KERNEL SMOOTHING TECHNIQUE FOR DIMENSIONALITY REDUCTION IN MARKOV CHAINS

Garajaÿewa Gunça A. & Hofreiter Milan

Institute of Instrumentation and Control Engineering,
Faculty of Mechanical Engineering,
Czech Technical University in Prague.
Technická 4, 166 07, Prague 6, Czech Republic

Abstract: The contribution concerns with the problems related to approximate identification of stochastic systems modelled by Markov chains. Although Markov chains are easily identifiable and adaptable their use is restricted because of extremely large dimension of the sufficient statistic. We introduce the approach, which helps to overcome this drawback. The proposed algorithm for dimensionality reduction in Markov chains is based on kernel smoothing technique. The applicability of the suggested methodology is presented in the Matlab programming environment

Keywords: Markov, chains, reduction, kernel, estimation, prediction.

1. INTRODUCTION

Most processes met in practice are uncertain in the sense that it is not possible to determine exactly the future output values of the process. Markov chain (MC) is an important class of universal black-box models suitable for the description of non-linear stochastic systems. When using the most general parameterisation their estimation as well as control design is simple. High dimensionality is their only but significant drawback.

Large variety of methodologies has been proposed to solve the dimensionality problem in parameterisation of MC (Kárný, *et al.*, 1994; Pavelková, 1994; Hofreiter, 1997a; Hofreiter, 1997b; Valečková, *et al.*, 2001).

In this paper we introduce the approach for dimensionality reduction, which uses the kernel smoothing technique (Hofreiter and Garajayewa, 2000). The proposed algorithm from the measured input-output data estimates parameters of transition probability matrix (TPM), which relate to the values of the so-called *regression vector*. These parameters are preserved and then used for the prediction of output signal. During the prediction time we (almost) always meet the situation, when the *unknown* regression vectors (which were not estimated before) have been occurred. In such a case the algorithm determines from the estimated TPM the set of neighbouring regression vectors, which are in close proximity to the measured unknown one. Then, according to the detailed description of their transition probabilities and distance information, the algorithm evaluates the resulting transition probability for the measured regression vector. By this way, the unknown row (regression vector) of TPM is estimated and the output prediction is determined. The result of this solution is considerable parameter reduction, which at least helps to overcome the mentioned above disadvantage of Markov chains. Furthermore, the proposed algorithm has been applied for one-step-ahead prediction of real ECG (Electro Cardio Gram) signal that is a basic instrument for a diagnosis of heart diseases in cardiology. Achieved results of this application confirm the feasibility of the proposed algorithm.

2. ESTIMATION OF FULLY PARAMETERISED MC

Bayesian estimation and prediction with fully parameterised MC is recalled. It introduces both the notation and the addressed problem.

A sequence of observable random discrete states $y_t \in S \equiv \{1,\ 2,\ \ldots,\ N\}$, $N < \infty$ is measured at discrete time $t = 1,2,\ldots$ on the inspected system. The value of y_{t+1} should be predicted using the related past $D(t)$, i.e. the collection of the states measured up to time t enriched by prior information $D(0)$ (including y_0) $D(t) \equiv \{y_1, y_2, \ldots, y_t\} \cup D(0)$. The relationship of y_{t+1} and $D(t)$ is modelled by the first-order time-invariant parameterised MC $p(y_{t+1} \mid D(t), \Theta) = p(y_{t+1} \mid y_t, \Theta)$. $p(* \mid \bullet)$ denotes probability (probability density function, pdf) of $*$ conditioned on $\bullet$ and Θ is an unknown parameter. Note, that $p(y_{t+1} \mid y_t, \Theta)$ is an (N,N) (transition probability) matrix. The complete information for constructing the desired predictor is contained in the predictive probability $p(y_{t+1} \mid D(t))$. Its evaluation needs the Bayesian parameter estimate (in a wide sense), i.e. the posterior pdf $p(\Theta \mid D(t))$ (Peterka, 1981).

The simplest Bayesian estimator and predictor with MC are obtained if the matrix of transition probabilities is taken as the unknown parameter

$$p(y_t \mid y_{t-1}, \Theta) = \Theta_{y_t|y_{t-1}} = \prod_{j\in Y}\prod_{i\in Y} \Theta_{i|j}^{\delta(y_t,i)\delta(y_{t-1},j)} . \tag{1}$$

Note the alternative form that uses Kronecker symbol δ simplifies derivation of the estimation and prediction results. Recall $\delta(i,j) = 1$ for $i = j$ and $\delta(i,j) = 0$ otherwise. $\Theta = [\Theta_{i|j}]$ is the matrix of unknown transition probabilities. Its entry $\Theta_{i|j}$ is the transition probability from the state j to the state i, $i,j \in Y^{[2]} = (\{1,\ldots,N\},\{1,\ldots,N\})$. The definition of probability restricts entries of $\Theta \in \Theta^* \equiv \left\{\Theta_{i|j} \geq 0,\ \sum_{i\in Y} \Theta_{i|j} = 1,\ \forall j \in Y\right\}$.

The advantageous conjugate prior pdf is considered:

$$p(\Theta) \propto \prod_{i,j\in Y^{[2]}} \left[\Theta_{i|j}^{n_{i|j}(0)-1}\right] \chi(\Theta, \Theta^*) \tag{2}$$

It is specified by $n_{i|j}(0) > 0$ and the indicator $\chi(\Theta,\Theta^*)$ of the set Θ^* equals to 1 on it and it is zero outside of Θ^*.

Both the estimator and predictor can be expressed in terms of the sufficient statistic $n_{i|j}(t) = \overline{n}_{i|j}(t) + n_{i|j}(0) = \sum_{\tau=1}^{t} \delta(y_\tau, i)\delta(y_{\tau-1}, j) + n_{i|j}(0)$, $i,j \in Y^{[2]} \equiv (Y,Y)$. Note that $\overline{n}_{i|j}(t)$

equals to the number of occurrences $y_\tau = i,\ y_{\tau-1} = j$ observed for $0 < \tau \leq t$. The numbers $n_{i|j}(0)$ can be interpreted as occurrences registered for $t = 0$.

In the Bayesian set-up prior information on Θ has to be quantified by a prior pdf $p(\Theta) \equiv p(\Theta \mid D(0))$. The posterior pdf $p(\Theta \mid D(t))$ is determined by the prior pdf, parameterised model and measured data through the Bayes formula

$$p(\Theta \mid D(t)) = \frac{p(y_t \mid y_{t-1}, \Theta) p(\Theta \mid D(t-1))}{\int_{\Theta^*} p(y_t \mid y_{t-1}, \Theta) p(\Theta \mid D(t-1)) d\Theta} \propto \prod_{\tau=1}^{t} p(y_\tau \mid y_{\tau-1}, \Theta) p(\Theta) \qquad (3)$$

where $\propto$ means equality up to a normalising factor.

Existence of conjugate prior pdf guarantees simplicity of the corresponding **Bayesian estimate**. It has exactly the same functional form

$$p(\Theta \mid D(t)) = \prod_{j \in Y} \left[\frac{\prod_{i \in Y} \Theta_{i|j}^{n_{i|j}(t)-1}}{\mathcal{B}(n_{*|j}(t))} \right] \chi(\Theta, \Theta^*) \qquad (4)$$

where $\mathcal{B}(n_{*|j}(t)) = \dfrac{\prod_{i \in Y} \Gamma(n_{i|j}(t))}{\Gamma(\sum_{i \in Y} n_{i|j}(t))}$ is the multivariate beta function. It is defined with the help of Γ function ($\Gamma(x+1) = x\Gamma(x)$).

Elementary rules for pdfs (Peterka, 1981) provide **Bayesian predictor** of the next state

$$p(y_{t+1} \mid D(t)) = \int_{\Theta^*} p(y_{t+1} \mid y_t, \Theta) p(\Theta \mid D(t)) d\Theta = \frac{n_{y_{t+1}|y_t}(t)}{\sum_{i \in Y} n_{i|y_t}(t)} \qquad (5)$$

The probability that y_{t+1} follows y_t is the relative frequency of this configuration observed in past.

3. KERNEL SMOOTHER FOR REDUCED PARAMETERISATION OF MC

The previous relations for estimation of the parameters and prediction of the output show that the estimation of Markov chain consists just in simple counting. However, as the dimension of the sufficient statistic is extremely large even for medium dimension of a regression model and small cardinality of data-value sets the applicability of Markov chains is restricted.

In real cases, it is not possible to assume, that the transition probability matrix will be known for all possible previous states and current inputs, which define the values of the regression vector and therefore a complete model of the system cannot be obtained in straightforward way. Nevertheless, if the parameters of the transition probability matrix are known for the regression vectors, which are in close proximity to the measured one, then it is possible to estimate unknown parameters of the measured regression vector through a kernel smoother.

Kernel smoother uses an explicitly defined set of local weights defined by the kernel to produce the estimate at each target value. Usually a kernel smoother uses weights that decrease in a smooth fashion as one moves away from the target point (Hastil and Tibshirani, 1997; Härdle, 1990). The weight given to the j-th point in producing the estimate at x_0 is defined by

$$w_{0j} = \frac{c_0}{\lambda} \cdot d\left(\left|\frac{x_0 - x_j}{\lambda}\right|\right) \qquad (6)$$

where $d(l)$ is an even function decreasing in $|\ l\ |$. The parameter λ is the window-width, also known as the bandwidth and the constant c_0 is usually chosen so that the weights sum to unity. Epanechnikov kernel $d(l)$ ranks among popular kernels (Epanechnikov, 1969). Recall $d(l) = 2/4 \cdot (1 - l^2)$ for $|\ l\ | \leq 1$ and $d(l) = 0$ otherwise.

We may use this technique for estimation unknown parameters $\Theta(r, y)$; $y \in S$ if some parameters $\Theta({}^j z, y)$; $y \in S_y$, ${}^j z \in S_r \equiv \{{}^j z : \left|r - {}^j z\right| < \lambda\}$, $j = 1,\ 2,\ \ldots,\ q$; $q > 0$ are known. In the previous relation was used the following notation:

r is the for the first time observed value of the regression vector,

q is the number of the known rows of the transition probability matrix Θ with the corresponding values of the regression vector in the set S_r,

${}^j z$ is the j-th value of the regression vector,

$|\ v\ |$ denotes the absolute value of v,

λ is chosen bandwidth,

In such a case, the suggested algorithm estimates the parameters $\Theta(r, y)$; $y \in S$ according to the following relation

$$\Theta(r, y) = \sum_{j=1}^{q} w_{r,j} \cdot \Theta({}^j z, y), \quad y \in S \qquad (7)$$

where $w_{r,\,j} = c_0 \cdot \left(1 - \frac{\|\ r\ -\ {}^j z\ \|^2}{\lambda^2}\right)$, c_0 is chosen so that $\sum_{j=1}^{q} w_{r,\,j} = 1$ and $\|\ v\ \|$ denotes Euclidean norm of the vector v.

Described algorithm for output prediction does not require to know all parameters of the transition probability matrix Θ and therefore it preserves only parameters of the transition probability table relating to the values of the regression vector that has occurred by the actual time which radically reduces memory demands.

4. APPLICATION

Described in the section 3 dimensionality reduction methodology using kernel smoothing technique was applied for one-step ahead prediction of real ECG (Electro Cardio Gram) signal that is a basic instrument for a diagnosis of heart diseases in cardiology. In Fig. 1 is shown the fragment of ECG-output signal, which was used to illustrate the application of the suggested methodology. Sampling interval was 0.003 s

Our task was to estimate Markov chain model and predict the output signal. Then, after applying the algorithm for parameter reduction in Markov chains, make a comparison of results and show the improvement of the prediction.

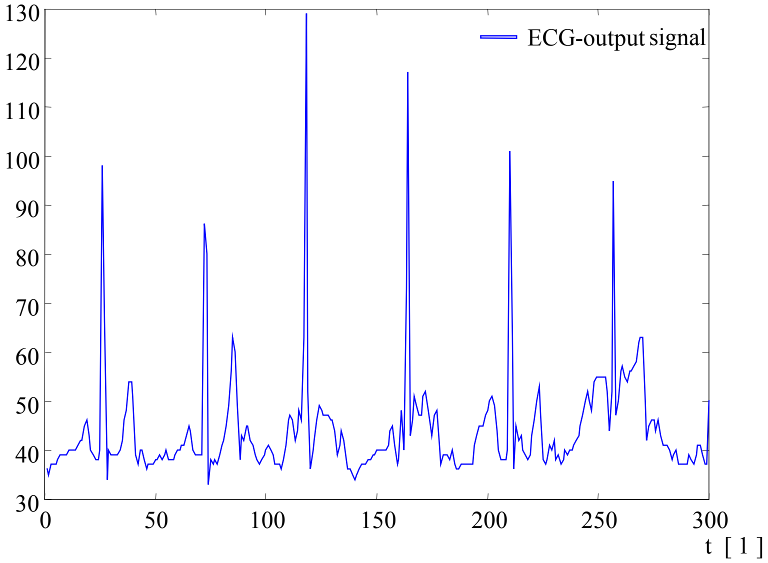

Fig. 1 Course of ECG output signal; t - discrete time

As we wanted to use Markov chains for modelling and output prediction, the output value interval was divided into 34 parts. The discretised output set is thus $S = \{1, 2, 3, \ldots, 34\}$. The structure of regression vector was determined $z(t) = [\ y_{t-1} \quad y_{t-2}],\ t = 1, 2, \ldots$.

Fig. 2 demonstrates the course of output y_t and the output prediction determined by the expected value Ey_t of the output y_t derived from the Markov model with the regression vector z. Results of this prediction were received before we applied suggested algorithm for the parameter reduction. Number of unknown regression vectors, which have occurred during the output prediction, is seventeen. In the Figure they are marked by symbol "○". It is seen, that existence of unknown regression vectors cause inaccurate output prediction.

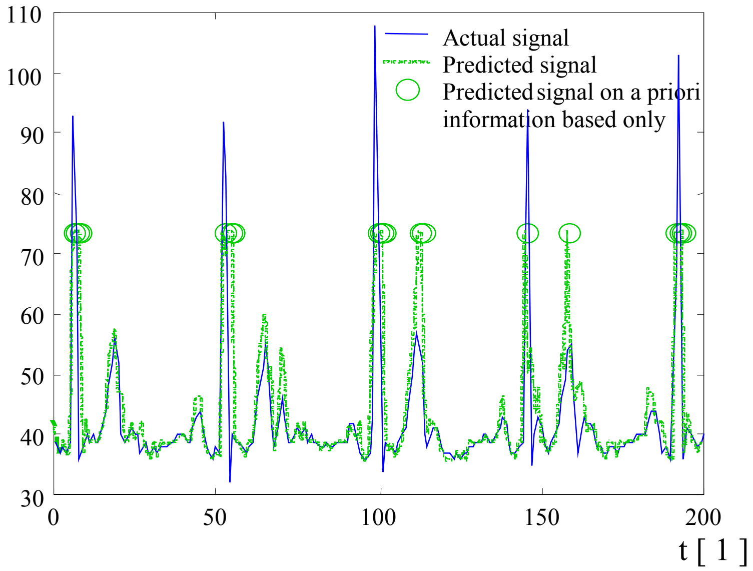

Fig. 2 Course of actual and predicted signals.
Situation before applying algorithm for dimensionality reduction

Improved results of the output prediction were received after we applied new suggested method, which is described in previous section. It is evident from Fig. 3, that mentioned previously algorithm is able to determine the set of neighbouring regression vectors, which are in close proximity to the measured unknown regression vector and according to the detailed description of their transition probabilities and distance information, the algorithm evaluates the resulting transition probability for the measured regression vector. By this way, the unknown rows of transition probability table are estimated and the output prediction is determined and improved.

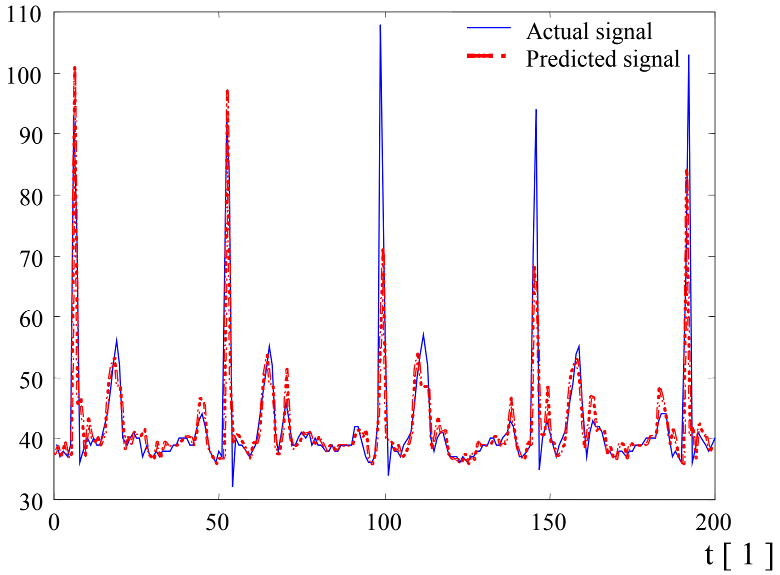

Fig. 3 Course of output and predicted signals.
Situation after applying algorithm for dimensionality reduction

In the Fig. 4 is illustrated output prediction, where the time interval [45, 65] was enlarged on purpose to accentuate the quality of prediction using mentioned algorithm.

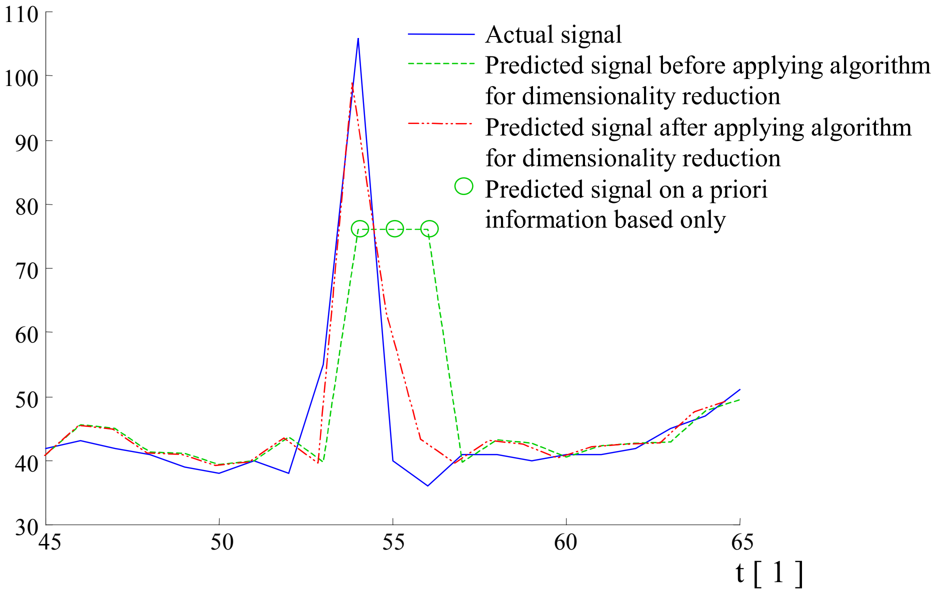

Fig. 4 Enlarged window time interval 45 - 65 s of Fig. 3

To show the computation precision, the criterion of Mean Absolute Deviation (MAD) was chosen:

$$MAD = \frac{1}{t}\sum_{t=1}^{T}\left|y_t - Ey_t\right|, \tag{8}$$

where Ey_t is the output prediction determined by the expected value of the output y_t. Fig. 5 demonstrates the typical course of the MAD for algorithm with (the curve *a*) and without (the curve *b*) considering the neighbouring regression vectors for output prediction.

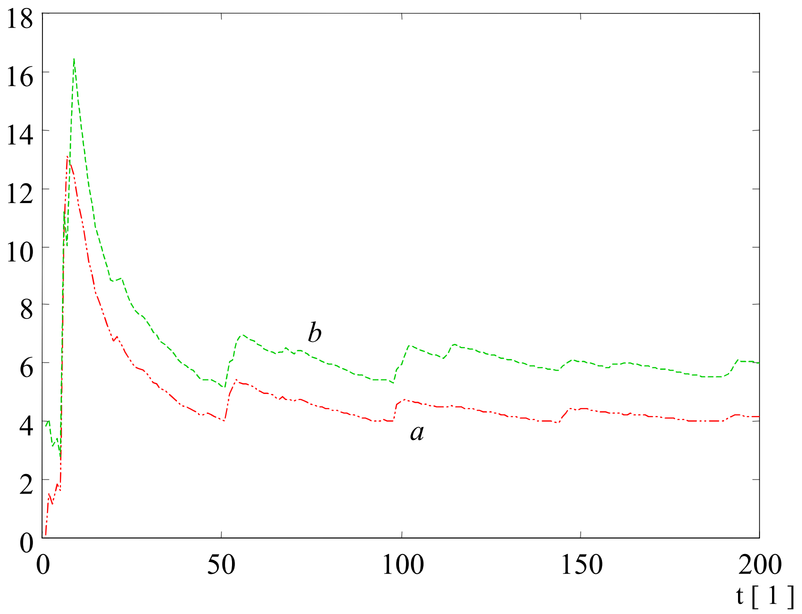

Fig. 5 Course of the MAD for algorithm with (curve a) and without (curve b) considering the neighbouring regression vectors

5 CONCLUSION

This contribution develops an approximate prediction methodology in order to combat the curse of dimensionality inherent in Markov chains. For this purpose Kernel smoothing technique has been used. The proposed algorithm from the measured input-output data estimates parameters of transition probability matrix (TPM), which relate to the values of the regression vector. These parameters are preserved and then used for the prediction of output signal. In case the unknown regression vectors have been occurred, algorithm determines from the estimated TPM the set of neighbouring regression vectors, which are in close proximity to the measured unknown one. Then, according to the detailed description of their transition probabilities and distance information, the algorithm evaluates the resulting transition probability for the measured regression vector. By this way, the unknown rows (regression vectors) of TPM are estimated and the output prediction is determined. Furthermore, the proposed algorithm has been applied for one-step-ahead prediction of real ECG (Electro Cardio Gram) signal that is a basic instrument for a diagnosis of heart diseases in cardiology. Achieved results of this application confirm the feasibility of the proposed algorithm.

Acknowledgement: The presented research has been supported by the:

1. *Research Project No. J04/98: 212200008 Development of Integrated Methods and Tools in Mechanical Engineering.*
2. *Foundation of Josef, Marie and Zdeňka Hlavkov.*
   Vodičkova 17, 110 00 Praha 1, Czech Republic.

REFERENCES

Epanechnikov, V. (1969). Nonparametric estimates of a multivariate probability density. Theory of Probability and Applications, 14, 153 – 8.

Härdle, W. (1990). *Applied nonparametric regression.* Cambridge University Press, United Kingdom, ISBN 0-521-42950-1.

Hastil, T.J. and Tibshirani, R.J. (1997). *Generalized Additive Models.* Chapman & Hall, London, ISBN 0-412-343908.

Hofreiter, M. (1997a). Approximate identification of continuous system. In: *Proceedings of WORKSHOP'97*, pp. 223-224, Czech Technical University, Prague.

Hofreiter, M. (1997b). Identification of Markov Chains Using Discretization with Decomposition. In: *Proceedings of 8th International DAAAM Symposium,* pp. 129–130, Croatia: ICCU, Dubrovnik, ISBN 3-901509-04-6.

Hofreiter, M. and Garajayewa, G.A. (2001). Parameter Reduction in Markov Chains. In: *Summaries 2, ICPR-16*, p. 139, Prague.

Kárný, M., Halousková, A.and Zornigová, L. (1994). On pooling of expert opinions. In: *Preprints SYSID'94*. **Vol. 2**, pp. 477-482, Kopenhagen.

Pavelková, L. (1994). Approximate Identification of Markov Chains. In: *Preprints of the European IEEE Workshop CMP´94,* pp. 335–340, UTIA AVCR, Prague.

Peterka,V. (1981). Bayesian system identification. In: *Trends and Progress in System Identification*. (Eykhoff P. (Ed)), pp. 239-304, Pergamon Press, Oxford.

Valečková, M., Kárný, M. and Sutanto, E.L. (2001). Bayesian M-T Clustering for Reduced Parameterisation of Markov Chains Used for Non-linear Adaptive Elements. *Automatica*, **Vol. 37**, pp. 1071-1078.