Proceedings of the 26th Conference
on Decision and Control
Los Angeles, CA • December 1987

Order Determination for Functions of Finite Markov Chains

Lorenzo Finesso*, and John S. Baras**

Electrical Engineering and Systems Research Center
University of Maryland, College Park MD 20742

1. Introduction

The simplest modeling problem for stochastic processes can be described as follows. We observe a scalar, finite valued, stationary stochastic process $\{Y_t\}$ (FVP) and try to model it as a probabilistic function of a finite Markov chain. The underlying assumption is that the probabilistic mechanism generating $\{Y_t\}$ can be reasonably well explained in the following way. There is an (unobserved) finite stationary Markov chain $\{X_t\}$ such that:

$$(1.1) \qquad P\left[X_{t+1}, Y_{t+1} | X^t, Y^t\right] = P\left[X_{t+1}, Y_{t+1} | X_t\right]$$

Without loss of generality we can assume that the chain $\{X_t\}$ satisfying (1.1) is such that:

$$(1.2) \qquad P\left[X_{t+1}, Y_{t+1} | X_t\right] = P\left[Y_{t+1} | X_{t+1}\right] \; P\left[X_{t+1} | X_t\right]$$

Once $\{X_t\}$ has been found it is easy to describe $\{Y_t\}$ as a deterministic function of a Markov chain. The standard construction is to define the new process $S_t := (X_t, Y_t)$ with values in the cartesian product $\mathcal{S} := \mathcal{X} \times \mathcal{Y}$ of the state spaces of $\{X_t\}$ and $\{Y_t\}$ and the function $f : \mathcal{S} \to \mathcal{Y}$ as $f(i, \varepsilon) = \varepsilon$. It follows from (1.1) that $\{S_t\}$ is a Markov chain and $Y'_t := f(S_t)$ has the same law as $\{Y_t\}$. We can now give the following interpretation: the process we observe, $\{Y_t\}$, is just a component of a process, $\{S_t\}$, which is Markovian. As one would expect the "completion" of $\{Y_t\}$ to a Markov chain $\{S_t\}$ is inherently non unique. Define as *size* of the model the *cardinality* of the state space of $\{X_t\}$. Between all possible models for $\{Y_t\}$ we would like to select one of minimal size *(order)*. The status of Realization Theory for FVP is rather unsatisfactory. The main result here is the characterization of the class of FVP that admit realization [1], [2] but characterization and construction of minimal ralizations are still poorly understood aspects of the theory [3]. The lack of a Realization Theory for FVP makes therefore impossible to follow Kalman's prescription [4]: Identification = Realization + Parametrization. To surmount the impasse and solve the identification problem we must follow the classical approach: select a parametric class of models $\mathcal{P}$ containing the observed process and estimate parameters on the basis of the data $\{y_1, y_2, \ldots, y_T\}$. In section 2 we define $\mathcal{P}$ as a class of probabilistic functions of Markov chains (the order considered as a parameter) satisfying conditions sufficient to guarantee identifiability. In section 3 we show how the order can be consistently estimated from the data. We present the results without proofs here. Further details and proofs can be found in [9]. Let $\{Y_t\}$ be a FVP. Denote by $\mathbf{r}$ its state space, where $\mathbf{r} := (1, 2, \ldots, r)$ and by $\mathbf{r}^*$ the set of finite sequences of elements of $\mathbf{r}$. We assume that r is known a priori (at most this will cost the elimination of a null subset from the state space of the observed process) and denote by $\hat{\mathcal{P}}$ the set of pdf of stationary $\mathbf{r}$-valued sequences.

2. The class of models

Our objective is to obtain a finite dimensional parametric representation for $\hat{\mathcal{P}}$. Finite dimensionality of the parameter space forces us to restrict attention to a subclass of $\hat{\mathcal{P}}$.

Definition 2.1

$$\mathcal{P} := \{p_\theta;\ \theta \in \Theta\}$$

where

$$\Theta := \left\{ \begin{array}{l} \theta := \ (k, A, B, \pi); k \in N,\ A \in \mathbb{R}^{k \times k}, \\ \qquad B \in \mathbb{R}^{k \times r},\ \pi \in \mathbb{R}^k \\ \qquad\qquad A, B, \pi \ \text{stochastic and} \ \pi = \pi A \end{array} \right\}$$

and, defining $\forall \varepsilon \in \mathbf{r}$

$$B_\varepsilon := \text{diag}\ \{b_{1\varepsilon}, b_{2\varepsilon}, \ldots, b_{k\varepsilon}\},$$

the pdf

$$p_\theta : \mathbf{r}^* \to [0, 1] \ \text{ is given by :}$$

$$(2.1)\ \varepsilon_1 \varepsilon_2 \ldots \varepsilon_t \longmapsto p_\theta(\varepsilon_1 \varepsilon_2 \ldots \varepsilon_t) := \pi A B_{\varepsilon_1} A B_{\varepsilon_2} \ldots A B_{\varepsilon_t} e$$

where $e := (1, 1, \ldots 1) \in \mathbb{R}^k$.

Remark 2.2

In terms of the underlying Markov chain $\{X_t\}$ we can make the following identifications:

$$\begin{array}{rcl} k & = & \text{cardinality of the state space of } \{X_t\}. \\ A & = & \text{transition matrix of } \{X_t\} \\ \pi & = & \text{invariant measure of } \{X_t\} \\ b_{j\varepsilon} & = & P[Y_t = \varepsilon | X_t = j] \end{array}$$

For notational convenience we define $M(\varepsilon) := AB_\varepsilon$. The interpretation of $M(\varepsilon)$ is:

$$(2.2) \qquad M(\varepsilon)_{ij} = P\left[Y_{t+1} = \varepsilon,\ X_{t+1} = j | X_t = i\right].$$

As one would expect the class $\mathcal{P}$ just defined is not identifiable.

Assumption 2.3

A and B in definition (2.1) are such that:

i) $a_{ij} > 0 \quad \forall i, j$
ii) $b_{j\varepsilon} > 0 \quad \forall j, \varepsilon$
iii) A is invertible
iv) B_ε has distinct diagonal elements for some $\varepsilon \in \mathbf{r}$

Definition 2.4

$$\begin{array}{rcl} \Theta' & := & \{\theta \in \Theta;\ \text{assumption 2.3 is satisfied}\} \\ \mathcal{P}' & = & \{p_\theta;\ \theta \in \Theta'\} \end{array}$$

The main result of this section is theorem 2.9 which guarantees the identifiability of a subset of Θ' of full (Lebesgue) measure.

* On leave from: LADSEB - C.N.R. Padova-Italy. Research supported in part by an SRC Fellowship.
** Research supported in part by NSF Grant CDR-85-00108, ONR grant NOO14-83-K-0731, and by a grant from UNISYS Corporation.

Definition 2.5

Let $p_\theta \in \mathcal{P}$. Define the set of compound sequence matrices of p_θ relative to $\varepsilon \in \mathbf{r}$ as:

$$C(\theta,\varepsilon) := \left\{M;\ m \in N,\ M \in \mathbb{R}^{m\times m},\ M_{ij} := p_\theta(s_i \varepsilon t_j)\ \ s_i, t_j \in \mathbf{r}^*\right\}.$$

Definition 2.6

Let $p_\theta \in \mathcal{P}$. Define the rank of p_θ relative to $\varepsilon \in \mathbf{r}$ as:

$$r_\theta(\varepsilon) := \sup_{M\in C(\theta,\varepsilon)} (rank\ M)$$

Remark 2.7 If $p_\theta \in \mathcal{P}$ and $\theta = (k, A, B, \pi)$ then:

$$p_\theta(s_i \varepsilon t_j) = \pi M(s_i) M(\varepsilon) M(t_j) e$$

and, from Sylvester's inequality, $r_\theta(\varepsilon) \le k$.

Definition 2.8

Let $p_\theta \in \mathcal{P}$ and $\theta = (k, A, B, \pi)$.

p_θ is said to be regular if $r_\theta(\varepsilon) = k, \quad \forall \varepsilon \in \mathbf{r}$.

θ is regular if $\theta \in \Theta'$ *and* p_θ *is regular.*

Definitions 2.6 and 2.8 are analogous to notions defined in [5] for deterministic functions of Markov chains.

Theorem 2.9

Regular points of Θ' are identifiable modulo permutations of the states.

Remark 2.10

It is interesting to observe that under the *same* hypotheses theorem 2.9 has been proved in [6] in the case of *fixed order* of the model. For identification it is convenient to have compactness of the parameter space. Define, for $\delta > 0$:

$$\Theta'_\delta := \{\theta \in \Theta';\ a_{ij} \ge \delta,\ b_{j\varepsilon} \ge \delta,\ \forall i, j, \varepsilon\}$$

In what follows we will assume that the observed process $\{Y_t\}$ has a regular pdf $p_Y(\cdot) \in \mathcal{P}'$, i.e. there exists a regular point $\theta_0 = (n, A^0, B^0) \in \Theta'$ such that $p_Y(\cdot) = p_{\theta_0}(\cdot)$.

In practice the choice of δ is reduced to the choice of an upper bound K to the order n of the observed process. If $n \le K$ and K is large enough taking $\delta := \frac{1}{K}$ we have $\theta_0 \in \Theta'_\delta$

3. Order Determination

We want to analyze the following question: given observations $(y_1, y_2, \dots y_T)$ with T arbitrarily large, can we determine the order n of $\{Y_t\}$? The answer is yes if $p_{\theta_0} \in \mathcal{P}'_\delta$ and is regular.

It is convenient to decompose Θ'_δ as follows: $\Theta'_\delta = U_{k=1}^K \Theta'_{\delta,k}$ where $\Theta'_{\delta,k}$ is the section of Θ'_δ along k. We identify $\theta = (k, A, B) \in \Theta'_{\delta,k}$ with its projection $\theta = (A, B)$.

Define the random variables on $\mathbf{r}^*$:

$$\forall \theta \in \Theta'_\delta \qquad f_1(\theta, Y(\cdot)) := P_\theta[Y_0]$$
$$\forall \theta \in \Theta'_\delta \text{ and } T \ge 2 \qquad f_T(\theta, Y(\cdot)) := P_\theta[Y_0 | Y_{-T-1}^{-1}]$$

Lemma 3.1[7]

$$f(\theta, Y(\cdot)) := \lim_{T\to\infty} f_T(\theta, Y(\cdot))$$

exists $\forall Y(\cdot) \in \mathbf{r}^*$ and is continuous on $\Theta'_{\delta,k}\ \forall k \in \mathbf{K}$.

Define:

$$h_k(\cdot) : \Theta'_{\delta,k} I\!R, \to h_k(\theta) := E_{\theta_o}[\log f(\theta, Y(\cdot))]$$
$$h_{k,T}(\cdot) : \Theta'_{\delta,k} \to I\!R, h_{k,T}(\theta) := \frac{1}{T} \log P[Y_1^T | \theta]$$
$$h_k : \mathbf{K} \to I\!R\ , h_k := \sup_{\theta\in\Theta_{\delta,k}} h_k(\theta)$$
$$h_{k,T} : \mathbf{K} \times N \to I\!R\ , h_{k,T} := \sup_{\theta\in\Theta_{\delta,k}} h_{k,T}(\theta)$$

Lemma 3.2[7]

i) $h_k(\cdot)$ is continuous on $\Theta'_{\delta,k}\ \forall k \in \mathbf{K}$

ii) $h_k(\cdot) = \lim_{T\to\infty} h_{k,T}(\cdot)\ \forall \theta \in \Theta'_{\delta,k}$ and a.e. P_{θ_0}

Theorem 3.3

i) $h_k(\theta) \le h_n(\theta_0) \quad \forall k \in \mathbf{K},\ \theta \in \Theta'_{\theta,k}$

ii) $h_k(\theta) = h_n(\theta_0)$ iff $k = n$

This is an important consequence of theorem 2.9. Compactness of Θ'_δ allows us to conclude with:

Corollary 3.4 $h_n > h_k \qquad \forall k \in \mathbf{K}, \qquad k \ne n.$

From the abstract point of view this solves the problem of order determination. To compute the sequence h_k we must start from the data and therefore from the sequence of normalized maximum log-likelihoods $h_{k,T}$. In our situation the asymptotic behavior of $h_{k,T}$ is easy to analyze.

Lemma 3.5: $\delta^T \le P\left[Y_1^T | \theta\right] \le (1-\delta)^T \quad \forall \theta \in \Theta'_\delta$

and therefore:

Lemma 3.6: $h_{k,\infty} := \lim_{T\to\infty} h_{k,T}$ exists $\forall k \in \mathbf{K}$

To be able to apply corollary 3.4 we need $h_{k,\infty} = h_k$ and this is true if:

$$(3.1) \qquad \lim_{T\to\infty} \sup_{\theta\in\Theta'_{\delta,k}} h_{k,T}(\theta) = \sup_{\theta\in\Theta'_{\delta,k}} \lim_{T\to\infty} h_{k,T}(\theta)$$

The interchange of limit operations (3.1) is allowed under uniform convergence (for $T \to \infty$) of $h_{k,T}(\theta)$ on $\Theta'_{\delta,k}$. Using the previous results one can verify that actually the sequence $h_{k,T}(\theta)$ satisfies Dini's criterion for uniform convergence.

Observe that (3.1) is verified for every trajectory $Y_t(\omega)$ and from Lemma 3.2 we conclude that $h_{k,T} \longrightarrow h_k$ for almost (P_{θ_o}) every trajectory.

Theorem 3.7

i) $arg\ max_k\{h_{k,T}\}$ is a strongly consistent estimator of n.

ii) For (almost) every trajectory there exists T_o such that:
$n = arg\ max_k\{h_{k,T}\} \qquad (\forall T \ge T_o)$

Remark The practical computation of $h_{k,T}$ (and of the maximizing $\theta = (A, B)$) can be done using the Baum-Eagon algorithm; see [8] for a description and the analysis of its numerical properties. p

References

[1] Heller, A., Math. Sys. Theory, 1, 1967, pp. 197-208.

[2] Picci, G., Springer-Verlag. Lecturer Notes in Economics and Math Sys.Vol. 162 (1978).

[3] Picci, G. van Schuppen, J. H. *On the Weak Finite Stochastic Realization Problem*

[4] Kalman, R. E., Developments in Statistics, Vol. 4, Academic Press, New York, 1983, pp. 97-136.

[5] Gilbert, E. J., Ann. Math. Stat., 30, 1959, pp. 688-697.

[6] Petrie, T., Ann. Math. Stat., 40, 1969, pp. 97-115.

[7] Baum, L., Petrie, T., Ann. Math. Stat., 37, 1966, pp. 1554-1563.

[8] Levinson, S. E., etal, Bell Tech. Jour., 62, 1983, pp. 1035-1074.

[9] Finesso, L., and Baras, J. S., Systems Research Center, TR-87-160, 1987.