AUSTRIAN JOURNAL OF STATISTICS
Volume 36 (2007), Number 1, 41–52

Observable Operator Models

Ilona Spanczér[1]
Dept. of Mathematics, Budapest University of Technology and Economics

Abstract: This paper describes a new approach to model discrete stochastic processes, called observable operator models (OOMs). The OOMs were introduced by Jaeger as a generalization of hidden Markov models (HMMs). The theory of OOMs makes use of both probabilistic and linear algebraic tools, which has an important advantage: using the tools of linear algebra a very simple and efficient learning algorithm can be developed for OOMs. This seems to be better than the known algorithms for HMMs. This learning algorithm is presented in detail in the second part of the article.

Zusammenfassung: Dieser Aufsatz beschreibt eine neue Vorgehensweise um diskrete stochastische Prozesse zu modellieren, so genannte Observable Operator Modelle (OOMs). Derartige OOMs wurden bereits von Jaeger als Verallgemeinerung der Hidden Markov Modelle (HMMs) eingeführt. Die Theorie der OOMs verwendet probabilistische wie auch lineare algebraische Hilfsmittel, was einen ganz wichtigen Vorteil hat: Mit den Werkzeugen der Linearen Algebra kann ein sehr einfacher und effizienter Lern-Algorithmus für OOMs hergeleitet werden. Dieser scheint besser zu sein als die bekannten Algorithmen für HMMs. Der Lern-Algorithmus wird im zweiten Teil dieser Arbeit in aller Genauigkeit präsentiert.

Keywords: Stochastic Process, Learning Algorithm.

1 Introduction

The theory of hidden Markov models (HMMs) was developed in the 1960's (Baum and Petrie, 1966). In the 1980's this model became very popular in applications, first of all in speech recognition. Since then the hidden Markov models proved to be very useful in nanotechnology (Hunter, Jones, Sagar, and Lafontaine, 1995), telecommunication (Shue, Dey, Anderson, and Bruyne, 1999), speech recognition (Huang, Ariki, and Jack, 1990), financial mathematics (Elliott, Malcolm, and Tsoi, 2002) and astronomy (Berger, 1997).

The observable operator models (OOMs) were introduced by Jaeger as a generalization of HMMs (Jaeger, 1997, 2000b). HMMs have a structure of hidden states and emission distributions whereas the theory of OOMs concentrates on the observations themselves. In OOM theory the model trajectory is seen as a sequence of linear operators and not as a sequence of states. This idea leads us to the linear algebra structure of OOMs, which provides efficient methods in estimation and learning. The core of the learning algorithm has a time complexity of $O(N+nm^3)$, where N is the size of the training data, n is the number of distinguishable outcomes and m is the model state space dimension which is the dimension of the vector space on which the observable operators act. Jaeger (2000a) presents a comprehensive study of the OOM learning algorithm.

[1] Research supported by NKFP 2/051/2004

Using OOMs instead of HMMs has advantages and disadvantages. In some cases the hidden states of HMMs can be interpreted in terms of the application domain. HMMs are widely known and have many applications including speech recognition. However, the class of OOMs is richer and combines the theory of linear algebra and stochastics. It seems that the training of OOMs can be done more effectively than that of HMMs. Moreover, the learning algorithm of OOMs is asymptotically correct, while the corresponding EM algorithms are not (Rabiner, 1989; Jaeger, 2000a).

Our purpose is to model text sources by OOMs. For instance, we want to predict the next characters and words in a written text if the past is given. The future conditional probability distributions of characters with respect to the past $P(b|a_0a_1\dots a_k)$ can be estimated from a sample of the text source and using these probabilities we can estimate the dimension of the OOM and the entries of observable operators. A practical application of this method can be found in Kretzschmar (2003). Kretzschmar chose the novel "Emma" written by Jane Austin and divided it into two parts. The first half was used as a training sample and the second half as a test sample. In this example, the dimension of the trained OOM turned out to be 120.

In the literature there are few publicly available comparisons between OOMs and HMMs. Ghizaru (2004) investigated the OOM learning algorithm and the EM algorithm how they performed in the same learning task. In this experiment a speech database was used for training both the HMM and the OOM. Ghizaru found that hidden Markov models were unable to learn the provided data to any useful extent, even though the data seemed to have a good deal of internal structure. The reason of this phenomenon could be that the EM algorithm easily encounters a local minimum for a certain data set and is unable to move away from it. On the other hand, the OOM learning algorithm is not an iterative improvement algorithm, so local minima or slow running times don't induce any problems.

We outline the structure of the remainder of the paper. In Section 2 the main concepts of the OOM theory will be presented. It will turn out that the class of HMMs is a proper subclass of OOMs. In Section 3 the dimension of OOMs and the learning algorithm will be discussed.

2 Hidden Markov Models and Observable Operator Models

2.1 Hidden Markov Models

Definition 1 *The pair (X_n, Y_n) is a hidden Markov process if (X_n) is a homogeneous Markov process with state space $\mathcal{X}$ and the observations Y_n are conditionally independent and given a fixed state s the distribution of Y_n is time invariant, i.e. $\mathbf{P}(Y_n = y|X_n = s) = \mathbf{P}(Y_{n+1} = y|X_{n+1} = s)$. If both the state space $\mathcal{X}$ and the observations Y_n are discrete we have*

$$\mathbf{P}(Y_n = y_n, \dots, Y_0 = y_o|X_n = x_n, \dots, X_0 = x_0) = \prod_{i=0}^{n} \mathbf{P}(Y_i = y_i|X_i = x_i)\,.$$

Assume that $\mathcal{X} = \{s_1, \dots, s_N\}$ and $\mathcal{Y} = \{v_1, \dots, v_M\}$. The state transition probability matrix of the Markov chain is $A = \{a_{ij}\}$, i.e.

$$a_{ij} = \mathbf{P}(X_{n+1} = s_i | X_n = s_j)\,, \qquad 1 \le i, j \le N\,.$$

The observation probability matrix is $B = \{b_j(k)\}$, i.e.

$$b_j(k) = \mathbf{P}(Y_t = v_k | X_t = s_j)\,, \qquad 1 \le j \le N\,, \quad 1 \le k \le M\,.$$

Finally, the initial state distribution of the Markov chain is π, so

$$\pi_i = \mathbf{P}(X_1 = s_i)\,.$$

Therefore, a hidden Markov model is given by the triple $\lambda = (A, B, \pi)$. For the useful applications of HMMs we have to solve three basic problems:

1. *Evaluation problem:* Given the observation sequence $y_1, \dots, y_n$ and the model $\lambda = (A, B, \pi)$, how do we efficiently compute the probability $\mathbf{P}(Y_i = y_i, 1 \le i \le n | \lambda)$?

2. *Specifying of the hidden states:* Given the observation sequence $Y_1, \dots, Y_n$ and the model $\lambda = (A, B, \pi)$, how do we choose a corresponding state sequence $s_1, \dots, s_n$ which best explains the observations?

3. *Parameter estimation:* How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize the probability $\mathbf{P}(Y_i = y_i, 1 \le i \le n | \lambda)$?

For detailed description and solutions of these problems see Rabiner (1989).

2.2 The OOMs as a Generalization of HMMs

The OOMs provide another point of view of of the HMM theory. When we use HMMs we consider a trajectory as a sequence of states whereas using OOM means that a trajectory is seen as a sequence of operators which generate the next outcome from the previous one.

In the followings we describe how a given HMM corresponds to an OOM in the sense that they generate the same observation process.

Let $(X_t)_{t \in \mathbb{N}}$ be a Markov chain with a finite state space $S = \{s_1, s_2, \dots, s_m\}$ which is given by the initial distribution w_0 and the state transition probability matrix M. When the Markov chain is in state s_j at time t, it produces an observable outcome Y_t with time-invariant probability. The set of the outcomes is $\mathcal{O} = \{a_1, a_2, \dots, a_n\}$. The distribution of the outcomes can be characterized by the emission probabilities

$$\mathbf{P}(Y_t = a_i | X_t = s_j)\,.$$

For every $a \in \mathcal{O}$ we define a diagonal matrix O_a which contains the probabilities $\mathbf{P}(Y_t = a | X_t = s_1), \dots, \mathbf{P}(Y_t = a | X_t = s_m)$ in its diagonal.

Figure 1 shows an example for an HMM with two hidden states and two outcomes. The hidden states are $\{s_1, s_2\}$, the outcomes are $\{a, b\}$. The initial state distribution of the

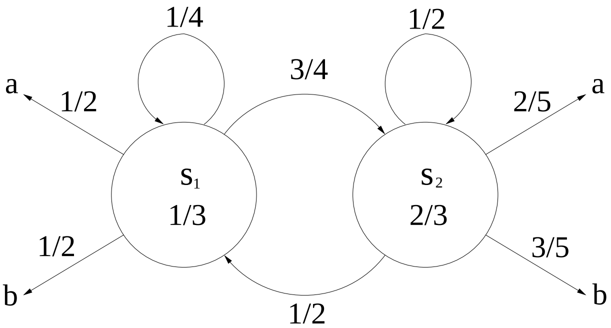

Figure 1: An example for an HMM with transition and emission probabilities.

Markov chain is $(1/3, 2/3)^T$. The state transition probabilities are indicated on the arrows between s_1 and s_2. The emission probabilities can be seen between the hidden states and the outcomes.

This HMM is uniquely characterized by the following matrices:

$$w_0 = \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}, \quad M = \begin{pmatrix} 1/4 & 3/4 \\ 1/2 & 1/2 \end{pmatrix}, \quad O_a = \begin{pmatrix} 1/2 & 0 \\ 0 & 2/5 \end{pmatrix}, \quad O_b = \begin{pmatrix} 1/2 & 0 \\ 0 & 3/5 \end{pmatrix}.$$

If we define the matrix $T_a = M^T O_a$ then we obtain that $\mathbf{P}(Y_0 = a) = \mathbf{1} T_a w_0$ where $\mathbf{1} = (1, \dots, 1)$ denotes the m-dimensional row vector of units. This equation is valid for any observation sequence in the sense that

$$\mathbf{P}(Y_0 = a_{i_0}, Y_1 = a_{i_1}, \dots, Y_k = a_{i_k}) = \mathbf{1} T_{a_{i_k}} T_{a_{i_{k-1}}} \dots T_{a_{i_0}} w_0 .$$

Therefore, the distribution of the process Y_t is specified by the operators T_{a_i} and the vector w_0 in as much as they contain the same information as the matrices M, O_a, and w_0. So, an HMM can be seen as a structure $(\mathbf{R}^m, (T_a)_{a \in \mathcal{O}}, w_0)$, where $\mathbf{R}^m$ is the domain of the operators T_a.

In our example we get the following structure:

$$(\mathbf{R}^m, (T_a, T_b), w_0) = \left(\mathbf{R}^2, \left(\begin{pmatrix} 1/8 & 1/5 \\ 3/8 & 1/5 \end{pmatrix}, \begin{pmatrix} 1/8 & 3/10 \\ 3/8 & 3/10 \end{pmatrix} \right) \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix} \right) .$$

Using this idea we get the definition of OOMs by weaker requirements. We use τ_a instead of T_a and μ instead of M^T.

Definition 2 *A m-dimensional observable operator model (OOM) is a triple $\mathcal{A} = (\mathbf{R}^m, (\tau_a)_{a \in \mathcal{O}}, w_0)$, where $w_0 \in \mathbf{R}^m$ and $\tau_a : \mathbf{R}^m \to \mathbf{R}^m$ are linear operators, satisfying*

1. $\mathbf{1} w_0 = 1$,
2. $\mu = \sum_{a \in \mathcal{O}} \tau_a$ *has column sums equal to 1,*
3. *for all sequences* $a_{i_0} \dots a_{i_k}$ *it holds that* $\mathbf{1} \tau_{a_{i_k}} \dots \tau_{a_{i_0}} w_0 \geq 0$.

These conditions are less stringent because negative entries are allowed in w_0 and μ, which w_0 is a probability vector and μ is a stochastic matrix in the theory of HMMs. The third condition ensures that a non-negative value will be obtained when we compute probabilities in the OOM.

2.3 Probability Clock

It has an obvious question whether there exist OOM which which cannot be modelled by a HMM. The next example is the answer to this question.

Let's take an OOM with outcomes $\mathcal{O} = \{a, b\}$. Let τ_φ be the linear mapping which rotates $\mathbf{R}^3$ by an angle φ around $(1, 0, 0)$ and φ is not a rational multiple of 2π. $\tau_a = \alpha\tau_\varphi$, where $0 < \alpha < 1$. Let τ_b be an operator which projects every vector on the direction of w_0. We have the following observable operators:

$$\tau_a = \alpha \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & \sin\varphi \\ 0 & -\sin\varphi & \cos\varphi \end{pmatrix}, \qquad \tau_b = w_0^T \begin{pmatrix} 1-\alpha \\ 1+\alpha(\sin\varphi - \cos\varphi) \\ 1-\alpha(\sin\varphi + \cos\varphi) \end{pmatrix}^T$$

Note that every occurrence of b "resets" the process to a multiple of the initial vector w_0. Therefore, only the conditional probabilities $\mathbf{P}(Y_t = a | Y_0 = a, Y_1 = a, \ldots, Y_{t-1} = a) = \mathbf{1}{\tau_a}^{t+1} w_0 / \mathbf{1}{\tau_a}^{t} w_0$ are of interest.

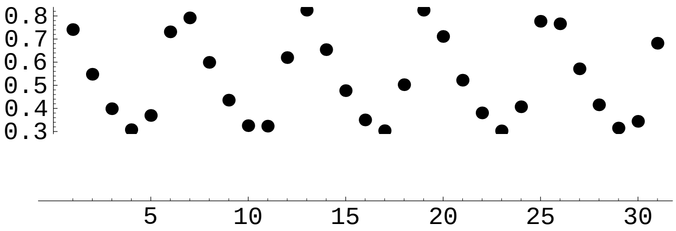

Figure 2: The horizontal axis represents t, the vertical axis represents the conditional probabilities $\mathbf{P}(Y_t = a | Y_0 = a, Y_1 = a, \ldots, Y_{t-1} = a)$ at the probability clock.

This process is called the probability clock. The probability clock cannot be modelled by an HMM, which shows that the class of OOMs is greater than the class of HMMs. This can be proved by means of convex analysis, see Jaeger (2000b). This result provides an equivalent condition for an OOM to be an HMM.

2.4 Generating and Prediction with OOMs

Suppose that the distribution of the process $(Y_t)_{t\in\mathbb{N}}$ is specified by the OOM $\mathcal{A} = (\mathbf{R}^m, (\tau_a)_{a\in\mathcal{O}}, w_0)$. The task is to generate the paths of Y_t. First, we have to generate Y_0. The distribution of Y_0 is known: $\mathbf{P}(Y_0 = a) = \mathbf{1}\tau_a w_0$. At time 0 we make a random decision for the value of Y_0 according to these probabilities. Assume that an initial realization $a_{i_0}, a_{i_1}, \ldots, a_{i_{t-1}}$ of the process is already known. Then we have the following expression for the conditional probabilities

$$\mathbf{P}(Y_t = b | Y_0 = a_{i_0}, Y_1 = a_{i_1}, \ldots, Y_{t-1} = a_{i_{t-1}}) = \frac{\mathbf{1}\tau_b \tau_{a_{i_{t-1}}} \cdots \tau_{a_{i_0}} w_0}{\mathbf{1}\tau_{a_{i_{t-1}}} \cdots \tau_{a_{i_0}} w_0} = \mathbf{1}\tau_b w_t\,, \quad (1)$$

where $w_t = \tau_{a_{t-1}} w_{t-1} / \mathbf{1}\tau_{a_{t-1}} w_{t-1}$. Hence, w_t can be computed from w_{t-1} recursively.

The prediction task is very similar to the generation task. After an initial realization $a_{i_0}, a_{i_1}, \ldots, a_{i_{t-1}}$ we would like to know the probability of the occurrence of b at the next step. This probability is calculated in Equation (1). Similarly, the probability of collective outcomes for multiple time steps can be computed as follows:

$$\mathbf{P}((Y_t,\dots,Y_{t+s}) \in B|(Y_0,\dots,Y_{t-1}) = \overline{a}) = \sum_{\overline{b}\in B} \mathbf{1}\tau_{\overline{b}} w_t = v_B w_t \,.$$

3 Learning OOMs from Data

In this section we follow the work of Jaeger (2000a).

Assume that a sequence $S = a_0 a_1 \dots a_N$ is given which is a path of an unknown stationary OOM $\mathcal{A} = (\mathbf{R}^m, (\tau_a)_{a\in\mathcal{O}}, w_0)$. An $\mathcal{A} = (\mathbf{R}^m, (\tau_a)_{a\in\mathcal{O}}, \omega_0)$ OOM is stationary if $\mu\omega_0 = \omega_0$ where $\mu = \sum_{a\in\mathcal{O}} \tau_a$. The learning task is to find an OOM $\tilde{\mathcal{A}}$ from S which is as close to $\mathcal{A}$ as possible. To this end, we first have to determine the dimension m and then we have to estimate the observable operators τ_a and the initial vector w_0.

To determine the correct dimension of the OOM is very important because if the dimension is too low then the data are underexploited. If the dimension is too high then the data are overfitted and the model is too expensive from a computational point of view.

3.1 Determination of the OOM Dimension

3.1.1 The Dimension of a Stochastic Process

In this section the general notion of a stochastic process will be introduced, which has a close connection with the dimension of the OOM of the process.

To define the dimension of a stochastic process we need some notations. Let $\mathcal{O}^*$ denote the set of all finite sequences whose elements come from $\mathcal{O}$ and the empty sequence (ε). To get a shorter form we write $\mathbf{P}((Y_t,\dots,Y_{t+s}) = \overline{a}|(Y_0,\dots,Y_{t-1}) = \overline{b}) = \mathbf{P}(\overline{a}|\overline{b})$. For every $\overline{b} \in \mathcal{O}^*$ we introduce the conditional probability function:

$$g_{\overline{b}} : \mathcal{O}^*\backslash\{\varepsilon\} \longrightarrow \mathbf{R}$$

$$g_{\overline{b}}(\overline{a}) = \begin{cases} \mathbf{P}(\overline{a}|\overline{b}), & \text{if } \mathbf{P}(\overline{b}) \neq 0 \\ 0, & \text{if } \mathbf{P}(\overline{b}) = 0, \end{cases}$$

and

$$g_\varepsilon(\overline{a}) = \mathbf{P}(\overline{a}|\varepsilon) = \mathbf{P}(\overline{a}) \,.$$

Let $\mathcal{B}$ denote the linear subspace spanned by the set $\{g_{\overline{b}}|\overline{b} \in \mathcal{O}^*\}$ in the linear space of functions $f : \mathcal{O}^*\backslash\{\varepsilon\} \to \mathbf{R}$. Choose $\mathcal{O}_0^* \subseteq \mathcal{O}^*$ so that $\mathcal{B}_0 = \{g_{\overline{b}}|\overline{b} \in \mathcal{O}_0^*\}$ is a basis of $\mathcal{B}$. The next step is to define a linear function for every $a \in \mathcal{O}$:

$$t_a(g_{\overline{b}}) = \mathbf{P}(a|\overline{b}) \cdot g_{\overline{b}a} \,. \tag{2}$$

Equation (2) carries over from basis elements to all $\overline{b} \in \mathcal{O}^*$.
It turns out from the definitions that if $\overline{a} = a_{i_0} a_{i_1} \dots a_{i_k}$ then it holds that

$$t_{\overline{a}} \circ g_\varepsilon = \prod_{l=1}^{k} \mathbf{P}(a_{i_l}|a_{i_0} a_{i_1} \dots a_{i_{l-1}}) \cdot \mathbf{P}(a_{i_0}) \cdot g_{\overline{a}} = \mathbf{P}(\overline{a}) \cdot g_{\overline{a}} \,. \tag{3}$$

Using Equation (3) we can compute probabilities of finite sequences:

$$\mathbf{P}(\overline{a}) = \sum_{c \in \mathcal{O}} \mathbf{P}(\overline{a}) \cdot g_{\overline{a}}(c) = \sum_{c \in \mathcal{O}} (t_{\overline{a}} \circ g_{\varepsilon})(c) \,.$$

These probabilities can be computed from the basis of $\mathcal{B}$ as the following theorem shows.

Theorem 1 *(Jaeger, 2000b) Let $\mathcal{B}_0 = \{g_{\overline{b}} | \overline{b} \in \mathcal{O}_0^*\} = \{g_{\overline{b}_1}, g_{\overline{b}_2}, \ldots, g_{\overline{b}_n}\}$ be a basis of $\mathcal{B}$. Let $\overline{a} = a_{i_0} a_{i_1} \ldots a_{i_k}$ be an initial realization of (Y_t). Let $t_{a_{i_k}} \circ t_{a_{i_{k-1}}} \circ \ldots \circ t_{a_{i_0}} \circ g_{\varepsilon} = \sum_{i=1}^{n} \alpha_i g_{\overline{b}_i}$ be the linear combination of $t_{a_{i_k}} \circ t_{a_{i_{k-1}}} \circ \ldots \circ t_{a_{i_0}} \circ g_{\varepsilon}$ from basis vectors. Then it holds that $\mathbf{P}(\overline{a}) = \sum_{i=1}^{n} \alpha_i$.*

This theorem states that the distribution of the process (Y_t) is uniquely characterized by the observable operators $(t_a)_{a \in \mathcal{O}}$ so the following definition makes sense.

Definition 3 *Let $(Y_t)_{t \in \mathbb{N}}$ be a stochastic process with values in a finite set $\mathcal{O}$. The structure $(\mathcal{B}, (t_a)_{a \in \mathcal{O}}, g_{\varepsilon})$ is called the predictor-space observable operator model of the process. The vector space dimension of $\mathcal{B}$ is called the dimension of the process.*

Consider the elements of $\mathcal{O}^*$ which are the finite realizations of the process Y_t. Let $\mathcal{O}^* = \{o_1, o_2, o_3, \ldots\}$ be ordered lexicographically and let D be an infinite matrix whose entries are the conditional probabilities $d_{ij} = \mathbf{P}(o_i | o_j)$. According to Definition 3 the dimension of the stochastic process Y_t is equal to the maximal number of linearly independent column vectors of the matrix D, which is the rank of D. Hence, determining the dimension of the process Y_t is equivalent to finding the rank of the matrix D.

3.1.2 Determination of the OOM Dimension

The following theorem shows the connection between the dimension of a process and the dimension of the ordinary OOM of this process. This will be the basic idea in the estimation procedure.

Theorem 2 *(Jaeger, 2000b)*

a) *If Y_t is a process with finite dimension m, then an m-dimensional ordinary OOM of this process exists.*

b) *A process Y_t whose distribution is described by a k-dimensional OOM has a dimension $m \leq k$.*

The correct dimension of the OOM is equal to the rank of the matrix D. Therefore, we have to estimate the conditional probabilities $d_{ij} = \mathbf{P}(o_i | o_j)$ from the realization $S = a_0 a_1 \ldots a_N$ then we have to estimate the true rank of the ”noisy” matrix $\tilde{D}$.

Choosing the correct model dimension of the OOM is very difficult. This is also a very hard task in the HMM theory and it was recently solved. In real life usually the task is not to learn models with the true process dimension because empirical processes that are generated by complex physical systems are quite likely very high-dimensional (even infinite-dimensional). However, one can hope that only few of the true dimensions are responsible for most of the stochastic phenomena that appear in the data. Given finite data, then, the question is to determine m such that learning an m-dimensional model reveals m "significant" process dimensions.

We present a method proposed in Jaeger (2000a), which finds the "significant" dimension (see the remark in the next paragraph). Let $\tilde{D}_k$ be a submatrix of $\tilde{D}$ which contains only the entries $\tilde{d}_{ij} = \tilde{\mathbf{P}}(o_i|o_j)$ where $|o_i| \leq k$ and $|o_j| \leq k$. For any matrix A there exists a diagonal matrix $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ and unitary matrices U, R such that $A = R\Sigma U^T$ is the singular value decomposition of A and $\sigma_1, \dots, \sigma_n$ are called the singular values of A. Let $\sigma_1, \dots, \sigma_p$ be the singular values of $\tilde{D}_k$, ordered by decreasing size. Compute the estimated average relative error of the entries of $\tilde{D}_k$:

$$\epsilon = \frac{1}{k^2} \sum_{i,j} \alpha_{ij} \,,$$

where

$$\alpha_{ij} = \begin{cases} \sqrt{\tilde{\mathbf{P}}(o_i|o_j)(1 - \tilde{\mathbf{P}}(o_i|o_j)/N)}/\tilde{\mathbf{P}}(o_i|o_j) \,, & \text{if } \tilde{\mathbf{P}}(o_i|o_j) > 0 \\ 0 \,, & \text{if } \tilde{\mathbf{P}}(o_i|o_j) = 0 \,. \end{cases}$$

Let the numerical rank m_k of $\tilde{D}_k$ be the minimal index m such that $\sigma_m > \epsilon \cdot ||\tilde{D}_k||_\infty$, where the ∞-norm of an $m \times m$ matrix (α_{ij}) is $\max_{1 \leq i,j \leq m} |\alpha_{ij}|$. Repeat the same procedure for $\tilde{D}_{k+1}$ to obtain numerical rank m_{k+1}. If $m_{k+1} = m_k$ the appropriate rank and the appropriate model dimension has been found. If $m_{k+1} > m_k$ continue by increasing k.

Unfortunately finding the "significant" processes dimension is closely related to the unsolved problem in numerical linear algebra of determining the distribution of singular values given the distribution of the matrix entries Jaeger (2000a).

The efficiency of the presented procedure was investigated for the Probability Clock model in Jaeger (2000a, Subsection 9.3). Outcomes of several lengths $N = 300$, 1000, 3000, 10000, 30000 were generated and the model dimension was estimated for these lengths $m = 1, 1, 2, 3, 3$, respectively.

We remark that in this paper several additional simplifications of this estimation method are suggested. Another practical example of this procedure can be found in Kretzschmar (2003, p.29-31). Here, the observed outcomes were generated by a three dimensional HMM. In this case the estimated model dimension coincides with the theoretical one.

Practical application of this method is the learning of the novel "Emma" written by Jane Austin Kretzschmar (2003). It was divided into two parts. The first half was used as a training sample and the second half as a test sample. The text was simplified into letters and blanks (without punctuations), so 27 different symbols were left. A 120-dimensional OOM was trained.

Beyond these examples more statistical results using matrix perturbation theory related to this procedure are presented in this article.

3.2 Estimation of the Operators

For the estimation of the matrices τ_a we introduce the so-called interpretable OOMs which will be very useful because their state space dimensions can be interpreted as probabilities of certain future outcomes.

3.2.1 Interpretable OOMs

Let $(Y_t)_{t\in\mathbf{N}}$ be an m-dimensional stochastic process described by an m-dimensional OOM. For a suitably large k let $\mathcal{O}^k = A_1 \cup \ldots \cup A_m$ be a partition of the set of sequences of length k into m disjoint nonempty subsets. The collective outcomes A_i are called characteristic events if some sequence $\bar{b}_1, \ldots, \bar{b}_m$ exists such that the $m \times m$ matrix $(\mathbf{P}(A_i|\bar{b}_j))_{i,j}$ is nonsingular. Every OOM has characteristic events, see Jaeger (2000a).

Let $\mathcal{A} = (\mathbf{R}^m, (\tau_a)_{a\in\mathcal{O}}, \omega_0)$ be an m-dimensional OOM of the process Y_t. Using the characteristic events $A_1, \ldots, A_m$ we will construct an equivalent, interpretable OOM $\mathcal{A}(A_1, \ldots, A_m)$ which describes the same process and has the property that the m state vector components represent the probabilities of the m characteristic events to occur.

We introduce some definition. We call an OOM $\mathcal{A}$ minimal-dimensional if the dimension of $\mathcal{A}$ is equal to the dimension of the stochastic process which it describes. Given two OOMs $\mathcal{A}$ and $\mathcal{B}$, when are they equivalent in the sense that they describe the same distribution? The answer to this question is given in the following proposition:

Proposition 1 *Two minimal-dimensional OOMs $\mathcal{A} = (\mathbf{R}^m, (\tau_a)_{a\in\mathcal{O}}, \omega_0)$, $\mathcal{B} = (\mathbf{R}^m, (\tau'_a)_{a\in\mathcal{O}}, \omega'_0)$ are equivalent, iff there exists a bijective linear map $\varrho : \mathbf{R}^m \to \mathbf{R}^m$, satisfying the following conditions:*

- $\varrho(\omega_0) = \omega'_0$,
- $\tau'_a = \varrho\tau_a\varrho^{-1}$ *for all* $a \in \mathcal{O}$,
- $\mathbf{1}v = \mathbf{1}\varrho v$ *for all vectors* $v \in \mathbf{R}^m$.

The proof can be found in Jaeger (2000b).

Now we turn to the construction of the equivalent, interpretable OOM. Define $\tau_{A_i} = \sum_{\bar{a}\in A_i} \tau_{\bar{a}}$. Define a mapping $\varrho : \mathbf{R}^m \to \mathbf{R}^m$ by

$$\varrho(x) = (\mathbf{1}\tau_{A_1}x, \ldots, \mathbf{1}\tau_{A_m}x)\,.$$

The mapping ϱ is linear and bijective since the matrix

$$\mathbf{P}(A_i|\bar{b}_j) = \mathbf{1}\tau_{A_i}\frac{\tau_{\bar{b}_j}\omega_0}{\mathbf{1}\tau_{\bar{b}_j}\omega_0} = \mathbf{1}\tau_{A_i}x_j$$

is nonsingular. Furthermore, ϱ preserves component sums of vectors. Hence, we obtain an OOM equivalent to $\mathcal{A}$ by putting

$$\mathcal{A}(A_1, \ldots, A_m) = (\mathbf{R}^m, (\varrho\tau_a\varrho^{-1})_{a\in\mathcal{O}}, \varrho\omega_0) = (\mathbf{R}^m, (\tau'_a)_{a\in\mathcal{O}}, \omega'_0)$$

and the proof as well as the proof of the next key result can be found in Jaeger (2000a).

Proposition 2 *In an interpretable OOM $\mathcal{A}(A_1, \ldots, A_m) = (\mathbf{R}^m, (\tau_a)_{a\in\mathcal{O}}, \omega_0)$, it holds that*

- $\omega_0 = (\mathbf{P}(A_1), \ldots, \mathbf{P}(A_m))$
- $\tau_{\bar{b}}\omega_0 = (\mathbf{P}(\bar{b}A_1), \ldots, \mathbf{P}(\bar{b}A_m))$.

3.2.2 The Learning Algorithm

Assume that a sequence $S = a_0 a_1 \ldots a_N$ is given which is a path of an unknown stationary process Y_t. Furthermore, assume that the dimension of Y_t is known to be m and the characteristic events $A_1, \ldots, A_m$ have already been selected. We would like to construct the interpretable OOM of the process.

In the first step, we estimate w_0. Proposition 2 states that $\omega_0 = (\mathbf{P}(A_1), \ldots, \mathbf{P}(A_m))$. Therefore, a natural estimate of w_0 is $\tilde{w}_0 = (\tilde{\mathbf{P}}(A_1), \ldots, \tilde{\mathbf{P}}(A_m))$, where

$$\tilde{\mathbf{P}}(A_i) = \frac{\text{number of } \bar{a} \in A_i \text{ occuring in } S}{N - k + 1} \,,$$

where k is the length of events A_i.

In the second step we estimate the operators τ_a. According to Proposition 2 for any sequence $\bar{b}_j$ it holds that

$$\tau_a(\tau_{\bar{b}_j}\omega_0) = (\mathbf{P}(\bar{b}_j a A_1), \ldots, \mathbf{P}(\bar{b}_j a A_m)) \,. \tag{4}$$

An m-dimensional linear operator is uniquely determined by the values it takes on m linearly independent vectors and this fact leads us to the estimation of τ_a using (4). We estimate m linearly independent vectors $v_j = \tau_{\bar{b}_j}\omega_0 = (\mathbf{P}(\bar{b}_j A_1), \ldots, \mathbf{P}(\bar{b}_j A_m))$ by putting

$$\tilde{\mathbf{P}}(\bar{b}_j A_i) = \frac{\text{number of } \bar{b}_j \bar{a} \ (\bar{a} \in A_i) \text{ occuring in } S}{N - k - l + 1} \,,$$

where l is the length of $\bar{b}_j$. We also estimate the results $\tau_a v_j = \tau_a(\tau_{\bar{b}_j}\omega_0)$:

$$\tilde{\mathbf{P}}(\bar{b}_j a A_i) = \frac{\text{number of } \bar{b}_j a \bar{a} \ (\bar{a} \in A_i) \text{ occuring in } S}{N - k - l} \,.$$

Thus we obtain estimates $(\tilde{v}_j, \widetilde{(\tau_a v_j)})$ of m argument-value pairs $(v_j, \tau_a v_j)$.

Finally, we can estimate τ_a using the following elementary fact. We collect the vectors $\tilde{v}_j$ as columns in a matrix $\tilde{V}$ and the vectors $\widetilde{(\tau_a v_j)}$ as columns in a matrix $\tilde{W}_a$. Then we obtain $\tilde{\tau}_a = \tilde{W}_a \tilde{V}^{-1}$.

Instead of the sequence $\bar{b}_j$ we can take collective events B_j $(1 \le j \le m)$ of some common length l to construct $\tilde{V}_{ij} = \tilde{\mathbf{P}}(B_j A_i)$ and $\tilde{W}_a = \tilde{\mathbf{P}}(B_j a A_i)$. We will call B_j indicative events. Now, the learning algorithm is the following:

1st step: $\tilde{V}_{ij} = \tilde{\mathbf{P}}(B_j A_i)$,

2nd step: $(\tilde{W}_a)_{ij} = \tilde{\mathbf{P}}(B_j a A_i)$,

3rd step: $\tilde{\tau}_a = \tilde{W}_a \tilde{V}^{-1}$.

As for the selection of characteristic events, we first remark that if we are given infinite training data then we may choose indicative and characteristic events virtually arbitrarily such that almost surely we arrive at a perfect model estimate. On the other hand, in real life we deal with finite training data. In this case the particular choice of indicative and characteristic events is most important as it will determine the quality of the estimated model. Although the learning algorithm is asymptotically correct, the speed of the convergence and possible bounds of model quality for finite training data are dependent on the particular choice of the indicative and characteristic events.

In the current state of the OOM theory one cannot provide a method for optimal choice but there are some helpful rules:

- The characteristic and indicative events should occur in the data with roughly equal frequencies. This minimizes the average relative error in $\tilde{V}$ and $\tilde{W}_a$.
- The inversion of the matrix $\tilde{V}$ should be as insensitive as possible against error in the matrix entries. So we have to be careful with the choice of R and U in the singular value decomposition of $V = RSU^T$.
- The sequences $\overline{a}$ contained in A_i should have high mutual correlation and members of different characteristic events should have low mutual correlation: if $\overline{a}_1, \overline{a}_2 \in A_i$ and $\overline{a}_1$ is likely to occur next, then so is $\overline{a}_2$.
- Indicative events should exhaust $\mathcal{O}^l$ for some l, i.e. $B_1 \cup \ldots \cup B_m = \mathcal{O}^l$. This warrants that when we move the inspection window over S, at every position we get at least one count for $\tilde{V}$, thus we exploit S best.

More selection criteria and explicit constructions can be found in the following papers. Jaeger (2000a) deals with the connection between the unbiased estimation and the selection of indicative and characteristic events. Moreover, it provides a sufficient condition for characteristic events such that using these characteristic events the OOM estimated from data is interpretable. In Kretzschmar (2003) the characteristic events are chosen such that it minimizes the model variance arising from the pseudoinverse operation $\tilde{V}^{-1}$ in the 3rd step of the learning procedure. Jaeger, Zhao, and Kolling (2006) provides a method to minimize the model variance by using *reverse* OOM. In contrast to the paper Kretzschmar (2003), they minimize the variance $\tilde{V}$ itself.

4 Conclusions

In this paper we presented an introduction to the theory of Observable Operator Models (OOM). The OOMs appeared as the generalization of HMMs. OOMs form a deep connection between linear algebra and stochastic processes. Using the tools of linear algebra a very simple and efficient learning algorithm can be developed for OOMs, which seems to be better than the known algorithms for HMMs. It turned out that the class of HMMs is a strict subset of the class of OOMs. As an example, the Probability Clock model was mentioned which is an OOM but cannot be modelled by an HMM.

In the second part of the paper the learning algorithm for OOMs was presented in details. First, we have to determine the dimension of the OOM and second, we have to estimate the observable operators. The learning method seems to be simple but there are some still partially solved problems. For example, how the characteristic and indicative events should be chosen so that we get a computationally simple algorithm that converges fast enough to the "real" OOM in the background.

References

Baum, L., and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, *37*, 1559-1563.

Berger, J. (1997). Some Recent Developments in Bayesian Analysis with Astronomical illustrations. In G. J. Babu and E. Feigelson (Eds.), *Statistical challenges in modern astronomy* (p. 15-39.). Springer.

Elliott, R. J., Malcolm, W. P., and Tsoi, A. (2002). HMM Volatility Estimation. In *Proceedings of the 41th IEEE Conference on Decision & Control, Las Vegas* (pp. TuA 12-6, 398-404).

Ghizaru, A. (2004). *Comparative experimental study of the OOM learning algorithm and of the EM algorithm for HMM* (Tech. Rep.). (http://www.cs.mcgill.ca/ aghiza/projects/cs652/report.pdf)

Huang, X., Ariki, Y., and Jack, M. (1990). *Hidden Markov Models for Speech Recognition*. New York: Columbia University Press.

Hunter, I., Jones, L., Sagar, M., and Lafontaine, P., S.R. andHunter. (1995). Opthalmic microsurgical robot and associated virtual environment. *Computers in Biology and Medicine*, *25*, 173-182.

Jaeger, H. (1997). *Observable operator models II: Interpretable models and model induction* (Tech. Rep. No. 1083). GMD.

Jaeger, H. (2000a). *Discrete-Time, Discrete-Valued Observable Operator Models: a Tutorial* (Tech. Rep.). (webber.physik.uni-freiburg.de/ hon/vorlss02/Literatur/jaeger/)

Jaeger, H. (2000b). Observable Operator Models for Discrete Stochastic Time Series. *Neural Computation*, *12*(6), 1371-1398.

Jaeger, H., Zhao, M., and Kolling, A. (2006). Efficient Estimation of OOMs. In Y. Weiss, B. Schölkopf, and J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 555–562). Cambridge, MA: MIT Press.

Kretzschmar, K. (2003). *Learning symbol sequences with Observable Operator Models* (Tech. Rep. No. 161). GMD. (http://omk.sourceforge.net/files/OomLearn.pdf)

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

Shue, L., Dey, S., Anderson, B., and Bruyne, F. D. (1999). Remarks on Filtering Error due to Quantisation of a 2-state Hidden Markov Model. In *Proceedings of the 40th IEEE Conference on Decision & Control* (pp. FrA05, 4123-4124.).

Author's address:

Ilona Spanczér
Department of Mathematics
Budapest University of Technology and Economics
Műegyetem rkp. 3, H ép. V em. Budapest, 1521
Hungary

E-mail: spanczer@math.bme.hu