**Abstract**

Because of "digitalization" of the real world the discrete-valued time series are necessary to model data sets in many applications, e.g. genetic analysis, information protection, forecasting of COVID-19 dynamics.

An universal model for discrete-valued time series is the Markov chain of the sufficiently large order $s$, but it leads to exponentially increasing number of model parameters. To avoid this "curse of dimensionality" we propose to use the parsimonious models [1] to construct algorithms for statistical analysis (parameters estimation, hypotheses testing, forecasting).

In this talk we present two main approaches to construction of parsimonious models and also results of statistical analysis for the following parsimonious models: Markov chain of order $s$ with $r$ partial connections MC$(s, r)$ and its generalization; MTD-model; Binomial conditionally nonlinear autoregression of order $s$ BiCNAR$(s)$, its special binary case and modifications; Poisson conditionally nonlinear autoregression of order $s$ and its modifications.

Theoretical results are illustrated by computer experiments on simulated and real statistical data.

*Keywords:* discrete-valued time series, parsimonious model, statistical estimator

# Statistical analysis of discrete-valued time-series based on parsimonious models

Yuriy Kharin, Valeriy Voloshko
Research Institute for Applied Problems of Mathematics and Informatics,
Belarusian State University, Minsk, Belarus

## 1 Two approaches to construction of parsimonious models

Let $\mathrm{A} = \{0, 1, \ldots, N-1\}$ be finite ($N = |\mathrm{A}| < \infty$) or countable ($N = |\mathrm{A}| = \infty$, $\mathrm{A} = \mathbb{N}_0$) state space, and let us consider an universal model for discrete-valued time series $\{x_t \in \mathrm{A}\}_{t\in\mathbb{Z}}$, that is the homogenous Markov chain of some sufficiently large order $s$, determined by the generalized Markov property:

$$\mathsf{P}\{x_t = i_t | x_{t-\tau} = i_{t-\tau}, \tau > 0\} = \mathsf{P}\{x_t = i_t | x_{t-\tau} = i_{t-\tau}, s \geq \tau > 0\} = p_{i_{t-s},\ldots,i_{t-1},i_t}. \qquad (1)$$

Here $s$ is the memory depth, $i_t \in \mathrm{A}$ is the value of the process at the discrete time moment $t \in \mathbb{Z}$, $\mathbf{P} = \left(p_{i_{t-s},\ldots,i_{t-1},i_t}\right)$ is an $(s+1)$-dimensional matrix of one-step transition probabilities. Number $D_{\mathrm{MC}(s)} = N^s(N-1)$ of independent parameters for the $\mathrm{MC}(s)$ model increases exponentially w.r.t. the memory depth $s$, and this "curse of dimensionality" makes it impossible to use the general long-memory Markov model (1) in practice: it needs the data sets and the computation work of size $O\left(N^{s+1}\right)$ to identify the model (1). To avoid the "curse of dimensionality" we propose two approaches to construction of parsimonious ("small-parametric") models of high-order Markov chains determined by small number of parameters $d \ll D_{\mathrm{MC}(s)}$ [1]. These two approaches are based on two ways of construction of parsimonious matrix $\mathbf{P}$:

1) "compression" of the set of different values of elements in matrix $\mathbf{P}$;

2) using generation equations for the conditional probability distribution (1) of the future state $x_t$ subject to its fixed prehistory.

Let us explain **the first approach**. Let $Q = \left(q_{j_1,\ldots,j_r,j_{r+1}}\right)$ be some stochastic $(r+1)$-dimensional matrix, $1 \leq r < s$, $\sum\limits_{j_{r+1}\in\mathrm{A}} q_{j_1,\ldots,j_r,j_{r+1}} \equiv 1$, $0 \leq q_{j_1,\ldots,j_r,j_{r+1}} \leq 1$, $j_1, \ldots, j_{r+1} \in \mathrm{A}$; $\mathrm{B}(\cdot) : \mathrm{A}^s \to \mathrm{A}^r$ be some discrete function. The $(s+1)$-dimensional matrix $\mathbf{P}$ is compressed to the $(r+1)$-dimensional matrix $Q$ by the transform:

$$p_{i_1,\ldots,i_s,i_{s+1}} = q_{\mathrm{B}(i_1,\ldots,i_s),i_{s+1}}. \qquad (2)$$

**The first approach** is represented by the following models: Markov chain of order $s$ with $r$ partial connections $\mathrm{MC}(s,r)$ [3, 4], Markov chain of conditional order $\mathrm{MCCO}(s,r)$ [5], variable length Markov chain [6].

**The second approach** is based on some generation equation for the conditional probability distribution of the future state $x_t \in \mathrm{A}$ subject its fixed prehistory $X_{t-s}^{t-1} = (x_{t-1}, \ldots, x_{t-s})'$:

$$p_{i_1,\ldots,i_s,i_{s+1}} = q_{i_{s+1}}\big(\theta(i_1,\ldots,i_s;a)\big),\ i_1,\ldots,i_{s+1} \in \mathrm{A}, \tag{3}$$

where $\{q_j(\theta) : j \in \mathrm{A}\}$ is some discrete probability distribution on A that is dependent on the parameter $\theta = (\theta_j) \in \Theta \subseteq \mathbb{R}^L$, $\theta(i_1,\ldots,i_s;a)$ is some parametric function a priori known up to some unknown vector parameter $a = (a_k) \in \mathbb{R}^m$. **The second approach** is represented by the following models: Jakobs – Lewis model [7], MTD-model [8], DAR($s$) [9], BCNAR($s$) [10], BiCNAR($s$) [11], PCNAR($s$) [12].

# 2 CNAR-family of parsimonious models

Let us describe here a family of models derived from **the second approach** and results of statistical analysis for these models. We call these models conditionally nonlinear autoregressions (CNAR). Let $\mathcal{E}$ be one-dimensional exponential family of probability distributions on A of the following form:

$$\mathcal{E}_x(\eta) = \exp(h(x) + \eta x - \phi(\eta)),\ x \in \mathrm{A},\ \phi(\eta) = \ln \sum_{x\in\mathrm{A}} \exp(h(x) + \eta x), \tag{4}$$

where the function $h(x) : \mathrm{A} \to \mathbb{R}$ determines the family $\mathcal{E}$, the function $\phi(\eta)$ is called potential [2], $\eta$ is called canonic parameter of the family $\mathcal{E}$. The mean value for a random variable $\xi$ with distribution (4) is called dual parameter of this family [2]:

$$\theta = \mathbf{E}_{\xi\sim\mathcal{E}(\eta)}\{\xi\} = \phi'(\eta).$$

For convenience let us use the two equivalent notations for the same probability distribution from the family $\mathcal{E}$: $\mathcal{E}(\eta) = \mathcal{E}[\theta]$, $\mathcal{E}_x(\eta) = \mathcal{E}_x[\theta]$, $x \in \mathrm{A}$. In other words, we use round brackets for canonic parameter $\eta$ and square brackets for dual parameter $\theta$. The range of the dual parameter: $\{\theta\} = \mathrm{A}^* ::= (0, N-1)$ is the interval for the case of finite state space ($N < \infty$) and the ray $\mathrm{A}^* = \mathbb{R}_+$ for the case of countable state space ($N = \infty$).

We say that time series $x_t \in \mathrm{A}$ is an $\mathcal{E}$-based conditionally nonlinear autoregression of order $s$ ($\mathcal{E}$-CNAR($s$)), if $x_t$ is a Markov chain of order $s$ with the following transition probabilities (1):

$$p_{x_{t-s},\ldots,x_{t-1},x_t} = \mathcal{E}_{x_t}[\theta_{X_{t-s}^{t-1}}],\ \theta_q = F\left(\sum_{j=1}^{m} \mathbf{a}_j \psi_j(q)\right),\ q \in \mathrm{A}^s, \tag{5}$$

where $\psi_j : \mathrm{A}^s \to \mathbb{R}$, $j = 1,\ldots,m$, are $m$ linearly independent base functions, $F : \mathbb{R} \to \mathrm{A}^*$ is a smooth bijection, and $\mathbf{a} = (\mathbf{a}_j)_{j=1}^m \in \mathbb{R}^m$ is the vector of $m \ll D_{\mathrm{MC}(s)}$ unknown model parameters.

Note, that for the finite state space ($N < \infty$) the function $F(\cdot)$ is representable as some smooth CDF multiplied by $(N-1)$. For instance, the Gaussian CDF may be used:

$$F(u) = (N-1)\Phi(u),\ \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-t^2/2} dt,\ u \in \mathbb{R}.$$

**Example 1.** Binomial family $\mathcal{E}[\theta] = \mathrm{Bi}(N-1, \theta/(N-1))$, $N < \infty$, leads to the BiCNAR model [11]. Its special binary case for $N = 2$ is called BCNAR model [10].

**Example 2.** Poisson family $\mathcal{E}[\theta] = \mathrm{Poisson}(\theta)$, $N = \infty$, leads to the PCNAR model [12].

# 3 The family of frequency-based estimators for CNAR-models

Maximum likelihood estimation (MLE) for the unknown parameters of the models is often a highly complicated problem due to many local maximums of the log-probability function and computational complexity of its numerical maximization in the cases when the MLE does not have an explicit form. Let us present here the family of frequencies-based estimators (FBE) for CNAR-models, that are free of these shortcomings, have explicit form, are asymptotically efficient, and fast recursively computable.

Let the time series $x_1, \ldots, x_T \in \mathrm{A}$ of the length $T > m$ be observed, and let

$$\hat{\theta}_q = \frac{\sum_{t=s+1}^{T} x_t \mathbb{1}\left\{X_{t-s}^{t-1} = q\right\}}{\sum_{t=s+1}^{T} \mathbb{1}\left\{X_{t-s}^{t-1} = q\right\}},\ q \in \mathbf{M}_T \subset \mathrm{A}^s, \tag{6}$$

be a sample (frequency-based) estimator for conditional mean $\mathbf{E}\left\{x_t | X_{t-s}^{t-1} = q\right\}$, where $\mathbb{1}\{B\}$ is the indicator function for the event $B$, $\mathbf{M}_T$ is the subset of $s$-prehistories for which the denominator is positive in (6)) and $0 < \hat{\theta}_q < N - 1$. Using (5), let us construct an asymptotically overdetermined system of $|\mathbf{M}_T|$ equations w.r.t. $\mathbf{a}$: $F^{-1}(\hat{\theta}_q) = \sum_{j=1}^{m} \mathbf{a}_j \psi_j(q)$, $q \in \mathbf{M}_T$. The quadratic error function for this system ($H = (H_{q,q'})_{q,q' \in \mathbf{M}_T}$ is positive definite symmetric matrix):

$$W(\mathbf{a}) = \sum_{q,q' \in \mathbf{M}_T} H_{q,q'} \left( F^{-1}(\hat{\theta}_q) - \sum_{j=1}^{m} \mathbf{a}_j \psi_j(q) \right) \left( F^{-1}(\hat{\theta}_{q'}) - \sum_{j=1}^{m} \mathbf{a}_j \psi_j(q') \right)$$

can be minimized w.r.t. $\mathbf{a}$ explicitly, which leads to the FBE:

$$\hat{\mathbf{a}} ::= \arg\min_{\mathbf{a} \in \mathbb{R}^m} W(\mathbf{a}). \tag{7}$$

In addition to the listed advantages, the FBE (7) has less restrictive uniqueness conditions on the function $F(\cdot)$ w.r.t. the MLE.

We have proved consistency, asymptotic normality and asymptotic efficiency of the FBE (7). Theoretical results are illustrated in computer experiments on simulated and real data.

# References


[1] Kharin Yu. (2013). *Robustness in Statistical Forecasting*. Springer, N.Y.

[2] Amari S., Nagaoka H. (2000). *Methods of Information Geometry*. Oxford University Press.

[3] Kharin Yu.S. (2004). Markov chains with $r$-partial connections and their statistical estimators. *Trans. Nat. Acad. Sci. Belarus*. Vol. **48**, No. 1, pp. 40-44.

[4] Kharin Yu.S., Piatlitski A.I. (2007). A Markov chain of order $s$ with $r$ partial connections and statistical inference on its parameters. *Discrete Mathematics and Applications*. Vol. **17**, No. 3, pp. 295-317.

[5] Kharin Yu., Maltsew M. (2014). Markov chain of conditional order: Properties and Statistical Analysis. *Austrian Journal of Statistics*. Vol. **43**, No. 3-4, pp. 205-217.

[6] Buhlmann P., Wyner A.J. (1999). Variable length Markov chains. *The Annals of Statistics*. Vol. **27**, No. 2, pp. 480-513.

[7] Jacobs P.A., Lewis P.A.W. (1978). Discrete time series generated by mixtures I: correlational and runs properties. *Journal of the Royal Statistical Society. Ser. B*. Vol. **40**, No. 1, pp. 94-105.

[8] Raftery A. (1985). A model for high-order Markov chains. *Journal of the RoyalStatistical Society. Ser. B*. Vol. **47**, No. 3, pp. 528-539.

[9] Kharin Yu., Maltsew M. (2017). Statistical analysis of high-order dependencies. *Acta et Commentationes Universitetas Tartuensis de Mathematica*. Vol. **21**, No. 1, pp. 37-45.

[10] Kharin Yu.S., Voloshko V.A., Medved E.A. (2019). Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series. *Mathematical Methods of Statistics*. Vol. **26**, No. 2, pp. 103-118.

[11] Kharin Yu.S., Voloshko V.A. (2019). Binomial conditionally nonlinear autoregressive model of discrete-valued time series and its probabilistic and statistical properties. *Transactions of the Institute of Mathematics of NAS Belarus*. Vol. **26**, No. 1, pp. 95-105.

[12] Kharin Y., Kislach M. (2019). Statistical analysis of Poisson conditionally nonlinear autoregressive time series by frequencies-based estimators. *Proceedings of the 14-th International Conference. Minsk, Belarus "Pattern Recognition and Information Processing"*, pp. 233-236.