Algebraic Geometry of Hidden Markov and Related Models

by

Andrew James Critch

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bernd Sturmfels, Chair
Professor Martin Olsson
Professor Lior Pachter
Professor Uros Seljak

Spring 2013

Algebraic Geometry of Hidden Markov and Related Models

Abstract

Algebraic Geometry of Hidden Markov and Related Models

by

Andrew James Critch

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Bernd Sturmfels, Chair

This thesis embodies a collection of algebraic techniques and results on hidden Markov models, related models in quantum physics called Matrix Product State models, and finally discrete directed acyclic graphical models.

Chapter 1 explores the statistical problems of model selection and parameter identifiability from the perspective of algebraic geometry, in the case of hidden Markov models (HMMs) where all the hidden random variables are binary. Its main contributions are (1) a new parametrization for every such HMM via a birational map with an explicit inverse for recovering the hidden parameters in terms of observables, (2) a semialgebraic model membership test to determine if a discrete probability distribution can arise from such an HMM, and (3) minimal defining equations for the set of probability distributions arising from the 4-node fully binary model, comprising 21 quadrics and 29 cubics, which were computed using Gröbner bases in the cumulant coordinates of Bernd Sturmfels and Piotr Zwiernik. The new model parameters in (1) are rationally identifiable in the sense of Seth Sullivant, Luis David Garcia-Puente, and Sarah Spielvogel, and each model's Zariski closure is therefore a rational projective variety of dimension 5. Gröbner basis computations for the model and its graph are found to be considerably faster using these parameters. In the case of two hidden states, (2) supersedes a previous algorithm of Alexander Schönhuth which is only generically defined, and the defining equations (3) yield new invariants for HMMs of all lengths ≥ 4. Such invariants have been used successfully in model selection problems in phylogenetics, and one can hope for similar applications in the case of HMMs.

In Chapter 2, we study the representational power of matrix product states (MPS) with binary virtual bonds for entangled qubit systems. We do this by giving polynomial expressions in a pure quantum state's amplitudes which hold if and only if the

state is a translation invariant matrix product state or a limit of such states. For systems with few qubits, we give these equations explicitly, considering both periodic and open boundary conditions. Using the classical theory of trace varieties and trace algebras, we explain the relationship between MPS and hidden Markov models and exploit this relationship to derive useful parameterizations of MPS. We present four conjectures on the identifiability of MPS parameters.

Chapter 3 develops new parameters for use with directed acyclic graphical (DAG) models on discrete variables, which can simplify symbolic computations for tree models with hidden variables having more than two states. This development is the first step toward generalizing work of Smith and Zwiernik on binary trees, and makes it possible for some of the techniques used in Chapters 1 and 2 to be applied to graphical models with variables having more than two states.

To My Family

For always caring, and always believing in me.

Contents

List of Figures

Acknowledgments

First, I would like to thank my advisor, Professor Bernd Sturmfels, for taking me as his student, and for being so nurturing of my evolving research interests. His willingness to talk, and genuine curiosity about my work kept me perpetually motivated and interested to continue, in a way I did not think previously possible. After less than two years of working with him, it is easy to say that I would not be the mathematician I am today without him.

I am grateful to Professor Martin Olsson, my previous PhD advisor, for all that he taught me about the world of arithmetic geometry and stacks, and for also encouraging me to pursue my interests in applied mathematics as I became certain of them. From Professor Olsson, I learned what it looks like to develop a new field, and how to be very clear and careful about what things mean. I never thought the kind of precision and abstraction he employed in his thinking and teaching would carry over so pervasively, as it has, in how I think about the applied world. Although time is always limited, I wish I could have learned more from him in my time at Berkeley.

I must thank Professor Arthur Ogus for teaching me algebraic geometry in such a clear and beautiful fashion; he surely sealed that deal that I would love and learn the subject. I also thank Professor David Eisenbud for making me feel so instantly welcome in the algebraic geometry community at Berkeley.

I have been very fortunate to have as my coauthors Jason Morton, Shaowei Lin, Piotr Zwiernik, and Luca Weihs. Jason was inspiring to me as a coauthor, and has given me a great deal of helpful career advice. Shaowei, who was my mentor during his time as a postdoc at Berkeley, helped me to clarify many of my earliest thoughts about the meaning of algebraic statistics, and as a friend and colleague gave me many reasons to continue with my research. Piotr was often incredibly patient and helpful in explaining his work to me when I had questions about it. Luca's early Macaulay2 experiments with tensor cumulants assured us all that they were worth studying.

My friends John DeIonno, Nisan Stiennon, David Kayvanfar, Marieke Kleemans, Mahendra Prasad, Anna Salamon, Mike Pacer, and Sebastian Benthall have been especially supportive and encouraging. They helped me to realize what I am passionate about. They never pressured me to abandon work when I was making progress, and were always there for a visit or a break when I needed one.

Finally, I would be nothing without my loving family. Throughout my childhood, they always made time for me to learn whatever I liked, and have tolerated many work-laden visits home since then. They never pressured or even asked me to choose any particular career, only to be sure that I am happy.

Thank you, everyone!

Chapter 0

Introduction

Algebraic statistics is the application of commutative algebra and algebraic geometry to the study of statistical models. Linear models in science and statistics — literally, equations and formulae which take a linear form — are constantly studied and manipulated using the techniques from linear algebra, such as matrix inversion, Gaussian elimination, singular value decomposition, hyperplane arrangements, and so on. Commutative algebra and algebraic geometry are together the analogue of linear algebra for studying models which involve quadratic or higher degree polynomials.

Many of the most commonly used statistical models are finite-dimensional families of probability distributions parametrized by polynomials, which are called *algebraic statistical models*. Gaussian models, exponential families, hidden Markov models, phylogenetic tree models, directed and undirected graphical models, structural equation models, and deep belief networks are all algebraic statistical models. For an introduction and overview of some biological applications, see *Algebraic Statistics for Computational Biology*, by Pachter and Sturmfels [28].

As an example, consider a coin A and parameters $a_i = \Pr(A = i)$ for $i = 0, 1$

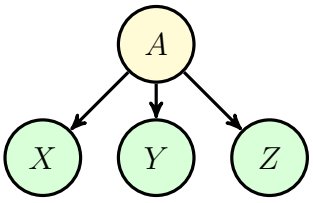


Figure 0.1: A naïve Bayes model with three observables

specifying its distribution. Suppose the outcome of A determines the distributions of three other coins, X, Y, and Z before they are flipped. That is, parameters $x_{ij} = \Pr(X = j|A = i)$, $y_{ik} = \Pr(Y = k|A = i)$, and $z_{i\ell} = \Pr(Z = \ell|A = i)$, determine the *effect* of A on each of the other coins. These dependencies are depicted in Figure 0, called a *causal diagram*. This particular diagram is called a *naïve Bayes model*. When this process runs, each outcome $(A, X, Y, Z) = (i, j, k, \ell)$ has some probability $p_{i,j,k,\ell}$ of occurring, given by the polynomial expression

$$p_{ijk\ell} = \Pr(A = i, X = j, Y = k, Z = \ell) = a_i x_{ij} y_{ik} z_{i\ell}.$$

Since $a_0 + a_1 = 1$ and $x_{j0} + x_{j1} = 1$, etc., if we choose the seven parameters a_1, x_{j1}, y_{k1}, and $z_{\ell 1}$ freely in $[0, 1]$, then the other seven parameters are uniquely determined. As we vary the free parameters, we obtain different probability tables p according to (1), thus defining a polynomial map $\phi : [0, 1]^7 \to \mathbb{R}^{2\times2\times2\times2}$. The set of $2 \times 2 \times 2 \times 2$ probability tables p which can be *explained* or *modeled* as arising from such a causal diagram of coins is hence the image of ϕ.

Many statistical properties of this model translate to algebraic or geometric properties of the map ϕ. For example, pardoning jargon for the moment, we have the following dictionary of non-trivial equivalences:

Algebraic geometry	**Statistics**
ϕ is injective.	The parameters can always be learned with sufficient data.
ϕ has smooth fibres.	The Bayesian Information Criterion (BIC) will accurately penalize this model in model selection algorithms.
The signed topological Euler characteristic of $\overline{\text{image}}(\phi)_+$ is 1.	There is 1 critical point in maximum likelihood estimation from generic data.

Most statistical models have more complicated geometry than this one, and have correspondingly more subtle statistical behavior. Such is the case for hidden Markov models, which I'll describe next.

Hidden Markov models (HMM), the topic of Chapter 1, are machine learning models with diverse applications, including natural language processing, gesture recognition, genomics, and Kalman filtering of physical measurements in robotics and aeronautics. An HMM treats a series of observed phenomena, such as words

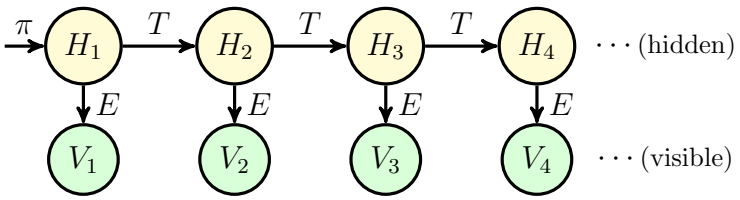


Figure 0.2: An HMM of length 4

being recorded by a microphone, as arising from a series of *hidden* or *unobserved* variables, such as the English text in the mind of a speaker that she is hoping to produce on-screen. An HMM-based learning algorithm updates its past beliefs about hidden variables based on present measurements of observables. For example, if a computer thinks you've said "ice cream", and you then say "loudly", for grammatical reasons, it may update its previous opinion to "I scream" while you are still speaking.

In more detail, an HMM of length n involves n hidden *nodes* (random variables) H_i and n visible nodes V_i. We will write $\mathrm{HMM}(\ell, m, n)$ for an HMM of length n where the hidden nodes have ℓ states and the visible nodes have m states. In any HMM, the variables *affect* each other according to a causal diagram with some parameter matrices π, T, and E, respectively of size $1 \times \ell$, $\ell \times \ell$, and $\ell \times m$. A diagram with $n = 4$ is depicted in Figure 2. The parameter matrices determine how the probabilistic effects work, according to the formulae

$$\pi_i = \Pr(H_i = 0), \quad T_{ij} = \Pr(H_t = j \,|\, H_{t-1} = i), \quad E_{ij} = \Pr(V_t = j \,|\, H_t = i).$$

Given the parameter matrices π, T, and E, any particular m-ary string $v = (v_1, \ldots, v_n)$ has a certain probability $p_v = P(V = v \mid \pi, T, E)$ of being *observed*, so we obtain an $m \times m \times \cdots \times m$ table of probabilities, p. The set of all tables p which can arise from such a causal process is denoted by $\mathrm{HMM}(\ell, m, n)$. The entries of such p are forced to satisfy some implicit polynomial equations, and one can ask for constructive description of the set of all such equations, which is called an *ideal*.

Since the initial work of Bray and Morton [8], it has remained an open question to construct this ideal of polynomial equations satisfied by a given HMM. The ideal of $(2, 2, 3)$ was determined by Schönhuth [32], and Chapter 1 of this thesis shows:

Theorem 1 (1.4.1)**.** *All but a measure-zero subset of* $\mathrm{HMM}(2, m, n)$ *can be parametrized by a single generically injective polynomial map* $U \to \Delta_p^{2^n-1}$ *with an explicitly known, rational inverse formula, where* $U \subseteq \mathbb{R}^5$ *is a* 5*-dimensional open set cut out by known*

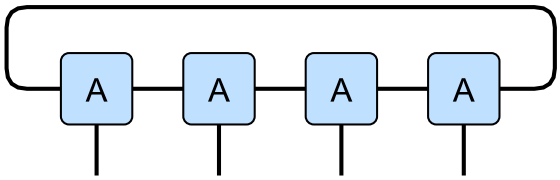


Figure 0.3: An MPS of length 4 with periodic boundary conditions

algebraic inequalities. In geometric terms, the Zariski closure of $\mathrm{HMM}(2,m,n)$ *in* $\mathbb{P}^{2^n-1}$ *is a rational projective variety.*

The proof makes use of classical invariant theory results of Sibirskii [33] and others on so-called *trace algebras*. Using this new parametrization, along with a new coordinate system called *cumulant coordinates* developed by Sturmfels and Zwiernik [37]. We also establish the following result:

Theorem 2 (1.3.1)**.** *Inside the hyperplane* $\sum_v p_v = 1$*, the ideal of polynomial equations satisfied by every* $p \in \mathrm{HMM}(2,2,4)$ *is minimally generated by 21 homogeneous quadric and 29 homogeneous cubic equations. Each of these* 50 *equations can be used to derive* $2^{m-1}[(n-3)+(n-6)+\cdots+(n-3\lfloor\frac{n}{3}\rfloor)]$ *polynomial equations satisfied by* $\mathrm{HMM}(2,m,n)$ *for each* $m \geq 2$ *and* $n \geq 4$*.*

HMM are an important test case for the application of algebraic techniques to statistical modeling. In particular, parameter estimation and model selection methods used in applications of HMM do not explicitly take into account their algebraic constraints, so there may be significant performance gains to be achieved by considering their geometry in this way.

In Chapter 2, our focus turns to matrix product state (MPS) models are used in condensed matter physics to express an entangled quantum state tensor in terms of a combination of simpler tensors connected by *virtual bonds*. For example, in Figure 0, if A is a $d \times D \times D$ tensor, the resulting MPS ψ is a $d \times d \times d \times d$ tensor Ψ — one d for each free "wire" – whose entries are given by

$$\psi_{i_1 i_2 i_3 i_4} = \sum_{j \in \{0,\ldots,D-1\}^4} A^{j_1}_{i_1 j_2} A^{j_2}_{i_2 j_3} A^{j_3}_{i_3 j_4} A^{j_4}_{i_4 j_1}.$$

MPS can be used to represent "stable" states of matter, and so classifying such states reduces to understanding the set of tensors representable as MPS (see Chen, Gu, and Wen [10]). The closure of this set is an algebraic variety, and in this paper, we study its geometry.

Using a reparametrization technique similar to that used on HMM above, I derived some polynomial constraints which must be satisfied by MPS of various formats.

Theorem 3 (2.1.3)**.** *A four-qubit state Ψ is a limit of binary periodic translation invariant MPS if and only if the following irreducible polynomial in its entries vanishes:*

$$\begin{aligned}
&\psi_{1010}^2\psi_{1100}^4 - 2\psi_{1100}^6 - 8\psi_{1000}\psi_{1010}\psi_{1100}^3\psi_{1110} + 12\psi_{1000}\psi_{1100}^4\psi_{1110} \\
&- 4\psi_{1000}^2\psi_{1010}^2\psi_{1110}^2 + 2\psi_{0000}\psi_{1010}^3\psi_{1110}^2 + 16\psi_{1000}^2\psi_{1010}\psi_{1100}\psi_{1110}^2 \\
&- 4\psi_{0000}\psi_{1010}^2\psi_{1100}\psi_{1110}^2 - 16\psi_{1000}^2\psi_{1100}^2\psi_{1110}^2 + 4\psi_{0000}\psi_{1010}\psi_{1100}^2\psi_{1110}^2 \\
&- 4\psi_{0000}\psi_{1100}^3\psi_{1110}^2 - 4\psi_{0000}\psi_{1000}\psi_{1010}\psi_{1110}^3 + 8\psi_{0000}\psi_{1000}\psi_{1100}\psi_{1110}^3 \\
&- \psi_{0000}^2\psi_{1110}^4 + 2\psi_{1000}^2\psi_{1010}^3\psi_{1111} - \psi_{0000}\psi_{1010}^4\psi_{1111} - 4\psi_{1000}^2\psi_{1010}^2\psi_{1100}\psi_{1111} \\
&+ 4\psi_{1000}^2\psi_{1010}\psi_{1100}^2\psi_{1111} + 2\psi_{0000}\psi_{1010}^2\psi_{1100}^2\psi_{1111} - 4\psi_{1000}^2\psi_{1100}^3\psi_{1111} \\
&+ \psi_{0000}\psi_{1100}^4\psi_{1111} - 4\psi_{1000}^3\psi_{1010}\psi_{1110}\psi_{1111} + 4\psi_{0000}\psi_{1000}\psi_{1010}^2\psi_{1110}\psi_{1111} \\
&+ 8\psi_{1000}^3\psi_{1100}\psi_{1110}\psi_{1111} - 8\psi_{0000}\psi_{1000}\psi_{1010}\psi_{1100}\psi_{1110}\psi_{1111} \\
&- 2\psi_{0000}\psi_{1000}^2\psi_{1110}^2\psi_{1111} + 2\psi_{0000}^2\psi_{1010}\psi_{1110}^2\psi_{1111} - \psi_{1000}^4\psi_{1111}^2 \\
&+ 2\psi_{0000}\psi_{1000}^2\psi_{1010}\psi_{1111}^2 - \psi_{0000}^2\psi_{1010}^2\psi_{1111}^2.
\end{aligned}$$

Theorem 4 (2.1.4)**.** *The ideal of constraints on binary translation invariant MPS with periodic boundary conditions is minimally generated by 3 quartics, 27 sextics, and possibly some higher degree polynomials.*

Aside from implicitly classifying states of matter, the proofs illustrate a connection between HMM and MPS first suggested by my collaborator Jason Morton, which we hope will begin a transfer of techniques between graphical statistical modeling and condensed matter physics.

In Chapter 3, we focus on a general method of reparametrizing directed acyclic graph (DAG) models on discrete variables. In a discrete DAG model, the nodes of the graph represent discrete random variables, and each variable is *affected* by its parent according to a conditional probability table, in a way similar to HMM. It is typical in applications that one can only observe a subset of the nodes, so one is interested in the joint marginal probability distribution induced on these nodes, called the *observed distribution*.

In the case when the DAG is a tree, Smith and Zwiernik [35] defined new coordinates called *tree cumulants* which allow for an extremely symbolically efficient expression of the observed distribution in terms of certain new parameters. The new parameters were at first somewhat mysterious and it was unclear whether they could be generalized for models with non-binary variables, i.e., variables taking on more

than two states. However, through discussions with Zwiernik, it became clear that they were in fact linear regression coefficients in a certain sense that would allow for their generalization. In Chapter 3, we develop the first step in this generalization, which in fact applies not only to trees but all discrete DAG models:

Theorem 5. *Given a discrete directed acyclic graph model G and a multiset S of observed nodes, the moment μ_S is given by the following automatic contraction equation:*

$$\mu_S \quad = \sum_{\substack{H \subset G \\ \text{sinks}(H) \subseteq S \subseteq \text{nodes}(H)}} \prod_{v \in \text{nodes}(H)} \beta_{v^{m(v;H,S)}}^{\text{pa}(v;H)}$$

This result will be used in future work with S. Lin, P. Zwiernik, and L. Weihs to generalize the tree cumulant parametrization of Smith and Zwiernik [35] for application to tree models where each variable can take on an arbitrary number of states.

Chapter 1

Hidden Markov models with two hidden states

1.1 Introduction

This chapter is based on my paper, *Binary hidden Markov models and varieties*, to appear in the *Journal of Algebraic Statistics* [1]. It is motivated primarily by the statistical problems of *model selection* and *parameter identifiability*, viewed from the perspective of algebraic geometry. Hidden Markov models (HMMs) are defined in Section 1.2, and here we focus on the simplest HMMs: those where all the hidden nodes are binary. Most questions about this case are answered by reducing to the case where the visible nodes are also binary. The hope is that eventually a very precise geometric understanding of HMMs can be attained that provides insight into their statistical properties.

The history of this and related problems has two main branches of historical lineage: that of hidden Markov models, and that of algebraic statistics.

In Section 1.2, we define hidden Markov models and related concepts needed for the remainder of the chapter. Their history began with a series of papers by Leonard E. Baum and others beginning with Baum and Petrie [4], after the description by Stratonovich [36] of the "forward-backward" algorithm that would be used for HMM parameter estimation. HMMs have been used extensively in natural language processing and speech recognition since the development of DRAGON by Baker [2]. As well, since Krogh, Mian, and Haussler [19] used HMM for finding genes in the DNA of E. coli bacteria, they have had many applications in genomics and biological sequence alignment; see also Yoon [42]. Now, HMM parameter estimation is built into the measurement of too minds kinds of measurements to reasonably count here.

Our second historical tributary, algebraic statistics, is the application of commutative algebra and algebraic geometry to the study of statistical models, especially those models involving non-linear relationships between parameters and observables. It was first described at length in the 2001 monograph *Algebraic Statistics* by Pistone, Riccomagno, and Wynn [30]. Subsequent introductions to the subject include *Algebraic Statistics for Computation Biology* by Pachter and Sturmfels [28], and *Lectures in Algebraic Statistics* by Drton, Sturmfels, and Sullivant [14]. Also notable is *Algebraic Geometry and Statistical Learning Theory* by Watanabe [41], for its focus on the problem of model selection.

The methods of algebraic statistics are much younger than hidden Markov models, and so the algebraic geometry of these models is far from fully explored. HMMs are hence an important early example for the theory to investigate. Here we focus on algebraic and geometric questions about HMMs coming from model selection and parameter identifiability.

The algebraic analogue of model selection is *implicitization*, i.e., finding polynomial defining equations for the Zariski closures of binary hidden Markov models. Here we use the term "model" synonymously with the set of probability distributions arising from the model. Polynomials vanishing on a model are called *invariants*: if a polynomial f is equal to a constant c at every point of the model (i.e., f does not vary with the model parameters), then we encode this equation by calling $f - c$ "an invariant". Model selection and implicitization are more than simply analogous; polynomial invariants have been used successfully in model selection by Casanellas and Fernandez-Sanchez [9] and Eriksson [15] for phylogenetic trees.

Such polynomial invariants have been difficult to specify for hidden Markov models, perhaps due to the high codimension of the models. One way to specify them is to exhibit a set of *defining equations* for the model, a finite collection of invariants which carve out the model in space, and combine to form the set of all its invariants, called an *ideal*. Bray and Morton [8] found many invariants using linear algebra, but did not exhibit defining equations for any model, and in fact their search was actually for invariants of a model that was slightly modified from the HMM proper. Schönhuth [32] found a large family of HMM invariants arising as minors of certain non-abelian Hankel matrices, and found that they constitute defining equations for the 3-node binary HMM, which is the simplest non-degenerate HMM. However, this seemed not to be the case for models with $n \geq 4$ nodes: Schönhuth reported on a computation of J. Hauenstein which verified numerically that the 4-node model was not cut out by the Hankel minors.

In Section 1.3, we will make use of moment and cumulant coordinates as exposited by Sturmfels and Zwiernik [37], as well as a new coordinate system on the parameter space, to find explicit defining equations for the 4-node binary HMM. The shortest

quadric and cubic equations are fairly simple; to give the reader a visual sense, they look like this:

$$g_{2,1} = m_{23}m_{13} - m_2 m_{134} - m_{13}m_{12} + m_1 m_{124}$$

$$g_{3,1} = m_{12}^3 - 2m_1 m_{12} m_{123} + m_\emptyset m_{123}^2 + m_1^2 m_{1234} - m_\emptyset m_{12} m_{1234}$$

Here each m is a moment of the observed probability distribution. These equations are not generated by Schönhuth's Hankel minors, and so provide a finer test for membership to any binary HMM of length $n \geq 4$ after marginalizing to any 4 equally spaced nodes.

The algebraic analogue of *parameter identifiability* is the generic or global injectivity or finiteness of a map of varieties that parametrizes the model, or in the case of identifying a single parameter, constancy of the parameter on the fibers of the parameterization. Sullivant, Garcia-Puente, and Spielvogel [38] provide an excellent discussion of this topic in the context of identifying causal effects; see also Meshkat, Eisenberg, and DiStefano [23] for a striking application to identification for ODE models in the biosciences.

In Section 1.4, for the purpose of parameter identification in binary hidden Markov models, we express the parametrization of a binary HMM as the composition of a dominant and generically finite monomial map $\mathfrak{q}$ and a birationally invertible map ψ. An explicit inverse to ψ is given, which allows for the easy recovery of hidden parameters in terms of observables. The components of the monomial map are *identifiable combinations* in the sense of Meshkat, Eisenberg, and DiStefano [23]. The formulae for recovering the hidden parameters are fairly simple when exhibited in a particular order, corresponding to a particular triangular set of generators in a union of lexicographic Gröbner bases for the model ideal. To show their simplicity, the most complicated recovery formula looks like this:

$$u = \frac{m_1 m_3 - m_2^2 + m_{23} - m_{12}}{2(m_3 - m_2)}$$

As a corollary, in Section 1.4 we find that the fibers of ϕ_n are generically zero-dimensional, each consisting of two points which are equivalent under a "hidden label swapping" operation.

Section 1.5 describes describes how the parametrization of every fully binary HMM, or "BHMM", can be factored through a particular 9-dimensional variety called a *trace variety*, which is the invariant theory quotient of the space of triples of 2×2 matrices under a simultaneous conjugation action by SL_2. As a quotient, the trace variety is not defined inside any particular ambient space. However, its

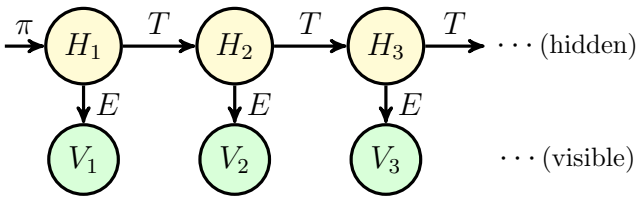


Figure 1.1: Hidden Markov models as Bayesian networks.

coordinate ring, a *trace algebra*, was found by Sibirskii [33] to be generated by 10 elements, which means we can embed the trace variety in $\mathbb{C}^{10}$. We prove the main results of Section 1.4 in the coordinates of this embedding. As a byproduct of this approach, in section Section 1.5 we find that the Zariski closures of all BHMMs with $n \geq 3$ are birational to each other.

Finally, Section 1.6 explores some applications of our results, including model membership testing, classification of identifiable parameters, a new grading on HMMs that can be used to find low-degree invariants, the geometry of equilibrium BHMMs, and HMMs with more than two visible states.

1.2 Definitions

Important note: In this chapter, we will work mostly with BHMMs — HMMs in which both the hidden and visible nodes are all binary — because, as will be explained, all our results will apply to models with any number $k \geq 2$ of visible states by reducing to this case.

Throughout, we will be referring to binary hidden Markov *processes, distributions, maps, models, varieties,* and *ideals*. Each of these terms is used with a distinct meaning, and effort is made to keep their usages consistent and separate.

Binary Hidden Markov processes and distributions

A binary hidden Markov process is a statistical process which generates random binary sequences. It is based on the simpler notion of a binary (and not hidden) Markov *chain* process, and can be depicted as a causal Bayesian network, or directed graphical model, as shown in Figure 1.1.

Definition 1.2.1. A **Binary Hidden Markov *process*** will comprise 5 data: π, T, E, and (H_t, V_t). The pair (H_t, V_t) denotes a jointly random sequence $(H_1, V_1, H_2, V_2, \ldots)$

of binary variables, also respectively called *hidden nodes* and *visible nodes*, with range $\{0,1\}$. Often a bound n on the (discrete) time index t is also given. The joint distribution of the nodes is specified by the following:

- A row vector $\pi = (\pi_0, \pi_1)$, called the *initial distribution*, which specifies a probability distribution on the first hidden node H_1 by $\Pr(H_1 = i) = \pi_i$;
- A matrix $T = \begin{bmatrix} T_{00} & T_{01} \\ T_{10} & T_{11} \end{bmatrix}$, called the *transition matrix*, which specifies conditional "transition" probabilities by the formula $\Pr(H_t = j \,|\, H_{t-1} = i) = T_{ij}$, read as the probability of "transitioning from hidden state i to hidden state j".[1] Together, π and T define what is traditionally called a *Markov chain process* on the hidden nodes H_i.
- A matrix $E = \begin{bmatrix} E_{00} & E_{01} \\ E_{10} & E_{11} \end{bmatrix}$, called the *emission matrix*, which specifies conditional "emission" probabilities by the formula $\Pr(V_t = j \,|\, H_t = i) = E_{ij}$, read as the probability that "hidden state i emits the visible state j".

To be precise, the parameter vector $\theta = (\pi, T, E)$ determines a probability distribution on the set of sequences of pairs $((H_1, V_1) \dots (H_n, V_n)) \in (\{0,1\}^2)^n$, or if no bound n is specified, a compatible sequence of such distributions as n grows. In applications, only the joint distribution on the visible nodes $(V_1, \dots, V_n) \in \{0,1\}^n$ is observed, and is called the *observed distribution*. This distribution is given by marginalizing (summing) over the possible hidden states of a BHM process:

$$\begin{aligned} \Pr(V = v \,|\, \theta = (\pi, T, E)) &= \sum_{h \in \{0,1\}^n} \Pr(h, v | \, \pi, T, E) = \sum_{h \in \{0,1\}^n} \Pr(h \,|\, \pi, T) \Pr(v \,|\, h, E) \\ &= \sum_{h \in \{0,1\}^n} \pi_{h_1} E_{h_1, v_1} \prod_{i=2}^{n} T_{h_{i-1} h_i} E_{h_i, v_i} \end{aligned} \tag{1.1}$$

Definition 1.2.2. A **Binary Hidden Markov *distribution*** is a probability distribution on sequences $v \in \{0,1\}^n$ of jointly random binary variables $(V_1, \dots, V_n)$ which arises as the observed distribution of *some* BHM *process* according to (1.1).

As we will see in Section 1.4, different processes (π, T, E, H_t, V_t) can give rise to the same observed distribution on the V_t, for example by permuting the labels of the hidden variables, or by other relations among the parameters.

Those already familiar with Markov models in some form may note that:

[1] Schönhuth [32] uses T for different matrices, which I will later denote by P.

- The matrices T and E are implicitly assumed to be *stationary*, meaning that they are not allowed to vary with the "time index" t of (H_t, V_t).
- The distribution π *is not* assumed to be *at equilibrium* with respect to T, i.e. we *do not* assume that $\pi T = \pi$. This allows for more diverse applications.

Remark 1.2.3. The term "stationary" is sometimes also used for a process that is at equilibrium; we will reserve the term "stationary" for the constancy of matrices T,E over time.

Binary Hidden Markov maps, models, varieties, and ideals

Statistical processes come in families defined by allowing their parameters to vary, and in short, the set of probability distributions that can arise from the processes in a given family is called a *statistical model*. The Zariski closure of such a model in an appropriate complex space is an algebraic variety, and the geometry of this variety carries information about the purely algebraic properties of the model.

In a binary hidden Markov process, π, T, and E must be *stochastic matrices*, i.e., each of their rows must consist of non-negative reals which sum to 1, since these rows are probability distributions. We denote by Θ_{st} the set of such triples (π, T, E), which is isometric to the 5-dimensional cube $(\Delta_1)^5$. We call Θ_{st} the space of *stochastic parameters*. It is helpful to also consider the larger space of triples (π, T, E) where the matrices can have arbitrary complex entries with row sums of 1. We write $\Theta_{\mathbb{C}}$ for this larger space, which is equal to the complex Zariski closure of Θ_{st}, and call it the space of *complex parameters*.

We will not simply replace Θ_{st} *by* $\Theta_{\mathbb{C}}$ *for convenience*, as has sometimes been done in algebraic phylogenetics. For the ring of polynomial functions on these spaces, we write

$$\mathbb{C}[\theta] := \mathbb{C}[\pi_j, T_{ij}, E_{ij}] \Big/ \left(1 = \sum_j \pi_j = \sum_j T_{ij} = \sum_j E_{ij} \text{ for } i = 0, 1\right)$$

so as to make the identification $\Theta_{\text{st}} \subseteq \Theta_{\mathbb{C}} = \operatorname{Spec} \mathbb{C}[\theta]$. Here Spec denotes the spectrum of a ring; see [12] for this and other fundamentals of algebraic geometry.

Now we fix a length $|v| = n$ for our binary sequences v, and write

$$R_{p,n} := \mathbb{C}[p_v \,|\, v \in \{0,1\}^n] \qquad \mathbb{C}_p^{2^n} := \operatorname{Spec}(R_{p,n})$$
$$\overline{R}_{p,n} := R_{p,n} / (1 - \sum_{|v|=n} p_v) \qquad \mathbb{C}_p^{2^n - 1} := \operatorname{Spec}(\overline{R}_{p,n})$$
$$\mathbb{P}_p^{2^n - 1} := \operatorname{Proj}(R_{p,n})$$

We will often have occasion to consider the natural inclusions,

$$\iota_n : \mathbb{C}_p^{2^n-1} \hookrightarrow \mathbb{C}_p^{2^n} \qquad\qquad \bar{\iota}_n : \mathbb{C}_p^{2^n-1} \hookrightarrow \mathbb{P}_p^{2^n-1}.$$

Convention 1.2.4. Complex spaces such as $\mathbb{C}^{2^n}$ will usually be decorated with a subscript to indicate the intended coordinates to be used on that space, like the p in $\mathbb{C}_p^{2^n}$ above. Likewise, a ring will usually be denoted by R with some subscripts to indicate its generators.

Definition 1.2.5. For each $n \geq 3$, we introduce the following objects:

- The **Binary Hidden Markov *map*** or *modeling map* on n nodes is the map which sends a parameter vector θ to the distribution p it induces on the vector of observable variables, according to (1.1). We denote this distribution by $\phi_{\mathrm{BHMM}(n)}$, or simply ϕ_n:

$$\phi_n : \Theta_{\mathbb{C}} \to \mathbb{C}_p^{2^n-1},$$

$$\phi_n^{\#}(p_v) := \sum_{h \in \{0,1\}^n} \pi_{h_1} E_{h_1,v_1} \prod_{i=2}^{n} T_{h_{i-1}h_i} E_{h_i,v_i}$$

  The word "model" is also frequently used for the map ϕ_n. This is a very reasonable usage of the term, but I reserve "model" for the image of the allowed parameter values:

- $\mathrm{BHMM}(n)$, the **Binary Hidden Markov *model*** on n nodes, is the image

$$\bar{\iota}_n \phi_n \left(\Theta_{\mathrm{st}}\right) \subseteq \mathbb{P}_p^{2^n-1},$$

  of the stochastic parameter space Θ_{st}, i.e., the set of observed distributions which can arise from *some* BHM process, considered as a subset of $\mathbb{P}_p^{2^n-1}$ via $\bar{\iota}_n$. Being the continuous image of the classically compact cube $\Theta_{\mathrm{st}} \simeq \Delta_1^5$, $\mathrm{BHMM}(n)$ is also classically compact and hence classically closed.

- $\overline{\mathrm{BHMM}}(n)$, the **Binary Hidden Markov *variety*** on n nodes, is the Zariski closure of $\mathrm{BHMM}(n)$, or equivalently the Zariski or classical closure of $\bar{\iota}_n \phi_n(\Theta_{\mathbb{C}})$, in $\mathbb{P}_p^{2^n-1}$.

- $\mathrm{I}_{\mathrm{BHMM}(n)}$, the **Binary Hidden Markov *ideal*** on n nodes, is the set of homogeneous polynomials which vanish on $\mathrm{BHMM}(n)$, i.e., the homogeneous defining ideal of $\overline{\mathrm{BHMM}}(n)$. Elements of $\mathrm{I}_{\mathrm{BHMM}(n)}$ are called **invariants** of the model.

In summary, probability distributions arise from processes according to modeling maps, models are families of distributions arising from processes of a certain type, and the Zariski closure of each model is a variety whose geometry reflects the algebraic properties of the model. The ideal of the model is the same as the ideal of the variety: the definition of Zariski closure is the largest set which has the same ideal of vanishing polynomials as the model. In a rigorous sense (namely, the anti-equivalence of the categories of affine schemes and rings), the variety encodes information about the "purely algebraic" properties of the model, i.e., properties that can be stated by the vanishing of polynomials.

The number of polynomials that vanish on any given set is infinite, but by the Hilbert Basis theorem, one can always find finitely many polynomials whose vanishing implies the vanishing of all the others, called a *generating set* for the ideal. To compute a generating set for $\mathrm{I}_{\mathrm{BHMM}(n)}$, we will need the following proposition:

Proposition 1.2.6. *The ideal* $\mathrm{I}_{\mathrm{BHMM}(n)}$ *is the homogenization of* $\ker(\phi_n^\# \circ \iota_n^\#)$ *with respect to* $p_\Sigma := \sum_{|v|=n} p_v$.

Proof. The affine ideal $\ker(\phi_n^\# \circ \iota_n^\#)$ cuts out the Zariski closure X of $\iota_n \circ \phi_n(\Theta_{\mathbb{C}})$ in $\mathbb{C}_p^{2^n}$, and this closure lies in the hyperplane $\{p_\Sigma = 1\} = \mathbb{C}_p^{2^n-1}$. Let X' be the projective closure of X in $\mathbb{P}_p^{2^n-1}$, so that $I(X')$ is the homogenization of $\ker(\phi_n^\# \circ \iota_n^\#)$ with respect to p_Σ.

The cube Θ_{st} is Zariski dense in $\Theta_{\mathbb{C}}$, so $\iota_n \circ \phi_n(\Theta_{\mathrm{st}})$ is Zariski dense in $\iota_n \circ \phi_n(\Theta_{\mathbb{C}})$, which is Zariski dense in X, which is Zariski dense in X'. Therefore $X' = \overline{\mathrm{BHMM}}(n)$, and $I(X') = \mathrm{I}_{\mathrm{BHMM}(n)}$, as required. □

HMMs with more visible states via BHMM(n)

All the results of this chapter apply also to HMMs with $k \geq 3$ visible states, provided the number of hidden states is $d = 2$. The idea is to encode such a hidden Markov process in a collection of BHM processes, and apply our methods to those.

Consider an HMM$(2, k, n)$ process P, which has 2 hidden states, k visible states $1 \ldots k$, and n (consecutive) visible nodes. As in Definition 1.2.1 and (1.9), P is given by a $2 \times k$ matrix E of emission probabilities, along with a 1×2 matrix π and a 2×2 matrix T describing the two-state hidden Markov chain. For $\ell \in \{1 \ldots, k\}$, we define a BHMM(n) process P^ℓ by defining binary variables V_t^ℓ, where $V_t^\ell = 1$ if $V_t = \ell$ and $V_t^\ell = 0$ otherwise. Figure 1.2 is a Bayesian network depicting this dependency.

Formally, as in Definition 1.2.1, the BHMM(n) process P^ℓ is defined to be the quintuple

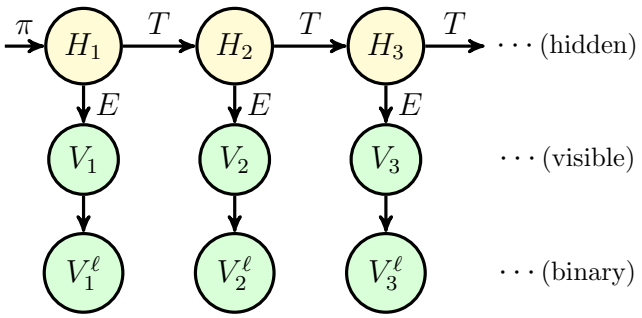


Figure 1.2: Converting an HMM with many visible states into HMMs with two visible states

$(\pi, T, E^\ell, H_t, V_t^\ell)$, where E^ℓ is the emission matrix from the H_t to the V_t^ℓ,

$$E^\ell = \begin{bmatrix} 1 - E_{0\ell} & E_{0\ell} \\ 1 - E_{1\ell} & E_{1\ell} \end{bmatrix}$$

The HMM$(2, k, n)$ process P is *encoded* in the processes P^ℓ in the sense that the value of V_t is recoverable as the unique ℓ such that $V_t^\ell = 1$. Thus any polynomial invariant constraining the joint distribution of the V_t^ℓ for some ℓ implies a constraint on the joint distribution of the V_t. In other words, for each ℓ, we can lift invariants for BHMM(n) to invariants for HMM$(2, k, n)$ by the substitution

$$p^\ell_{i_1 \dots i_n} \mapsto \sum \{p_{j_1 \dots j_n} \mid j_t = \ell \text{ if and only if } i_t = 1 \text{ for each } 1 \le t \le n\}$$

For example, $p^\ell_{1010} \mapsto p_{\ell * \ell *}$, where "$*$" denotes an index to be summed over $\{1, \dots, n\} \setminus \{\ell\}$. In this way, any invariant for BHMM(n) yields k invariants for HMM$(2, k, n)$.

As well, as ℓ varies, we obtain all the entries of E as entries of some E^ℓ. So any formula for identifying the parameters of BHMM(n) in terms of observables can be applied to identify the parameters of an HMM$(2, k, n)$ process P as they arise as parameters of the BHMM(n) processes P^ℓ.

We shall remark throughout when results for the model BHMM(n) can also be applied to the model HMM$(2, k, n)$ using this encoding method.

1.3 Defining equations of $\overline{\text{BHMM}}(3)$ and $\overline{\text{BHMM}}(4)$

Theorem 1.3.1. *The homogeneous ideal* $\text{I}_{\overline{\text{BHMM}}(4)}$ *of the binary hidden Markov variety* $\overline{\text{BHMM}}(4)$ *is minimally generated by 21 homogeneous quadrics and 29 homogeneous cubics.*

Since Schönhuth [32] found numerically that his Hankel minors did not cut out BHMM(4) even set-theoretically, these equations are genuinely new invariants of the model. Moreover, they are not only applicable to BHMM(4), because a BHM process of length $n > 4$ can be marginalized to any 4 evenly-spaced hidden-visible node pairs to obtain a BHM process of length 4. When there are n node pairs, there are $f(n) = (n-3) + (n-7) + \cdots + (n - 3\lfloor\frac{n-1}{3}\rfloor)$ equally-spaced sequences of four node pairs, and thus $f(n)$ linear maps from BHMM(n) to BHMM(4), each of which allows us to write 21 quadrics and 29 cubics which vanish on BHMM(n). Thus we obtain a super-exponential number, $50 \cdot f(n)$, of invariants for BHMM(n) as n grows. Finally, using the encoding of Section 1.2, we can even obtain $50 \cdot k \cdot f(n)$ invariants of HMM$(2, k, n)$ via the k different reductions to BHMM(n).

Our fastest derivation of Theorem 1.3.1 in Macaulay2 (see Grayson and Stillman [22]) uses the birational parametrization of Section 1.4, but in only a single step, so we defer the lengthier discussion of the parametrization until then. Modulo this dependency, the proof is described in Section 1.3, using moment coordinates (Section 1.3) and cumulant coordinates (Section 1.3).

In probability coordinates, the generators found for $\mathrm{I}_{\mathrm{BHMM(4)}}$ have the following sizes:

- Quadrics $g_{2,1}, \ldots, g_{2,21}$: respectively $8, 8, 12, 14, 16, 21, 24, 24, 26, 26, 28, 32, 32, 41, 42, 43, 43, 44, 45, 72, 72$ probability terms.
- Cubics $g_{3,1}, \ldots, g_{3,29}$: respectively $32, 43, 44, 44, 44, 52, 52, 56, 56, 61, 69, 71, 74, 76, 78, 81, 99, 104, 109, 119, 128, 132, 148, 157, 176, 207, 224, 236, 429$ probability terms.

As a motivation for introducing moment coordinates, we note here that these generators have considerably fewer terms when written in terms of moments:

- Quadrics $g_{2,1}, \ldots, g_{2,21}$: respectively $4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 6, 6, 8, 8, 8, 8, 8, 10, 10, 10, 17$ moment terms.
- Cubics $g_{3,1}, \ldots, g_{3,29}$: respectively $5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 8, 8, 8, 8, 10, 10, 10, 10, 10, 12, 12, 13, 14, 16, 18, 21, 27, 35$ moment terms.

To give a sense of how they look in terms of moments, the shortest quadric and cubic are

- $g_{2,1} = m_{23}m_{13} - m_2m_{134} - m_{13}m_{12} + m_1m_{124}$, and
- $g_{3,1} = m_{12}^3 - 2m_1m_{12}m_{123} + m_\emptyset m_{123}^2 + m_1^2m_{1234} - m_\emptyset m_{12}m_{1234}$.

Let us compare this ideal with $\mathrm{I}_{\mathrm{BHMM}(3)}$, the homogeneous defining ideal of $\overline{\mathrm{BHMM}}(3)$. Schönhuth [32] found that $\mathrm{I}_{\mathrm{BHMM}(3)}$ is precisely the ideal of 3×3 minors of the matrix

$$A_{3,3} = \begin{bmatrix} p_{000}+p_{001} & p_{000} & p_{100} \\ p_{010}+p_{011} & p_{001} & p_{101} \\ p_{100}+p_{101} & p_{010} & p_{110} \\ p_{110}+p_{111} & p_{011} & p_{111} \end{bmatrix} \tag{1.2}$$

Schönhuth defines an analogous matrix $A_{n,3}$ for $\overline{\mathrm{BHMM}}(n)$, but then remarks that J. Hauenstein has found, using numerical rank deficiency testing (see Bates et al. [3]) with the algebraic geometry package Bertini (see Bates et al. [6]), that $\mathrm{minors}_3(A_{n,3})$ does not cut out $\overline{\mathrm{BHMM}}(n)$ when $n = 4$. In general, Schönhuth shows that $\mathrm{I}_{\mathrm{BHMM}(n)} = (\mathrm{minors}_3(A_{n,3}) : \mathrm{minors}_2(B_{n,2}))$ for a particular 2×3 matrix $B_{n,2}$, but computing generators for this colon ideal is a costly operation, and so no generating set for $\mathrm{I}_{\mathrm{BHMM}(n)}$ was found for any $n \geq 4$ by this method. Instead, we will use the *moment coordinates* and *cumulant coordinates* exposited by Sturmfels and Zwiernik [37].

Moment coordinates

When the length n of the observed binary sequences v is understood, for each subset I of $[n] = \{1, \ldots, n\}$, we define the I^{th} *moment coordinate* by

$$m_I := \sum \{p_v \mid v_t = 1 \text{ for all } t \in I\} \tag{1.3}$$

This is just an expression for the probability that $V_t = 1$ for all $t \in I$. For example, when $n = 5$, $m_{14} = p_{1{+}{+}1{+}}$, and $m_\emptyset = p_{{+}{+}{+}{+}{+}} = 1$, following the usual convention that "+" denotes an index to be summed over. Thus, moments are particular linear expressions in probabilities. They can also be derived from a moment generating function as in Sturmfels and Zwiernik [37], which in our case reduces to the above. The $m_I \in R_{p,n}$ provide alternative linear coordinates on $\mathbb{P}_p^{2^n-1}$ in which some previously intractable BHM computations become feasible.

We will sometimes use the notation m_v for the moment m_I when v is the indicator string of I. For example, $m_{01010} = m_{24}$, $m_{10000} = m_1$, and $m_{00000} = m_\emptyset$. The string notation has the advantage of making the value $n = 5$ apparent, and is also convenient in the Baum formula for moments (Proposition 1.5.1) as explained in Section 1.5.

When working with BHMMs, the expression for m_I in terms of the parameters (π, T, E) is in fact independent of n. That is, if $I \subseteq [n]$ and I' denotes I considered as a subset of $[n']$ for some $n' > n$, then

$$\phi_n^\#(m_I) = \phi_{n'}^\#(m_{I'}) \tag{1.4}$$

This can be seen in many ways, for example using the Baum formula for moments (Proposition 1.5.1), or by noting from the interpretation of m_I as a marginal probability that it can be computed without considering the states of nodes H_t and V_t where t is larger than the largest element of I.

Just as for probabilities, for moments we can define rings and spaces

$$\begin{aligned} R_{m,n} &:= \mathbb{C}[m_I \,|\, I \subseteq [n]] & \mathbb{C}_m^{2^n} &:= \operatorname{Spec}(R_{m,n}), \\ \overline{R}_{m,n} &:= R_{m,n}/\langle 1 - m_\emptyset \rangle & \mathbb{C}_m^{2^n-1} &:= \operatorname{Spec}(\overline{R}_{m,n}), \\ & & \mathbb{P}_m^{2^n-1} &:= \operatorname{Proj}(R_{m,n}). \end{aligned} \tag{1.5}$$

Convention 1.3.2. To avoid having notation for too many ring isomorphisms, we will use (1.3) to treat m_I as an element of $R_{p,n}$, thus creating literal identifications

$$\begin{aligned} R_{m,n} &= R_{p,n} & \mathbb{C}_m^{2^n} &= \mathbb{C}_p^{2^n}, \\ \overline{R}_{m,n} &= \overline{R}_{p,n} & \mathbb{C}_m^{2^n-1} &:= \mathbb{C}_p^{2^n-1}, \\ & & \mathbb{P}_m^{2^n-1} &:= \mathbb{P}_p^{2^n-1}. \end{aligned} \tag{1.6}$$

Note also that we obtain natural ring inclusions $R_{m,n} \subseteq R_{m,n'}$ whenever $n < n'$, which respect the BHM maps ϕ_n because of (1.4).

As a first application of moment coordinates, we have

Proposition 1.3.3. *The homogeneous ideal* $\mathrm{I}_{\mathrm{BHMM}(3)}$ *is generated in moment coordinates by the* 3×3 *minors of the matrix*

$$A'_{3,3} = \begin{bmatrix} m_{000} & m_{000} & m_{100} \\ m_{100} & m_{010} & m_{110} \\ m_{010} & m_{001} & m_{101} \\ m_{110} & m_{011} & m_{111} \end{bmatrix} = \begin{bmatrix} m_\emptyset & m_\emptyset & m_1 \\ m_1 & m_2 & m_{12} \\ m_2 & m_3 & m_{13} \\ m_{12} & m_{23} & m_{123} \end{bmatrix}$$

In particular, the projective variety $\overline{\mathrm{BHMM}}(3)$ *is cut out by these minors.*

Proof. Observe that Schönhuth's matrix $A_{3,3}$ in (1.2) is equivalent under elementary row/column operations to $A'_{3,3}$, so $\operatorname{minors}_3(A'_{3,3}) = \operatorname{minors}_3(A_{3,3}) = \mathrm{I}_{\mathrm{BHMM}(3)}$. □

Proposition 1.3.4. *The ideal* $\mathrm{I}_{\mathrm{BHMM}(n)}$ *is the homogenization of* $\ker(\phi_n^\#)$ *with respect to* $m_\emptyset$.

Proof. From Proposition 1.2.6 we know that $\mathrm{I}_{\mathrm{BHMM}(n)}$ is the homogenization of $\ker(\phi_n^\# \circ \iota_n^\#)$ with respect to $m_\emptyset = \sum_{|v|=n} p_v$. From (1.6), we can identify $\overline{R}_{m,4}$ with the polynomial subring of $R_{m,4}$ obtained by omitting $m_\emptyset$, so that $\ker(\phi_4^\# \circ \iota_4^\#) = \ker(\phi_4^\#) + \langle 1 - m_\emptyset \rangle$. Since the additional generator $1 - m_\emptyset$ homogenizes to 0, $\ker(\phi_4^\#)$ has the same homogenization as $\ker(\phi_4^\# \circ \iota_4^\#)$, hence the result. □

Cumulant coordinates

Cumulants are non-linear expressions in moments or probabilities which seem to allow even faster computations with binary hidden Markov models. Let

$$\begin{aligned} R_{k,n} &:= \mathbb{C}[k_I \,|\, I \subseteq [n]] \\ \overline{R}_{k,n} &:= R_{k,n}/\langle k_\emptyset \rangle \\ \mathbb{C}_k^{2^n-1} &:= \mathrm{Spec}(\overline{R}_{k,n}) \end{aligned}$$

where, as with moments, we may freely alternate between writing k_v and writing k_I, where I is the set of positions where 1 occurs in v. For building generating functions, let $x_1, \ldots, x_n$ be indeterminates, and write $x^v = x^I$ for $x_1^{v_1} \cdots x_n^{v_n} = \prod_{i \in I} x_i$. Let J be the ideal generated by all the squares x_i^2. Following Sturmfels and Zwiernik [37], we introduce the *moment* and *cumulant generating functions*, respectively, as

$$f_m(x) := \sum_{I \subseteq [n]} m_I x^I \in \overline{R}_{m,n}[x]/J \qquad f_k(x) := \sum_{I \subseteq [n]} k_I x^I \in \overline{R}_{k,n}[x]/J$$

We now define changes of coordinates

$$\kappa_n : \mathbb{C}_m^{2^n-1} \to \mathbb{C}_k^{2^n-1} \qquad \kappa_n^{-1} : \mathbb{C}_k^{2^n-1} \to \mathbb{C}_m^{2^n-1}$$

by the formulae

$$\begin{aligned} \kappa_n^{\#}(f_k) = \log(f_m) &= \frac{(f_m - 1)}{1} + \cdots + (-1)^{n+1} \frac{(f_m - 1)^n}{n} \\ \kappa_n^{-\#}(f_m) = \exp(f_k) &= 1 + \frac{(f_k)}{1} + \cdots + \frac{(f_k)^n}{n!} \end{aligned} \tag{1.7}$$

That is, we let $\kappa_n^{\#}(k_I)$ be the coefficient of x^I in the Taylor expansion of $\log f_m$ about 1, and let $\kappa_n^{-\#}(m_I)$ be the coefficient of x^I in the Taylor expansion of $\exp f_k$ about 0. Note that in the relevant coordinate rings $\overline{R}_{m,n}$ and $\overline{R}_{k,n}$, $m_\emptyset = 1$ and $k_\emptyset = 0$. This is why we only need to compute the first n terms of each Taylor expansion: the higher terms all vanish modulo the ideal J.

Proposition 1.3.5. *The expressions $\kappa_n^{\#}(k_I)$ and $\kappa_n^{-\#}(m_I)$, i.e., writing of cumulants in terms of moments and conversely, do not depend on n.*

Proof. By Sturmfels and Zwiernik [37], these formulae are re-expressed using Möbius functions, which do not depend on the generating functions above, and in particular do not depend on n. □

Deriving $\mathrm{I}_{\mathrm{BHMM}(4)}$ in Macaulay2

This section describes the proof of Theorem 1.3.1 using Macaulay2. These computations were carried out on a Toshiba Satellite P500 laptop running Ubuntu 10.04, with an Intel Core i7 Q740 .73 GHz CPU and 8gb of RAM. In light of Proposition 1.3.4, we will aim to compute $\ker(\phi_4^\# \circ \iota_4^\#)$, which can be understood geometrically as the (non-homogeneous) ideal of the standard affine patch of $\overline{\mathrm{BHMM}}(4)$ where $m_\emptyset = \sum_{|v|=4} p_v = 1$. To reduce the number of variables, as in Proposition 1.3.4 we continue to make the identification

$$\overline{R}_{m,4} = \mathbb{C}[m_I | \emptyset \neq I \subseteq [4]] \subseteq R_{m,4}$$

We begin by providing Macaulay2 with the map $\phi_4^\# : \overline{R}_{m,4} \to \mathbb{C}[\theta]$ in moment coordinates (Section 1.3), because probability coordinates result in longer, higher degree expressions. This can be done by composing the expression of $\phi_n^\#(p_v)$ in Definition 1.2.5 with the expression of $m_v = m_I$ in (1.3), or alternatively using the Baum formula for moments (Proposition 1.5.1) , which involves many fewer arithmetic operations.

Macaulay2 runs out of memory (8gb) trying to compute $\ker(\phi_4^\#)$, and as expected, this memory runs out even sooner in probability coordinates, so we use cumulant coordinates instead (Section 1.3). We input

$$\kappa_4^\# : \overline{R}_{k,4} \to \overline{R}_{m,4}$$

using coefficient extraction from (1.7), and compute the composition $\phi_4^\# \circ \kappa_4^\#$. Then, it is possible to compute

$$\mathbf{I}_{k,4} := \ker(\phi_4^\# \circ \kappa_4^\#)$$

which takes **around 1.5 hours**. Alternatively, we can compute $\mathbf{I}_{k,4}$ by expressing the cumulants in terms of the birational parameters of Section 1.4 (i.e. using ψ_4 in place of ϕ_4), which takes **less than 1 second** and yields 100 generators for $\mathbf{I}_{k,4}$.

Subsequent computations run out of memory with this set of 100 generators, so we must take some steps to simplify it. Macaulay2's **trim** command reduces the number of generators of $\mathbf{I}_{k,4}$ to 46 in **under 1 second**. We then order these 46 generators lexicographically, first by degree and then by number of terms, and eliminate redundant generators in reverse order, which takes **19 seconds**. The result is an inclusion-minimal, non-homogeneous generating set for $\mathbf{I}_{k,4}$ with 35 generators: 24 quadrics and 11 cubics.

Now we compute $\mathbf{I}_{m,4} := \kappa^\#(I_{k,4}) = \kappa^\#(\ker(\phi_4^\# \circ \kappa_4^\#)) = \ker(\phi_4^\#)$, i.e., we push forward the 35 generators for $\mathbf{I}_{k,4}$ under the non-linear ring isomorphism $\kappa_4^\#$ to obtain 35 generators for $\mathbf{I}_{m,4} = \ker(\phi_4^\#)$: 2 quadrics, 7 cubics, 16 quartics, 5 quintics, and

5 sextics. In **under 1 second**, Macaulay2's **trim** command computes a new set of 39 generators for $\mathbf{I}_{m,4}$ with lower degrees: 21 quadrics, 14 cubics, and 4 quartics, which turns out to **save around 1 hour** of computing time in what follows. These generators have many terms each, and eliminating redundant generators as in the previous paragraph turns out to be too slow to be worth it here, taking more than 2 hours, so we omit this step.

Finally, we apply Proposition 1.3.4 to compute $\mathrm{I}_{\mathrm{BHMM}(4)}$ as the homogenization of $\mathbf{I}_{m,4}$ with respect to $m_\emptyset$. In Macaulay2, this is achieved by homogenizing the 39 generators for $\mathbf{I}_{m,4}$ with respect to $m_\emptyset$ and then saturating the ideal they generate with respect to $m_\emptyset$. This saturation operation takes about **29 minutes**, and yields a minimal generating set of 50 polynomials: 21 quadrics and 29 cubics. Since probabilities are linear in moments, their degrees are the same in probability coordinates. Moreover, since these are homogeneous generators for a homogeneous ideal, they are minimal in a very strong sense:

Corollary 1.3.6. *Any inclusion-minimal homogeneous generating set for* $\mathrm{I}_{\mathrm{BHMM}(4)}$ *in probability or moment coordinates must contain exactly 21 quadrics and 29 cubics.*

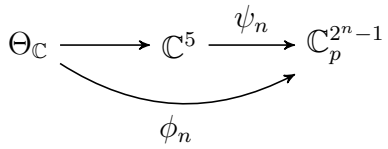
We still do not know a generating set for $\mathrm{I}_{\mathrm{BHMM}(5)}$. Macaulay2 runs out of memory (8gb) attempting to compute $\mathbf{I}_{k,5}$, even using the birational parametrization of Section 1.4. The author has also attempted this computation using the *tree cumulants* of Smith and Zwiernik [35] in place of cumulants, but again Macaulay2 runs out of memory trying to compute the first kernel. Presumably the subsequent saturation step would be even more computationally difficult.

1.4 Birational parametrization of BHMMs

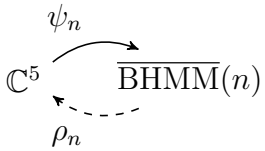
We say that two parameter vectors $\theta = (\pi, T, E)$ and $\theta' = (\pi', T', E')$ for a BHMM(n) are *equivalent* if they give rise to the same probability distribution p on the observed nodes, i.e., if $\phi_n(\theta) = \phi_n(\theta')$. A generic θ has only one other θ' in its equivalence class, obtainable essentially by swapping the labels $\{0, 1\}$ of all the H_t simultaneously, as we'll describe in Section 1.4. But some θ have larger equivalence classes, as we'll see in Section 1.4 in the cases where $det(T) = 0$ or $det(E) = 0$.

It turns out to be possible to algebraically "glue together" some, but not all, pairs of equivalent θ's to obtain a new parameter space which still lives in $\mathbb{C}^5$, but such that the map from our new parameters to observables is *generically injective*, and has an inverse which is a rational function. This is the intuition behind the following:

Theorem 1.4.1 (Birational Parameter Theorem)**.** *There is a generically two-to-one, dominant morphism $\Theta_{\mathbb{C}} \to \mathbb{C}^5$ such that, for each $n \geq 3$, the binary hidden Markov map ϕ_n factors uniquely as follows,*



and each $\psi_n : \mathbb{C}^5 \to \overline{\mathrm{BHMM}}(n)$ has a birational inverse map ρ_n:



In particular, $\overline{\mathrm{BHMM}}(n)$ is always a rational projective variety of dimension 5, i.e., birationally equivalent to $\mathbb{P}^5$.

Using the encoding of Section 1.2, the same is true if we allow $k > 2$ visible states in the model and replace $\mathbb{C}^5$ by $\mathbb{C}^{3+k}$. This theorem will be proven in Section 1.5 using trace algebras and the Baum formula for moments. In the course of this section and Section 1.5 we will exhibit formulae for ψ_n and their inverses ρ_n. The inverse map ρ_3 has a number of practical uses, to be explored in Section 1.6.

Our first step toward Theorem 1.4.1 is to re-parametrize $\Theta_{\mathbb{C}}$.

A linear reparametrization of $\Theta_{\mathbb{C}}$

Since the hidden variables H_t are never observed, there is no change in the final expression of p_v in Definition 1.2.5 if we swap the labels $\{0, 1\}$ of all the H_t simultaneously. This swapping is equivalent to an action of the elementary permutation matrix $\sigma = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$:

$$\begin{aligned} \mathbf{sw} : \Theta_{\mathbb{C}} &\to \Theta_{\mathbb{C}} \\ \theta = (\pi, T, E) &\mapsto (\pi\sigma,\, \sigma^{-1}T\sigma,\, \sigma^{-1}E) \end{aligned} \tag{1.8}$$

(In our case $\sigma^{-1} = \sigma$, but the form above generalizes to permutations of larger hidden alphabets.) Hence we have that $\Pr(v \,|\, \pi, T, E) = \Pr(v \,|\, \mathbf{sw}(\pi, T, E))$, i.e., $\phi_n = \phi_n \circ \mathbf{sw}$.

We will make essential use of a linear parametrization of $\Theta_{\mathbb{C}}$ in which $\mathbf{sw}$ has a simple form. Our new parameters will be $\eta_0 := (a_0, b, c_0, u, v_0)$, with subscript 0's

to be explained shortly. Although we have already used the letter v at times to represent visible binary strings, we hope that the context will be clear enough to avoid confusion between these usages. We let

$$\begin{aligned} \pi &= \frac{1}{2}\begin{bmatrix} 1-a_0, & 1+a_0 \end{bmatrix} \\ T &= \frac{1}{2}\begin{bmatrix} 1+b-c_0, & 1-b+c_0 \\ 1-b-c_0, & 1+b+c_0 \end{bmatrix} \qquad E = \begin{bmatrix} 1-u+v_0, & u-v_0 \\ 1-u-v_0, & u+v_0 \end{bmatrix} \end{aligned} \tag{1.9}$$

(The rightmost column of E is made intentionally homogeneous in the new parameters.) We can linearly solve for η_0 in terms of θ by $a_0 = \pi_1 - \pi_0$ etc., so in fact (a_0, b, c_0, u, v_0) generate the parameter ring $\mathbb{C}[\theta]$. In these coordinates, **sw** acts by

$$a_0 \mapsto -a_0, \qquad b \mapsto b, \qquad c_0 \mapsto -c_0, \qquad u \mapsto u, \qquad v_0 \mapsto -v_0$$

In other words, swapping the signs of the subscripted variables a_0, c_0, v_0 has the same effect as acting on the matrices π, T, E by σ as in (1.8), i.e., relabeling the hidden alphabet.

Introducing the birational parameters

Since $\phi_n \circ \mathbf{sw} = \phi_n$, the ring map $\phi_n^\# : \overline{R}_{p,n} \to \mathbb{C}[\theta]$ must land in the subring of invariants $\mathbb{C}[\theta]^{\mathbf{sw}} = \mathbb{C}[b, u, a_0^2, c_0^2, v_0^2, a_0c_0, a_0v_0, c_0v_0]$ (otherwise there would be a map f such that $f \circ \phi_n \circ \mathbf{sw} \neq f \circ \phi_n$, a contradiction).

However, $\phi_n^\#$ in fact factors through a smaller subring, conveniently generated by 5 elements:

Lemma 1.4.2 (Parameter Subring Lemma)**.** *For all $n \geq 3$, the ring map $\phi_n^\#$ lands in the subring $\mathbb{C}[\eta] := \mathbb{C}[a, b, c, u, v]$ of $\mathbb{C}[\theta]$, where $a = a_0v_0$, $c = c_0v_0$, $v = v_0^2$.*

The proof of this key lemma will be given in Section 1.5 after introducing trace algebras. To interpret its geometric consequences, write $\mathfrak{q}^\#$ for the subring inclusion

$$\mathfrak{q}^\# : \mathbb{C}[\eta] \hookrightarrow \mathbb{C}[\theta]$$

$$a \mapsto a_0v_0, \quad b \mapsto b, \quad c \mapsto c_0v_0, \quad u \mapsto u, \quad v \mapsto v_0^2,$$

write $\psi_n^\# : \overline{R}_{p,n} \to \mathbb{C}[\eta]$ for the factorization of $\phi_n^\#$ through $\mathfrak{q}^\#$, and write $\Theta'_{\mathbb{C}} := \operatorname{Spec} \mathbb{C}[\eta]$, so $\Theta'_{\mathbb{C}} \simeq \mathbb{C}^5$. The result:

Corollary 1.4.3. *The following diagram of dominant maps commutes*

$$\Theta_{\mathbb{C}} \xrightarrow{\mathfrak{q}} \Theta'_{\mathbb{C}} \xrightarrow{\psi_n} \overline{\mathrm{BHMM}}(n)$$
$$\Theta_{\mathbb{C}} \xrightarrow{\phi_n} \overline{\mathrm{BHMM}}(n)$$

and $\mathfrak{q}$ *is generically two-to-one.*

This corollary in particular implies the first part of the Birational Parameter Theorem (Theorem 1.4.1), by taking $\mathfrak{q} : \Theta_{\mathbb{C}} \to \Theta'_{\mathbb{C}} \simeq \mathbb{C}^5$ as the generically $2:1$ map.

Remark 1.4.4. The map $\mathfrak{q}$ is only dominant, and not surjective; for example, it misses the point $(1,0,0,0,0)$.

Corollary 1.4.5. *For all* $n \geq 3$, $\overline{\mathrm{BHMM}}(n) = \overline{\mathrm{image}}(\bar{\iota}_n \psi_n)$.

Proof. Since $\mathfrak{q}$ is dominant, $\overline{\mathrm{image}}(\bar{\iota}_n \psi_n) = \overline{\mathrm{image}}(\bar{\iota}_n \psi_n \mathfrak{q}) = \overline{\mathrm{image}}(\bar{\iota}_n \phi_n) = \overline{\mathrm{BHMM}}(n)$. □

The unique factorization map $\psi_n^{\#}$ can be computed directly in Macaulay2 for small n. The expressions in moment coordinates are simpler than in probabilities, so we present these in the following proposition.

Proposition 1.4.6. *The map* $\psi_3^{\#}$ *is given in moment coordinates by*

$$\begin{aligned}
m_\emptyset = m_{000} &\mapsto 1 \\
m_1 = m_{100} &\mapsto a + u \\
m_2 = m_{010} &\mapsto ab + c + u \\
m_3 = m_{001} &\mapsto ab^2 + bc + c + u \\
m_{12} = m_{110} &\mapsto abu + ac + au + cu + u^2 + bv \\
m_{13} = m_{101} &\mapsto ab^2u + abc + bcu + b^2v + ac + au + cu + u^2 \\
m_{23} = m_{011} &\mapsto ab^2u + abc + abu + bcu + c^2 + 2cu + u^2 + bv \\
m_{123} = m_{111} &\mapsto ab^2u^2 + 2abcu + abu^2 + bcu^2 + b^2uv + ac^2 + 2acu \\
&\qquad + c^2u + au^2 + 2cu^2 + u^3 + abv + bcv + 2buv
\end{aligned}$$

We will eventually prove the Birational Parameter Theorem (Theorem 1.4.1) by marginalization to the case $n = 3$, which we can prove here:

Proposition 1.4.7. *The following triangular set of equations hold on the graph of* ψ_3*, after clearing denominators, and can thus be used to recover parameters from*

observed moments where the denominators are non-zero:

$$\begin{aligned} b &= \frac{m_3 - m_2}{m_2 - m_1} \\ u &= \frac{m_1 m_3 - m_2^2 + m_{23} - m_{12}}{2(m_3 - m_2)} \\ a &= m_1 - u \\ c &= a - ba + m_2 - m_1 \\ v &= a^2 - \frac{m_1 m_2 - m_{12}}{b} \end{aligned}$$

(This proposition and the following corollary actually hold for all ϕ_n with $n \geq 3$, because of Proposition 1.5.2, and by the encoding of Section 1.2, these same formulae can be used to recover parameters for HMM$(2, k, n)$ when $k > 2$ as well.)

Proof. These equations can be checked with direct substitution by hand from Proposition 1.4.6. Regarding the derivation, they can be obtained in Macaulay2 by computing two Gröbner bases of the elimination ideal $I = \langle m_v - \phi_3(m_v) | v \in \{0,1\}^3 \rangle$ over the ring $R_{m,3}$, in Lex monomial order: once in the ring $R_{m,3}[v, c, a, b, u]$, and once in $R_{m,3}[v, c, u, b, a]$. Each variable occurs in the leading term of some generator in one of these two bases with a simple expression in moments as its leading coefficient. We solve each such generator (set to 0) for the desired parameter. □

Corollary 1.4.8. *The map $\psi_3 : \mathbb{C}^5 \to \overline{\text{BHMM}}(3)$ has a birational inverse ρ_3. The map $\rho_3^{\#}$ on moment coordinate functions is given by:*

$$a \mapsto \frac{m_2^2 + m_3 m_1 - 2m_2 m_1 - m_{23} + m_{12}}{2(m_3 - m_2)} \qquad u \mapsto \frac{-m_2^2 + m_3 m_1 + m_{23} - m_{12}}{2(m_3 - m_2)}$$

$$b \mapsto \frac{m_3 - m_2}{m_2 - m_1} \qquad v \mapsto \frac{\text{num}(v)}{4(m_3 - m_2)^2}$$

$$c \mapsto \frac{\text{num}(c)}{2(m_2 - m_1)(m_3 - m_2)}, \text{ where}$$

$$\begin{aligned} \text{num}(c) = &- m_1 m_2^2 + m_1^2 m_3 + m_2^2 m_3 - m_1 m_3^2 - m_1 m_{12} \\ &+ 2m_2 m_{12} - m_3 m_{12} + m_1 m_{23} - 2m_2 m_{23} + m_3 m_{23}, \text{ and} \\ \text{num}(v) = &\ m_2^4 - 2m_1 m_2^2 m_3 + m_1^2 m_3^2 - 2m_2^2 m_{12} - 2m_1 m_3 m_{12} + 4m_2 m_3 m_{12} \\ &+ 4m_1 m_2 m_{23} - 2m_2^2 m_{23} - 2m_1 m_3 m_{23} + m_{12}^2 - 2m_{12} m_{23} + m_{23}^2. \end{aligned}$$

Proof. This can be derived by substituting the solutions for u, a, and b in the previous propositions into the subsequent solutions for a, c, and v. Alternatively, it can be checked by direct substitution in Macaulay2, i.e., one computes that $\psi_3^{\#} \circ \rho^{\#}(\theta) = \theta$ for each birational parameter $\theta \in \{a, b, c, u, v\}$. □

The expressions in Corollary 1.4.8 are considerably simpler in moment coordinates than in probabilities. Comparing the number of terms, the numerators for a, b, c, u, v respectively have sizes 5, 2, 10, 4, and 12 in moment coordinates, versus sizes 22, 4, 56, 22, and 190 in probability coordinates. This explains in part why Macaulay2's Gröbner basis computations execute in moment coordinates with much less time and memory.

Statistical interpretation of the birational inverse ρ_3

It turns out that the factors appearing in the denominators of Corollary 1.4.8 defining ρ_3 have simple factorizations in terms of the rational and birational parameters:

- $m_3 - m_2$ appears in the denominator of all $\rho_3(\theta)$ except $\rho_3(b)$, and

$$m_3 - m_2 \overset{\psi_3}{\mapsto} (b)(ab - a + c) \overset{\mathfrak{q}}{\mapsto} (b)(v_0)(a_0 b - a_0 + c_0)$$

- $m_2 - m_1$ appears in the denominator of $\rho_3(b)$ and $\rho_3(c)$, and

$$m_2 - m_1 \overset{\psi_3}{\mapsto} ab - a + c \overset{\mathfrak{q}}{\mapsto} (v_0)(a_0 b - a_0 + c_0)$$

Let us pause to reflect on the meaning of these factors.

- The factor v_0 occurs in $det(E) = 2v_0$, hence $v = v_0^2 = 0$ if and only if the hidden Markov chain has "no effect" on the observed variables. The image locus $\phi_3(\{v_0 = 0\})$ can thus be modeled by a sequence of IID coin flips with distribution $E_0 = E_1 = (1 - u, u)$, so the BHMM is an unlikely model choice. This is a **one-dimensional submodel**, parametrizable by $u \in [0, 1]$, with a regular (everywhere-defined) inverse given simply by $u = m_1$. Denote this model by BIID(n).

- The factor b occurs in $det(T) = b$, hence $b = 0$ if and only if each hidden node has "no effect" on the subsequent hidden nodes. In this case, the observed process can be modeled as a sequence of independent coin flips, the first flip having distribution $(1 - \alpha, \alpha) := \pi E$ and subsequent flips being IID having distribution $(1 - \beta, \beta) := T_0 E = T_1 E$. The image locus $\phi_3(\{b = 0\})$ is hence

a **two-dimensional submodel**, parametrizable by $(\alpha, \beta) \in [0,1]^2$, with a regular inverse given by $\alpha = m_1, \beta = m_2$. Denote this model by BINID(n), for "binary independent nearly identically distributed" model, and note that BINID(n) $\supseteq$ BIID(n) by setting $\alpha = \beta$.

- The factor $a_0 b - a_0 + c_0$ occurs in $\pi T - \pi = \frac{1}{2}(-a_0 b + a_0 - c_0,\ a_0 b - a_0 + c_0)$. Hence $a_0 b - a_0 + c_0 = 0$ if and only if π is a fixed point of T, i.e., the hidden Markov chain is *at equilibrium*. We may define the Equilibrium Binary Hidden Markov model by restricting ϕ_n to the locus $\{a_0 b - a_0 + c_0 = 0\}$, which turns out to yield a **four-dimensional submodel** for each $n \geq 3$, paramterizable by a generically $2:1$ map from (a_0, b, u, v_0). Denote this submodel by EBHMM(n).

It can be easily shown, with the same methods used here for BHMM(n), that EBHMM(n) itself has a birational parametrization by $(a_0 v_0, b, u, v_0^2) = (a, b, u, v)$, where $a_0, b \in [-1, 1]$, $c_0 := a_0(1-b) \in [|b|-1, 1-|b|]$, $v_0 \in [0,1]$, and $u \in [|v_0|, 1-|v_0|]$, with an inverse parametrization given by

$$b = \frac{m_1^2 - m_{13}}{m_1^2 - m_{12}} \qquad u = \frac{2m_1 m_{12} - m_1 m_{13} - m_{123}}{2(m_1^2 - m_{13})}$$

$$a = m_1 - u \qquad v = \frac{a^2 b - m_1^2 + m_{12}}{b}$$

The newly occurring denominators here are $m_1^2 - m_{12} = (b)(a^2 - v) = (b)(v_0)^2(a_0^2 - 1)$ and $m_1^2 - m_{13} = (b)^2(a^2 - v) = (b)(v_0)^2(a_0^2 - 1)$. It easy to check that the only points of EBHMM(n) where these expressions vanish are points that lie in BINID(n). Thus, for $n \geq 3$, BHMM(n) can be stratified as a union of three statistically meaningful submodels

$$\begin{aligned} \text{BHMM}(n) &= \text{BINID}(n) && \leftarrow \text{2 dimensional} \\ &\cup \big(\text{EBHMM}(n) \setminus \text{BINID}(n)\big) && \leftarrow \text{4 dimensional} \\ &\cup \big(\text{BHMM}(n) \setminus \big(\text{EBHMM}(n) \cup \text{BINID}(n)\big)\big) && \leftarrow \text{5 dimensional} \end{aligned}$$

each of which has an everywhere-defined inverse parametrization.

Computational advantages of moments, cumulants, and birational parameters

Our approach has been to work with moments m_v and cumulants k_v instead of probabilities p_v, and the birational parameters a, b, c, u, v instead of the matrix entries

π_1, t_{i1}, e_{i1}. Other than the theoretical advantage that the model map is generically injective on the birational parameter space, significant computation gains in Macaulay2 also result from these choices (see Section 1.3 for laptop specifications):

- Computing $\ker \psi_3 = \ker \phi_3$, the affine defining ideal of $\overline{\mathrm{BHMM}}(3)$, took less than 1 second in Macaulay2 when using the birational parameters, compared to 25 seconds when using the matrix entries and moments, and 15 minutes when using the matrix entries and probabilities.

- Computing $\ker \psi_4 = \ker \phi_4$, the affine defining ideal of $\overline{\mathrm{BHMM}}(4)$ took less than 1 second in Macaulay2 when using the birational parameters and *cumulant coordinates* (see Sturmfels and Zwiernik [37]), compared to 1.5 hours when using the matrix entries and cumulant coordinates, and running out of memory (8gb) when using the matrix entries and probabilities.

Despite these advantages, we have been unsuccessful in computing a full generating set of invariants for $\overline{\mathrm{BHMM}}(5)$. We hope that further investigation into reparametrization methods will eventually lead to a solution in this and subsequent cases.

1.5 Parametrizing BHMMs though a trace variety

In this section, we exhibit a parametrization of every BHMM through a particular *trace variety* called $\operatorname{Spec} C_{2,3}$, which itself can be embedded in $\mathbb{C}^{10}$. We use these coordinates to prove the Birational Parameter Theorem (Theorem 1.4.1) and the Parameter Subring Lemma (1.4.2), which were stated without proof.

For this, we will define a map ϕ_∞ through which all the ϕ_n factor, and using a version of the Baum formula for moments, we factor this map further through $\operatorname{Spec} C_{2,3}$. Then we use a finite set 10 of generators of the ring $C_{2,3}$ exhibited by (see Sibirskii [33]) to show that the image of ϕ_∞ lands in the desired subring $\mathbb{C}[\eta]$, and write ψ_∞ for the factorization. Finally, by marginalizing to the case $n = 3$, we obtain a birational inverse for ψ_n from the map ρ_3 given in Corollary 1.4.8.

Marginalization maps

For each pair of integers $n' \geq n \geq 1$, the *marginalization map* $\mu_n^{n'} : \mathbb{C}_p^{2^{n'}} \to \mathbb{C}_p^{2^n}$ is given by

$$\mu_n^{n'\#}(p_v) := \sum_{|w|=n'-n} p_{vw}.$$

These restrict to maps $\mu_n^{n'} : \mathbb{C}_p^{2^{n'}-1} \to \mathbb{C}_p^{2^n-1}$, and define *rational* maps $\mu_n^{n'} : \mathbb{P}_p^{2^{n'}-1} \dashrightarrow \mathbb{P}_p^{2^n-1}$. In moment coordinates, these maps are actually coordinate projections: $\mu_n^{n'\#}(m_v) = m_{v\bar{0}}$ where $\bar{0}$ denotes a sequence of $n'-n$ zeros. In fact, using the subset notation for moments m_I, the corresponding ring maps are literal inclusions: $\mu_n^{n'\#}(m_I) = m_I$. In other words, $\mu_n^{n'} : \mathbb{C}_m^{2^{n'}} \to \mathbb{C}_m^{2^n}$ is just the map which forgets those m_I where $I \not\subseteq [n]$.

The Baum formula for moments

The evaluation of equation (1.1) requires $O(2^n)$ addition operations. There is a faster way to compute $\phi_n^{\#}(p_v)$, using $O(n)$ arithmetic operations, by treating the BHM process as a *finitary process* (see Schönhuth [32]). We define two new matrices[2]

$$(P_i)_{jk} := E_{ji}T_{jk} = \Pr(V_t = i \text{ and } H_{t+1} = k \,|\, H_t = j \text{ and } \pi, E, T), \text{ that is,}$$

$$P_0 := \begin{bmatrix} T_{00}E_{00} & T_{01}E_{00} \\ T_{10}E_{10} & T_{11}E_{10} \end{bmatrix} \quad \text{and} \quad P_1 := \begin{bmatrix} T_{00}E_{01} & T_{01}E_{01} \\ T_{10}E_{11} & T_{11}E_{11} \end{bmatrix}.$$

Writing $\mathbb{1}$ for the vector $\left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$ we obtain the matrix expression $\phi^{\#}(p_v) = \pi P_{v_1} P_{v_2} \cdots P_{v_v} \mathbb{1}$ whose evaluation requires only $4n+2$ multiplications and $2n+1$ additions. This is known as the Baum formula. We can rewrite this formula as a trace product of 2×2 matrices, where the first trace is actually the trace of a 1×1 matrix:

$$\phi^{\#}(p_v) = \text{trace}(\pi P_{v_1} P_{v_2} \cdots P_{v_n} \mathbb{1}) = \text{trace}((\mathbb{1}\pi) P_{v_1} P_{v_2} \cdots P_{v_n})$$

To create an analogue of this formula in moment coordinates, we let

$$M_0 := P_0 + P_1 = T \qquad M_1 := P_1 \qquad M_2 := \mathbb{1}\pi = \begin{bmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{bmatrix}.$$

[2] P can be thought of naturally as a $2 \times 2 \times 2$ tensor, but we will not make use of this interpretation.

Proposition 1.5.1 (Baum formula for moments)**.** *The binary hidden Markov map ϕ_n can be written in moment coordinates as*

$$\phi_n^{\#}(m_v) = \operatorname{trace}(M_2 M_{v_1} M_{v_2} \cdots M_{v_n}).$$

For example, $\phi_n^{\#}(m_{01001}) = \operatorname{trace}(M_2 M_0 M_1 M_0 M_0 M_1)$.

Proof. By our definition of m_v (1.3), we have

$$\begin{aligned}\phi_n^{\#}(m_v) &= \sum_{w \geq v} \phi_n^{\#}(p_w) = \sum_{w \geq v} \operatorname{trace}((\mathbb{1}\pi) P_{w_1} P_{w_2} \cdots P_{w_n}) \\ &= \operatorname{trace}\left((\mathbb{1}\pi)\left(\sum_{w_1 \geq v_1} P_{w_1}\right)\left(\sum_{w_2 \geq v_2} P_{w_2}\right) \cdots \left(\sum_{w_n \geq v_n} P_{w_n}\right)\right) \\ &= \operatorname{trace}(M_2 M_{v_1} M_{v_2} \cdots M_{v_n}).\end{aligned}$$

□

Truncation and ϕ_∞

Proposition 1.5.2. *The binary hidden Markov maps ϕ_n form a directed system of maps under marginalization, meaning that, for each $n' \geq n \geq 1$, the following diagrams commute:*

Proof. This can be seen directly from the definition of ϕ_n using (1.1) and of m_v in (1.3). Alternatively, observe that because $M_0 = T$ is stochastic, $M_0 M_2 = M_2$, so for any sequence $\overline{0}$ of length $n' - n$, the Baum formula for moments (Proposition 1.5.1) implies that

$$\phi_{n'}^{\#}(m_{v\overline{0}}) = \phi_n^{\#}(m_v) \tag{1.10}$$

□

Thus, to compute ϕ_n for all n, it is only necessary to compute those $\phi_n^{\#} m_{v'}$ where v' ends in 1. Motivated by this observation, let $\overline{R}_{m,\infty} := \mathbb{C}[m_{v1} \mid v \in$

$\{0,1\}^n$ for some $n \geq 0] = \mathbb{C}[m_1, m_{01}, m_{11}, m_{001}, m_{101}, m_{011}, \ldots]$, which in subset index notation is simply

$$\begin{aligned}\overline{R}_{m,\infty} &:= \mathbb{C}[m_I \mid I \subseteq [n] \text{ for some } n \geq 0] \\ &= \mathbb{C}[m_1, m_2, m_{12}, m_3, m_{13}, m_{23}, \ldots]\end{aligned}$$

Then we define $\phi_\infty : \Theta_\mathbb{C} \to \operatorname{Spec} \overline{R}_{m,\infty}$ and $\phi_\infty^\# : \mathbb{C}[\theta] \leftarrow \overline{R}_{m,\infty}$ by the formula $\phi_\infty^\#(m_{v1\bar{0}}) := \phi^\#_{\text{length}(v1)}(m_{v1})$, i.e.

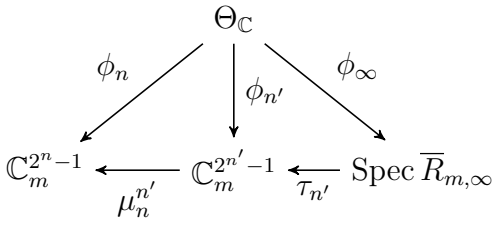
$$\phi_\infty^\#(m_I) := \phi^\#_{\text{size}(I)}(m_I) \tag{1.11}$$

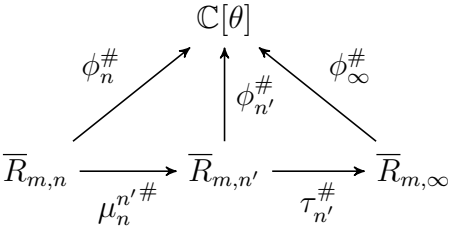
Note that by locating the position of the last 1 in a binary sequence $v' \neq 0 \ldots 0$, we can write v' in the form $v1\bar{0}$ for a unique string v (possibly empty if $v' = 1$), so this map is well-defined. By the same principle, for each n, we can also define a "truncation" map $\tau :$ $\operatorname{Spec} \overline{R}_{m,\infty} \to \mathbb{C}_m^{2^n-1}$ by $\tau^\#(m_{v1\bar{0}}) := m_{v1}$, which, in subset index notation, is a literal ring inclusion:

$$\tau^\#(m_I) := m_I \tag{1.12}$$

With this definition, $\phi_n^\#$ factorizes as $\phi_n^\# = \phi_\infty^\# \circ \tau_n^\#$. We can summarize this and Proposition 1.5.2 as follows:

Proposition 1.5.3. *For all $n' \geq n \geq 1$, the following diagrams commute:*





Remark 1.5.4. These diagrams exhibit the rings $\overline{R}_{m,n}$ and maps $\phi_n^\#$ as a directed system under the inclusion maps $\mu_n^{n'\#}$, such that $\overline{R}_{m,\infty} = \operatorname{colim}_{n\to\infty} \overline{R}_{m,n}$ and $\phi_\infty^\# = \lim_{n\to\infty} \phi_n^\#$.

Now, to prove that ϕ_n factors through $\mathfrak{q}$, we need only show that ϕ_∞ does.

Factoring ϕ_∞ through a trace variety

Let X_0, X_1, X_2 be 2×2 matrices of indeterminates,

$$X_0 = \begin{bmatrix} x_{000} & x_{001} \\ x_{010} & x_{011} \end{bmatrix} \qquad X_1 = \begin{bmatrix} x_{100} & x_{101} \\ x_{110} & x_{111} \end{bmatrix} \qquad X_2 = \begin{bmatrix} x_{200} & x_{201} \\ x_{210} & x_{211} \end{bmatrix}$$

and following the notation of Drensky [13], $\Omega_{2,3} := \mathbb{C}[\text{entries of } X_0, X_1, X_2]$ denotes the polynomial ring on the entries x_{ijk} of these three 2×2 matrices. The *trace algebra* $C_{2,3}$ is defined as the subring of $\Omega_{2,3}$ generated by the traces of products of these matrices, $C_{2,3} := \mathbb{C}[\text{trace}(X_{i_1}X_{i_2}\cdots X_{i_r}) \mid r \geq 1] \subseteq \Omega_{2,3}$ and we refer to $\operatorname{Spec} C_{2,3}$ as a *trace variety*. We write

$$\nu : \operatorname{Spec}\Omega_{2,3} \to \operatorname{Spec} C_{2,3} \qquad \text{and} \qquad \nu^{\#} : C_{2,3} \hookrightarrow \Omega_{2,3}$$

for the natural dominant map and corresponding ring inclusion. To relate these varieties to binary HMMs , we define two new maps $\omega^{\#} : \Omega_{2,3} \to \mathbb{C}[\theta]$ and $\xi^{\#} : \overline{R}_{m,\infty} \to C_{2,3}$ by

$$\omega^{\#}(X_i) := M_i \qquad \text{and} \qquad \xi^{\#}(m_{v1}) := \operatorname{trace}\left(\left(X_2 \prod_{i\in v} X_i\right) X_1\right).$$

Proposition 1.5.5 (Baum factorization)**.** *The ring map $\phi^{\#}_{\infty}$ factorizes as $\phi^{\#}_{\infty} = \omega^{\#} \circ \nu^{\#} \circ \xi^{\#}$, i.e., the following diagram commutes:*

$$\begin{array}{ccc} \Theta_{\mathbb{C}} & \xrightarrow{\phi_\infty} & \operatorname{Spec}\overline{R}_{m,\infty} \\ {\scriptstyle\omega}\downarrow & & \uparrow{\scriptstyle\xi} \\ \operatorname{Spec}\Omega_{2,3} & \xrightarrow[\nu]{} & \operatorname{Spec} C_{2,3} \end{array}$$

Proof. This is just a restatement of the Baum formula for moments (Proposition 1.5.1) :

$$\begin{aligned} \omega^{\#}(\nu^{\#}(\xi^{\#}(m_{v1}))) &= \omega^{\#} \operatorname{trace}\left(X_2 \prod_{i\in v1} X_i\right) = \operatorname{trace}\left(M_2 \prod_{i\in v1} M_i\right) \\ &= \phi^{\#}_{length(v1)}(m_{v1}) = \phi^{\#}_{\infty}(m_{v1}) \end{aligned}$$

□

Proving the Parameter Subring Lemma

We begin by seeking a factorization of the map $\omega^{\#} \circ \nu^{\#}$. For this we apply the following commutative algebra result of Sibirskii on the trace algebras $C_{2,r}$:

Proposition 1.5.6 (Sibirskii, 1968)**.** *The trace algebra $C_{2,r}$ is generated by the elements*

$$\begin{aligned} \operatorname{trace}(X_i) &: 0 \le i \le r, \\ \operatorname{trace}(X_i X_j) &: 0 \le i \le j \le r, \\ \operatorname{trace}(X_i X_j X_k) &: 0 \le i < j < k \le r. \end{aligned}$$

Corollary 1.5.7. *The algebra $C_{2,3}$ is generated by the 10 elements*

$$\begin{aligned} &\operatorname{trace}(X_0), \operatorname{trace}(X_1), \operatorname{trace}(X_2), \\ &\operatorname{trace}(X_0^2), \operatorname{trace}(X_1^2), \operatorname{trace}(X_2^2), \operatorname{trace}(X_0X_1), \operatorname{trace}(X_0X_2), \operatorname{trace}(X_1X_2), \\ &\operatorname{trace}(X_0X_1X_2). \end{aligned}$$

Proposition 1.5.8. *The ring map $\omega^\# \circ \nu^\#$ factors through the inclusion*

$$\mathfrak{q}^\# \ : \ \mathbb{C}[\eta] := \mathbb{C}[a,b,c,u,v] \ \hookrightarrow \ \mathbb{C}[\theta] := \mathbb{C}[a_0,b,c_0,u,v_0],$$

i.e., we can write $\omega^\# \circ \nu^\# = \mathfrak{q}^\# \circ \mathfrak{r}^\#$ so that the following diagram commutes:

$$\begin{array}{ccc} \Theta_{\mathbb{C}} & \xrightarrow{\ \mathfrak{q}\ } & \Theta'_{\mathbb{C}} \\ {\scriptstyle\omega}\downarrow & & \downarrow{\scriptstyle\mathfrak{r}} \\ \operatorname{Spec}\Omega_{2,3} & \xrightarrow[\ \nu\ ]{} & \operatorname{Spec} C_{2,3} \end{array}$$

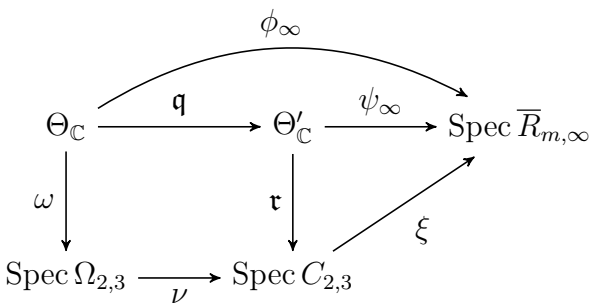
Proof. We apply $\omega^\#$ to the ten generators of $C_{2,3}$ given in Corollary 1.5.7 and check that they land in $\mathbb{C}[\eta]$. Explicit, we find that:

$$\begin{array}{lll} \operatorname{trace}(M_0) = b+1 & \operatorname{trace}(M_1) = bu+c+u & \operatorname{trace}(M_2) = 1 \\ \operatorname{trace}(M_0^2) = b^2+1 & \operatorname{trace}(M_1^2) = b^2u^2+2bcu{+}c^2+2cu+u^2+2bv & \\ \operatorname{trace}(M_2^2) = 1 & \operatorname{trace}(M_0M_1) = b^2u+bc+c+u & \operatorname{trace}(M_0M_2) = 1 \\ \operatorname{trace}(M_1M_2) = a+u & \operatorname{trace}(M_0M_1M_2) = ab+c+u & \square \end{array}$$

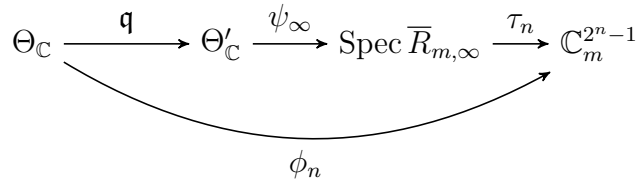
Now, by letting $\psi_\infty^\# := \mathfrak{r}^\# \circ \xi^\#$ we may factor the ring map $\phi_\infty^\#$ as

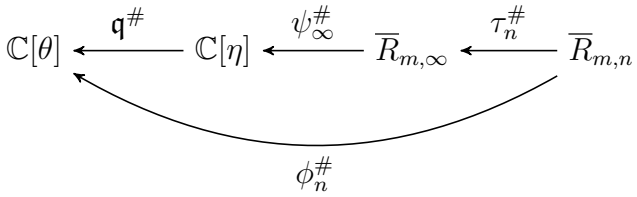
$$\phi_\infty^\# = \omega^\# \circ \nu^\# \circ \xi^\# = \mathfrak{q}^\# \circ \mathfrak{r}^\# \circ \xi^\# = \mathfrak{q}^\# \circ \psi_\infty^\#.$$

Corollary 1.5.9. *The following diagram commutes:*



Proof of the Parameter Subring Lemma. Proposition 1.5.3 and Corollary 1.5.9 together imply that the following diagrams commute:





In particular, the map $\phi_n^\#$ factors through $\mathbb{C}[\eta]$, as required. □

Proving the Birational Parameter Theorem (Theorem 1.4.1)

Recall that Corollary 1.4.3 implies the first part of the Birational Parameter Theorem (Theorem 1.4.1), by taking

$$\mathfrak{q} : \Theta_{\mathbb{C}} \longrightarrow \Theta'_{\mathbb{C}}$$

as the generically $2:1$ map. Thus, it remains to show that the maps

$$\psi_n : \Theta'_{\mathbb{C}} \longrightarrow \overline{\mathrm{BHMM}}(n)$$

have birational inverses ρ_n. The inverse map ρ_3 was already exhibited in Corollary 1.4.8, and we obtain ρ_n by marginalization: let

$$\rho_n = \rho_3 \circ \mu_3^n.$$

Let $U \subseteq \Theta'_{\mathbb{C}}$ be the Zariski open set on which ψ_3 is an isomorphism with inverse ρ_3. Consider the set $\psi_n(U) \subseteq \overline{\text{BHMM}}(n)$. It is Zariski dense in $\overline{\text{BHMM}}(n)$, and by Chevalley's theorem (EGAIV, 1.8.4), it is constructible, so it must contain a dense open set $W' \subseteq \overline{\text{BHMM}}(n)$. Now let $W = \psi_n^{-1}(W')$, so we have $\psi_n(W) = W' \subseteq \psi_n(U)$.

Proposition 1.5.10. *$\rho_n \circ \psi_n = Id$ on W and $\psi_n \circ \rho_n = Id$ on W'.*

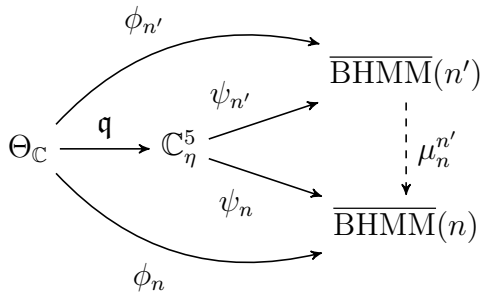
Proof. Suppose $\widehat{\eta} \in W$. Then $\rho_n \circ \psi_n(\widehat{\eta}) = \rho_3 \circ \mu_3^n \circ \psi_n(\widehat{\eta}) = \rho_3 \circ \psi_3(\widehat{\eta}) = \widehat{\eta}$ since $\widehat{\eta} \in U$. Now suppose $\widehat{p} \in W'$, so $\widehat{p} = \psi_n(\widehat{\eta})$ for some $\widehat{\eta} \in W$. Then, applying Proposition 1.5.2,

$$\begin{aligned}\psi_n \circ \rho_n(\widehat{p}) &= \psi_n \circ \rho_n \circ \psi_n(\widehat{\eta}) = \psi_n \circ \rho_3 \circ \mu_3^n \circ \psi_n(\widehat{\eta}) \\ &= \psi_n \circ \rho_3 \circ \psi_3(\widehat{\eta}) = \psi_n(\widehat{\eta}) = \widehat{p}\end{aligned}$$

□

This completes the proof of the Birational Parameter Theorem (Theorem 1.4.1). In fact we have also proven the following:

Theorem 1.5.11. *For any $n' \geq n \geq 3$, there is a commutative diagram of dominant maps:*



1.6 Applications and future directions

Besides attempting to compute a set of generators for $\text{I}_{\text{BHMM}(5)}$, there are many other questions to be answered about HMMs that can be approached immediately with the techniques of this chapter.

A nonnegative distribution in $\overline{\text{BHMM}}(3)$ but not $\text{BHMM}(3)$

It turns out that not all of the probability distributions (non-negative real points) of $\overline{\text{BHMM}}(n)$ lie in the model $\text{BHMM}(n)$. In other words, $\overline{\text{BHMM}}(n) \cap \Delta_p^{2^n-1} \neq \text{BHMM}(n)$, so the model must be cut out by some non-trivial inequalities inside the simplex. To illustrate this, the following real point $\widehat{\theta}$ of $\Theta_{\mathbb{C}}$ does not lie in Θ_{st}, but maps under ϕ_3 to a point $\widehat{p}$ of Δ_p^7:

$$\widehat{\theta} = (\widehat{\pi}, \widehat{T}, \widehat{E}) = \left(\begin{bmatrix} -\frac{1}{8} & \frac{9}{8} \end{bmatrix}, \quad \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}, \quad \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \right) \tag{1.13}$$

Moreover, the analysis of Section 1.4 reveals that the fiber $\phi_3^{-1}(\widehat{p})$ consists only of the point $\widehat{\theta}$ and the "swapped" point

$$\widehat{\theta'} = (\widehat{\pi}', \widehat{T}', \widehat{E}') = \left(\begin{bmatrix} \frac{9}{8} & -\frac{1}{8} \end{bmatrix}, \quad \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}, \quad \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix} \right) \tag{1.14}$$

which is also not in Θ_{st}. Hence the image point $\widehat{p} = \phi_3(\widehat{\theta}) = \phi_3(\widehat{\theta'})$ is a non-negative point of $\overline{\text{BHMM}}(3)$ that does not lie in $\text{BHMM}(3)$.

A semialgebraic model membership test

In light of the fact that not every nonnegative distribution in $\overline{\text{BHMM}}(n)$ is in $\text{BHMM}(n)$, the defining equations of $\overline{\text{BHMM}}(n)$ are not sufficient to test a probability distribution for membership to the model. Using the encoding of Section 1.2, membership to $\text{HMM}(2, k, n)$ can be tested by reductions to the $k = 2$ case to recover the parameters.

So, suppose we are given a distribution $p \in \Delta_p^{2^n-1}$ and asked to determine whether $p \in \text{BHMM}(n)$. The following procedure yields either

(1) a proof by contradiction that $p \notin \text{BHMM}(n)$,

(2) a parameter vector $\theta \in \Theta_{\text{st}}$ such that $\phi_n(\theta) = p \in \text{BHMM}(n)$, or

(3) a reduction of the question to whether p lies in one of the lower-dimensional submodels of $\text{BHMM}(n)$ discussed in Section 1.4.

How to proceed from (3) is essentially the same as what follows, using the birational parameterizations of the respective submodels given in Section 1.4, which will ultimately lead to case (1) or (2).

To begin, we let $p' = \mu_3^n(p) \in \Delta_p^{2^3-1}$, i.e., we marginalize p to the distribution p' it induces on the first three visible nodes. Note that if $p \in \mathrm{BHMM}(n)$ then $p' \in \mathrm{BHMM}(3)$. Observing the moments m_I of p', if any denominators in the formulae of Corollary 1.4.8 vanish, then we end in case (3).

Otherwise, we let $(a, b, c, u, v) = \psi_3^{-1}(p')$, choose v_0 to be either square root of v, and let $a_0 = a/v_0$, $c_0 = c/v_0$. If p were due to some BHM process, then by Theorem 1.5.11, these would be its parameters, up to a simultaneous sign change of (a_0, b_0, v_0). With this in mind, we define $\theta = (\pi, T, E)$ using (1.9). If (π, T, E) are not non-negative stochastic matrices, then $p \notin \mathrm{BHMM}(n)$ and we end in case (1). If they are, we compute $p'' = \phi_n(\theta)$, and if $p = p''$ then we end in case (2). Otherwise p must not have been in $\mathrm{BHMM}(n)$, so we end in case (1).

Note that since all the criteria in this test are algebraic equalities and inequalities, this procedure implicitly describes a semialgebraic characterization of $\mathrm{BHMM}(n)$ for all $n \geq 3$.

Identifiability of parameters

By a *rational map* on a possibly non-algebraic subset $\Theta \subseteq \mathbb{C}^k$, we mean any rational map on the Zariski closure of Θ, which will necessarily be defined as a function on a Zariski dense open subset of Θ. We define polynomial maps on Θ similarly.

Let $\phi : \Theta \to \mathbb{C}^n$ be an algebraic statistical model, where as usual we assume $\Theta \subseteq \mathbb{C}^k$ is Zariski dense, and therefore Zariski irreducible. A (rational) parameter of the model is any rational map $s : \Theta \to \mathbb{C}$. Such parameters form a field, $K \simeq \mathrm{Frac}(\mathbb{C}^k)$. In applications, for example in the work of Meshkat, Eisenberg, and DiStefano [23], it is important to know to what extent a parameter can be identified from observational data alone. In other words, given $\phi(\theta)$, what can we say about $s(\theta)$? This leads to the following notions of parameter identifiability, as discussed by Sullivant, Garcia-Puente, and Spielvogel [38], each of which implies the next:

Definition 1.6.1. We say that a rational parameter $s \in K$ is

- *(set-theoretically) identifiable* if $s = \sigma \circ \phi$ for some set-theoretic function $\sigma : \phi(\Theta) \to \mathbb{C}$. In other words, for all $\theta, \theta' \in \Theta$, if $\phi(\theta) = \phi(\theta')$ then $s(\theta) = s(\theta')$.
- *rationally identifiable* if $s = \sigma \circ \phi$ for some *rational* map $\sigma : \phi(\Theta) \to \mathbb{C}$ (this notion is used without a name by Sullivant, Garcia-Puente, and Spielvogel [38]).
- *generically identifiable* if there is a (relatively) Zariski dense open subset $U \subseteq \Theta$ such that $s|_U = \sigma \circ \phi|_U$ for some set-theoretic function $\sigma : \phi(U) \to \mathbb{C}$.

- *algebraically identifiable* if there is a polynomial function $g(p,q) := \sum_i g_i(p_1,\ldots,p_n)q^i$ on $\phi(\Theta)\times\mathbb{C}$ of degree $d>0$ in q (so that g_d is not identically 0 on $\phi(\Theta)$) such that $g(\phi(\theta),s(\theta))=0$ for all $\theta\in\Theta$ (and hence all $\theta\in\mathbb{C}^k$).

For example, from Proposition 1.4.7, we see that for BHMM(n) with $n\geq 3$, the parameter b is rationally identifiable, and therefore generically and algebraically identifiable. The parameter v_0 is algebraically identifiable, because $v_0^2=v$ and v is rationally identifiable. In general, we can ask question:

Question 1.6.2. What combinations of BHM parameters are rationally identifiable, generically identifiable, or algebraically identifiable?

To answer this question we introduce a lemma on algebraic statistical models in general:

Lemma 1.6.3. *For any algebraic statistical model ϕ as above, the sets K_{ri}, K_{gi}, and K_{ai}, of rationally, generically, and algebraically identifiable parameters, respectively, are all fields.*

Proof. Since Θ is Zariski irreducible, so is $\phi(\Theta)$. Hence the set of rational maps on $\phi(\Theta)$ is simply the fraction field of its Zariski closure (an irreducible variety), and K_{ri} is the image of this field under $\phi^{\#}$, which must be a field.

For K_{gi}, the crux is to show that if $s,s'\in K_{gi}$ and $s\neq 0$ then $s'/s\in K_{gi}$. Let $U\subseteq\Theta$ and $\sigma:\phi(U)\to\mathbb{C}$ be as in the definition for s, and likewise $U'\subseteq\Theta$ and $\sigma:\phi(U')\to\mathbb{C}$ for s'. Let $U''=\{\theta\in U\cap U' \mid s(\theta)\neq 0\}$, which, being an intersection of three Zariski dense open subsets of Θ, is a dense open. We have $\sigma\neq 0$ on $\phi(U'')\subseteq\phi(U)\cap\phi(U')$, so we can let $\sigma''=\sigma'/\sigma:\phi(U'')\to\mathbb{C}$, and then $\sigma''\circ\phi=s'/s$, so $s'/s\in K_{gi}$. Thus K_{gi} is stable under division, and simpler arguments show it is stable stable under $+,-$, and $\cdot$.

Finally, K_{ai} is expressly the relative algebraic closure in K of the image under $\phi^{\#}$ of the coordinate ring of $\phi(\Theta)$, which is therefore a field. □

Proposition 1.6.4. *For any algebraic statistical model ϕ as above, $K_{ri}\subseteq K_{gi}\subseteq K_{ai}\subseteq K$.*

Proof. This is now just a restatement of Proposition 3 by Sullivant, Garcia-Puente, and Spielvogel [38]. □

Now, the answer to our identifiability question for BHM parameters can be given easily in the coordinates of Section 1.4. Here ϕ is the BHM map ϕ_n. The field K_{ri} is simply the image $\mathfrak{q}^{\#}(\mathrm{Frac}(\Theta'_{\mathbb{C}}))$ because by Theorem 1.4.1,

$$\psi^{\#}:\mathrm{Frac}(\overline{\mathrm{BHMM}}(n))\to\mathrm{Frac}(\Theta'_{\mathbb{C}})$$

is an isomorphism. Hence the rationally identifiable parameters are precisely the field of rational functions in $(a, b, c, u, v) = (a_0v_0, b, c_0v_0, u, v_0^2)$ (see (1.9) for the meanings of these parameters). Since K is a quadratic field extension of K_{ri} given by adjoining $v_0 = \sqrt{v}$, and K_{ai} is the algebraic closure of K_{ri} in K (almost by definition), it follows that $K_{ai} = K$, i.e., *all parameters* are algebraically identifiable. Finally, we observe that, by the action of **sw** in Section 1.4, there are generically two possible values of $v_0 = \frac{1}{2}(E_{11} - E_{01})$ for a given observed distribution, namely $\pm\sqrt{v}$. Hence $v_0 \notin K_{gi}$, and since a quadratic field extension has no intermediate extensions, it follows that $K_{ri} = K_{gi}$, i.e., all generically identifiable parameters are in fact rationally identifiable. In summary,

Proposition 1.6.5. *For* BHMM(n) *where* $n \geq 3$,

$$\mathbb{C}(a, b, c, u, v) = K_{ri} = K_{gi} \subsetneq K_{ai} = \mathbb{C}(a_0, b, c_0, u, v_0)$$

Remark 1.6.6. In a sense, the only obstruction between K_{ai} and K_{gi} is the label swapping ambiguity, in the sense that K_{gi} is the fixed field of K_{ai} under the action of **sw**.

A new grading on BHMM invariants

The re-parametrized model map ψ_n is homogeneous in cumulant and moment coordinates, with respect to a $\mathbb{Z}$-grading where $deg(m_v) = deg(k_v) = sum(v)$, $deg(b) = 0$, $deg(a) = deg(c) = deg(u) = 1$, and $deg(v) = 2$. This grading allows for fast linear algebra techniques that solve for low degree model invariants as in Bray and Morton [8], except that this grading is intrinsic to the model. Bray and Morton's grading, which is in *probability* coordinates, is not on the binary HMM proper, but on a larger variety obtained by relaxing the parameter constraints that the transition and emission matrix row sums are 1. The invariants obtained in their search are hence invariants of this larger variety, and exclude some invariants of BHMM(n). The grading presented here can thus be used to complete their search for invariants up to any finite degree.

Equilibrium BHM processes

In Section 1.4 we found that if a BHM process is at equilibrium, our formula for ψ_3^{-1} is undefined. We may define Equilibrium Binary Hidden Markov Models, EBHMMs, by restricting ϕ_n to the locus $\{a_0b - a_0 + c_0 = 0\}$, which turns out to yield a four-dimensional submodel of BHMM(n) for each $n \geq 3$. The same techniques used here to study BHMMs have revealed that the EBHMMs, too, have birational

parameterizations, and the ideal of EBHMM(3) is generated by the equations $m_1 = m_2 = m_3$ and $m_{12} = m_{13}$. The geometry of EBHMMs will need to be considered explicitly in future work to identify the learning coefficients of BHMM fibers.

Larger hidden Markov models

As we have remarked throughout, many results on BHMM(n) can be readily applied to HMM($2, k, n$), i.e., HMMs with two hidden states and k visible states. For example, consider the parameter identification problem. We may specify the process by a $2 \times k$ matrix E of emission probabilities, along with a triple (a_0, b, c_0) defining the π and T of the two-state hidden Markov chain as in (1.9). As in Section 1.2, to obtain $E_{0\ell}$ and $E_{1\ell}$ from the observed probability distribution for any fixed ℓ, we simply define a BHM process where $V_t^\ell = 1$ if $V_t = \ell$ and $V_t^\ell = 0$ otherwise. Applying Proposition 1.4.7 to the moments of the distribution yields values for (a, b, c, u, v) provided that the denominators involved do not vanish. Letting $v_0 = \sqrt{v}$, $a_0 = a/v_0$, and $c_0 = c/v_0$, we obtain (a_0, b, c_0, u, v_0) up to a simultaneous sign change on (a_0, c_0, v_0) corresponding to swapping the hidden alphabet as in Section 1.4. Then $E_{0\ell} = u - v$ and $E_{1\ell} = u + v$, and we get π, T as well from (a_0, b, c_0). We can repeat this for each $\ell = 1, \ldots, k$ to obtain all the emission parameters, and hence identify all the process parameters modulo the swapping operation.

As well, as described in Section 1.3 we can obtain

$$50k[(n-3) + (n-7) + \cdots + (n - 3\lfloor \frac{n-1}{3} \rfloor)]$$

polynomial invariants of HMM($2, k, n$) by reducing to BHMM(n) as above, and marginalizing to collections of 4 equally-spaced visible nodes to obtain points of BHMM(4) at which we know the invariants of Theorem 1.3.1 will vanish.

Given these extensions, one can hope that algebraic techniques similar to those used here could elucidate the geometry of HMMs with any number of hidden states as well.

Chapter 2

Matrix product state models with two-dimensional virtual bonds

This chapter is based on joint work with Jason Morton.

Matrix product states (MPS) provide a useful model of 1-D quantum spin systems which approximate the ground states of gapped local Hamiltonians [39]. Accordingly the problem of classifying phases of matter for such chains has been reduced to understanding equivalence classes (such as under LU operations) in the space of quantum states representable as matrix product states [40, 17, 10].

With periodic or open boundary conditions, we describe the closure of this space of states representable by translation invariant binary MPS as an algebraic variety. Our description is given as an ideal of polynomials in the state's amplitudes that vanish if and only if the state is a limit of MPS with N spins and $D = d = 2$ dimensional virtual and physical bonds. In small cases our description is complete. In Section 2.1, we exhibit a polynomial which vanishes on a pure state if and only if it is a limit of binary translation invariant, periodic boundary MPS with $N = 4$, and a set of 30 polynomials which vanish when $N = 5$. We also obtain many linear equations which are satisfied for N up to 12. In Section 2.2, Theorem 2.2.1 gives an analogous result for MPS with open boundary conditions and $N = 3$. Finally we examine cases where $N \gg 0$.

Matrix product states bear a close relationship to probabilistic graphical models known as a *hidden Markov models* (HMM) [26]. In Section 2.3, we make this relationship precise by modifying the parametrization of HMM to obtain MPS. We review the invariant theory of trace identities and trace varieties that was used to study HMM in Chapter 1, and how these results apply to varieties of MPS. In particular we obtain a nice parametrization for translation invariant binary MPS with periodic

boundary conditions. Finally in Section 2.4 we suggest other such relationships between probabilistic graphical models and tensor network state models. Our results are complimentary to the connection between invariant theory and diagrammatic representations explored in [7] and the approaches to quantum state tomography for MPS developed in [5, 18].

2.1 Representability by translation invariant matrix product states

First consider a translation-invariant matrix product state with *periodic boundary* conditions. Suppose the inner (virtual) bond dimension is D, the outer (physical) bond dimension is d, and there are N spins. Fix $D \times D$ complex parameter matrices $A_0, \ldots, A_{d-1}$, defining the same $D \times D \times d$ parameter tensor at each site. This defines the tensor network state, for $i_j \in \{0, \ldots, d-1\}$,

$$\Psi = \sum_{i_1,\ldots,i_N} \operatorname{tr}(A_{i_1} \cdots A_{i_N}) |i_1 i_2 \cdots i_N\rangle. \tag{2.1}$$

Question 2.1.1. Fixing virtual and physical bond dimension, which states are matrix product states?

Including states which are limits of MPS, a precise answer to this question could be given as a constructive description of the set of polynomials f in the coefficients of Ψ such that $f(\psi_{i_1,\ldots i_N}) = 0$ if and only if ψ is a limit of MPS. This would describe the (closure of the) set of MPS as an algebraic variety. See [11] for background on varieties.

Such a description is possible because of the way MPS are defined. Each coefficient $\psi_{i_1,\ldots,i_N}$ is a polynomial function of the parameters a_{rst} in the $D \times D \times d$ tensor A. Thus (2.1) defines a regular map $\Psi : \mathbb{C}^{D^2 d} \to \mathbb{C}^{d^N}$, whose image we denote by $\mathrm{PB}(D, d, N)$, the set of tensors representable by translation-invariant matrix product states with periodic boundary conditions. Its closure $\overline{\mathrm{PB}}(D, d, N)$ in either the Zariski or classical topology is an irreducible algebraic variety consisting of those tensors which can be approximated *arbitrarily well* by MPS. We can thus refine Question 2.1.1 as follows.

Question 2.1.2. Fixing, D, d, and N, what polynomial relations must the coefficients of a matrix product state satisfy: what is the defining ideal of $\overline{\mathrm{PB}}(D, d, N)$?

We primarily examine the fully binary case $D = d = 2$. The invariance of trace under cyclic permutations of the matrices $A_{i_1}, \ldots, A_{i_N}$ means we can immediately

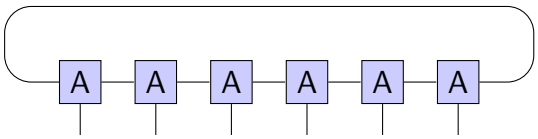


Figure 2.1: Translation-invariant MPS with periodic boundary.

restrict to the subspace spanned by *binary necklaces* (equivalence classes of binary strings under cyclic permutation). For $N = 3$ physical legs, this is the coordinate subspace $(\psi_{000} : \psi_{100} : \psi_{110} : \psi_{111})$ and all three-qubit states with cyclic symmetry are matrix product states. For $N = 4$ it is the six-dimensional coordinate subspace $(\psi_{0000} : \psi_{1000} : \psi_{1100} : \psi_{1010} : \psi_{1110} : \psi_{1111})$ and not all states are MPS (Theorem 2.1.3). In the $N = 5$ case the 8 equivalence classes of coefficients under cyclic permutation are ψ_{00000}, ψ_{10000}, ψ_{11000}, ψ_{11100}, ψ_{11110}, ψ_{11111}, ψ_{10100}, and ψ_{11010}.

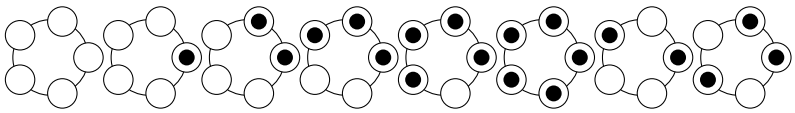


Figure 2.2: The eight binary necklaces for $N = 5$.

For $N = 6, \ldots, 15$ the dimensions of this "necklace space" are 14, 20, 36, 60, 108, 188, 352, 632, 1182, and 2192. In general the number of d-ary necklaces of length N is given by the following formula of Moreau [24]:

$$n_d(N) = \frac{1}{N} \sum_{\ell | N} \varphi(\ell) d^{N/\ell},$$

where φ is Euler's totient function. Thus translation invariant MPS with periodic boundary of length N and physical bond dimension d live in a linear space isomorphic to $\mathbb{C}^{n_d(N)}$.

Naïvely we have 8 parameters in our 2×2×2 tensor A, but on each virtual bond we can apply a gauge transformation $P(\,\cdot\,)P^{-1}$ for $P \in SL_2$ without changing the state [29]. Since SL_2 is 3-dimensional, we expect $\overline{\mathrm{PB}}(2, 2, 3)$ to be 5-dimensional. Counting this way, our *expected dimension* of $\overline{\mathrm{PB}}(D, d, N)$ is $\min\{D^2(d-1)+1, n_d(N)\}$. We expect $\overline{\mathrm{PB}}(D, d, N)$ to be a hypersurface when this equals $n_d(N)$, which happens first when $(D, d, N) = (2, 2, 4)$. In this case our expectation holds:

Theorem 2.1.3. *A four-qubit state Ψ is a limit of binary periodic translation invariant MPS with $N = 4$ if and only if the following irreducible polynomial vanishes:*

$$\begin{aligned}
&\psi_{1010}^2\psi_{1100}^4 - 2\psi_{1100}^6 - 8\psi_{1000}\psi_{1010}\psi_{1100}^3\psi_{1110} + 12\psi_{1000}\psi_{1100}^4\psi_{1110} \\
&- 4\psi_{1000}^2\psi_{1010}^2\psi_{1110}^2 + 2\psi_{0000}\psi_{1010}^3\psi_{1110}^2 + 16\psi_{1000}^2\psi_{1010}\psi_{1100}\psi_{1110}^2 \\
&- 4\psi_{0000}\psi_{1010}^2\psi_{1100}\psi_{1110}^2 - 16\psi_{1000}^2\psi_{1100}^2\psi_{1110}^2 + 4\psi_{0000}\psi_{1010}\psi_{1100}^2\psi_{1110}^2 \\
&- 4\psi_{0000}\psi_{1100}^3\psi_{1110}^2 - 4\psi_{0000}\psi_{1000}\psi_{1010}\psi_{1110}^3 + 8\psi_{0000}\psi_{1000}\psi_{1100}\psi_{1110}^3 \\
&- \psi_{0000}^2\psi_{1110}^4 + 2\psi_{1000}^2\psi_{1010}^3\psi_{1111} - \psi_{0000}\psi_{1010}^4\psi_{1111} - 4\psi_{1000}^2\psi_{1010}^2\psi_{1100}\psi_{1111} \\
&+ 4\psi_{1000}^2\psi_{1010}\psi_{1100}^2\psi_{1111} + 2\psi_{0000}\psi_{1010}^2\psi_{1100}^2\psi_{1111} - 4\psi_{1000}^2\psi_{1100}^3\psi_{1111} \\
&+ \psi_{0000}\psi_{1100}^4\psi_{1111} - 4\psi_{1000}^3\psi_{1010}\psi_{1110}\psi_{1111} + 4\psi_{0000}\psi_{1000}\psi_{1010}^2\psi_{1110}\psi_{1111} \\
&+ 8\psi_{1000}^3\psi_{1100}\psi_{1110}\psi_{1111} - 8\psi_{0000}\psi_{1000}\psi_{1010}\psi_{1100}\psi_{1110}\psi_{1111} \\
&- 2\psi_{0000}\psi_{1000}^2\psi_{1110}^2\psi_{1111} + 2\psi_{0000}^2\psi_{1010}\psi_{1110}^2\psi_{1111} - \psi_{1000}^4\psi_{1111}^2 \\
&+ 2\psi_{0000}\psi_{1000}^2\psi_{1010}\psi_{1111}^2 - \psi_{0000}^2\psi_{1010}^2\psi_{1111}^2.
\end{aligned}$$

Hence, up to closure, the the set $\mathrm{PB}(2,2,4)$ of tensors that can be represented in the form (2.1) where A_0 and A_1 are arbitrary 2×2 matrices, is a sextic hypersurface in the space of $2 \times 2 \times 2 \times 2$ tensors invariant under cyclic permutations of the indices. The 30-term hypersurface equation was found using a parametrization of the matrices that is similar to the birational parametrization of binary hidden Markov models given in Chapter 1.

An example of a pure state on four qubits on which the polynomial f of Theorem 2.1.3 is nonvanishing, and so cannot be arbitrarily well approximated by such a matrix product state, is given by letting $\psi_{1010} = \psi_{1110} = -1/4$ and $\psi_{0000} = \psi_{1000} = \psi_{1100} = \psi_{1111} = 1/4$. In this example, $f(\Psi) = 2^{-5}$, which is the maximal value of $f(\Psi)$ attained on corners of the 6-D hypercube.

The other cases with $N \leq 15$ when we expect $\overline{\mathrm{PB}}$ to be a hypersurface are when $(D, d, N) = (2, 4, 6)$, $(3, 3, 7)$, $(5, 15, 12)$, $(3, 71, 13)$, and $(2, 296, 14)$. In general, we will need many more polynomials to define the space of matrix product states as their zero locus. As an example, consider $\overline{\mathrm{PB}}(2, 2, 5)$, which we expect to be a five-dimensional variety in the necklace space $\mathbb{C}^8 = \mathbb{C}^{N_2(5)}$.

Theorem 2.1.4. *Any homogeneous minimal generating set for the ideal of $\overline{\mathrm{PB}}(2,2,5)$ must contain exactly 3 quartic and 27 sextic polynomials, possibly some higher degree polynomials, but none of degree 1, 2, 3, or 5.*

Proof. Using the bi-grading of Theorem 2.1.5, we decompose the ideal I into vector spaces $I_{r,s}$. For each (r, s) with $\frac{1}{5}(r + s) \leq 6$, we select a large number of parameter values $\widehat{A}$ at random, and use Gaussian elimination to compute a basis for the vector

space $\widehat{I_{r,s}}$ of polynomials vanishing at their images $\Psi(\widehat{A})$, which is certain to contain $I_{r,s}$. We then substitute indeterminate entries for A symbolically into the polynomials to ensure that they lie in $I_{r,s}$. This yields a bihomogeneous basis for I in total degree ≤ 6. □

This is interesting, because the variety only has codimension 3, but requires *at least* 30 equations to cut it out ideal-theoretically. Such a collection of 3 quartics and 27 quadrics was found and verified symbolically. Exact numerical tests (intersection with random hyperplanes) indicate that the top dimensional component of the ideal they generate is reduced and irreducible of dimension 5, and is therefore equal to $\overline{\mathrm{PB}}(2,2,5)$.

Homogeneity and GL_d-invariance

Note that the equation of Theorem 2.1.3 is homogeneous of degree 6, and every monomial has the same total number of 1s appearing in its subscripts. Every MPS variety will be homogeneous in such a grading:

Proposition 2.1.5. *For any D, d, N, the space of translation-invariant MPS limits with periodic boundary conditions is cut out by polynomials in which each monomial has the same total number of 0s, 1s, $\ldots (d-1)$s appearing in its subscripts.*

Proof. In fact we claim that the ideal of $\overline{\mathrm{PB}}(D,d,N)$ is $\mathbb{Z}^d$-homogeneous with respect to d different $\mathbb{Z}$-gradings $\deg_i$ for $0 \leq i \leq d-1$, where $\deg_i(\Psi_J)$ is the number of occurrences of i in J. Since $\deg(\psi_J) := \frac{1}{n}\sum_{i=0}^{N-1} \deg_i(\psi_J) = 1$, the ideal of $\overline{\mathrm{PB}}(D,d,N)$ is also homogeneous in the standard grading.

The usual parametrization Ψ, where $A_0, \ldots, A_{d-1}$ have generic entries, is $\mathbb{Z}^d$-homogeneous with respect to the grading above along with letting $\deg_i(A_j) = 1$ when $i = j$ and 0 when $i \neq j$. Being a homogeneous map, its kernel, the defining ideal of $\overline{\mathrm{PB}}(D,d,N)$, is homogeneous in each of these gradings as well. □

In fact, the variety is homogeneous in a stronger sense because of an action of GL_d on the parameter space of Ψ. In the example above, the action is given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} (A_0, A_1) = (aA_0 + bA_1, cA_1 + dA_1)$$

which descends to an action on Ψ by

$$\begin{pmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{pmatrix} \cdot \psi_{ijkl} = \sum_{pqrs} g_{ip} g_{jq} g_{kr} g_{ls} \psi_{pqrs}.$$

The embedding $(\mathbb{C}^*)^d \subset GL_d$ as diagonal matrices gives rise to the $\mathbb{Z}^d$ homogeneity of the proposition above.

Linear invariants and reflection symmetry

There are additional symmetries peculiar to the case $D = d = 2$. For a generic pair of 2×2 matrices A_0, A_1, there is a one-dimensional family of matrices $P \in SL_2$ such that $P^{-1}A_iP$ are symmetric. Thus, a generic point $\Psi \in \mathrm{PB}(2, 2, N)$ can be written as $\Psi(A_0, A_1)$ with $A_i^T = A_i$, and then $\Psi_J = \mathrm{tr}(\prod_{j\in J} A_j) = \mathrm{tr}((\prod_{j\in J} A_j)^T) = \mathrm{tr}(\prod_{j\in \mathrm{reverse}(J)} A_j) = \Psi_{\mathrm{reverse}(J)}$. This implies

Proposition 2.1.6. *If an N-qubit state Ψ is a limit of binary periodic translation invariant matrix product states, then it has reflection symmetry: $\psi_J = \psi_{\mathrm{reverse}(J)}$ for all J.*

For $N \geq 6$, N-bit strings can be equivalent under reflection but not cyclic permutation, so then $\mathrm{PB}(2, 2, N)$ admits additional linear invariants, i.e. linear polynomials vanishing on the model. For $N = 6, 7$ these are:

$$\begin{aligned} \mathrm{PB}(2,2,6): &\quad \psi_{110100} - \psi_{110010} \\ \mathrm{PB}(2,2,7): &\quad \psi_{1110100} - \psi_{1110010} \text{ and } \psi_{1101000} - \psi_{1100010} \end{aligned}$$

For small N we can find *all* the linear invariants of $\mathrm{PB}(2, 2, N)$ using the bigrading of Theorem 2.1.5 as in the proof of Theorem 2.1.4. Modulo the cyclic and reflection invariants, there are no further linear invariants for $N \leq 7$, but $\mathrm{PB}(2, 2, 8)$ has a single "non-trivial" linear invariant,

$$\psi_{11010010} + \psi_{11001100} - \psi_{11001010} + \psi_{11101000} - \psi_{11011000} - \psi_{11100100}.$$

For $N = 9$, 10, 11, and 12, $\mathrm{PB}(2, 2, N)$ admits 6, 17, 44, and 106 such non-trivial invariants, in each case unique up to change of basis on the vector space they generate.

2.2 MPS with open boundary conditions

We now consider matrix product states with open boundary conditions, which are even more similar to hidden Markov models than the periodic version. Here the state is determined by two boundary state vectors $b_0, b_1 \in \mathbb{C}^D$, along with the $D \times D$

parameter matrices $A_0, \ldots, A_{d-1}$ of the MPS, by

$$\Psi = \sum_{i_1,\ldots,i_N} b_0^T A_{i_1} \cdots A_{i_N} b_1 |i_1 i_2 \cdots i_N\rangle \tag{2.2}$$

$$= \sum_{i_1,\ldots,i_N} \operatorname{tr}(B A_{i_1} \cdots A_{i_N}) |i_1 i_2 \cdots i_N\rangle \tag{2.3}$$

where $B = b_1 b_0^T$ is a rank 1 matrix. We denote the set of states obtainable in this way by $\mathrm{OB}(D,d,N)$, and its closure (Zariski or classical) by $\overline{\mathrm{OB}}(D,d,N)$. We do not have the cyclic symmetries of the PB model here, so we consider $\overline{\mathrm{OB}}(D,d,N)$ as a subvariety of $\mathbb{C}^{d^N}$. If the A_i and b_0^T have non-negative entries with row sums equal to 1, and b_1 is a vector of 1's, then (2.2) is exactly the Baum formula for HMM, so in fact the model $\mathrm{HMM}(D,d,N)$ studied in Chapter 1 is contained in $\mathrm{PB}(D,d,N)$.

The expression (2.3) for Ψ is invariant under the action of SL_D on the A_i and B by simultaneous conjugation. Thus, we may assume B is in Jordan normal form, i.e. a matrix of all zeroes except possibly in the top left corner. As well, the map $(B, A_1, \ldots, A_d) \mapsto (t^{-N}B, tA_1, \ldots, tA_d)$ preserves Ψ, so discarding the case $B = 0$ (which will not change $\overline{\mathrm{OB}}$) we can assume that the top left entry of B is 1. Thus Ψ is determined by dD^2 parameters, the entries of the A_i. In particular, $\overline{\mathrm{OB}}(2,2,3)$ is parametrized by (a dominant map from) 8 parameters, and lives in an 8-dimensional space. This parametrization still turns out still to be degenerate:

Theorem 2.2.1. *A three-qubit state Ψ is a limit of $N = 3$ binary translation invariant MPS with open boundary conditions if and only if the following 22-term quartic polynomial vanishes:*

$$\begin{aligned}
&\psi_{011}^2\psi_{100}^2 - \psi_{001}\psi_{011}\psi_{100}\psi_{101} - \psi_{010}\psi_{011}\psi_{100}\psi_{101} + \psi_{000}\psi_{011}\psi_{101}^2 \\
&+ \psi_{001}\psi_{010}\psi_{011}\psi_{110} - \psi_{000}\psi_{011}^2\psi_{110} - \psi_{010}\psi_{011}\psi_{100}\psi_{110} \\
&+ \psi_{001}\psi_{010}\psi_{101}\psi_{110} + \psi_{001}\psi_{100}\psi_{101}\psi_{110} - \psi_{000}\psi_{101}^2\psi_{110} - \psi_{001}^2\psi_{110}^2 \\
&+ \psi_{000}\psi_{011}\psi_{110}^2 - \psi_{001}\psi_{010}^2\psi_{111} + \psi_{000}\psi_{010}\psi_{011}\psi_{111} + \psi_{001}^2\psi_{100}\psi_{111} \\
&+ \psi_{010}^2\psi_{100}\psi_{111} - \psi_{000}\psi_{011}\psi_{100}\psi_{111} - \psi_{001}\psi_{100}^2\psi_{111} - \psi_{000}\psi_{001}\psi_{101}\psi_{111} \\
&+ \psi_{000}\psi_{100}\psi_{101}\psi_{111} + \psi_{000}\psi_{001}\psi_{110}\psi_{111} - \psi_{000}\psi_{010}\psi_{110}\psi_{111}
\end{aligned}$$

That is, the variety $\overline{\mathrm{OB}}(2,2,3)$ is a quartic hypersurface in $\mathbb{C}^8$ cut out by the polynomial above. This polynomial previously appeared in the context of the HMM [27].

Proof. The map Ψ and its image are homogeneous in the same grading as described in Theorem 2.1.5, which we can use as in the proof of Theorem 2.1.4 to search for low degree polynomials vanishing on the variety. When $(D,d,N) = (2,2,3)$ the quartic

from the theorem appears in this search. The quartic is prime, and therefore defines a 7-dimensional irreducible hypersurface in $\mathbb{C}^8$. On the other hand, the Jacobian of the map Ψ at a random point, e.g. the point where A_0, A_1 have entries $1, 2, 3, 4, 5, 6, 7, 8$ in that order, has rank 7. Therefore $\overline{\mathrm{OB}}(2, 2, 3)$ is of dimension at least 7, and contained in the quartic hypersurface above, so they must be equal. □

From Theorem 2.2.1, we can derive conditions on $\mathrm{OB}(2, 2, N)$ for $N \geq 4$ as well. There is a *marginalization map* from $\mathrm{OB}(2, 2, N)$ to $\mathrm{OB}(2, 2, 3)$ given by $\Psi_I \mapsto \sum_{|J|=N-3} \Psi_{IJ}$ for each I of length 3, which commutes with the assignment $b_1 \mapsto \sum_{j_3 \ldots j_N} A_{j_3} \cdots A_{j_N} b_1$. In fact there are $N-2$ such marginalization maps, each given by choosing 3 consecutive indices I to marginalize to (summing over the remaining indices J). By composing these maps with the quartic polynomial above, we can obtain $N-2$ quartic polynomials vanishing on $\mathrm{OB}(2, 2, N)$.

By analogy to the case of hidden Markov models discussed in the next section, we make the following

Conjecture 2.2.2. *For $N \geq 4$, a generic N-qubit state can be recovered from its marginalization to any three consecutive states. That is, each marginalization map $\overline{\mathrm{OB}}(2, 2, N) \to \overline{\mathrm{OB}}(2, 2, 3)$ is a birational equivalence of varieties.*

The analogous statement with the variety $\overline{\mathrm{HMM}}$ in place of $\overline{\mathrm{OB}}$ is shown to be true in Chapter 1.

Conjecture 2.2.3. *A generic N-qubit (D=d=2) translation invariant matrix product state Ψ with open boundary conditions is determined up to phase by a reduced density operator which traces out all but a chain of three adjacent states, but no fewer.*

When the three adjacent states are qubits 1, 2, and 3 (the first three legs of the diagram), this amounts to saying that the group S^1 of unit-modulus complex numbers acts transitively on generic fibres of the real-algebraic map

$$\Psi \mapsto \left(\sum_{i_4, \ldots, i_N} \Psi_{j_1 j_2 j_3 i_4 \ldots i_N} \Psi^{\dagger}_{k_1 k_2 k_3 i_4 \ldots i_N} \right)_{j_1 j_2 j_3 k_1 k_2 k_3}$$

when restricted to $\mathrm{OB}(2, 2, N)$. Here the right hand side denotes an order 6 tensor with indices j_1, j_2, j_3, k_1, k_2, k_3, and $\Psi^{\dagger}$ denotes complex conjugation.

Conjecture 2.2.4. *A generic N-qubit (D=d=2) periodic translation invariant matrix product state Ψ is determined up to phase by a reduced density operator which traces out all but a chain of four adjacent states, but no fewer.*

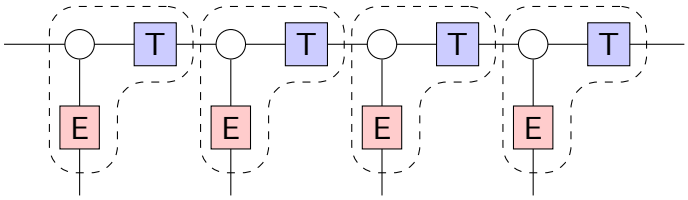


Figure 2.3: Parameterization of an MPS model as a complex HMM using complex E and T matrices with all row sums equal to $z \in \mathbb{C}$ and copy dot (comultiplication) tensor (circle). Contraction of a region of the tensor network enclosed by a dashed line yields an A tensor.

Similarly, this amounts to saying that S^1 acts transitively on generic fibres of the map

$$\Psi \mapsto \left(\sum_{i_5,\ldots,i_N} \Psi_{j_1 j_2 j_3 j_4 i_5 \ldots i_N} \Psi^{\dagger}_{k_1 k_2 k_3 k_4 i_5 \ldots i_N} \right)_{j_1 j_2 j_3 j_4 k_1 k_2 k_3 k_4}$$

when restricted to $\mathrm{PB}(2,2,N)$.

2.3 Matrix product states as complex valued hidden Markov models

We now explain how the polynomial in Theorem 2.1.3 was obtained, and connect the classical *hidden Markov model* and matrix product states through a reparametrizing rational map. The parametrization of the state Ψ is analogous to that of the moment tensor of a binary hidden Markov model used in Chapter 1.

Let T be a 2×2 *transition* matrix and E a 2×2 *emission* matrix. For a (classical) hidden Markov model, T and E are nonnegative stochastic matrices (their rows sum to one), representing a four-dimensional parameter space. For PB, T and E will be complex with row sums all equal to some constant $z \in \mathbb{C}$, so they form parameter space isomorphic to $\mathbb{C}^5$. We parametrize the A_i in terms of (T, E) by

$$A_0 = T, \qquad A_1 = \begin{pmatrix} e_{01} & 0 \\ 0 & e_{11} \end{pmatrix} \begin{pmatrix} t_{00} & t_{01} \\ t_{10} & t_{11} \end{pmatrix}.$$

This is shown in Figure 2.3; grouping and contracting the E, T, and copy dot tensors into an A tensor yields a dense parameterization of an MPS as depicted in Figure

2.1. We then parameterize E and T with the five parameters u, v_0, b, c_0, z by setting

$$E = \begin{pmatrix} z-u+v_0 & u-v_0 \\ z-u-v_0 & u+v_0 \end{pmatrix} \text{ and}$$
$$T = \begin{pmatrix} z+u-c_0 & z-b+c_0 \\ z-b-c_0 & z+b+c_0 \end{pmatrix}.$$

Composing these formulae with the map $(A_0, A_1) \mapsto \Psi$ yields a restricted parametrization $\rho_N : \mathbb{C}^5 \to \mathbb{C}^{2^N}$, whose image lies inside $\mathrm{PB}(2,2,N)$.

Proposition 2.3.1. *The variety* $\overline{\mathrm{PB}}(2,2,N)$ *is at most* 5*-dimensional, and the image of our restricted parametrization* ρ_N *is dense in it.*

Proof. Suppose $\Psi = \Psi(A_0, A_1)$ for A_0, A_1 generic. First, we will transform the A_i by simultaneous conjugation with an element P of SL_2 to a new pair of matrices A_0', A_1' such that A_0' has equal row sums and $A_1' = DA_0'$ for a diagonal matrix D. Generically, A_0 is invertible, and we can diagonalize the matrix $A_1A_0^{-1}$, so we write $U^{-1}A_1A_0^{-1}U = D_0$, and then $U^{-1}A_1U = DU^{-1}A_0U$. Next we find another diagonal matrix $D_1 \in SL_2$ such that $D_1^{-1}U^{-1}A_0UD_1$ has equal row sums. Then let $P = UD_1$ and $A_i' = D_1^{-1}U^{-1}A_iUD_1$, and we are done with our transformation. Now $\Psi = \Psi(A_0', A_1')$ since simultaneous conjugation does not change trace products. But now letting z be the common row sum of A_0', we can solve linearly for u, v_0, b, and c_0 to obtain $\Psi = \rho(u, v_0, b, c_0, z)$. □

In fact we know from exact computations in Macaulay2 [22] that for $4 \leq N \leq 100$, $\dim \mathrm{PB}(2,2,N) = 5$. This is proven by checking that the Jacobian of ρ attains rank 5 at some point with randomly chosen integer coordinates, giving a lower bound of 5 on the dimension of its image.

When parametrized using ρ, there are sufficiently few parameters and the entries of Ψ are sufficiently short expressions that Macaulay2 is also able to compute the exact kernel of the parametrization, i.e. defining equations for the model. It is by this method that we obtain the hypersurface equation of Theorem 2.1.3 as the only ideal generator for $\mathrm{PB}(2,2,4)$.

Identifying parameters of MPS

Determining the parameters of an MPS is related to *quantum state tomography*, and represents a quantum analog to the identifiability problem in statistics. The extent to which the parameters can be identified can be addressed algebraically.

Given $D \times D$ matrices $A_1 \dots A_d$ with indeterminate entries, we write $\mathcal{C}_{D,d}$ for the algebra of polynomial expressions in their entries that are invariant under simultaneous conjugation of the matrices by GL_2.

Sibirskii [34], Leron [21], and Procesi [31] showed that the algebra $\mathcal{C}_{D,d}$ is generated by the traces of products $tr(A_{i_0} \cdots A_{i_n})$ as $n \geq 0$ varies. For this reason, $\mathcal{C}_{D,d}$ is called a *trace algebra*. Its spectrum, $\operatorname{Spec} \mathcal{C}_{D,d}$, is a *trace variety*. Since the coordinate ring of $\overline{\mathrm{PB}}(D,d,N)$ is a subring of $\mathcal{C}_{D,d}$, we have a map $\operatorname{Spec} \mathcal{C}_{D,d} \to \mathbb{C}^{d^N}$ parametrizing a dense open subset of $\overline{\mathrm{PB}}(D,d,N)$.

In the case $D = 2$, Sibirskii showed further that the trace algebra $\mathcal{C}_{2,d}$ is minimally generated by the elements $tr(A_i)$ and $tr(A_i^2)$ for $1 \leq i \leq d$, $tr(A_i A_j)$ for $1 \leq i < j \leq d$, and $tr(A_i A_j A_k)$ for $1 \leq i < j < k \leq d$.

For $d = 1 \dots 6$, the number of such generators is $2, 5, 10, 18, 30, 47$. In particular, when $d = 2$, the number of generators equals the transcendence degree of the ring, $5 = 8 - 3$. This means $\operatorname{Spec} C_{2,2}$ is isomorphic to $\mathbb{C}^5$, yielding for each N a dominant parametrization $\phi_N : \mathbb{C}^5 \to \overline{\mathrm{PB}}(2,2,N)$. Gröbner bases for randomly chosen fibers indicate that for $N = 4 \dots 10$, the map ϕ_N is generically k-to-one, where $k = 8, 5, 6, 7, 8, 9, 10$, respectively. Continuing this sequence suggests the following.

Conjecture 2.3.2. *Using the trace parameterization ϕ_N, for $N \geq 5$, almost every periodic boundary MPS has exactly N choices of parameters that yield it.*

In other words, for $N \geq 5$, the parametrization $\phi_N : \mathbb{C}^5 \simeq \operatorname{Spec} C_{2,2} \to \overline{\mathrm{PB}}(2,2,N)$ is generically N-to-1. Generically, the points of $\operatorname{Spec} C_{2,2}$ are in bijection with the SL_2-orbits of the tensors A. The conjecture implies that, up to the action of SL_2, the parameters of a binary, $D = d = 2$ translation invariant matrix product state with periodic boundary are algebraically identifiable from its entries.

2.4 Conclusion

A conjectured dictionary between tensor network state models and classical probabilistic graphical models was presented in [25]. In this dictionary, matrix product states correspond to hidden Markov models, the density matrix renormalization group (DMRG) algorithm to the forward-backward algorithm, tree tensor networks to general Markov models, projected entangled pair states (PEPS) to Markov or conditional random fields, and the multi-scale entanglement renormalization ansatz (MERA) loosely to deep belief networks. In this chapter, we formalize the first of these correspondences and use it to algebraically characterize quantum states representable by MPS and study their identifiability.

Chapter 3

Tensors and models with more hidden states

This chapter is based on joint work with Shaowei Lin, Luca Weihs, and Piotr Zwiernik.

In Chapter 1, the hidden variables of the hidden Markov models were all assumed to be binary, i.e., they could each take on *only two possible states*. In Chapter 2, the virtual bonds of the matrix product state models — which are algebraically very similar to hidden variables — were also assumed to be binary. In each case, this assumption allowed for a particularly nice reparametrization of the model which made symbolic computations more tractable.

In this chapter, we study directed acyclic graphical (DAG) models where each random variable, or *node*, is allowed to take on *an arbitrary finite number* of states. Two things are accomplished. First, we introduce a notational method called *automatic tensor contraction* which provides a convenient and index-free way to represent the large but structured polynomials which arise in the study of DAG models. Second, using this notation, we state and prove a new and symbolically efficient parametrization (Theorem 3.8.3) of discrete DAG models.

Initially, this work began as an effort to generalize work of Smith and Zwiernik [35], but turned out to be more broadly applicable. In the case of certain tree-shaped models on binary variables, Smith and Zwiernik [35] defined new coordinates called *tree cumulants* which allow for an extremely symbolically efficient reparametrization. The new parameters were at first somewhat mysterious and it was unclear whether they could be generalized for models with non-binary variables, i.e., variables taking on more than two states. However, through discussions with Zwiernik, it became clear that they could be viewed as linear regression coefficients in a sense that would allow for their generalization. This chapter, among other uses, serves as a first step

in this generalization, which in fact applies not only to trees but all discrete DAG models.

3.1 Introduction

In a directed acyclic graph (DAG) model, the nodes of the graph represent random variables, and each variable is *affected* by its parent according to a some conditional probability distribution. Thus, DAG models on discrete variables are typically parametrized using conditional probabilities. For example, the DAG model $B \leftarrow A \rightarrow C$ where B and C are observed is parametrized as

$$\Pr(BC) = \sum_A \Pr(A)\ \Pr(B \mid A)\ \Pr(C \mid A).$$

These parameterizations are somewhat redundant in that not all the parameters are free: the rows of a conditional probability matrix must sum to one. In this chapter, we introduce a more compact parametrization of DAGs using moment tensors μ and linear regression coefficients β which result in shorter expression lengths and simpler characterization of statistical independence. Using this, we first derive the following symbolically efficient expression of the moments of a the sinks of a DAG model, to be stated and explained more precisely in section 3.8:

Theorem 3.1.1 (3.8.3)**.** *Given a discrete DAG model G and a set $S \subseteq \text{nodes}(G)$, the joint moment μ_S of the nodes S is given by the following formula:*

$$\mu_S = \sum_{\substack{H \subseteq G \\ \text{sinks}(H) \subseteq S \subseteq \text{nodes}(H)}} \prod_{v \in \text{nodes}(H)} \beta^{\text{pa}(v;H)}_{v^{m(v;H,S)}}$$

The definition of moments μ_S is given in Section 3.6, and the tensors $\beta^{\text{pa}(v;H)}_{v^{m(v;H,S)}}$ are defined in Section 3.8. The product here is an operation called *automatic contraction* to be defined in section 3.3. Without saying much more, we can note here that as S varies among all the subsets of of nodes(G), the total number of entries the tensors $\beta^{\text{pa}(v;H)}_v$ is exactly equal to the dimension of the standard parameter space of the model. They also determine the entries of the larger tensors $\beta^{\text{pa}(v;H)}_{v^{m(v;H,S)}}$, and so they indeed parametrize the model as one would hope.

With this result in place, there are many possible subsequent directions. Models where not all nodes are observed abound in statistics, for example hidden Markov models (HMM), phylogenetic trees, restricted Boltzmann machines (RBM) and other

neural networks. The above formula gives a more efficient way to encode the probability distribution on the observed nodes of these models. The new parametrization can be used to study the defining equations of the model, or the inequalities defining the model inside its Zariksi closure, or to compute the learning coefficients to be used in the sBIC (singular Bayesian information criterion) for model selection. We can generalize previous work of Smith and Zwiernik [35] on *tree cumulants* of binary tree models. Namely, nearly identical formulae apply for tree models models where each node can take an arbitrary finite number of states.

3.2 Tensors

In this section, we first give a brief introduction to *tensors* and lay out the notations used in this chapter. In the second part, we define *affine distributions* which generalize probability distributions in the finite discrete case. Loosely speaking, they are probabilities which sum to one but are allowed to take negative or complex values.

Vectors

There are many standard ways of writing and talking about vectors. Here we illustrate very briefly *which* standard notations and terminology we are using.

Given a vector space $\mathcal{V} \simeq \mathbb{C}^r$, let $e_1, \ldots, e_r$ be a basis for $\mathcal{V}$ and denote the coordinates of a vector $v \in \mathcal{V}$ with respect to this basis by writing $v = v_1 e_1 + \cdots + v_r e_r = \sum_i e^i(v) e_i$, where e^i is the function $\mathcal{V} \to \mathbb{C}$ that returns the i-th coordinate of a vector. Now, let the *dual space* $\mathcal{V}^*$ be the vector space of linear maps $\mathcal{V} \to \mathbb{C}$. Every $\ell \in \mathcal{V}^*$ can be written as a linear combination $\ell = \ell^1 e^1 + \cdots + \ell^r e^r$ so the coordinate functions $e^1, \ldots, e^r$ form a basis that is dual to $e_1, \ldots, e_r$, i.e. $e^i(e_j) = \delta^i_j$ where $\delta = (\delta^i_j)$ is the Kronecker delta

$$\delta^i_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Note that we denote the coordinates v_i of $v \in \mathcal{V}$ using subscripts while those of a vector $\ell = (\ell^i)$ in the dual space $\mathcal{V}^*$ are denoted using superscripts.

Tensors

Here we illustrate our notation and terminology for tensors, which aims to avoid the cumbersome use of indices when possible. There are many ways to define the tensor

product $\mathcal{V} \otimes \mathcal{W}$ of vector spaces $\mathcal{V}$ and $\mathcal{W}$. If $\mathcal{V} \simeq \mathbb{C}^r$ and $\mathcal{W} \simeq \mathbb{C}^s$, then we can represent a tensor $\alpha \in \mathcal{V} \otimes \mathcal{W}$ as an array in $\mathbb{C}^{r \times s}$. The tensor product $v \otimes w$ of vectors $v = (v_i) \in \mathcal{V}$ and $w = (w_j) \in \mathcal{W}$ is the array

$$\alpha = (\alpha_{ij}) = (v_i w_j) \ \in \ \mathcal{V} \otimes \mathcal{W}$$

of products of the coordinates of v and w. Thus $\otimes$ is a bilinear operation, i.e.

$$\begin{aligned}(v + v') \otimes w &= v \otimes w + v' \otimes w \\ v \otimes (w + w') &= v \otimes w + v \otimes w' \\ (cv) \otimes w &= v \otimes (cw) = c(v \otimes w)\end{aligned}$$

for all $v, v' \in \mathcal{V}$, $w, w' \in \mathcal{W}$ and $c \in \mathbb{C}$. Let $e_1, \ldots, e_r$ and $e'_1, \ldots, e'_s$ be basis vectors for $\mathcal{V}$ and $\mathcal{W}$. Then $\mathcal{V} \otimes \mathcal{W}$ is space of all linear combinations $\sum_{i,j} \alpha_{ij}\, e_i \otimes e'_j$. When working in a tensor product of vector spaces and their duals, we denote the coordinates by

$$v = (v^{j_1 \cdots j_n}_{i_1 \cdots i_m}) \ \in \ \mathcal{V}_{a_1} \otimes \cdots \otimes \mathcal{V}_{a_m} \otimes \mathcal{V}^*_{b_1} \otimes \cdots \otimes \mathcal{V}^*_{b_n}$$

where the superscripts and subscripts indicate the vector spaces involved.

The *order* of a tensor is the dimension of its array of coordinates. For instance, vectors are tensors of order one while matrices have order two.

3.3 Automatic tensor contractions

In this section, we take a lesson from computer science to simplify our tensor notation. Tensor equations often involve large numbers of subscripts and superscripts which can be difficult to read and manipulate, but the form of these formulae are often determined by the *type* of particular tensors involved. By carefully defining types of tensors and how they should interact, many complicated expressions can be written and manipulated much more simply.

For example, suppose that $\beta \in \mathcal{U} \otimes \mathcal{U} \otimes \mathcal{V}^*$ and $\gamma \in \mathcal{V}$. It is "natural" to evaluate the expression $\beta \cdot \gamma$ by temporarily considering β as a map $\mathcal{V} \to \mathcal{U} \otimes \mathcal{U}$ and computing $\beta(\gamma)$. This can be written by equating and summing some indices, commonly known as *contracting* indices. Explicitly, say $\{u_0, \ldots, u_\ell\}$ is a basis for $\mathcal{U}$ and $\{v_0, \ldots, v_m\}$ is a basis for $\mathcal{V}$ with dual basis $\{v^0, \ldots, v^m\}$, so there are unique constants β^j_i and γ_k such that

$$\beta = \sum_{ijk} \beta^k_{ij} u_i \otimes u_j \otimes v^k \text{ and } \gamma = \sum_{\ell} \gamma_\ell v_\ell.$$

Evaluating "$\beta(\gamma)$" amounts to contracting the index pair $\{k, \ell\}$, i.e., replacing the symbol ℓ by k and summing over k as follows:

$$\sum_{ij} \left(\sum_k \alpha_{ij}^k \beta_k \right) u_i \otimes u_j$$

To define an operation "$\cdot$" such that $\beta \cdot \gamma$ evaluates this way automatically, we can declare that a pair of indices in an expression involving "$\cdot$" should be contracted if and only if they correspond to dual vector spaces. Some conditions will be needed for this operation to be well defined; for example, if $\delta = (\delta_m) \in \mathcal{U}^*$, then in the expression $\beta \cdot \delta$ it is ambiguous whether to contract m with i or j, because they both correspond to $\mathcal{U}$.

In general, given a symbolic array expression for a tensor $\alpha = (\alpha_{i_1,\ldots,i_n}) \in \mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_n$, we say that the index symbol i_k *corresponds* to the vector space $\mathcal{V}_k$, and write $\mathrm{Corr}(i_k) = \mathcal{V}_k$. In the example above, $\mathrm{Corr}(i) = \mathrm{Corr}(j) = \mathcal{U}$, $\mathrm{Corr}(k) = \mathcal{V}^*$, and $\mathrm{Corr}(\ell) = \mathcal{V}$. Thus, determining whether k should be contracted with ℓ in the expression $\beta \cdot \gamma$ amounts to checking the *equality* of vector spaces $\mathrm{Corr}(k) = \mathrm{Corr}(\ell)^*$. So if we are careful about what me mean by vector space equality (as opposed to mere isomorphism), then we can decide how indices should be contracted by carefully deciding in advance what vector spaces they will correspond to. This is analogous to defining types in a programming language in order to simplify subsequent code.

Notation 3.3.1 (Vector space equality conventions)**.** We distinguish between *literal equality* and *isomorphism* of vector spaces. For example, if v_1, v_2, w_1, w_2 are distinct symbols, $\mathcal{V} = \langle v_1, v_2 \rangle_{\mathbb{C}}$ and $\mathcal{W} = \langle w_1, w_2 \rangle_{\mathbb{C}}$, then $\mathcal{V} \neq \mathcal{W}$ although $\mathcal{V} \simeq \mathcal{W}$. We consider

- $\mathcal{V}^{**} = \mathcal{V}$ for every finite vector space $\mathcal{V}$. We have implemented this in computations by having $\mathcal{V}^*$ "remember" that it is a dual space, and declare that taking its dual "removes the $*$".
- $(\mathcal{U} \otimes \mathcal{V}) \otimes \mathcal{W} = \mathcal{U} \otimes (\mathcal{V} \otimes \mathcal{W})$. We have implemented this computationally by having a tensor product of vector spaces "remember" which vector spaces it is a tensor product of, so that both of the above expressions can be told to "drop parentheses" and equal to $\mathcal{U} \otimes \mathcal{V} \otimes \mathcal{W}$.
- $\mathcal{V} \otimes \mathcal{U} \neq \mathcal{U} \otimes \mathcal{V}$ in general unless $\mathcal{V} = \mathcal{U}$.

Now we are ready to state some symmetric conditions under which will make our automatic contractions well-defined:

Definition 3.3.2 (Sufficiently symmetric tensors)**.** We say that a tensor α is $\mathcal{V}$-symmetric if it is invariant under permutation of indices corresponding to $\mathcal{V}$. For example, if $\alpha = (\alpha_{ijk}) \in \mathcal{V} \otimes \mathcal{W} \otimes \mathcal{V}$, then α is $\mathcal{V}$-symmetric if and only if $\alpha_{ijk} = \alpha_{kji}$ for all i, j, k. If α has no $\mathcal{V}$ indices or exactly one $\mathcal{V}$ index, and we consider it trivially $\mathcal{V}$-symmetric. We say that α is *sufficiently symmetric* if for every vector space $\mathcal{V}$ such that α has both $\mathcal{V}$ and $\mathcal{V}^*$ indices, either

- α is $\mathcal{V}$-symmetric and has at least as many $\mathcal{V}$-indices as $\mathcal{V}^*$-indices, or
- α is $\mathcal{V}^*$-symmetric and has at least as many $\mathcal{V}^*$ indices as $\mathcal{V}$ indices.

Definition 3.3.3. If α is a sufficiently symmetric tensor, we define the *automatic contraction* $\mathrm{Auto}(\alpha)$ to be the tensor obtained from α by contracting every pair of indices (i, j) corresponding to a pair of dual vector spaces $(\mathcal{V}_i, \mathcal{V}_j)$. The sufficient symmetry condition ensures that this definition is well-defined.

Notation 3.3.4. If $\alpha = \alpha_1 \otimes \cdots \otimes \alpha_n$ is sufficiently symmetric then we write

$$\alpha_1 \cdot \alpha_2 \cdot \cdots \cdot \alpha_{n-1} \cdot \alpha_n$$

in place of $\mathrm{Auto}(\alpha)$. For example, if $\alpha = (\alpha_{ijk}) \in \mathcal{U} \otimes \mathcal{V} \otimes \mathcal{W}$, $\beta = (\beta^\ell_m) \in \mathcal{U}^* \otimes \mathcal{X}$, and $\gamma = (\gamma^r_s) \in \mathcal{V}^* \otimes \mathcal{X}$, then

$$\alpha \cdot \beta \cdot \gamma := \sum_{ij} (\alpha_{ijk} \beta^i_m \gamma^j_s) \in \mathcal{W} \otimes \mathcal{X} \otimes \mathcal{X}.$$

We will take much advantage of this notation, especially to simplify formulae in which an unspecified number of tensors are being contracted.

The following lemma illustrates how matrix inversion interacts with automatic contraction, which will be useful later:

Lemma 3.3.5. *If $A \in \mathcal{U} \otimes \mathcal{V}$ is invertible (considered either as a map $\mathcal{U}^* \to \mathcal{V}$ or $\mathcal{V}^* \to \mathcal{U}$) with inverse $A^{-1} \in \mathcal{U}^* \otimes \mathcal{V}^*$, then for any $B \in \mathcal{U}^* \otimes \mathcal{W}$, and $C \in \mathcal{U} \otimes \mathcal{X}$,*

$$B \cdot C = (A \cdot B) \cdot (A^{-1} \cdot C).$$

Example 3.3.6. Here is a good place to illustrate some examples of automatic contraction. Suppose $A \in \mathcal{U} \otimes \mathcal{U}$ is symmetric and invertible, and denote its inverse by $A^{-1} \in \mathcal{U}^* \otimes \mathcal{U}^*$. First, note that the automatic contraction $A \cdot A^{-1}$ is well-defined and equal to the scalar $\dim(U)$, not an identity matrix, because we end up summing

over *both* indices of A, yielding the trace of the identity. Next, if $B \in \mathcal{U}^* \otimes \mathcal{W}$, and $C \in \mathcal{U} \otimes \mathcal{X}$, observe that

$$\begin{aligned} B \cdot C &= (B \cdot A) \cdot (A^{-1} \cdot C) \\ &\neq B \cdot (A \cdot A^{-1}) \cdot C = B \cdot \dim(\mathcal{U}) \cdot C = \dim(\mathcal{U}) B \cdot C \end{aligned}$$

In particular, the automatic contraction expression $B \cdot A \cdot A^{-1} \cdot C$ is not well-defined, and indeed, $\alpha = B \otimes A \otimes A^{-1} \otimes C$ does not satisfy the sufficient symmetry condition because there are more $\mathcal{U}$ indices than $\mathcal{U}^*$ indices and the $\mathcal{U}$ indices in A cannot be permuted with the $\mathcal{U}$ index in C. The reader is invited to please think critically about our use of automatic contraction throughout this chapter, and observe that our use of symmetry and occasional parentheses are adequate.

3.4 Affine and probability distributions

A finite measure space can be specified by a finite set $\mathcal{I}$, called the set of *outcomes* or *indices*, along with a map $p : \mathcal{I} \to \mathbb{C}$ called its *distribution* or *mass* function. Since p is simply an element of the vector space $\mathbb{C}^{\mathcal{I}}$, for each $\imath \in \mathcal{I}$ we write $p_\imath$ for $p(\imath)$, and call the $p_\imath$ *entries* of p.

Definition 3.4.1 (Affine and probability distributions)**.** If $\sum_{\imath \in \mathcal{I}} p_\imath = 1$, we say p is an *affine distribution*, and if also the entries $p_\imath$ are all nonnegative real numbers, we say p is a *probability distribution*. This corresponds to the usual definition of a probability distribution in the finite discrete case.

In this chapter, we study affine distributions $p \in \mathbb{C}^{\mathcal{I}}$ where $\mathcal{I}$ is the finite discrete set

$$\mathcal{I} = \mathcal{I}_1 \times \cdots \times \mathcal{I}_n = \{0, \ldots, k_1\} \times \cdots \times \{0, \ldots, k_n\} \tag{3.1}$$

for some positive integers $n, k_1, \ldots, k_n$. We write

$$\mathcal{U} := \mathbb{C}^{\mathcal{I}} \simeq \mathbb{C}^{(k_1+1)\times\cdots\times(k_n+1)} \quad \text{and} \quad \mathcal{U}_r := \mathbb{C}^{\mathcal{I}_r},$$

and make the natural identification $\mathcal{U} = \mathcal{U}_1 \otimes \cdots \otimes \mathcal{U}_n$. For index classification reasons, we consider the sets $\mathcal{I}_r$ to be disjoint, so that the vector spaces $\mathcal{U}_r$ are considered distinct, even if they are sometimes isomorphic.

Notation 3.4.2 (Multiset subscripts)**.** Given an index $\imath = (\imath_1, \ldots, \imath_n) \in \mathcal{I}$ and a multiset $A \subseteq [n]$ we write

$$\imath_A := (\imath_j)_{j \in A} \ \in \ \mathcal{I}_A := \textstyle\prod_{j \in A} \mathcal{I}_j$$

$$\mathcal{U}_A := \bigotimes_{j\in A} \mathcal{U}_r = \mathbb{C}^{\mathcal{I}_A}$$

For example, if $n = 4$, $\iota = (5, 6, 7, 8)$, and $A = (1, 2, 2, 2, 4)$, then $\iota_A = (5, 6, 6, 6, 8)$. In computations, we implement a multiset A as a list, so that it is easy to take sums and products indexed by A as written above. In writing, we use the language of multisets rather than lists because it is easier to read and write $\prod_{r\in A} f(r)$ instead of $\prod_{i=1}^{|A|} f(A_i)$.

Next we introduce marginal distributions. Consider a matrix $p = (p_{ij}) \in \mathbb{C}^{(k_1+1)\times(k_2+1)}$ as a table. One would typically write its row sums and column sums in the *margins* of the table, so we call the vector of row sums and the vector of column sums *marginal* distributions. We write $P_1 := [p_{0+}, \ldots, p_{k_1+}]$ for the result of *marginalizing to* the first index, i.e., summing over the second index. Likewise we write $P_2 = [p_{+0}, \ldots, p_{+k_2}]$ for the marginal distribution on the second index. More generally, given any multiset $A \subseteq [n]$ and distribution $p \in \mathcal{U} = \mathcal{U}_{[n]}$, the *marginal distribution* $P_A \in \mathcal{U}_A$ is the tensor with entries

$$(P_A)_a \quad := \quad \sum_{\iota\in\mathcal{I}:\, \iota_A = a} p_\iota \qquad \text{for all } a \in \mathcal{I}_A. \tag{3.2}$$

In particular $P_{[n]} = p$, and if p is affine, then $P_\emptyset = 1$ and every marginal distribution is also affine.

The reader is hereby reassured that we will use subscripts consistently with this convention throughout this chapter: whenever the symbol P has two subscripts, e.g. $(P_A)_a$, the first subscript always stands for a subset of (or multiset of elements in) $[n]$, and the second subscript a always stands for an index in $\mathcal{I}_A$. For example, if we write $(P_1)_2$, the 1 is short for the subset $\{1\}$ of $[n]$ and the 2 is taken as an index in $\mathcal{I}_{\{1\}}$, so that $(P_1)_2 = p_{2++\cdots}$.

We also note that the operation of marginalization $P \mapsto P_A$ can be written as a contraction. For any vector space $\mathcal{V}$ define $\mathbf{1} : \mathcal{V} \to \mathbb{C}$ as the map that maps any vector v to the sum of its coefficients. For any $r \in [n]$ we write $\mathbf{1}(r) : \mathcal{U}_r \to \mathbb{C}$ and $\mathbf{1}(A) := \otimes_{r\in A}\mathbf{1}(r)$. Then for any subset $A \subseteq [n]$, the marginalization P_A is

$$P_A = \mathbf{1}([n] \setminus A) \cdot P.$$

3.5 Independence and conditional independence

In statistics and algebraic geometry, independence and conditional independence show up in many places. Families of distributions satisfying these properties are

respectively called independence models and mixture models. Meanwhile, in geometry these point sets are called Segre varieties and secant varieties. The geometry of many classical tensor spaces is studied at length in the book *Tensors: Geometry and Applications* by Landsberg [20].

Independence

Given disjoint subsets $A, B \subset [n]$ we say that A is *independent* of B if, using Notation 3.4.2,

$$(P_{AB})_{ab} = (P_A)_a (P_B)_b \qquad \text{for every } a \in \mathcal{I}_A \text{ and } b \in \mathcal{I}_B$$

and we denote this by $A \perp\!\!\!\perp B$. In other words, P_{AB} can be written as a tensor product

$$P_{AB} = P_A \otimes P_B.$$

It follows that for any subset I of $A \cup B$, we have

$$P_I = P_{I\cap A} \otimes P_{I \cap B}.$$

This notion of independence can be generalized to the joint independence of a partition.

Definition 3.5.1. Given a partition $B_1|\cdots|B_r$ of $B \subseteq [n]$ we say that $B_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp B_r$ if

$$P_B = P_{B_1} \otimes \ldots \otimes P_{B_r}.$$

Consequently, for any other subset $I \subset B$, we have $P_I = P_{B_1 \cap I} \otimes \cdots \otimes P_{B_r \cap I}$. □

Example 3.5.2. The Segre variety $\Sigma_{\mathcal{I}}$ is the subvariety of all rank one tensors x in $\mathbb{P}^{(k_1+1)\cdots(k_n+1)-1}$. It is parametrized by $t_r \in \mathbb{P}^{k_j}$ for $r = 1, \ldots, n$ by $x = t_1 \otimes \cdots \otimes t_n$. In particular,

$$\Sigma_{\mathcal{I}} = \{1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp n\}.$$

Conditional independence

Let $\mathcal{U}$ be the vector space $\mathbb{C}^{\mathcal{I}}$ and $\mathcal{U}_A$ the subspace $\mathbb{C}^{\mathcal{I}_A}$. Let A, B be two disjoint subsets of $[n]$. In what follows we write AB for $A \cup B$ and a for $\{a\}$. Define the *conditional distribution* of B given A to be the tensor $P_{B|A} \in \mathcal{U}_B \otimes \mathcal{U}_A^*$ such that

$$(P_A)_a (P_{B|A})^a_b = (P_{AB})_{ab} \qquad \text{for every } a \in \mathcal{I}_A, b \in \mathcal{I}_B. \tag{3.3}$$

This tensor is uniquely defined if the marginal distribution P_A is *nondegenerate*, i.e. $(P_A)_b \neq 0$ for all $a \in \mathcal{I}_A$. Note that if p is affine (has entry sum 1) then the "columns" $(P_{B|A})^a_\bullet$ are affine for every $a \in \mathcal{I}_A$

Using multiset subscript notation, the tensor $P_{AA} \in \mathcal{U}_A \otimes \mathcal{U}_A$ has entries $(P_{AA})_{aa'} = \delta_{aa'}(P_A)_a$. This allows us to rewrite the definition of conditional independence (3.3) as a very simple automatic contraction:

$$P_{B|A} = P_{AB} \cdot P_{AA}^{-1}. \tag{3.4}$$

Example 3.5.3. Let $\mathcal{I} = \mathcal{I}_1 \times \mathcal{I}_2 = \{0,1\} \times \{0,1\}$ so that $\mathcal{U} \simeq \mathbb{C}^{2\times 2}$ and $\mathcal{U}_1 \simeq \mathcal{U}_2 \simeq \mathbb{C}^2$. Given a distribution $p = P_{12} \in \mathcal{U}$, the conditional distribution $P_{2|1} \in \mathcal{U}_2 \otimes \mathcal{U}_1^*$ is the tensor with

$$\begin{aligned} (P_{2|1})^0_0 &= \tfrac{p_{00}}{p_{0+}}, \quad (P_{2|1})^0_1 = \tfrac{p_{10}}{p_{0+}}, \\ (P_{2|1})^1_0 &= \tfrac{p_{01}}{p_{1+}}, \quad (P_{2|1})^1_1 = \tfrac{p_{11}}{p_{1+}}. \end{aligned}$$

Hence, the conditional distribution is well defined as long as the distribution p lies outside the hyperplane arrangement described by $p_{0+}p_{1+} = 0$. □

Given pairwise disjoint subsets $A, B, C \subset [n]$ we say that B is *conditionally independent* of C given A, or $B \perp\!\!\!\perp C|A$, if for each $a \in \mathcal{I}_A$ such that $(P_A)_a \neq 0$,

$$(P_{BC|A})^a_{bc} = (P_{B|A})^a_b (P_{C|A})^a_c \quad \text{for every } b \in \mathcal{I}_B, c \in \mathcal{I}_C.$$

In terms of tensors, this can be written as

$$(P_{BC|A})^a_\bullet = (P_{B|A})^a_\bullet \otimes (P_{C|A})^a_\bullet.$$

More generally, we have the following definition using tensors.

Definition 3.5.4. Given disjoint subsets $A, B \subset [n]$ and a partition $B_1|\cdots|B_r$ of B, we say that $B_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp B_r|A$ if for each $a \in \mathcal{I}_A$,

$$(P_{B|A})^a_\bullet = (P_{B_1|A})^a_\bullet \otimes \cdots \otimes (P_{B_r|A})^a_\bullet.$$

Consequently, given any subset $I \subset B$, we also have $(P_{I|A})^a_\bullet = (P_{B_1 \cap I|A})^a_\bullet \otimes \cdots \otimes (P_{B_r \cap I|A})^a_\bullet$.

Mixtures and secants

Note that if $B \perp\!\!\!\perp C|A$ then we can express the marginal distribution P_{BC} as

$$(P_{BC})_{bc} = \sum_{a \in \mathcal{I}_A} (P_A)_a (P_{B|A})^a_b (P_{C|A})^a_b.$$

Using automatic contraction, this equation can be written succinctly as

$$P_{BC} = P_{AA} \cdot P_{B|A} \cdot P_{C|A}$$

3.6 Moment tensors

In chapter 1, we took advantage of *moment coordinates* in the symbolic computation of the defining ideals of certain models with binary variables. Using tensors, here we develop a generalization of those moment coordinates for models with variables having more than two states.

Suppose temporarily that $\mathcal{I} = \mathcal{I}_1 \times \mathcal{I}_2$, so that $p = (p_{ij}) \in \mathbb{C}^{\mathcal{I}} = \mathbb{C}^{\mathcal{I}_1 \times \mathcal{I}_2}$. As usual, we write $\mathcal{I}_i = \{0, 1, \ldots, k_i\}$. By adding all rows of p to its first row and all its columns to its first column we obtain new matrix

$$\bar{\mu}_{12} = \left[\begin{array}{c|ccc} 1 & p_{+1} & \cdots & p_{+k_2} \\ \hline p_{1+} & p_{11} & \cdots & p_{1k_2} \\ \vdots & & \cdots & \vdots \\ p_{k_1+} & p_{k_1 1} & \cdots & p_{k_1 k_2} \end{array}\right]. \tag{3.5}$$

which we call a *total moment* matrix. (The statistical interpretation that justifies this terminology is explained at the end of this section.) We denote the top left corner of this matrix by $\mu_\emptyset = 1 \in \mathbb{C}$, the bottom left block by $\mu_1 \in \mathbb{C}^{k_1}$, the top right block by $\mu_2 \in \mathbb{C}^{k_2}$, and the remaining block by $\mu_{12} \in \mathbb{C}^{k_1 \times k_2}$. We refer to these blocks as *(non-central) moments*.

To extend this construction in a notationally compact way to the general case where p is a tensor in $\mathbb{C}^{\mathcal{I}} = \mathbb{C}^{\mathcal{I}_1 \times \cdots \times \mathcal{I}_n}$, we can represent the row and column operations performed above via automatic contractions with elementary matrices that we will call $M(1)$ and $M(2)$, defined as follows. For $r = 1, \ldots, n$, we define new and distinct vector spaces

$$\overline{\mathcal{V}}_r := \langle v_0, \ldots, v_{k_r} \rangle_{\mathbb{C}} \simeq \mathbb{C}^{k_r+1} \quad \text{and} \quad \mathcal{V}_r := \langle v_1, \ldots, v_{k_r} \rangle_{\mathbb{C}} \subseteq \overline{\mathcal{V}}_r,$$

and for each r, we define a map $M(r) : \mathcal{U}_r \to \overline{\mathcal{V}}_r$ by $u_0 \mapsto v_0$ and $u_i \mapsto v_0 + v_i$ when $i \neq 0$. As matrices,

$$M(r) = \left[\begin{array}{c|c} 1 & \mathbf{1}^T \\ \hline \mathbf{0} & I_{k_r} \end{array}\right] \in \overline{\mathcal{V}}_r \otimes \mathcal{U}_r^*, \quad \text{so} \quad M(r)^{-1} = \left[\begin{array}{c|c} 1 & -\mathbf{1}^T \\ \hline \mathbf{0} & I_{k_r} \end{array}\right] \in \overline{\mathcal{U}}_r \otimes \mathcal{V}_r^*. \tag{3.6}$$

Here $\mathbf{0}, \mathbf{1} \in \mathbb{C}^k$ denote the vector of zeros and ones respectively, and I_k is the $k \times k$ identity matrix. Then the matrix $\bar{\mu}$ in (3.5), where $n = 2$, is equal to $M(1) \cdot M(2) \cdot p$. For the case of general n where p is a higher order tensor, for each multiset $A \subseteq [n]$ we let $M(A) := \otimes_{r \in A} M(r)$, and define a *total moment* tensor,

$$\bar{\mu}_A := M(A) \cdot P_A \in \overline{\mathcal{V}}_A.$$

Definition 3.6.1. We define the *moment* tensor, $\mu_A \in \mathcal{V}_A$, as the block of $\bar{\mu}_A$ consisting of those entries with no 0 in their index.

Example 3.6.2. Suppose $n = 2$, $\mathcal{I}_1 = \{0, 1, 2\}$ and $\mathcal{I}_2 = \{0, 1\}$. Then for instance

$$\bar{\mu}_{12} = \left[\begin{array}{c|c} 1 & p_{+1} \\ \hline p_{1+} & p_{11} \\ p_{2+} & p_{21} \end{array}\right] = \left[\begin{array}{c|c} 1 & \mu_2 \\ \hline \mu_1 & \mu_{12} \end{array}\right], \quad \bar{\mu}_1 = \left[\begin{array}{c} 1 \\ \hline p_{1+} \\ p_{2+} \end{array}\right] = \left[\begin{array}{c} 1 \\ \hline \mu_1 \end{array}\right], \text{ and}$$

$$\bar{\mu}_{11} = \left[\begin{array}{c|cc} 1 & p_{1+} & p_{2+} \\ \hline p_{1+} & p_{1+} & 0 \\ p_{2+} & 0 & p_{2+} \end{array}\right] = \left[\begin{array}{c|c} 1 & \mu_1 \\ \hline \mu_1 & \mu_{11} \end{array}\right], \quad \text{so}$$

$$\mu_{12} = \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix}, \quad \mu_1 = \begin{bmatrix} p_{1+} \\ p_{2+} \end{bmatrix} \quad \text{and} \quad \mu_{11} = \begin{bmatrix} p_{1+} & 0 \\ 0 & p_{2+} \end{bmatrix}.$$

Remark 3.6.3. We avoid using transpose notation μ_2^T in the block decomposition for $\bar{\mu}_{12}$ above because $\bar{\mu}_{12} \in \overline{\mathcal{V}}_1 \otimes \overline{\mathcal{V}}_2$ is covariant in both indices, so we remain agnostic as to which direction in the array is "horizontal". As well, with higher order tensors, there is no single notion of transposition.

In Example 3.6.2 above, observe how $\bar{\mu}_1$ appears as a sub-tensor of $\bar{\mu}_{12}$. This phenomenon conveniently generalizes as follows:

Lemma 3.6.4. *Let $B \subseteq A \subseteq [n]$ then for every $a \in \mathcal{I}_A$ with support in B*

$$(\bar{\mu}_A)_a \quad = \quad (\bar{\mu}_B)_{a_B}.$$

Proof. We have

$$(\bar{\mu}_A)_a = (\bar{\mu}_{B \cup (A \setminus B)})_{a_B \mathbf{0}} = M(B)_{a_B} \cdot M(A \setminus B)_{\mathbf{0}} \cdot p_{B \cup (A \setminus B)}.$$

Since $M(A \setminus B)^{\kappa}_{\mathbf{0}} = 1$ for every $\kappa \in \mathcal{I}_{A \setminus B}$ then the above is equal to

$$M(B)_{a_B} \cdot \mathbf{1}(A \setminus B) \cdot p_{B \cup (A \setminus B)} = M(B)_{a_B} \cdot p_B = (\bar{\mu}_B)_{a_B}. \quad \square$$

Remark 3.6.5. Unlike P_{11}, the tensor $\bar{\mu}_{11}$ is not diagonal, but it is symmetric, and the block μ_{11} is in fact diagonal.

By the above lemma, the block structure observed in the $n = 2$ case above generalizes as follows: for each $a \in \mathcal{I}_A$ and $b \in \mathcal{I}_B$,

$$(\bar{\mu}_{AB})_{a\mathbf{0}} = (\bar{\mu}_A)_a \quad \text{and} \quad (\bar{\mu}_{AB})_{\mathbf{0}b} = (\bar{\mu}_B)_b. \tag{3.7}$$

Thus, as an array, any total moment tensor $\bar{\mu}_C$ is a disjoint union of the moments μ_A where $A \subseteq C$ (as multisets if they are multisets). In particular, as an array the tensor $\mu_{[n]}$ is the union of the tensors μ_A as A ranges over all subsets of $[n]$.

Next, we define *central moments*. We let $M'(r) : \mathcal{U}_r \to \overline{\mathcal{V}}_r$ be the matrices

$$M'(r) := \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline -\mu_r & I_{l_r} \end{array}\right] \circ M(r), \qquad \text{so} \qquad M'(r)^{-1} = M(r)^{-1} \circ \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mu_r & I_{l_r} \end{array}\right].$$

where $\circ$ denotes usual matrix multiplication (as a composition of linear maps). Then let

$$M'(A) := \bigotimes_{r \in A} M'(r) \tag{3.8}$$

to define a *total central moment* tensor

$$\bar{\mu}'_A := M'(A) \cdot P_A \in \overline{\mathcal{V}}_A.$$

As for non-central moments, we define the *(central) moment* μ'_A as the block of $\bar{\mu}'_A$ entries with no 0 in their index, and analogously to (3.7) we have

$$(\bar{\mu}'_{AB})_{a\mathbf{0}} = (\bar{\mu}'_A)_a \quad \text{and} \quad (\bar{\mu}'_{AB})_{\mathbf{0}b} = (\bar{\mu}'_B)_b. \tag{3.9}$$

To get a sense for how these quantities behave, observe that $\mu'_r = 0$ for $r = 1, \ldots, n$. As well, for any $r, s \in [n]$, using the block decomposition arising from $\overline{\mathcal{V}}_r = \mathbb{C} \oplus \mathcal{V}_r$ and $\overline{\mathcal{V}}_s = \mathbb{C} \oplus \mathcal{V}_s$ we have

$$\bar{\mu}_{rs} = \left[\begin{array}{c|c} 1 & \mu_s \\ \hline \mu_r & \mu_{rs} \end{array}\right] \in \overline{\mathcal{V}}_{rs} \qquad \textit{and} \qquad \bar{\mu}'_{rs} = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mu_{rs} - \mu_r \otimes \mu_s \end{array}\right] \in \overline{\mathcal{V}}_{rs}$$

so we see that $\mu'_{rs} = \mu_{rs} - \mu_r \otimes \mu_s$, and in particular,

$$\operatorname{rank}(\bar{\mu}_{rs}) = \operatorname{rank}(\bar{\mu}'_{rs}) = 1 + \operatorname{rank}(\mu'_r s).$$

This is convenient, because $\operatorname{rank}(\bar{\mu}_{rs}) = \operatorname{rank}(P_{rs}) \leq 1$ if and only if $r \perp\!\!\!\perp s$. We record this as a proposition:

Proposition 3.6.6. *For any two $r, s \in [n]$, we have $r \perp\!\!\!\perp s \iff \mu'_{rs} = 0$.*

Statistical interpretation of moments

The tensor $M(A)$ can be interpreted as an operator which computes the expectation of a certain formal random variable. Although this interpretation will not be necessary for proofs, it was the motivation behind most of our approach.

To make this interpretation precise, we define a *formal random variable* to be a map X with domain equal to a complex measure space. In our case the measure space is $\mathcal{I}$ with measure distribution p. If the measure space is a probability space then X is a random variable in the usual sense. When a formal random variable is vector-valued, say $X : \mathcal{I} \to \mathcal{V}$, it makes sense to talk about expressions like the *mean* or *expectation* of X as defined by

$$\mathbb{E}[X] := \sum_{\iota \in \mathcal{I}} p_\iota X(\iota) \in \mathcal{V}$$

even though the values p_ι are not probabilities. Other expressions typically defined in statistics also translate directly to this context.

To interpret μ_A as an expectation, for each $r \in [n]$ we define a special formal random vector $\mathcal{C}_r : \mathcal{I} \to \mathcal{I}_r \to \mathcal{V}_r$ with values $\mathcal{C}_r(\iota) = 0$ if $\iota_r = 0$ and otherwise $\mathcal{C}_r(\iota) = v_{\iota_r}$. The values $\{0, v_1, v_2, \ldots, v_{k_r}\}$ are called the *states* of $\mathcal{C}_r$, which form the shape of a "corner" in the positive orthant. Then letting $\mathcal{C}_A := \bigotimes_{r \in A} \mathcal{C}_r$, we have the alternative definition

$$\mu_A = \mathbb{E}(\mathcal{C}_A)$$

as the A^{th} moment of the system of random vectors $\mathcal{C}$. We can likewise define random variables $\overline{\mathcal{C}}_r := \mathcal{C}_r + v_0 \in \overline{\mathcal{V}}_r$ and $\overline{\mathcal{C}}_A := \bigotimes_{r \in A} \overline{\mathcal{C}}_r$ so that

$$\bar{\mu}_A = \mathbb{E}(\overline{\mathcal{C}}_A).$$

As for central moments, defining random variables $\mathcal{C}'_r := \mathcal{C}_r - \mu_r$ and $\mathcal{C}'_A := \bigotimes_{r \in A} \mathcal{C}'_r$ gives us the interpretation

$$\mu'_A = \mathbb{E}(\mathcal{C}'_A)$$

of μ'_A as a central moment in the usual sense (i.e., as the expectation of a random variable whose mean is 0).

3.7 Regression tensors

Linear regression is fundamental in the analysis of classical Gaussian random variables. For example, directed Gaussian graphical models are defined by regressing each node on the collection of its parent nodes. We now explore linear regression for formal discrete random variables as a way to reparametrize directed discrete graphical models. We will introduce our various regression coefficients at first in a form that is most convenient for our proofs, and explain their statistical meaning in Section 3.7.

For any two disjoint multisets $A, B \subseteq [n]$, we define the *total regression tensor*

$$\begin{aligned}\bar{\mu}_{B|A} &:= \bar{\mu}_{AB} \cdot \bar{\mu}_{AA}^{-1} \qquad (3.10)\\ &= (M(A) \cdot M(B) \cdot P_{AB}) \cdot (M(A) \cdot M(A) \cdot P_{AA})^{-1} \\ &= (M(A) \cdot M(B) \cdot P_{AB}) \cdot (M(A)^{-1} \cdot M(A)^{-1} \cdot P_{AA}^{-1}) \\ &= M(B) \cdot P_{BA} \cdot P_{AA}^{-1} \cdot M(A)^{-1} \qquad \text{(by lemma 3.3.5)} \\ &= M(B) \cdot P_{B|A} \cdot M(A)^{-1}. \qquad (3.11)\end{aligned}$$

Here, $\bar{\mu}_{AA}^{-1} \in \mathcal{V}_A^* \otimes \mathcal{V}_A^*$ denotes the inverse of the tensor $\mu_{AA} \in \mathcal{V}_A \otimes \mathcal{V}_A$, as in Lemma 3.3.5, and similarly for $P_{AA}^{-1} \in \mathcal{U}_A^* \otimes \mathcal{U}_A^*$ and $M(A)^{-1} \in \mathcal{U}_A \otimes \overline{\mathcal{V}}_A^*$.

For example, when $n = 2$, in the block decomposition arising from $\overline{\mathcal{V}}_r = \mathbb{C} \oplus \mathcal{V}_r$ we can use the definition $\bar{\mu}_{2|1} := \bar{\mu}_{12} \cdot \bar{\mu}_{11}^{-1}$ directly to compute that

$$\bar{\mu}_{2|1} = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline \frac{p_{01}}{p_{0+}} & \frac{p_{11}}{p_{1+}} - \frac{p_{01}}{p_{0+}} & \cdots & \frac{p_{k_1 1}}{p_{k_1+}} - \frac{p_{01}}{p_{0+}} \\ \frac{p_{02}}{p_{0+}} & \frac{p_{12}}{p_{1+}} - \frac{p_{02}}{p_{0+}} & \cdots & \frac{p_{k_1 2}}{p_{k_1+}} - \frac{p_{02}}{p_{0+}} \\ \vdots & & \cdots & \vdots \\ \frac{p_{0k_2}}{p_{0+}} & \frac{p_{1k_2}}{p_{1+}} - \frac{p_{0k_2}}{p_{0+}} & \cdots & \frac{p_{k_1 k_2}}{p_{k_1+}} - \frac{p_{0k_2}}{p_{0+}} \end{array}\right] \in \overline{\mathcal{V}}_2 \otimes \overline{\mathcal{V}}_1^* \qquad (3.12)$$

The above block formation generalizes for larger n:

Lemma 3.7.1. *Writing $0 \in \mathcal{I}_B$ for a sequences of all zeroes, we have for all $a \in I_A$,*

$$(\bar{\mu}_{B|A})_0^a = \delta_0^a.$$

Proof.

$$(\bar{\mu}_{B|A})_0^a = \sum_{a' \in \mathcal{I}_A} (\bar{\mu}_{AA}^{-1})^{aa'} (\bar{\mu}_{AB})_{a'0}$$

Since $(\bar{\mu}_{AB})_{a'0} = (\bar{\mu}_A)_{a'} = (\bar{\mu}_{AA})_{a'0}$ by (3.7), the above equations can be rewritten as

$$(\bar{\mu}_{B|A})_0^a = \sum_{a' \in \mathcal{I}_A} (\bar{\mu}_{AA}^{-1})^{aa'} (\bar{\mu}_{AA})_{a'0} = \delta_0^a$$

where the last equation follows by the definition of the inverse tensor. □

Statistical interpretation of regression tensors

The motivation for this terminology is that, following the statistical interpretation of moments given in Section 3.6, $\bar{\mu}_{B|A}$ is actually the coefficient of linear regression expressing $\overline{\mathcal{C}}_B$ in terms of $\overline{\mathcal{C}}_A$ defined in that section:

$$\mathbb{E}(\overline{\mathcal{C}}_B|\overline{\mathcal{C}}_A) = \bar{\mu}_{B|A}\,\overline{\mathcal{C}}_A.$$

That is, for every $a \in \mathcal{I}_A$, $\mathbb{E}[\overline{\mathcal{C}}_B|\overline{\mathcal{C}}_A(a)] = \bar{\mu}_{B|A}\overline{\mathcal{C}}_A(a)$, which explains why $\bar{\mu}_{B|A}$ is called a regression tensor.

A useful observation is that the regression tensor holds information about independence and conditional independence.

Lemma 3.7.2. *If μ_{AA}^{-1} is a well defined tensor then $A \perp\!\!\!\perp B$ if and only if the block $(\bar{\mu}_{B|A})_{\neq \mathbf{0}}^{\neq \mathbf{0}}$ is a zero tensor.*

3.8 Directed acyclic graphical models

Directed acyclic graphical (DAG) models on discrete variables are traditionally described by parametrizing their joint probability distribution in terms of conditional and root probabilities. Often, only a subset S of the nodes in graph represent *observed* random variables, and in this section, we study how these models can be described more efficiently by parametrizing their moments in terms of linear regression coefficients. This parametrization can be expressed combinatorially in terms of subgraphs H of the graph such that $\text{sinks}(H) \subseteq S \subseteq \text{nodes}(H)$.

Given a directed acyclic graph $G = (V, E)$ and a finite set $\mathcal{I}_v = \{0, 1, \ldots, k_v\}$ for each node $v \in V$, we study a special family of affine distributions in $\mathcal{U} := \mathbb{C}^{\mathcal{I}}$ where $\mathcal{I} := \prod_{v \in V} \mathcal{I}_v$.

Definition 3.8.1. For any subgraph $H \subseteq G$, we write

$$\begin{aligned} \text{pa}(v; H) &:= \text{ the set of parents of } v \text{ in } H, \text{ and} \\ \text{ch}(v; H) &:= \text{ the set of children of } v \text{ in } H. \end{aligned}$$

When $H = G$, we simply write $\text{pa}(v)$ and $\text{ch}(v)$ respectively for the sets of parents and children of v in G.

An affine distribution p is said to belong to the model G if there are conditional probability tensors $P_{v|\text{pa}(v)}$ such that:

$$p_\iota = \prod_{v \in V} (P_{v|\text{pa}(v)})_{\iota_v}^{\iota_{\text{pa}(v)}} \qquad \text{for all } \iota \in \mathcal{I}. \tag{3.13}$$

By treating the conditional probabilities $P_{v|\text{pa}(v)}$ as parameters we are able to characterize all distributions $P_{[n]}$ that can be factored according to G as above. We refer to this set of distributions as the *fully observed* graphical model for G.

These models are very well understood algebraically; for example, see [16]. However, as soon as some of the nodes in the graph represent portion of data that is not observed, they become much harder to analyze. If only a subset $S \subseteq V$ of the nodes are observed, this corresponds to taking the marginal distribution P_S, where $p = P_V$ satisfies (3.13).

An example that occurs frequently in applications is that of a *rooted tree*, that is, a directed tree with a unique inner source node called the *root*. In this case, we assume that the leaves of the tree are observed; these leaves are the sinks. Meanwhile all the other nodes are assumed to be hidden

Now suppose $G = (V, E)$ is any DAG. The parametrization in (3.13) can be written as an automatic contraction (Section 3.3),

$$p = P_V = \prod_{v \in V} P_{v^{|\mathrm{ch}(v)|+1}\mathrm{pa}(v)}$$

where $v^{|\mathrm{ch}(v)|+1}$ denotes $\underbrace{vv\ldots v}_{|\mathrm{ch}(v)|+1 \text{ times}}$, so that each

$$P_{v^{|\mathrm{ch}(v)|+1}\mathrm{pa}(v)} \in \mathcal{U}_v^{\otimes(|\mathrm{ch}(v)|+1)} \otimes \mathcal{U}^*_{\mathrm{pa}(v)}.$$

For example, if $G = \boxed{2 \leftarrow 1 \rightarrow 3}$, then we have

$$P_{123} = P_{111} \cdot P_{2|1} \cdot P_{3|1}, \quad i.e.,$$
$$(P_{123})_\iota = (P_1)_{\iota_1} (P_{2|1})^{\iota_1}_{\iota_2} (P_{3|1})^{\iota_1}_{\iota_3}. \tag{3.14}$$

To generalize this formula, for any set of observed nodes S and subgraph $H \subseteq G$ we define the *multiplicity* of v in S and H to be

$$m(v; H, S) := |\mathrm{ch}(v; H)| + \begin{cases} 0 & \text{if } v \notin S \\ 1 & \text{if } v \in S \end{cases} \tag{3.15}$$

For example, if $G = \boxed{2 \leftarrow 1 \rightarrow 3}$ and $S = \{2, 3\}$ then the marginal distribution P_{23} is obtained by summing over all possible values of ι_1 in equation (3.14) above:

$$P_{23} = P_{11} \cdot P_{2|1} \cdot P_{3|1}.$$

Graphical linear regression coefficients

Our next result provides a formula for the moments of the observed nodes in terms of blocks of regression tensors which arise from regressing each variable on its parents. We continue to work with a DAG model $G = (V, E)$, and for any $\kappa \in I_{\mathrm{pa}(v)}$ and $v \in \mathrm{nodes}(G)$ we will write $\kappa(v)$ for the value of κ indexed by v and similarly for $\kappa \in I_{\mathrm{pa}(v;H)}$.

Definition 3.8.2 (Graphical linear regression coefficients)**.** For each subset of nodes $A \subseteq \mathrm{pa}(v)$, we define the regression coefficient

$$\beta_v^A \in \mathcal{V}_v \otimes \mathcal{V}_A^*$$

by extracting entries of $\bar{\mu}_{v|\mathrm{pa}(v)}$ as follows. For each $\kappa \in I_A$, we define a new index $\kappa' \in I_{\mathrm{pa}(v)}$ by

$$\kappa'(u) = \begin{cases} \kappa(u) & \text{if } u \in A \\ 0 & \text{if } u \notin A \end{cases}$$

and then for every $(i\kappa) \in \mathcal{I}_v \times \mathcal{I}_A$ where $i \neq 0$ we let

$$(\beta_v^A)_i^\kappa := (\bar{\mu}_{v|\mathrm{pa}(v)})_i^{\kappa'}$$

In other words, β_v^A "forgets" the upper indices of $\bar{\mu}_{v|\mathrm{pa}(v)}$ which correspond to vertices that are not in A by always using a 0 at that index position. We sometimes refer to the tensor $\beta_v^\emptyset$ as a *residual mean*, because it is the mean of the residual in the linear regression of $\mathcal{C}_v$ on its parents (see Section 3.6).

More generally, for each integer $m > 0$ we define the tensor $\beta_{v^m}^A \in (\mathcal{V}_v)^{\otimes m} \otimes \mathcal{V}_A^*$ such that for each $(\iota, \kappa) \in (\mathcal{I}_v)^m \times \mathcal{I}_A$ with $0 \notin \iota$, we have

$$(\beta_{v^m}^A)_\iota^\kappa := (\bar{\mu}_{v^m|\mathrm{pa}(v)})_\iota^{\kappa'} = (\bar{\mu}_{m'|\mathrm{pa}(v)})_{\iota\mathbf{0}}^{\kappa'}, \tag{3.16}$$

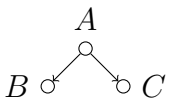
where the last equation holds for any $m' > m$ by Lemma 3.6.4. Thus, for each $\kappa \in \mathcal{I}_A$ with $0 \notin \kappa$, $(\beta_{v^m}^A)_\bullet^\kappa$ is a diagonal tensor with the entries of $(\beta_v^A)_\bullet^\kappa$ along its diagonal.

We will show that these β tensors allow us to express μ_S efficiently as a summation over certain subgraphs of G.

Theorem 3.8.3. *Given a discrete directed acyclic graph model G and a multiset S of observed nodes, the moment μ_S is given by the following automatic contraction equation:*

$$\mu_S \quad = \sum_{\substack{H \subset G \\ \mathrm{sinks}(H) \subseteq S \subseteq \mathrm{nodes}(H)}} \prod_{v \in \mathrm{nodes}(H)} \beta_{v^{m(v;H,S)}}^{\mathrm{pa}(v;H)}$$

Example 3.8.4. Before proving the result, let us walk through it in a very simple example (more examples follow after the proof). Consider a model where G is the graph



supposing each variable has 3 states. The fully observed joint distribution of this model is $p \in \mathbb{C}^{\mathcal{I}_A \times \mathcal{I}_B \times \mathcal{I}_C} \simeq \mathbb{C}^{3\times 3\times 3}$, whose entries we denote by p_{abc}.

The moments μ_S of this model will live in tensor products of the vector spaces $\mathcal{V}_A \simeq \mathbb{C}^2$, $\mathcal{V}_B \simeq \mathbb{C}^2$, and $\mathcal{V}_C \simeq \mathbb{C}^2$. To compute the moment $\mu_B \in \mathcal{V}_B$, we apply Theorem 3.8.3 where $S = \{B\}$, which needs the only the three tensors $\beta_A^\emptyset \in \mathcal{V}_A$, $\beta_B^A \in \mathcal{V}_B \otimes \mathcal{V}_A^*$, and $\beta_B^\emptyset \in \mathcal{V}_B$. These tensors arise as blocks of the larger tensors $\bar{\mu}_{A|\emptyset} = \bar{\mu}_A \in \overline{\mathcal{V}}_A \simeq \mathbb{C}^3$ and $\bar{\mu}_{B|A} \in \overline{\mathcal{V}}_B \otimes \mathcal{V}_A^* \simeq \mathbb{C}^{3\times 3}$ as follows:

$$\bar{\mu}_{A|\emptyset} = \left[\begin{array}{c} 1 \\ \hline p_{1++} \\ p_{2++} \end{array}\right] = \left[\begin{array}{c} 1 \\ \hline \beta_A^\emptyset \end{array}\right], \text{ and}$$

$$\bar{\mu}_{B|A} = \left[\begin{array}{c|cc} 1 & 0 & 0 \\ \hline \frac{p_{01+}}{p_{0++}} & \frac{p_{11+}}{p_{1++}} - \frac{p_{01+}}{p_{0++}} & \frac{p_{21+}}{p_{2++}} - \frac{p_{01+}}{p_{0++}} \\ \frac{p_{02+}}{p_{0++}} & \frac{p_{12+}}{p_{1++}} - \frac{p_{02+}}{p_{0++}} & \frac{p_{22+}}{p_{2++}} - \frac{p_{02+}}{p_{0++}} \end{array}\right] = \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \beta_B^\emptyset & \beta_B^A \end{array}\right]$$

Note that the tensor $\beta_A^\emptyset$ is essentially a vector, and the tensor β_B^A is essentially a matrix.

There are two subgraphs $H \subseteq G$ such that $\text{sinks}(H) \subseteq \{B\} \subseteq \text{nodes}(H)$: the graph $\boxed{A \to B}$, which gives rise to automatic contraction term $\beta_B^A \cdot \beta_A^\emptyset$, and $\boxed{B}$, giving rise to the term $\beta_B^\emptyset$. Hence the theorem simply says that

$$\mu_B = \beta_B^A \cdot \beta_A^\emptyset + \beta_B^\emptyset = \left[\begin{array}{cc} \frac{p_{11+}}{p_{1++}} - \frac{p_{01+}}{p_{0++}} & \frac{p_{21+}}{p_{2++}} - \frac{p_{01+}}{p_{0++}} \\ \frac{p_{22+}}{p_{1++}} - \frac{p_{02+}}{p_{0++}} & \frac{p_{22+}}{p_{2++}} - \frac{p_{02+}}{p_{0++}} \end{array}\right] \left[\begin{array}{c} p_{1++} \\ p_{2++} \end{array}\right] + \left[\begin{array}{c} \frac{p_{01+}}{p_{0++}} \\ \frac{p_{02+}}{p_{0++}} \end{array}\right] = \left[\begin{array}{c} p_{+1+} \\ p_{+2+} \end{array}\right]$$

where in the last step we used the fact that $p_{1++} + p_{2++} = 1 - p_{0++}$. Similarly we have $\mu_C = \beta_C^A \cdot \beta_A^\emptyset + \mu_C^\emptyset$. The moment μ_{BC} is more interesting: the formula involves four subgraphs, namely $\boxed{B \leftarrow A \to C}$, $\boxed{B \leftarrow A \quad C}$, $\boxed{B \quad A \to C}$, and $\boxed{B \quad C}$, which in that order give rise to the sum

$$\mu_{BC} = \beta_{AA}^\emptyset \cdot \beta_B^A \cdot \beta_C^A + \beta_A^\emptyset \cdot \beta_B^A \cdot \beta_C^\emptyset + \beta_A^\emptyset \cdot \beta_B^\emptyset \cdot \beta_C^A + \beta_B^\emptyset \cdot \beta_C^\emptyset$$

Here the tensor $\beta_{AA}^\emptyset$ is defined according to Definition 3.6.1 as a block of $\bar{\mu}_{AA|\emptyset}$ as follows:

$$\bar{\mu}_{AA|\emptyset} = \left[\begin{array}{c|cc} 1 & p_{1++} & p_{2++} \\ \hline p_{1++} & p_{1++} & 0 \\ p_{2++} & 0 & p_{2++} \end{array}\right] = \left[\begin{array}{c|c} 1 & \beta_A^\emptyset \\ \hline \beta_A^\emptyset & \beta_{AA}^\emptyset \end{array}\right], \quad \text{so} \quad \beta_{AA}^\emptyset = \left[\begin{array}{cc} p_{1++} & 0 \\ 0 & p_{2++} \end{array}\right]$$

It is interesting to note that considered as matrix, it turns out that $\beta^{\emptyset}_{AA} = \mathrm{diag}(\beta^{\emptyset}_{\mathrm{A}})$, while it is not true that $\bar{\mu}^{\emptyset}_{AA} \neq \mathrm{diag}(\mu^{\emptyset}_{\mathrm{A}})$. This is a nice feature of the β tensors as we have defined them. As well, note that here $\beta^{\emptyset}_{A} = \mu_A$ because A has no parents, but generically $\beta^{\emptyset}_{B} \neq \mu_B$ and $\beta^{\emptyset}_{C} \neq \mu_C$.

Note also the the model of all $p = P_{ABC}$ arising from this DAG model for some assignment of conditional probabilities $p_{v|\mathrm{pa}(v)}$ is a 14-dimensional model, which is parametrized generically injectively (and hence birationally) by the 14 entries of the tensors of $\beta_v^{\mathrm{pa}(v)}$ considered as indeterminates. This is always the case for fully observed DAG models, because the tensors $\beta_v^{\mathrm{pa}(v)}$ incorporate all the non-constant entries of the conditional moments $\bar{\mu}_{v|\mathrm{pa}(v)}$, as can be seen for instance in Example 3.8.4, and the moments $\bar{\mu}_{v|\mathrm{pa}(v)}$ are in bijection with the usual model parameters $P_{v|\mathrm{pa}(v)}$ via (3.11).

We now proceed with the proof of the theorem. The reader is invited to work through the meaning of the proof in terms of the example above, or Example 3.8.5 immediately following the proof.

Proof of Theorem 3.8.3. For brevity, let us write $\beta(v; H) = \beta^{\mathrm{pa}(v;H)}_{v^{m(v;H,S)}}$, $m(v) = m(v; G, S)$, and $m(v; H) = m(v; H, S)$. We continue to write $V = \mathrm{nodes}(G)$ and $E = \mathrm{edges}(G)$.

First observe by (3.11) and (3.8) and that

$$\prod_{v\in V} \bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)} = \prod_{v\in V} \left(M(v)^{\otimes m(v)} \cdot P_{v^{m(v)}|\mathrm{pa}(v)} \cdot \prod_{w\in\mathrm{pa}(v)} M(w)^{-1} \right)$$

Here, for every $v \in V$, $M(v)$ occurs as many times as $M(v)^{-1}$, except once more for each time v occurs in S, so most of them cancel and so the above expression is

$$\prod_{v\in V} \bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)} = \left(\prod_{s\in S} M(s) \right) \cdot \left(\prod_{v\in V} P_{v^{m(v)}|\mathrm{pa}(v)} \right) = M(S) \cdot P_S = \bar{\mu}_S$$

using (3.13) and the definition of $\bar{\mu}_S$. In summary,

$$\bar{\mu}_S = \prod_{v\in V} \bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)} \tag{3.17}$$

We will now manipulate the above contraction in coordinate form to arrive at the desired result.

The first step is essentially to rewrite the contraction as a summation which uses one summation symbol for each edge in the graph G, where the symbol for an edge

ranges over the states of the source node of that edge. To make this precise, we define

$$\mathcal{I}_E := \prod_{uv \in E} \mathcal{I}_u$$

For any edge $uv \in E$ and $\kappa \in \mathcal{I}_E$ we write $\kappa(uv) \in \mathcal{I}_u$ for the entry of κ indexed by uv. For any vertex v of G we define subsequences $\kappa(\mathrm{pa}(v)) := (\kappa(uv) \mid u \in \mathrm{pa}(v)) \in \mathcal{I}_{\mathrm{pa}(v)}$ and $\kappa(\mathrm{ch}(v)) := (\kappa(vu) \mid u \in \mathrm{ch}(v)) \in \mathcal{I}_v^{|\mathrm{ch}(v)|}$. Then we can rewrite the contraction (3.17) in terms of scalars by saying that for each $\iota \in \mathcal{I}_S$

$$(\bar{\mu}_S)_\iota = \sum_{\kappa \in \mathcal{I}_E} \underbrace{\left(\prod_{v \in V \setminus S} (\bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)})_{\kappa(\mathrm{ch}(v))}^{\kappa(\mathrm{pa}(v))} \cdot \prod_{v \in S} (\bar{\mu}_{v|\mathrm{pa}(v)})_{\iota(v)\kappa(\mathrm{ch}(v))}^{\kappa(\mathrm{pa}(v))} \right)}_{(\bigstar)} \tag{3.18}$$

Now for every $\kappa \in \mathcal{I}_E$, we define a corresponding subgraph H_κ as follows. Let $E_\kappa := \{uv \in E \mid \kappa_{uv} \neq 0\}$, and V_κ be the union of S and the endpoints of all the edges $uv \in E_\kappa$. We then let H_κ be the subgraph of G with vertices V_κ and edges E_κ. Note that $S \subseteq \mathrm{nodes}(H_\kappa) = V_\kappa$ by construction.

The point now is that the product $(\bigstar)$ in equation (3.18) is equal to 0 unless $\mathrm{sinks}(H_\kappa) \subseteq S$. To see this, suppose there is some $u \in \mathrm{sinks}(H_k) \setminus S$. This means that $\kappa_{\mathrm{ch}(v)} = \mathbf{0}$, so by Lemma 3.7.1, $(\bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)})_{\kappa_{\mathrm{ch}(v)}}^{\kappa_{\mathrm{pa}(v)}} = \delta_{\mathbf{0}}^{\kappa_{\mathrm{pa}(v)}}$ However, by the construction of H_κ, since $v \notin S$ and v is not the source of an edge in E_κ (it has no children in H_κ), it must be the target of an edge in E_κ. That is, we must have $\kappa_{\mathrm{pa}(v)} \neq \mathbf{0}$, so the Kronecker $\delta_{\mathbf{0}}^{\kappa_{\mathrm{pa}(v)}} = 0$, and hence $(\bigstar) = 0$. Hence the only non-zero terms in equation (3.18) must occur when $\mathrm{sinks}(H_\kappa) \subseteq S$, so we can rewrite it as

$$(\bar{\mu}_S)_\iota = \sum_{\substack{H \subset G \\ \mathrm{sinks}(H) \subseteq S \subseteq \mathrm{nodes}(H)}} \sum_{\substack{\kappa \in \mathcal{I}_E \\ H_\kappa = H}} (\bigstar).$$

Also, if $v \notin V_\kappa$ then we have $\kappa(\mathrm{pa}(v)) = \mathbf{0}$ and $\kappa(\mathrm{ch}(v)) = \mathbf{0}$, so

$$(\bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)})_{\kappa(\mathrm{ch}(v))}^{\kappa(\mathrm{pa}(v))} = (\bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)})_{\mathbf{0}}^{\mathbf{0}} = 1$$

and we may remove these factors from the product $(\bigstar)$ without changing it, i.e.,

$$(\bigstar) = \prod_{v \in V_k \setminus S} (\bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)})_{\kappa(\mathrm{ch}(v))}^{\kappa(\mathrm{pa}(v))} \cdot \prod_{v \in S} (\bar{\mu}_{v|\mathrm{pa}(v)})_{\iota(v)\kappa(\mathrm{ch}(v))}^{\kappa(\mathrm{pa}(v))}.$$

Now, the sequence $\kappa(\mathrm{ch}(v; H_\kappa))$ is equal to $\kappa(\mathrm{ch}(v))$ with all occurrences of 0 removed, so by equation (3.16) we can write

$$(\bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)})_{\kappa(\mathrm{ch}(v))}^{\kappa(\mathrm{pa}(v))} = \beta(v; H)_{\kappa(\mathrm{ch}(v;H_\kappa))}^{\kappa(\mathrm{pa}(v))}$$

and assuming now that ι contains no zeroes, we can also write

$$(\bar{\mu}_{v^{m(v)}|\mathrm{pa}(v)})^{\kappa(\mathrm{pa}(v))}_{\iota(v)\kappa(\mathrm{ch}(v))} = \beta(v;H)^{\kappa(\mathrm{pa}(v))}_{\iota(v)\kappa(\mathrm{ch}(v;H_\kappa))}.$$

Putting these together, we have

$$(\bigstar) = \prod_{v\in V_\kappa\setminus S} \beta(v;H)^{\kappa(\mathrm{pa}(v))}_{\kappa(\mathrm{ch}(v;H_\kappa))} \cdot \prod_{v\in S} \beta(v;H)^{\kappa(\mathrm{pa}(v))}_{\iota(v)\kappa(\mathrm{ch}(v;H_\kappa))}$$

and observe that

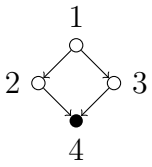
$$\sum_{\substack{\kappa\in\mathcal{I}_E\\ H_\kappa=H}} (\bigstar) = \left(\prod_{v\in\mathrm{nodes}(H)} \beta(v;H)\right)_\iota$$

where the latter product is an automatic tensor contraction. Finally, we assumed ι has no zeroes, so have $(\mu_S)_\iota = (\bar{\mu}_S)_\iota$, hence summing over H we obtain as required

$$(\mu_S)_\iota = \left(\sum_{\substack{H\subset G\\ \mathrm{sinks}(H)\subseteq S\subseteq\mathrm{nodes}(H)}} \prod_{v\in\mathrm{nodes}(H)} \beta(v;H)\right)_\iota$$

□

Example 3.8.5. Consider the following directed acyclic graph G.



and consider observing $S = \{4\}$. The joint probability

$$P = P_{111}\cdot P_{22|1}\cdot P_{33|1}\cdot P_{4|23}, \quad i.e., \quad p_{ijk\ell} = (P_1)_i (P_{2|1})^i_j (P_{3|1})^i_k (P_{4|23})^{jk}_\ell$$

is a product of conditional and root probabilities. Meanwhile, the marginal probability P_4 is

$$P_4 = P_{11}\cdot P_{2|1}\cdot P_{3|1}\cdot P_{4|23}.$$

Short formulae such as this one often have a long expansion in terms of conditional and root probabilities. For instance, when all the random variables have two states, the model has 9 free parameters: $(P_1)_1$, $(P_{2|1})^0_1$, $(P_{2|1})^1_1$, $(P_{3|1})^0_1$, $(P_{3|1})^1_1$, $(P_{4|23})^{00}_1$,

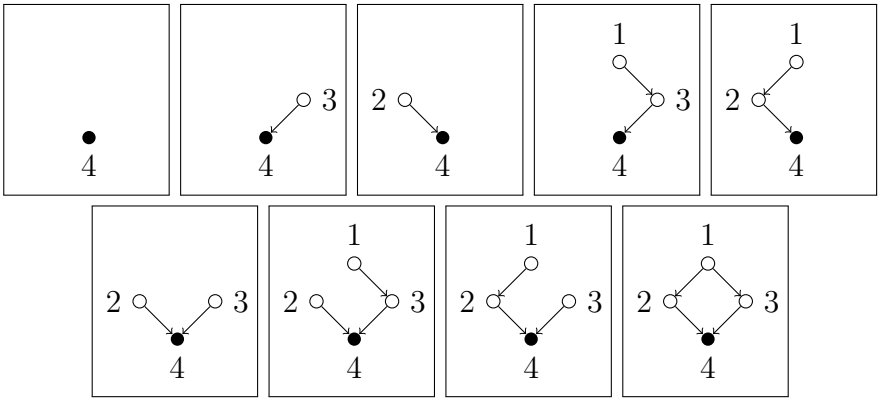
$(P_{4|23})^{01}_1$, $(P_{4|23})^{10}_1$, and $(P_{4|23})^{11}_1$. The probability $(P_4)_1$ expands into a quartic polynomial with 25 terms after substituting

$$\begin{aligned}(P_1)_0 &= 1 - (P_1)_1,\\ (P_{2|1})^i_0 &= 1 - (P_{2|1})^i_1 \quad \text{for } i \in \{0,1\},\\ (P_{3|1})^i_0 &= 1 - (P_{3|1})^i_1 \quad \text{for } i \in \{0,1\},\\ (P_{4|23})^{jk}_0 &= 1 - (P_{4|23})^{jk}_1 \quad \text{for } j,k \in \{0,1\}.\end{aligned}$$

Alternatively, by equation (3.17) we can parametrize this model using moments and regression coefficients as

$$\bar{\mu}_4 = \bar{\mu}_{11} \cdot \bar{\mu}_{2|1} \cdot \bar{\mu}_{3|1} \cdot \bar{\mu}_{4|23}.$$

To apply Theorem 3.8.3, observe that the subgraphs $\{H \mid \text{sinks}(H) \subseteq \{4\} \subseteq \text{nodes}(H)\}$ are:



which respectively give rise to the terms of the following sum for μ_r (ommitting $\cdot$'s):

$$\begin{aligned}\mu_4 \quad = \quad & \beta^\emptyset_4 + \beta^\emptyset_3\beta^3_4 + \beta^\emptyset_2\beta^2_4 + \beta^\emptyset_1\beta^1_3\beta^3_4 + \beta^\emptyset_1\beta^1_2\beta^2_4 \\ & + \beta^\emptyset_2\beta^\emptyset_3\beta^{23}_4 + \beta^\emptyset_1\beta^1_3\beta^\emptyset_2\beta^{23}_4 + \beta^\emptyset_1\beta^1_2\beta^\emptyset_3\beta^{23}_4 + \beta^\emptyset_{11}\beta^1_2\beta^1_3\beta^{23}_4.\end{aligned} \tag{3.19}$$

When $k_1 = k_2 = k_3 = k_4 = 1$, all of the tensors' array dimensions are 1, so they are effectively scalars. Thus, the model can be parametrized by the 9 scalar parameters

$$\beta^\emptyset_1, \beta^\emptyset_2, \beta^\emptyset_3, \beta^\emptyset_4, \beta^1_2, \beta^1_3, \beta^2_4, \beta^3_4, \beta^{23}_4$$

using the 9-term quartic polynomial (3.19). Recall that the entries of $\beta^\emptyset_{11}$ are determined by the entries of $\beta^\emptyset_1$, and in this situation they are both just the same scalar. The transformation between conditional probabilities and regression coefficients is the linear map

$$\beta^\emptyset_1 = \beta^\emptyset_{11} = (P_1)^1,$$

$$\begin{aligned}
\beta_2^{\emptyset} &= (P_{2|1})_1^0, & \beta_2^1 &= (P_{2|1})_1^1 - (P_{2|1})_1^0, \\
\beta_3^{\emptyset} &= (P_{3|1})_1^0, & \beta_3^1 &= (P_{3|1})_1^1 - (P_{3|1})_1^0, \\
\beta_4^{\emptyset} &= (P_{4|23})_1^{00}, & \beta_4^2 &= (P_{4|23})_1^{10} - (P_{4|23})_1^{00}, \\
& & \beta_4^3 &= (P_{4|23})_1^{01} - (P_{4|23})_1^{00}, \\
\beta_4^{23} &= (P_{4|23})_1^{11} + (P_{4|23})_1^{00} - (P_{4|23})_1^{01} - (P_{4|23})_1^{10}.
\end{aligned}$$

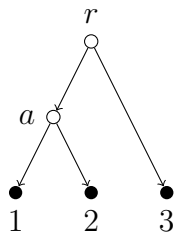
Moreover, the moment μ_4 is equal to the probability $(P_4)_1$. □

When the graph G is a rooted tree, this graphical model can also be parametrized using the means μ_v instead of the residual means $\beta_v^{\emptyset}$ where v varies over all the nodes of G. More specifically, because each node has at most one parent, we have

$$\mu_{v^d} = \beta_{v^d}^{\emptyset} + \beta_{v^d}^{\mathrm{pa}(v)} \mu_{\mathrm{pa}(v)}.$$

This equation lets us express each $\beta_{v^d}^{\emptyset}$ in terms of μ_{v^d} which in turn only depends on μ_v. Thus, we can alternatively parametrize the model using parameters μ_v and $\beta_v^{\mathrm{pa}(v)}$ for each node v.

Example 3.8.6. Consider the following rooted tree with three leaves.



There are 14 sink-connected subgraphs with $\{1, 2, 3\}$ as sinks, giving us the expansion

$$\begin{aligned}
\mu_{123} = {} & \beta_1^{\emptyset}\beta_2^{\emptyset}\beta_3^{\emptyset} + \beta_1^{\emptyset}\beta_2^{\emptyset}\beta_r^{\emptyset}\beta_3^r + \beta_a^{\emptyset}\beta_1^a\beta_2^{\emptyset}\beta_3^{\emptyset} + \beta_a^{\emptyset}\beta_1^a\beta_2^{\emptyset}\beta_r^{\emptyset}\beta_3^r + \beta_r^{\emptyset}\beta_a^r\beta_1^a\beta_2^{\emptyset}\beta_3^{\emptyset} \\
& + \beta_{rr}^{\emptyset}\beta_a^r\beta_1^a\beta_3^r\beta_2^{\emptyset} + \beta_a^{\emptyset}\beta_2^a\beta_1^{\emptyset}\beta_3^{\emptyset} + \beta_1^{\emptyset}\beta_a^{\emptyset}\beta_2^a\beta_r^{\emptyset}\beta_3^r + \beta_1^{\emptyset}\beta_r^{\emptyset}\beta_a^r\beta_2^a\beta_3^{\emptyset} + \beta_{rr}^{\emptyset}\beta_a^r\beta_2^a\beta_3^r\beta_1^{\emptyset} \\
& + \beta_{aa}^{\emptyset}\beta_1^a\beta_2^a\beta_3^{\emptyset} + \beta_{aa}^{\emptyset}\beta_1^a\beta_2^a\beta_r^{\emptyset}\beta_3^r + \beta_r^{\emptyset}\beta_{aa}^r\beta_1^a\beta_2^a\beta_3^{\emptyset} + \beta_{rr}^{\emptyset}\beta_{aa}^r\beta_1^a\beta_2^a\beta_3^r.
\end{aligned}$$

By expressing the residual means $\beta_v^{\emptyset}$ in terms of the means μ_v via the formulas

$$\begin{gathered}
\beta_1^{\emptyset} = \mu_1 - \mu_a\beta_1^a, \quad \beta_2^{\emptyset} = \mu_2 - \mu_a\beta_2^a, \quad \beta_3^{\emptyset} = \mu_3 - \mu_r\beta_3^r, \\
\beta_a^{\emptyset} = \mu_a - \mu_r\beta_a^r, \quad \beta_{aa}^{\emptyset} = \mu_{aa} - \mu_r\beta_{aa}^r,
\end{gathered}$$

we get the following 11-term expansion

$$\begin{aligned}
\mu_{123} = {} & \mu_3(\mu_{aa} - \mu_a^2)\beta_1^a\beta_2^a + \mu_2(\mu_{rr} - \mu_r^2)\beta_a^r\beta_1^a\beta_3^r + \mu_1(\mu_{rr} - \mu_r^2)\beta_a^r\beta_2^a\beta_3^r \\
& + (\mu_{rr} - \mu_r^2)(\beta_{aa}^r - 2\beta_a^r\mu_a)\beta_1^a\beta_2^a\beta_3^r + \mu_1\mu_2\mu_3.
\end{aligned}$$

In terms of central moments,

$$\mu'_{123} = \mu'_{rr}\beta^r_{aa}\beta^a_1\beta^a_2\beta^r_3.$$

These methods allow one to generalize previous work of Smith and Zwiernik [35] on *binary tree cumulants* to apply to trees where nodes can take any number of states. □

3.9 Conclusion

In general, the large polynomials involved in parameterizing discrete models are more easily and compactly expressed using the method of automatic tensor contraction developed here, which we hope will facilitate further investigation into the algebraic structure of these models.

As well, using the new parameterization in Theorem 3.8.3 of discrete DAG models in terms of regression coefficients, the symbolic expression length of the parameterizations are smaller, which facilitate their study by methods such as Gb̈ner bases.

Finally, with this new parametrization we are ready to generalize the work of Smith and Zwiernik [35] on *tree cumulants* for application to trees with more states at each node, as tree cumulants are defined in terms of the moments of the observed distribution as defined here. Indeed, in future work we will show a derivation of the tree cumulants formula directly from Theorem 3.8.3 by specializing to the case of binary trees on binary variables. I have been joined by S. Lin, P. Zwiernik, and L. Weihs on this project, and preliminary computational and experiments show that the new parametrization method, which generalizes tree cumulants, is more efficient than using conditional probabilities, as we expect.

Bibliography

[1] Andrew J. Critch. *Binary hidden Markov models and varieties*. arXiv:1206.0500, to appear in *Journal of Algebraic Statistics*, Vol . 4, No. 1, 2013, Special Volume dedicated to Algebraic Statistics in the Alleghenies, II. 2012.

[2] J. Baker. "The DRAGON system – An overview". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 23.1 (Feb. 1975), pp. 24–29.

[3] D. J. Bates et al. "Numerical decomposition of the rank-deficiency set of a matrix of multivariate polynomials". In: *Approximate commutative agebra*. Ed. by L. Robbiano and J. Abbott. Texts and Monographs in Symbolic Computation. Spinger-Verlag, 2010, pp. 55–77.

[4] L. Baum and T. Petrie. "Statistical inference for probabilistic functions of finite state Markov chains." English. In: *Ann. Math. Stat.* 37 (1966), pp. 1554–1563.

[5] T. Baumgratz et al. "Scalable reconstruction of density matrices". In: *arXiv preprint arXiv:1207.0358* (2012).

[6] D. J. Bates et al. *Bertini: Software for Numerical Algebraic Geometry*. Available at http://www.nd.edu/∼sommese/bertini.

[7] J. Biamonte, V. Bergholm, and M. Lanzagorta. *Invariant Theory for Matrix Product States*. arXiv:1209.0631. 2012.

[8] N. Bray and J. Morton. "Equations defining hidden Markov models". In: *Algebraic Statistics for Computational Biology*. Cambridge Univerisy Press, 2005. Chap. 11.

[9] M. Casanellas and J. Fernandez-Sanchez. "Performance of a new invariants method on homogeneous and non-homogeneous quartet trees". In: *Molecular Biology and Evolution* 24.1 (2006), pp. 288–293.

[10] X. Chen, Z. Gu, and X. Wen. "Classification of gapped symmetric phases in one-dimensional spin systems". In: *Physical Review B* 83.3 (2011), p. 035107.

[11] D. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms*. Second. New York: Springer-Verlag, 1997, pp. xiv+536. ISBN: 0-387-94680-2.

[12] D. A. Cox, J. B. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms, Third Edition*. Springer New York, 2007.

[13] V. Drensky. "Computing with matrix invariants". In: *Math. Balkanica* 21.1-2 (2007), pp. 141–172.

[14] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*. English. Oberwolfach Seminars 39, 2009.

[15] N. Eriksson. "Using invariants for phylogenetic tree construction". In: *Emerging Applications of Algebraic Geometry*. I.M.A. Volumes in Mathematics and its Applications, 2008.

[16] L. Garcia, M. Stillman, and B. Sturmfels. "Algebraic geometry of Bayesian networks". In: *Journal of Symbolic Computation* 39.3 (2005), pp. 331–355.

[17] M. B. Hastings. "Entropy and Entanglement in Quantum Ground States". In: *Phys. Rev. B* 76 (2007), pp. 35–114.

[18] R. Hübener, A. Mari, and J. Eisert. "Wick's theorem for matrix product states". In: *arXiv preprint arXiv:1207.6537* (2012).

[19] A. Krogh, I. S. Mian, and D. Haussler. "A Hidden Markov Model that finds genes in E. coli DNA". In: *Nucleic Acids Research* (1994), pp. 4768–4778.

[20] J. M. Landsberg. *Tensors: Geometry and Applications*. American Mathematical Society Graduate Studies in Mathematics, 2012.

[21] U. Leron. "Trace identities and polynomial identities of $n \times n$ matrices". In: *J. Algebra* 42 (1976), pp. 369–377.

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

[23] N. Meshkat, M. Eisenberg, and J. J. DiStefano. "An algorithm for finding globally identifiable parameter combinations of nonlinear ODE models using Gröbner bases". In: *Mathematical Biosciences* 222.2 (2009), pp. 61–72.

[24] C. Moreau. "Sur les permutations circulaires distinctes". In: *Nouvelles annales de mathématiques, journal des candidats aux écoles polytechnique et normale* 11 (1972), pp. 309–314.

[25] J. Morton. *Tensor Networks in Algebraic Geometry and Statistics*. Lecture at Networking Tensor Networks, Centro de Ciencias de Benasque Pedro Pascual, Benasque, Spain. May 2012.

[26] N. Bray and J. Morton. "Equations defining hidden Markov models". In: *Algebraic Statistics for Computational Bio*. Ed. by L. Pachter and B. Sturmfels. Camb. Univ. Press, 2005. Chap. 11, pp. 237–249.

[27] L. Pachter and B. Sturmfels. "Tropical geometry of statistical models". In: *PNAS* 101.46 (2004), pp. 16132–16137.

[28] L. Pachter and B. Sturmfels. *Algebraic Statistics for Computational Biology.* Cambridge Univerisy Press, 2005.

[29] D. Perez-Garcia et al. "Matrix product state representations". In: *Quantum Information & Computation* 7.5 (2007), pp. 401–430.

[30] G. Pistone, E. Riccomagno, and H. P. Wynn. *Computational commutative algebra in discrete statistics.* Chapman and Hall / CRC, 2001.

[31] C. Procesi. "The invariant theory of $n \times n$ matrices". In: *Adv. Math.* 19.1 (1976), pp. 306–381.

[32] A. Schönhuth. *Generic identification of binary-valued hidden Markov processes.* arXiv:1101.3712. 2011.

[33] K. Sibirskii. "Algebraic invariants for a set of matrices". In: *Siberian Mathematical Journal* 9.1 (1968), pp. 115–124. ISSN: 0037-4466.

[34] K. Sibirskii. "Algebraic invariants for a set of matrices". In: *Siberian Mathematical Journal* 9.1 (1968), pp. 115–124. ISSN: 0037-4466.

[35] J. Q. Smith and P. Zwiernik. "Binary Cumulant Varieties". In: *Bernoulli* 18.1 (2013), pp. 290–321.

[36] R. L. Stratonovich. "Conditional Markov Processes". In: *Theory of Probability and its Applications* 5 (1960), pp. 156–178.

[37] B. Sturmfels and P. Zwiernik. "Binary Cumulant Varieties". In: *Annals of Combinatorics* 17.1 (2013), pp. 229–250.

[38] S. Sullivant, L. D. Garcia-Puente, and S. Spielvogel. *Identifying Causal Effects with Computer Algebra.* Proceedings of the 26th Conference of Uncertainty in Artificial Intelligence. 2010.

[39] F. Verstraete and J. Cirac. "Matrix product states represent ground states faithfully". In: *Physical Review B* 73.9 (2006), p. 094423.

[40] F. Verstraete et al. "Renormalization-group transformations on quantum states". In: *Physical review letters* 94.14 (2005), p. 140601.

[41] S. Watanabe. *Algebraic Geometry and Statistical Learning Theory (Cambridge Monographs on Applied and Computational Mathematics).* Cambridge University Press, 2009.

[42] B.-J. Yoon. "Hidden Markov Models and their Applications in Biological Sequence Analysis". In: *Current Genomics* 10.6 (2009), pp. 402–415.