Observable operator models II: Interpretable models and model induction[1]

Herbert Jaeger
GMD, St. Augustin
herbert.jaeger@gmd.de

June 6, 1997

[1]This article appeared in the series Arbeitspapiere der GMD, GMD, Sankt Augustin, Nr. 1083, June 1997

Abstract: This article is a direct continuation of the tech report *Observable Operator Models and Conditioned Continuation Representations, Arbeitspapiere der GMD 1043, 1997.* While the former paper described the mathematical theory of OOMs, the present article presents techniques for practical applications. A standardized graphical representation of OOM-generated processes is developed, which helps a lot to gain an intuitive grasp on relevant phenomena. The main contribution is an efficient constructive algorithm for the induction of OOMs from data.

Zusammenfassung: Dieser Artikel ist eine direkte Fortsetzung des Reports *Observable Operator Models and Conditioned Continuation Representations, Arbeitspapiere der GMD 1043, 1997.* Während jener Aufsatz die Grundzüge der formalen Theorie von OOMs entwarf, beschreibt die vorliegende Arbeit Techniken für praktische Anwendungen. Eine standardisierte graphische Darstellung für OOM-generierte Prozesse fördert einen intuitiven Zugang zu relevanten Phänomenen. Der Hauptbeitrag besteht in einem effizienten konstruktiven Algorithmus für die Induktion von OOMs aus Daten.

1 Introduction

This paper is a direct continuation of [1], which is a prerequisite for the present one. References to that prior paper are made by "(I)"; e.g. "prop. 3 (I)" refers to proposition 3 in the first paper.

In section 2 of the present article, I describe a "standardized" subclass of OOMs, in which the axes of the state space can be interpreted in terms of certain probabilities of future events. These *interpretable* OOMs support a standardized graphical representation of state sequences. The (typically fractal) geometries of such sequences can be investigated, and different OOMs can be compared in terms of their geometrical properties (section 3). In section 4, I present a constructive procedure for the induction of OOMs from empirical time series. Section 5 concludes with a brief discussion.

2 Interpretable OOMs

Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0)$ be a minimal-dimension OOM. According to prop. 16 (I), we obtain the family of all equivalent minimal-dimension OOMs essentially by applying any internal-sum preserving vector space isomorphism on $\mathcal{A}$. There is a canonical 1-1 correspondence between such isomorphisms and the regular $m \times m$ matrices with column sumns equal to 1. Thus, one could map the family of equivalent minimal-dimensional OOMs on the space of such matrices. This space has dimension $m(m-1)$. In other words, uncountably many equivalent minimal-dimension OOMs co-exist.

In this section, it is shown how from that unwieldy family of equivalent minimal-dimension OOMs one can single out a few (countably many) "standardized" OOMs. As we shall see, these particular OOMs yield an easy-to-use basis for practical applications of various sorts.

The crucial requirement for "standardized" OOMs is that they are *interpretable*: the unit vectors of their state spaces represent probabilities of certain future events ("characteristic events"). This idea is worked out in this section.

Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0)$ be a minimal-dimension OOM, which generates a process (X_t). Let $\Sigma^k \subset \Sigma^*$ be the set of all words of length k. We start our way into interpretable OOMs by defining characteristic events:

Definition 1 *1. Any subset $B \subseteq \Sigma^k$ is called a* k-event.

2. Let $\bar{a} \in \Sigma^*$. The conditioned probability of a k-event B given $\bar{a}$ *is defined by*

$$P[B \mid \bar{a}] = \sum_{\bar{b} \in B} P[\bar{b} \mid \bar{a}] \tag{1}$$

(Recall that $P[\bar{b} \mid \bar{a}]$ is the probability that $\bar{b}$ is observed directly after $\bar{a}$, cf. def. 9 (I)).

3. *Let $\Sigma^k = A_1 \dot\cup \ldots \dot\cup A_m$ be a disjoint partitioning of Σ^k into m non-empty k-events. $A_1, \ldots, A_m$ are called* characteristic events *of (X_t) if some $\bar{a}_1, \ldots, \bar{a}_m \in \Sigma^*$ exist such that the $m \times m$ matrix*

$$(P[A_j \mid \bar{a}_i])_{i,j}$$

is regular.

For the remainder of this section, let $A_1, \ldots, A_m$ denote characteristic events of a process (X_t) generated by an OOM $\mathcal{A}$. Disjoint unions of sets will be denoted by $\dot\cup$.

Characteristic events do exist:

Proposition 1 *Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in \Sigma}, w_0)$ be a minimal-dimension OOM, which generates a process (X_t). Then there exists some $k \geq 1$, and a partitioning $\Sigma^k = A_1 \dot\cup \ldots \dot\cup A_m$ of Σ^k into characteristic events.*

Proof. First we observe (cf. prop. 8 (I)) that there exist m words $\bar{a}_1, \ldots, \bar{a}_m \in \Sigma^*$ such that $\mathfrak{g}_{\bar{a}_1}, \ldots, \mathfrak{g}_{\bar{a}_m} : \Sigma^* \to [0,1]$ are m linearly independent functions.

Therefore, m words $\bar{c}_1, \ldots, \bar{c}_m$ exist such that the $m \times m$ matrix $(\mathfrak{g}_{\bar{a}_i} \bar{c}_j)_{i,j}$ is regular.

Therefore (cf. def. 9 (I)), the matrix $(P[\bar{c}_j \mid \bar{a}_i])_{i,j}$ is regular.

Let $k := \max\{\mid \bar{c}_1 \mid, \ldots, \mid \bar{c}_m \mid\}$ be the maximal word length among the $\bar{c}_j$. We define the k-event B_j to consist of all words of length k beginning with $\bar{c}_j$:

$$B_j := \{\bar{c}_j \bar{a} \mid \bar{a} \in \Sigma^{k - |\bar{c}_j|}\}.$$

It holds that $P[B_j \mid \bar{a}_i] = P[\bar{c}_j \mid \bar{a}_i]$, which implies:

$$\text{The matrix } (P[B_j \mid \bar{a}_i])_{i,j} \text{ is regular.} \tag{2}$$

Now we transform the events B_j in two steps in order to arrive at characteristic events. In the first step, we make them disjoint.

We say that a word $\bar{a}$ *properly starts* a word $\bar{c}$ if there exists a non-empty word $\bar{b}$ such that $\bar{c} = \bar{a}\bar{b}$. We write $\bar{a} < \bar{c}$ if $\bar{a}$ properly starts $\bar{c}$. Note that $<$ is a partial ordering on Σ^*.

First we observe that for $B_r \cap B_s \neq \emptyset$ (where $r \neq s$), it holds that either $\bar{c}_r < \bar{c}_s$ or $\bar{c}_s < \bar{c}_r$. The first case implies $B_r \supset B_s$, the second case implies $B_s \supset B_r$ (note that $r \neq s$ implies $B_r \neq B_s$ because of (2)). If we define

$$B_r < B_s \quad :\text{iff} \quad \bar{c}_r < \bar{c}_s,$$

the partial ordering on words carries over to a partial ordering on the B_j's. It holds that

$$B_r < B_s \text{ iff } B_r \supset B_s. \tag{3}$$

Furthermore, for $r \neq s$ it holds that

$$B_r \cap B_s \neq \emptyset \Rightarrow B_r > B_s \vee B_s > B_r. \tag{4}$$

We now show that B_j is not exhausted by the B_s contained in it, e.g.

$$\forall j = 1, \ldots, m : \; B_j \neq \bigcup_{B_s > B_j} B_s \tag{5}$$

Assume that (5) is wrong, i.e. that for some j_0, $B_{j_0} = \bigcup_{B_s > B_{j_0}} B_s$.

Define

$$S := \{s \in \mathbb{N} \mid B_s > B_{j_0} \; \wedge \neg \exists B_r \; B_s > B_r > B_{j_0}\}.$$

Then it holds (observe (3)) that $B_{j_0} = \bigcup_{s \in S} B_s$. This union is even disjoint, i.e. $B_{j_0} = \dot{\bigcup}_{s \in S} B_s$, because of (4).

Therefore, it holds that $P[B_{j_0} \mid \bar{a}_i] = \sum_{s \in S} P[B_s \mid \bar{a}_i]$. This implies that the matrix $(P[B_j \mid \bar{a}_i])_{i,j}$ is not regular, which is a contradiction to (2). Therefore the assumption was wrong, i.e. (5) is true.

Now we define new B'_j, by taking away from B_j the B_s contained in them:

$$B'_j := B_j \setminus \bigcup_{B_s > B_j} B_s.$$

Because of (5), it holds that $B'_j \neq \emptyset$.

Define $S_j := \{s \mid B_s > B_j \; \wedge \; \neg \exists B_r : \; B_s > B_r > B_j\}$. Then it holds that

$$B'_j = B_j \setminus \dot{\bigcup}_{s \in S_j} B_s,$$

i.e. we get B'_j from B_j by taking away some *disjoint* B_s contained in it. As a consequence, for $i, j = 1, \ldots, m$ it holds that

$$P[B'_j \mid \bar{a}_i] = P[B_j \mid \bar{a}_i] - \sum_{s \in S_j} P[B_s \mid \bar{a}_i],$$

which in turn implies that the matrix $(P[B'_j \mid \bar{a}_i])_{i,j}$ is regular. In this matrix, the k-events B'_j are disjoint.

In the second step, we blow up the B'_j to make events A_j which exhaust Σ^k, i.e. we arrive at a situation where $A_1 \dot{\cup} \ldots \dot{\cup} A_m = \Sigma^k$.

Put $B'_0 := \Sigma^k \setminus (B'_1 \cup \ldots \cup B'_m)$. If $B'_0 = \emptyset$, we are done, since then we can put $A_j := B'_j$. In the case $B'_0 \neq \emptyset$, consider the $(m+1) \times m$ matrix

$$M = \begin{pmatrix} P[B'_0 \mid \bar{a}_1] & P[B'_1 \mid \bar{a}_1] & \cdots & P[B'_m \mid \bar{a}_1] \\ \cdots & \cdots & & \cdots \\ P[B'_0 \mid \bar{a}_m] & P[B'_1 \mid \bar{a}_m] & \cdots & P[B'_m \mid \bar{a}_m] \end{pmatrix}$$

Call the column vectors of M $v_0, v_1, \ldots, v_m$. M has rank m, since the matrix $(v_1, \ldots, v_m)$ has rank m. Therefore, v_0 is a linear combination of $v_1, \ldots, v_m$. We distinguish two cases.

Case 1: v_0 is the null vector. We put $A_1 := B'_1 \cup B'_0$, and $A_j := B'_j$ for $j > 1$. Then $(P[A_j \mid \bar{a}_i])_{i,j} = (P[B'_j \mid \bar{a}_i])_{i,j}$, which we know to be regular, i.e. the A_j's are characteristic events.

Case 2: $v_0 \neq 0$. Let $v_0 = \sum_{s=1,\ldots,m} \alpha_s v_s$. Since all v_s are non-null vectors with non-negative components only, some α_{j_0} must be properly greater than 0. This implies that the matrix $M' := (v_1, \ldots, v_{j_0} + v_0, \ldots, v_m)$ still has rank m, i.e. is regular. We put $A_{j_0} := B'_{j_0} \cup B'_0$, and $A_j := B'_j$ for $j \neq j_0$. Then $(P[A_j \mid \bar{a}_i])_{i,j} = M'$, which is regular, i.e. the A_j's are characteristic events. □

This somewhat painstaking proof should not leave the reader with the impression that characteristic events are a rare commodity. Quite to the contrary, any random partition of Σ^k into m events of non-zero probability is exceedingly likely to produce characteristic events (since, roughly speaking, it is almost certain that an essentially random matrix $(P[A_j \mid \bar{a}_i])_{i,j}$ is regular).

For the remainder of this section, let (X_t) be the process generated by $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in \Sigma}, w_0)$, where m is the dimension of the process, and let $\bar{a}_i \in \Sigma^*, A_j \subset \Sigma^k$ be words and characteristic events such that $M = (P[A_j \mid \bar{a}_i])_{i,j}$ is regular.

We will now show how $\mathcal{A}$ can be transformed into an equivalent, minimal-dimension OOM $\tilde{\mathcal{A}}$, which is interpretable in the sense that the m coordinate

axes of $\tilde{\mathcal{A}}$'s state space directly represent the probabilities that the characteristic events $A_1, \ldots, A_m$ will be observed next.

First, we generalize observable operators to cover the observation of k-events:

Definition 2 *Let $B \in \Sigma^k$ be a k-event. Then*

$$\tau_B := \sum_{\bar{b} \in B} \tau_{\bar{b}}$$

is the operator corresponding to the observation of B.

The next definition introduces a convenient notational abbreviation, which we will use in the next proposition to come:

Definition 3 *Let $\bar{a}_i$ be one of the m words occuring in M. Then*

$$x_i := \frac{\tau_{\bar{a}_i} w_0}{\sigma \tau_{\bar{a}_i} w_0}$$

denotes the state vector obtained after an application of $\tau_{\bar{a}_i}$ on w_0, renormalized to internal sum 1 (for the intuitive meaning of this renormalization, cf. the introduction section in (I)).

Note that the well-definedness of M implies that $\sigma \tau_{\bar{a}_i} w_0 \neq 0$, i.e. x_i is well-defined.

We collect some properties of τ_{A_j} and x_i:

Proposition 2 *1. $P[A_j \mid \bar{a}_i] = \sigma \tau_{A_j} x_i$.*

2. $\forall y \in H : \sum_{j=1,\ldots,m} \sigma \tau_{A_j} y = 1$.

3. $\forall i = 1, \ldots, m : \; x_i \in H$.

4. $x_1, \ldots, x_m$ are linearly independent.

5. Define a mapping $\varrho : \mathbb{R}^m \to \mathbb{R}^m$ by putting

$$\varrho(x) := (\sigma \tau_{A_1} x, \ldots, \sigma \tau_{A_m} x).$$

ϱ is a bijective, linear, internal-sum preserving mapping.

Proof. (1)

$$\begin{aligned} P[A_j \mid \bar{a}_i] \quad &= \\ &= \sum_{\bar{b} \in A_j} P[\bar{b} \mid \bar{a}_i] \;=\; \sum_{\bar{b} \in A_j} \frac{P[\bar{a}_i \bar{b}]}{P[\bar{a}_i]} \\ &= \sum_{\bar{b} \in A_j} \frac{\sigma \tau_{\bar{a}_i \bar{b}} w_0}{\sigma \tau_{\bar{a}_i} w_0} \;=\; \sum_{\bar{b} \in A_j} \sigma \tau_{\bar{b}} \frac{\tau_{\bar{a}_i} w_0}{\sigma \tau_{\bar{a}_i} w_0} \\ &= \sum_{\bar{b} \in A_j} \sigma \tau_{\bar{b}} x_i \;=\; \sigma \tau_{A_j} x_i. \end{aligned}$$

(2) $\sum_{j=1,\ldots,m} \sigma \tau_{A_j} y = \sum_{\bar{b} \in \Sigma^k} \sigma \tau_{\bar{b}} y = 1$.

(3) Obvious.

(4) Assume that $x_1, \ldots, x_m$ are linearly dependent. Use (1) and linearity of σ and τ_{A_j} to conclude that M is not regular, a contradiction.

(5) Linearity of ϱ is a consequence of the linearity of σ and τ_{A_j}. In order to show bijectivity and preservation of internal sums, consider $x_1, \ldots, x_m$. Because of (3) and (4), they are in H and they are linearly independent. Because of (2), they are mapped by ϱ into H. Because of (1) and regularity of M, the ϱ-images of $x_1, \ldots, x_m$ are linearly independent. Therefore, ϱ bijectively maps H onto itself, which implies that ϱ is bijective and preserves internal sums. $\Box$

According to the last statement of this proposition, ϱ has all the properties required in prop. 16 (I), which states that we obtain an OMM $\tilde{\mathcal{A}} = (\mathbb{R}^k, (\tilde{\tau}_a)_{a \in \Sigma}, \tilde{w}_0)$ equivalent to $\mathcal{A}$ by putting

$$\tilde{\mathcal{A}} = (\mathbb{R}^k, (\varrho \tau_a \varrho^{-1})_{a \in \Sigma}, \varrho w_0).$$

The $m \times m$ matrix corresponding to ϱ can easily be obtained from the original OOM $\mathcal{A}$:

Proposition 3

$$\varrho = (\sigma \tau_{A_i} \mathfrak{e}_j)_{ij},$$

where $\mathfrak{e}_j$ is the j-th unit vector.

Proof. Follows directly from

$$\begin{pmatrix} \varrho_{1j} \\ \cdots \\ \varrho_{mj} \end{pmatrix} = \varrho \mathfrak{e}_j = \begin{pmatrix} \sigma \tau_{A_1} \mathfrak{e}_j \\ \cdots \\ \sigma \tau_{A_m} \mathfrak{e}_j \end{pmatrix}. \quad \Box$$

$\tilde{\mathcal{A}}$ has a remarkable property, which provides the key for all further results reported in this paper:

Proposition 4 *Let $v \in H, v = (v_1, \ldots, v_m)$. Then it holds that the m components of v represent the probabilities that the corresponding characteristic events will be observed when the system is in state v:*

$$\forall j = 1, \ldots, m : \; v_j = \sigma \tilde{\tau}_{A_j} v \tag{6}$$

A notational variant of (6) is

$$(v_1, \ldots, v_m) = (P[A_1 \mid v], \ldots, P[A_m \mid v]),$$

which displays the core idea of (6) more clearly.

Proof. Select $x \in H$ such that $v = \varrho x$, i.e.

$$v = (\sigma \tau_{A_1} x, \ldots, \sigma \tau_{A_m} x).$$

Since $\sigma \tilde{\tau}_{A_j} v = \sigma \tau_{A_j} x$, this directly implies (6). $\square$

Within the family of OOM's generating (X_t), there exists at most one OOM which has the property (6) (exercise). I.e., characteristic events uniquely determine a special OOM, which does not depend on the OOM $\mathcal{A}$ and the words $\bar{a}_i$ which we used here for a constructive proof of existence. This justifies the following definition:

Definition 4 *Let $A_1, \ldots, A_m$ be a set of characteristic events of (X_t). Then $\mathcal{A}(A_1, \ldots, A_m)$ denotes the OOM which has property (6). It is called the OOM* interpretable by *$A_1, \ldots, A_m$.*

A neat, albeit somewhat informal way of characterizing the "knack" of interpretable OOMs is to say that the states of $\mathcal{A}(A_1, \ldots, A_m)$ have the form $(P[A_1 \mid \cdot], \ldots, P[A_m \mid \cdot])$. In informal discussions I will sometimes use this notation.

An interpretable OOM has the following properties, which highlight the intimate connection between states and operators on the one hand, and probabilities of characteristic events on the other:

Proposition 5 *In an interpretable OOM $\mathcal{A}(A_1, \ldots, A_m) = (\mathbb{R}^m, (\tau_a)_{a \in \Sigma}, w_0)$, the following statements hold:*

1. $w_0 = (P[A_1], \ldots, P[A_m])$,

2. If $P[\bar{b}] \neq 0$, *then* $\tau_{\bar{b}} w_0 = (P[\bar{b}A_1], \ldots, P[\bar{b}A_m])$.

Proof. (1) Follows directly from $\sigma\tau_{A_i} w_0 = P[A_i]$ and (6).
(2) According to (6), it holds that

$$\frac{\tau_{\bar{b}} w_0}{\sigma(\tau_{\bar{b}} w_0)} = (P[A_1 \mid \frac{\tau_{\bar{b}} w_0}{\sigma(\tau_{\bar{b}} w_0)}], \ldots, P[A_m \mid \frac{\tau_{\bar{b}} w_0}{\sigma(\tau_{\bar{b}} w_0)}]).$$

Observing that $\frac{\tau_{\bar{b}} w_0}{\sigma(\tau_{\bar{b}} w_0)}$ is the renormalized state vector obtained after an application of $\tau_{\bar{b}}$, i.e. after an observation of $\bar{b}$, this is equivalent to stating

$$\frac{\tau_{\bar{b}} w_0}{\sigma(\tau_{\bar{b}} w_0)} = (P[A_1 \mid \bar{b}], \ldots, P[A_m \mid \bar{b}]).$$

Using $\sigma(\tau_{\bar{b}} w_0) = P[\bar{b}]$ and $P[A_i \mid \bar{b}]P[\bar{b}] = P[\bar{b}A_i]$, the statement is obtained immediately. □

Since a process (X_t) has many different sets of characteristic events, there are many different, equivalent, interpretable OOMs. Thus, this is not a "normal form" representation in the usual sense of the word. However, after the selection of a particular set of characteristic events, we have gained, for all practical purposes, the main benefits usually afforded by normal form generators. In particular, we can *compare* non-equivalent OOMs, and we can *construct* an OOM from time series data. These two applications are the themes of the next two sections.

3 A standardized visualization of OOM-generated processes

In this section, instead of furthering mathematical theory, I will "play" a bit with interpretable OOMs, highlighting the almost palpable access to OOM-generated processes they afford. We will visualize state sequences of paths of (X_t). Such states are vectors $v \in H$. Since we can best represent graphically hyperplanes H when they are 2-dimensional, we will stick to that case. This implies that we will be dealing essentially with 3-dimensional processes (X_t) in this section (recall that H is an $(m-1)$-dimensional hyperplane in the m-dimensional state space of an OOM).

Thus, for the remainder of this section, we consider an interpretable, 3-dimensional OOM $\mathcal{A} = \mathcal{A}(A_1, A_2, A_3) = (\mathbb{R}^3, (\tau_a)_{a\in\Sigma}, w_0)$.

An elementary property of $\mathcal{A}$ is that state vectors always lie in the completely non-negative part of H. More precisely, define $H^{\geq 0} := \{(v_1, v_2, v_3) \in H \mid v_i \geq 0 \ (i = 1, 2, 3)\}$. Then the following proposition is a direct implication of definition 3 (I), property 3, and of proposition 6:

Proposition 6 *1.*

$$w_0 \in H^{\geq 0}$$

2.

$$\forall \bar{a} \in \Sigma^* : \ \sigma\tau_{\bar{a}} w_0 = 0 \ \vee \ \frac{\tau_{\bar{a}} w_0}{\sigma\tau_{\bar{a}} w_0} \in H^{\geq 0} \quad \square$$

Intuitively, this means that if we consider some path $a_0 a_1 a_2 \ldots$ of (X_t), the corresponding sequence of hidden states is confined to $H^{\geq 0}$, i.e. the "triangle" area depicted in figure 1a. This has the useful practical consequence that we can graphically represent state sequences of interpretable (3-dimensional) OOMs in a standardized fashion. We use H as the drawing plane, in which we place, for our orientation, the contours of $H^{\geq 0}$. This is an equilateral triangle whose edges have length $\sqrt{2}$. If $v = (v_1, v_2, v_3) \in H^{\geq 0}$ is a state vector, an elementary geometrical argument tells us that its components can be recovered from its position within this triangle, by exploiting $v_i = \sqrt{2/3} d_i$, where d_i $(i = 1, 2, 3)$ are the distances to the edges of the triangle (compare fig.1b).

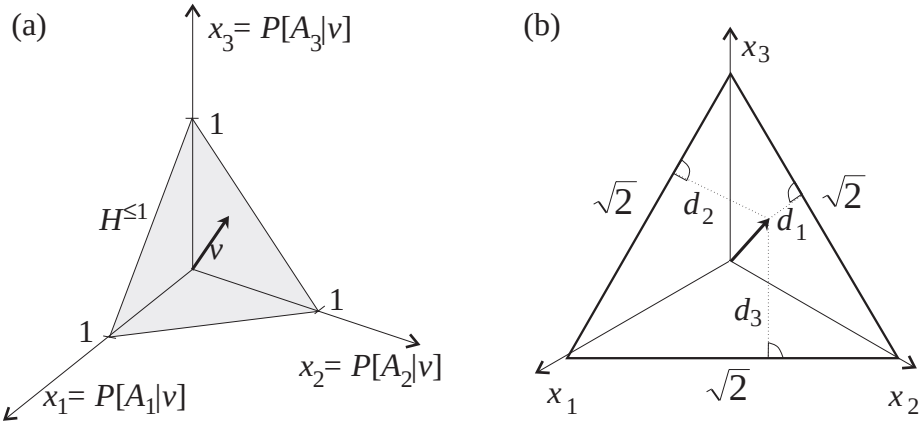

Figure 1: (a) The positioning of $H^{\geq 0}$ within state space. (b) The setup of standardized plots, and how the components v_1, v_2, v_3 of a state vector v can be recovered from the graphical representation, exploiting $v_i = \sqrt{2/3} d_i$.

Quite frequently one will encounter OOMs where the dimension m coincides with the number of observable operators, i.e. where $\Sigma = \{a_1, \ldots, a_m\}$,

and where the operators are linearly independent. In that case, the operators themselves can be interpreted as characteristic events. Said more precisely, the singleton sets $(A_j)_{j=1,\ldots,m} = (\{a_j\})_{j=1,\ldots,m}$ are characteristic events.

We will graphically investigate an OOM of this kind in the remainder of this section. The purpose is to illustrate the usefulness of the standardized representation.

Consider the following HMM (where $\Sigma = \{a, b, c\}$), which is specified by

$$M = \begin{pmatrix} .1 & .1 & .8 \\ 0 & .8 & .2 \\ .8 & 0 & .2 \end{pmatrix} \tag{7}$$

$$O_a = \begin{pmatrix} 0 & 0 & 0 \\ 0 & .3 & 0 \\ 0 & 0 & .5 \end{pmatrix} \quad O_b = \begin{pmatrix} .7 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & .3 \end{pmatrix} \quad O_c = \begin{pmatrix} .3 & 0 & 0 \\ 0 & .7 & 0 \\ 0 & 0 & .2 \end{pmatrix} \tag{8}$$

This HMM can be interpreted as a 3-dimensional OOM (cf. (I), section 2), which we shall call $\mathcal{B}$. We wish to graphically render state sequences of the corresponding interpretable OOM $\mathcal{B}(\{a\}, \{b\}, \{c\})$.

Since we wish to plot *interpretable* states, one might think that it is necessary to first transform $\mathcal{B}$ into $\mathcal{B}(\{a\}, \{b\}, \{c\})$. Actually, this is not necessary. We can employ the original (non-interpretable) OOM $\mathcal{B}$, using the generator procedure described in (I), section 2. Recall that in this procedure, at each time step t, where the generator $\mathcal{B}$ is in (hidden) state s, the probabilities $P[a \mid s], P[b \mid s], P[c \mid s]$ are computed, which specify the chances that a, b, or c is selected. Now, the triple $(P[a \mid s], P[b \mid s], P[c \mid s])$ is exactly the (interpretable) state (v_1, v_2, v_3) of the interpretable OOM $\mathcal{B}(\{a\}, \{b\}, \{c\})$. Thus, here we get the interpretable state sequence for free even when we use a non-interpretable OOM as a generator.

Note that this is possible because we are dealing with characteristic events that coincide with the observable operators. In the general case, when we wish to plot interpretable state sequences derived from other characteristic events, we would have to construct the corresponding interpretable OOM.

According to the scheme just outlined, $\mathcal{B}$ was run (using Mathematica on a MacIntosh) as a generator for 230 time steps. The initial 30 steps were discarded, and the 200 remaining interpretable states were plotted in the reference triangle described in fig. 1. Figure 2 shows the resulting plot (left) and an enlarged portion (right).

Allow me to make a few informal remarks on what we see in these plots. The operators τ_a and τ_b have (regarded as matrices) rank 2. Accordingly, states produced by an application of either of these operators lie on a single

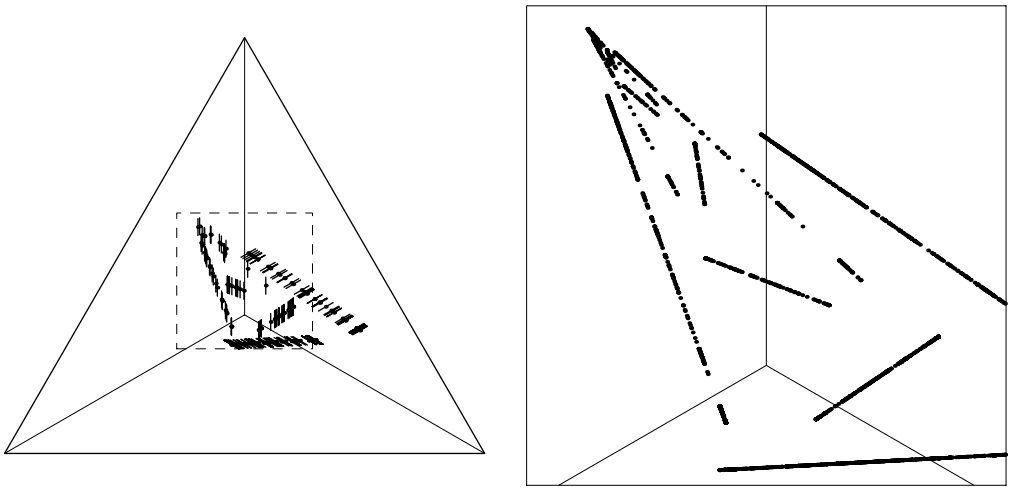

Figure 2: *Left*: A 200 step sequence of states of $\mathcal{B}(\{a\},\{b\},\{c\})$, an interpretable OOM equivalent to the one specified in (7), (8). The operators from whose application states originate are indicated by bars which are parallel to the corresponding axes: / signifies that the state is the result of an application of τ_a, \ indicates τ_b, and | means τ_c. *Right*: Enlarged section from the left diagram. The process has been run 5000 steps to produce this figure.

straight line in H. By contrast, τ_c is regular. Thus, states produced by this operator are not confined to a single line. In this example, one finds several line segments on which τ_c-produced states fall. These lines are the (iterated) τ_c-images of the single lines produced by τ_a and τ_b.

Let us now modify the example a bit and see what happens. We leave the Markov transition matrix (7) unchanged. Likewise, τ_c is not touched (i.e., O_c is not modiefied). We make τ_b regular by replacing the 0 on the diagonal of O_b by 0.2. This modification is compensated by changing the entry .3 in O_a to .1 (recall from (I) that $O_a + O_b + O_c$ must be the identity matrix):

$$O_a = \begin{pmatrix} 0 & 0 & 0 \\ 0 & .1 & 0 \\ 0 & 0 & .5 \end{pmatrix} \quad O_b = \begin{pmatrix} .7 & 0 & 0 \\ 0 & .2 & 0 \\ 0 & 0 & .3 \end{pmatrix} \quad O_c = \begin{pmatrix} .3 & 0 & 0 \\ 0 & .7 & 0 \\ 0 & 0 & .2 \end{pmatrix} \tag{9}$$

Figure 3 shows state sequences of (the interpretable version of) this modified OOM. They are computed in the same way as in fig. 2.

Comparing fig. 2 with fig. 3, one finds (among other things) that the states generally have shifted a bit to the right. This means that the overall selection probability of τ_a has decreased and that of τ_b has increased. This

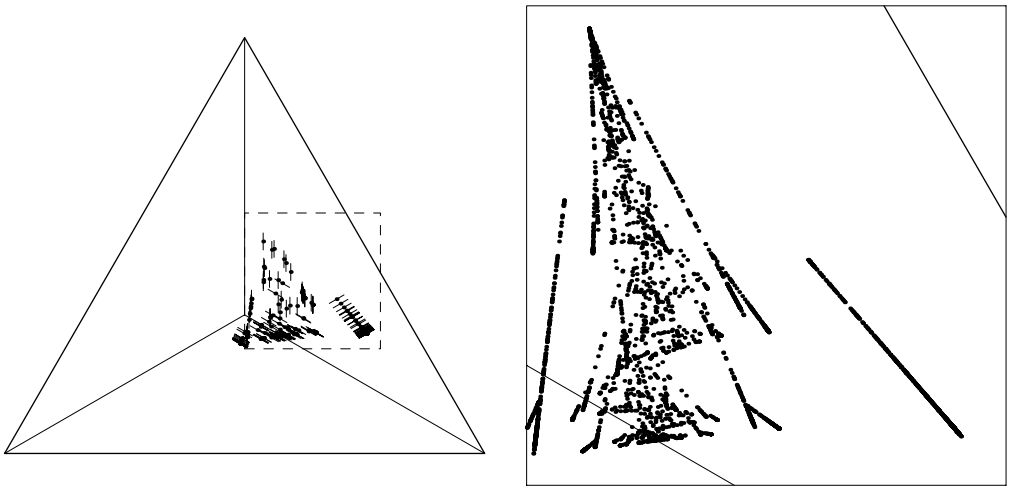

Figure 3: The analogue of fig. 2 for the modified OOM (9).

is easily (if superficially) explained: the sum of elements of the matrix τ_b has increased, while that of τ_a has decreased.

Another conspicuous visual difference between the two examples is that fig. 3 looks more "fractal" than fig. 2. The fractal appearance becomes still stronger after the following modification of the O_i, which results in three regular observable operators:

$$O_a = \begin{pmatrix} .2 & 0 & 0 \\ 0 & .1 & 0 \\ 0 & 0 & .6 \end{pmatrix} \quad O_b = \begin{pmatrix} .5 & 0 & 0 \\ 0 & .2 & 0 \\ 0 & 0 & .3 \end{pmatrix} \quad O_c = \begin{pmatrix} .3 & 0 & 0 \\ 0 & .7 & 0 \\ 0 & 0 & .1 \end{pmatrix} \quad (10)$$

The process determined by (10) is visualized in fig. 4.

The fractal structure of state sequences as revealed in these graphics comes not as a surprise – fractal attractors of this kind arise naturally from mixed iterations of several linear mappings. However, I will not enter into this line of investigation here. All I wanted to show is that a handy visualization method is a powerful help. Although it does not in itself yield explanations, it does direct our attention to interesting phenomena, and allows to "play" with them.

The visualization techniques described in this section can be applied to processes of dimension greater than 3. In such cases, the n-dimensional state space must be projected on a 3-dimensional one, in a way which preserves internal sums. A particularly transparent projection of this kind can be obtained by merging characteristic events. If $A_1, \ldots, A_n$ are characteristic

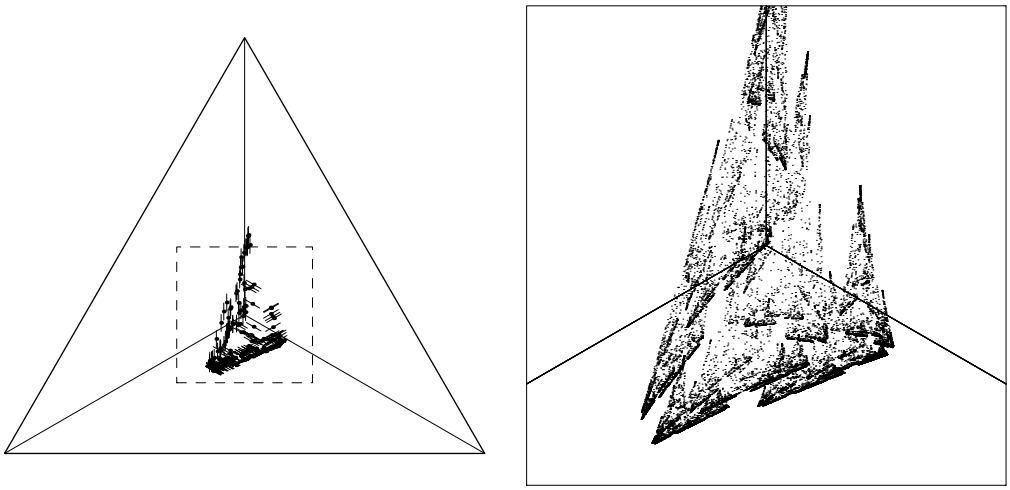

Figure 4: The analogue of fig. 2 for the modified OOM (9). 30000 time steps were executed to obtain the right figure.

events of a high-dimensional process (i.e. where $n > 3$), define 3 "semi-characteristic" events B_1, B_2, B_3 by merging the A_i into three sets, e.g. $B_1 = A_1 \cup \ldots \cup A_{r_1}, B_2 = A_{r_1+1} \cup \ldots \cup A_{r_2}, B_3 = A_{r_2+1} \cup \ldots \cup A_n$. Then instead of plotting the n-dimensional states $(P[A_1 \mid \cdot], \ldots, P[A_n \mid \cdot])$ of $\mathcal{A}(A_1, \ldots, A_n)$, plot the 3-dimensional "semi-states" $(P[B_1 \mid \cdot], P[B_1 \mid \cdot], P[B_3 \mid \cdot])$.

4 Reconstructing OOMs from data

Assume that a sequence of observations $S = a_0, a_1, \ldots, a_N$ is given which has been generated by an unknown OOM $\mathcal{U}$. In this section we will learn how to reconstruct from S an OOM $\mathcal{A}$ which is equivalent to $\mathcal{U}$. More precisely, we will learn how to compute $\mathcal{A}$ from a finite number of conditioned continuation probabilities $P[\bar{a} \mid \bar{b}]$.

Such conditioned continuation probabilities can be easily estimated from S by a simple counting of occurences of substrings in S, exploiting that $P[\bar{a} \mid \bar{b}] \approx \frac{N_S(\bar{b}\bar{a})}{N_S(\bar{b})}$, where $N_S\bar{c}$ is the number of occurcences of a subsequence $\bar{c}$ in S.

One caveat I must mention at the outset. The reconstruction of an OOM equivalent to $\mathcal{U}$ cannot be perfect, since the finite sequence S contains only a finite amount of information, whereas $\mathcal{U}$ contains an infinite amount of information (being specified in terms of real numbers). Therefore, the reconstruction procedure will come up with only an *estimate* $\tilde{\mathcal{A}}$ of an OOM $\mathcal{A}$

equivalent to $\mathcal{U}$. More precisely, from S we can only derive *estimated* conditioned probabilities $\tilde{P}[\bar{a} \mid \bar{b}]$, which we have to use in the reconstruction procedure instead of the correct ones. A proper treatment of this situation would require a statistical theory of distributions of estimated $\tilde{P}[\bar{a} \mid \bar{b}]$, which would allow us to calculated how strongly $\tilde{\mathcal{A}}$ is expected to deviate from $\mathcal{A}$, given a length N of observation. Such a statistical theory I cannot offer here.

The best I can offer for the time being is a reconstruction procedure which perfectly reconstructs $\mathcal{U}$ in the limit of $N \to \infty$.

The reconstruction proceeds in two steps. First, the dimension m of the process (X_t) (of which S is a finite path) is calculated, along with characteristic events $A_1, \ldots, A_m$. In the second step, an interpretable OOM $\mathcal{A}(A_1, \ldots, A_m)$ is constructed. A subsection is devoted to each of the steps.

4.1 Calculation of process dimension

The calculation of the process dimension m relies on two technical propositions (props. 7 and 8). Since these propositions are most conveniently proven using conditioned continuation representations (CCRs), we shall adopt that framework here (recall from (I), section 3, that $\mathfrak{g}_{\bar{b}}\bar{a} = P[\bar{a} \mid \bar{b}]$).

Proposition 7 *Let $n \in \mathbb{N}$, and $\mathfrak{g}_{\bar{b}_1}, \ldots, \mathfrak{g}_{\bar{b}_n} \in \mathfrak{G}$. Let $\Sigma^{\leq r} = \{\bar{c} \in \Sigma^* \mid \mid \bar{c} \mid \leq r\}$ denote the words of lenght at most r. Let $\bar{a}_1, \ldots, \bar{a}_{k_r}$ be the alphabetical enumeration of $\Sigma^{\leq r}$. Let M_r be the $n \times k_r$ matrix*

$$M_r = \begin{pmatrix} \mathfrak{g}_{\bar{b}_1}\bar{a}_1 & \cdots & \mathfrak{g}_{\bar{b}_1}\bar{a}_{k_r} \\ \vdots & & \vdots \\ \mathfrak{g}_{\bar{b}_n}\bar{a}_1 & \cdots & \mathfrak{g}_{\bar{b}_n}\bar{a}_{k_r} \end{pmatrix}$$

The following statements hold for M_r:

1. $\operatorname{rk} M_r = \operatorname{rk} M_{r+1} \;\Rightarrow\; \operatorname{rk} M_{r+1} = \operatorname{rk} M_{r+2}$,
2. $\operatorname{rk} M_r = \operatorname{rk} M_{r+1} \;\Rightarrow\; r = dim[\mathfrak{g}_{\bar{b}_1}, \ldots, \mathfrak{g}_{\bar{b}_n}]$,

where rk *denotes the rank of a matrix, and $[\mathfrak{g}_{\bar{b}_1}, \ldots, \mathfrak{g}_{\bar{b}_n}]$ is the linear subspace of $\mathfrak{G}$ spanned by $\mathfrak{g}_{\bar{b}_1}, \ldots, \mathfrak{g}_{\bar{b}_n}$.*

Proof. (1). For $\bar{d} \in \Sigma^*$, let $x_{\bar{d}}$ denote the column vector

$$x_{\bar{d}} := \begin{pmatrix} \mathfrak{g}_{\bar{b}_1}\bar{d} \\ \vdots \\ \mathfrak{g}_{\bar{b}_n}\bar{d} \end{pmatrix}.$$

Using that M_r actually consists of some initial column vectors of M_{r+1}, and that $\operatorname{rk} M_r = \operatorname{rk} M_{r+1}$, we can conclude that for $\bar{c} \in \Sigma^{r+1}$, $x_{\bar{c}}$ can be written as a linear combination from column vectors $x_{\bar{a}_i}$, where the $\bar{a}_i$ are the words from $\Sigma^{\leq r}$:

$$x_{\bar{c}} = \sum_{i=1,\ldots,k_r} \alpha_i^{\bar{c}} x_{\bar{a}_i}.$$

This means that for $j = 1, \ldots, n$, and $c_1 \ldots c_{r+1} \in \Sigma^{r+1}$ it holds that

$$\mathfrak{g}_{\bar{b}_j} c_1 \ldots c_{r+1} = \sum_{i=1,\ldots,k_r} \alpha_i^{c_1 \ldots c_{r+1}} \mathfrak{g}_{\bar{b}_j} \bar{a}_i. \tag{11}$$

Now we consider some $c_1 \ldots c_{r+1} c_{r+2} \in \Sigma^{r+2}$. Using the equation $\mathfrak{t}_a(\mathfrak{g}_{\bar{c}}) = P[a \mid \bar{c}]\mathfrak{g}_{\bar{c}a}$ (cf. eqn. (12)(I)), elementary transformations reveal that

$$\mathfrak{g}_{\bar{b}_j} c_1 \ldots c_{r+1} c_{r+2} = \mathfrak{t}_{c_1} \mathfrak{g}_{\bar{b}_j} c_2 \ldots c_{r+2}.$$

.

Utilizing this fact and (11), we can rewrite $\mathfrak{g}_{\bar{b}_j} c_1 \ldots c_{r+1} c_{r+2}$ as follows:

$$\begin{aligned}
&\mathfrak{g}_{\bar{b}_j} c_1 \ldots c_{r+1} c_{r+2} = \\
&= \mathfrak{t}_{c_1} \mathfrak{g}_{\bar{b}_j} c_2 \ldots c_{r+2} = \mathfrak{t}_{c_1} \sum_{i=1,\ldots,k_r} \alpha_i^{c_2 \ldots c_{r+2}} \mathfrak{g}_{\bar{b}_j} \bar{a}_i \\
&= \sum_{i=1,\ldots,k_r} \alpha_i^{c_2 \ldots c_{r+2}} \mathfrak{t}_{c_1} \mathfrak{g}_{\bar{b}_j} \bar{a}_i = \sum_{i=1,\ldots,k_r} \alpha_i^{c_2 \ldots c_{r+2}} \mathfrak{g}_{\bar{b}_j c_1} \bar{a}_i.
\end{aligned}$$

For column vectors, this implies

$$\begin{aligned}
\begin{pmatrix} \mathfrak{g}_{\bar{b}_1} c_1 \ldots c_{r+2} \\ \vdots \\ \mathfrak{g}_{\bar{b}_n} c_1 \ldots c_{r+2} \end{pmatrix} &= \begin{pmatrix} \sum_{i=1,\ldots,k_r} \alpha_i^{c_2 \ldots c_{r+2}} \mathfrak{g}_{\bar{b}_1 c_1} \bar{a}_i \\ \vdots \\ \sum_{i=1,\ldots,k_r} \alpha_i^{c_2 \ldots c_{r+2}} \mathfrak{g}_{\bar{b}_n c_1} \bar{a}_i \end{pmatrix} \\
&= \sum_{i=1,\ldots,k_r} \alpha_i^{c_2 \ldots c_{r+2}} \begin{pmatrix} \mathfrak{g}_{\bar{b}_1 c_1} \bar{a}_i \\ \vdots \\ \mathfrak{g}_{\bar{b}_n c_1} \bar{a}_i \end{pmatrix},
\end{aligned}$$

i.e., column vectors $x_{\bar{c}}$, where $\bar{c} \in \Sigma^{r+2}$, can be linearly combined from column vectors from M_{r+1}. This implies (1).

(2): First observe that (1) directly implies $\operatorname{rk} M_r = \operatorname{rk} M_{r+1} \Rightarrow \operatorname{rk} M_{r+1} = \operatorname{rk} M_{r+s}$ for all $s \in \mathbb{N}$. This in turn implies (2), if one exploits that

$$dim[\mathfrak{g}_{\bar{b}_1}, \ldots, \mathfrak{g}_{\bar{b}_n}] = dim[\{(\mathfrak{g}_{\bar{b}_1}\bar{a}, \ldots, \mathfrak{g}_{\bar{b}_n}\bar{a}) \in \mathbb{R}^n \mid \bar{a} \in \Sigma^*\}],$$

where the rhs. denotes the dimension of the linear subspace of $\mathbb{R}^n$ which is spanned by the vectors of the kind $(\mathfrak{g}_{\bar{b}_1}\bar{a}, \ldots, \mathfrak{g}_{\bar{b}_n}\bar{a})$. $\Box$

Proposition 8 *Let $d_p = dim[\{\mathfrak{g}_{\bar{b}} \mid \bar{b} \in \Sigma^{\leq p}\}]$ denote the dimension of the subspace of $\mathfrak{G}$ spanned by the $\mathfrak{g}_{\bar{b}}$, where the length of $\bar{b}$ is at most p. Then it holds that*

1. *$d_p = d_{p+1} \Rightarrow d_p = d_{p+s}$ for all $s \in \mathbb{N}$,*
2. *$d_p = d_{p+1} \Rightarrow d_p = dim(X_t)$.*

Proof. (1). Let $\mathfrak{G}_p = [\{\mathfrak{g}_{\bar{b}} \mid \bar{b} \in \Sigma^{\leq p}\}]$ be the linear subspace of $\mathfrak{G}$ spanned by the $\mathfrak{g}_{\bar{b}}$, where the length of $\bar{b}$ is at most p. Then obviously

$$\mathfrak{G}_p \subseteq \mathfrak{G}_{p+1}. \tag{12}$$

Furthermore, if $\Sigma = \{a_1, \ldots, a_k\}$, for any $q \in \mathbb{N}$ it holds that

$$\mathfrak{G}_{q+1} = [\bigcup\{\mathfrak{t}_{a_1}\mathfrak{G}_q, \ldots, \mathfrak{t}_{a_k}\mathfrak{G}_q\} \cup \{mathfrakg_\varepsilon\}], \tag{13}$$

where the rhs. denotes the linear subspace of $\mathfrak{G}$ spanned by the union of the $\mathfrak{t}_{a_i}$-images of $\mathfrak{G}_q$.

(13) can be derived as follows:

$$\begin{aligned}
\mathfrak{G}_{q+1} &= [\{\mathfrak{g}_{\bar{a}b} \mid \bar{a} \in \Sigma^{\leq q}, b \in \Sigma\} \cup \{mathfrakg_\varepsilon\}] \\
&= [\{\mathfrak{g}_{\bar{a}b} \mid \bar{a} \in \Sigma^{\leq q}, b \in \Sigma, \mathfrak{g}_{\bar{a}b} \neq 0\} \cup \{mathfrakg_\varepsilon\}] \\
&= [\{\mathfrak{g}_{\bar{a}b} \mid \bar{a} \in \Sigma^{\leq q}, b \in \Sigma, P[b \mid \bar{a}] \neq 0\} \cup \{mathfrakg_\varepsilon\}] \\
&= [\{\frac{\mathfrak{t}_b\mathfrak{g}_{\bar{a}}}{P[b \mid \bar{a}]} \mid \bar{a} \in \Sigma^{\leq q}, b \in \Sigma, P[b \mid \bar{a}] \neq 0\} \cup \{mathfrakg_\varepsilon\}] \quad \text{(cf. (12)(I))} \\
&= [\bigcup\{\mathfrak{t}_{a_1}\mathfrak{G}_q, \ldots, \mathfrak{t}_{a_k}\mathfrak{G}_q\} \cup \{mathfrakg_\varepsilon\}]
\end{aligned}$$

Using $d_p = d_{p+1}$, from (12) it follows that $\mathfrak{G}_p = \mathfrak{G}_{p+1}$. Using (13), this implies $\mathfrak{G}_p = \mathfrak{G}_{p+s}$ for all $s \in \mathbb{N}$, i.e. (1).

(2). Follows from (1) and $\dim(X_t) = \dim\mathfrak{G}$ (cf. definition 14 (I)). $\Box$.

Propositions 7 and 8 yield the following algorithm for calculating $m = \dim(X_t)$ from conditioned continuation probabilities $P[\bar{a} \mid \bar{b}]$:

1. For $p = 1, 2, \ldots$, compute d_p as follows:

   (a) Let $\bar{b}_1, \ldots, \bar{b}_n$ be an enumeration of $\Sigma^{\leq p}$. Use $\mathfrak{g}_{\bar{b}_1}, \ldots, \mathfrak{g}_{\bar{b}_n}$ to compute M_r as defined in proposition 7, for $r = 1, 2, \ldots$ (note that $\mathfrak{g}_{\bar{b}}\bar{a} = P[\bar{a} \mid \bar{b}]$). For each r, determine the rank rk of M_r.

   (b) If $\operatorname{rk} M_r = \operatorname{rk} M_{r+1}$, return $d_p = \operatorname{rk} M_r$ (justified by prop. 7).

2. If $d_p = d_{p+1}$, return $m = d_p$ (justified by prop. 8).

This algorithm requires matrices of the kind $M = (\mathfrak{g}_{\bar{b}_i}\bar{a}_j)_{ij} = (P[\bar{a}_j \mid \bar{b}_i])_{ij}$ to be computed repeatedly. Since it is solely the rank of these matrices which is of interest, they can be replaced by the matrices $M' = (P[\bar{b}_i\bar{a}_j])_{ij}$, which are easier to handle in practice. Observing that $P[\bar{b}_i\bar{a}_j] = P[\bar{a}_j \mid \bar{b}_i]P[\bar{b}_i]$, it is easy to see that $\operatorname{rk} M = \operatorname{rk} M'$. Matrices of the form M' can be estimated from S simply by counting occurences of subsequences $\bar{b}_i\bar{a}_j$.

Once the process dimension m is established, it is easy to obtain a set of characteristic events $A_1, \ldots, A_m$. A safe, if rather complicated, method would be to replay the (constructive) proof of proposition 1. However, for practical purposes it is much more appropriate to simply randomly select some partition $\Sigma^k = A_1 \dot{\cup} \ldots \dot{\cup} A_m$ of events A_i which occur in S with nonzero probability. The word length k may be chosen minimal under the constraint that $\mid \Sigma^k \mid \geq m$. It is exceedingly likely that the events thus obtained are characteristic ones. A simple test for characteristicity is to compute some matrix fo the form

$$M_{test} = \begin{pmatrix} P[A_1 \mid w_1] & \cdots & P[A_1 \mid w_m] \\ \vdots & & \vdots \\ P[A_m \mid w_1] & \cdots & P[A_m \mid w_m] \end{pmatrix}$$

where the w_i are arbitrary different words from Σ^*. If M_{test} is regular, $A_1, \ldots, A_m$ are characteristic events. The case may occur that M_{test} is not regular although $A_1, \ldots, A_m$ are indeed characteristic events. Again, this is exceedingly unlikely to happen. If however one finds that M_{test} is not regular, one can select some new partition of Σ^k and some new test words w_i and test again. This can be iterated. Since it is certain that characteristic events do exist, this iterated random search is certain to spot characteristic events eventually. Furthermore, since characterisic events abound, and since most test word selections will yield regular test matrices for characteristic events, this iterated random search will come up with an immediate hit almost always.

4.2 Reconstruction of observable operators

Once the process dimension m is established, and once characteristic events $A_1, \ldots, A_m$ are available, it is easy to reconstruct the observable operators τ_a of the interpretable OOM $\mathcal{A}(A_1, \ldots, A_m) = (\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0)$.

Recall that $\tau_{\bar{b}} w_0 = (P[\bar{b}A_1], \ldots, P[\bar{b}A_m])$ (prop. 5(2)). This means that we can estimate from S the vectors $\tau_{\bar{b}} w_0$, which result from an application of $\tau_{\bar{b}}$ on w_0 by putting $\tilde{\tau}_{\bar{b}} w_0 = (\tilde{P}\bar{b}A_1], \ldots, \tilde{P}[\bar{b}A_m])$, where $\tilde{P}[\bar{b}A_m])$ denotes probability estimates gained from counting occurences of $\bar{b}A_m$ in S.

Using this fact, each τ_a can be reconstructed as follows. First select m different words $\bar{b}_1, \ldots, \bar{b}_m \in \Sigma^*$. Then estimate the m vectors $\tau_{\bar{b}_1} w_0, \ldots, \tau_{\bar{b}_m} w_0$ by putting $\tilde{\tau}_{\bar{b}_j} w_0 = (\tilde{P}\bar{b}_j A_1], \ldots, \tilde{P}[\bar{b}_j A_m])$. Collect these vectors as columns in a matrix $\tilde{V}$:

$$\tilde{V} = \begin{pmatrix} \tilde{P}[\bar{b}_1 A_1] & \cdots & \tilde{P}[\bar{b}_m A_1] \\ \vdots & & \vdots \\ \tilde{P}[\bar{b}_1 A_m] & \cdots & \tilde{P}[\bar{b}_m A_m] \end{pmatrix} \tag{14}$$

Test whether $\tilde{V}$ is regular (i.e. whether its determinant is nonzero). If it is not, repeat this construction with different words $\bar{b}_1, \ldots, \bar{b}_m$. Since the A_i are characteristic events, a selection of words exists such that the resulting matrix $\tilde{V}$ is regular; therefore, a systematic search through possible selections of words is guaranteed to eventually yield a regular $\tilde{V}$. In fact, it is exceedingly probable for a random selection of such words that $\tilde{V}$ is regular; therefore, in practice the first attempt will almost always be successful.

Next, construct the matrix $\tilde{W}_a$:

$$\tilde{W}_a = \begin{pmatrix} \tilde{P}[\bar{b}_1 a A_1] & \cdots & \tilde{P}[\bar{b}_m a A_1] \\ \vdots & & \vdots \\ \tilde{P}[\bar{b}_1 a A_m] & \cdots & \tilde{P}[\bar{b}_m a A_m] \end{pmatrix} \tag{15}$$

Now observe that $\tilde{V}$ is an estimate for

$$V = \begin{pmatrix} P[\bar{b}_1 A_1] & \cdots & P[\bar{b}_m A_1] \\ \vdots & & \vdots \\ P[\bar{b}_1 A_m] & \cdots & P[\bar{b}_m A_m] \end{pmatrix} = ((\tau_{\bar{b}_1} w_0)^T, \ldots, (\tau_{\bar{b}_m} w_0)^T),$$

i.e. of a matrix whose colums are the vectors $\tau_{\bar{b}_j} w_0$ ($\cdot^T$ denotes transposes). Analogically, $\tilde{W}_a$ is an estimate for a matrix W_a whose columns are the vectors $\tau_{\bar{b}_j a} w_0 = \tau_a \circ \tau_{\bar{b}_i} w_0$. I.e., the j-th column of W_a is the result of an application of τ_a on the j-th column of V. This implies that $\tau_a V = W_a$. Since V was selected to be regular, this implies $\tau_a = W_a V^{-1}$.

Returning to estimates, this means that we obtain an estimate of τ_a by putting

$$\tilde{\tau}_a = \tilde{W}_a \tilde{V}^{-1} \tag{16}$$

This is the core result of this paper.

4.3 An example

In this subsection, I demonstrate the model induction techniques described above by going through an example.

Consider the HMM $\mathcal{H}$, where $\Sigma = \{a, b\}$, which is specified by the following Markov matrix M and the diagonal matrices O_a, O_b (cf. section 2 in (I) for terminology):

$$M = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & .5 & .5 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad O_a = \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & .5 & \\ 0 & & & .2 \end{pmatrix} \quad O_b = \begin{pmatrix} 0 & & & 0 \\ & 0 & & \\ & & .5 & \\ 0 & & & .8 \end{pmatrix}$$

Figure 5 gives a graphical representation of this HMM.

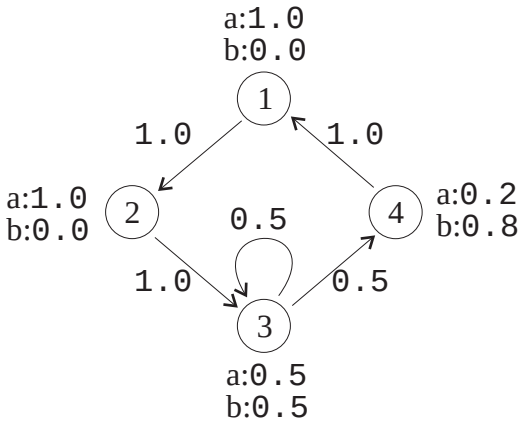

Figure 5: The example HMM. Circles correspond to hidden states, numbers at arrows indicate state transition probabilities. Emission probabilities of events a and b are noted besides states.

This HMM was used to generate a sample sequence S of length 5000.

In the remainder of this subsection, I describe how an estimate $\tilde{\mathcal{A}} = (\mathbb{R}^m, (\tilde{\tau}_a, \tilde{\tau}_b), \tilde{w}_0)$ of an interpretable OOM $\mathcal{A} = (\mathbb{R}^m, (\tau_a, \tau_b), w_0)$ equivalent to $\mathcal{H}$ is reconstructed from S.

The software package Mathematica was used for implementing the algorithms[1].

The first step is to estimate the dimension m of the process of which S is a sample path. To this end, the algorithm described at the end of subsection 4.1 was used. Recall that this algorithm requires the rank of certain $k \times l$ matrices M_r to be calculated. A problem arises from the fact that only estimates $\tilde{M}_r$ of these matrices are available. Such empirical estimates are "noisy" and are exceedingly likely to have maximal rank (i.e., their rank will be $\min\{k, l\}$). Therefore, even if the "correct" matrix M_r has a non-maximal rank, a straightforward, precise computation of the rank of $\tilde{M}_r$ will most likely return a maximal rank.

This difficulty was circumvented in the following way. If the correct matrix M_r has rank d, then there exist $d \times d$ submatrices whose determinant is nonzero, while all $d' \times d'$ submatrices, where $d' > d$, have zero determinant. Assuming that $\tilde{M}_r$ is actually M_r plus some noise, we would expect all $d' \times d'$ submatrices of $\tilde{M}_r$ to have *nearly* zero determinants, while some $d \times d$ submatrices exist whose determinant is markedly different from zero.

This phenomenon was exploited to determine the "actual" rank of the estimated matrices $\tilde{M}_r$. There are many ways how this basic idea can be put to practice. I implemented a somewhat "quick and dirty" rank estimation algorithm, which was largely dictated by the non-availability of advanced statistical linear algebra procedures in my copy of Mathematica. Given a $k \times l$-matrix, where (say) $k \geq l$, this algorithm first randomly selected 300 submatrices of size $l \times l$, and calculated their determinant. If some determinant was found which differed from zero by more than .005, this was taken to be an instance of a "truly" non-zero determinant, and l was returned as an estimate of the rank of M_r. If no such determinant was found, 300 submatrices of size $l - 1 \times l - 1$ were investigated, etc., until at some size $l - x \times l - x$, a submatrix was encountered that met the .005 criterion.

Using this subprocedure for estimating ranks of noisy matrices, the algorithm from subsection 4.1 returned $m = 4$, which is the correct value.

The next step was to select four characteristic events. I arbitrarily opted for the simplest possible choice, wich is $A_1, A_2, A_3, A_4 = \{aa\}, \{ab\}, \{ba\}, \{bb\}$, blindly relying on the almost certain chance that these events indeed would be characteristic (it later turned out that they were).

According to proposition 5(1), the invariant vector w_0 can be estimated from S by putting $\tilde{w}_0 = (\tilde{P}[aa], \tilde{P}[ab], \tilde{P}[ba], \tilde{P}[bb])$. This yields

[1] The algorithms and the Mathematica "notebooks" containing the computations can be fetched from my webpages at http://www.gmd.de/People/Herbert.Jaeger/Resources.html

$$\tilde{w} = (.409, .231, .231, .129) \tag{17}$$

For estimating the observable operators, I calculated matrices $\tilde{V}, \tilde{W}_a, \tilde{W}_b$ according to (14) and (15). For the words $\bar{b}_1, \ldots, \bar{b}_4$ required according to that procedure, I arbitrarily chose aa, ab, ba, bb. I.e., I calculated the matrices

$$\tilde{V} = \begin{pmatrix} \tilde{P}[aaaa] & \tilde{P}[abaa] & \tilde{P}[baaa] & \tilde{P}[bbaa] \\ \tilde{P}[aaab] & \tilde{P}[abab] & \tilde{P}[baab] & \tilde{P}[bbab] \\ \tilde{P}[aaba] & \tilde{P}[abba] & \tilde{P}[baba] & \tilde{P}[bbba] \\ \tilde{P}[aabb] & \tilde{P}[abbb] & \tilde{P}[babb] & \tilde{P}[bbbb] \end{pmatrix}$$

and

$$\tilde{W}_a = \begin{pmatrix} \tilde{P}[aaaaa] & \tilde{P}[abaaa] & \tilde{P}[baaaa] & \tilde{P}[bbaaa] \\ \tilde{P}[aaaab] & \tilde{P}[abaab] & \tilde{P}[baaab] & \tilde{P}[bbaab] \\ \tilde{P}[aaaba] & \tilde{P}[ababa] & \tilde{P}[baaba] & \tilde{P}[bbaba] \\ \tilde{P}[aaabb] & \tilde{P}[ababb] & \tilde{P}[baabb] & \tilde{P}[bbabb] \end{pmatrix}$$

$$\tilde{W}_b = \begin{pmatrix} \tilde{P}[aabaa] & \tilde{P}[abbaa] & \tilde{P}[babaa] & \tilde{P}[bbbaa] \\ \tilde{P}[aabab] & \tilde{P}[abbab] & \tilde{P}[babab] & \tilde{P}[bbbab] \\ \tilde{P}[aabba] & \tilde{P}[abbba] & \tilde{P}[babba] & \tilde{P}[bbbba] \\ \tilde{P}[aabbb] & \tilde{P}[abbbb] & \tilde{P}[babbb] & \tilde{P}[bbbbb] \end{pmatrix}.$$

Note that $\tilde{V}$ is regular, which in retrospect justifies the arbitrary selection of events $\{aa\}, \{ab\}, \{ba\}, \{bb\}$ (which now turn out to be, indeed, characteristic events) and "test words" aa, ab, ba, bb.

$\tilde{V}$ and $\tilde{W}_a, \tilde{W}_b$ yield the following estimates $\tilde{\tau}_a = \tilde{W}_a \tilde{V}^{-1}, \tilde{\tau}_b = \tilde{W}_b \tilde{V}^{-1}$:

$$\tilde{\tau}_a = \begin{pmatrix} .535 & -.137 & .030 & .141 \\ .465 & .137 & -.029 & -.145 \\ -.027 & .391 & .112 & .237 \\ .027 & .608 & -.112 & -.237 \end{pmatrix} \tag{18}$$

$$\tilde{\tau}_b = \begin{pmatrix} .006 & -.018 & 1.018 & -.283 \\ -.006 & .018 & -.018 & .283 \\ -.032 & .017 & .070 & .680 \\ .032 & -.017 & -.070 & .320 \end{pmatrix}$$

This finishes the reconstruction. We have obtained $\tilde{\mathcal{A}}(\{aa\}, \{ab\}, \{ba\}, \{bb\}) = (\mathbb{R}^4, (\tilde{\tau}_a, \tilde{\tau}_b), \tilde{w}_0)$, where the observable operators and the invariant vector take the values given in (17) and (18).

How "good" is this estimate? For a first impression, we compare $\tilde{\mathcal{A}}(\{aa\},\{ab\},\{ba\},\{bb\})$ with the original interpretable OOM $\mathcal{A}(\{aa\},\{ab\},\{ba\},\{bb\}) = (\mathbb{R}^4,(\tau_a,\tau_b),w_0)$, which can be directly obtained from $\mathcal{H}$ using proposition 3:

$$w_0 = (.410, .230, .230, .130) \tag{19}$$

$$\tau_a = \begin{pmatrix} .500 & -.150 & .125 & .101 \\ .500 & .150 & -.125 & -.101 \\ .000 & .350 & .000 & .400 \\ .000 & .650 & .000 & -.400 \end{pmatrix} \tag{20}$$

$$\tau_b = \begin{pmatrix} .000 & .000 & 1.000 & -.250 \\ .000 & .000 & .000 & .250 \\ .000 & .000 & .000 & .750 \\ .000 & .000 & .000 & .250 \end{pmatrix}$$

The average absolute difference between entries in $\tilde{w}_0$ vs. w_0 is .001, and the average absolute difference between matrix entries in $\tilde{\tau}_a, \tilde{\tau}_b$ vs. τ_a, τ_b is .049.

However, it is not easy to interpret these differences – are they "good" or "not so good"? In order to gain better insight into the quality of the estimates, we can compare the distributions of words in the processes generated by $\tilde{\mathcal{A}}$ vs. $\mathcal{A}$. Let $P_{\tilde{\mathcal{A}}}[\bar{b}]$, $P_{\mathcal{A}}[\bar{b}]$ denote the probability of observing the word $\bar{b}$ (relative to observing any other word of equal length) in the processes generated by $\tilde{\mathcal{A}}$ and $\mathcal{A}$, respectively. Let $D_n(\mathcal{A},\tilde{\mathcal{A}})$ denote the absolute difference of these probabilities, averaged over all words of length n, i.e. put $D_n(\mathcal{A},\tilde{\mathcal{A}}) = \mid \Sigma^n \mid^{-1} \sum_{\bar{b}\in\Sigma^n} \mathrm{abs}(P_{\tilde{\mathcal{A}}}[\bar{b}] - P_{\mathcal{A}}[\bar{b}])$. For $n = 5$ and $n = 10$ we obtain $D_5(\mathcal{A},\tilde{\mathcal{A}}) = .0012$ and $D_{10}(\mathcal{A},\tilde{\mathcal{A}}) = .00021$. Putting these values in relation with the average probabilities of words of length 5 and 10, which (in a two-letter alphabet) are 1/16 and 1/1024, we get relative average deviations of word probabilities in the orignal vs. the estimated OOM of $16D_5(\mathcal{A},\tilde{\mathcal{A}}) = .020$ and $1024D_{10}(\mathcal{A},\tilde{\mathcal{A}}) = .22$, i.e. we find average deviations of 2 % and 22 %, respectively, for word lenghts 5 and 10.

These figures have to be judged on the background of the "imprecision" of S. To what degree is the deviation between $\mathcal{A}$ and $\tilde{\mathcal{A}}$ due to shortcomings of our reconstruction procedure, and to what degree is it caused by the inevitable imprecision of S, whose finite length allows but imprecise estimates of word probabilities?

This is not a very precisely stated question. However, a kind of answer can be given if we look at how much the empirical frequences of words in S deviate

from the correct probabilities according to $\mathcal{A}$. I.e., define $D_n(S,\mathcal{A}) = |\,\Sigma^n\,|^{-1}$ $\sum_{\bar{b}\in\Sigma^n} \mathrm{abs}(\tilde{P}[\bar{b}] - P_{\mathcal{A}}[\bar{b}])$. We obtain $D_5(S,\mathcal{A}) = .0040$ and $D_{10}(S,\mathcal{A}) = .00047$, which means deviations of 6 % and 49 %, computed like above.

Thus, we find that the statistics of the sample sequence S deviate considerably more from the "correct" statistics than the statistics of the reconstructed OOM! In other words, the reconstruction procedure has cleaned away from S some noise. The reason for this to be possible is, of course, that the reconstruction procedure "knows" that the source of S is a 4-dimensional OOM, and thus filters out all noise components which are not compatible with this premise.

All in all, this appears to be a completely satisfying account of the quality of $\tilde{\mathcal{A}}$. It seems possible that the reconstruction procedure can be still improved by exploiting further examples of applications of observable operators to interpretable states. In our procedure, we used only the minimal required number of such examples, namely, m examples per operator. One way to include further examples would be to use several sets of "test words", compute several estimates of observable operators, and then average between them. It is however not self-evident how, exactly, this averaging should best be done. I feel that this kind of improvement will not lead very far: the quality of estimates derived from long test words is likely to be quite poor, since the average frequencies of test words drop exponentially with their length. I will not further pursue questions of this kind here.

A somewhat riddling fact is that matrix entries in the estimated vs. original observable operators (18), (20) on the average differ quite considerably from each other. As noted above, this average difference is about .049. Considering that the average value of matrix entries is 1/8, this implies a relative average deviation of $8 \times .049 = .39$, or 39 %. If we compare this with the deviations of estimated vs. original statistics from above, which were 2 % and 22 %, somehow this looks as if the matrices are more different from each other than the processes they generate.

This riddle can be resolved by a closer inspection of $\tilde{V}$. The determinant of this matrix is $\det \tilde{V} = .000016$, which is a very small number even considering that the entries in $\tilde{V}$ average only about .0625. Intuitively, $\tilde{V}$ is "almost" degenerate, in the sense that it describes a mapping which projects $\mathbb{R}^4$ on a very flat, "almost" 3-D subspace. This implies that small changes in $\tilde{V}$ will lead to large changes in $\tilde{V}^{-1}$, and therefore, in $\tilde{W}_a$ and $\tilde{W}_b$. Conversely, this means that certain relatively large changes in the latter will have only small effect on the resulting process statistics. This explains the small riddle.

The fact that $\tilde{V}$ is almost degenerate is not, in this example, due to an unlucky choice of characteristic events and/or test words. Quite to the contrary, if one tries out some alternative choices, one will observe that typically

the determinants of the resulting $\tilde{V}'$ are even much smaller than the value obtained here.

This makes one supect that the process generated by $\mathcal{A}$ "almost" has dimension 3. Therefore, it should be possible to obtain a good 3-dimensional approximation to $\mathcal{A}$. This is what we shall try next.

The way to get a 3-D OOM which models "most" of S is simply to go through the above procedures once again, but assuming that $m = 3$. Selecting as characteristic events $(A_1, A_2, A_3) = (\{aa\}, \{ab\}, \{ba, bb\})$, and as test words aa, ab, ba, we obtain an estimate $\tilde{\mathcal{A}}'(A_1, A_2, A_3) = (\mathbb{R}^3, (\tilde{\tau}'_a, \tilde{\tau}'_b), w'_0)$, where

$$\tilde{w}'_0 = (.409, .230, .360))$$

and

$$\tilde{\tau}'_a = \begin{pmatrix} .509 & -.116 & .082 \\ .491 & .115 & -.083 \\ .000 & 1.000 & .000 \end{pmatrix} \quad \tilde{\tau}'_b = \begin{pmatrix} .305 & -.259 & .411 \\ -.075 & .074 & .122 \\ -.229 & .184 & .466 \end{pmatrix}.$$

The quality of $\tilde{\mathcal{A}}'$ as an estimate of $\mathcal{A}$ can be checked like above. We find that the probabilities of words of length 5, computed with $\tilde{\mathcal{A}}'$, deviate about 9 % from the correct probabilities. For words of length 10, the deviation rises to 53 %. Considering the deviations of word frequencies in S from the correct ones (6 % and 49 % for the two word lengthes), $\tilde{\mathcal{A}}'$ appears to be a reasonably good approximation to $\mathcal{A}$.

Figure 6 displays the processes generated by $\mathcal{A}$, $\tilde{\mathcal{A}}$, and $\tilde{\mathcal{A}}'$ in the standardized fashion described in the previous section. The events whose posterior probabilities define the axes are $\{aa\}, \{ab\}, \{ba, bb\}$. Note that some points in the diagram belonging to $\tilde{\mathcal{A}}'$ fall outside the triangle, i.e. represent state vectors which cannot be interpreted in terms of probabilities. This implies that $\tilde{\mathcal{A}}'$ actually is not a valid OOM (condition 3 of definition 3 (I) is not satisfied). Destroying the property of being OOM is a possible (although not a necessary) consequence of reconstructing a process in too few dimensions.

4.4 Modeling non-OOM sources with OOMs

In the previous subsections, we demonstrated how (in the limit of infinite-length sample sequences) an OOM can be correctly reconstructed from data produced by itself. However, empirical time series, as encountered e.g. in speech signals or in neural event dynamics, are very unlikely to be produced by an underlying OOM. Therefore, in most practical applications it does not make sense to try finding the "correct" dimension of the process in the first

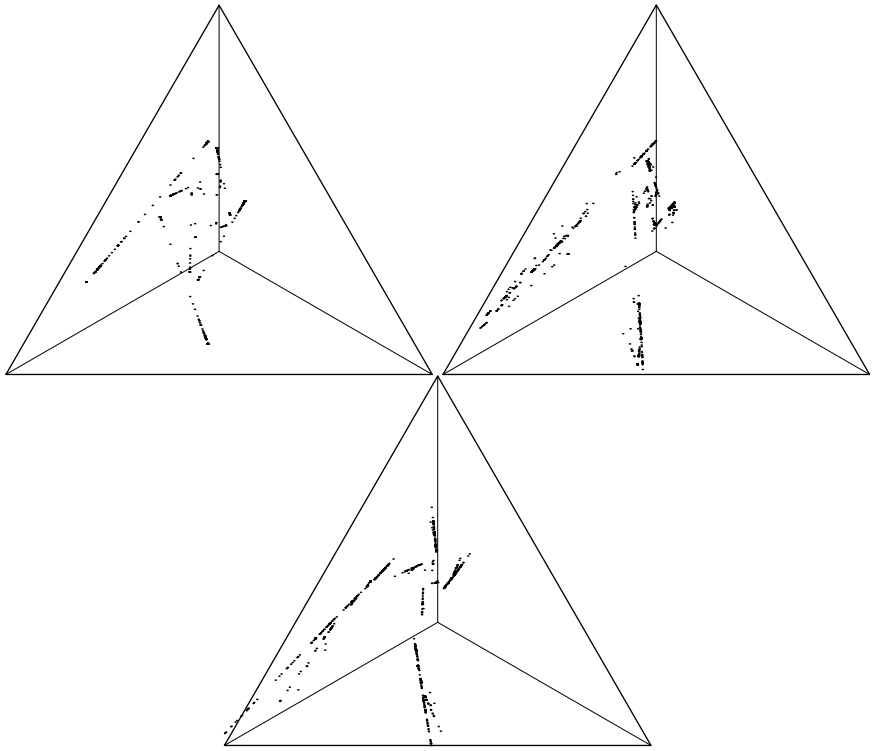

Figure 6: Graphical representations of the processes generated by the original OOM $\mathcal{A}$ (left), its reconstruction $\tilde{\mathcal{A}}$ (right), and the 3-D approximation $\tilde{\mathcal{A}}'$ (bottom).

place. Instead, what one will find is that reconstructing the empirical process with OOMs of increasing dimension will yield increasingly good (but never perfect) approximations.

Given any stationary symbolic process with finite alphabet size, there exist HMMs which perfectly model all cylinder distributions up to any fixed maximal length. In this sense, HMMs can be regarded as universal approximators for symbolic processes. However, in practice "good" HMM approximations to empirical processes can result in large numbers of hidden states (i.e., dimensions). The same holds for OOMs, although due to their greater generality an equal quality of approximation is likely to be obtainable with models of lesser dimension. The current theory of OOMs is, however, not sufficiently developed to allow more specific statements about this important issue.

The reconstruction of an approximate OOM from a sequence is computationally extraordinarily cheap, once one has settled for an OOM dimension.

Reading in a symbol sequence of, say, the length of Proust's *A la recherche du temps perdu*, with a local reading window of, say, length 10, and collecting the statistics of say, 150 characteristic events into $26 + 1$ matrices of size 150×150 (corresponding to the 26 matrices $(\tilde{W}_a)_{a\in\Sigma}$, where Σ is our Roman alphabet, plus the matrix $\tilde{V}$), might take no longer than some 30 seconds on a modern workstation. Inverting $\tilde{V}$ would take another 20 seconds, and the rest (the multiplications $\tilde{W}_a\tilde{V}$) is done in a flash. I.e., we obtain a model containing $26 \times 150 \times 150 \approx .5$ Mio. parameters from a sample sequence of length *pages* $\times$ *lines per page* $\times$ *letters per line* $\approx 4100 \times 35 \times 60 \approx 8.6$ Mio. in less than a minute.

Given this basic computational efficiency, the following strategy for obtaining a useful model of an empirical process seems promising:

1. Select a small OOM dimension m to start with. A reasonable value is $m = |\,\Sigma\,|$.
2. Reconstruct an OOM $\mathcal{A}_m$ of dimension m.
3. Test the quality of $\mathcal{A}_m$, e.g. by comparing word frequencies in S with word probabilities obtained from $\mathcal{A}_m$.
   (a) If the quality is sufficient, stop and return $\mathcal{A}_m$.
   (b) Else
      i. if m is so large that further increasing it would become too costly, stop and return "Process cannot be satisfyingly approximated by OOM given available resources",
      ii. else put $m = m +$ increment and return to 2.

The efficiency of OOM estimations also allows an incremental model update. This is useful e.g. in speech understanding systems for keeping track with a shifting source (caused be changing speakers or changing listening conditions), or in robot navigation where probabilistic internal world models of the robot's environment have to be adapted to changes in the environment. A simple way to obtain OOM models that adaptively follow shifting sources is to continuously update $\tilde{V}$ and the $\tilde{W}_a$'s (e.g. by leaky integration of event frequencies derived from incoming data), and recompute the model OOM when the current model starts to show deficiencies.

5 Discussion

In this article I have furthered the mathematical theory of OOMs by introducing interpretable OOMs, and I have shown how the latter are useful in

practice. In particular, an efficient constructive algorithm for reconstructing OOMs from data has been presented.

The algorithms currently used for estimating HMMs [2] [4] [3] essentially are hill-climbing procedures. They can get trapped in local optima. Being in itself computationally expensive, hill-climbing procedures have to be augmented by costly meta-routines (e.g. simulated annealing or random retries) to cope with local optima. Since hill-climbing is feasible only when the number of parameters to be estimated is not too great, in a typical HMM estimation procedure the actual parameter estimation is preceded by an estimation of model "structure". This amounts to finding out how many hidden states are appropriate, and which of them should be linked by transitions (from the viewpoint of computational resources, this amounts to fixing the greatest possible number of parameters at zero). While there are standard hill-climbing procedures available for the parameter estimation part, finding of an appropriate model structure requires some subtelty. Several techniques have been proposed, and the art of picking the right one and employing it properly requires considerable training.

By contrast, the model induction procedure presented in this article is constructive, and thus avoids the pitfalls of local optima. It is computationally extremely cheap. This allows the reconstruction of models with huge numbers of parameters and renders superfluous a stage of model structure estimation, i.e. a preparatory stage where it is decided which parameters can be harmlessly put to zero. Last but not least, there is a clear and quite elementary mathematical model underlying the algorithm, which makes it simple to understand and allows a routine application.

However, this is only a start. HMM practitioners have developed many extensions of the basic HMM model, e.g. higher-order HMMs or compound HMMs which consist of several specialized HMMs. These augmentations are motivated by the fact that single basic HMMs are often too weak to account for relevant stochastic regularities. It remains to be seen in which cases the greater expressiveness of OOMs renders them applicable where until now augmented versions of HMMs were required. It is likely that OOMs need to be augmented, too, in many cases.

I shall close with propositions for further OOM-related research:

1. Find interesting non-HMM classes of OOMs. Until now, the only models which are provably OOMs are HMMs and the single non-HMM example described in section 5 (I). (For instance, I suspect that whenever an observable operator contains a non-rational rotational component, the OOM is non-HMM.)

2. Generalize OOMs to continuous values and, if possible, continuous

time. (Idea for the latter: for observable operators, take linear differential operators instead of linear operators.)

3. Develop a statistical theory of word frequencies in OOM-generated processes which allows one to judge the goodness of models beyond the ad hoc comparison of word probabilities used in this article.

4. Investigate substructures and projections of OOMs.

I am devoting my present investigations mainly to the fourth topic. The other three fields of work are as yet completely unploughed.

Acknowledgments I feel deeply grateful toward Thomas Christaller for confidence and great support. The results described in this article have been obtained while working under a postdoctoral grant from GMD, Sankt Augustin.

References

[1] H. Jaeger. Observable operator models and conditioned continuation representations. Arbeitspapiere der GMD 1043, GMD, Sankt Augustin, 1997.

[2] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 267–296. Morgan Kaufmann, San Mateo, 1990. Reprinted from Proceedings of the IEEE 77 (2), 257-286 (1989).

[3] P. Smyth, D. Heckerman, and M.I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–270, 1997.

[4] A. Stolcke and S. Omohundro. Hidden Markov model induction by Bayesian model merging. In S.J. Hanson, J.D. Cowan, and Giles. C.L., editors, *Advances in Neural Information Processing Systems*, volume 5, pages 11–18. Morgan Kaufmann, San Mateo, 1993.