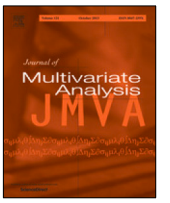
Journal of Multivariate Analysis 188 (2022) 104805

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



Statistical analysis of multivariate discrete-valued time series

Konstantinos Fokianos [a], Roland Fried [b], Yuriy Kharin [c,*], Valeriy Voloshko [c]

[a] Department of Mathematics & Statistics, University of Cyprus, PO BOX 20537, 1678 Nicosia, Cyprus
[b] Department of Statistics, TU Dortmund, 44221 Dortmund, Germany
[c] Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, Minsk, Republic of Belarus

ARTICLE INFO

Article history:
Received 27 June 2021
Received in revised form 31 July 2021
Accepted 31 July 2021
Available online 18 August 2021

AMS 2020 subject classifications:
primary 62M10
secondary 60G10

Keywords:
Autoregression
Categorical time series
Integer-valued time series
Markov chains
Maximum likelihood estimation
Multivariate discrete distributions
Robust estimation

ABSTRACT

This work gives an overview of statistical analysis for some models for multivariate discrete-valued (MDV) time series. We present observation-driven models and models based on higher-order Markov chains. Several extensions are highlighted including non-stationarity, network autoregressions, conditional non-linear autoregressive models, robust estimation, random fields and spatio-temporal models.

1. Introduction

Multivariate time series are sequentially taken observations in time for multiple units. For instance, consider rainfall precipitation in different locations measured every hour or the price of several stocks observed over time. The main investigated problems in this context are modeling, inference and prediction. It is rather common to analyze such data by so-called Vector Autoregressive (VAR) models. They are easily understood by a wide audience, can be fitted by simple statistical procedures and they provide predictions; see the excellent textbooks by Lütkepohl [97] and Tsay [125] for an introduction to VAR models. Though VAR models are well understood from theoretical and methodological points of view and are quite useful for the analysis of continuous-valued data, they become inappropriate when dealing with multivariate time series with integer-valued components (daily number of patient admissions to a hospital, number of transactions of some stocks, absence or presence of a daily characteristic). This contribution aims to give an overview of statistical analysis of multivariate discrete-valued (MDV) time series.

Consider, for instance, the case of multivariate count time series. This is an active research topic as such data can be observed in several applications; see [112] for a medical application, [114] for a financial application and more recently [118] for a marketing application and [96] for an environmental study. A review of current statistical methods for multivariate count time series is given by [43]. In summary, three main approaches are presented: integer

* Corresponding author.
E-mail address: Kharin@bsu.by (Y. Kharin).

https://doi.org/10.1016/j.jmva.2021.104805
0047-259X/

autoregressive (INAR) models (see [93,113,114], among others), parameter-driven models (see [57,67,68,129], among others) and observation-driven models. This class of models is the topic discussed in Section 2.

Section 3 gives an overview of multivariate discrete-valued time series with emphasis on high-order Markov chains and their properties, construction of parsimonious models, and methods and algorithms for their fitting. Such data occur when the observation space A is a discrete (finite or countable) subset of the d-dimensional Euclidean space and the corresponding models are increasingly more important in applications. For example, genetic data (modeling and analysis of genetic time series with $|A| = 4$), economics (number of transactions), sociology (modeling of social behavior), medicine (diagnostics in personalized medicine) and information protection (analysis of binary sequences with $|A| = 2$). Finally, Section 4 lists several possible research directions based on personal research interests.

2. Observation-driven models for multivariate integer-valued time series

The analysis of multivariate integer-valued data is based on Integer Autoregressive Processes (INAR) for count data and on observation-driven and parameter-driven models for quantitative and qualitative data. This section reviews the second approach because such models naturally generalize the standard ARMA theory [13]. Indeed, observation-driven model dynamics evolve past values of the process plus some noise. This is the case of the usual autoregressive models. In particular, observation driven models for count time series have been studied in [26,29,45,47], among others. There is a growing recent literature for the multivariate case; see [2,4,46,60,94,95], for instance. Foundations of these models are based on generalized linear models (GLM) methodology; see [27,28,71,102,128], for more references. Covariates are easily included in observation driven models. Maximum likelihood and/or estimating equations theories provide a solid framework for the analysis of quantitative as well as qualitative time series data. Estimation, diagnostics, model assessment, and forecasting are implemented easily and the computations are carried out by standard softwares. All these methodologies call for the definition of a joint conditional distribution, which, in general, is a challenging problem. Some choices for multivariate count vectors are discussed in the next section.

2.1. Multivariate distribution for count data

In connection to the choice of multivariate count distribution some up-to-date references are given by [43,62,91,133], among others. We denote by $\mathbf{Y} = (Y_1, \ldots, Y_d)^T$ a d-dimensional vector of counts whose components are not necessarily independent.

A natural extension of the univariate Poisson distribution is given by the multivariate Poisson distribution (see [66,89]). Denote by $Y_i = W_i + W$, $i \in \{1, \ldots, d\}$ with $W_i \sim \text{Poisson}(\lambda_i)$ and $W \sim \text{Poisson}(\lambda_0)$ and the random variables $\{W, W_1, \ldots, W_d\}$ are mutually independent. Then, the joint probability mass function (p.m.f) of the vector $\mathbf{Y}$ is given by

$$\mathsf{P}[\mathbf{Y} = \mathbf{y}] = \exp\left(-\sum_{i=0}^{d} \lambda_i\right)\left(\prod_{i=1}^{d} \frac{\lambda_i^{y_i}}{y_i!}\right) \sum_{k=0}^{\min_i y_i} \left(\prod_{i=1}^{d} \binom{y_i}{k}\right) k! \left(\frac{\lambda_0}{\prod_{i=1}^{d} \lambda_i}\right)^k, \ \mathbf{y} = (y_1, \ldots, y_d)^T. \tag{1}$$

It is easy to see that Y_i is Poisson distributed with mean $\lambda_i + \lambda_0$, for $i \in \{1, \ldots, d\}$, and it holds that $\mathsf{Cov}(Y_i, Y_j) = \lambda_0 > 0$. The parameter λ_0 determines all possible pairwise correlations. In addition, the functional form of (1) makes estimation a challenging task. Therefore, regression models based on (1) are suitable for low dimensions.

Mixed models [101] provide a richer class of multivariate count distributions. If the random variables $\{Y_i\}_{i=1}^d$ are assumed to be independent Poisson distributed, conditional on λ_i, with $\mathsf{E}[Y_i|\lambda_i] = \lambda_i$, $i \in \{1, \ldots, d\}$, and the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^T$ is distributed according to some distribution $G(\cdot)$, the mixed Poisson distribution is given by

$$\mathsf{P}[\mathbf{Y} = \mathbf{y}] = \int_{(\mathbb{R}^+)^d} \left[\prod_{i=1}^{n} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}\right] dG(\boldsymbol{\lambda}). \tag{2}$$

It can be shown that (provided that appropriate moments exist)

$$\mathsf{E}[\mathbf{Y}] = \mathsf{E}[\boldsymbol{\lambda}], \quad \mathsf{Cov}[\mathbf{Y}, \mathbf{Y}] = \mathrm{diag}(\mathsf{E}[\boldsymbol{\lambda}]) + \mathsf{Cov}[\boldsymbol{\lambda}, \boldsymbol{\lambda}] = \mathrm{diag}(\mathsf{E}[\mathbf{Y}]) + \mathsf{Cov}[\boldsymbol{\lambda}, \boldsymbol{\lambda}],$$

because of the first equality and $\mathrm{diag}(\mathbf{x})$ denotes a diagonal matrix whose elements are given by a vector $\mathbf{x}$.

Several choices for the mixing distribution $G(\cdot)$ exist. If $G(\cdot)$ is the Dirac distribution placing all its mass at $\boldsymbol{\lambda}$, then $\mathbf{Y}$ reduces to a vector consisting of independent Poisson random variables. Finite mixtures of multivariate Poisson distributions, with application to clustering, have been discussed in [69]. For instance, the multivariate negative-multinomial distribution belongs to the mixed-Poisson class of distributions; see [66, Ch. 36] and [65, Ch. 7.2]. A multivariate negative-multinomial distribution has p.m.f. defined by

$$\mathsf{P}[\mathbf{Y} = \mathbf{y}] = \frac{(n + \sum_{i=1}^{d} y_i)!}{(\prod_{i=1}^{d} y_i!)(n-1)!} p_0^n \prod_{j=1}^{d} p_j^{y_j}, \ \mathbf{y} \in \mathbb{N}^d, \tag{3}$$

where $0 < p_j < 1, j \in \{0, \ldots, d\}$, such that $p_0 = 1 - \sum_{j=1}^{d} p_j$. The parameter n is allowed to take real values because it represents a dispersion parameter. For univariate count time series, overdispersion implies that the marginal variance exceeds the marginal mean but this notion needs further study for multivariate count data. For (3), all pairwise correlations between the components of $\mathbf{Y}$ are positive. Finally this p.m.f. is derived as mixed Poisson model (2) by assuming that, conditionally on a Gamma distributed random variable θ, say $\theta \sim \text{Gamma}(\beta, \beta)$, random variable Y_j is conditionally Poisson distributed with mean $\lambda_j\theta, j \in \{1, \ldots, d\}$. Then

$$\mathsf{P}[\mathbf{Y} = \mathbf{y}] = \frac{\Gamma(\beta + \sum_{i=1}^{d} y_i)}{(\prod_{i=1}^{d} y_i!)\Gamma(\beta)} \left(\frac{\beta}{\beta + \sum_{i=1}^{d} \lambda_i}\right)^{\beta} \prod_{j=1}^{d} \left(\frac{\lambda_j}{\beta + \sum_{i=1}^{d} \lambda_i}\right)^{y_j}, \tag{4}$$

where $\Gamma(\cdot)$ is the Gamma function. In this case, the random variable θ describes common unobserved heterogeneity; see [15]. Obviously p.m.f. (3) and (4) are identical. Finally, (3) can be employed for regression modeling. Given a covariate vector $\mathbf{x}$, a multinomial logistic regression model (see [1]) is employed to link $\mathbf{x}$ with the probabilities $p_j, j \in \{1, \ldots, d+1\}$. Furthermore, the model can be extended to include a log-linear model for n (which can be a positive real number in general, see the previous case) — for more details see [133].

Copula-based construction of multivariate count distributions is another topic of research which deserves special attention; see [62,108] for excellent surveys. Joint distributions can be constructed after employing a copula function (d-dimensional distribution with standard uniform marginals) because of Sklar's theorem [122]. Nelsen [106] provides several examples of copula functions. Though this is an attractive approach for defining properly multivariate distribution functions (see also [64,110,123,130] for related work in the context of integer-valued data, among others), it suffers from two drawbacks. First, for a discrete random vector, with cumulative distribution function F, its p.m.f. involves 2^d finite differences of F, i.e.,

$$\mathsf{P}[\mathbf{Y} = \mathbf{y}] = \sum_{l_1=0,1} \cdots \sum_{l_d=0,1} (-1)^{l_1+\cdots+l_d}\, \mathsf{P}[Y_1 \le y_1 - l_1, \ldots, Y_d \le y_d - l_d],$$

so the likelihood function cannot be obtained easily. Additionally, ties in count data (several zeros, for example) make the copula non-identifiable ([53], in particular pp. 507–508, illustrate the lack of identifiability). Approaches to bypass the second issue have been studied by [32,119]. Recently, [46] introduced a particular data generating process for multivariate count time series analysis which keeps all marginal distributions of the vector $\mathbf{Y}$ to be Poisson distributed and, at the same time, it allows for arbitrary dependence among them. Further studies should be done for this methodology but the main idea is to employ elementary properties of Poisson processes. The algorithm generates i.i.d. random vectors and introduces dependence among their components by a copula structure on the waiting times of the Poisson process. In other words, the copula is imposed on the uniform random variables generating the exponential waiting times. The result gives a sample of multivariate discrete random variables. The methodology can be extended to other discrete marginal distributions provided that they can be generated by continuous inter arrival times.

2.2. Models for multivariate integer-valued time series

We review some regression models for multivariate count time series. We discuss some properties of linear and log-linear models and we give suitable estimating functions to obtain consistent estimators of their parameters. Several models have been proposed for the univariate case, see [27,28,40,71,128] for references.

2.2.1. Linear models for count time series

Denote by $\mathcal{F}_t$ the σ-field generated by all past values of the process $\{\mathbf{Y}_s,\ s \le t\}$. Let $\{\boldsymbol{\lambda}_t = (\lambda_{i,t}),\ i \in \{1, \ldots, d\},\ t \in \mathbb{Z}\}$ be the corresponding d-dimensional intensity process: $\mathsf{E}[\mathbf{Y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{\lambda}_t$. Assuming the data generating process, which is based on copula functions and properties of Poisson process (see [46] for more details), a direct extension of the univariate linear observation-driven model [42,45,59,121] is given by the Vector Integer Autoregressive Conditional Heteroscedastic model, abbreviated by V-INARCH(p) model,

$$Y_{i,t} \mid \mathcal{F}_{t-1} \ \text{ is marginally } \ \text{Poisson}(\lambda_{i,t}), \quad \boldsymbol{\lambda}_t = \boldsymbol{\omega} + \sum_{i=1}^{p} \mathbf{B}_i \mathbf{Y}_{t-i},$$

or the V-INGARCH(p, q) (where "G" stands for Generalized)

$$Y_{i,t} \mid \mathcal{F}_{t-1} \ \text{ is marginally } \ \text{Poisson}(\lambda_{i,t}), \quad \boldsymbol{\lambda}_t = \boldsymbol{\omega} + \sum_{i=1}^{p} \mathbf{B}_i \mathbf{Y}_{t-i} + \sum_{j=1}^{q} \mathbf{A}_j \boldsymbol{\lambda}_{t-j}, \tag{5}$$

where $(\mathbf{A}_j)_{j=1}^{q}$, $(\mathbf{B}_i)_{i=1}^{p}$ are $d \times d$ unknown matrices and all the elements of $\boldsymbol{\omega}$, $(\mathbf{A}_j)_{j=1}^{q}$, $(\mathbf{B}_i)_{i=1}^{p}$ are positive such that $\lambda_{i,t} > 0$ for all i and t.

It is easy to show that (5) can be represented by a VARMA(max$(p,q),q$) process. Set $\boldsymbol{\zeta}_t = \mathbf{Y}_t - \boldsymbol{\lambda}_t$. Assuming first order stationarity of $\mathbf{Y}_t$ and taking expectations on both sides of (5), we see that $\boldsymbol{\mu} = \mathsf{E}[\mathbf{Y}_t] = (\mathbf{I}_d - \sum_{i=1}^{p}\mathbf{A}_i - \sum_{j=1}^{q}\mathbf{B}_j)^{-1}\boldsymbol{\omega}$, provided that $\rho(\sum_{i=1}^{p}\mathbf{A}_i + \sum_{j=1}^{q}\mathbf{B}_j) < 1$. Rearrangement and manipulation of (5) shows that

$$\mathbf{Y}_t - \mathsf{E}[\mathbf{Y}_t] = \sum_{i=1}^{\max(p,q)}\Big(\mathbf{A}_i + \mathbf{B}_i\Big)\Big(\mathbf{Y}_{t-i} - \mathsf{E}[\mathbf{Y}_t]\Big) + \boldsymbol{\zeta}_t - \sum_{j=1}^{q}\mathbf{A}_j\boldsymbol{\zeta}_{t-j}. \tag{6}$$

In the case $p = q = 1$, the corresponding one-sided MA(∞) is given by $\mathbf{Y}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty}\boldsymbol{\Phi}_j\boldsymbol{\zeta}_{t-j}$ with $\boldsymbol{\Phi}_0 = \mathbf{I}_d$ and $\boldsymbol{\Phi}_j = (\mathbf{A}_1 + \mathbf{B}_1)^{j-1}\mathbf{B}_1$, for $j \geq 1$. Hence, for any $h > 0$, we obtain

$$\Gamma_{\mathbf{Y}}(h) \equiv \mathsf{Cov}[\mathbf{Y}_t, \mathbf{Y}_{t+h}] = \sum_{j=0}^{\infty}\big(\mathbf{A}_1 + \mathbf{B}_1\big)^{j+h-1}\mathbf{B}_1\mathsf{E}[\boldsymbol{\Sigma}_t]\mathbf{B}_1^T\big(\mathbf{A}_1^T + \mathbf{B}_1^T\big)^{j+h-1},$$

by using properties of the linear multivariate processes (see [125], for instance).

2.2.2. Log-linear models

A multivariate analogue of the univariate log-linear model proposed by [47] is given by

$$Y_{i,t} \mid \mathcal{F}_{t-1} \ \text{ is marginally } \ \text{Poisson}(\lambda_{i,t}), \quad \boldsymbol{\nu}_t = \boldsymbol{\omega} + \sum_{i=1}^{p}\mathbf{B}_i\log(\mathbf{Y}_{t-i} + \mathbf{1}_d) + \sum_{j=1}^{q}\mathbf{A}_j\boldsymbol{\nu}_{t-j}, \tag{7}$$

where $\boldsymbol{\nu}_t \equiv \log\boldsymbol{\lambda}_t$ is defined componentwise ($\nu_{i,t} = \log\lambda_{i,t}$) and $\mathbf{1}_d$ denotes the d-dimensional vector which consists of ones. Though its analysis is more complicated than that of the linear model, (7) enjoys several advantages which makes it more suitable for applications. First, the coefficients' matrices $(\mathbf{A}_j)_{j=1}^{q}$, $(\mathbf{B}_i)_{i=1}^{p}$ and the vector $\boldsymbol{\omega}$ can take any real values. And more importantly, the log-linear models include covariates in a straightforward way. Indeed, if $\mathbf{Z}_t$ is a covariate vector of dimension d, then the second equation of (7) becomes $\boldsymbol{\nu}_t = \boldsymbol{\omega} + \sum_{i=1}^{p}\mathbf{B}_i\log(\mathbf{Y}_{t-i} + \mathbf{1}_d) + \sum_{j=1}^{q}\mathbf{A}_j\boldsymbol{\nu}_{t-j} + \mathbf{C}\mathbf{Z}_t$ for a $d\times d$ matrix $\mathbf{C}$. Note that for the log-linear model, lagged observations of the response $\mathbf{Y}_t$ are fed into the autoregressive equation for $\boldsymbol{\nu}_t$ via the term $\log(\mathbf{Y}_{t-1} + \mathbf{1}_d)$. This is a one-to-one transformation of $\mathbf{Y}_{t-1}$ which handles zero data values and it maps zeros of $\mathbf{Y}_{t-1}$ into zeros of $\log(\mathbf{Y}_{t-1} + \mathbf{1}_d)$. It is a hard problem to obtain formulas for the mean and autocovariance function of model (7), some approximations are discussed in [46].

Remark 1. Stability conditions for model (5) have been developed by [95] (under the framework of multivariate Poisson distribution (1)) and [46] using the copula construction for the case $p = q = 1$. Recently [30] has improved these conditions considering the copula-based data generating process. Without introducing any further notation, we note that the condition $\rho(\sum_i \mathbf{A}_i + \sum_j \mathbf{B}_j) < 1$ guarantees stability of the process. For the log-linear model (7) the desired conditions are more complicated; see [46] where they consider the case $p = q = 1$ and prove that either $|||\mathbf{A}_1|||_2 + |||\mathbf{B}_1|||_2 < 1$ or $|||\mathbf{A}_1|||_1 + |||\mathbf{B}_1|||_1 < 1$, where $|||\mathbf{A}|||_d = \max_{\|\mathbf{x}\|_d=1}\|\mathbf{A}\mathbf{x}\|_d$, guarantee ergodicity of the process. Related stability conditions are discussed in [30] but they are not directly comparable with those obtained in [46].

2.2.3. Inference

Suppose that $\mathbf{Y}_t$, $t \in \{1,\ldots,n\}$, is an available sample from a count time series and for the sake of presentation assume model (5) for $p = q = 1$. Inference is analogously developed to the case of log-linear model and for $p, q > 1$. Denote by $\boldsymbol{\theta} = (d^T, \text{vec}^T(\mathbf{A}_1), \text{vec}^T(\mathbf{B}_1))$, with $\dim(\boldsymbol{\theta}) = d(1 + 2d)$. Following Fokianos et al. [46], consider the "independence" conditional quasi-likelihood function, given a starting value λ_0,

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n}\prod_{i=1}^{d}\Big\{\frac{\exp(-\lambda_{i,t}(\boldsymbol{\theta}))\lambda_{i,t}^{y_{i,t}}(\boldsymbol{\theta})}{y_{i,t}!}\Big\}.$$

This strong assumption simplifies computation of estimators and their respective standard errors and yields consistency and asymptotic normality of the maximizer. The dependence structure in (5) and (7) is taken into account through the dependence of the likelihood function on the matrices $\mathbf{A}_1$ and $\mathbf{B}_1$; see Fokianos et al. [46] for more information. The quasi log-likelihood function is equal to

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{d}\Big(y_{i,t}\log\lambda_{i,t}(\boldsymbol{\theta}) - \lambda_{i,t}(\boldsymbol{\theta})\Big).$$

We denote by $\widehat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$, the QMLE of $\boldsymbol{\theta}$. The score function is given by

$$S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{d}\Big(\frac{y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta})} - 1\Big)\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{t=1}^{n}\frac{\partial\boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{D}_t^{-1}(\boldsymbol{\theta})\Big(\mathbf{Y}_t - \boldsymbol{\lambda}_t(\boldsymbol{\theta})\Big), \tag{8}$$

where $\partial\boldsymbol{\lambda}_t/\partial\boldsymbol{\theta}^T$ is a $d\times\kappa$ matrix and $\mathbf{D}_t$ is the $d\times d$ diagonal matrix with the ith diagonal element equal to $\lambda_{i,t}(\boldsymbol{\theta})$, $i\in\{1,\dots,p\}$. The Hessian matrix is given by

$$H_n(\boldsymbol{\theta})=\sum_{t=1}^{n}\sum_{i=1}^{p}\frac{y_{i;t}}{\lambda_{i,t}^2(\boldsymbol{\theta})}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}-\sum_{t=1}^{n}\sum_{i=1}^{p}\Big(\frac{y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta})}-1\Big)\frac{\partial^2\lambda_{i,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}. \tag{9}$$

Therefore, the conditional information matrix is equal to

$$G_n(\boldsymbol{\theta})=\sum_{t=1}^{n}\frac{\partial\boldsymbol{\lambda}_t^T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{D}_t^{-1}(\boldsymbol{\theta})\boldsymbol{\Sigma}_t(\boldsymbol{\theta})\mathbf{D}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}, \tag{10}$$

where the matrix $\boldsymbol{\Sigma}_t(\cdot)$ denotes the true covariance matrix of the vector $\mathbf{Y}_t$. In case when the process $\{\mathbf{Y}_t\}$ consists of uncorrelated components then $\boldsymbol{\Sigma}_t(\boldsymbol{\theta})=\mathbf{D}_t(\boldsymbol{\theta})$. Under suitable conditions, it can be shown that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{\mathbf{D}}\mathcal{N}(0,\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}),$$

where the matrices $\mathbf{H}$ and $\mathbf{G}$ are defined by the limits (in probability) of (9) and (10), respectively. The same result is true for the log-linear model (7); details are omitted. To estimate the copula parameter, it is desirable to compare the conditional distribution of $\mathbf{Y}_t\mid\boldsymbol{\lambda}_t$ to that of $\mathbf{Y}_t^*\mid\boldsymbol{\lambda}_t$, where $\mathbf{Y}_t^*$ is a count time series generated by a suitable choice of a copula. There are several ways of comparing such distributions and this topic is still under investigation. In [46] an initial approach, based on the newly developed concept of local Gaussian correlation (see [9]) was shown to be satisfactory. However, the problem of estimating the copula parameter remains unexplored.

3. Analytical overview of models for MDV time series based on high-order Markov chains

3.1. High-order Markov model of MDV and its probabilistic properties

A large family for MDV time series

$$\mathbf{Y}_t=(Y_{i,t}),\ \ t\in\mathbb{Z},\ \ Y_{i,t}\in\mathsf{A}_i,\ \ i\in\{1,\dots,d\}, \tag{11}$$

where d is the number of components, A_i is a discrete state space of power $N_i=|\mathsf{A}_i|$ for the ith component, is based on the generalized Markovian property [33]:

$$\mathsf{P}[\mathbf{Y}_{t+1}=\mathbf{J}_{s+1}|\mathbf{Y}_t=\mathbf{J}_s,\mathbf{Y}_{t-1}=\mathbf{J}_{s-1},\dots]=\mathsf{P}[\mathbf{Y}_{t+1}=\mathbf{J}_{s+1}|\mathbf{Y}_t=\mathbf{J}_s,\dots,\mathbf{Y}_{t-s+1}=\mathbf{J}_1]=p_{\mathbf{J}_1,\dots,\mathbf{J}_s;\mathbf{J}_{s+1}}, \tag{12}$$

where $\{\mathbf{J}_k\}\subset\mathsf{A}_1\times\cdots\times\mathsf{A}_d$ are arbitrary states of the process, s is a positive integer. The MDV time series defined by (11) and (12) is called d-dimensional (vector) Markov chain of order s and it is abbreviated by ($d-\mathrm{MC}(s)$). Furthermore, it is called homogeneous if the $(s+1)$-dimensional 1-step transition probabilities matrix $\mathbf{P}=\big(p_{\mathbf{J}_1,\dots,\mathbf{J}_s;\mathbf{J}_{s+1}}\big)$ does not depend on time t, otherwise, it is a non-homogeneous chain. The number of parameters for this model is given by

$$D_{d-\mathrm{MC}(s)}=N^s(N-1),\ \ N=N_1\cdots N_d. \tag{13}$$

If the state space $\mathsf{A}_1\times\cdots\times\mathsf{A}_d$ is finite ($N<+\infty$), then we obtain a finite Markov chain, otherwise we encounter a countable Markov chain. Because of the discreteness of $\{\mathsf{A}_i\}$ we can relabel the states so that the $d-\mathrm{MC}(s)$ model is equivalently represented by an 1-dimensional Markov chain of order s ($\mathrm{MC}(s)$) $y_t\in\mathsf{A}$ with the state space A of the power $|\mathsf{A}|=N$ and the $(s+1)$-dimensional 1-step transition probabilities matrix $\mathbf{P}=\big(p_{j_1,\dots,j_s;j_{s+1}}\big)$,

$$p_{j_1,\dots,j_s;j_{s+1}}=\mathsf{P}[y_{t+1}=j_{s+1}|y_t=j_s,\dots,y_{t-s+1}=j_1],\ j_1,\dots,j_{s+1}\in\mathsf{A}. \tag{14}$$

For example, if $N<+\infty$, $\mathsf{A}_i=\{0,\dots,N_i-1\}$, then the renumbering of $\mathbf{Y}_t=\big(Y_{i,t}\big)$ is given by the formula ($N_0=1$)

$$y_t=\sum_{i=1}^{d}Y_{i,t}\prod_{k=0}^{i-1}N_k,\quad y_t\in\{0,\dots,N-1\}.$$

The number of parameters for the model (14) is equal to the number of parameters for (13).

Let us introduce one more useful equivalent transformation (for more, see [6,11]) of a 1-dimensional Markov chain $y_t\in\{0,\dots,N-1\}$ of order s ($\mathrm{MC}(s)$) into the 1-dimensional Markov chain $w_t\in\mathsf{B}=\{0,\dots,N^s-1\}$ of order 1 ($\mathrm{MC}(1)$), $|\mathsf{B}|=N^s$:

$$w_t=\varrho(y_{t+1},\dots,y_{t+s}),\quad \varrho(z_1,\dots,z_s)=\sum_{i=1}^{s}z_iN^{i-1}, \tag{15}$$

and the significantly sparse $(N^s\times N^s)$-square matrix of 1-step transition probabilities $\mathbf{Q}=(q_{K,L})$, $K=\varrho(k_1,\dots,k_s)$, $L=\varrho(l_1,\dots,l_s)\in\mathsf{B}$ given by

$$q_{K,L}=\begin{cases}p_{k_1,\dots,k_s;\,l_s},\ \ l_1=k_2,l_2=k_3,\dots,l_{s-1}=k_s,\\ 0,\ \ \text{otherwise.}\end{cases} \tag{16}$$

Because, in this case the number of parameters is $D=N^s(N^s-1)$, we note that (15)–(16) are not useful in practice.

The MDV time series $\mathbf{Y}_t$ (transformed into w_t) is ergodic [33] iff there exists a finite positive integer $\nu < +\infty$ such that all elements of the matrix $\mathbf{Q}^{\nu}$ are positive, i.e from $(\mathbf{Y}_{t-s+1} = \mathbf{J}_1, \ldots, \mathbf{Y}_t = \mathbf{J}_s)$ the process moves to any future sequence of states $(\tilde{\mathbf{J}}_1, \ldots, \tilde{\mathbf{J}}_s)$ with positive probability by at most ν steps. Because of this property, a sufficient condition of ergodicity is that all conditional 1-step transition probabilities (12), (14) are positive.

For ergodic Markov chain MC(s), the stationary s-dimensional probability distribution $\boldsymbol{\pi} = (\pi_{j_1,\ldots,j_s} > 0), j_1, \ldots, j_s \in \mathrm{A}$ exists, and it satisfies the system of equations $(j_2, j_3, \ldots, j_{s+1} \in \mathrm{A})$:

$$\sum_{j_1 \in \mathrm{A}} \pi_{j_1,\ldots,j_s} p_{j_1,\ldots,j_s;j_{s+1}} = \pi_{j_2,\ldots,j_s,j_{s+1}}. \tag{17}$$

Ergodicity conditions and stationary distributions are necessary to prove asymptotic properties of statistical estimators for the parameters of MDV time series models.

Construction of MDV high-order Markov models for time series has been discussed by several authors. For example, [18] gives a brief overview on some recent models and results in high-dimensional Markov chain models for categorical data sequences, and also introduces a number of high-dimensional Markov chain models with applications, including high-order Markov chain models, multivariate Markov chain models, and high-order multivariate Markov chain models. Furthermore, [10] considers the problem of existence and construction of multivariate Markov chains such that their components are Markov chains with given laws. Sufficient and necessary conditions, in terms of semimartingale characteristics, are provided for a component of a multivariate Markov chain to be a Markov chain in its own filtration (the corresponding property is called weak Markov consistency). The concept of weak Markov copulas is introduced and discussed. The relationship between the concepts of weak Markov consistency and weak Markov copulas (and strong versions of these concepts) is examined. Finally, [37] studies graphical time series models for the analysis of dynamic relationships among variables in multivariate time series. The modeling approach is based on the notion of strong Granger causality and can be applied to time series with non-linear dependencies. The models are constructed from ordinary time series models using constraints that are encoded by mixed graphs. In the graph each component series is represented by a single vertex, directed edges indicate possible Granger-causal relationships between variables, undirected edges are used to map the contemporaneous dependence structure. Various notions of Granger causal Markov properties are introduced. Relationships among these notions and other relevant Markov properties are discussed in this context.

Note also that the actual transition probabilities matrix $\mathbf{P}$ can be constructed as a functional distortion of some parsimonious matrix by an approach similar to [98].

3.2. Construction of parsimonious high-order Markov models for MDV time series

As it is seen from (13), the number of parameters for MDV time series increases exponentially with respect to the memory parameter s. To fit successfully (11), (12) one needs to have huge amounts of data and the computational time associated with this task is of the order $\mathcal{O}\left(N^{s+1}\right)$. To avoid this "curse of dimensionality" it is appropriate to use parsimonious models of high-order Markov chains that are determined by small number of parameters, say $M \ll D_{\mathrm{MC}(s)}$ [72]. We review three main approaches [74] for construction of convenient parametrization of the matrix $\mathbf{P}$ characterizing the MDV time series by (12).

Approach I — reduction of the set of values of elements in the matrix $\mathbf{P}$: Let $\mathbf{Q} = \left(q_{j_1,\ldots,j_r;j_{r+1}}\right)$ be a stochastic $(r+1)$-dimensional matrix, $1 \le r < s$, $\sum_{j_{r+1} \in \mathrm{A}} q_{j_1,\ldots,j_r;j_{r+1}} \equiv 1$, $0 \le q_{j_1,\ldots,j_r;j_{r+1}} \le 1$, $j_1, \ldots, j_{r+1} \in \mathrm{A}$; $\mathrm{B}(\cdot) : \mathrm{A}^s \to \mathrm{A}^r$ be a discrete function. The $(s+1)$-dimensional matrix $\mathbf{P}$, defined by (14), is reduced (squeezed) to the $(r+1)$-dimensional matrix $\mathbf{Q}$ by the transformation:

$$p_{i_1,\ldots,i_s;i_{s+1}} = q_{\mathrm{B}(i_1,\ldots,i_s);i_{s+1}}, \;\; i_1, \ldots, i_{s+1} \in \mathrm{A}, \tag{18}$$

with parsimony coefficient

$$\varkappa \equiv M / D_{\mathrm{MC}(s)} = N^{r-s} < 1.$$

As it is seen from (18), the matrix $\mathbf{P}$ has many identical elements. Examples of this approach include Markov chains of order s with r partial connections MC(s, r) [77], Markov chains of conditional order MCCO(s, r) [76] and variable length Markov chain [14].

Approach II — parametrization of the generation equation for the conditional probability distribution (14) of the future state under its s-past values: In this case, assume that

$$p_{j_1,\ldots,j_s;j_{s+1}} = q_{j_{s+1}}(\boldsymbol{\theta}), \;\; \boldsymbol{\theta} = \boldsymbol{\theta}(j_1, \ldots, j_s; \mathbf{a}), \;\; j_1, \ldots, j_{s+1} \in \mathrm{A}, \tag{19}$$

where $\{q_j(\boldsymbol{\theta}) : j \in \mathrm{A}\}$ is a discrete probability distribution on A that is dependent on the parameter $\boldsymbol{\theta} = (\theta_i) \in \Theta \subseteq \mathbb{R}^L$, $\boldsymbol{\theta}(j_1, \ldots, j_s; \mathbf{a})$ is a parametric function that is a priori known up to an unknown vector parameter $\mathbf{a} = (\mathbf{a}_k) \in \mathbb{R}^m$. The parsimony coefficient for this approach is $\varkappa = m/(N^s(N-1)) \le 1$. This approach has been employed by several authors including [3,40,44,48,63,75,76,79,104,116,128,131].

Approach III — reduction of the initial number of states and the Markovian dependence order: As the computational complexity of the high-order Markov chain is $O(N^{s+1})$, the idea of this trivial approach is in minimization of the power N of the state space (state aggregation) and of the Markov chain order s. Some theory on these problems can be found in [11], for instance.

3.3. Methods, algorithms and software for statistical analysis (estimation, hypotheses testing, forecasting) of MDV time series based on parsimonious high-order Markov chains

3.3.1. The case of models constructed by Approach I

Markov chain MC(s, r) of order s with r partial connections [72,77] is determined by the parsimonious presentation (18) of the $(s+1)$-dimensional transition probability matrix with the function

$$\mathrm{B}(j_1, \dots, j_s) = (j_{m_1^0}, \dots, j_{m_r^0}),\ j_1, \dots, j_s \in \mathrm{A},$$

where r is the number of connections ($1 \le r \le s$); $\mathbf{M}_r^0 = \left(m_1^0, \dots, m_r^0\right) \in M$ is an integer-valued vector with r ordered components, $1 = m_1^0 < m_2^0 < \cdots < m_r^0 \le s$, called the template of connections; $\mathbf{Q} = \left(q_{j_1^{r+1}}\right)_{j_1^{r+1} \in \mathrm{A}^{r+1}}$ is an $(r+1)$-dimensional stochastic matrix. If $r = s$, we have the general MC(s)-model (14). Using maximum likelihood methodology, [72,77] calculate ML estimators for $\mathbf{Q}$ and for $\mathbf{M}_r^0$, r, s and they develop testing theory. A generalization of the MC(s, r) model called Markov chain of conditional order (MCCO) is considered in [76], where ML estimators of the parameters for MCCO-model are given. Another useful class is variable length Markov chains (VLMC) defined on a finite state space, where the Markov property is retained with variable order, see [14] where model (18) is considered with the function $\mathrm{B}(j_1, \dots, j_s) = (j_{s-l+1}, \dots, j_s)$, where $l = l(j_1, \dots, j_s) : \mathrm{A}^s \to \{1, \dots, s\}$ is a discrete function. It is equivalent to assignment of the so-called context function $\mathrm{B}(j_1, \dots, j_s)$ that determines the context tree:

$$\tau = \left\{u : u = \mathrm{B}(j_1, \dots, j_s),\ (j_s, \dots, j_s) \in \mathrm{A}^s\right\}.$$

The ML-estimator for $\mathrm{B}(\cdot)$ is constructed in [14] and used by [120] while more recent paper [90] develops Bayesian modeling in the VLMC framework.

3.3.2. The case of models constructed by Approach II

Historically, a first parsimonious high-order Markov model was introduced in [63] as a discrete-valued time series generated by a stochastic difference equation:

$$y_t = \mu_t y_{t-\eta_t} + (1-\mu_t)\xi_t,\ t > s,\ k \in \mathrm{A},\ \mathrm{A} = \{0, \dots, N-1\},$$

where $\{\xi_t, \eta_t, \mu_t\}$ are jointly independent random variables with probability distributions: $\mathrm{P}\{\mu_t = 1\} = 1 - \mathrm{P}\{\mu_t = 0\} = \rho$; $\mathrm{P}\{\xi_t = k\} = \pi_k$, $k \in \mathrm{A}$; $\mathrm{P}\{\eta_t = i\} = \lambda_i$, $i \in \{1, \dots, s\}$, $\lambda_s \ne 0$; the parsimony coefficient is given by $\varkappa = (N+s-1)/(N^s(N-1)) \le 1$. Parametric representation of the conditional probability distribution (19) takes the form:

$$p_{i_1,\dots,i_s;i_{s+1}} = (1-\rho)\pi_{i_{s+1}} + \rho \sum_{j=1}^{s} \lambda_j\, 1\{i_{s-j+1} = i_{s+1}\}, \quad i_1, \dots, i_{s+1} \in \mathrm{A},$$

where $1\{.\}$ is the indicator function; see [63] where the moments and stationary distribution are studied and [72] where ML estimation theory is developed.

A widely used class of models is Mixture Transition Distribution models (MTD) introduced in [116]. This class is defined by the parsimonious representation (19) of conditional probability distribution:

$$p_{i_1,\dots,i_s;i_{s+1}} = \sum_{j=1}^{s} \lambda_j q_{i_j,i_{s+1}}, \quad i_1, \dots, i_{s+1} \in \mathrm{A},$$

where $\mathbf{Q} = (q_{i,k})$ is a stochastic $(N \times N)$-matrix, $0 \le q_{i,k} \le 1$, $\sum_{k \in \mathrm{A}} q_{i,k} \equiv 1$, $i, k \in \mathrm{A}$, $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_s)^T$ is a discrete probability distribution, $\lambda_1 > 0$. Generalized MTD-model (MTDg) has the following conditional probability distribution:

$$p_{i_1,\dots,i_s;i_{s+1}} = \sum_{j=1}^{s} \lambda_j q^{(j)}_{i_{s-j+1},i_{s+1}}, \quad i_1, \dots, i_{s+1} \in \mathrm{A},$$

where $\mathbf{Q}^{(j)} = \left(q^{(j)}_{i,k}\right)$ is a stochastic matrix for the jth lag; the parsimony coefficient for the MTDg is equal to $\varkappa = \mathcal{O}\left(s \cdot N^{2-s}\right)$. ML-estimators and numerical algorithms for their computation can be found in [117]. In [72] a useful property of the s-dimensional stationary probability distribution (17) for the MTDg-model is proved:

$$\pi_{i_1,\dots,i_s} = \prod_{l=0}^{s-1} \left(\pi_{i_{s-l}} + \sum_{j=l+1}^{s} \lambda_j \left(q^{(j)}_{i_{j-l},i_{s-l}} - \sum_{r=0}^{N-1} q^{(j)}_{r,i_{s-l}} \pi_r\right)\right), \quad i_1, \dots, i_s \in \mathrm{A}.$$

Using this property consistent statistical estimators for parameters $\boldsymbol{\lambda}$, $\mathbf{Q}$ of the MTD-model are constructed in explicit form and used for iterative computation of the ML-estimators [72]. An interesting computational approach to statistical estimation of parameters $\{\boldsymbol{\lambda}, \mathbf{Q}^{(j)}\}$ of the MTDg-model, proposed in [17], is based on linear programming for minimization

of difference (in L_∞- or L_1-norm) between left and right sides of Eqs. (17) where the estimators $\hat{\pi}_{j_1,\dots,j_s}$ are used instead of true values $\pi_{j_1,\dots,j_s}$ ($j_1,\dots,j_s \in \mathrm{A}$). This approach is generalized also for the multivariate high-order Markov chains. In [124] the authors study MTD-model by means of algebraic statistics. In [107] the authors propose a new MTD-Probit model for multivariate high-order Markov chains based on MTD-model. The proposed model has two features: it is completely free of constraints, facilitating the estimation procedure, and it gives more precise estimators for the transition probabilities of higher-order Markov chain than the standard MTD-model.

In [127] a class of ARMA-type models for stationary binary time series (BinARMA) is investigated. New results on the autocorrelation structure of BinARMA models are presented: these results simplify in the BinMA(q) case, while the known results concerning BinAR(p) models are included as a special case. Binomial conditionally nonlinear autoregressive model BiCNAR(s) was introduced in [79] as a special (binomial) case of Eq. (19):

$$p_{j_1,\dots,j_s;j_{s+1}} = \binom{N-1}{j_{s+1}} \theta^{j_{s+1}} (1-\theta)^{N-1-j_{s+1}},\ j_{s+1} \in \mathrm{A},\ \mathrm{A} = \{0,\dots,N-1\},$$

$$\theta = \theta\left(J_1^s\right) = F\left(\mathbf{a}'\Psi\left(J_1^s\right)\right),\ J_1^s = (j_1,\dots,j_s)^T \in \mathrm{A}^s,$$

where $\Psi\left(J_1^s\right) = \left(\psi_1\left(J_1^s\right),\dots,\psi_m\left(J_1^s\right)\right)^T : \mathrm{A}^s \to \mathbb{R}^m$ is a column-vector of $m \le N^s$ linearly independent functions, e.g., polynomials; $F(\cdot)$: $\mathbb{R}^1 \to [0,1]$ is a fixed cumulative distribution function, e.g., logistic, normal or Cauchy; $\mathbf{a} = (\mathbf{a}_1,\dots,\mathbf{a}_m)^T$ is column-vector of m unknown model parameters; parsimony coefficient for this model $\varkappa = m(N^s(N-1))^{-1} \le 1$. Instead of traditional ML-approach having known computational shortages (induced by multi-modality of the loglikelihood function and iterative computer procedures) the Frequencies Based Estimators (FBE) were developed.

We introduce some notation: $F^{-1}(\cdot)$ is the quantile function; $Y_1^T = (Y_1,\dots,Y_n)^T \in \mathrm{A}^T$ is the observed time series of length n;

$$\hat{\pi}_{J_1^s} = \frac{1}{n-s+1}\sum_{t=s}^{n} 1\{Y_{t-s+1}^t = J_1^s\},\ \ \hat{p}_{J_1^s;j_{s+1}} = \begin{cases} \hat{\pi}_{J_1^{s+1}}/\hat{\pi}_{J_1^s},\ \hat{\pi}_{J_1^s} > 0, \\ 1/N,\ \text{otherwise}, \end{cases}$$

are frequencies used for estimation of parameters $\mathbf{a} \in \mathbb{R}^m$,

$$\hat{\theta}\left(J_1^s\right) = \frac{1}{N-1}\sum_{j_{s+1}\in\mathrm{A}} j_{s+1}\hat{p}_{J_1^s;j_{s+1}},$$

$\mathbf{B} = (b_{J_1^s})$ is $(N^s \times 1)$-vector-column, $b_{J_1^s} = F^{-1}\left(\hat{\theta}\left(J_1^s\right)\right)$, $\mathbf{H} = (h_{JJ'})$ is a fixed $(N^s \times N^s)$-symmetric non-negative definite matrix, $\Psi = (\Psi(J_1^s))$ is $(m \times N^s)$-matrix, O_m is the zero m-vector.

If $F(\cdot)$ satisfies the smoothness assumptions: $0 < F(\zeta) < 1$, $0 < F'(\zeta) < +\infty$, $F(\cdot)$ and $F^{-1}(\cdot)$ are twice differentiable, and $|\Psi\mathbf{H}\Psi'| \neq 0$, then it is proved that the FBE $\hat{\mathbf{a}} = (\Psi\mathbf{H}\Psi')^{-1}\Psi\mathbf{H}\mathbf{B}$ is consistent and asymptotically normal when $T \to +\infty$:

$$\hat{\mathbf{a}} \xrightarrow{P} \mathbf{a},\ \ \sqrt{T}(\hat{\mathbf{a}} - \mathbf{a}) \xrightarrow{D} \mathcal{N}_m(\mathbf{0}_m, \Sigma_{\mathbf{H}}),\ \ \Sigma_{\mathbf{H}} = (\Psi\mathbf{H}\Psi')^{-1}\Psi\mathbf{H}\mathbf{H}_*^{-1}\mathbf{H}(\Psi\mathbf{H}\Psi')^{-1},$$

where $\mathbf{H}_*$ is the Fisher information matrix. For the weight matrix $\mathbf{H} = \mathbf{H}_*$ the FBE $\hat{\mathbf{a}}$ is asymptotically efficient.

The FBE estimator $\hat{\mathbf{a}}$ has the following significant advantages w.r.t. the MLE: (i) explicit expression of FBE w.r.t. the iterative computation of MLE; (ii) fast iterative computation of FBE if the basis $\{\psi_i(\cdot)\}$ needs to be extended; (iii) possibility to control the computational complexity of FBE by variation of the matrix $\mathbf{H}$.

The FBE approach is successively developed for some other types of discrete-valued time series (19): semibinomial [80], binary [78,84], Poisson, geometric, negative binomial [75], regression time series [81], and also for Binomial conditionally nonlinear model for spatio-temporal data [85]; it is used for statistical forecasting of COVID-19, see [83], for more references.

3.3.3. The case of models constructed by Approach III

The work in [52] presents a method for reducing a regular, discrete-time Markov chain to another one with a smaller number of states. The cost of reduction is the Kullback–Leibler divergence rate between a projection of the original process through a partition function and a Markov chain on the correspondingly partitioned state space. Since finding the reduced model with minimal cost is computationally expensive, in [52] it is proposed, to minimize an upper bound on the reduction cost instead of minimizing the exact cost. The proposed upper bound is fast computable and it is tight if the original chain is lumpable (see [70]) w.r.t. the partition. The problem is expressed in the form of information bottleneck optimization, and solved using the agglomerative information bottleneck algorithm for searching a suboptimal partition greedily, rather than exhaustively. The theory is examined on bio-molecular data.

In [8] the use of vine copulas is discussed for building flexible semiparametric models for stationary multivariate higher-order Markov chains. The authors propose a new vine structure called M-vine, and explore the use of vine copulas in the multivariate framework. In particular, the M-vine is shown to be well suited for modeling stationary multivariate higher-order Markov chains. M-vine copula specification is estimated using existing semiparametric techniques that are nonparametric w.r.t. univariate marginal behavior and parametric w.r.t. dependence between variables.

Finally, [111] study the Conditional Mutual Information (CMI) for the estimation of the Markov chain order. For a Markov chain of K symbols CMI $I_c(m)$ of order m is the mutual information of two variables in the chain being m time steps apart, conditioning on the intermediate variables of the chain. Approximate analytic significance limits based on the estimation bias of CMI are found and a randomization significance test of $I_c(m)$ is developed, where the randomized s estimated by the last order for which the null hypothesis is rejected. The order criterion of CMI-testing is compared with the Akaike and Bayesian information criteria, the maximal fluctuation method (Peres-Shields estimator) and a likelihood ratio test for increasing orders using φ-divergence. CMI-testing criterion turns out to be superior for orders larger than one, but its effectiveness for large orders depends on data availability.

3.3.4. Other problems of statistical analysis for high-order Markov models of MDV time series

In [41] the authors introduce a generalized linear mixed model of multivariate counts in which all random effects may evolve over time. Random effects have a discrete support and follow a first-order Markov chain. Constraints control the size of the parameter space and possibly yield blocks of time-constant random effects. The theory is illustrated by an application to the relationship between health education and depression in a panel of adolescents, where the random effects are high dimensional and separately evolve over time.

In [34] a parametrized family of partially observed bivariate Markov chains is considered. It is shown that, under very mild assumptions, the limit of the normalized log-likelihood function is maximized when the parameters belong to the equivalence class of the true parameter, which is a key feature for obtaining the consistency of the maximum likelihood estimators (MLEs) in well-specified models. This result is obtained by means of the general framework of partially dominated models. Two specific cases of interest are examined: hidden Markov models (HMMs), and observation-driven time series models. The distinct feature of the proposed approach is that the identifiability is addressed by relying on the uniqueness of the stationary distribution of the Markov chain associated to the complete data, regardless its rate of convergence to the equilibrium.

In [54] the authors study optimality properties of decision procedures for the quickest detection of a change-point of parameters in autoregressive and other Markov type sequences. The limit of the normalized conditional log-likelihood ratios is shown to exist for Markov chains satisfying the ergodic theorem of information theory. The explicit expressions for this limit are derived from the time average rate of Kullback–Leibler divergence.

4. Directions for future research

4.1. Extensions of observation-driven models for applications

Recent work in the area of observation-driven models in [126] studies continuous-time processes for multivariate count time series such that their marginal distribution is infinitely divisible. This construction enables separate modeling of serial correlation and the cross-sectional dependence. The recent article by [56] considers high-dimensional count time series and studies the issues of inference for autoregressive parameters and the corresponding network structure by developing a sparsity-regularized maximum likelihood estimator. The work in [31] studies a general approach for modeling the dynamics of multivariate time series when the data are of mixed type (binary/count/continuous) building on the generating process introduced in [46].

Non-stationarity: Though stationarity is an important notion for the theoretical development of time series methods, applications from neurobiology, financial and social sciences data call for the development of non-stationary models. Indeed, models that allow the inclusion of periodic and trend covariates will give further insight to real data problems and address important issues like heterogeneity and time-varying effects. Promising approaches towards this development can be based on parameter-driven models (see [43] for some references) or on the notion of local stationarity (see [25,132], for instance).

Diagnostics: Diagnostics for models with and without covariates, along the lines of [19], serve as an additional tool to assess models. A major problems will be to develop a Probability Integral Transform (PIT), see [23], for multivariate count autoregression. Such a tool will enable identification of a suitable marginal response distribution for modeling count time series. Other topics include development of graphical methods for examining suitably defined residuals and studying standard portmanteau test statistics under this setup.

Network Autoregression: Measuring the impact of a network structure to a multivariate time series process has attracted considerable attention over the last years, mainly due to the growing availability of network data, recorded over a given timespan, in several areas of study (social networks, GPS data, epidemics, air pollution monitoring systems and more generally environmental wireless sensor networks, among many other applications). Development of multivariate time series models for a vector of variables $\mathbf{Y}_t \in \mathbb{R}^d$, for $t \in \{1, \ldots, n\}$, using a VAR model is a standard approach in time series analysis. However, if d is the size of the network, then the number of parameters to be estimated is $\mathcal{O}(d^2)$ which is much larger than the temporal sample size n. Hence, other approaches have been considered to reduce the number of parameters. One way is based on sparsity, see for example [61], among others. An alternative method is dimension reduction using factor models. A dimension reduction via factors which accounts for network impact has been developed by [135], where the so-called Network vector Autoregressive model (NAR) is introduced. Such models are tailored for continuous network data. The parameters of the NAR model are estimated via ordinary least squares method, under two

asymptotic regimes (i) with increasing time sample size $n \to \infty$ and fixed network dimension d (which is standard in multivariate time series analysis) and (ii) with both n, d increasing, i.e., $\min\{n, d\} \to \infty$.

Since discrete responses are commonly encountered in real applications and are strongly connected to network data, [5] developed linear and log-linear multivariate count-valued versions of the NAR model, called Poisson Network Autoregression (PNAR), for the two related types of asymptotic inference discussed (i)–(ii), under the α-mixing property of the innovation term, see [35]. The marginal distribution of the count process is Poisson whereas the dependence among them is captured by the copula construction as described before. For this case, the OLS inference is not applicable, so that inference for the PNAR model relies on the quasi MLE.

Closely related to this topic is the work by [103] who study space–time models for infectious disease surveillance data capturing the dynamics of disease spread by incorporating a power-law network structure. Furthermore [12] study the endemic-epidemic framework, a class of autoregressive models for infectious disease surveillance counts, and replace the default autoregression on counts from the previous time period with more flexible weighting schemes inspired by discrete-time serial interval distributions, see also Section 4.4 for further discussion on this topic.

4.2. Extensions for MDV time series based on high-order Markov chains

4.2.1. Extension of conditionally nonlinear autoregressive (CNAR) models

Define L-parametric exponential family $\mathcal{E}$ of probability distributions on A (recall (14))

$$p(x;\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{\lambda})}\exp\left(h_0(x) + \sum_{i=1}^{L}\boldsymbol{\lambda}_i h_i(x)\right),\ Z(\boldsymbol{\lambda}) = \sum_{x\in \mathrm{A}}\exp\left(h_0(x) + \sum_{i=1}^{L}\boldsymbol{\lambda}_i h_i(x)\right),\ \boldsymbol{\lambda} = (\boldsymbol{\lambda}_i)_{i=1}^{L} \in \Lambda \subset \mathbb{R}^L. \tag{20}$$

Here $\boldsymbol{\lambda}$ is L-dimensional canonical parameter of the exponential family $\mathcal{E}$, $h_i(x) : \mathrm{A} \to \mathbb{R}$, $i \in \{0, \dots, L\}$, are $L+1$ base functions that define (non bijectively) family $\mathcal{E}$, $Z(\boldsymbol{\lambda})$ is a partition function, Λ is the set of admissible values of canonical parameter, i.e., the values $\boldsymbol{\lambda}$ providing convergence of the sum (20) for partition function $Z(\boldsymbol{\lambda})$. The dual parameter for family $\mathcal{E}$ is

$$\boldsymbol{\theta} = \phi(\boldsymbol{\lambda}) ::= \frac{d}{d\boldsymbol{\lambda}}\ln Z(\boldsymbol{\lambda}),\ \boldsymbol{\theta} \in \Theta = \mathrm{conv}\{(h_i(x))_{i=1}^{L}, x \in \mathrm{A}\} \subset \mathbb{R}^L.$$

Here $\mathrm{conv}\{\cdot\}$ means convex hull for subset of $\mathbb{R}^L$. Canonical and dual parameters $\boldsymbol{\lambda} \in \Lambda$ and $\boldsymbol{\theta} = \phi(\boldsymbol{\lambda}) \in \Theta$ both specify the same distribution from the family $\mathcal{E}$. Let us denote this distribution $\mathcal{E}(\boldsymbol{\lambda}) = \mathcal{E}[\boldsymbol{\theta}] \in \mathcal{E}$, where the type of brackets depends on the type of parameter used.

Consider now discrete-valued time series $Y_t \in \mathrm{A}$, $t \in \mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$, and probabilistic models for it. Let us define model $\mathcal{E}$-MC(s) as a stationary Markov chain of order $s \in \mathbb{N}$ such that all the conditional distributions of observations, given the past of the process, belong to $\mathcal{E}$-family:

$$\mathcal{L}\{Y_t | Y_\tau, \tau < t\} = \mathcal{E}[\boldsymbol{\theta}(Y_{t-s}^{t-1})],\ t \in \mathbb{Z},\ Y_{t-s}^{t-1} = (Y_i)_{t-s}^{t-1}.$$

Model $\mathcal{E}$-MC(s) is determined by the order s, the family $\mathcal{E}$, and the function $\boldsymbol{\theta}(q) : \mathrm{A}^s \to \Theta$. The following special case of the latter function leads to the so-called Conditionally Nonlinear AutoRegressive (CNAR) model $\mathcal{E}$-CNAR(s) [55,80,84]:

$$\boldsymbol{\theta}(q) = F\left(\sum_{i=1}^{m}\mathbf{a}_i\psi_i(q)\right),\ q \in \mathrm{A}^s,$$

where $\mathbf{a} = (\mathbf{a}_i)_{i=1}^{m} \in \mathbb{R}^m$ is an unknown model parameter, $\psi_i(q) : \mathrm{A}^s \to \mathbb{R}^L$, $i \in \{1, \dots, m\}$, are an m linearly independent base functions, $F : \mathbb{R}^L \to \Theta$ is a bijection.

This model generates the models considered in Section 3.3.2, e.g., for Binomial CNAR-model BiCNAR (see [80,81]): $\mathrm{A} = \{0, \dots, N\}$, $\mathcal{E} = \mathrm{Binomial}(N, \cdot)$ with $h_0(x) = \ln\binom{N}{x}$, $h_1(x) = x$, $\Theta = [0, N]$. Note that standard multivariate probability distributions [66] usually have more than one parameter, so if we use them as conditional distributions of $\mathcal{E}$-CNAR MDV time series, we should take $L > 1$.

4.2.2. Extensions of high-order Markov models for MDV time series based on information geometry

In [58] the authors develop informational geometric tools for construction of parsimonious Markov models and their statistical analysis. It is shown that the set of probability measures of stationary first order irreducible Markov chains with a finite state space A forms an exponential family, if all the probability transition matrices $\mathbf{W} \in \mathbb{R}^{|\mathrm{A}|\times|\mathrm{A}|}$ of these Markov measures have the same common support $\{(i,j) \in \mathrm{A}^2 : \mathbf{W}(i,j) > 0\}$. In particular, the set of probability measures of stationary irreducible Markov chains of any order $s \in \mathbb{N}$ form an exponential family: vectorization (15) transforms MC(s) to MC(1) with larger state space A^s, and the support of the corresponding probability transition matrices $\mathbf{W} \in \mathbb{R}^{|\mathrm{A}^s|\times|\mathrm{A}^s|}$ consists of the pairs $(\mathbf{K}, \mathbf{L})$ satisfying condition (16) (left $(s-1)$-subword of $\mathbf{L} \in \mathrm{A}^s$ equals right $(s-1)$-subword of $\mathbf{K} \in \mathrm{A}^s$). Low-dimensional exponential and curved exponential subfamilies of the exponential family MC(s) may be considered as parsimonious models for MDV time series based on high-order Markov chains.

4.2.3. Statistical analysis of discrete-valued Markov fields and MDV spatio-temporal data

MDV time series $\mathbf{Y}_t = (Y_{i,t})_{i=1}^d$, $t \in \mathbb{Z}$, may be considered as a Markov random field (MRF) on a stripe $\{1, \dots, d\} \times \mathbb{Z}$, and the methods of discrete random field theory may be used for the statistical analysis of the MDV time series, and vice versa. For instance, Lars Onsager derived his famous solution for square lattice Ising model [109] by representing binary field $Y_{i,j} \in \{0, 1\}$, $(i, j) \in \{1, \dots, m\} \times \{1, \dots, n\}$, as a Markov chain $\mathbf{W}_j = (Y_{1,j}, \dots, Y_{m,j})'$, $j \in \{1, \dots, n\}$, with 2^m states, and analyzing the spectrum of the corresponding probability transition $(2^m \times 2^m)$-matrix. Discrete spatio-temporal data as a kind of discrete field may be in a similar way transformed into MDV time series, when the space data at the moment $t \in \mathbb{Z}$ is aggregated into a multivariate observation $\mathbf{Y}_t$.

4.2.4. Extensions based on artificial neural networks

Artificial neural networks (ANN) and deep learning are widely used for statistical analysis of time series [50]: in prediction [16,49], classification [134], outliers detection [100]. The standard use of ANN in time series prediction is often called sliding window technique [49] and has a form:

$$\hat{Y}_t = \mathfrak{N}(Y_{t-1}, \dots, Y_{t-s}), \ t \in \mathbb{Z}, \tag{21}$$

where $\mathfrak{N}(\cdot)$ is an ANN function (transforms the input of ANN into its output). In the case of MDV time series the right-hand side of (21) may not belong to discrete state space A, so the following modification may be used:

$$\mathsf{E}\{Y_t | Y_\tau, \tau < t\} = \mu_t = \mathfrak{N}(Y_{t-1}, \dots, Y_{t-s}), \ \hat{Y}_t = \delta(\mu_t), \ t \in \mathbb{Z}, \tag{22}$$

where the function $\delta(\cdot)$ rounds continuous values of ANN output to discrete values from A: for instance, $\delta(z) = [z]$, $z \in \mathbb{R}$, $\mathrm{A} = \mathbb{Z}$. The main equation of GLM-based time series:

$$\mathsf{E}\{Y_t | Y_\tau, \tau < t\} = \mu_t = F\left(\sum_{i=1}^{s} Y_{t-i}\mathbf{a}_i\right), \ t \in \mathbb{Z},$$

is a special case of (22) for one-layer ANN. Using multilayer ANNs instead of one-layer ANNs leads to a wide class of extensions for the GLM-based MDV time series.

4.3. Robust estimation

Research on robust inference for multivariate count time series is scarce so far. An exception is [92] where the authors apply a conditional version of the minimum density power divergence estimation (MDPDE) approach [7] for fitting the bivariate Poisson INGARCH model (recall Section 2.2) discussed in [22]:

$$(Y_{1,t}, Y_{2,t})^T \sim BP(\lambda_{1,t}, \lambda_{2,t}, \delta), \quad \boldsymbol{\lambda}_t = (\lambda_{1,t}, \lambda_{2,t})^T = \boldsymbol{\omega} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}\mathbf{Y}_{t-1}.$$

Hereby, $BP(\lambda_{1,t}, \lambda_{2,t}, \delta)$ is the bivariate Poisson distribution with mean $(\lambda_{1,t}, \lambda_{2,t})^T$ and joint probability density function (pdf) which equals the product of the marginal pdfs multiplied by the correction factor $1 + \delta(e^{-y_1} - e^{-c\lambda_{1,t}})(e^{-y_2} - e^{-c\lambda_{2,t}})$ with $c = 1 - 1/e$ (compare with (1) which is of different functional form). This distribution supports both positive and negative correlations as $\mathrm{Cov}[Y_{1,t}, Y_{2,t}|\mathcal{F}_{t-1}] = \delta c^2 \lambda_{1,t}\lambda_{2,t} \exp(-c(\lambda_{1,t} + \lambda_{2,t}))$. In [92] the authors focus on the case of a diagonal matrix $\mathbf{A}$ and show the strong consistency and asymptotic normality of the MDPDE for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, $\boldsymbol{\theta}_i = (\omega_i, a_i, b_{i,1}, b_{i,2})^T$, which is

$$\hat{\boldsymbol{\theta}}_{\alpha,n} = \arg_{\boldsymbol{\theta}} \min n^{-1} \sum_{t=1}^{n} \tilde{h}_{\alpha,t}(\boldsymbol{\theta}),$$

where

$$\tilde{h}_{\alpha,t}(\boldsymbol{\theta}) = \begin{cases} \sum_{y_1}^{\infty} \sum_{y_2}^{\infty} f_{\boldsymbol{\theta}}^{1+\alpha}((y_1, y_2)'|\tilde{\boldsymbol{\lambda}}_t) - (1 + 1/\alpha) f_{\boldsymbol{\theta}}^{\alpha}(\mathbf{Y}_t|\tilde{\boldsymbol{\lambda}}_t), & \alpha > 0, \\ -\ln f_{\boldsymbol{\theta}}(\mathbf{Y}_t|\tilde{\boldsymbol{\lambda}}_t), & \alpha = 0. \end{cases}$$

Hereby, $f_{\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\lambda}_t)$ is the conditional pdf, and $\tilde{\boldsymbol{\lambda}}_t$ is recursively defined as

$$\tilde{\boldsymbol{\lambda}}_t = \boldsymbol{\omega} + \mathbf{A}\tilde{\boldsymbol{\lambda}}_{t-1} + \mathbf{B}\mathbf{Y}_{t-1}, \quad t \geq 2,$$

using a suitable initialization $\tilde{\boldsymbol{\lambda}}_1$. For $\alpha = 0$ we obtain the conditional MLE, while values of the tuning parameter α larger than 0 lead to some robustness at the expense of a loss of some efficiency for data without outliers. Small values like $\alpha = 0.3$ yield a reasonable compromise between efficiency and robustness.

More generally, multivariate INGARCH models can be fitted making use of the VARMA representation of the auto- and cross-covariance structure, as robust estimators are available for VARMA models. An inherent problem of this approach is that many robust multivariate covariance estimators are designed for continuous elliptical distributions. See [21] for the popular minimum covariance determinant estimator [24]. Such estimators often have problems with collinearities which may occur in discrete data [36,39]. The MDPDE mentioned above has also been suggested for estimation of autocovariance matrices but this approach seems to have been developed only for Gaussian processes [86].

Alternatively, we could apply robust regression-based estimators, but in time series autoregression these suffer from the propagation of outlier effects to the residuals of subsequent observations. In case of an V-INARCH(p) model e.g., an outlying $\mathbf{y}_t$ contaminates the residuals at p subsequent time points, $\mathbf{a}_{t+h}(\boldsymbol{\theta}) = \mathbf{y}_{t+h} - \boldsymbol{\omega} - \mathbf{B}_1\mathbf{y}_{t+h-1} - \cdots - \mathbf{B}_p\mathbf{y}_{t+h-p}$, $h \in \{1, \ldots, p\}$. So called MM-estimators based on a bounded innovation propagation autoregressive model are proposed in [105], but these rely on Mahalanobis distances and thus assume an elliptical distribution at least implicitly. Residual adjustment (RA) estimators are proposed in [51]: they can be seen as CMLEs of a process generated by modified innovations. While the normal CMLE minimizes

$$\sum_{t=1}^{n} \mathbf{a}_t(\hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}} \mathbf{a}_t(\hat{\boldsymbol{\theta}}), \;\; \hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{t=1}^{n} \mathbf{a}_t(\hat{\boldsymbol{\theta}}) \mathbf{a}_t(\hat{\boldsymbol{\theta}})^T,$$

the RA estimator introduces robust weights $w(x) = \psi(x)/x$ based on the Huber or the Tukey score function ψ to define modified residuals

$$\tilde{\mathbf{a}}_t(\boldsymbol{\theta}) = \mathbf{a}_t(\hat{\boldsymbol{\theta}}) w(\{\hat{\mathbf{a}}_t(\hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}} \hat{\mathbf{a}}_t(\hat{\boldsymbol{\theta}})\}^{1/2}).$$

Use of the Huber or the Tukey ψ-function needs the choice of a tuning constant, which should be chosen larger in case of conditional Poisson [38] than in case of conditional Gaussian distributions, and even larger than this in case of overdispersed conditional distributions like the negative binomial to achieve satisfactory efficiency [39]. RA estimators are qualitatively robust only for VAR but not for VARMA models, but Monte Carlo studies suggest that they are generally more stable than the CMLE in the presence of outliers.

Robust estimation seems to be even harder for multivariate log-linear models than for multivariate linear models as there is no VARMA representation available. Solutions for the univariate case are provided in [87,88], but so far nothing seems to be done in the multivariate case.

Some results on robust inference based on Markov models for observed MDV time series are in [82] where the authors develop robust estimators for Binomial conditionally nonlinear autoregressive time series of order s, under innovation outliers with arbitrary discrete probability distribution having some fixed known expectation. In [73], the robustness of sequential testing of parametric hypotheses for M-valued Markov chains is analyzed under Tukey–Huber distortions.

4.4. Spatio-temporal models

When analyzing, e.g., numbers of new infections on a county or district level in a whole country we are confronted with high-dimensional count time series which may consist of several hundred components. Fitting a linear V-IN(G)ARCH or a log-linear model as described in Sections 2.2.1 and 2.2.2 would need estimation of several thousands or even tens of thousands parameters if we do not impose additional restrictions. In an unpublished master thesis, Maletz (2021) transfers the STARMA-modeling approach developed by [20,115] to the multivariate models for count time series reviewed in Section 2.2. This approach parameterizes the matrices $\mathbf{A}_i = \sum_{k=0}^{K} \alpha_{i,k}\mathbf{W}_k$ and $\mathbf{B}_j = \sum_{k=0}^{K} \beta_{j,k}\mathbf{W}_k$ in terms of predefined weight matrices $\mathbf{W}_k$, $k \in \{0, \ldots, K\}$, capturing neighborhoods of different orders. While $\mathbf{W}_0$ is the identity matrix, $\mathbf{W}_1 \ldots, \mathbf{W}_K$ describe first to Kth order neighborhoods and are row-normalized, i.e., the elements of each row add up to 1. In the simple case of a regular 2×2 grid we may define each point to have two direct (first order) neighbors and one second order neighbor, leading to the neighborhood matrices

$$\mathbf{W}_1 = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix}, \quad \mathbf{W}_2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

In this scenario we would achieve a parameter reduction from four to three for each of the matrices $\mathbf{A}_i$ and $\mathbf{B}_j$, or possibly to two if second order neighborhoods are not significant. Parameter reductions can be huge in case of larger grids or areas consisting of many districts. A recent thesis [99] derives rather simple expressions for the conditions of stability and ergodicity described in Section 2.2 in terms of the parameters α_k and β_k and investigates several estimation procedures for fitting such models to high-dimensional count time series.

Acknowledgments

The authors gratefully acknowledge the helpful suggestions of the Guest Editors of the JMVA Jubilee Issue during the preparation of the paper.

References

[1] A. Agresti, Categorical Data Analysis, second ed., John Wiley & Sons, New York, 2002.
[2] A. Ahmad, Contributions à l'économétrie des Séries Temporelles à Valeurs Entières (PhD thesis), University Charles De Gaulle-Lille III, France, 2016.
[3] M. Al-Osch Alzaid, An integer-valued pth-order autoregressive structure (INAR(p)) process, J. Appl. Probab. 27 (1990) 314–324.

[4] C.M. Andreassen, Models and Inference for Correlated Count Data (Ph.D. thesis), Aarhus University, Denmark, 2013.
[5] M. Armillotta, K. Fokianos, Poisson network autoregression, 2021, Available at https://arxiv.org/abs/2104.06296.
[6] I. Basawa, Statistical Inference for Stochastic Processes, Academic Press, London, 1980.
[7] A. Basu, I.R. Harris, N.L. Hjort, M.C. Jones, Robust and efficient estimation by minimizing a density power divergence, Biometrika 85 (1998) 549–559.
[8] B. Beare, J. Seo, Vine copula specifications for stationary multivariate Markov chains, J. Time Series Anal. 36 (2015) 228–246.
[9] G.D. Berentsen, B. Støve, D. Tjøstheim, T. Nordbø, Recognizing and visualizing copulas: an approach using local Gaussian approximation, Insurance Math. Econom. 57 (2014) 90–103.
[10] T. Bielecki, J. Jakubowski, M. Nieweglowski, Intricacies of dependence between components of multivariate Markov chains: weak Markov consistency and weak Markov copulae, Electron. J. Probab. 18 (2013) 21.
[11] P. Billingsley, Statistical methods in Markov chains, Ann. Math. Stat. 30 (1961) 417–437.
[12] J. Bracher, L. Held, Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction, Int. J. Forecast. (2020).
[13] P.J. Brockwell, R.A. Davis, Time Series: Data Analysis and Theory, second ed., Springer, New York, 1991.
[14] P. Bühlmann, A.J. Wyner, Variable length Markov chains, Ann. Statist. 27 (1999) 480–513.
[15] A.C. Cameron, P.K. Trivedi, Regression Analysis of Count Data, second ed., in: Econometric Society Monographs, vol. 53, Cambridge University Press, Cambridge, 2013.
[16] K. Chakraborty, K. Mehrotra, C. Mohan, S. Ranka, Forecasting the behavior of multivariate time series using neural networks, Neural Netw. 5 (1992) 961–970.
[17] W. Ching, M. Ng, Markov Chains: Algorithms and Applications, Springer Science + Business Media, 2006.
[18] W. Ching, D. Zhu, On high-dimensional Markov chain models for categorical data sequences with applications, in: Recent Advances in Scientific Computing and Matrix Analysis, Int. Press, Somerville, MA, 2011, pp. 15–34.
[19] V. Christou, K. Fokianos, On count time series prediction, J. Stat. Comput. Simul. 2 (2015) 357–373.
[20] A. Cliff, J. Ord, Spatial Autocorrelation, Pion, London, 1973.
[21] C. Croux, K. Joossens, Robust estimation of the vector autoregressive model by a least trimmed squares procedure, Compstat 2008 (2008) 489–501.
[22] Y. Cui, F. Zhu, A new bivariate integer-valued garch model allowing for negative cross-correlation, Test 27 (2018) 428–452.
[23] C. Czado, T. Gneiting, L. Held, Predictive model assessment for count data, Biometrics 65 (2009) 1254–1261.
[24] M. D. M. Hubert, P. Rousseeuw, Minimum covariance determinant and extensions, WIREs Comput. Stat. 10 (2018) e1421.
[25] R. Dahlhaus, A likelihood approximation for locally stationary processes, Ann. Statist. 28 (2000) 1762–1794.
[26] R.A. Davis, W.T.M. Dunsmuir, Y. Wang, On autocorrelation in a Poisson regression model, Biometrika 87 (2000) 491–505.
[27] R.A. Davis, K. Fokianos, S.H. Holan, H. Joe, J. Livsey, R. Lund, V. Pipiras, N. Ravishanker, Count time series: A methodological review, J. Amer. Statist. Assoc. (2021) (in press).
[28] R.A. Davis, S.H. Holan, R. Lund, N. Ravishanker (Eds.), Handbooks of Modern Statistical Methods, in: Handbook of Discrete-Valued Time Series, Chapman & Hall/CRC, London, 2016.
[29] R.A. Davis, H. Liu, Theory and inference for a class of nonlinear models with application to time series of counts, Statist. Sinica 26 (2016) 1673–1707.
[30] Z.M. Debaly, L. Truquet, Stationarity and moment properties of some multivariate count autoregressions, 2019, arXiv preprint arXiv:1909.11392.
[31] Z.M. Debaly, L. Truquet, Multivariate time series models for mixed data, 2021, Available at https://arxiv.org/abs/2104.01067.
[32] M. Denuit, P. Lambert, Constraints on concordance measures in bivariate discrete data, J. Multivariate Anal. 93 (2005) 40–57.
[33] J. Doob, Stochastic Processes, Wiley, New York, 1953.
[34] R. Douc, S.T.F. Roueff, The maximizing set of the asymptotic normalized log-likelihood for partially observed Markov chains, Ann. Appl. Probab. 26 (2016) 2357–2383.
[35] P. Doukhan, Mixing: Properties and Examples, in: Lecture Notes in Statistics, vol. 85, Springer-Verlag, New York, 1994.
[36] A.F.R. Dürre, T. Liboschik, Robust estimation of (partial) autocorrelation, WIREs Comput. Stat. 7 (2015) 205–222.
[37] M. Eichler, Graphical modelling of multivariate time series, Probab. Theory Related Fields 153 (2012) 233–268.
[38] H. Elsaied, R. Fried, Robust fitting of inarch models, J. Time Series Anal. 35 (2014) 517–535.
[39] H. Elsaied, R. Fried, On robust estimation of negative binomial inarch models, Metron (2021) (in press).
[40] L. Fahrmeir, G. Tutz, Multivariate Statistical Modelling Based on Generalized Linear Models, second ed., Springer, New York, 2001.
[41] A. Farcomeni, Generalized linear mixed models based on latent Markov heterogeneity structures, Scand. J. Stat. 42 (2015) 1127–1135.
[42] R. Ferland, A. Latour, D. Oraichi, Integer–valued GARCH processes, J. Time Series Anal. 27 (2006) 923–942.
[43] K. Fokianos, Multivariate Count Time Series Modelling, Technical Report, 2021, Available at https://arxiv.org/abs/2103.08028.
[44] K. Fokianos, B. Kedem, Regression theory for categorical time series, Statist. Sci. 18 (2003) 357–376.
[45] K. Fokianos, A. Rahbek, D. Tjøstheim, Poisson autoregression, J. Amer. Statist. Assoc. 104 (2009) 1430–1439.
[46] K. Fokianos, B. Støve, D. Tjøstheim, P. Doukhan, Multivariate count autoregressions, Bernoulli 26 (2020) 471–499.
[47] K. Fokianos, D. Tjøstheim, Log–linear Poisson autoregression, J. Multivariate Anal. 102 (2011) 563–578.
[48] K. Fokianos, L. Truquet, On categorical time series models with covariates, Stochastic Process. Appl. 129 (2019).
[49] R. Frank, N. Davey, S. Hunt, Time series prediction and neural networks, J. Intell. Robot. Syst. 31 (2001) 91–103.
[50] J. Gamboa, Deep learning for time series analysis, 2017, arXiv:1701.01887.
[51] M.M.E.J. Garcia Ben, V. Yohai, Robust estimation in vector autoregressive moving average models, J. Time Series Anal. 20 (1999) 381–399.
[52] B. Geiger, T. Petrov, G. Kubin, H. Koeppl, Optimal Kullback-Leibler aggregation via information bottleneck, IEEE Trans. Automat. Control 60 (2015) 1010–1022.
[53] C. Genest, J. Nešlehová, A primer on copulas for count data, Astin Bull. 37 (2007) 475–515.
[54] V. Girardin, V. Konev, S. Pergamenchtchikov, Kullback-Leibler approach to CUSUM quickest detection rule for Markovian time series, Sequential Anal. 37 (2018) 322–341.
[55] G.K. Grunwald, R.J. Hyndman, L. Tedesco, R.L. Tweedie, Non-Gaussian conditional linear AR(1) models, Aust. N.Z. J. Stat. 42 (2000) 479–495.
[56] E.C. Hall, G. Raskutti, R.M. Willett, Learning high-dimensional generalized linear autoregressive models, IEEE Trans. Inform. Theory 65 (2019) 2401–2422.
[57] A.C. Harvey, C. Fernandes, Time series models for count or qualitative observations, J. Bus. Econom. Statist. 7 (1989) 407–422, With discussion.
[58] M. Hayashi, S. Watanabe, Information geometry approach to parameter estimation in Markov chains, Ann. Statist. 44 (2016) 1495–1535.
[59] A. Heinen, Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model, Technical Report MPRA Paper 8113, University Library of Munich, Germany, 2003, Availabel at http://mpra.ub.uni-muenchen.de/8113/.
[60] A. Heinen, E. Rengifo, Multivariate autoregressive modeling of time series count data using copulas, J. Empir. Financ. 14 (2007) 564–583.
[61] N.-J. Hsu, H.-L. Hung, Y.-M. Chang, Subset selection for vector autoregressive processes using lasso, Comput. Statist. Data Anal. 52 (2008) 3645–3657.

[62] D.I. Inouye, E. Yang, G.I. Allen, P. Ravikumar, A review of multivariate distributions for count data derived from the Poisson distribution, WIREs Comput. Stat. 9 (2017).
[63] P. Jacobs, P. Lewis, Discrete time series generated by mixtures I: correlational and runs properties, J. R. Stat. Soc. Ser. B Stat. Methodol. 40 (1978) 94–105.
[64] Y. Jia, S. Kechagias, J. Livsey, R. Lund, V. Pipiras, Count time series modeling with Gaussian copulas, 2020.
[65] H. Joe, Multivariate Models and Dependence Concepts, Chapman & Hall, London, 1997.
[66] N.L. Johnson, S. Kotz, N. Balakrishnan, Discrete Multivariate Distributions, John Wiley, New York, 1997.
[67] B. Jørgensen, S. Lundbye-Christensen, P.X.-K. Song, L. Sun, A state space model for multivariate longitudinal count data, Biometrika 86 (1999) 169–181.
[68] R. Jung, R. Liesenfeld, J.-F. Richard, Dynamic factor models for multivariate count data: an application to stock–market trading activity, J. Bus. Econom. Statist. 29 (2011) 73–85.
[69] D. Karlis, L. Meligkotsidou, Finite mixtures of multivariate Poisson distributions with application, J. Statist. Plann. Inference 137 (2007) 1942–1960.
[70] M.N. Katehakis, L.C. Smit, A successive lumping procedure for a class of Markov chains, Probab. Engrg. Inform. Sci. 26 (4) (2012) 483–508.
[71] B. Kedem, K. Fokianos, Regression Models for Time Series Analysis, Wiley, Hoboken, New Jersey, 2002.
[72] Y. Kharin, Robustness in Statistical Forecasting, Springer, New York, 2013.
[73] A. Kharin, Robustness of sequential testing of hypotheses on parameters of m-valued random sequences, J. Math. Sci. 189 (2013) 924–931.
[74] Y. Kharin, Statistical analysis of discrete-valued time series by parsimonious high-order Markov chains, Aust. J. Stat. 49 (2020) 76–88.
[75] Y. Kharin, M. Kislach, Statistical analysis of Poisson conditionally nonlinear autoregressive time series by frequencies-based estimators, Pattern Recognit. Image Anal. 30 (2020) 22–26.
[76] Y. Kharin, M. Maltsew, Statistical analysis of high-order dependencies, Acta Comment. Univ. Tartu. Math. 21 (2017) 37–45.
[77] Y. Kharin, A. Piatlitski, A Markov chain of order s with r partial connections and statistical inference on its parameters, Discrete Math. Appl. 17 (2007) 295–317.
[78] Y. Kharin, E. Vecherko, Statistical estimation of parameters for binary Markov chain models with embeddings, Discrete Math. Appl. 23 (2013) 153–169.
[79] Y. Kharin, V. Voloshko, Binomial conditionally nonlinear autoregressive model of discrete-valued time series and its probabilistic and statistical properties, Trans. Inst. Math. NAS Belarus 26 (2019) 95–105.
[80] Y. Kharin, V. Voloshko, Semibinomial conditionally nonlinear autoregressive models of discrete random sequences; probabilistic properties and statistical parameter estimation, Discrete Math. Appl. 30 (2020) 417–437.
[81] Y. Kharin, V. Voloshko, Statistical analysis of conditionally bionomial nonlinear regression time series with discrete regressors, Theory Probab. Math. Statist. 100 (2020) 181–190.
[82] Y. Kharin, V. Voloshko, Robust estimation for Binomial conditionally nonlinear autoregressive time series based on multivariate conditional frequencies, J. Multivariate Anal. 185 (2021) 104777.
[83] Y. Kharin, V. Voloshko, O. Dernakova, V. Malugin, A. Kharin, Statistical forecasting of the dynamics of epidemiological indicators for COVID-19 incidence in the Republic of Belarus, J. Belarus. State Univ. Math. Inform. 3 (2020) 36–50.
[84] Y. Kharin, V. Voloshko, E. Medved, Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series, Math. Methods Statist. 26 (2018) 103–118.
[85] Y. Kharin, M. Zhurak, Statistical analysis of spatio-temporal data based on Poisson conditional autoregressive model, Informatica 26 (2015) 67–87.
[86] B. Kim, S. Lee, Robust estimation for the covariance matrix of multi-variate time series, J. Time Series Anal. 32 (2011) 469–481.
[87] S. Kitromilidou, K. Fokianos, Mallows' quasi-likelihood estimation for log-linear Poisson autoregressions, Stat. Inference Stoch. Process. 19 (2016) 337–361.
[88] S. Kitromilidou, K. Fokianos, Robust estimation methods for a class of log-linear count time series models, J. Stat. Comput. Simul. 86 (2016) 740–755.
[89] S. Kocherlakota, K. Kocherlakota, Bivariate Discrete Distributions, Marcel Dekker, Inc., New York, 1992.
[90] I. Kontoyiannis, L. Mertzanis, A. Panotopoulou, I. Papageorgiou, M. Skoularidou, Bayesian context trees: Modelling and exact inference for discrete time series, 2020, Available at https://arxiv.org/abs/2007.14900.
[91] P. Koochemeshkian, N. Zamzami, N. Bouguila, Flexible distribution-based regression models for count data: Application to medical diagnosis, Cybern. Syst. 51 (2020) 442–466.
[92] B. L.-S. Kim, D. Kim, Robust estimation for bivariate Poisson INGARCH models, Entropy 23 (2021) 367.
[93] A. Latour, The multivariate GINAR(p) process, Adv. Appl. Probab. 29 (1997) 228–248.
[94] Y. Lee, S. Lee, D. Tjøstheim, Asymptotic normality and parameter change test for bivariate Poisson INGARCH models, Test 27 (2018) 52–69.
[95] H. Liu, Some Models for Time Series of Counts (Ph.D. thesis), Columbia University, USA, 2012.
[96] J. Livsey, R. Lund, S. Kechagias, V. Pipiras, Multivariate integer-valued time series with flexible autocovariances and their application to major hurricane counts, Ann. Appl. Stat. 12 (2018) 408–431.
[97] H. Lütkepohl, New Introduction to Multiple Time Series Analysis, Springer-Verlag, Berlin, 2005.
[98] V. Maevskii, Y.S. Kharin, Robust regressive forecasting under functional distortions in a model, Autom. Remote Control 63 (2002) 1803–1820.
[99] S. Maletz, Spatio-Temporal Models for Count Data (Master's thesis), TU Dortmund, Germany, 2021.
[100] P. Malhotra, L. Vig, G. Shroff, P. Agarwal, Long short term memory networks for anomaly detection in time series, in: European Symposium on Artificial Neural Networks (ESANN), 2015.
[101] A.W. Marshall, I. Olkin, Families of multivariate distributions, J. Amer. Statist. Assoc. 83 (1988) 834–841.
[102] P. McCullagh, J.A. Nelder, Generalized Linear Models, second ed., Chapman & Hall, London, 1989.
[103] S. Meyer, L. Held, Power-law models for infectious disease spread, Ann. Appl. Stat. 8 (2014) 1612–1639.
[104] T. Moysiadis, K. Fokianos, On binary and categorical time series models with feedback, J. Multivariate Anal. 131 (2014) 209–228.
[105] N. Muler, V. Yohai, Robust estimation for vector autoregressive models, Comput. Statist. Data Anal. 65 (2013) 68–79.
[106] R.B. Nelsen, An Introduction to Copulas, in: Lecture Notes in Statistics, vol. 139, Springer-Verlag, New York, 1999.
[107] J. Nicolau, A new model for multivariate Markov chains, Scand. J. Stat. 41 (2014) 1124–1135.
[108] A.K. Nikoloulopoulos, On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood, J. Statist. Plann. Inference 143 (2013) 1923–1937.
[109] L. Onsager, Crystal statistics. I. a two-dimensional model with an order–disorder transition, Phys. Rev. II 65 (1944) 117–149.
[110] A. Panagiotelis, C. Czado, H. Joe, Pair copula constructions for multivariate discrete data, J. Amer. Statist. Assoc. 107 (2012) 1063–1072.
[111] M. Papapetrou, D. Kugiumtzis, Markov chain order estimation with conditional mutual information, Physica A 392 (2013) 1593–1601.
[112] M. Paul, L. Held, A.M. Toschke, Multivariate modelling of infectious disease surveillance data, Stat. Med. 27 (2008) 6250–6267.
[113] X. Pedeli, D. Karlis, On composite likelihood estimation of a multivariate INAR(1) model, J. Time Series Anal. 34 (2013) 206–220.
[114] X. Pedeli, D. Karlis, Some properties of multivariate INAR(1) processes, Comput. Statist. Data Anal. 67 (2013) 213–225.

[115] P. Pfeifer, S. Deutsch, A three-stage iterative procedure for space–time modelling, Technometrics 22 (1980) 35–47.
[116] A. Raftery, A model for high-order Markov chains, J. R. Stat. Soc. Ser. B Stat. Methodol. 47 (1985) 528–539.
[117] A. Raftery, S. Tavare, Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model, J. Appl. Stat. B 43 (1994) 179–199.
[118] N. Ravishanker, R. Venkatesan, S. Hu, Dynamic models for time series of counts with a marketing application, in: R. Davis, S. Holan, R. Lund, N. Ravishanker (Eds.), Handbook of Discrete-Valued Time Series, in: Handbooks of Modern Statistical Methods, Chapman & Hall, London, 2015, pp. 425–446.
[119] L. Rüschendorf, Copulas, Sklar's theorem, and distributional transform, in: Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 3–34.
[120] B. Ryabko, J. Astola, M. Malytov, Compression-Based Methods of Statistical Analysis and Prediction of Time Series, Springer, 2016.
[121] T.H. Rydberg, N. Shephard, A modeling framework for the prices and times of trades on the New York stock exchange, in: W.J. Fitzgerlad, R.L. Smith, A.T. Walden, P.C. Young (Eds.), Nonlinear and Nonstationary Signal Processing, Isaac Newton Institute and Cambridge University Press, Cambridge, 2000, pp. 217–246.
[122] A. Sklar, Fonctions de répartition à n dimensions et leurs marges, Publ. Inst. Stat. Univ. Paris 8 (1959) 229–231.
[123] P.X.-K. Song, M. Li, Y. Yuan, Joint regression analysis of correlated data using Gaussian copulas, Biometrics 65 (2009) 60–68.
[124] B. Sturmfels, Geometry of higher-order Markov chains, J. Algebr. Stat. 3 (2012) 1–10.
[125] R.S. Tsay, Multivariate Time Series Analysis, John Wiley & Sons, Inc., Hoboken, New Jersey, 2014.
[126] A.E. Veraart, Modeling, simulation and inference for multivariate time series of counts using trawl processes, J. Multivariate Anal. 169 (2019) 110–129.
[127] C.H. Weiss, Properties of a class of binary ARMA models, Statistics 43 (2009) 131–138.
[128] C.H. Weiss, An Introduction to Discrete-Valued Time Series, Jonh Wiley, Hoboken, New Jersey, 2018.
[129] M. West, Bayesian forecasting of multivariate time series: scalability, structure uncertainty and decisions, Ann. Inst. Statist. Math. 72 (2020) 1–31.
[130] L. Yang, E.W. Frees, Z. Zhang, Nonparametric estimation of copula regression models with discrete outcomes, J. Amer. Statist. Assoc. 115 (2020) 707–720.
[131] S.L. Zeger, B. Qaqish, Markov regression models for time series: a quasi-likelihood approach, Biometrics 44 (1988) 1019–1031.
[132] D. Zhang, W.B. Wu, Gaussian approximation for high dimensional time series, Ann. Statist. 45 (2017) 1895–1919.
[133] Y. Zhang, H. Zhou, J. Zhou, W. Sun, Regression models for multivariate count data, J. Comput. Graph. Statist. 26 (2017) 1–13.
[134] Y. Zheng, Q. Liu, E. Chen, Y. Ge, J. Zhao, Time Series Classification using Multi-Channels Deep Convolutional Neural Networks, in: Time Series Classification using Multi-Channels Deep Convolutional Neural Networks, Springer, 2014, pp. 298–310.
[135] X. Zhu, R. Pan, G. Li, Y. Liu, H. Wang, Network vector autoregression, Ann. Statist. 45 (2017) 1096–1123.