem 1 says that $\mathcal{M}$ is the image of $Y_{A,\sigma}^*$ under $\phi_{(\Delta,\mathbf{m})}$. The MLE is $\Phi = \phi_{(\Delta,\mathbf{m})} \circ H$.

We now come to the definitions needed for Theorem 1. Let $H = (h_{ij})$ be an $m \times (n+1)$ integer matrix whose columns sum to zero, that is, $\sum_{i=1}^m h_{ij} = 0$ for $j = 0, \ldots, n$. We call such a matrix a *Horn matrix* and denote its columns by $h_0, h_1, \ldots, h_n$. The following alternating products of linear forms are rational functions of degree zero:

$$(Hu)^{h_j} := \prod_{i=1}^{m} (h_{i0}u_0 + h_{i1}u_1 + \cdots + h_{in}u_n)^{h_{ij}} \quad \text{for } j = 0, 1, \ldots, n.$$

We use the notation $v^h := \prod_i v_i^{h_i}$ for two vectors v, h of the same size. The Horn matrix H is *friendly* if there exists a real vector $\lambda = (\lambda_0, \ldots, \lambda_n)$ with $\lambda_i \neq 0$ for all i such that the following identity holds in the rational function field $\mathbb{R}(u_0, u_1, \ldots, u_n)$:

$$\lambda_0(Hu)^{h_0} + \lambda_1(Hu)^{h_1} + \cdots + \lambda_n(Hu)^{h_n} = 1. \tag{3}$$

If this holds, then we call (H, λ) a *friendly pair*, and we consider the rational function

$$\mathbb{R}^{n+1} \to \mathbb{R}^{n+1}, \qquad u \mapsto \left(\lambda_0(Hu)^{h_0}, \lambda_1(Hu)^{h_1}, \ldots, \lambda_n(Hu)^{h_n}\right). \tag{4}$$

The friendly pair (H, λ) is called a *Horn pair* if the function (4) is defined for all positive vectors, and it maps these to positive vectors. If these conditions hold, then we write $\varphi_{(H,\lambda)} : \mathbb{R}_{>0}^{n+1} \to \mathbb{R}_{>0}^{n+1}$ for the restriction of (4) to the positive orthant. We call $\varphi_{(H,\lambda)}$ the *Horn map* associated to the Horn pair (H, λ).

The difference between our Horn pairs and the more general pairs considered by Huh in [10] is the positivity condition we just introduced, along with the "friendliness" condition. These conditions guarantee that the image of the Horn map lies in the probability simplex, which is necessary for its interpretation as a statistical model. They also imply special properties for the Horn pair; see Propositions 22 and 23 in Section 4. The examples in Section 5 show that only a fraction of Huh's pairs (H, λ) are Horn pairs.

Different Horn pairs may give rise to the same Horn map. For example, the Horn pair

$$H' = \begin{pmatrix} 0 & 2 & 2 \\ 2 & 1 & 0 \\ 0 & -1 & -1 \\ -2 & -2 & -1 \end{pmatrix} \quad \text{and} \quad \lambda' = \left(1, -\frac{1}{4}, \frac{1}{4}\right)$$

also gives the map in Example 2. This is because the first and third rows of H' are collinear, causing the cancellation of linear factors in the Horn map. Following [2], a Horn pair (H, λ) is *minimal* if the matrix H has no zero rows and no pair of collinear rows.

Lemma 3. *Let (H', λ') be a Horn pair arising from the Horn pair (H, λ) by replacing two collinear rows r_k and r_ℓ in H such that $r_\ell = \mu r_k$ with their sum $r_k + r_\ell$ and setting*

$$\lambda'_j = \frac{\lambda_j \mu^{\mu \cdot h_{kj}}}{(1+\mu)^{(1+\mu)h_{kj}}} \quad \text{for all } j = 0, \ldots, n.$$

Then the Horn maps $\varphi_{(H',\lambda')}$ and $\varphi_{(H,\lambda)}$ are equal.

Proof. Let w_k and w_ℓ be the linear forms associated to the rows r_k and r_ℓ, respectively. Fix a column index j. We have $w_\ell = \mu w_k$ and $h_{\ell j} = \mu h_{kj}$. The factors of the jth coordinates of the Horn maps $\varphi_{(H,\lambda)}$ and $\varphi_{(H',\lambda')}$ that have changed after the operation are $\lambda_j w_k^{b_{kj}} w_\ell^{b_{\ell j}} = \lambda_j \mu^{\mu \cdot h_{kj}} w_k^{(1+\mu)h_{kj}}$ for (H, λ) and $\lambda'_j (w_k + w_\ell)^{(1+\mu)h_{kj}} = \lambda'_j (1+\mu)^{(1+\mu)h_{kj}} w_k^{(1+\mu)h_{kj}}$ for (H', λ'). Equating these two gives the desired formula. □

Every Horn map is represented by a unique minimal Horn pair. This follows by unique factorization; see also [2], Proposition 6.11. To make a Horn pair minimal, while retaining the Horn map, we can use Lemma 3 repeatedly, deleting zero rows as they appear.

Example 4. We illustrate the equivalence of (1) and (2) in Theorem 1 for the model described in [11], Example 3.11. Here, $n = 3$ and $m = 4$ and the Horn matrix equals

$$H = \begin{pmatrix} -1 & -1 & -2 & -2 \\ 1 & 0 & 3 & 2 \\ 1 & 3 & 0 & 2 \\ -1 & -2 & -1 & -2 \end{pmatrix}. \tag{5}$$

This Horn matrix is friendly because the following vector satisfies the identity (3):

$$\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_3) = \left(\frac{2}{3}, -\frac{4}{27}, -\frac{4}{27}, \frac{1}{27}\right). \tag{6}$$

The pair (H, λ) is a Horn pair, with associated Horn map

$$\varphi_{(H,\lambda)} : \quad \mathbb{R}^4_{>0} \to \mathbb{R}^4_{>0},$$

$$\begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \end{pmatrix} \mapsto \begin{pmatrix} \dfrac{2(u_0+3u_2+2u_3)(u_0+3u_1+2u_3)}{3(u_0+u_1+2u_2+2u_3)(u_0+2u_1+u_2+2u_3)} \\ \dfrac{4(u_0+3u_1+2u_3)^3}{27(u_0+u_1+2u_2+2u_3)(u_0+2u_1+u_2+2u_3)^2} \\ \dfrac{4(u_0+3u_2+2u_3)^3}{27(u_0+u_1+2u_2+2u_3)^2(u_0+2u_1+u_2+2u_3)} \\ \dfrac{(u_0+3u_2+2u_3)^2(u_0+3u_1+2u_3)^2}{27(u_0+u_1+2u_2+2u_3)^2(u_0+2u_1+u_2+2u_3)^2} \end{pmatrix}. \tag{7}$$

Indeed, this rational function takes positive vectors to positive vectors. The image of the map $\varphi_{(H,\lambda)}$ is a subset $\mathcal{M}$ of the tetrahedron $\Delta_3 = \{p \in \mathbb{R}^4_{>0} : p_0 + p_1 + p_2 + p_3 = 1\}$. We regard the subset $\mathcal{M}$ as a discrete statistical model on the state space $\{0, 1, 2, 3\}$. The model $\mathcal{M}$ is the curve of degree 4 inside Δ_3 defined by the two quadratic equations

$$9p_1p_2 - 8p_0p_3 = p_0^2 - 12p_3 = 0.$$

As in [11], Example 3.11, one verifies that $\mathcal{M}$ has rational MLE, namely $\Phi = \varphi_{(H,\lambda)}$.

We next define all the terms used in part (3) of Theorem 1. Fix a matrix $A = (a_{ij}) \in \mathbb{Z}^{r\times m}$ of rank r that has the vector $(1, \ldots, 1)$ in its row span. The connection to part (2) of Theorem 1 will be that the rows of A span the left kernel of H. We identify the columns of A with Laurent monomials in r unknowns $t_1, \ldots, t_r$. The associated monomial map is

$$\gamma_A : \quad (\mathbb{R}^*)^r \to \mathbb{RP}^{m-1}, \qquad (t_1, \ldots, t_r) \mapsto \left(\prod_{i=1}^r t_i^{a_{i1}} : \prod_{i=1}^r t_i^{a_{i2}} : \cdots : \prod_{i=1}^r t_i^{a_{im}}\right). \tag{8}$$

Here, $\mathbb{R}^* = \mathbb{R}\backslash\{0\}$ and $\mathbb{RP}^{m-1}$ denotes the real projective space of dimension $m-1$. Let Y_A be the closure of the image of γ_A. This is the projective toric variety given by A.

Every point $x = (x_1 : \cdots : x_m)$ in the dual projective space $(\mathbb{RP}^{m-1})^\vee$ corresponds to a hyperplane H_x in $\mathbb{RP}^{m-1}$. The *dual variety* Y_A^* to the toric variety Y_A is the closure of

$$\left\{x \in \left(\mathbb{RP}^{m-1}\right)^\vee \mid \gamma_A^{-1}(H_x \cap Y_A) \text{ is singular}\right\}.$$

Here, the term *singular* means that the variety $\gamma_A^{-1}(H_x \cap Y_A)$ has a singular point in $(\mathbb{R}^*)^r$. A general point x in Y_A^* hence corresponds to a hyperplane H_x that is tangent to the toric variety Y_A at a point $\gamma_A(t)$ with nonzero coordinates. We identify sign vectors $\sigma \in \{-1, +1\}^m$ with orthants in $\mathbb{R}^m$. These map in a 2-to-1 manner to orthants in $\mathbb{RP}^{m-1}$. If we intersect them with Y_A^*, then we get the *orthants* of the dual toric variety:

$$Y_{A,\sigma}^* = \left\{x \in Y_A^* : \sigma_i \cdot x_i > 0 \text{ for } i = 1, 2, \ldots, m\right\} \subset \mathbb{RP}^{m-1}. \tag{9}$$

One of these is the distinguished orthant in Theorem 1, part (3).

Example 5. Fix $m = 4$ and $r = 2$. The following matrix has $(1, 1, 1, 1)$ in its row span:

$$A = \begin{pmatrix} 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix}. \tag{10}$$

As in [11], Example 3.9, the toric variety of A is the *twisted cubic curve* in 3-space:

$$Y_A = \overline{\left\{\left(t_1^3 : t_1^2t_2 : t_1t_2^2 : t_2^3\right) \in \mathbb{RP}^3 : t_1, t_2 \in \mathbb{R}^*\right\}}.$$

The dual toric variety Y_A^* is a surface in $(\mathbb{RP}^3)^\vee$. Its points x represent planes in $\mathbb{RP}^3$ that are tangent to the curve Y_A. Such a tangent plane corresponds to a cubic $x_1t^3 + x_2t^2 + x_3t + x_4$ with a double root. Just as we recognize quadrics with a double root by the vanishing of the quadratic discriminant, a cubic with coefficients (x_1, x_2, x_3, x_4) has a double root if and only if the following discriminant vanishes:

$$\Delta_A = \underline{27x_1^2x_4^2} - 18x_1x_2x_3x_4 + 4x_1x_3^3 + 4x_2^3x_4 - x_2^2x_3^2. \tag{11}$$

Hence, Y_A^* is the surface of degree 4 in $(\mathbb{RP}^3)^\vee$ defined by Δ_A. All eight orthants $Y_{A,\sigma}^*$ are nonempty. The coefficient vectors of the following eight cubics lie on different orthants:

$$(t+1)^2(t+3), \qquad (t+5)^2(t-1), \qquad (t-1)^2(t+3), \qquad (t+5)^2(t-8),$$
$$(t-3)^2(t+1), \qquad (t-1)^2(t-3), \qquad \underline{(t-2)^2(t+3)}, \qquad (t+1)^2(t-3).$$

For instance, the underlined cubic corresponds to the point $x = (1, -1, -8, 12)$ in the orthant $Y_{A,\sigma}^*$ associated with the sign vector $\sigma = (+1, -1, -1, +1)$.

Let Δ be a homogeneous polynomial in m variables with $n+2$ monomials and $\mathbf{m}$ one of these monomials. There is a one-to-one correspondence between such pairs $(\Delta, \mathbf{m})$ and pairs (H, λ) where H is a Horn matrix of size $m \times (n+1)$ and λ is a coefficient vector. Namely, for $k = 0, \ldots, n$ write h_k^+, respectively, h_k^- for the positive, respectively, negative part of the column vector h_k, so that $h_k = h_k^+ - h_k^-$. In addition, let $\max_k(h_k^-)$ be the entrywise maximum of the h_k^-. We pass from pairs (H, λ) to pairs $(\Delta, \mathbf{m})$ as follows:

$$\mathbf{m} = x^{\max_k(h_k^-)} \quad \text{and} \quad \Delta = \mathbf{m} \cdot \left(1 - \sum_{k=0}^{n} \lambda_k x^{h_k}\right). \tag{12}$$

For the converse, from pairs $(\Delta, \mathbf{m})$ to pairs (H, λ), we divide Δ by $\mathbf{m}$ and use the same equations to determine the pair (H, λ). Note that the polynomial Δ being homogeneous and the matrix H being a Horn matrix are equivalent conditions using the equations (12). Given a pair $(\Delta, \mathbf{m})$ with associated pair (H, λ), we define the monomial map

$$\phi_{(\Delta,\mathbf{m})}: \quad \left(\mathbb{R}^*\right)^m \to \mathbb{R}^{n+1}, \qquad x \mapsto \left(\lambda_0 x^{h_0}, \lambda_1 x^{h_1}, \ldots, \lambda_n x^{h_n}\right).$$

We now present the definition that is needed for part (3) of Theorem 1.

Definition 6. A *discriminantal triple* $(A, \Delta, \mathbf{m})$ consists of

1. an $r \times m$ integer matrix A of rank r having $(1, 1, \ldots, 1)$ in its row span,
2. an A-homogeneous polynomial Δ that vanishes on the dual toric variety Y_A^*,
3. a distinguished term $\mathbf{m}$ among those that occur in the polynomial Δ,

such that the pair (H, λ) associated to $(\Delta, \mathbf{m})$ is a Horn pair. Here, the polynomial Δ being *A-homogeneous* means that $Av = Aw$ for any two exponent vectors v and w of Δ.

All definitions are now complete. We illustrate Definition 6 for our running example:

Example 7. Let A be the 2×4 matrix in (10), $\Delta = \Delta_A$ its discriminant in (11), and $\mathbf{m} = 27x_1^2x_4^2$ the special term. Then $(A, \Delta, \mathbf{m})$ is a discriminantal triple with associated sign vector $\sigma = (+1, -1, -1, +1)$. The orthant $Y_{A,\sigma}^*$, highlighted in Example 5, is a semialgebraic surface in $Y_A^* \subset \mathbb{RP}^3$. This surface is mapped into the tetrahedron Δ_3 by

$$\phi_{(\Delta,\mathbf{m})}: \quad (x_1, x_2, x_3, x_4) \mapsto \left(\frac{2}{3}\frac{x_2x_3}{x_1x_4}, -\frac{4}{27}\frac{x_3^3}{x_1x_4^2}, -\frac{4}{27}\frac{x_2^3}{x_1^2x_4}, \frac{1}{27}\frac{x_2^2x_3^2}{x_1^2x_4^2}\right). \tag{13}$$

The image of this map is a curve in Δ_3, namely the model $\mathcal{M}$ in Example 4. We verify (1) by comparing (7) with (13). The former is obtained from the latter by setting $x = Hu$.

3. Staged trees

We consider contingency tables $u = (u_{i_1 i_2 \cdots i_m})$ of format $r_1 \times r_2 \times \cdots \times r_m$. Following [4,14], these represent joint distributions of discrete statistical models with $n+1 = r_1 r_2 \cdots r_m$ states. Namely, the contingency table u represents the probability distribution $p := u/|u|$. For any subset $C \subset \{1, \ldots, m\}$, one considers the marginal table u_C that is obtained by summing out all indices not in C. The entries of the marginal table u_C are sums of entries in u. To obtain the entry $u_{I,C}$ of u_C for any state $I = (i_1, i_2, \ldots, i_m)$, we fix the indices of the states in C and sum over the indices not in C. For example, if $m = 4$, $C = \{1, 3\}$, $I = (i, j, k, l)$, then u_C is the $r_1 \times r_3$ matrix with entries

$$u_{I,C} = u_{i+k+} = \sum_{j=1}^{r_2} \sum_{l=1}^{r_4} u_{ijkl}.$$

Such linear forms are the basic building blocks for familiar models with rational MLE.

Consider an undirected graph G with vertex set $\{1, \ldots, m\}$ which is assumed to be *chordal*. The associated *decomposable graphical model* $\mathcal{M}_G$ in Δ_n has the rational MLE

$$\hat{p}_I = \frac{\prod_C u_{I,C}}{\prod_S u_{I,S}}, \tag{14}$$

where the product in the numerator is over all maximal cliques C of G, and the product in the denominator is over all separators S in a junction tree for G. See [14], §4.4.1. We shall regard G as a directed graph, with edge directions given by a perfect elimination ordering on the vertex set $\{1, \ldots, m\}$. This turns $\mathcal{M}_G$ into a Bayesian network. More generally, a *Bayesian network* $\mathcal{M}_G$ is given by a directed acyclic graph G. We write $\mathrm{pa}(j)$ for the set of parents of the node j. The model $\mathcal{M}_G$ in Δ_n has the rational MLE

$$\hat{p}_I = \prod_{j=1}^{m} \frac{u_{I,\mathrm{pa}(j)\cup\{j\}}}{u_{I,\mathrm{pa}(j)}}. \tag{15}$$

If G comes from an undirected chordal graph, then (14) arises from (15) by cancellations.

Example 8 ($m = 4$). We revisit two examples from on page 36 in [4], §2.1. The *star graph* $G = [14][24][34]$ is chordal. The MLE for $\mathcal{M}_G$ is the map Φ with coordinates

$$\hat{p}_{ijkl} = \frac{u_{i++l} \cdot u_{+j+l} \cdot u_{++kl}}{u_{++++} \cdot u_{+++l}^2} = \frac{u_{i+++}}{u_{++++}} \cdot \frac{u_{+j+l}}{u_{+++l}} \cdot \frac{u_{++kl}}{u_{+++l}} \cdot \frac{u_{i++l}}{u_{i+++}}.$$

The left expression is (14). The right is (15) for the directed graph $1 \to 4, 4 \to 2, 4 \to 3$.

The *chain graph* $G = [12][23][34]$ is chordal. Its MLE is the map Φ with coordinates

$$\hat{p}_{ijkl} = \frac{u_{ij++} \cdot u_{+jk+} \cdot u_{++kl}}{u_{+j++} \cdot u_{++k+} \cdot u_{++++}} = \varphi_{(H,\lambda)}(u)_{ijkl}.$$

This is the Horn map given by the matrix H in Figure 1 and $\lambda = (1, \ldots, 1)$.

The formulas (14) and (15) are familiar to statisticians. Theorem 1 places them into a larger context. However, some readers may find our approach too algebraic and too general. Our aim in this section is to lay out a useful middle ground: staged tree models.

Staged trees were introduced by Smith and Anderson [16] as a generalization of discrete Bayesian networks. They furnish an intuitive representation of many situations that the above graphs G cannot capture. In spite of their wide scope, staged tree models are appealing because of their intuitive formalism for encoding events. For an introduction, see the textbook [3]. In what follows, we study parts (1) and (2) in Theorem 1 for staged trees.

To define a *staged tree model*, we consider a directed rooted tree $\mathcal{T}$ with at least two edges emanating from each nonleaf vertex, a label set $S = \{s_i \mid i \in I\}$ and a labeling $\theta\colon \mathrm{E}(\mathcal{T}) \to S$ of the edges of the tree. Each vertex of $\mathcal{T}$ has a corresponding *floret*, which is the multiset of edge labels emanating from it. The labeled tree $\mathcal{T}$ is a *staged tree* if any two florets are either equal or disjoint. Two vertices in $\mathcal{T}$ are in the same stage if their corresponding florets are the same. From now on, F denotes the set of florets of $\mathcal{T}$.

Definition 9. Let J be the set of root-to-leaf paths in the tree $\mathcal{T}$. We set $|J| = n + 1$. For $i \in I$ and $j \in J$, let μ_{ij} denote the number of times edge label s_i appears in the jth root-to-leaf path. The *staged tree model* $\mathcal{M}_{\mathcal{T}}$ is the image of the parametrization

$$\phi_{\mathcal{T}}\colon \quad \Theta \to \Delta_n, \qquad (s_i)_{i\in I} \mapsto (p_j)_{j\in J},$$

where the parameter space is $\Theta := \{(s_i)_{i\in I} \in (0,1)^{|I|} : \sum_{s_i \in f} s_i = 1 \text{ for all florets } f \in F\}$, and $p_j = \prod_{i\in I} s_i^{\mu_{ij}}$ is the product of the edge parameters on the jth root-to-leaf path.

In the model $\mathcal{M}_{\mathcal{T}}$, the tree $\mathcal{T}$ represents possible sequences of events. The parameter s_i associated to an edge vv' is the transition probability from v to v'. All parameter labels in a floret sum to 1. The fact that distinct nodes in $\mathcal{T}$ can have the same floret of parameter labels enables staged tree models to encode conditional independence statements [16]. This allows us to represent any discrete Bayesian network or decomposable model as a staged tree model. Our first staged tree was seen in Example 2. Here is another specimen.

Example 10 ($n = 15$)**.** Consider the decomposable model for binary variables given by the 4-chain $G = [12][23][34]$ as in Example 8. Figure 1 shows a realization of $\mathcal{M}_G$ as a staged tree model $\mathcal{M}_{\mathcal{T}}$. The leaves of $\mathcal{T}$ represent the outcome space $\{0,1\}^4$. Nodes with the same color have the same associated floret. The blank nodes all have different florets. The seven florets of $\mathcal{T}$ are

$$f_1 = \{s_0, s_1\}, \qquad f_2 = \{s_2, s_3\}, \qquad f_3 = \{s_4, s_5\}, \qquad f_4 = \{s_6, s_7\}, \qquad f_5 = \{s_8, s_9\},$$
$$f_6 = \{s_{10}, s_{11}\}, \qquad f_7 = \{s_{12}, s_{13}\}.$$

Next, we show that staged tree models have rational MLE, so they satisfy part (1) of Theorem 1. Our formula for Φ uses the notation for I, J and μ_{ij} introduced in Definition 9. This formula is known in the literature on chain event graphs (see, e.g., [15]).

Proposition 11. *Let $\mathcal{M}_{\mathcal{T}}$ be a staged tree model, and let $u = (u_j)_{j\in J}$ be a vector of counts. For $i \in I$, let f be the floret containing the label s_i, and define the estimates*

$$\hat{s}_i := \frac{\sum_j \mu_{ij} u_j}{\sum_{s_\ell \in f} \sum_j \mu_{\ell j} u_j} \quad \text{and} \quad \hat{p}_j := \prod_{i\in I} (\hat{s}_i)^{\mu_{ij}}.$$

The rational function Φ that sends $(u_j)_{j\in J}$ to $(\hat{p}_j)_{j\in J}$ is the MLE of the model $\mathcal{M}_{\mathcal{T}}$.

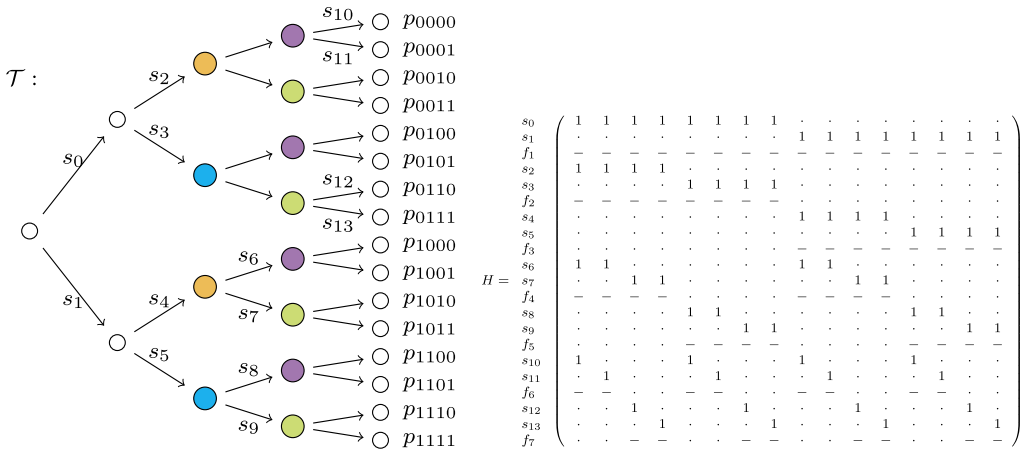


Figure 1. A staged tree $\mathcal{T}$ and its Horn matrix H from Proposition 11. Entries $-$ indicate -1.

Proof. We prove that the likelihood function $L(p,u)$ has a unique maximum at $p = (\hat{p}_j)_{j\in J}$. For a floret $f \in F$, we fix the vector of parameters $s_f = (s_i)_{s_i\in f}$, and we define the local likelihood function $L_f(s_f,u) = \prod_{s_i\in f} s_i^{\alpha_i}$, where $\alpha_i = \sum_j \mu_{ij}u_j$. We have

$$L(p,u) = \prod_j p_j^{u_j} = \prod_j \prod_i s_i^{u_j\mu_{ij}} = \prod_i s_i^{\alpha_i} = \prod_{f\in F} L_f(s_f,u).$$

Since the L_f depend on disjoint sets of unknowns, maximizing L is achieved by maximizing the factors L_f separately. But L_f is the likelihood function of the full model $\Delta_{|f|-1}$, given the data vector $(\alpha_i)_{s_i\in f}$. The MLE of that model is $\hat{s}_i = \alpha_i / \sum_{s_\ell\in f}\alpha_\ell$, where $s_i \in f$. We conclude that $\operatorname{argmax}_{s_f}(L_f(s_f,u)) = (\hat{s}_i)_{s_i\in f}$ and $\operatorname{argmax}_p(L(p,u)) = (\hat{p}_j)_{j\in J}$. □

Remark 12. Here is a method for evaluating the MLE in Proposition 11. Let $[v] \subset J$ be the set of root-to-leaf paths through a node v in the tree $\mathcal{T}$ and define $u_{[v]} = \sum_{j\in[v]} u_j$. The ratio $\frac{u_{[v']}}{u_{[v]}}$ is the empirical transition probability from v to v' given arrival at v. To obtain $\hat{s}_i$ we first compute the quotients $\frac{u_{[v']}}{u_{[v]}}$ for all edges vv' with parameter label s_i. We aggregate them by adding their numerators and denominators separately. This gives $\hat{s}_i = (\sum u_{[v']})/(\sum u_{[v]})$, where both sums range over all edges vv' with parameter label s_i.

Proposition 11 yields an explicit description of the Horn pair (H,λ) associated to $\mathcal{M}_\mathcal{T}$.

Corollary 13. *Fix a staged tree model $\mathcal{M}_\mathcal{T}$ as above. Let H be the $(|I|+|F|)\times|J|$ matrix whose rows are indexed by the set $I \sqcup F$ and entries are given by*

$$h_{ij} = \mu_{ij} \quad \text{for } i \in I, \quad \text{and}$$
$$h_{fj} = -\sum_{s_\ell\in f} \mu_{\ell j} \quad \text{for } f \in F.$$

Define $\lambda \in \{-1,+1\}^{|J|}$ by $\lambda_j = (-1)^{\sum_f h_{fj}}$. Then (H,λ) is a Horn pair for $\mathcal{M}_\mathcal{T}$.

Given a staged tree $\mathcal{T}$, we call the matrix H in Corollary 13 the *Horn matrix* of $\mathcal{T}$.

Remark 14. In Corollary 13, for a floret f, let H_f be the submatrix of H with row indices $\{i : s_i \in f\} \cup \{f\}$. Then H is the vertical concatenation of the matrices H_f for $f \in F$.

Example 15. For the tree $\mathcal{T}$ in Example 10, the Horn matrix H of $\mathcal{M}_\mathcal{T}$ is given in Figure 1. The vector λ of the Horn pair (H, λ) is the vector of ones $(1, \ldots, 1) \in \mathbb{R}^{16}$. The rows of H are indexed by the florets and labels

$$(s_0, s_1, f_1, s_2, s_3, f_2, s_4, s_5, f_3, s_6, s_7, f_4, s_8, s_9, f_5, s_{10}, s_{11}, f_6, s_{12}, s_{13}, f_7).$$

Note that (H, λ) is not minimal. Following the recipe in Lemma 3, we can delete the rows s_0, s_1, f_2, f_3 of the matrix H by summing the pairs (s_0, f_2) and (s_1, f_3) and deleting zero rows. The result is the minimal Horn pair (H', λ'), where $\lambda' = (-1, \ldots, -1)$.

Two staged trees $\mathcal{T}$ and $\mathcal{T}'$ are called *statistically equivalent* in [8] if there exists a bijection between the sets of root-to-leaf paths of $\mathcal{T}$ and $\mathcal{T}'$ such that, after applying this bijection, $\mathcal{M}_\mathcal{T} = \mathcal{M}_{\mathcal{T}'}$ in the open simplex Δ_n. A staged tree model may have different but statistically equivalent tree representations. In [8], Theorem 1, it is shown that statistical equivalence of staged trees can be determined by a sequence of operations on the trees, named *swap* and *resize*. One of the advantages of describing a staged tree model via its Horn pair is that it gives a new criterion to decide whether two staged trees are statistically equivalent. This is simpler to implement than the criterion given in [8].

Corollary 16. *Two staged trees are statistically equivalent if and only if their associated Horn pairs reduce to the same minimal Horn pair.*

One natural operation on a staged tree $\mathcal{T}$ is identifying two florets of the same size. This gives a new staged tree $\mathcal{T}'$ whose Horn matrix is easy to get from that of $\mathcal{T}$.

Corollary 17. *Let $\mathcal{T}'$ be a staged tree arising from $\mathcal{T}$ by identifying two florets f and f', say by the bijection $(-)' : f \to f'$. The Horn matrix H' of $\mathcal{M}_{\mathcal{T}'}$ arises from the Horn matrix H of $\mathcal{M}_\mathcal{T}$ by replacing the blocks H_f and $H_{f'}$ in H by the block H'_f defined by*

$$h'_{ij} = h_{ij} + h_{i'j} \quad \text{for } s_i \in f,$$
$$h'_{fj} = h_{fj} + h_{f'j}.$$

Proof. This follows from the definition of the Horn matrices for $\mathcal{M}_\mathcal{T}$ and $\mathcal{M}_{\mathcal{T}'}$. □

Example 18. Let $\mathcal{T}'$ be the tree obtained from Example 10 by identifying florets f_4 and f_5 in $\mathcal{T}$. Then $\mathcal{M}_{\mathcal{T}'}$ is the independence model of two random variables with four states.

Now we turn to part (3) of Theorem 1. We describe the triple $(A, \Delta, \mathbf{m})$ for a staged tree model $\mathcal{M}_\mathcal{T}$. The pair (H, λ) was given in Corollary 13. Let A be any matrix whose rows span the left kernel of H, set $m = |I| + |F|$, and write s for the m-tuple of parameters $(s_i, s_f)_{i \in I, f \in F}$. From the Horn matrix in Corollary 13, we see that

$$\Delta = \mathbf{m} \cdot \left(1 - \sum_j (-1)^{\epsilon_j} \prod_i \left(\frac{s_i}{s_f}\right)^{\mu_{ij}}\right),$$

where f depends on i, $\mathbf{m} = \operatorname{lcm}(\prod_i s_f^{\mu_{ij}} : f \in F)$ and $\epsilon_j = \sum_i \mu_{ij}$. The sign vector σ for the triple $(A, \Delta, \mathbf{m})$ is given by $\sigma_i = +1$ for $i \in I$ and $\sigma_f = -1$ for $f \in F$. Then $Y^*_{A,\sigma}$ gets mapped to $\mathcal{M}_\mathcal{T}$ via $\phi_{(\Delta,\mathbf{m})}$. Moreover, the map $\phi_\mathcal{T}$ from Definition 9 factors through $\phi_{(\Delta,\mathbf{m})}$. Indeed, if we define $\iota : \Theta \to Y^*_{A,\sigma}$ by $(s_i)_{i\in I} \mapsto (s_i, -1)_{i\in I, f\in F}$, then $\phi_\mathcal{T} = \phi_{(\Delta,\mathbf{m})} \circ \iota$. The following derivation is an extension of that in [11], Example 3.13.

Example 19. Let $\mathcal{M}_\mathcal{T}$ be the 4-chain model in Example 10. Here, the discriminant is

$$\begin{aligned}\Delta = {} & f_1 f_2 f_3 f_4 f_5 f_6 f_7 \\ & - s_0 s_2 s_6 s_{10} f_3 f_5 f_7 - s_0 s_2 s_6 s_{11} f_3 f_5 f_7 - s_0 s_2 s_7 s_{12} f_3 f_5 f_6 - s_0 s_2 s_7 s_{13} f_3 f_5 f_6 \\ & - s_0 s_3 s_8 s_{10} f_3 f_4 f_7 - s_0 s_3 s_8 s_{11} f_3 f_4 f_7 - s_0 s_3 s_9 s_{12} f_3 f_4 f_6 - s_0 s_3 s_9 s_{13} f_3 f_4 f_6 \\ & - s_1 s_4 s_6 s_{10} f_2 f_5 f_7 - s_1 s_4 s_6 s_{11} f_2 f_5 f_7 - s_1 s_4 s_7 s_{12} f_2 f_5 f_6 - s_1 s_4 s_7 s_{13} f_2 f_5 f_6 \\ & - s_1 s_5 s_8 s_{10} f_2 f_4 f_7 - s_1 s_5 s_8 s_{11} f_2 f_4 f_7 - s_1 s_5 s_9 s_{12} f_2 f_4 f_6 - s_1 s_5 s_9 s_{13} f_2 f_4 f_6.\end{aligned}$$

Our notation for the parameters matches the row labels of the Horn matrix H in Figure 1. This polynomial of degree 7 is irreducible, so it equals the A-discriminant: $\Delta = \Delta_A$. The underlying matrix A has format 13×21, and we represent it by its associated toric ideal

$$\begin{aligned}I_A = \langle & s_{10} - s_{11}, s_1 s_5 f_2 - s_0 s_3 f_3, s_1 s_4 f_2 - s_0 s_2 f_3, s_5 s_9 f_4 - s_4 s_7 f_5, s_3 s_9 f_4 - s_2 s_7 f_5, \\ & s_{12} - s_{13}, s_5 s_8 f_4 - s_4 s_6 f_5, s_3 s_8 f_4 - s_2 s_6 f_5, s_9 s_{13} f_6 - s_8 s_{11} f_7, s_7 s_{13} f_6 - s_6 s_{11} f_7, \\ & s_0 s_2 s_6 s_{11} - f_1 f_2 f_4 f_6, s_0 s_2 s_7 s_{13} - f_1 f_2 f_4 f_7, s_0 s_3 s_8 s_{11} - f_1 f_2 f_5 f_6, \\ & s_0 s_3 s_9 s_{13} - f_1 f_2 f_5 f_7, s_1 s_4 s_6 s_{11} - f_1 f_3 f_4 f_6, s_1 s_4 s_7 s_{13} - f_1 f_3 f_4 f_7, \\ & s_1 s_5 s_9 s_{13} - f_1 f_3 f_5 f_7, s_1 s_5 s_8 s_{11} - f_1 f_3 f_5 f_6 \rangle.\end{aligned}$$

The toric variety $Y_A = \mathcal{V}(I_A)$ has dimension 12 and degree 141. It lives in a linear space of codimension 2 in $\mathbb{P}^{20}$, where it is defined by eight cubics and eight quartics. The dual variety $Y^*_A = \mathcal{V}(\Delta_A)$ is the above hypersurface of degree seven. We have $\mathbf{m} = f_1 f_2 f_3 f_4 f_5 f_6 f_7$, and σ is the vector in $\{-1, +1\}^{21}$ that has entry $+1$ at the indices corresponding to the s_i and entry -1 at the indices corresponding to the f_i.

It would be interesting to study the combinatorics of discriminantal triples for staged tree models. Our computations suggest that, for many such models, the polynomial Δ is irreducible and equals the A-discriminant Δ_A of the underlying configuration A. However, this is not true for all staged trees, as seen in equation (2) of Example 2. We close this section with a familiar class of models with rational MLE whose associated Δ factor.

Example 20. The *multinomial distribution* encodes the experiment of rolling a k-sided die m times and recording the number of times one observed the jth side, for $j = 1, \ldots, k$. The associated model $\mathcal{M}$ is the independence model for m identically distributed random variables on k states. We have $n + 1 = \binom{k+m-1}{m}$. The Horn matrix H is the $(k + 1) \times (n + 1)$ matrix whose columns are the vectors $(-m, i_1, i_2, \ldots, i_k)^T$ where $i_1, i_2, \ldots, i_k$ are nonnegative integers whose sum equals m. Here, $A = (1\ 1\ \cdots\ 1)$, so the A-discriminant equals $\Delta_A = x_0 + x_1 + \cdots + x_k$. The following polynomial is a multiple of Δ_A:

$$\Delta = (-x_0)^m - (x_1 + x_2 + \cdots + x_k)^m.$$

This Δ, with its marked term $\mathbf{m} = (-x_0)^m$, encodes the MLE for the model $\mathcal{M}$:

$$\hat{p}_{(i_1,\ldots,i_k)} = \prod_{j=1}^{k} \left(\frac{\sum_{|I|=m} u_I \cdot I_j}{m \sum_{|I|=m} u_I} \right)^{i_j}$$

Here, I ranges over all vectors in $\mathbb{N}^k$ that sum to m, and I_j denotes the jth entry of I.

4. Proof of the main theorem

In this section, we prove Theorem 1. For a pair (H, λ) consisting of a Horn matrix H and a coefficient vector λ, let φ be the rational map defined in (4). We use φ and $\varphi_{(H,\lambda)}$ interchangeably in this section, as well as ϕ and $\phi_{(\Delta,\mathbf{m})}$. Recall that its jth coordinate is

$$\varphi_j(v) = \lambda_j \prod_{i=1}^{m} \left(\sum_{k=0}^{n} h_{ik} v_k \right)^{h_{ij}}. \tag{16}$$

For a fixed data vector $u \in \mathbb{N}^{n+1}$, we define the likelihood function for the image of φ:

$$L_u : \quad \mathbb{R}^{n+1} \to \mathbb{R}, \qquad v \mapsto \prod_{j=0}^{n} \varphi_j(v)^{u_j}. \tag{17}$$

Lemma 21. *Let $H = (h_{ij})$ be a Horn matrix, λ a vector satisfying (3) and $u \in \mathbb{N}^{n+1}$. Then u is a critical point of its own likelihood function L_u. Furthermore, if u' is another critical point of L_u, then $\varphi(u) = \varphi(u')$.*

Proof. We compute the partial derivatives of L_u. For $\ell = 0, \ldots, n$, we find

$$\begin{aligned}
\frac{\partial}{\partial v_\ell} L_u(v) &= \sum_{j=0}^{n} u_j \frac{L_u(v)}{\varphi_j(v)} \frac{\partial}{\partial v_\ell} \varphi_j(v) \\
&= \sum_{j=0}^{n} u_j \frac{L_u(v)}{\varphi_j(v)} \sum_{i=1}^{m} h_{ij} \frac{\varphi_j(v)}{\sum_{k=0}^{n} h_{ik} v_k} h_{i\ell} \\
&= L_u(v) \sum_{i=1}^{m} \sum_{j=0}^{n} \frac{u_j h_{ij} h_{i\ell}}{\sum_{k=0}^{n} h_{ik} v_k} = L_u(v) \sum_{i=1}^{m} \frac{h_{i\ell} \sum_{j=0}^{n} h_{ij} u_j}{\sum_{k=0}^{n} h_{ik} v_k}.
\end{aligned}$$

For $v = u$, this evaluates to zero, since the sums in the fraction cancel and the ℓth column of H sums to zero. This shows that u is a critical point.

Next, let u' be another critical point of L_u. Using terminology from [10], Theorem 1, this means that $\varphi(u')$ is a critical point of the likelihood function $L(p, u)$ of the model $\mathcal{M}$ defined as the image of φ. The same holds for $\varphi(u)$. By the implication (ii) to (i) in [10], Theorem 1, the model $\mathcal{M}$ has ML degree one. This implies $\varphi(u) = \varphi(u')$. □

We use [10] to explain the relation between models with rational MLE and Horn pairs.

Proof of Theorem 1, Equivalence of (1) and (2). Let $\mathcal{M}$ be a model with rational MLE Φ. The Zariski closure of $\mathcal{M}$ is a variety whose likelihood function has a unique critical point. By [10], Theorem 1, there is a Horn matrix H and a coefficient vector λ such that $\varphi_{(H,\lambda)} = \Phi$. Now, the required sum-to-one and positivity conditions for $\varphi_{(H,\lambda)}$ are satisfied because they are satisfied by the MLE Φ. Indeed, the MLE of any discrete statistical model maps positive vectors u in $\mathbb{R}^{n+1}_{>0}$ into the simplex Δ_n. Conversely, we claim that every Horn pair (H, λ) specifies a nonempty model $\mathcal{M}$ with rational MLE. Indeed, define $\mathcal{M}$ to be the image of $\varphi_{(H,\lambda)}$. By the defining properties of the Horn pair, we have $\mathcal{M} \subset \Delta_n$. Lemma 21 shows that $\varphi_{(H,\lambda)}$ is the MLE of $\mathcal{M}$. □

Next, we relate Horn pairs to discriminantal triples.

Proof of Theorem 1, Equivalence of (2) and (3). We already exhibited a bijection between pairs (H, λ) and pairs $(\Delta, \mathbf{m})$ given by Equation (12). The matrix A is the left kernel of H and forms the triple $(A, \Delta, \mathbf{m})$. It is a matrix of size $r \times m$ of rank r. When H is a Horn matrix, A contains $(1, \ldots, 1)$ in its row span. This implies that the polynomial Δ is homogeneous, which in turn implies that it is A-homogeneous by $AH = 0$.

Next, we show that the pair (H, λ) being friendly corresponds to the polynomial Δ vanishing on Y_A^*. This is part of the desired equivalence.

Claim. *The pair (H, λ) is friendly if and only if the A-homogeneous polynomial Δ vanishes on the dual toric variety Y_A^*.*

Proof of Claim. Let (H, λ) be friendly and A as above. The Laurent polynomial $q := \Delta/\mathbf{m}$ is a rational function on $\mathbb{P}^{m-1}$ that vanishes on the dual toric variety Y_A^*. To see this, consider the exponentiation map $\varphi_2 \colon \mathbb{P}^{m-1} \to \mathbb{R}^{n+1}$, $x \mapsto \lambda * x^H$, where $*$ is the entrywise product and $x^H := (x^{h_0}, \ldots, x^{h_n})$. Let $f = 1 - (p_0 + \cdots + p_n)$. We have $q = f \circ \varphi_2$. By [10], Theorems 1 and 2, the function φ_2 maps an open dense subset of Y_A^* dominantly to the closure $\overline{\mathcal{M}}$ of the image of $\varphi_{(H,\lambda)}$. Since $f = 0$ on $\overline{\mathcal{M}}$, we have $f \circ \varphi_2 = 0$ on an open dense subset of Y_A^*, hence $q = 0$ on Y_A^*, so $\Delta = 0$ there as well.

Conversely, let Δ vanish on Y_A^*. We claim that $q(x)$ is zero for all $x = Hu$ in the image of the linear map H. We may assume $\mathbf{m}(x) \neq 0$. We only need to show that x is in the dual toric variety Y_A^*, since Δ vanishes on it. So, let $x_i = \sum_{j=0}^n h_{ij} u_j$ for $i = 1, \ldots, m$. We claim that $t = (1, \ldots, 1)$ is a singular point of the hypersurface

$$\gamma_A^{-1}(H_x \cap Y_A) = \left\{ t \in \mathbb{C}^r \;\middle|\; \sum_{i=1}^m x_i t^{a_i} = 0 \right\}.$$

First, the point t lies on that hypersurface since the columns of H sum to zero:

$$\sum_{i=1}^m x_i = \sum_{i=1}^m \sum_{j=0}^n h_{ij} u_j = \sum_{j=0}^n u_j \sum_{i=1}^m h_{ij} = 0.$$

For $s = 1, \ldots, r$ we have $\frac{\partial}{\partial t_s} t^{a_i} = a_{si} t^{a_i - e_s}$, with e_s the standard basis vector of $\mathbb{Z}^r$, and

$$\frac{\partial}{\partial t_s} \sum_{i=1}^m x_i t^{a_i} = \sum_{i=1}^m \sum_{j=0}^n h_{ij} u_j a_{si} t^{a_i - e_s} = \sum_{j=0}^n u_j \sum_{i=1}^m a_{si} h_{ij} t^{a_i - e_s}.$$

This is zero at $t = (1, \ldots, 1)$ because $AH = 0$. □

We now prove the rest of the equivalence. Let (H, λ) be a Horn pair, let φ be its Horn map and let ϕ be the associated monomial map. Let $\mathcal{M}$ be the statistical model with MLE φ, so $\mathcal{M} = \varphi(\mathbb{R}^{n+1}_{>0})$. We have $\varphi = \phi \circ H$. By Proposition 23, there exists a unique sign vector σ such that $\operatorname{im} H|_{\mathbb{R}^{n+1}_{>0}} \subseteq \mathbb{R}^m_\sigma$. From the proof of the above claim, we know that $\operatorname{im} H \subseteq Y_A^*$. Together, we have

$$\mathcal{M} = \varphi\big(\mathbb{R}^{n+1}_{>0}\big) = \phi(\operatorname{im} H|_{\mathbb{R}^{n+1}_{>0}}) \subseteq \phi\big(Y^*_{A,\sigma}\big).$$

By [10], Theorems 1 and 2, we have $\phi(Y_A^*) \subseteq \mathcal{M}'$, where $\mathcal{M}'$ is the real part of $\overline{\varphi(\mathbb{C}^{n+1})}$. We also have $\phi(Y^*_{A,\sigma}) \subseteq \mathbb{R}^{n+1}_{>0}$ by definition of the orthant. Thus $\phi(Y^*_{A,\sigma}) \subseteq \mathcal{M}' \cap \mathbb{R}^{n+1}_{>0}$. Every element in the latter set is a fixed point of the rational function φ, by a similar argument as in Lemma 21 for complex space. Hence $\mathcal{M}' \cap \mathbb{R}^{n+1}_{>0} = \mathcal{M}$, so $\phi(Y^*_{A,\sigma}) \subseteq \mathcal{M}$.

Finally, if $(A, \Delta, \mathbf{m})$ is a discriminantal triple then (H, λ) is a Horn pair by definition. This completes the proof of Theorem 1. □

In the next two propositions, we formulate simple criteria to decide whether the image of the map $\varphi_{(H,\lambda)}$ associated to a Horn matrix H and a coefficient vector λ is a statistical model. These are essential for constructing models with rational MLE in Algorithm 1.

Proposition 22. *Let (H, λ) be a friendly pair. If there exists a vector $u_0 \in \mathbb{R}^{n+1}$ such that $\varphi(u_0) > 0$, then we have $\varphi(u) > 0$ for all u in $\mathbb{R}^{n+1}_{>0}$ where it is defined.*

Proof. The function φ is homogeneous of degree zero. It suffices to prove that each coordinate of $\varphi(u)$ is a positive real number, for all vectors u with positive integer entries. Indeed, every positive u in $\mathbb{R}^{n+1}$ can be approximated by rational vectors, which can be scaled to be integral. The open subset $U = \varphi^{-1}(\Delta_n)$ of $\mathbb{R}^{n+1}$ contains u_0 by our assumptions. If $U = \mathbb{R}^{n+1}$, then we are done. Else, U has a nonempty boundary ∂U. By continuity, $\partial U \subseteq \varphi^{-1}(\partial \Delta_n)$. The likelihood function L_u for the data vector u vanishes on ∂U.

We claim that L_u has a critical point in U. The closed subset $\overline{U}$ is homogeneous. Seen in projective space $\mathbb{P}^n$, it becomes compact. The likelihood function L_u is well defined on this compact set in $\mathbb{P}^n$, since it is homogeneous of degree zero, and L_u vanishes on the boundary. Hence the restriction $L_u|_U$ is either identically zero or it has a critical point in U. But, since $u_0 \in U$ is a point with $L_u(u_0) \neq 0$, the second statement must be true.

Pick such a critical point u'. Since U is open in $\mathbb{R}^{n+1}$, the point u' is also critical point of L_u. By Lemma 21 and since $u' \in U$, we have $\varphi(u) = \varphi(u') > 0$. □

Proposition 23. *Let (H, λ) be a friendly pair, with no zero or collinear rows in H. Then (H, λ) is a Horn pair if and only if for every row r_i of H all nonzero entries of r_i have the same sign σ_i, and the sign vector $\sigma = (\sigma_i)$ satisfies $\lambda_j \sigma^{h_j} > 0$ for all columns j.*

Proof. Let (H, λ) be a Horn pair. Let $\ell_1, \ldots, \ell_k$ be the linear forms corresponding to the rows in H that have both positive and negative entries. Since ℓ_1 has positive and negative coefficients, there exists a positive vector u such that $\ell_1(u) = 0$. Since (H, λ) is minimal, we may choose $u > 0$ such that $\ell_1(u) = 0$ but $\ell_{k'}(u) \neq 0$ for all $k' \neq 1$. The form ℓ_1 appears in the numerator of some coordinate of φ, making this coordinate zero at u. But this contradicts the fact that (H, λ) is a Horn pair. Therefore, we cannot have rows with both positive and negative entries. The inequalities $\lambda_j \sigma^{h_j} > 0$ then follow from the definition of a Horn pair by evaluating $\varphi(u)$ for some positive vector u.

Conversely, if the sign vector σ is well-defined, the inequalities $\lambda_j \sigma^{h_j} > 0$ imply that $\varphi(u) > 0$ for all positive u. Hence (H, λ) is a Horn pair. □

Every model with rational MLE *arises from* a toric variety Y_A. In some cases, the model *is itself* a toric variety Y_C. It is crucial to distinguish the two matrices A and C. The two toric structures are very different. For instance, every undirected graphical model is toric [4], Proposition 3.3.3. The toric varieties Y_C among staged tree models $\mathcal{M}_\mathcal{T}$ were classified in [5]. The 4-chain model $\mathcal{M}_\mathcal{T} = Y_C$ *is itself* a toric variety of dimension 7 in $\mathbb{P}^{15}$. But it *arises from* a toric variety Y_A of dimension 12 in $\mathbb{P}^{20}$, seen in Example 19.

Toric models with rational MLE play an important role in *geometric modeling* [2,6]. Given a matrix $C \in \mathbb{Z}^{r\times(n+1)}$ and a vector of weights $w \in \mathbb{R}^{n+1}_{>0}$, one considers the *scaled projective toric variety* $Y_{C,w}$ in $\mathbb{RP}^n$. This is defined as the closure of the image of

$$\gamma_{C,w}: \quad \left(\mathbb{R}^*\right)^r \to \mathbb{RP}^n, \qquad (t_1,\ldots,t_r) \mapsto \left(w_0\prod_{i=1}^r t_i^{c_{i0}}, w_1\prod_{i=1}^r t_i^{c_{i1t}}, \ldots, w_n\prod_{i=1}^r t_i^{c_{in}}\right). \tag{18}$$

The set $\mathcal{M}_{C,w}$ of positive points in $Y_{C,w}$ is a statistical model in Δ_n. There is a natural homeomorphism from the toric model $\mathcal{M}_{C,w}$ onto the polytope of C. This is known in geometry as the *moment map*. For a reference from algebraic statistics, see [4], Proposition 2.1.5. In geometric modeling, the pair (C, w) defines *toric blending functions* [13].

It is desirable for the toric blending functions to have *rational linear precision* [2,13]. The property is rare and it depends in a subtle way on (C, w). Garcia-Puente and Sottile [6] established the connection to algebraic statistics. They showed that rational linear precision holds for (C, w) if and only if the statistical model $\mathcal{M}_{C,w}$ has rational MLE.

Example 24. The most classical blending functions with rational linear precision live on the triangle $\{x \in \mathbb{R}^3_{>0} : x_1+x_2+x_3 = 1\}$. They are the *Bernstein basis polynomials*

$$\frac{m!}{i!j!(m-i-j)!}x_1^i x_2^j x_3^{m-i-j} \quad \text{for } i, j \geq 0, i+j \leq m. \tag{19}$$

Here, C is the $3 \times \binom{m+1}{2}$ matrix whose columns are the vectors $(i, j, m-i-j)$. The weights are $w_{(i,j)} = \frac{m!}{i!j!(m-i-j)!}$. The toric model $\mathcal{M}_{C,w}$ is the multinomial family, where (19) is the probability of observing i times 1, j times 2 and $m-i-j$ times 3 in m trials. This model has rational MLE, as seen in Example 20. Again, notice the distinction between the two toric varieties. Here, Y_A is a point in $\mathbb{P}^m$, whereas Y_C is a surface in $\mathbb{P}^{\binom{m}{2}-1}$.

Clarke and Cox [2] raise the problem of characterizing all pairs (C, w) with rational linear precision. This was solved by Duarte and Görgen [5] for pairs arising from staged trees. While the problem remains open in general, our theory in this paper offers new tools. We may ask for a characterization of discriminantal triples whose models are toric.

5. Constructing models with rational MLE

Part (3) in Theorem 1 allows us to construct models with rational MLE starting from a matrix A that defines a projective toric variety Y_A. To carry out this construction effectively, we propose Algorithm 1. In most cases, the dual variety Y_A^* is a hypersurface, and we can compute its defining polynomial Δ_A, the *discriminant* [7]. The polynomial Δ in a discriminantal triple can be any homogeneous multiple of Δ_A, but we just take $\Delta = \Delta_A$. For all terms $\mathbf{m}$ in Δ_A, we check whether $(A, \Delta_A, \mathbf{m})$ is a discriminantal triple. We implemented this algorithm in `Macaulay2` [9], and our code is available online at [18].

Algorithm 1: From toric varieties to statistical models

Input : An integer matrix A of size $r \times m$ with $(1, \ldots, 1)$ in its row span

Output: An integer n and a collection of statistical models $\mathcal{M}^{(\ell)} = (\Phi^{(\ell)}, I^{(\ell)})$, where $\Phi^{(\ell)} \colon \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is a rational MLE for $\mathcal{M}^{(\ell)}$, and $I^{(\ell)} \subseteq \mathbb{R}[p_0, \ldots, p_n]$ is the defining prime ideal of $\mathcal{M}^{(\ell)}$.

```
1  Compute the A-discriminant Δ_A ∈ ℤ[x_1, ..., x_m];
2  n ← #terms(Δ_A) − 2;
3  models ← {};
4  for 0 ≤ ℓ ≤ n + 1 do
5  |  m ← terms(Δ_A)_ℓ;
6  |  q ← 1 − Δ_A/m;
7  |  for 0 ≤ j ≤ n do
8  |  |  λ_j ← coefficients(q)_j;
9  |  |  h_j ← exponent_vectors(q)_j;
10 |  |  Φ_j^(ℓ) ← (u ↦ λ_j ∏_{i=1}^m (∑_{k=0}^n h_ik u_k)^{h_ij});
11 |  end
12 |  H ← (h_ij);
13 |  Choose any positive vector v in ℝ_{>0}^{n+1};
14 |  if Φ_j^(ℓ)(v) > 0 for j = 0, 1, ..., n then
15 |  |  Compute the ideal I^(ℓ) of the image of Φ^(ℓ);
16 |  |  models ← models ∪ {(Φ^(ℓ), I^(ℓ))};
17 |  end
18 end
19 return models;
```

Lines 1 and 15 of Algorithm 1 are computations with Gröbner bases. Executing Line 15 can be very slow. It may be omitted if one is satisfied with obtaining the parametric description and MLE $\Phi^{(\ell)}$ of the model $\mathcal{M}_\ell$. For the check in Line 14, we rely on Proposition 22 for correctness. A check based on the criterion in Proposition 23 is also possible.

Example 25 ($r = 2$, $m = 4$)**.** For distinct integers $\alpha, \beta, \gamma > 0$ with $\gcd(\alpha, \beta, \gamma) = 1$ let

$$A_{\alpha,\beta,\gamma} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & \alpha & \beta & \gamma \end{pmatrix}.$$

We ran Algorithm 1 for all 613 such matrices with $0 < \alpha < \beta < \gamma \leq 17$. Line 1 computes the discriminant Δ_A of the univariate polynomial $f(t) = x_1 + x_2 t^\alpha + x_3 t^\beta + x_4 t^\gamma$. The number $n + 2$ of terms of these discriminants equals $7927/613 = 12.93$ on average. Thus a total of 7927 candidate triples $(A, \Delta_A, \mathbf{m})$ were tested in Lines 12 to 21. Precisely, 123 of these were found to be discriminantal triples. This is a fraction of 1.55%. Hence, only 1.55% of the resulting complex varieties permitted by [10] are actually statistical models.

Here is a typical model that was discovered. Take $\alpha = 1$, $\beta = 4$, $\gamma = 7$. The discriminant

$$\begin{aligned} \Delta_A = {} & 729x_2^4x_3^6 - 6912x_1^3x_3^7 - 8748x_2^5x_3^4x_4 + 84\,672x_1^3x_2x_3^5x_4 + 34\,992x_2^6x_3^2x_4^2 \\ & - 351\,918x_1^3x_2^2x_3^3x_4^2 - 46\,656x_2^7x_4^3 + 518\,616x_1^3x_2^3x_3x_4^3 \underline{- 823\,543x_1^6x_4^4} \end{aligned}$$

has 9 terms, so $n = 7$. The term $\mathbf{m}$ is underlined. The associated model is a curve of degree ten in Δ_7. Its prime ideal $I^{(\ell)}$ is generated by 18 quadrics. Among them are 15 binomials that define a toric surface of degree six: $49p_1p_2 - 48p_0p_3, 3p_0p_4 - p_2^2, \ldots, 361p_3p_7 - 128p_5^2$. Inside that surface, our curve is cut out by three quadrics, like $26\,068p_2^2 + 73\,728p_0p_5 + 703\,836p_0p_6 + 234\,612p_2p_6 + 78\,204p_4p_6 + 612\,864p_0p_7 + 212\,268p_2p_7 + 78\,204p_4p_7 - 8379p_7^2$.

Example 26 ($r = 3$, $m = 6$)**.** For any positive integers $\alpha, \beta, \gamma, \varepsilon$, we consider the matrix

$$A = \begin{pmatrix} 0 & \alpha & \beta & 0 & \gamma & \varepsilon \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

The discriminant Δ_A is the *resultant* of two trinomials $x_1 + x_2t^\alpha + x_3t^\beta$ and $x_4 + x_5t^\gamma + x_6t^\varepsilon$. We ran Algorithm 1 for all 138 such matrices with $0 < \alpha < \beta \le 17$, $0 < \gamma < \varepsilon \le 17$, $\gcd(\alpha, \beta) = \gcd(\gamma, \varepsilon) = 1$. The number $n + 2$ of terms of these discriminants equals $2665/138 = 19.31$ on average. Thus a total of 2665 candidate triples $(A, \Delta_A, \mathbf{m})$ were tested in Line 13. Precisely, 93 of these are discriminantal triples. This is only 3.49%.

We now shift gears by looking at polynomials Δ that are multiples of the A-discriminant.

Example 27 ($r = 1$, $m = 4$)**.** We saw in Examples 2 and 20 that interesting models arise from the matrix $A = (1\ 1\ \cdots\ 1)$ whose toric variety is just a point. Any homogeneous multiple Δ of the linear form $\Delta_A = x_1 + x_2 + \cdots + x_m$ can be the input in Line 1 of Algorithm 1. Here, taking $\Delta = \Delta_A$ results in the model given by the full simplex Δ_{m-2}.

Let $m = 4$ and abbreviate $x^a = x_1^{a_1}x_2^{a_2}x_3^{a_3}x_4^{a_4}$ and $|a| = a_1+a_2+a_3+a_4$ for $a \in \mathbb{N}^4$. We conducted experiments with two families of multiples. The first uses binomial multipliers:

$$\Delta = \bigl(x^a + x^b\bigr)\Delta_A \quad \text{or} \quad \Delta = \bigl(x^a - x^b\bigr)\Delta_A,$$

where $|a| = |b| \in \{1, 2, \ldots, 8\}$ and $\gcd(x^a, x^b) = 1$. This gives 1028 polynomials Δ. The numbers of polynomials of degree 2, 3, 4, 5, 6, 7, 8, 9 is 6, 21, 46, 81, 126, 181, 246, 321. For the second family, we use the trinomial multiples

$$\Delta = \bigl(x^a + x^b + x^c\bigr)\Delta_A \quad \text{or} \quad \Delta = \bigl(x^a + x^b - x^c\bigr)\Delta_A,$$

where $|a| = |b| = |c| \in \{1, 2, 3\}$ and $\gcd(x^a, x^b, x^c) = 1$. Each list contains 4 quadrics, 104 cubics and 684 quartics. We report our findings in Table 1.

All 12 Horn pairs in the first family represent the same model, up to permuting coordinates. All are coming from the six quadrics of the family. The model is the surface in Δ_4 defined by the 2×2 minors

Table 1. Horn pairs from families of multiples of $\Delta_A = x_1 + \cdots + x_m$

Family	Pairs $(\Delta, \mathbf{m})$	Horn pairs	Percentage
$(x^a - x^b)\Delta_A$	8212	12	0.15%
$(x^a + x^b)\Delta_A$	8218	0	0%
$(x^a + x^b - x^c)\Delta_A$	8678	8	0.01%
$(x^a + x^b + x^c)\Delta_A$	8968	0	0%

of the matrix $\left(\begin{smallmatrix} p_0 & p_1 & p_2 \\ p_0+p_1+p_2 & p_3 & p_4 \end{smallmatrix}\right)$. This is a staged tree model similar to Example 2, but now with three choices at each blue node instead of two. The eight Horn pairs in the third family represent two distinct models. Four of the eight Horn pairs represent a surface in Δ_5 and the rest represent a surface in Δ_6.

Our construction of models with rational MLE starts with families where r and m are fixed. However, as the entries of the matrix A go up, the number $n+1$ of states increases. This suggests the possibility of listing all models for fixed values of n. Is this list finite?

Problem. Suppose that n is fixed. Are there only finitely many models with rational MLE in the simplex Δ_n? Can we find absolute bounds, depending only on n, for the dimension, degree and number of ideal generators of the associated varieties in $\mathbb{P}^n$?

Algorithm 1 is a tool for studying these questions experimentally. At present, we do not have any clear answers, even for $n = 3$, where the models are curves in a triangle.

Acknowledgments

The first author was supported by the Deutsche Forschungsgemeinschaft DFG under grant 314838170, GRK 2297 MathCoRe.

Eliana Duarte: Partial affiliation with Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig. Bernd Sturmfels: Partial affiliation with University of California, Berkeley.

References

[1] Ay, N., Jost, J., Lê, H.V. and Schwachhöfer, L. (2017). *Information Geometry*. Cham: Springer. MR3701408 https://doi.org/10.1007/978-3-319-56478-4

[2] Clarke, P. and Cox, D.A. (2020). Moment maps, strict linear precision, and maximum likelihood degree one. *Adv. Math*. **370** 107233. MR4103774 https://doi.org/10.1016/j.aim.2020.107233

[3] Collazo, R.A., Görgen, C. and Smith, J.Q. (2018). *Chain Event Graphs*. *Chapman & Hall/CRC Computer Science and Data Analysis Series*. Boca Raton, FL: CRC Press. MR3752634

[4] Drton, M., Sturmfels, B. and Sullivant, S. (2009). *Lectures on Algebraic Statistics*. *Oberwolfach Seminars* **39**. Basel: Birkhäuser. MR2723140 https://doi.org/10.1007/978-3-7643-8905-5

[5] Duarte, E. and Görgen, C. (2020). Equations defining probability tree models. *J. Symbolic Comput*. **99** 127–146. MR4051945 https://doi.org/10.1016/j.jsc.2019.04.001

[6] Garcia-Puente, L.D. and Sottile, F. (2010). Linear precision for parametric patches. *Adv. Comput. Math*. **33** 191–214. MR2659586 https://doi.org/10.1007/s10444-009-9126-7

[7] Gelfand, I.M., Kapranov, M.M. and Zelevinsky, A.V. (1994). *Discriminants*, *Resultants*, *and Multidimensional Determinants*. *Mathematics*: *Theory & Applications*. Boston, MA: Birkhäuser, Inc. MR1264417 https://doi.org/10.1007/978-0-8176-4771-1

[8] Görgen, C. and Smith, J.Q. (2018). Equivalence classes of staged trees. *Bernoulli* **24** 2676–2692. MR3779698 https://doi.org/10.3150/17-BEJ940

[9] Grayson, D. and Stillman, M. Macaulay2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/.

[10] Huh, J. (2014). Varieties with maximum likelihood degree one. *J. Algebr. Stat*. **5** 1–17. MR3279951 https://doi.org/10.18409/jas.v5i1.22

[11] Huh, J. and Sturmfels, B. (2014). Likelihood geometry. In *Combinatorial Algebraic Geometry*. *Lecture Notes in Math*. **2108** 63–117. Cham: Springer. MR3329087 https://doi.org/10.1007/978-3-319-04870-3_3

[12] Kapranov, M.M. (1991). A characterization of A-discriminantal hypersurfaces in terms of the logarithmic Gauss map. *Math. Ann*. **290** 277–285. MR1109634 https://doi.org/10.1007/BF01459245

[13] Krasauskas, R. (2002). Toric surface patches. *Adv. Comput. Math.* **17** 89–113. MR1902537 https://doi.org/10.1023/A:1015289823859
[14] Lauritzen, S.L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. New York: Oxford University Press. MR1419991
[15] Silander, T. and Leong, T.-Y. (2013). A dynamic programming algorithm for learning chain event graphs. In *Discovery Science* (J. Fürnkranz, E. Hüllermeier and T. Higuchi, eds.). *Lecture Notes in Computer Science* **8140** 201–216. Berlin: Springer.
[16] Smith, J.Q. and Anderson, P.E. (2008). Conditional independence and chain event graphs. *Artificial Intelligence* **172** 42–68. MR2388535 https://doi.org/10.1016/j.artint.2007.05.004
[17] Sullivant, S. (2018). *Algebraic Statistics*. *Graduate Studies in Mathematics* **194**. Providence, RI: Amer. Math. Soc. MR3838364
[18] https://github.com/emduart2/DiscreteStatisticalModelsWithRationalMLE.

Received March 2019 and revised April 2020