Geometry of Higher-Order Markov Chains

Bernd Sturmfels*

Abstract

We determine an explicit Gröbner basis, consisting of linear forms and determinantal quadrics, for the prime ideal of Raftery's mixture transition distribution model for Markov chains. When the states are binary, the corresponding projective variety is a linear space, the model itself consists of two simplices in a cross-polytope, and the likelihood function typically has two local maxima. In the general non-binary case, the model corresponds to a cone over a Segre variety.

1 Introduction

In this note we investigate Adrian Raftery's *mixture transition distribution model* (MTD) from the perspective of algebraic statistics [4, 8]. The MTD model, which was first proposed in [9], has a wide range of applications in engineering and the sciences [10]. The article by Berchtold and Raftery [2] offers a detailed introduction and review.

The point of departure for this project was a conjecture due to Donald Richards [11], stating that the likelihood function of an MTD model can have multiple local maxima. We establish this conjecture for the case of binary states in Proposition 6.

Our main result, to be derived in Section 4, gives an explicit Gröbner basis for the MTD model. Here, both the sequence length and the number of states are arbitrary.

We begin with an algebraic description of the model in [2, 9]. Fix a pair of positive integers l and m, and set $N = m^{l+1} - 1$. We define the statistical model $\mathrm{MTD}_{l,m}$ whose state space is the set $[m]^{l+1}$ of sequences $i_0 i_1 \cdots i_l$ of length $l+1$ over the alphabet $[m] = \{1, 2, \ldots, m\}$. The model has $(m-1)m + l - 1$ parameters, given by the entries of an $m \times m$-transition matrix (q_{ij}) and a probability distribution $\lambda = (\lambda_1, \ldots, \lambda_l)$ on the set $[l] = \{1, 2, \ldots, l\}$ of the hidden states. Thus the parameter space is the product of simplices $(\Delta_{m-1})^m \times \Delta_{l-1}$. The model $\mathrm{MTD}_{l,m}$ will be a semialgebraic subset of the simplex Δ_N. That simplex has its coordinates $p_{i_0 i_1 \cdots i_l}$ indexed by sequences in $[m]^{l+1}$.

The model $\mathrm{MTD}_{l,m}$ is the image of the bilinear map

$$\phi_{l,m} \,:\, (\Delta_{m-1})^m \times \Delta_{l-1} \,\rightarrow\, \Delta_N$$

*Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720, USA, bernd@math.berkeley.edu. This research project was supported in part by the National Science Foundation (DMS-0968882) and the DARPA Deep Learning program (FA8650-10-C-7020).

which is defined by the formula

$$p_{i_0 i_1 \ldots i_{l-1} i_l} \quad = \quad \frac{1}{m^l} \cdot \sum_{j=1}^{l} \lambda_j q_{i_{j-1}, i_l} \tag{1}$$

As is customary in algebraic statistics, we pass to a simpler object of study by considering the Zariski closure $\overline{\mathrm{MTD}}_{l,m}$ of our model in the complex projective space $\mathbb{P}^N$, and we seek to compute the homogeneous prime ideal of all polynomials in the $N+1$ unknowns $p_{i_0 i_1 \ldots i_l}$ that vanish on $\overline{\mathrm{MTD}}_{l,m}$. This particular goal will be reached in our Theorem 8.

The following probabilistic interpretation of the formula (1) makes it evident that $\sum p_{i_0 i_1 \cdots i_l} = 1$ holds on the image of $\phi_{l,m}$. We generate a sequence of length $l+1$ on m states as follows. First we select from the uniform distribution on all m^l sequences $i_0 i_1 \cdots i_{l-1}$ of length l. All that remains is to determine the state i_l in position l. The mixture distribution λ determines which of the earlier states gets used in the transition. With probability λ_j, we select position $j-1$ for that. The character in the last position l is determined from the state i_{j-1} in position $j-1$ using the transition matrix (q_{ij}).

The model $\mathrm{MTD}_{l,m}$ is known to be identifiable [2, §4.2]. Consequently, the dimension of the projective variety $\overline{\mathrm{MTD}}_{l,m}$ is equal to the number $(m-1)m+l-1$ of model parameters. A geometric characterization of this variety will be given in Corollary 11.

Equations defining Markov chains and Hidden Markov Models have received considerable attention in algebraic statistics [3, 5, 6, 12]. We contribute to this literature by studying the algebraic geometry of a fundamental model for higher order Markov chains. In addition to our theoretical results in Theorems 1 and 8, readers from statistics will find in Section 3 an analysis of the behavior of the EM algorithm for binary MTD models.

2 Binary States

Our first result concerns the geometry of the model in the case $m = 2$ of binary states.

Theorem 1. *The variety $\overline{\mathrm{MTD}}_{l,2}$ is a linear subspace of dimension $l+1$ in the projective space $\mathbb{P}^N$. This variety intersects the probability simplex Δ_N in a regular cross-polytope of dimension $l+1$. The model $\mathrm{MTD}_{l,2}$ is the union of two $(l+1)$-simplices spanned by vertices of the cross-polytope $\overline{\mathrm{MTD}}_{l,2} \cap \Delta_N$. The two simplices meet along a common edge.*

The *cross-polytope* is the free object in the category of centrally symmetric polytopes [13]. It can be represented as the convex hull of all signed unit vectors e_i and $-e_i$ where $i = 0, 1, \ldots, l$, so it is an $(l+1)$-dimensional polytope with $2l+2$ vertices and 2^{l+1} facets.

Before we come to the proof Theorem 1, let us first see some examples to illustrate it. In what follows we abbreviate the model parameters by $q_{11} = a$, $q_{21} = b$ and $\lambda_2 = \lambda$.

Example 2. Theorem 1 also applies in the trivial case $l = 1$, where (1) reads

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \quad = \quad \begin{pmatrix} a/2 & (1-a)/2 \\ b/2 & (1-b)/2 \end{pmatrix}. \tag{2}$$

The variety $\overline{\mathrm{MTD}}_{1,2}$ is the plane in $\mathbb{P}^3$ given by $p_{11}+p_{12}=p_{21}+p_{22}$. Its intersection with the tetrahedron Δ_3 coincides with the model $\mathrm{MTD}_{1,2}$, which is a regular square:

$$\mathrm{MTD}_{1,2} \;=\; \overline{\mathrm{MTD}}_{1,2} \cap \Delta_3 \;=\; \mathrm{conv}\left\{ \begin{pmatrix} 1/2 & 0 \\ 1/2 & 0 \end{pmatrix}, \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}, \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1/2 \\ 0 & 1/2 \end{pmatrix} \right\}.$$

The first three and last three matrices in this list form the two triangles referred to in Theorem 1. Their common edge consists of all transition matrices (2) of rank 1. □

Example 3. Our first non-trivial example arises for $l=m=2$. The map $\phi_{2,2}$ is given by

$$(a,b,\lambda) \mapsto p = \frac{1}{4}\begin{bmatrix} ae_{111}+(\lambda b+(1-\lambda)a)e_{121}+(\lambda a+(1-\lambda)b)e_{211}+be_{221}+(1-a)e_{112} \\ +(\lambda(1-b)+(1-\lambda)(1-a))e_{122}+(\lambda(1-a)+(1-\lambda)(1-b))e_{212}+(1-b)e_{222} \end{bmatrix}$$

Here $\{e_{111}, e_{112}, \ldots, e_{222}\}$ denotes the standard basis in the space of $2\times 2\times 2$-tensors. The variety $\overline{\mathrm{MTD}}_{2,2}$ is the 3-dimensional linear subspace of $\mathbb{P}^7$ defined by

$$\begin{aligned} p_{111}+p_{112}=p_{121}+p_{122}, \quad p_{211}+p_{212}=p_{221}+p_{222}, \\ p_{121}+p_{122}=p_{221}+p_{222}, \quad p_{111}+p_{221}=p_{121}+p_{211}. \end{aligned}$$

The intersection of this linear space with the simplex Δ_7 is the regular octahedron whose vertices are the images under $\phi_{2,2}$ of the vertices of the cube $(\Delta_1)^2\times\Delta_1$. The model $\mathrm{MTD}_{2,2}$ consists of two tetrahedra formed by vertices of the octahedron. Their common edge is the segment between $\frac{1}{4}(e_{111}+e_{121}+e_{211}+e_{221})$ and $\frac{1}{4}(e_{112}+e_{122}+e_{212}+e_{222})$. □

Example 4. The statement of Theorem 1 does not extend to $m\geq 3$. Consider the case $l=2, m=3$. The 7-dimensional variety $\overline{\mathrm{MTD}}_{2,3}$ lives in $\mathbb{P}^{26}$, and it is not a linear space. The linear span of $\overline{\mathrm{MTD}}_{2,3}$ is 10-dimensional. Inside this $\mathbb{P}^{10}$, the variety $\overline{\mathrm{MTD}}_{2,3}$ has codimension 3, degree 4, and it is cut out by six quadrics. In Example 10 we shall display a Gröbner basis consisting of 16 linear forms and six quadrics for its prime ideal. □

Proof of Theorem 1. It is known by [2, §4.2] that the model is identifiable, so $\mathrm{MTD}_{l,2}$ is a semi-algebraic set of dimension $l+1$ in Δ_N. Its Zariski closure $\overline{\mathrm{MTD}_{l,2}}$ is a variety of dimension $l+1$ in $\mathbb{P}^N$. That variety is irreducible because it is defined by way of a rational parametrization. For any binary sequence $i_0 i_1 \cdots i_{l-1}$, the identity

$$p_{i_0 i_1 \cdots i_{l-1} 2} \;=\; 2^{-l} - p_{i_0 i_1 \cdots i_{l-1} 1} \tag{3}$$

holds on $\mathrm{MTD}_{l,2}$, so it suffices to consider relations on probabilities of sequences that end with 1. On our model, these probabilities satisfy the linear equations

$$p_{i_0 i_1 \cdots i_r \cdots i_s \cdots i_{l-1} 1} + p_{i_0 i_1 \cdots \tilde{i}_r \cdots \tilde{i}_s \cdots i_{l-1} 1} \;=\; p_{i_0 i_1 \cdots i_r \cdots \tilde{i}_s \cdots i_{l-1} 1} + p_{i_0 i_1 \cdots \tilde{i}_r \cdots i_s \cdots i_{l-1} 1}. \tag{4}$$

In other words, the l-dimensional $2\times 2\times\cdots\times 2$-tensor $(p_{i_0 i_1 \cdots i_{l-1} 1})$ has tropical rank 1. The set of such tensors is a classical linear space of dimension $l+1$.

Solving the linear equations (3) and (4) on the simplex Δ_N, we obtain an $(l+1)$-dimensional polytope P that contains the model $\mathrm{MTD}_{l,2}$. Its Zariski closure in $\mathbb{P}^N$ is an $(l+1)$-dimensional linear space that contains the variety $\overline{\mathrm{MTD}_{l,2}}$. Being irreducible varieties of the same dimension, they must be equal. This proves the first assertion.

We next claim that the polytope P of all non-negative real solutions to (3) and (4) is a regular cross-polytope. For $r \in \{0, 1, \ldots, l-1\}$ and $s \in \{1, 2\}$ define the $2l$ points

$$E_{rs} \;=\; \frac{1}{2^l} \cdot \left[\sum \{\, e_{i_0 i_1 \cdots i_{l-1} 1} \,|\, i_r = s \,\} \,+\, \sum \{\, e_{i_0 i_1 \cdots i_{l-1} 2} \,|\, i_r \neq s \,\} \right] \quad \in \; \Delta_N.$$

These are extreme non-negative solutions of (3) and (4). They form the vertices of an l-dimensional cross-polytope, since $\frac{1}{2}(E_{r1} + E_{r2})$ is equal to the uniform distribution $\frac{1}{2^{l+1}} e_{++\cdots++}$ for all r. In addition to the $2l$ vertices E_{rs}, the polytope P has two more vertices, namely, $\frac{1}{2^l} e_{++\cdots+1}$ and $\frac{1}{2^l} e_{++\cdots+2}$. Hence P is a bipyramid over the l-dimensional cross-polytope, so it is an $(l+1)$-dimensional cross-polytope.

It remains to identify the model $\mathrm{MTD}_{l,2}$ inside P. The parameter polytope is the product $(\Delta_1)^2 \times \Delta_{l-1}$, and, as before, we chose coordinates (a, b) on the square $(\Delta_1)^2$. The map $\phi_{l,2}$ contracts the simplex $\{(0,0)\} \times \Delta_{l-1}$ onto the vertex $\frac{1}{2^l} e_{++\cdots+2}$ of P, and it contracts the simplex $\{(1,1)\} \times \Delta_{l-1}$ onto the vertex $\frac{1}{2^l} e_{++\cdots+1}$ of P. The vertex $(0,1) \times e_r$ is mapped to the vertex $E_{r,2}$, and the vertex $(1,0) \times e_r$ is mapped to the vertex $E_{r,1}$. The parameter points with $a = b$ are contracted onto the line segment $S = [\frac{1}{2^l} e_{++\cdots+1}, \frac{1}{2^l} e_{++\cdots+2}]$. The parameter points with $a < b$ are mapped bijectively onto the $(l+1)$-simplex formed by S and $\{E_{0,2}, E_{1,2}, \ldots, E_{l-1,2}\}$, but with S removed. The parameter points with $a > b$ are mapped bijectively onto the $(l+1)$-simplex formed by S and $\{E_{0,1}, E_{1,1}, \ldots, E_{l-1,1}\}$, but with S removed. Hence $\mathrm{MTD}_{l,2}$ equals the union of two $(l+1)$-simplices glued along the special diagonal S of the cross-polytope P. □

Corollary 5. *For large l, there are far fewer distributions in the model $\mathrm{MTD}_{l,2}$ than distributions in its Zariski closure. Namely, with respect to Lebesgue measure, we have*

$$\frac{\mathrm{vol}(\mathrm{MTD}_{l,2})}{\mathrm{vol}(\overline{\mathrm{MTD}}_{l,2} \cap \Delta_N)} \;=\; \frac{1}{2^{l-1}}.$$

Proof. We can triangulate the cross-polytope P into 2^l simplices, all of the same volume and containing the special diagonal S. The model $\mathrm{MTD}_{l,2}$ consists of two of them. Hence $2/2^l$ is the fraction of the volume of $P = \overline{\mathrm{MTD}}_{l,2} \cap \Delta_N$ that is occupied by $\mathrm{MTD}_{l,2}$. □

3 Likelihood inference

We next discuss maximum likelihood estimation (MLE) for the mixture transition distribution model $\mathrm{MTD}_{l,m}$. Any data set is represented by a function $u : [m]^{l+1} \to \mathbb{N}$ that records the frequency counts of the observed sequences. Given such a function u, our objective is to maximize the corresponding log-likelihood function

$$L_u \quad = \quad \sum_{i_0 i_1 \cdots i_l} u_{i_0 i_1 \cdots i_l} \cdot \log(p_{i_0 i_1 \cdots i_l}) \tag{5}$$

over all probability distributions that lie in the model $\mathrm{MTD}_{l,m}$. A standard method for solving this optimization problem is the expectation-maximization (EM) algorithm. Other algorithms for the same task can be found in [1, 10].

A general version of the EM algorithm for algebraic models with discrete data is described in [8, §1.3], while the specific case of the MTD model is treated in [2, §4.5]. Richards [11] conjectured that the EM algorithm for the MTD model may get stuck in local maxima. Our next result confirms that this is indeed the case, even for $m = 2$.

Proposition 6. *The log-likelihood function L_u on the binary model $\mathrm{MTD}_{l,2}$ has either one or two local maxima. With probability one, there will be two local maxima, and both of these will be reached by the EM algorithm for different choices of initial parameters.*

Here the statement about "probability one" in the second sentence refers to any absolutely continuous probability distribution that is positive on the simplex Δ_N.

Proof. We saw in Theorem 1 that $\mathrm{MTD}_{l,2}$ is the union of two convex polytopes. The log-likelihood function L_u is strictly concave on the ambient simplex Δ_N, so it attains a unique maximum on each of the two polytopes. This proves the first statement.

For the second statement consider the empirical distribution $u/|u|$ which is a point in Δ_N. Its log-likelihood function L_u has a unique maximum p^* in the interior of the cross-polytope P. With probability one, this maximum p^* will not lie in the segment S, so let us assume that this is the case. Then either p^* lies in precisely one of the two $(l+1)$-simplices that make up $\mathrm{MTD}_{l,2}$, or p^* does not lie in $\mathrm{MTD}_{l,2}$. In the former case, p^* is the MLE, and the maximum over the other simplex is in the boundary of that simplex and constitutes a second local maximum. In the latter case, each of the two simplices has a local maximum in its boundary. When choosing starting parameter values near either of these local maxima, the EM algorithm converges to that local maximum. □

The point p^* in the cross-polytope P at which L_u attains its maximum is an algebraic function of the data u. The degree of this algebraic function is the *ML degree* (see [7]) of the linear subvariety $\overline{\mathrm{MTD}}_{l,2}$ of $\mathbb{P}^N$. By Varchenko's Formula [8, Theorem 1.5], this ML degree coincides with the number of bounded regions in an arrangement of hyperplanes. This arrangement lives inside the affine space that is cut out by (3) and (4) and it consists of the restrictions of the coordinate hyperplanes $\{p_\bullet = 0\}$.

Computations show that the ML degree equals 9 for $l = 3$, and it equals 209 for $l = 4$. It would be interesting to find a general formula for that ML degree as a function of l.

The local maxima that occur on the boundary of the two simplices of $\mathrm{MTD}_{l,2}$ have ML degree 1, that is, they are expressed as rational functions in the data u. Indeed, these local maxima are precisely the estimates for the Markov chain obtained by fixing $\lambda_i = 1$ for some i. Hence, if $p^* \notin \mathrm{MTD}_{l,2}$, then the MLE is a rational expression in u. The next example illustrates the behavior of the EM algorithm for $m = 2$ and $l = 3$.

Example 7. The data consists of eight positive integers, here written as a matrix

$$U \;=\; \begin{pmatrix} u_{111} & u_{121} & u_{211} & u_{221} \\ u_{112} & u_{122} & u_{212} & u_{222} \end{pmatrix}.$$

The MLE $\hat{p}$ will be either

$$p' \;=\; \frac{1}{2|u|}\begin{pmatrix} u_{111}+u_{211} & u_{121}+u_{221} & u_{111}+u_{211} & u_{121}+u_{221} \\ u_{112}+u_{212} & u_{122}+u_{222} & u_{112}+u_{212} & u_{122}+u_{222} \end{pmatrix}$$

or

$$p'' \quad = \quad \frac{1}{2|u|}\begin{pmatrix} u_{111}+u_{121} & u_{111}+u_{121} & u_{211}+u_{221} & u_{211}+u_{221} \\ u_{112}+u_{122} & u_{112}+u_{122} & u_{212}+u_{222} & u_{212}+u_{222} \end{pmatrix},$$

or it will be the unique probability distribution satisfying (3), (4), and

$$\operatorname{rank}\begin{pmatrix} u_{111} & u_{112} & u_{121} & u_{122} & u_{211} & u_{212} & u_{221} & u_{222} \\ p_{111} & p_{112} & p_{121} & p_{122} & p_{211} & p_{212} & p_{221} & p_{222} \\ p_{111} & p_{112} & -p_{121} & -p_{122} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{211} & p_{212} & -p_{221} & -p_{222} \\ 0 & 0 & p_{121} & p_{122} & 0 & 0 & -p_{221} & -p_{222} \\ p_{111} & 0 & -p_{121} & 0 & -p_{211} & 0 & p_{221} & 0 \end{pmatrix} \leq 5. \qquad (6)$$

This is the matrix denoted $\begin{bmatrix} u \\ \tilde{J} \end{bmatrix}$ in [7, §3]. The rank constraint (6) represents Proposition 2 in [7]. The unique probability distribution that lies in our model and also satisfies (6) was called p^* in the proof of Proposition 6. Its defining constraints (3), (4) and (6) form a system of polynomial equations that has 9 complex solutions. The distribution p^* is the unique solution to that system whose coordinates are both real and positive.

The trichotomy in this example is best explained by the following observations: For almost all data matrices U, the three points p', p'', p^* are distinct, one of them coincides with the global maximum $\hat{p}$ of L_u over $\mathrm{MTD}_{l,2}$, and another one is a local maximum. □

It would be interesting to extend the findings in Proposition 6 to $m \geq 3$. The algebraic tools that may be needed for such an analysis are developed in the next section.

4 Non-linear Models

In this section we examine the geometry of model $\mathrm{MTD}_{l,m}$ and the variety $\overline{\mathrm{MTD}}_{l,m}$ for an arbitrary number m of states. In particular, we prove that its prime ideal is minimally generated by linear forms and quadrics. These minimal generators form a Gröbner basis.

Theorem 8. *The variety $\overline{\mathrm{MTD}}_{l,m}$ spans a linear space of dimension $(m-1)(lm-l+1)$ in $\mathbb{P}^N$. In this linear space, its prime ideal is given by the 2×2-minors of an $l\times(m-1)^2$-matrix of linear forms. The linear and quadratic ideal generators form a Gröbner basis.*

This theorem explains our earlier result that the model is linear for binary states. Indeed, for $m = 2$, the dimension $(m-1)m+l-1$ of the model coincides with the dimension $(m-1)(lm-l+1)$ of the ambient linear space, and there are no 2×2-minors.

Proof. We shall present an explicit Gröbner basis consisting of linear forms and quadrics. The term order we choose is the reverse lexicographic term order induced by the lexicographic order on the states $i_0 i_1 \cdots i_l$ of the model. We first consider the linear relations

$$\underline{p_{i_0 i_1 i_2 \cdots i_{l-1} i_l}} - \sum_{j=0}^{l-1} p_{m\cdots m i_j m \cdots m i_l} + (l-1) p_{mm\cdots mmi_l}. \qquad (7)$$

This linear form is non-zero and has the underlined leading term if and only if at least two of the entries of the l-tuple $(i_0, i_1, \ldots, i_{l-1})$ are not equal to m. Thus the number of distinct Gröbner basis elements (7) equals $m^{l+1} - m(1 + l(m-1))$.

Our second class of Gröbner basis elements consists of the linear relations

$$\begin{array}{rc} & \underline{p_{m\cdots mi_jm\cdots m1}} + p_{m\cdots mi_jm\cdots m2} + \cdots + p_{m\cdots mi_jm\cdots mm} \\ - & p_{m\cdots mmm\cdots m1} - p_{m\cdots mmm\cdots m2} - \cdots - p_{m\cdots mmm\cdots mm}. \end{array} \tag{8}$$

These linear forms are non-zero with the underlined leading term provided $0 \leq j \leq l-1$ and $1 \leq i_j \leq m-1$. The number of distinct linear forms (8) equals $l(m-1)$, and the set of their leading terms is disjoint from the set of leading terms in (7).

The number of unknowns $p_\bullet$ not yet underlined equals $l(m-1)^2+(m-1)+1$. We use these unknowns to form $m-1$ matrices $A_2, A_3, \ldots, A_m$, each having format $l \times (m-1)$, as follows. Define the matrix A_r by placing the following entry in row j and column i_j:

$$\underline{p_{m\cdots mi_jm\cdots mr}} \;-\; p_{m\cdots mmm\cdots mr}. \tag{9}$$

We finally form an $l \times (m-1)^2$ matrix by concatenating these $m-1$ matrices:

$$A \;=\; \big(\, A_2\, A_3 \;\cdots\; A_m\big). \tag{10}$$

The third and last group of polynomials in our Gröbner basis is the set of 2×2-minors of A. The entries of A have distinct leading terms, underlined in (9), and the leading term of each 2×2-minor is the product of the leading terms on the main diagonal.

Note that we could also define the matrix A_1 and include it when forming (10). This would not change the ideal, but it would lead to a generating set that is not minimal.

It is well-known that the 2×2-minors of a matrix of unknowns form a Gröbner basis for the prime ideal they generate. Since no unknown $p_\bullet$ underlined in (7) or (8) appears in the matrix A, it follows that these linear relations together with the 2×2-minors of (10) generate a prime ideal and form a Gröbner basis for that prime ideal.

The ideal of 2×2 minors of A has codimension $l(m-1)^2 - l - (m-1)^2 + 1$. Subtracting this quantity from the number $l(m-1)^2+(m-1)+1$ of unknowns not underlined in (7) or (8), we obtain $l + (m-1)^2 - 1 + (m-1) + 1 \;=\; (m-1)m + l$. This is the dimension of the affine variety defined by our prime ideal. The corresponding irreducible projective variety has dimension $(m-1)m + l - 1$. This is precisely the dimension of $\overline{\mathrm{MTD}}_{l,m}$.

It hence suffices to prove that our variety contains the model $\mathrm{MTD}_{l,m}$, or, equivalently, that the linear forms (7) and (8) are mapped to 0 by the parameterization (1), and that the specialized matrix $\phi_{l,m}(A)$ has rank 1. For (8) this is obvious because, for fixed i_j,

$$\sum_{r=1}^{m} \phi_{l,m}^*\big(p_{m\cdots mi_jm\cdots mr}\big) \;=\; \frac{1}{m^l}.$$

Here $\phi_{l,m}^*$ denotes the homomorphism of polynomial rings induced by the map $\phi_{l,m}$.

The indices of the unknowns in the linear form (7) all have the same letter i_l in the end. The formula (1) for the corresponding probabilities can thus be written as

$$\phi_{l,m}^*(p_{i_0 i_1 \cdots i_{l-1} i_l}) \;=\; u + x_{i_0} + y_{i_1} + \cdots + z_{i_{l-1}}.$$

In other words, for any fixed i_l, the resulting l-dimensional tensor has tropical rank 1. This representation implies linear relations like (4), and these are equivalent to (7).

Finally, if we apply our ring homomorphism to (9) then we get

$$\phi^*_{l,m}(p_{m\cdots mi_jm\cdots mr}) \,-\, \phi^*_{l,m}(p_{m\cdots mmm\cdots mr}) \;=\; \lambda_j \cdot (q_{i_j,r} - q_{m,r}). \tag{11}$$

Thus, the matrix $\phi_{k,l}(A)$ is the product of the column vector $(\lambda_1, \ldots, \lambda_l)$ and a row vector of length $(m-1)^2$ whose entries are $q_{i_j,r} - q_{m,r}$ for $2 \leq r \leq m$ and $1 \leq i_j \leq m-1$. In particular, the matrix $\phi^*_{l,m}(A)$ has rank ≤ 1. This completes the proof of Theorem 8. □

Remark 9. The prime ideal in Theorem 8 is the kernel of $\phi^*_{l,m}$, so it characterizes the image of the model parametrization $\phi_{l,m}$. On the model $\mathrm{MTD}_{l,m}$, the map $\phi_{l,m}$ can be inverted as long as the rows of the transition matrix (q_{ij}) are distinct. Indeed, q_{ij} equals $2^l \phi^*_{l,m}(p_{ii\cdots iij})$, and the coordinates of λ are identified from (11). Thus, our result refines the well-known fact that MTD models are identifiable [2, §4.2].

Example 10. We illustrate Theorem 8 for the case $l = 2, m = 3$, by presenting the Gröbner basis promised in Example 4. Note that $N = 26$. Here the ambient linear space has dimension $(m-1)(lm - l + 1) = 10$, and our Gröbner basis for that linear space consists of twelve linear forms (7) and four linear forms (8). These are respectively,

$$\underline{p_{111}} - p_{311} - p_{131} + p_{331},\ \underline{p_{121}} - p_{321} - p_{131} + p_{331},\ \underline{p_{211}} - p_{311} - p_{231} + p_{331},\ \underline{p_{221}} - p_{321} - p_{231} + p_{331},$$
$$\underline{p_{112}} - p_{312} - p_{132} + p_{332},\ \underline{p_{122}} - p_{322} - p_{132} + p_{332},\ \underline{p_{212}} - p_{312} - p_{232} + p_{332},\ \underline{p_{222}} - p_{322} - p_{232} + p_{332},$$
$$\underline{p_{113}} - p_{313} - p_{133} + p_{333},\ \underline{p_{123}} - p_{323} - p_{133} + p_{333},\ \underline{p_{213}} - p_{313} - p_{233} + p_{333},\ \underline{p_{223}} - p_{323} - p_{233} + p_{333}.$$

and
$$\underline{p_{311}} + p_{312} + p_{313} - p_{331} - p_{332} - p_{333}\,,\ \underline{p_{321}} + p_{322} + p_{323} - p_{331} - p_{332} - p_{333}\,,$$
$$\underline{p_{131}} + p_{132} + p_{133} - p_{331} - p_{332} - p_{333}\,,\ \underline{p_{231}} + p_{232} + p_{233} - p_{331} - p_{332} - p_{333}.$$

The remaining $l(m-1)^2 + (m-1) + 1 = 8 + 2 + 1 = 11$ not yet underlined unknowns are $p_{132}, p_{232}, p_{312}, p_{322}, p_{133}, p_{233}, p_{313}, p_{323}, p_{332}, p_{333}, p_{331}$. These represent coordinates on the linear subspace $\mathbb{P}^{10}$ of $\mathbb{P}^{26}$ that is cut out by these linear forms. Inside that linear subspace $\mathbb{P}^{10}$, our variety $\overline{\mathrm{MTD}}_{2,3}$ has codimension 3, and it is defined ideal-theoretically by the 2×2-minors of the 2×4-matrix

$$A \;\;=\;\; \begin{pmatrix} A_2 & A_3 \end{pmatrix} \;=\; \begin{pmatrix} \underline{p_{132}} - p_{332} & \underline{p_{232}} - p_{332} & \underline{p_{133}} - p_{333} & \underline{p_{233}} - p_{333} \\ \underline{p_{312}} - p_{332} & \underline{p_{322}} - p_{332} & \underline{p_{313}} - p_{333} & \underline{p_{323}} - p_{333} \end{pmatrix}.$$

These six quadrics, together with the 16 linear forms, form a reduced Gröbner basis. □

Our proof of Theorem 8 gives rise to the following geometric description:

Corollary 11. *The projective variety $\overline{\mathrm{MTD}}_{l,m}$ is a cone with base $\mathbb{P}^{m-1}$ over the Segre variety $\mathbb{P}^{l-1} \times \mathbb{P}^{m^2-2m}$. If $m \geq 3$, then this variety is singular and its singular locus is the $\mathbb{P}^{m-1}$ that forms the base of that cone. The degree of $\overline{\mathrm{MTD}}_{l,m}$ equals $\binom{l+(m-1)^2-2}{l-1}$.*

Proof. The ideal of singular locus of $\overline{\mathrm{MTD}}_{l,m}$ is generated by the entries of the matrix A together with the linear forms (7) and (8). Together, these linear equations are equivalent to requiring that the value of $p_{i_0 i_1 \cdots i_{l-1} r}$ depends only on r. It does not on $i_0 i_1 \cdots i_{l-1}$. These constraints define a linear space $\mathbb{P}^{m-1}$ in $\mathbb{P}^N$. The 2×2-minors of an $l \times (m-1)^2$ matrix define the Segre variety $\mathbb{P}^{l-1} \times \mathbb{P}^{m^2-2m}$, whose degree is known to be the binomial coefficient. □

References

[1] A. Berchtold: Estimation in the mixture transition distribution model, *J. Time Ser. Anal.* **22** (2001) 379–397.

[2] A. Berchtold and A. Raftery: The mixture transition distribution model for high-order Markov chains and non-Gaussian time series, *Statistical Science* **17** (2002) 328–356.

[3] A. Critch: Binary hidden Markov models and varieties, `arXiv:1206.0500`.

[4] M. Drton, B. Sturmfels and S. Sullivant: *Lectures on Algebraic Statistics*, Oberwolfach Seminars **39**, Birkhäuser Verlag, Basel, 2009.

[5] H. Hara and A. Takemura: A Markov basis for two-state toric homogeneous Markov chain model without initial parameters, *J. Japan Statist. Soc.* **41** (2011) 33–49.

[6] D. Haws, A. Martin Del Campo and R. Yoshida: Degree bounds for a minimal Markov basis for the three-state toric homogeneous Markov chain model, in T. Hibi: *Harmony of Gröbner Bases and the Modern Industrial Society*, 2012, pp. 63–98.

[7] S. Hoşten, A. Khetan and B. Sturmfels: Solving the likelihood equations, *Foundations of Computational Mathematics* **5** (2005) 389–407.

[8] L. Pachter and B. Sturmfels: *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.

[9] A. Raftery: A model for high-order Markov chains, *J. Roy. Statist. Soc. Ser. B* **47** (1985) 528–539.

[10] A. Raftery and S. Taveré: Estimation and modelling repeated patterns in high order Markov chains with the Mixture Transition Distribution Model, *Applied Statistics* (1994) 179–199.

[11] D. Richards: Counting and locating the solutions of polynomial systems of ML equations, presentation at the International Workshop in Applied Probability, University of Connecticut, 2006.

[12] A. Schönhuth: Generic identification of binary-valued hidden Markov processes, `arXiv:1101.3712`.

[13] G. Ziegler: *Lectures on Polytopes*, Graduate Texts in Mathematics, 152, Springer-Verlag, New York, 1995.