Statistical analysis of high-order Markov dependencies

Yuriy Kharin and Michail Maltsew

Belarusian State University, Minsk, Belarus, kharin@bsu.by, maltsew@bsu.by

Keywords: high-order Markov chain, parsimonious model, estimator, test

A universal model for real-world processes with discrete time t, finite state space $A = \{0, 1, \ldots, N-1\}$ and high-order dependence $s \gg 1$ (in genetics, computer networks, financial markets, meteorology and other fields) is the order s homogeneous Markov chain ($MC(s)$) x_t on some probability space $(\Omega, \mathcal{F}, \mathrm{P})$ determined by an $(s+1)$-dimensional matrix of one-step transition probabilities $P = (p_{i_1,\ldots,i_{s+1}})$, $p_{i_1,\ldots,i_{s+1}} = \mathbf{P}\{x_{t+1} = i_{s+1} | x_t = i_s, \ldots, x_{t-s+1} = i_1\}$. Unfortunately, the number of independent parameters for the $MC(s)$ increases exponentially w.r.t. the order s: $D_{MC(s)} = N^s(N-1)$, and we need data of huge size to identify this model.

To avoid this "curse of dimensionality" we propose to use parsimonious (or "small-parametric" [1]) models for $MC(s)$ that are determined by small number of parameters $d \ll D_{MC(s)}$. Three known examples of parsimonious models: the Jacobs-Lewis model with $d_{JL} = N + s - 1$ parameters; the MTD-model proposed by A. Raftery with $d_{MTD} = N^2 + s - 1$ parameters; the variable length Markov chain model proposed by P. Buhlmann.

We give short analysis of previous results and propose two new parsimonious models of $MC(s)$.

Markov chain of order s with r partial connections $MC(s, r)$ is determined by the following small-parametric form of P:

$$p_{i_1,\ldots,i_{s+1}} = q_{i_{m_1},\ldots,i_{m_r},i_{s+1}},\ i_1, \ldots, i_{s+1} \in A,$$

where $r \in \{1, \ldots, s\}$ is the number of connections; $M_r = (m_1, \ldots, m_r)$ is the integer-valued vector with r ordered components $1 = m_1 < m_2 < \cdots < m_r \le s$, called the template of connections; $Q = (q_{i_1,\ldots,i_r,i_{r+1}})$ is a stochastic $(r+1)$-dimensional matrix. We need $N^r(N-1)$ parameters to completely determine $MC(s, r)$.

Markov chain of conditional order $MCCO(s, L)$ is determined by the equation:

$$p_{i_1,\ldots,i_{s+1}} = \sum_{k=0}^{N^L-1} \mathrm{I}\{< i_{s-L+1}, \ldots, i_s >= k\} q^{(m_k)}_{i_{s-s_k+1},i_{s+1}},$$

where $\mathrm{I}\{C\}$ is the indicator of event C; $< i_{s-L+1}, \ldots, i_s >= \sum_{k=s-L+1}^{s} N^{k-s+L-1} i_k$ is the numeric representation of the sequence $(i_{s-L+1}, \ldots, i_s)$, called the base memory fragment of length $L \in \{1, \ldots, s-1\}$; $Q^{(1)}, \ldots, Q^{(M)}$ are $M \in \{1, \ldots, N^L\}$ different stochastic square matrices of the order N: $Q^{(m_k)} = (q^{(m_k)}_{i,j})$, $1 \le m_k \le M$; the value $s_k \in \{L+1, \ldots, s\}$ is called the conditional order. The transition matrix for the $MCCO(s, L)$ is determined by $2(N^L + 1) + MN(N-1)$ parameters.

We present theoretical results on probabilistic properties and statistical inferences for these models and results of computer experiments on simulated and real data.

References

[1] Kharin, Yu. (2013). *Robustness in Statistical Forecasting.* Springer, Heidelberg/Dordrecht/New York/London.