JOURNAL OF ALGEBRAIC STATISTICS
Vol. 5, No. 1, 2014, 72-99
ISSN 1309-3452 – www.jalgstat.com

Generic Identification of Binary-Valued Hidden Markov Processes

Alexander Schönhuth

Centrum Wiskunde & Informatica, Amsterdam, Netherlands

Abstract. The generic identification problem is to decide whether a stochastic process (X_t) is a hidden Markov process and if yes to infer its parameters for all but a subset of parametrizations that form a lower-dimensional subvariety in parameter space. Partial answers so far available depend on extra assumptions on the processes, which are usually centered around stationarity. Here we present a general solution for binary-valued hidden Markov processes. Our approach is rooted in algebraic statistics hence it is geometric in nature. We find that the algebraic varieties associated with the probability distributions of binary-valued hidden Markov processes are zero sets of determinantal equations which draws a connection to well-studied objects from algebra. As a consequence, our solution allows for algorithmic implementation based on elementary (linear) algebraic routines.

2000 Mathematics Subject Classifications: 62M05, 62M99, 14Q99, 68W30

Key Words and Phrases: Algebraic Statistics, Hidden Markov Processes, Generic Identification

1. Introduction

Hidden Markov processes (HMPs) have gained widespread interest in statistics, predominantly due to their striking successes in applications. Central theoretical concerns have revolved around the fundamental problems of *identifiability* and *complete identification*. Here and in the following, stochastic processes (X_t) take values in a finite set (*alphabet*) Σ where *binary-valued* refers to the case $|\Sigma| = 2$.

Problem 1.1 (Complete Identification)**.** Decide whether a stochastic process (X_t) is a hidden Markov process. If it is, infer its parameters.

The problem was already raised in the late 50s. A representative list of references is [6, 26, 13, 14, 15, 24, 35, 20]. See also the more recent contributions [3, 2, 39, 23] and the (near-exhaustive) list of references in [19]. See also [3, 4] for HMM parameter estimation from data and [10] for a textbook on related practical issues. In terms of practical arguments one can argue that it is reasonable to solve Problem 1.1 for all but a

Email address: as@cwi.nl (A. Schönhuth)

null set of parametrizations which also explains that the most recent contributions [39, 23] provide *generic* solutions. That is, solutions apply for all, but a subset of parametrizations which form a lower-dimensional subvariety in parameter space.

The above-mentioned treatments all raise extra assumptions on the processes, usually centered around stationarity. The only exception is Heller who provided a polyhedral cone-based characterization of arbitrary, also non-stationary HMPs [29]. This, however, was exposed as a reformulation rather than a solution [2] in the sense that it does not give rise to an algorithmic solution of Problem 1.1. To date, Problem 1.1 has still not yet been fully resolved.

The fact that one can assign every probability distribution $\mathbf{P} : \Sigma^n \to [0, 1]$ over finite-length strings to a HMP on $|\Sigma|^n$ states (which is a well-known exercise, the hidden states of the HMP form a de Bruijn graph over Σ^n, together with the obvious transition probabilities), introduces further complications when aiming at algorithmic solutions. We therefore turn our attention to the following finite reformulation of Problem 1.1.

Problem 1.2 (Finite Identification)**.** Let $\mathbf{P} : \Sigma^n \to [0, 1]$ be a probability distribution over strings of finite length n. Decide whether $\mathbf{P}$ is due to a HMP on d hidden states. If it is, infer its parameters.

In the course of this paper, we provide a generic solution of Problem 1.2 for binary-valued alphabets in the case of $d \leq \frac{n+1}{2}$. Our solution is rooted in algebraic statistics where we draw in particular from the concept of an algebraic statistical model, as described in [33, 17]. See for example [25, 38] for discussions on Bayesian networks, which, as latent variable models, are related to hidden Markov models. Since HMPs are uniquely determined by their distributions over strings of length $2d - 1$ [34], a solution of Problem 1.2 also gives rise to a solution of the original Problem 1.1:

1. For each $n \in \mathbb{N}$ determine $d(n)$ as the minimal number of hidden states such that the answer in the 'Decision' part of Problem 1.2 is 'Yes'. In case that there is no $d \leq \frac{n+1}{2}$ set $d(n) := \infty$.
2. If $d(n), n \in \mathbb{N}$ converges, output 'Yes' and infer corresponding parameters. If not, output 'No'.

Since one can extend any probability distribution $\mathbf{P} : \Sigma^n \to [0, 1]$ that stems from a HMP, to a full, non-HMP stochastic process (X_t) taking values in Σ (that is a density $f : \Sigma^{\mathbb{N}} \to \mathbb{R}$), an infinite-runtime solution of the kind from above is all one can expect.

We denote the set of parameters of HMPs with d hidden states by $\mathcal{H}_{d,+}$. By (3.6) below, $\mathcal{H}_{d,+}$ is a full-dimensional subset of the positive orthant of real affine space $\mathbb{R}^{d^2+d-1}$. In the form of a theorem, our solution to Problem 1.2 reads as follows.

Theorem 1.3. *Let* $|\Sigma| = 2$*,* $d \leq \frac{n+1}{2}$ *and* $\mathbf{P} : \Sigma^n \to [0, 1]$ *be a probability distribution. There is a an algebraic variety* $\mathcal{N}_d \subset \mathbb{R}^{d^2+d-1}$ *such that* $\dim \mathcal{N}_d < d^2 + d - 1$ *and an* *algorithmic routine A which, when given $\mathbf{P}$ as input, outputs*

$$A(\mathbf{P}) = \begin{cases} \textit{'HMP on d hidden states'} & \mathbf{P} \in \mathbf{f}_{n,d}((\mathcal{H}_{d,+} \setminus \mathcal{N}_d) \\ \textit{'Cannot decide'} & \mathbf{P} \in \mathbf{f}_{n,d}(\mathcal{N}_d) \\ \textit{'No HMP on d hidden states'} & \textit{otherwise} \end{cases} .$$

In the first case, A also outputs the parametrization, which is unique up to permutation of hidden states.

In the course of proving Theorem 1.3, we provide an ideal-theoretic characterization of the varieties associated with finitary processes with arbitrary output alphabets. Based on dimension arguments, we point out that the varieties of finitary and hidden Markov processes coincide for binary alphabets. Relationships between finitary processes and HMPs have been noted already in seminal work on identification of HMPs (e.g. [6, 26, 13, 14, 29]). Here we review them from the vantage point of algebraic statistics. We summarize the corresponding results into the ideal-theoretic Theorem 6.10, which, in turn, is based on the set-theoretic Lemma 6.12. Note that the ideals we encounter are determinantal in nature; corresponding relationships for latent variable models have also been noted in [7, 12, 40].

Short Summary of Contributions.

- We provide an algebraic statistical treatment of Problem 1.1, based on treating the algebraically more convenient Problem 1.2. Thereby, we lay the foundations for a treatment that does not require extra assumptions on the processes, such as stationarity.
- We present a generic solution of Problem 1.2, and hence of Problem 1.1. The solution is *algorithmic in nature* (unlike the only earlier solution from [29]).
- We provide an ideal-theoretic characterization of the varieties associated with the probability distributions of finitary processes and binary-valued HMPs.

All of this is novel, to the best of our knowledge.

Organization of Chapters. In section 2, we give the basic definition of an algebraic statistical model and also the definition of an algebraic process model, which serves the general purpose to treat stochastic processes in algebraic statistical settings. In section 3, we give formal definitions of finitary and hidden Markov processes. In section 4, we give the definitions of their algebraic statistical counterparts, the finitary and the hidden Markov process model. Along with these definitions, we provide a brief list of fundamental relationships. In section 5 we compute the dimensions of the algebraic varieties associated with finitary and hidden Markov process models. The crucial observation is that the varieties of both models coincide for binary-valued output alphabets. In section 6 we provide a Hankel-matrix-based characterization of finitary models hence also of binary-valued hidden Markov models. The ideal-theoretic formulation of this is Theorem 6.10. In section 7 we present the algorithm on which Theorem 1.3 from above is based.

Major Notation. We denote by $\Sigma^* := \cup_{t\geq 0}\Sigma^t$ the set of all strings over the alphabet Σ where $\Sigma^0 = \{\epsilon\}$ with ϵ the empty string. We write $a, b \in \Sigma$ for single letters, $v, w \in \Sigma^*$ for strings and vw, va etc. for concatenation of letters and strings. Throughout this paper, if $v = a_1...a_n \in \Sigma^n$

$$p_X(v) := \mathbf{P}(\{X_1 = a_1, ..., X_n = a_n\}) \tag{1.1}$$

refers to the probability that the stochastic process (X_t) generates the string $v \in \Sigma^n$ (for technical convenience we let stochastic processes start at $t = 1$). We simply write $p = p_X$ if this cannot lead to confusion. We write $'$ for matrix transposition throughout. None of our algebraic arguments exceed an elementary level, see [11] for an appropriate textbook.

2. Algebraic Statistical Models

Definition 2.1. Following [33], an *algebraic statistical model* with m parameters for strings of length n over an alphabet Σ is a map

$$\begin{array}{rccc} \mathbf{f}: & \mathbb{C}^m & \longrightarrow & \mathbb{C}^{|\Sigma|^n} \\ & \mathbf{z} = (z_1, ..., z_m) & \mapsto & \mathbf{f}(\mathbf{z}) = (f_v(z_1, ..., z_m))_{v\in\Sigma^n} \end{array}$$

where $f_v \in \mathbb{C}[Z_1, ..., Z_m], v \in \Sigma^n$ are polynomials in the indeterminates $Z_1, ..., Z_m$ and there is a parameter set $\mathcal{S} \subset \mathbb{C}^m$ (usually $\mathcal{S} \subset \mathbb{R}^m$) such that for $\mathbf{z} \in \mathcal{S}$

$$\begin{array}{rccc} p_{\mathbf{z}}: & \Sigma^n & \longrightarrow & [0,1] \\ & v & \mapsto & f_v(\mathbf{z}) \end{array} \tag{2.1}$$

is a probability distribution and such that $\mathbb{C}^m$ is the natural extension of the parameter set $\mathcal{S}$ to a complex affine space.

We recall that varieties $V \subset \mathbb{C}^n$ correspond to radical ideals $I \subset \mathbb{C}[X_1, ..., X_n]$ insofar as V is the set of zeros of all polynomials in I [11]. We also recall that an ideal I is prime iff $xy \in I$ implies $x \in I$ or $y \in I$ and that, in terms of the above-mentioned correspondence, prime ideals have irreducible varieties as counterparts. $\mathbf{f}(\mathbb{C}^m)$, as the image of a complex-valued polynomial map is a Boolean combination of varieties (e.g. [33, Th. 3.14]). In particular, its topological closure $V_{\mathbf{f}} = \overline{\mathbf{f}(\mathbb{C}^m)}$ is an irreducible algebraic variety in $\mathbb{C}^{|\Sigma|^n}$ which corresponds to the prime ideal $I_{\mathbf{f}} \subset \mathbb{C}[p_v \mid v \in \Sigma^n]$. We write p_v for indeterminates to stress that they are associated with probability distributions over strings $v \in \Sigma^n$. We will write $\mathbf{P}$ or $(p(v))_{v\in\Sigma^n}$ for the points in complex affine space $\mathbb{C}^{\Sigma^n}$. Polynomials $g \in I_{\mathbf{f}}$ are referred to as *(model) invariants* and the goal of an algebraic statistical treatment is usually to characterize or even explicitly list these invariants. See [33, 16, 17] for related textbooks

2.1. Algebraic Stochastic Process Models

When dealing with stochastic processes (X_t), the auxiliary, helpful observation is that

$$p_X(a_1...a_m) = \sum_{b_1...b_{n-m}\in\Sigma^{n-m}} p_X(a_1...a_m b_1...b_{n-m}). \tag{2.2}$$

As a consequence, one can make use of (virtual) indeterminates p_u for strings u of length m shorter than n when dealing with $\mathbb{C}[p_v, v \in \Sigma^n]$:

$$p_u = \sum_{w \in \Sigma^{n-m}} p_{uw} \tag{2.3}$$

reveals p_u as polynomials in the $p_v, v \in \Sigma^n$. This means in particular that there is no elimination necessary, which is crucial for this work.

We emphasize these facts with a definition.

Definition 2.2 (Algebraic Stochastic Process Model)**.** A family of algebraic statistical models

$$(\mathbf{f}_n : \mathbb{C}^d \longrightarrow \mathbb{C}^{\Sigma^n})_{n \in \mathbb{N}} \tag{2.4}$$

is called an *algebraic (stochastic) process model* if for all $1 \leq m \leq n$ and $u \in \Sigma^m$:

$$\mathbf{f}_m(z)_u = \sum_{w \in \Sigma^{n-m}} \mathbf{f}_n(z)_{uw}. \tag{2.5}$$

2.2. Note on Stationarity

A process (X_t) which takes values in Σ is stationary iff for all $v \in \Sigma^*$

$$\sum_{a \in \Sigma} p_X(va) = \sum_{a \in \Sigma} p_X(av) \tag{2.6}$$

which implies (the more common) $p(v) = \sum_{w \in \Sigma^n} p(wv)$ for all n, v, see (2.2). Let $\mathcal{X}$ be a class of parameterized processes associated with the process model $(\mathbf{f}_{\mathcal{X},n})_{n \in \mathbb{N}}$. Let

$$V_{\mathbf{f}_{\mathcal{X},\mathbf{n}}} = V(I_{\mathbf{f}_{\mathcal{X},n}}) \tag{2.7}$$

be the variety associated with the string length n probability distributions. Let $\langle f_j, j \in J \rangle$ be, as usual, the ideal generated by polynomials $f_j, j \in J$ and $+$ denote addition of ideals. The stationary distributions in $V_{\mathbf{f}_{\mathcal{X},\mathbf{n}}}$ then give rise to the subvariety

$$V(I_{\mathbf{f}_{\mathcal{X},n}} + \langle \sum_{a \in \Sigma} p_{va} - \sum_{a \in \Sigma} p_{av}, v \in \Sigma^{n-1} \rangle). \tag{2.8}$$

This, unless the processes $\mathcal{X}$ are stationary by definition establishes that stationary processes form a lower-dimensional subvariety in $V_{\mathbf{f}_{\mathcal{X},n}}$.

The extent to which earlier work depends on stationarity often remains unclear: for [39], for example, this is difficult to determine, whereas [23] base their approach on Kullback-Leibler divergence computations, which is definitely only possible in case of stationarity. As above-mentioned, [29] is the only contribution that clearly does not depend on stationarity.

Stationarity has geometric implications: by (2.8), stationary HMPs only form a null set among all HMPs. Stationarity also has technical advantages. For example, it introduces certain symmetries among row and column conditions in the Hankel matrices, which is discussed in section 6, see Remark 6.8.

In practical applications, it is very often essential to assume that processes are not stationary. This becomes evident in particular in application domains where HMPs or their close derivatives have established "gold standards", for example speech recognition [36], protein classification (through profile HMMs) [18], gene finding [9] and gene expression time-course analysis [27]. Therefore a general treatment of HMP identification is certainly desirable.

3. Processes

3.1. Finitary Processes

Finitary processes emerged in the above-mentioned early work on HMP identification [6, 26, 13, 14, 15, 29] and have remained a core concept also in recent work on identifiability [39, 23]. Finitary processes were later also referred to as *linearly dependent* [30], *observable operator models* [31] or as *finite-dimensional* [21, 37]. In their possibly most prevalent application they served to determine equivalence of hidden Markov processes (HMPs) in 1992 [30] whose exponential runtime algorithm was later improved to polynomial runtime [22].

Definition 3.1 (Finitary Process)**.** A stochastic process (X_t) is said to be *finitary* iff there are matrices $T_a \in \mathbb{R}^{d\times d}$ for all $a \in \Sigma$ with $(\sum_{a\in\Sigma} T_a)\mathbf{1} = \mathbf{1}$ (that is $(\sum_{a\in\Sigma} T_a)$ has unit row sums) and a vector $\pi \in \mathbb{R}^d$ whose entries sum up to one ($\pi'\mathbf{1} = 1$) such that

$$\mathbf{P}(\{X_1 = a_1, ..., X_n = a_n\}) = \pi' T_{a_1} \cdot \ldots \cdot T_{a_n}\mathbf{1} \tag{3.1}$$

where $\mathbf{1} = (1, ..., 1)' \in \mathbb{R}^d$ is the vector of all ones. The parametrization $((T_a)_{a\in\Sigma}, x)$ is referred to as *d-dimensional* in case of $\pi \in \mathbb{R}^d$ and $T_a \in \mathbb{R}^{d\times d}$ for all $a \in \Sigma$.

It is an immediate observation that a finitary process which admits a d-dimensional parametrization also admits a parametrization of dimension $d + 1$.

Definition 3.2 (Rank of a Finitary Process)**.** The *rank* of a finitary process (X_t) is the minimal dimension of a parametrization that it admits.

We conclude by providing a condition that is necessary for rank d finitary processes. For further reference, we use the notation

$$T_v := T_{a_1} T_{a_2} \ldots T_{a_{n-1}} T_{a_n} \in \mathbb{R}^{d\times d} \tag{3.2}$$

for any $v = a_1 \ldots a_n \in \Sigma^n$.

Proposition 3.3. *Let (X_t) be a finitary process of rank d and let $n \in \mathbb{N}$ be an arbitrary integer. Then it holds that*

$$\text{rk } [p_X(v_i w_j)]_{1 \leq i,j \leq n} \leq d \tag{3.3}$$

for all choices of strings $v_1, ..., v_n, w_1, ..., w_n \in \Sigma^*$.

Proof. Let $((T_a)_{a \in \Sigma}, \pi)$ be a d-dimensional parametrization of (X_t). We observe that

$$p_X(v_i w_j) = \langle \pi' T_{v_i}, T_{w_j} \mathbf{1} \rangle. \tag{3.4}$$

Since $\pi' T_{v_i} \in \mathbb{R}^{1 \times d}, T_{w_j} \mathbf{1} \in \mathbb{R}^{d \times 1}$ the claim becomes obvious. □

3.2. Hidden Markov Processes

Definition 3.4 (Hidden Markov process)**.** A *hidden Markov process (HMP)* (X_t) on d hidden states which takes values in Σ is parametrized by a tuple $\Theta = (M, E, \pi)$ where

1. $M = [m_{s\bar{s}}] \in \mathbb{R}^{d \times d}$ is a non-negative *transition probability matrix* with unit row sums $\sum_{\bar{s}=1}^{n} m_{s\bar{s}} = 1$ (*i.e.* the row vectors of M are probability distributions over the hidden states)

2. $E = [e_{sa}] \in \mathbb{R}^{d \times \Sigma}$ is a non-negative *emission probability matrix* with unit row sums $\sum_{a \in \Sigma} e_{sa} = 1$, (*i.e.* the row vectors of E are probability distributions over Σ)

3. π is an *initial probability distribution* over the hidden states

We write

$$\mathcal{H}_{d,+} := \{(M, E, \pi) \mid \sum_{\bar{s}} m_{s\bar{s}} = \sum_{a \in \Sigma} e_{sa} = \sum_{s} \pi_s = 1\} \subset \mathbb{R}_+^{d^2 + d(|\Sigma|) + d} \tag{3.5}$$

for the set of HMP parametrizations. We refer to $\mathcal{H}_{d,+}$ as the *stochastic* parametrizations.

Remark 3.5. If more convenient and not leading to clashes with other indices, we write i, j, instead of $s, \bar{s}$, for hidden states.

The naming *stochastic* parametrizations is to distinguish them from more relaxed, complex-valued parameter sets whose definition will follow. Note that

$$\dim \mathcal{H}_{d,+} = d(d-1) + d(|\Sigma| - 1) + (d-1) = d^2 + d(|\Sigma| - 1) - 1 \tag{3.6}$$

which means that $\mathcal{H}_{d,+}$ can be considered a full-dimensional subset of $\mathbb{R}^{d^2 + d(|\Sigma|-1)-1}$. A HMP (X_t) on d hidden states as parametrized by (M, E, π) proceeds by initially moving to a state $s \in \{1, ..., d\}$ with probability π_s and emitting the symbol $X_1 = a$ with probability e_{sa}. Then one moves from s to a state $\bar{s}$ with probability $m_{s\bar{s}}$ and emits the symbol $X_2 = b$ with probability $e_{\bar{s}b}$ and so on.

We further observe that M decomposes as $M = \sum_{a \in \Sigma} T_a$ where

$$(T_a)_{s\bar{s}} := e_{sa} \cdot m_{s\bar{s}} \tag{3.7}$$

which reflect the probabilities to emit symbol a from state s and subsequently to move on to state $\bar{s}$. In addition, we use the notation

$$O_a := \text{diag }(e_{1a}, ..., e_{da})T_a = O_a M. \tag{3.8}$$

which yields $T_a = O_a M$. Correspondingly, we write $\Theta = (M, (O_a)_{a\in\Sigma}, \pi)$ for HMP parametrizations. In analogy to finitary process notation, we write

$$T_v := T_{a_1}T_{a_2}\dots T_{a_{n-1}}T_{a_n} = O_{a_1}MO_{a_2}M\dots O_{a_{n-1}}MO_{a_n}M \in \mathbb{R}^{d\times d} \tag{3.9}$$

for any $v = a_1 \dots a_n \in \Sigma^n$. Standard technical computations then reveal that, for $v = a_1...a_n \in \Sigma^n$

$$p(v) = \pi' T_{a_1}...T_{a_n}\mathbf{1} = \pi' T_v \mathbf{1}, \tag{3.10}$$

where $\mathbf{1} = (1, ..., 1)' \in \mathbb{R}^d$ is the vector of all ones.

Remark 3.6. Computation of vectors $\pi' T_v \in \mathbb{R}^{1\times d}$ and $T_v\mathbf{1} \in \mathbb{R}^{d\times 1}$ reflects the well-known Forward and Backward algorithms (*e.g.* [19]) for computation of HMP probabilities. In this respect, entries of these vectors are just the common Forward and Backward variables. That is

$$(\pi' T_v)'_s = \Pr(S_{n+1} = s \mid X_1 = a_1, ..., X_n = a_n) \tag{3.11}$$
$$(T_v\mathbf{1})_s = \Pr(S_t = s \mid X_{t+1} = a_1, ..., X_{t+n} = a_n) \tag{3.12}$$

where (S_t) is the (non-observable) Markov process which takes values in the hidden states $\{1, ..., d\}$.

(3.10) makes it obvious that a HMP on d hidden states is a finitary processes which admits a d-dimensional parametrization. This allows the following definition.

Definition 3.7 (Rank of a Hidden Markov Process)**.** The *rank* of a hidden Markov process (X_t) is its rank as a finitary process.

The definition gives rise to the following trivial proposition.

Proposition 3.8. *A hidden Markov process acting on d hidden states is a finitary process of rank at most d.*

Example 3.9 (HMPs on d hidden states of rank d)**.** Let Σ such that $|\Sigma| \geq 2$. Let $\lambda_1, ..., \lambda_d \in (0, 1)$ be pairwise different. Consider HMP parametrizations (M, E, π) where

$$M = \text{Id } \in \mathbb{R}^{d\times d}, \quad \pi = (\frac{1}{d}, ..., \frac{1}{d}) = \frac{1}{d}\mathbf{1} \in \mathbb{R}^d \tag{3.13}$$

where there is $a \in \Sigma$ such that

$$O_a = \text{diag }(\lambda_1, ..., \lambda_d). \tag{3.14}$$

The $O_b = \text{diag}\,(e_{1b}, ..., e_{db}), b \in \Sigma \setminus \{a\}$ can be chosen arbitrarily. Observe that

$$S(\lambda) := (\mathbf{1}'\text{Id}\ , ..., \mathbf{1}'O_a^{d-1}) = \begin{pmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_d \\ \vdots & \ddots & \vdots \\ \lambda_1^{d-1} & \cdots & \lambda_d^{d-1} \end{pmatrix} = [\lambda_j^{i-1}]_{1\leq i,j\leq d} \in \mathbb{R}^{d\times d} \qquad (3.15)$$

forms a Vandermonde matrix and hence is invertible. Writing $a^i := a...a \in \Sigma^i$, it follows that

$$[p(a^{i-1}a^{j-1})]_{1\leq i,j\leq d} = [\frac{1}{d}\cdot \mathbf{1}'O_a^{i-1}O_a^{j-1}\mathbf{1}]_{1\leq i,j\leq d} = \frac{1}{d}\cdot S(\lambda)S(\lambda)' \in \mathbb{R}^{d\times d} \qquad (3.16)$$

is an invertible matrix. By Proposition 3.3, the hidden Markov process with parametrization $(M, (O_a)_{a\in\Sigma}, \pi)$ has rank d.

4. Models

4.1. Finitary Models

Finitary models $\mathbf{f}_{\mathcal{M}_{d,n}}$ are the algebraic statistical equivalent of *finitary processes* that admit d-dimensional parametrizations.

Definition 4.1. *Finitary models* are polynomial maps

$$\begin{array}{cccc} \mathbf{f}_{\mathcal{M}_d,n}: & \mathcal{M}_d & \longrightarrow & \mathbb{C}^{\Sigma^n} \\ & ((T_a)_{a\in\Sigma}), \pi) & \mapsto & (\pi' T_v \mathbf{1})_{v\in\Sigma^n}. \end{array} \qquad (4.1)$$

where

$$\mathcal{M}_d := \{((T_a)_{a\in\Sigma}), \pi) \in \mathbb{C}^{|\Sigma|d^2+d} \mid \sum_{a\in\Sigma} T_a\mathbf{1} = \mathbf{1}\} \cong \mathbb{C}^{|\Sigma|d^2} \qquad (4.2)$$

We write

$$V_{\mathcal{M}_{d,n}} := \overline{\text{Im }\mathbf{f}_{\mathcal{M}_{d,n}}}$$

for the variety that is associated with $\mathbf{f}_{\mathcal{M}_{d,n}}$ and

$$I_{\mathcal{M}_{d,n}} := I_{\mathbf{f}_{\mathcal{M}_{d,n}}}$$

for the ideal of its invariants. Unlike in the definition of finitary processes, we do not require that $\pi'\mathbf{1} = 1$ which would add the (technically inconvenient) inhomogeneous invariant $\sum_v p_v = 1$ to $I_{\mathcal{M}_{d,n}}$. The relationship $\sum_a T_a\mathbf{1} = \mathbf{1}$ yields that the family $(\mathbf{f}_{\mathcal{M}_{d,n}})_{n\in\mathbb{N}}$ is an algebraic process model.

Proposition 4.2. *The family* $(\mathbf{f}_{\mathcal{M}_{d,n}})_{n\in\mathbb{N}}$ *is an algebraic process model.*

Proof. Let $v \in \Sigma^m$. Writing $M := \sum_a T_a$ we observe

$$\mathbf{f}_{\mathcal{M}_{d,m}}(z)_v = \pi' T_v \mathbf{1} \overset{(4.2)}{=} \pi' T_v M^{n-m} \mathbf{1} = \sum_{u \in \Sigma^{n-m}} \mathbf{f}_{\mathcal{M}_{d,n}}(z)_{vu}. \tag{4.3}$$

□

By the definition of finitary process models one can further register:

Proposition 4.3. *For all $d, n \in \mathbb{N}$ it holds that* $\mathrm{Im}\, \mathbf{f}_{\mathcal{M}_{d,n}} \subset \mathrm{Im}\, \mathbf{f}_{\mathcal{M}_{d+1,n}}$.

Proof. This is because one can extend d-dimensional matrices by zero entries to obtain a $d+1$-dimensional parametrization and reflects that every finitary process with a d-dimensional parametrization also admits a $d+1$-dimensional parametrization. □

4.2. Hidden Markov Models

We obtain an algebraic statistical treatment of HMPs on d hidden states by allowing that parameters in M, E and π are complex. We write

$$\mathcal{H}_d := \{(M, E, \pi) \in \mathbb{C}^{d^2+d|\Sigma|+d} \mid \sum_{j=1}^{n} m_{ij} = 1, \sum_{a \in \Sigma} e_{ia} = 1\} \cong \mathbb{C}^{d^2+d(|\Sigma|-1)} \tag{4.4}$$

for the resulting set of parameters. We still require unit row sums in both M and E, but we do not make any such assumption for π. The unit row sum assumption for E implies that still

$$M = \sum_{a \in \Sigma} T_a \quad \text{where} \quad (T_a)_{s\bar{s}} = e_{sa} m_{s\bar{s}} \tag{4.5}$$

while the unit row sum assumption on M implies that $M\mathbf{1} = \mathbf{1}$ hence (let $v \in \Sigma^m$ and $m < n$)

$$p(v) = \pi' T_v \mathbf{1} = \pi' T_v M^m \mathbf{1} = \sum_{u \in \Sigma^{n-m}} p(vu) \tag{4.6}$$

a relationship which holds for stochastic processes in general. Note that

$$\dim \mathcal{H}_d = d^2 + d(|\Sigma| - 1) \overset{(3.6)}{=} \dim \mathcal{H}_{d,+} + 1. \tag{4.7}$$

The increase in dimension for $\mathcal{H}_d$ follows from not requiring that π is a unit vector—in analogy to finitary models we avoid the non-homogeneous invariant $\sum_v p_v = 1$ for technical convenience.

Definition 4.4. We recall the notation (3.9) and say that

$$\begin{array}{cccc} \mathbf{f}_{\mathcal{H}_{d,n}}: & \mathcal{H}_d & \longrightarrow & \mathbb{C}^{|\Sigma|^n} \\ & (M, (O_a)_{a \in \Sigma}, \pi) & \mapsto & (\pi' T_v \mathbf{1})_{v \in \Sigma^n} \end{array} \tag{4.8}$$

is a *hidden Markov model* for d hidden states and string length n.

The relationship (4.6) yields further:

Proposition 4.5. *The family* $(\mathbf{f}_{\mathcal{H}_d,n})_{n\in\mathbb{N}}$ *of hidden Markov models for* d *hidden states is an algebraic process model.*

We write

$$V_{\mathcal{H}_d,n} := \overline{\mathbf{f}_{\mathcal{H}_d,n}(\mathcal{H}_d)} \tag{4.9}$$

for the algebraic variety that is associated with $\mathbf{f}_{\mathcal{H}_d,n}$ and

$$I_{\mathcal{H}_d,n} := I_{\mathbf{f}_{\mathcal{H}_d,n}}$$

for the ideal of its invariants.

Proposition 4.6. *For all* $d, n \in \mathbb{N}$*:*

(a) $\operatorname{Im} \mathbf{f}_{\mathcal{H}_d,n} \subset \operatorname{Im} \mathbf{f}_{\mathcal{H}_{d+1},n}$

(b) $\operatorname{Im} \mathbf{f}_{\mathcal{H}_d,n} \subset \operatorname{Im} \mathbf{f}_{\mathcal{M}_d,n}$.

(c) $V_{\mathcal{H}_d,n} \subset V_{\mathcal{M}_d,n}$.

(d) $I_{\mathcal{M}_d,n} \subset I_{\mathcal{H}_d,n}$.

While (a) reflects that HMPs on $d+1$ hidden states encompass the HMPs on d hidden states, (d) translates to the fact that each invariant of a finitary model applies for the corresponding hidden Markov model. (d) is a key observation for this work.

Proof. (a) holds because one can extend matrices by zero entries thereby obtaining higher-dimensional parametrizations, (b) is obvious by the definitions of hidden Markov and finitary process models while (c) immediately follows from (b). (c) and (d) finally are equivalent, due to elementary algebraic geometric arguments [11]. □

5. Dimension

5.1. Finitary Models

In this section we compute the dimension of the variety $V_{\mathcal{M}_d,n}$ for $n \geq 2d-1$. The key insight to this computation is the following lemma.

Lemma 5.1. *Let* $n \geq 2d-1$ *and let* $\Theta := ((T_a)_{a\in\Sigma}, x), \tilde{\Theta} := ((\tilde{T}_a)_{a\in\Sigma}, \tilde{x}) \in \mathcal{M}_d$ *be two parameterizations giving rise to finitary processes. Consider the following two statements:*

(i)

$$\mathbf{f}_{\mathcal{M}_d,n}(\Theta) = \mathbf{f}_{\mathcal{M}_d,n}(\tilde{\Theta}) \tag{5.1}$$

(ii) There exists an invertible linear map $S : \mathbb{C}^d \to \mathbb{C}^d$ *such that*

$$S\mathbf{1} = \mathbf{1}, \qquad \tilde{x}' = x'S \qquad \text{and} \qquad \forall a \in \Sigma : \quad \tilde{T}_a = S^{-1}T_aS \tag{5.2}$$

Then (ii) implies (i) and the two statements are equivalent if both $\Theta, \tilde{\Theta}$ give rise to processes of rank d.

Proof. While $(ii) \Rightarrow (i)$ is obvious, $(i) \Rightarrow (ii)$ is a straightforward generalization of statements presented in previous works (e.g. [30, 31]) to complex-valued parameters $\Theta, \tilde{\Theta}$. □

Lemma 5.1 enables application of a well-known theorem [28, Th. 11.12] for computing dimensions of varieties.

Theorem 5.2. *Let $\mathbf{f}_{\mathcal{M}_d,n}$ as in Definition 4.1 such that $n \geq 2d-1$. Then*

$$\dim V_{\mathcal{M}_d,n} = \begin{cases} 1 & |\Sigma| = 1 \\ (|\Sigma|-1)d^2 + d & |\Sigma| \geq 2 \end{cases}. \tag{5.3}$$

Proof. The case $|\Sigma| = 1$ is trivial: Im $\mathbf{f}_{\mathcal{M}_{n,d}} = \mathbb{C}^1$ for all n, d. For the case $|\Sigma| \geq 2$, we proceed by plugging $\mathcal{M}_d, \mathbf{f}_{\mathcal{M}_d,n}$ here into X, π in [28, Th. 11.12]. Therefore we first have to observe that $\overline{\mathbf{f}_{\mathcal{M}_d,n}(\mathcal{M}_d)}$ is a quasi-projective variety, which follows from standard arguments. Applying [28, Th. 11.12] then yields

$$\dim V_{\mathcal{M}_d,n} = \dim \overline{\mathbf{f}_{\mathcal{M}_d,n}(\mathcal{M}_d)} = \dim \mathcal{M}_d - \dim \mathbf{f}^{-1}_{\mathcal{M}_d,n}(\Theta) \tag{5.4}$$

where Θ is chosen such that $\dim \mathbf{f}^{-1}_{\mathcal{M}_d,n}(\Theta)$ is minimal. Since $\dim \mathcal{M}_d = |\Sigma|d^2$, it remains to show that

$$\min_{\Theta \in \mathcal{M}_d} \dim \mathbf{f}^{-1}_{\mathcal{M}_d,n}(\Theta) = d(d-1). \tag{5.5}$$

Therefore, we first observe that, by Example 3.9, stochastic processes of rank d exist. That is, there is Θ such that $\mathbf{f}_{\mathcal{M}_d,n}(\Theta) \notin$ Im $\mathbf{f}_{\mathcal{M}_{d-1},n}$. Lemma 5.1 then states that $\mathbf{f}_{\mathcal{M}_d,n}(\Theta) = \mathbf{f}_{\mathcal{M}_d,n}(\bar{\Theta})$ if and only if there is an invertible linear map $S \in \mathbb{C}^{d \times d}$ with $S\mathbf{1} = \mathbf{1}$ by which to transform Θ into $\bar{\Theta}$ as further described in Lemma 5.1. This yields that the fiber $\mathbf{f}^{-1}_{\mathcal{M}_d,n}(\Theta)$ has dimension equal to that of the space of invertible linear maps S with $S\mathbf{1} = \mathbf{1}$ which is $d(d-1)$.

If $\mathbf{f}_{\mathcal{M}_d,n}(\Theta) \in$ Im $\mathbf{f}_{\mathcal{M}_{d-1},n}(\Theta)$, Lemma 5.1 states that the existence of invertible linear maps S with $S\mathbf{1} = \mathbf{1}$ that transform Θ into another point $\bar{\Theta} \in \mathbf{f}^{-1}_{\mathcal{M}_d,n}(\Theta)$ is only a sufficient condition, which implies $\dim \mathbf{f}^{-1}_{\mathcal{M}_d,n}(\Theta) \geq d(d-1)$. In summary, we obtain (5.5), which concludes the proof. □

5.2. Hidden Markov Models

Let $\mathcal{H}_{d,0} \subset \mathcal{H}_d$ encompass all parametrizations $\Theta = (M, (O_a)_{a \in \Sigma}, \pi)$ such that

- M is not invertible, *or*
- there is no $a \in \Sigma$ such that the eigenvalues of O_a are pairwise different.

Note that $\mathbf{f}^{-1}_{\mathcal{H}_d,n}(\mathcal{M}_{d-1,n})$ are the HMM parametrizations on d hidden states whose rank is less than d. We set

$$\mathcal{N}_d := \mathbf{f}^{-1}_{\mathcal{H}_d,n}(\mathcal{M}_{d-1,n}) \cup \mathcal{H}_{d,0}. \tag{5.6}$$

Lemma 5.3. *$\mathcal{N}_d$ forms a variety of dimension*

$$\dim \mathcal{N}_d < \dim \mathcal{H}_d = d^2 + d(|\Sigma| - 1) \tag{5.7}$$

and for $\Theta = (M, E, \pi) \in \mathcal{H}_d \setminus \mathcal{N}_d$ it holds that

$$\text{card } \mathbf{f}^{-1}_{\mathcal{H}_d,n}(\mathbf{f}_{\mathcal{H}_d,n}(\Theta)) = d! < \infty \tag{5.8}$$

The cardinality of the generic fiber in (5.8) reflects that permutation of the d hidden states yields an equivalent HMP.

Proof. Theorems 3.1 and 3.2 in [3] prove this for stationary processes. Our proof consists in observing that the stationarity assumption in [3] is not used. Moreover, it is straightforward to replace real values by complex values. □

Remark 5.4. [1] provide alternative arguments to prove identifiability of stationary HMPs. Although formulated only for stationary HMPs in [1], the arguments can be easily extended to non-stationary HMPs and also to complex values. [1] particularly focus on generic identifiability of HMPs from their distributions over strings of length $n < 2d-1$ for alphabets $|\Sigma| > 2$. As they do not explicitly name the generic subsets, application of results from the earlier [3] yields a more convenient treatment here.

Corollary 5.5. *As real varieties,*

$$\dim (\mathcal{N}_d \cap \mathcal{H}_{d,+}) < \dim \mathcal{H}_{d,+} = \dim \mathcal{H}_d - 1 \tag{5.9}$$

and

$$\text{card } \mathbf{f}^{-1}_{\mathcal{H}_d,n}(\mathbf{f}_{\mathcal{H}_d,n}(\Theta)) = d! \tag{5.10}$$

for $\Theta \in \mathcal{H}_{d,+} \setminus \mathcal{N}_d$.

Proof. The proof is analogous to that for Lemma 5.3. The reduction in dimension by 1 for $\mathcal{H}_{d,+}$ is due to not requiring $\sum_{i=1}^d \pi_i = 1$ for $\Theta \in \mathcal{H}_d$, see (3.6). □

Identification Algorithm: Workflow. We pause for a moment and relate the results obtained so far with the statements of Theorem 1.3. Given a probability distribution $\mathbf{P} : \Sigma^n \to [0, 1]$ as input, the algorithm of Theorem 1.3 will proceed in three steps:

1. Determine whether $\mathbf{P} \in \text{Im } \mathbf{f}_{\mathcal{H}_d,n}$.

2. If yes, determine $\Theta \in \mathcal{H}_d$ such that $\mathbf{f}_{\mathcal{H}_d,n}(\Theta) = \mathbf{P}$.

3. If $\Theta \in \mathcal{H}_d \setminus \mathcal{N}_d$ determine whether Θ is real non-negative.

From this outer perspective, Lemma 5.3 and Corollary 5.5 are key to performing the third step. We will provide the ingredients for the first two steps in the subsequent sections 6 and 7. We create the necessary link to these sections with the main theorem of this section.

Theorem 5.6. *Let $\mathbf{f}_{\mathcal{H}_d,n}$ be as in Definition 4.4 where $n \geq 2d-1$. Then it holds that*

$$\dim V_{\mathcal{H}_d,n} = \begin{cases} 1 & |\Sigma| = 1 \\ d^2 + (|\Sigma|-1)d & |\Sigma| \geq 2 \end{cases}. \tag{5.11}$$

Proof. The proof again is an application of [28, Th.11.12]. Let $\Theta = ((T_a = O_aM)_{a\in\Sigma}, \pi) \in \mathcal{H}_d \setminus \mathcal{N}_d$. (5.8) implies that

$$\dim \mathbf{f}^{-1}_{\mathcal{H}_d,n}(\Theta) = 0. \tag{5.12}$$

Applying [28, Th. 11.12] in the way of the proof for Theorem 5.2 yields

$$\begin{aligned} \dim V_{\mathcal{H}_d,n} &= \dim \overline{\mathrm{Im}\ \mathbf{f}_{\mathcal{H}_d,n}} \\ &= \dim \mathbb{C}^{d^2+(|\Sigma|-1)d} - \dim \mathbf{f}^{-1}_{\mathcal{H}_d,n}(\Theta) \\ &= d^2 + (|\Sigma|-1)d - 0. \end{aligned} \tag{5.13}$$

□

Binary-Valued HMMs In case of a two-letter alphabet Σ we find

$$\dim V_{\mathcal{H}_d,n} = (|\Sigma|-1)d + d^2 = d + d^2 = d + (|\Sigma|-1)d^2 = \dim V_{\mathcal{M}_d,n}.$$

Since $V_{\mathcal{H}_d,n} \subset V_{\mathcal{M}_d,n}$ and both varieties are irreducible, $V_{\mathcal{H}_d,n}$ and $V_{\mathcal{M}_d,n}$ coincide, which is a standard conclusion from algebraic geometry [11, Prop. 10, p. 463]. Therefore, we obtain the following key insight.

Corollary 5.7. *If $|\Sigma| = 2$*

$$V_{\mathcal{H}_d,n} = V_{\mathcal{M}_d,n}. \tag{5.14}$$

□

6. Invariants

Computation of invariants for finitary models is made possible by a *Hankel matrix* based characterization of finitary processes, corollaries of which will also shed light on the relationship $n \geq 2d-1$ in the formulation of Problem 1.2.

6.1. The Hankel Matrix

Definition 6.1. A string function $p : \Sigma^* \to \mathbb{C}$ such that

$$\forall v \in \Sigma^* : \sum_{a \in \Sigma} p(va) = p(v) \tag{6.1}$$

is called a *process function.*

$\sum_a p(va) = p(v)$ implies $\sum_{u \in \Sigma^m} p(vu) = p(v)$ for all $m \in \mathbb{N}$ which parallels the definition of a process model. By standard arguments, string functions $p : \Sigma^* \to \mathbb{C}$ are associated with stochastic processes if and only if

$$\forall v \in \Sigma^* : \sum_{a \in \Sigma} p(va) = p(v), \quad \sum_{a \in \Sigma} p(a) = 1 \quad \text{and} \quad p(\Sigma^*) \subset [0,1]. \tag{6.2}$$

Omitting $\sum_a p(a) = 1$, $p(\Sigma^*) \subset [0,1]$ in the definition of process function is for compatibility with algebraic process models, see Def. 2.2.

Definition 6.2. Let $p : \Sigma^* \to \mathbb{C}$ be a string function.

- $$\mathcal{P}_p := [p(vw)_{v,w \in \Sigma^*}] \in \mathbb{C}^{\Sigma^* \times \Sigma^*} \tag{6.3}$$

  is called the *Hankel matrix* of p (also called *prediction matrix* in case of a process function p, see e.g. [37]).

- We define

  $$\text{rk } p := \text{rk } \mathcal{P}_p \tag{6.4}$$

  to be the *rank* of the string function p.

- In case of rk $p < \infty$ the string function p is said to be *finitary.*

Example 6.3. Let $p : \Sigma^* \to \mathbb{C}$ be a string function over the binary alphabet $\Sigma = \{0,1\}$. Using lexicographical order on finite strings, the Hankel matrix is

$$\mathcal{P}_p = \begin{pmatrix} p(\epsilon) & p(0) & p(1) & p(00) & p(01) & p(10) & p(11) & \dots \\ p(0) & p(00) & p(01) & p(000) & p(001) & p(010) & p(011) & \dots \\ p(1) & p(10) & p(11) & p(100) & p(101) & p(110) & p(111) & \dots \\ p(00) & p(000) & p(001) & p(0000) & p(0001) & p(0010) & p(0011) & \dots \\ p(01) & p(010) & p(011) & p(0100) & p(0101) & p(0110) & p(0111) & \dots \\ p(10) & p(100) & p(101) & p(1000) & p(1001) & p(1010) & p(1011) & \dots \\ p(11) & p(110) & p(111) & p(1100) & p(1101) & p(1110) & p(1111) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

See also [23] for examples.

Example 6.4 (Rank 1 Hankel matrices: i.i.d. processes)**.** Let (X_t) be an i.i.d. stochastic process taking values in Σ. That is, there are $\rho_a \in [0,1], a \in \Sigma$ with $\sum_a \rho_a = 1$ such that

$$p_X(a_1...a_n) = \rho_{a_1} \cdot ... \cdot \rho_{a_n} \tag{6.5}$$

for all $a_1...a_n \in \Sigma^*$. We observe that rk $\mathcal{P}_{p_X} = 1$ in that case. In fact, rk $\mathcal{P}_{p_X} = 1$ is a characterization of i.i.d. processes.

Example 6.5. Revisiting Example 3.9 (there Σ was $\{a,b\}$) yields that the finite submatrix

$$\begin{pmatrix} p(\epsilon) & p(a) & \cdots & p(a^{d-1}) \\ p(a) & p(aa) & \cdots & p(aa^{d-1} = a^d) \\ \vdots & \vdots & \ddots & \vdots \\ p(a^{d-1}) & p(a^{d-1}a = a^d) & \cdots & p(a^{d-1}a^{d-1} = a^{2d-2}) \end{pmatrix} \in [0,1]^{d \times d} \tag{6.6}$$

of $\mathcal{P}_{p_X}$, as an invertible matrix, has rank d.

For finitary processes one may ask if their rank as process (see Definition 3.2) agrees with their rank as string function.

Generalizing [21] one can show that this is the case and therefore $\mathcal{M}_d$ consists precisely of parameterizations with process functions of rank $\leq d$.

Theorem 6.6. *Let $p : \Sigma^* \to \mathbb{C}$ be a process function. Then the following conditions are equivalent.*

(i) p is finitary of rank at most d.

(ii) There exist vectors $x, y \in \mathbb{C}^d$ as well as matrices $T_a \in \mathbb{C}^{d\times d}$ for all $a \in \Sigma$ such that

$$\forall a_1...a_n \in \Sigma^* : \quad p(a_1...a_n) = x'T_{a_1}...T_{a_n}y \quad \textit{and} \quad (\sum_{a\in\Sigma} T_a)y = y. \tag{6.7}$$

(iii) There exists a vector $x \in \mathbb{C}^d$ as well as matrices $T_a \in \mathbb{C}^{d\times d}$ for all $a \in \Sigma$ such that

$$\forall a_1...a_n \in \Sigma^* : \quad p(a_1...a_n) = x'T_{a_1}...T_{a_n}\mathbf{1} \quad \textit{and} \quad (\sum_{a\in\Sigma} T_a)\mathbf{1} = \mathbf{1}. \tag{6.8}$$

where $\mathbf{1} = (1,...,1)' \in \mathbb{C}^d$ is the vector of all ones.

Proof. $(ii) \Leftrightarrow (iii)$ (where (iii) trivially implies (ii)) follows from the observation that, given an invertible linear map $S : \mathbb{C}^d \to \mathbb{C}^d$ such that $S\mathbf{1} = y$ yields

$$x'T_{a_1}...T_{a_n}y = x'SS^{-1}T_{a_n}SS^{-1}...SS^{-1}T_{a_1}SS^{-1}y = \tilde{x}'\tilde{T}_{a_n}...\tilde{T}_{a_1}\mathbf{1} \tag{6.9}$$

where $\tilde{T}_{a_i} = S^{-1}T_{a_i}S, \tilde{x} = S'x$. $(iii) \Rightarrow (i)$ is because the arguments from [31, 37] work for arbitrary fields. $(i) \Rightarrow (ii)$ follows because the arguments from [31, 37] do not require $\sum_{v\in\Sigma^n} p(v) = 1$, which is missing here, for a proof. □

Finite Algebraic Relationships In the following, we write

$$\mathcal{P}_{p,m,n} := [p(vw)]_{|v|\le m, |w|\le n} \in \mathbb{C}^{\frac{(|\Sigma|^{m+1}-1)}{|\Sigma|-1}\times\frac{(|\Sigma|^{n+1}-1)}{|\Sigma|-1}}. \tag{6.10}$$

for the upper left submatrices of $\mathcal{P}$ which refer to prefixes and suffixes of length at most m and n. Well-known arguments (e.g. [37, Lemma 2.4]) show that

$$\text{rk } \mathcal{P}_p = \text{rk } \mathcal{P}_{p,d-1,d-1} \tag{6.11}$$

for a process function p of rank $\le d$. It follows that a process function of rank $\le d$ is uniquely determined by the values

$$p(v), \quad |v| = 2d-1 \tag{6.12}$$

which applies for d-state HMPs. Combining this with Lemma 5.3 yields that HMPs are generically identifiable from their string-length $2d-1$ probabilities.

Remark 6.7. [1] demonstrate that for $|\Sigma| > 2$, HMPs are generically identifiable already from distributions over strings of length smaller than $2d-1$. However, hidden Markov processes are no longer uniquely determined by their distributions on strings of length smaller than $2d-1$. We conjecture that we could, with some extra work, also employ [1] for our arguments. However, the bounds provided by [1] on string length agree with the ones in use here for $|\Sigma| = 2$ in any case. To date, for binary-valued alphabets, $2d-1$ is the lowest bound presented in the literature.

Remark 6.8 (Stationarity)**.** Let p represent a stochastic process and let $(\mathcal{P}_p)_v : \Sigma^* \to \mathbb{C}$ be the v-row in $\mathcal{P}_p$ (that is $(\mathcal{P}_p)_v(w) = p(vw)$) resp. $(\mathcal{P}_p)^w : \Sigma^* \to \mathbb{C}$ be the w-column of $\mathcal{P}_p$ (that is $(\mathcal{P}_p)^w(v) = p(vw)$). Because p is a process, we have

$$\sum_{a\in\Sigma} (\mathcal{P}_p)^{wa} = (\mathcal{P}_p)^w \tag{6.13}$$

which is a reformulation of the recurring theme (2.3). In case that p is a stationary process, (2.6) translates to

$$\sum_{a\in\Sigma} (\mathcal{P}_p)_{av} = (\mathcal{P}_p)_v. \tag{6.14}$$

This introduces a “symmetry” in $\mathcal{P}_p$ insofar as (6.14) is the condition for the rows that is analogous to the column condition (6.13).

To summarize, Theorem 6.6 states that the finitary processes of rank $\le d$ are precisely the ones whose process functions give rise to Hankel matrices of rank at most d. This translates to the fact that rk $p \le d$ if and only all $(d+1)\times(d+1)$-minors of $\mathcal{P}_p$ are zero, which yields a polynomial characterization of rank $\le d$ processes.

This characterization, however, requires usage of probabilities for strings of arbitrary length. Since we aim at obtaining polynomial equations for probabilities of strings of a fixed length n alone (where $n \ge 2d-1$), we need to collect further insights. Therefore, immediately note that (2.3) reveals probabilities of strings of length $m < n$ as sums of probabilities of length-n strings. We will demonstrate in the next section how to avoid probabilities for strings of length $m > n$.

6.2. Ideals and Varieties

Let

$$R := \mathbb{C}[p_v \mid v \in \Sigma^*] \tag{6.15}$$

be the polynomial ring with (infinitely many) indeterminates p_v. Let further

$$R_n := \mathbb{C}[p_v \mid v \in \Sigma^n] \tag{6.16}$$

(2.3) reveals that R_m can be regarded as a subring of R_n for $m < n$, which is crucial in the following.

We define

$$\mathcal{P}_R := [p_{vw}]_{v,w \in \Sigma^*} \in R^{\Sigma^* \times \Sigma^*} \tag{6.17}$$

as a matrix of Hankel type whose entries are indeterminates p_{vw}, which is analogous to $\mathcal{P}_p$. As for $\mathcal{P}_p$, we also write

$$\mathcal{P}_{R,m,n} := [p_{vw}]_{|v| \leq m, |w| \leq n} \in R^{\frac{(|\Sigma|^{m+1}-1)}{|\Sigma|-1} \times \frac{(|\Sigma|^{n+1}-1)}{|\Sigma|-1}}. \tag{6.18}$$

Let

$$I_{d+1,n} := \langle\ f \mid f \text{ (d+1)-minor of } \mathcal{P}_{R,\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}\ \rangle + \langle\ f \mid f \text{ (d+1)-minor of } \mathcal{P}_{R,\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor}\ \rangle \tag{6.19}$$

$I_{d+1,n}$ is the ideal of all $(d+1)$-minors in either $\mathcal{P}_{R,\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$ or $\mathcal{P}_{R,\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor}$. Let

$$J_{d,n} := \langle\ g \mid\ g \text{ d-minor of } \mathcal{P}_{R,d-1,d-1}\ \rangle \tag{6.20}$$

$J_{d,n}$ is the ideal of all d-minors in $\mathcal{P}_{R,d-1,d-1}$. Due to the comment following (6.16), one can view both $I_{d+1,n}$ and $J_{d,n}$ as ideals of R_n.

Remark 6.9. Elementary insights [8] point out that determinantal ideals are prime if matrix entries represent independent indeterminates. This is not the case here—as just outlined, $p_u = \sum_{w \in \Sigma^{n-m}} p_{uw}$ for $|u| = m < n = |uw|$ reveals p_u, for $|u| = m < n$, as a sum of indeterminates referring to strings of length n. Indeed, computations with Bertini [5] confirm that $I_{3,4}$, for example, is not prime.

Let rad I be the radical of I and $I : J$ the quotient ideal of I with respect to J.

Theorem 6.10. *Let $n \geq 2d-1$. Then*

$$I_{\mathcal{M}_{d,n}} = \text{rad } I_{d+1,n} : J_{d,n}.$$

For $|\Sigma| = 2$, also

$$I_{\mathcal{H}_{d,n}} = \text{rad } I_{d+1,n} : J_{d,n}. \tag{6.21}$$

Theorem 6.10 provides an ideal-theoretic characterization of the variety associated with the finitary model $\mathbf{f}_{\mathcal{M}_{d,n}}$. If $|\Sigma| = 2$, this also applies for the hidden Markov model $\mathbf{f}_{\mathcal{H}_{d,n}}$, due to Corollary 5.7.

Remark 6.11. As pointed out in Remark 6.9, the quotient operation is necessary. However, it remains an open problem whether the radical operation is. Macaulay [32] computations reveal that it is not for $d = 2, n = 3$. Macaulay and Bertini computations also confirm Theorem 6.10 in terms of dimension computations.

The proof of Theorem 6.10 is based on a set-theoretic lemma which makes use of the insights assembled in the earlier chapters. For the following, we recall the notation

$$T_v = T_{a_1}...T_{a_n} \quad \text{for } v = a_1...a_n \tag{6.22}$$

see (3.2).

Lemma 6.12. *Let* $n \geq 2d-1$ *and* $(p(v))_{v\in\Sigma^n} \in \mathbb{C}^{\Sigma^n}$. *The following statements are equivalent:*

(i)

$$(p(v))_{v\in\Sigma^n} \in \text{Im } \mathbf{f}_{\mathcal{M}_d,n} \setminus \text{Im } \mathbf{f}_{\mathcal{M}_{d-1},n} \tag{6.23}$$

(ii)

$$\text{rk } \mathcal{P}_{p,d-1,d-1} = \text{rk } \mathcal{P}_{p,\lfloor\frac{n}{2}\rfloor,\lceil\frac{n}{2}\rceil} = \text{rk } \mathcal{P}_{p,\lceil\frac{n}{2}\rceil,\lfloor\frac{n}{2}\rfloor} = d \tag{6.24}$$

In case of (6.24), *one can choose parameters for* $(p(v))_{v\in\Sigma^n}$ *by determining an invertible submatrix*

$$V = [p(v_iw_j)]_{1\leq i,j\leq d} \in \mathbb{C}^{d\times d} \tag{6.25}$$

from $\mathcal{P}_{p,d-1,d-1}$ *and setting*

$$x' := (p(w_1), ..., p(w_d)) \tag{6.26}$$

$$y := V^{-1}\begin{pmatrix} p(v_1) \\ \vdots \\ p(v_d)\end{pmatrix} \tag{6.27}$$

$$T_a := V^{-1}W_a := V^{-1}[p(v_iaw_j)]_{1\leq i,j\leq d} \tag{6.28}$$

which yields that $p(v) = x'T_vy$ *so that we obtain*

$$p(v) = \pi\tilde{T}_v\mathbf{1} \tag{6.29}$$

by further application of Theorem 6.6. Note that probabilities in W_a *may refer to strings* v_iaw_j *of length up to* $2d-1$. *This explains the necessity of the assumption* $n \geq 2d-1$.

$n \geq 2d-1$ implies that $d-1 < \lceil\frac{n}{2}\rceil$. Writing $A \subsetneq B$ for a submatrix A of a matrix B which is strictly smaller than B shows that

$$\mathcal{P}_{p,d-1,d-1} \subsetneq \mathcal{P}_{p,\lfloor\frac{n}{2}\rfloor,\lceil\frac{n}{2}\rceil} \quad \text{and} \quad \mathcal{P}_{p,d-1,d-1} \subsetneq \mathcal{P}_{p,\lceil\frac{n}{2}\rceil,\lfloor\frac{n}{2}\rfloor} \tag{6.30}$$

Example 6.13. Let $n = 3, d = 2$ and $\Sigma = \{0, 1\}$. Hence $\lceil \frac{n}{2} \rceil = 2$ and $\lfloor \frac{n}{2} \rfloor = 1$ such that we have

$$\mathcal{P}_{p,\lceil \frac{n}{2} \rceil,\lfloor \frac{n}{2} \rfloor} = \mathcal{P}_{p,2,1} = \begin{pmatrix} p(\epsilon) & p(0) & p(1) \\ p(0) & p(00) & p(10) \\ p(1) & p(01) & p(11) \\ p(00) & p(000) & p(100) \\ p(01) & p(001) & p(101) \\ p(10) & p(010) & p(110) \\ p(11) & p(011) & p(111) \end{pmatrix} \in \mathbb{C}^{7\times 3}$$

$$\mathcal{P}_{p,\lfloor \frac{n}{2} \rfloor,\lceil \frac{n}{2} \rceil} = \mathcal{P}_{p,1,2} =$$
$$\begin{pmatrix} p(\epsilon) & p(0) & p(1) & p(00) & p(01) & p(10) & p(11) \\ p(0) & p(00) & p(10) & p(000) & p(010) & p(100) & p(110) \\ p(1) & p(01) & p(11) & p(001) & p(011) & p(101) & p(111) \end{pmatrix} \in \mathbb{C}^{3\times 7}$$

and

$$\mathcal{P}_{p,d-1,d-1} = \mathcal{P}_{p,1,1} = \begin{pmatrix} p(\epsilon) & p(0) & p(1) \\ p(0) & p(00) & p(10) \\ p(1) & p(01) & p(11) \end{pmatrix} \in \mathbb{C}^{3\times 3}.$$

We recall the relationship $p(v) = \sum_{w\in\Sigma^{3-|v|}} p(vw)$ (6.1). For example,

$$\begin{aligned} p(00) &= p(000) + p(001) \\ p(1) &= p(100) + p(101) + p(110) + p(111) \\ p(\epsilon) &= p(000) + p(001) + ... + p(110) + p(111) \end{aligned}$$

which yields expressions in strings of length $n = 3$ only. As $\mathcal{P}_{p,d-1,d-1}$ is a submatrix of both $\mathcal{P}_{p,\lfloor \frac{n}{2} \rfloor,\lceil \frac{n}{2} \rceil}$ and $\mathcal{P}_{p,\lceil \frac{n}{2} \rceil,\lfloor \frac{n}{2} \rfloor}$ we can decompose (6.24) into

$$\begin{aligned} \text{rk } \mathcal{P}_{p,1,1} &\geq 2 && (6.31) \\ \text{rk } \mathcal{P}_{p,1,2} &\leq 2 && (6.32) \\ \text{rk } \mathcal{P}_{p,2,1} &\leq 2. && (6.33) \end{aligned}$$

Proof of Lemma 6.12. (i) ⇒ (ii): Let $(p(v))_{v\in\Sigma^n}$ be in the image of $\mathbf{f}_{\mathcal{M}_d,n}$, but not in the image of $\mathbf{f}_{\mathcal{M}_{d-1},n}$. Combining Theorem 6.6 with (6.11) reveals that

$$d = \text{rk } p \overset{\text{Th.6.6}}{=} \text{rk } \mathcal{P}_p \overset{(6.11)}{=} \text{rk } \mathcal{P}_{p,d-1,d-1} \qquad (6.34)$$

where the second equation is just the definition of the rank of a string function. Since $\text{rk } \mathcal{P}_{p,d-1,d-1} \leq \text{rk } \mathcal{P}_{p,\lfloor \frac{n}{2} \rfloor,\lceil \frac{n}{2} \rceil}, \text{rk } \mathcal{P}_{p,\lceil \frac{n}{2} \rceil,\lfloor \frac{n}{2} \rfloor} \leq \text{rk } \mathcal{P}_p$ (see (6.30)), we obtain the claim.

(ii) ⇒ (i): Let $P := (p(u))_{u\in\Sigma^n} \in \mathbb{C}^{\Sigma^n}$ such that (6.24) applies. By Theorem 6.6, $P \in \mathbf{f}_{\mathcal{M}_{d-1},n}$ would imply $\text{rk } \mathcal{P}_{p,d-1,d-1} \leq d - 1$, a contradiction! In order to show that $P \in \text{Im } \mathbf{f}_{\mathcal{M}_d,n}$ we will demonstrate that determining $V, x, y, (T_a)_{a\in\Sigma}$ according to (6.25),(6.26),(6.27),(6.28) yields

$$\begin{aligned} p(u) &= x'T_uy \quad \text{for all } u \in \Sigma^* && (6.35) \\ (\sum_{a\in\Sigma} T_a)y &= y. && (6.36) \end{aligned}$$

Applying $(ii) \Rightarrow (iii)$ from Theorem 6.6 to x, y and $T_a, a \in \Sigma$ then proves the claim.

The proof concludes by means of the following two elementary sublemmata 6.14, 6.15.

Lemma 6.14. *Let $v = a_1...a_m \in \Sigma^*$ such that $|v| = m \leq \lceil \frac{n}{2} \rceil$. Then*

$$x'T_v = (p(vw_1), ..., p(vw_d)). \tag{6.37}$$

Proof of Lemma 6.14. By induction on $|v|$, we obtain a proof by showing

$$(p(vw_1), ..., p(vw_d))T_a = (p(vaw_1), ..., p(vaw_d)) \tag{6.38}$$

for all $v \in \Sigma^*, a \in \Sigma$ with $|v| < \frac{n}{2}$. Therefore, $|w_j| \leq d-1 < \frac{n}{2}$ implies $|aw_j| \leq \lceil \frac{n}{2} \rceil$. Hence both $|va|, |aw_j| \leq \lceil \frac{n}{2} \rceil$. Furthermore, $|v| < n/2$ implies $|v| \leq \lfloor \frac{n}{2} \rfloor$ and rk $\mathcal{P}_{p,d-1,d-1} = \mathcal{P}_{p,\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$ from (6.24) implies that the v-row $(\mathcal{P}_p)_v$ in $\mathcal{P}_{p,\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$ is contained in the span of the rows $(\mathcal{P}_p)_{v_i}$, by choice of the v_i (6.25). Accordingly, we determine $\alpha_i, i = 1, ..., d$ such that $(\mathcal{P}_p)_v = \sum_{i=1}^d \alpha_i (\mathcal{P}_p)_{v_i}$ which, by definition of $\mathcal{P}_{p,\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$, yields $p(vw) = \sum_{i=1}^d \alpha_i p(v_i w)$ for all $w, |w| \leq \lceil \frac{n}{2} \rceil$. As $|aw_j| \leq \frac{n}{2}$ for all $j = 1, ..., d$, we obtain in particular

$$(p(vaw_1), ..., p(vaw_d)) = \sum_{i=1}^d \alpha_i (p(v_i aw_1), ..., p(v_i aw_d)). \tag{6.39}$$

This is the key insight. We finally compute

$$\begin{aligned} (p(vw_1), ..., p(vw_d))T_a &= \sum_{i=1}^d \alpha_i (p(v_i w_1), ..., p(v_i w_d))T_a \\ &= \sum_{i=1}^d \alpha_i (p(v_i w_1), ..., p(v_i w_d))V^{-1}W_a = \sum_{i=1}^d \alpha_i e_i' W_a \qquad (6.40) \\ &= \sum_{i=1}^d \alpha_i (p(v_i aw_1), ..., p(v_i aw_d)) \overset{(6.39)}{=} (p(vaw_1), ..., p(vaw_d)). \end{aligned}$$

□

Lemma 6.15. *For all $v, w \in \Sigma^*$ such that $|v| \leq \lceil \frac{n}{2} \rceil, |w| \leq \lfloor \frac{n}{2} \rfloor$ (which implies $|vw| \leq n$):*

$$(p(vw_1), ..., p(vw_d))T_w y = p(vw). \tag{6.41}$$

Proof of Lemma 6.15. We do this by induction on $|w|$, starting with $|w| = 0$, that is $w = \epsilon$ and $T_\epsilon = V^{-1}W_\epsilon = Id$. Due to rk $\mathcal{P}_{p,d-1,d-1} =$ rk $\mathcal{P}_{p,\lceil\frac{n}{2}\rceil,\lfloor\frac{n}{2}\rfloor}$, by (6.24), the row $(\mathcal{P}_p)_v$ in $\mathcal{P}_{p,\lceil\frac{n}{2}\rceil,\lfloor\frac{n}{2}\rfloor}$ is contained in the span of the rows $(\mathcal{P}_p)_{v_i}$, by choice of the v_i (6.25). Therefore, it suffices to show the statement for $v = v_i$. Writing V_i for the i-th row of V and e_i for the i-th canonical basis vector, we get

$$(p(v_iw_1), ..., p(v_iw_d))T_\epsilon y \quad = \quad V_iV^{-1}\begin{pmatrix} p(v_1) \\ \vdots \\ p(v_d) \end{pmatrix} \quad = \quad e_i'\begin{pmatrix} p(v_1) \\ \vdots \\ p(v_d) \end{pmatrix} \quad = \quad p(v_i). \tag{6.42}$$

For the step $|w| \to |w| + 1$, let $\tilde{w} = aw$ with $a \in \Sigma$. By arguments which are analogous to those for $|w| = 0$, it suffices to consider $v = v_i$ referring to one of the row space generators $(\mathcal{P}_p)_{v_i}$ (while the induction hypothesis already holds for *all* $v, |v| \leq \lceil\frac{n}{2}\rceil$)

$$\begin{aligned} &(p(v_iw_1), ..., p(v_iw_d))T_{\tilde{w}}y = V_iT_aT_wy = V_iV^{-1}W_aT_wy \\ &\quad = e_i'W_aT_wy = (p(v_iaw_1), ..., p(v_iaw_d))T_wy \overset{(*)}{=} p(v_iaw) = p(v_i\tilde{w}) \end{aligned} \tag{6.43}$$

where $(*)$ is the induction hypothesis with $v = v_ia$, which applies because of $|v_ia| \leq d \leq \lceil\frac{n}{2}\rceil$. □

Proof of Lemma 6.12 cont. Let $u \in \Sigma^*$ such that $|u| \leq n$. Split $u = vw$ into two strings v, w such that $|v| \leq \lceil\frac{n}{2}\rceil, |w| \leq \lfloor\frac{n}{2}\rfloor$. We compute

$$x'T_uy = x'T_vT_wy \overset{L.6.14}{=} (p(vw_1), ..., p(vw_d))T_wy \overset{L.\ 6.15}{=} p(vw) = p(u). \tag{6.44}$$

This yields (6.35). For (6.36) we compute

$$\begin{aligned} &(p(v_iw_1), ..., p(v_iw_d))\sum_{a\in\Sigma}T_ay = \sum_{a\in\Sigma}(p(v_iw_1), ..., p(v_iw_d))T_ay \\ &\overset{(L.6.15)}{=} \sum_{a\in\Sigma}p(v_ia) = p(v_i) \overset{(L.6.15)}{=} (p(v_iw_1), ..., p(v_iw_d))y \end{aligned} \tag{6.45}$$

which yields the claim since span$\{(p(v_iw_1), ..., p(v_iw_d)) \mid i = 1, ..., d\} = \mathbb{C}^d$. □

The step from the set-theoretic Lemma 6.12 to the our ideal-theoretic Theorem 6.10 now follows from standard arguments, as e.g. listed in [11]. In the following, $\overline{A}$ denotes the Zariski closure of a set A, which is the smallest affine algebraic variety which contains the set A, see [11], sec. 4.4, def. 2.

In the following, we use

$$F_d := \text{Im } \mathbf{f}_{\mathcal{M}_d,n}$$

as a simpler notation for the image of $\mathbf{f}_{\mathcal{M}_d,n}$.

Proof of Theorem 6.10: We first compute

$$\begin{aligned} V_{\mathcal{M}_{d,n}} = \overline{F_d} &= \overline{(F_d \setminus F_{d-1}) \,\cup\, (F_{d-1} \setminus F_{d-2}) \,\cup\, ... \,\cup\, (F_1 \setminus F_0) \,\cup\, F_0} \\ &= \overline{F_d \setminus F_{d-1}} \,\cup\, \overline{F_{d-1} \setminus F_{d-2}} \,\cup\, ... \,\cup\, \overline{F_1 \setminus F_0} \,\cup\, \overline{F_0} \end{aligned} \tag{6.46}$$

where the last equation is an obvious consequence of the definition of the Zariski closure: the Zariski closure agrees with the topological closure if the latter one already is a variety. The irreducibility of $V_{\mathcal{M}_{d,n}}$ implies that $V_{\mathcal{M}_{d,n}}$ agrees with one of the components $\overline{F_0}, \overline{F_1 \setminus F_0}, ..., \overline{F_d \setminus F_{d-1}}$. By Theorem 5.2, $\dim V_{\mathcal{M}_{d,n}} = (|\Sigma| - 1)d^2 + d$. Because of $\dim \overline{F_e \setminus F_{e-1}} \leq \overline{F_e} \leq \dim e^2(|\Sigma| - 1) + e$, which also follows from Theorem 5.2, we conclude that

$$V_{\mathcal{M}_{d,n}} = \overline{F_d \setminus F_{d-1}}. \tag{6.47}$$

By (6.30), it follows that (6.24) is equivalent to

$$\text{rk } \mathcal{P}_{p,\lfloor\frac{n}{2}\rfloor,\lceil\frac{n}{2}\rceil},\ \text{rk } \mathcal{P}_{p,\lceil\frac{n}{2}\rceil,\lfloor\frac{n}{2}\rfloor} \leq d \quad \text{and} \quad \text{rk } \mathcal{P}_{p,d-1,d-1} \geq d. \tag{6.48}$$

Application of Lemma 6.12 reveals that

$$F_d \setminus F_{d-1} = A_{d+1,n} \setminus B_d \tag{6.49}$$

where

$$\begin{aligned} A_{d+1,n} \quad &:= \quad \{(p(v))_{v\in\Sigma^n} \mid \det\,(p(u_iv_j))_{1\leq i,j\leq d+1} = 0, \\ &\qquad\ \forall\, 0 \leq |u_i|, |v_j| \leq \lceil\frac{n}{2}\rceil, |u_iv_j| \leq n\} \\ B_d \quad &:= \quad \{(p(v))_{v\in\Sigma^n} \mid \det\,(p(u_iv_j))_{1\leq i,j\leq d} = 0, \\ &\qquad\ \forall\, 0 \leq |u_i|, |v_j| \leq d-1\} \end{aligned}$$

since $A_{d+1,n}$ consists of all p such that all $(d+1)$-minors in $\mathcal{P}_{p,\lfloor\frac{n}{2}\rfloor,\lceil\frac{n}{2}\rceil}$ and $\mathcal{P}_{p,\lceil\frac{n}{2}\rceil,\lfloor\frac{n}{2}\rfloor}$ are zero whereas B_d encompasses all p such that not all d-minors in $\mathcal{P}_{p,d-1,d-1}$ are zero.

As zero sets of determinantal (hence polynomial) equations, both $A_{d+1,n}$ and B_d are varieties, and recalling the definition (6.19),(6.20) of $I_{d+1,n}$ and J_d, we can conclude that these are just the ideals associated with $A_{d+1,n}$ and B_d. By Hilbert's Nullstellensatz (see [11, p. 174, theorem 6]):

$$I(A_{d+1,n}) = \text{rad } I_{d+1,n} \quad \text{and} \quad I(B_d) = \text{rad } J_d. \tag{6.50}$$

The claim of Theorem 6.10 now follows from the interrelationship between quotients of ideals and differences of varieties, as explicitly expressed by plugging rad $I_{d+1,n}$ and J_d into I and J of the second statement of [11, p. 192, th. 7] (the algebraically closed k there becomes $\mathbb{C}$ here). □

7. Algorithm

Let $\Sigma := \{a, b\}$ be a binary-valued alphabet. The following algorithm determines whether a probability distribution $\mathbf{P} : \Sigma^n \to [0, 1]$ is due to a HMP on at most $d^* \leq \frac{n+1}{2}$ hidden states, as supported by

Theorem 7.1. *Algorithm 7.2 below correctly decides and infers a HMP parametrization with at most d hidden states for all but a lower-dimensional subvariety in $\mathcal{H}_{d,+}$.*

That is, Theorem 7.1 establishes a *generic* solution for Problem 1.2.

Algorithm 7.2.

```
IdentifyHMP($\mathbf{P} = (p(v))_{v \in \Sigma^n}$)
 1: $e \leftarrow 1$
 2: while $e \leq d := \lfloor \frac{n+1}{2} \rfloor$ do
 3:   if rk $\mathcal{P}_{p,e-1,e-1}$ = rk $\mathcal{P}_{p,\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$ = rk $\mathcal{P}_{p,\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor} = e$ then
 4:     $T_a, T_b, x \leftarrow$ InferFinitaryParam($\mathbf{P}, e$)
 5:     if det $[T_a + T_b] > 0$ and $T_a[T_a + T_b]^{-1}$ is diagonalizable
        such that all eigenvalues are different then
 6:       $M, O_a, O_b, \pi \leftarrow$ InferHMMParam($T_a, T_b, x$)
 7:       if $(M, O_a, O_b, \pi)$ is stochastic then
 8:         print 'HMP on $e$ hidden states'
 9:         return $M, O_a, O_b, \pi$ as parametrization
10:       else
11:         print 'No HMP on $d$ hidden states'
12:         return
13:       end if
14:     end if
15:   end if
16:   $e \leftarrow e + 1$
17: end while
18: print 'No HMP on $d$ hidden states'
```

InferFinitaryParam$(\mathbf{P}, e)$ is a routine that computes an e-dimensional parametrization $(T_a, T_b, x) \in \mathcal{M}_e$ for a finitary process. It works by computing T_a, T_b and x according to (6.25,6.26,6.27,6.28) and subsequent application of Theorem 6.6 (note that any invertible S with $S^{-1}y = \mathbf{1}$ applies). According to Lemma 6.12 this works if

$$\text{rk } \mathcal{P}_{p,e-1,e-1} = \text{rk } \mathcal{P}_{p,\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil} = \text{rk } \mathcal{P}_{p,\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor} = e$$

which is guaranteed by step 3.

InferHMMParam(T_a, T_b, x) works if $[T_a + T_b]$ is invertible and $T_a[T_a + T_b]^{-1}$ is diagonalizable such that all eigenvalues $\lambda_1, ..., \lambda_e$ are different, by Lemma 5.3. In this case, one chooses (note that this is possible!) $S \in \mathbb{C}^{e \times e}$ such that $S\mathbf{1} = \mathbf{1}$ and

$$S^{-1} T_a [T_a + T_b]^{-1} S = \text{diag } (\lambda_1, ..., \lambda_e).$$

One then computes

$$\begin{aligned} M &= S^{-1}[T_a + T_b]S, \\ O_a, O_b &= T_a M^{-1}, T_b M^{-1} \\ \pi &= S'x. \end{aligned}$$

The proof of Theorem 7.1 is based on the following lemma for which we recall the definition of $\mathcal{N}_d$, see (5.6).

Lemma 7.3. *Algorithm 7.2 can decide incorrectly only if*

$$\mathbf{P} \in \mathbf{f}_{\mathcal{H}_e,n}(\mathcal{N}_e)$$

for some $e = 1, ..., d$ where it may mistakenly output 'No HMP on at most d hidden states'.

Using Lemma 7.3 a proof of Theorem 7.1 is easy:

Proof of Theorem 7.1. By Lemma 5.3, $\mathcal{N}_d$ forms a lower-dimensional variety in $\mathcal{H}_d$ and further, by Corollary 5.5, $\mathcal{N}_d \cap \mathcal{H}_{d,+}$ also forms a lower-dimensional semialgebraic set in $\mathcal{H}_{d,+}$. □

Proof of Lemma 7.3. Let $\Theta \in \mathcal{H}_{d,+}$ and

$$\mathbf{P} = \mathbf{f}_{\mathcal{H}_d,n}(\Theta)$$

such that $\mathbf{P}$ is incorrectly classified as 'No HMP' by Algorithm 7.2. We have to show that

$$\Theta \in \mathcal{N}_e \quad \text{for some } e = 1, ..., d.$$

We recall the fundamental relationship (see Propositions 4.3,4.6)

$$\mathbf{f}_{\mathcal{H}_e,n}(\mathcal{H}_{e,+}) \subset \text{Im } \mathbf{f}_{\mathcal{H}_e,n} \subset \text{Im } \mathbf{f}_{\mathcal{M}_e,n} \tag{7.1}$$

In relation to (7.1), Algorithm 7.2 tests for membership from right to left in the e-th iteration of the while loop, thereby stepwise approving or rejecting that $\mathbf{P} \in \mathbf{f}_{\mathcal{H}_e,n}(\mathcal{H}_{e,+})$ $[\subset \mathbf{f}_{\mathcal{H}_d,n}(\mathcal{H}_{d,+})]$. First, by Lemma 6.12, *step* 3 tests for

$$\mathbf{P} \in (\text{Im } \mathbf{f}_{\mathcal{M}_e,n} \setminus \text{Im } \mathbf{f}_{\mathcal{M}_{e-1},n}). \tag{7.2}$$

Note that the case $\mathbf{P} \in \text{Im } \mathbf{f}_{\mathcal{M}_{e-1},n}$ was excluded in the iteration before. This allows to infer an e-dimensional parametrization for the respective finitary process in *step* 6 (see the description of InferFinitaryParam above). The **if** condition in *step* 7 finally is the critical point; it determines whether

$$\mathbf{P} \in \mathbf{f}_{\mathcal{H}_e,n}(\mathcal{H}_{e,+} \setminus \mathcal{N}_e) \tag{7.3}$$

see the description of InferHMMParam. If not, the algorithm issues the output 'No HMP' which can be mistakenly due to either $\mathbf{P} \in \mathbf{f}_{\mathcal{H}_e,n}(\mathcal{H}_{e,+} \cap \mathcal{N}_e) \subset \mathbf{f}_{\mathcal{H}_e,n}(\mathcal{N}_e)$ or correctly due to either $\mathbf{P} \in \mathbf{f}_{\mathcal{H}_e,n}(\mathcal{N}_e \setminus \mathcal{H}_{e,+})$ or $\mathbf{P} \in \text{Im } \mathbf{f}_{\mathcal{M}_e,n} \setminus \text{Im } \mathbf{f}_{\mathcal{H}_e,n}$.

By Lemma 5.3, the parameters inferred in step 7 are unique, up to permutations of rows and columns. Therefore, steps 8 and 11 decide correctly. □

Acknowledgements

First and foremost, the author would like to thank Bernd Sturmfels who made the author aware of the algebraic challenges in this line of research and has been a great source of inspiration throughout the author's postdoctoral stay at UC Berkeley. Special thanks to Jon Hauenstein for Bertini computations and to Anton Leykin for advice on both Macaulay and Bertini computations. Preliminary versions of that work were presented on a workshop on Algebraic Statistics at the MSRI, Berkeley, December 2008 and the annual AMS meeting in Lexington, Kentucky, March 2010. The author would like to thank the organizers and participants of those meetings for support and stimulating discussions. The author also would like to thank David DesJardins for a private donation that made the postdoctoral stay at UC Berkeley possible.

References

[1] Allman, E.S., Matias, C. and Rhodes, J.A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics* **37** 3099–3132. `http://arxiv.org/abs/0809.5032`

[2] Anderson, B.D.O. (1999). The realization problem for hidden Markov models, *Mathematics of Control, Signal and Systems* **12** 80–120.

[3] Baras, J.S. and Finesso, L. (1992). Consistent estimation of the order of hidden Markov chains. *Lecture Notes in Control and Information Sciences* **184** 26–39.

[4] Baum, L.E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite Markov chains. *Annals of Mathematical Statistics* **37** 1559–1563.

[5] Bates, D.J., Hauenstein, J.D., Sommese, A.J. and Wampler, C.W.. Bertini: Software for Numerical Algebraic Geometry, `http://www.nd.edu/~sommese/bertini/`.

[6] Blackwell, D. and Koopmans, L. (1957). On the identifiability problem for functions of finite markov chains. *Annals of Mathematical Statistics* **28** 1011–1015.

[7] Bray, N. and Morton, J. (2005). Equations defining hidden Markov models, in *Algebraic Statistics for Computational Biology* (Pachter, L. and Sturmfels, B., eds), Cambridge University Press, 235—247.

[8] Bruns, W. and Vetter, U. (1988). *Determinantal Rings*, Lecture Notes in Mathematics 1327, Springer, Berlin, Heidelberg.

[9] Burge, C. and Karlin, S. (1987). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268** 78–94.

[10] Cappe, O., Moulines, E. and Ryden, T. (2005). *Inference in Hidden Markov Models*, Springer Series in Statistics, Berlin, Heidelberg.

[11] Cox, D., Little, J. and O'Shea, D. (2005). *Ideals, Varieties and Algorithms*, Springer, Berlin, Heidelberg.

[12] Cueto, M.A., Morton, J. and Sturmfels, B. (2010). Geometry of the restricted Boltzmann Machine. *Algebraic Methods in Statistics and Probability*, (M. Viana and H. Wynn, eds), American Mathematical Society, Contemporary Mathematics. `http://front.math.ucdavis.edu/0908.4425`

[13] Dharmadhikari, S.W. (1963). Functions of finite markov chains. *Annals of Mathematical Statistics* **34** 1022–1032.

[14] Dharmadhikari, S.W. (1963). Sufficient conditions for a stationary process to be a function of a finite markov chain. *Annals of Mathematical Statistics*, **34** 1033–1041.

[15] Dharmadhikari, S.W. (1965). A characterization of a class of functions of finite markov chains. *Annals of Mathematical Statistics* **36** 524–528.

[16] Drton, M., Sturmfels, B. and Sullivant, S. (2009). *Lectures on Algebraic Statistics*, Oberwolfach Seminar Series 39, Birkhäuser.

[17] Drton, M. and Sullivant, S. (2007). Algebraic Statistical Models. *Statistica Sinica*, **17** 1273–1297. Available at `http://arxiv.org/abs/math/0703609`.

[18] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological sequence analysis*, Cambridge University Press, Cambridge, England.

[19] Ephraim, Y. and Merhav, N. (2002). Hidden Markov Processes. *IEEE Transactions on Information Theory*, **48** 1518–1569.

[20] Erickson, R.V. (1970). Functions of Markov Chains. *Annals of Mathematical Statistics* **41** 843–850.

[21] Faigle, U. and Schoenhuth, A. (2007). Asymptotic mean stationarity of sources with finite evolution dimension. *IEEE Transactions on Information Theory* **53(7)** 2342–2348.

[22] Faigle, U. and Schönhuth, A. (2011). Efficient tests for equivalence of hidden Markov processes and quantum random walks. *IEEE Transactions on Information Theory*, **57(3)** 1746–1753. `http://arxiv.org/abs/0808.2833`

[23] Finesso, L., Grassi, A. and Spreij, P. (2010). Approximation of stationary processes by hidden Markov models. *Mathematics of Control, Signals and Systems* **22** 1–22. Available at `http://arxiv.org/abs/math/0606591`.

[24] Fox, M. and Rubin, H. (1968). Functions of processes with Markovian states. *Annals of Mathematical Statistics*, **39** 938–946.

[25] Garcia, L.D., Stillman, M. and Sturmfels, B. (2005). Algebraic geometry of Bayesian networks. *Journal of Symbolic Computation* **39** 331–355. `http://arxiv.org/abs/math/0301255`

[26] Gilbert, E.J. (1959). On the identifiability problem for functions of finite Markov chains. *Annals of Mathematical Statistics* **30** 688–697.

[27] Hafemeister, C., Costa, I., Schönhuth, A. and Schliep, A. (2011). Classifying short gene expression time-courses with Bayesian estimation of piecewise constant functions. *Bioinformatics* **27(7)** 946–952.

[28] Harris, J. (1992). *Algebraic Geometry*, Springer, New York.

[29] Heller, A. (1965). On stochastic processes derived from Markov chains. *Annals of Mathematical Statistics*, **36(4)** 1286–1291.

[30] Ito, H., Amari, S.-I. and Kobayashi, K. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Transactions on Information Theory*, **38(2)** 324–333.

[31] Jaeger, H. (2000). Observable operator models for discrete stochastic time series. *Neural Computation* **12(6)** 1371–1398.

[32] A software system for research in algebraic geometry, `http://www.math.uiuc.edu/Macaulay2/`

[33] Pachter, L. and Sturmfels, B. (2005). *Algebraic Statistics for Computational Biology*, Cambridge University Press, Cambridge, .

[34] Paz, A. (1971). *Introduction to Probabilistic Automata.* Academic Press Inc, London, New York, San Diego.

[35] Petrie, T. (1969). Probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, **40(1)** 97–115.

[36] L. Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77(2)** 257–286.

[37] Schönhuth, A. and Jaeger, H. (2009). Characterization of ergodic hidden Markov sources. *IEEE Transactions on Information Theory*, **55(5)** 2107–2118.

[38] Sullivant, S. (2008). Algebraic geometry of Gaussian Bayesian networks. *Advances in Applied Mathematics* **40(4)** 482–518. `http://arxiv.org/abs/0704.0918`

[39] Vidyasagar, M. (2011). The complete realization problem for hidden Markov models: A survey and some new results. *Mathematics of Control, Signals and Systems*, **23(1)** 1–65.

[40] Zwiernik, P. and Smith, J.Q. (2012). Tree cumulants and the geometry of binary tree models. *Bernoulli*, **18(1)** 290–321. `http://arxiv.org/abs/1004.4360`