Statistics
A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: http://www.tandfonline.com/loi/gsta20

Time series analysis of categorical data using auto-odds ratio function

Raju Maiti & Atanu Biswas

To cite this article: Raju Maiti & Atanu Biswas (2018): Time series analysis of categorical data using auto-odds ratio function, Statistics, DOI: 10.1080/02331888.2017.1421196

To link to this article: https://doi.org/10.1080/02331888.2017.1421196

Published online: 23 Jan 2018.

STATISTICS, 2018
https://doi.org/10.1080/02331888.2017.1421196

Time series analysis of categorical data using auto-odds ratio function

Raju Maiti[a] and Atanu Biswas[b]

[a]Centre for Quantitative Medicine, Duke-NUS Graduate Medical School, Singapore; [b]Applied Statistics Unit, Indian Statistical Institute, Kolkata, India

ABSTRACT

In this paper, we consider the auto-odds ratio function (AORF) as a measure of serial association for a stationary time series process of categorical data at two different time points. Numerical measures such as the autocorrelation function (ACF) have no meaningful interpretation, unless the time series data are numerical. Instead, we use the AORF as a measure of association to study the serial dependency of the categorical time series for both ordinal and nominal categories. Biswas and Song [Discrete-valued ARMA processes. Stat Probab Lett. 2009;79(17):1884–1889] provided some results on this measure for Pegram's operator-based AR(1) process with binary responses. Here, we extend this measure to more general set-ups, i.e. for AR(p) and MA(q) processes and for a general number of categories. We discuss how this method can effectively be used in parameter estimation and model selection. Following Weiß [Empirical measures of signed serial dependence in categorical time series. J Stat Comput Simul. 2011;81(4):411–429], we derive the large sample distribution of the estimator of the AORF under independent and identically distributed (iid) set-up. Some simulation results and two categorical data examples (one is ordinal and other nominal) are presented to illustrate the proposed method.

ARTICLE HISTORY
Received 18 June 2014
Accepted 31 August 2015

KEYWORDS
Pegram's operator; auto-odds ratio function; autocorrelation function; categorical time series

1. Introduction

Time series of discrete-valued data can broadly be classified into two classes, namely count time series and categorical time series. Time series of categorical data can again be of two types: one ordinal, where the categories are of ordinal nature (such as 'poor', 'mediocre', 'good' and 'excellent' for marks), and the other nominal, where the categories cannot be ordered (such as a, c, t and g of a genome structure). A process $\{Y_t\}_{t\in\mathbb{N}}$ is categorical means each random variable Y_t takes one of the finite number of categories. Categorical time series arise in various fields such as biological sequence analysis, speech recognition, part-of-speech tagging, network monitoring and many further areas (see, e.g. [1–3]). However, the problem lying with the categorical time series analysis is the assignment of numerical values to the categories. Few attempts have been made towards building categorical data analysis, especially towards categorical time series analysis. For example, Jacobs and Lewis [4–6] in a series of papers introduced a simple method for obtaining a stationary sequence of dependent random variables with a specified marginal distribution and dependence structure chosen independently. It was perhaps the first attempt to obtain a general class of simple models for discrete variate time series data including categorical time series data. These models are structurally based on the well-known autoregressive moving average models and are referred to as DARMA models.

CONTACT Raju Maiti raju.maiti@duke-nus.edu.sg, rajumaiti@gmail.com

In another attempt, Raftery [7] proposed a mixture transition distribution model (MTDM), a class of models based on a time homogeneous Markov Chain for modelling categorical time series. Later, it has been generalized by Berchtold and Raftery [8]. The major problem with the MTDM is that the estimation of the parameters are to be done by computation-intensive methods like numerical maximization of the log likelihood or the EM algorithm, which often has problems due to the presence of multiple local optima. On the other hand, Pegram [9] proposed a discrete AR(p) model which is a special kind of Markovian model applicable to both count and categorical time series. Later, Biswas and Song [10] extended this model to MA(q) and ARMA(p, q) models. It is to be noted that the extension by Biswas and Song [10] is equivalent to the NDARMA model of Jacobs and Lewis [11]. The autocorrelation structure of these models resembles the representation of Box and Jenkins' AR(p) models and allows some of the correlations to be negative. The major advantages of these models are its distribution free nature and very few parameters involved. It is important to note that the AR(p) model proposed by Pegram [9] is equivalent to the DAR(p) model introduced by Jacobs and Lewis [6,11]. In particular, the DAR(1) model of Jacobs and Lewis [4] is exactly the same as that of Pegram's AR(1) model.

When the concept of correlation is appropriate, Biswas and Song [10] used Pegram's operator to construct a stationary ARMA(p,q) model, a model equivalent to the NDARMA model of Jacobs and Lewis [11]. However, it is important to note that the concept of correlation may not be appropriate in many categorical set-ups. In particular, if the categorical data are nominal type, the correlation cannot be defined. Even if the categories are ordinal, the value of the correlation will change dramatically with numbering of categories. Instead, we may use some popular measures such as Cohen's κ, Cramer's ν, Goodman and Kruskal's τ (see, e.g. [2,12,13]) and the auto-mutual information proposed by Biswas and Guha [3] or some general association measures of serial dependence given in [14]. On the other hand, to reduce the amount of arbitrariness incurred by numerical scaling to the categories, Fokianos and Kedem [15] expressed the categories by a vector representation and used multiple ACF plots to study the serial association, which has recently been used by Maiti and Biswas [16]. Heagerty and Zeger [17], while studying the dependence in longitudinal data, used a different serial measure of dependence based on the auto-odds ratio function. In the recent past, Biswas and Song [10] used the same idea while studying the serial dependence in the case of categorical time series data. However, Biswas and Song's work was restricted to the stationary Pegram's AR(1) model with two categories.

In this article, we extend this AORF, denoted by $\theta(h)$, to stationary Pegram's AR(p) and MA(q) models which are equivalent to DAR(p) and DMA(q) models of Jacobs and Lewis [4–6], respectively. In view of our previous arguments on arbitrariness of numerical scaling for categorical time series, moments and autocorrelation have no meaningful interpretation and therefore Yule-Walker (YW)-type estimation cannot be used to estimate the model parameters. However, we show that our measure can effectively be used to estimate model parameters in such cases. Performance of the proposed method is compared with the maximum likelihood estimation and the estimation based on Cohen's κ measure using some simulated data. Thereafter, we study the finite sample behaviour of the AORF for various lagged values h. Since the AORF can take any value between $(0, \infty)$, we can transform this measure to logarithmic scale in order to make its range to the whole real line $(-\infty, \infty)$. We may also make it to a finite measure like auto-Yule's Q measure by transforming $(\theta(h) - 1)/(\theta(h) + 1)$, which lies between -1 and $+1$. Some simulated data and two categorical data examples, namely infant sleep status data and Champions league data show that the proposed method can effectively be used for categorical time series data.

The article is presented as follows. In Section 2, we define the AORF $\theta(h)$ along with its logarithmic transformation and auto-Yule's Q measure. The auto-odds ratio measure for a stationary AR(p) process is studied in Section 3. In Section 4, the measure is studied for a stationary MA(q) process. In Section 5, we employ the proposed measure to estimate model parameters and to select the order of the AR(p) process. Under the assumption that data are generated from iid distribution, we derive a large sample distribution of the estimator of $\theta(h)$ in Section 6. Some simulation experiments are carried out in Section 7. In Section 8, we analyse two categorical time series to illustrate the proposed method. We conclude with some discussions in Section 8. All the proofs are relegated to the appendix.

2. Auto-odds ratio measure

For studying dependency in categorical longitudinal responses, Heagerty and Zeger [17] probably were the first to use the measure of dependency based on marginal pairwise log-odds ratio which they termed as lorelogram. They used parametric and nonparametric additive models to study the log-odds measure. However, for a univariate stationary time series of categorical data, such as a, c, t and g of a genome sequence or performance of a student observed over a certain period of time, Biswas and Song [10] used the same measure to study the serial association. In order to do that, they considered Pegram's operator-based AR(1) model with binary response. Here, we generalize the work of Biswas and Song [10] from binary to finite number of categories. Not only that, we extend it to a general AR(p) and MA(q) processes.

Let $\theta(h)$ denote the AORF between Y_t and Y_{t+h}. Then the function, $\theta(h)$, of lag h for a process, say $\{Y_t\}_{t\in\mathbb{N}}$, with only two categories $\{C_0, C_1\}$ can be defined as

$$\theta(h) = \frac{P(Y_t = C_0, Y_{t-h} = C_0)P(Y_t = C_1, Y_{t-h} = C_1)}{P(Y_t = C_0, Y_{t-h} = C_1)P(Y_t = C_1, Y_{t-h} = C_0)}, \quad h \geq 1. \tag{1}$$

Generalizing it to the case with a finite number of categories (three or larger), here we define the measure as follows. For a pair of categories indexed by (C_i, C_j), $i \neq j$,

$$\begin{aligned} \theta_{ij}(h) &= \frac{P(Y_t = C_i, Y_{t-h} = C_i)P(Y_t = C_j, Y_{t-h} = C_j)}{P(Y_t = C_i, Y_{t-h} = C_j)P(Y_t = C_j, Y_{t-h} = C_i)}, \quad h \geq 1,\ i,j = 0,1,\dots,k, \\ &= \frac{p_{ii}(h)p_{jj}(h)}{p_{ij}(h)p_{ji}(h)}, \end{aligned} \tag{2}$$

where $p_{ij}(h) = P(Y_t = C_i, Y_{t-h} = C_j)$. Combining all the $\theta_{ij}(h)$ for all $i,j,\ i \neq j$ with some weight functions $w_{ij}(h)$, we define a weighted auto-odds ratio function which is as follows:

$$\theta(h) = \sum\sum_{0\leq i\neq j\leq k} \theta_{ij}(h) w_{ij}(h), \tag{3}$$

where $w_{ij}(h)$ is some normalized weights. Throughout this article, we take $w_{ij}(h) = C \cdot P(Y_t = C_i, Y_{t-h} = C_j)$, where C is the normalizing constant such that $\sum\sum_{0\leq i\neq j\leq k} w_{ij}(h) = 1$, and therefore $w_{ij}(h)$ has the form:

$$w_{ij}(h) = \frac{P(Y_t = C_i, Y_{t-h} = C_j)}{\sum\sum_{0\leq s\neq l\leq k} P(Y_t = C_s, Y_{t-h} = C_l)}. \tag{4}$$

Note that the AORF defined in Equation (3), can take any value between 0 and ∞ and therefore we can transform this measure to $\log\theta(h)$, which is the initial measure considered by Heagerty and Zeger [17] for studying longitudinal data, to reduce the scale value. Furthermore, we can also make it a finite measure, namely auto-Yule's Q measure which is defined as follows:

$$Q(h) = \frac{\theta(h) - 1}{\theta(h) + 1} \in (-1, 1), \quad h \geq 1. \tag{5}$$

However, in all our simulation experiments and data analysis, we use the logarithm of $\theta(h)$. In the next few sections, we derive the exact form of $\theta(h)$ for Pegram's operator-based AR(p) and MA(q) models, and apply the measure to estimate the model parameters and to select the order of the AR(p) process. We study its finite sample behaviour using Monte Carlo simulations.

3. Pegram's AR(p) process

Pegram's operator *, when operated on U and V , say, defines a new random variable Z as a mixture of U and V with mixing coefficients ϕ and $1-\phi$. This is defined as follows:

$$Z = (U, \phi) * (V, 1-\phi),$$

where the marginal probability function of Z is given by

$$P(Z=j) = \phi P(U=j) + (1-\phi)P(V=j), \quad j = 0, 1, \ldots .$$

Pegram's construction has been extended to the ARMA(p,q) model by Biswas and Song [10] and Biswas and Guha [3]. The key advantage of Pegram's operator is that it provides a flexible mixing operation that enables us to define the mixture among a finite number of probability distributions of categorical random variables.

Pegram's operator-based AR(p) model due to Pegram [9] was first developed for time series with finite number of categories. Biswas and Song [10] extended this model from a finite number of categories to count time series. Here we present the model for finite number of categories which is the original model proposed by Pegram [9]. Let $\{Y_t\}_{t\in\mathbb{N}}$ be a categorical time series with $(k+1)$ categories, namely $\{C_0, C_1, \ldots, C_k\}$. Then Pegram's operator-based AR(p) model, which is equivalent to the DAR(p) model of Jacobs and Lewis [4], is defined as follows:

$$Y_t = (I(Y_{t-1}), \phi_1) * (I(Y_{t-2}), \phi_2) * \cdots * (I(Y_{t-p}), \phi_p) * (\epsilon_t, 1-\phi_1-\phi_2-\cdots-\phi_p), \qquad (6)$$

which is a mixture of $(p+1)$ discrete distributions, where $P(\epsilon_t = C_i) = \pi_i$, $i = 0, 1, \ldots, k$, and it is denoted by $\epsilon_t \sim D((C_i, \pi_i),\ i = 0, 1, \ldots, k)$, with respective mixing weights being $\phi_1, \ldots, \phi_p$ with $\phi_i \in (0,1)$, $i = 1, 2, \ldots, p$, and $\sum_{i=1}^{p} \phi_i \in (0,1)$. For every $t = 0, \pm 1, \pm 2, \ldots,$ the conditional probability function takes the form

$$\begin{aligned} P(Y_t = C_i \mid Y_{t-1}, \ldots, Y_{t-p}) &= \phi_1 I(Y_{t-1} = C_i) + \cdots + \phi_p I(Y_{t-p} = C_i) \\ &\quad + (1-\phi_1-\phi_2-\cdots-\phi_p)\pi_i, \end{aligned} \qquad (7)$$

where ϕ_i, $i = 1, \ldots, p$, are chosen such that the polynomial equation $1-\phi_1 z-\cdots-\phi_p z^p = 0$ has roots lying outside of the unit disc. Here $I(A)$ is the indicator function such that $I(A) = 1$ or 0 accordingly A occurs or not.

Taking expectation in both sides of Equation (7), we observe that $P(Y_{t-h} = C_i) = \pi_i$ for $h = 1, \ldots, p$, resulting in $P(Y_t = C_i) = \pi_i$, which implies the marginal stationarity of the process, i.e. marginally $Y_t \sim D((C_i, \pi_i), i = 0, 1, \ldots, k)$ for all t.

Theorem 3.1: *For a stationary Pegram's AR(1) process with $(k+1)$ categories, the AORF between Y_t and Y_{t-h} can be derived as follows:*

$$\begin{aligned} \theta(h) &= 1 + \frac{k(k+1)\phi^{2h} + 2k\phi^h(1-\phi^h)}{(1-\phi^h)^2(1-\boldsymbol{\pi}^T\boldsymbol{\pi})} \\ &= g(\phi^h), \quad h = 1, 2, \ldots, \end{aligned} \qquad (8)$$

where $g(\phi^h)$ is an increasing function of ϕ.

Proof: Proof of this result is presented in Appendix 1. ■

To derive this measure for the AR(p) process, we use the following result derived from Maiti and Biswas [16]. For a stationary AR(p) process, the h-step ahead conditional distribution of Y_{n+h} given $\mathbf{Y}_n = (Y_n, Y_{n-1}, \ldots, Y_1)^{\mathrm{T}}$ is given by

$$\begin{aligned} p_h(C_i \mid \mathbf{Y}_n, \boldsymbol{\phi}) &= P(Y_{n+h} = C_i \mid Y_n, \ldots, Y_{n-p+1}) \\ &= \eta_{h1} I(Y_n = C_i) + \cdots + \eta_{hp} I(Y_{n-p+1} = C_i) + (1 - \eta_{h1} - \cdots - \eta_{hp})\pi_i, \end{aligned} \tag{9}$$

where the vector of h-step ahead parameters $\boldsymbol{\eta}_h = (\eta_{h1}, \eta_{h2}, \ldots, \eta_{hp})^{\mathrm{T}}$ is given by

$$\boldsymbol{\eta}_h = \boldsymbol{\Phi}^{h-1}\boldsymbol{\phi},$$

with

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & \phi_2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \\ 0 & 0 & 0 & \cdots & \phi_{p-1} & 1 \\ 0 & 0 & 0 & \cdots & 0 & \phi_p \end{pmatrix}, \quad \boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\Phi}^{h-1} = \underbrace{\Phi \times \Phi \times \cdots \times \Phi}_{h-1}.$$

Using this result, we derive the AORF for a stationary AR(p) process as follows.

Theorem 3.2: *For a stationary AR(p) process defined in Equation* (6) *with* $(k+1)$ *many categories, the AORF between* Y_t *and* Y_{t-h} *can be derived as follows:*

$$\theta(h) = 1 + \frac{k f_2(\boldsymbol{\phi})}{f_1^2(\boldsymbol{\phi})(1 - \boldsymbol{\pi}'\boldsymbol{\pi})}\{(k+1)f_2(\boldsymbol{\phi}) + 2f_1(\boldsymbol{\phi})\}, \tag{10}$$

where

$$f_1(\boldsymbol{\phi}) = \boldsymbol{\eta}_h^{(2)'}(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{b} + (1 - \mathbf{1}'\boldsymbol{\eta}_h), \quad f_2(\boldsymbol{\phi}) = \eta_{h1} + \boldsymbol{\eta}_h^{(2)'}(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\delta}_{p-1}$$

with

$$\boldsymbol{A} = \begin{pmatrix} \eta_{12} & \eta_{13} & \cdots & \eta_{1p} \\ \eta_{22} & \eta_{23} & \cdots & \eta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{(p-1)2} & \eta_{(p-1)3} & \cdots & \eta_{(p-1)p} \end{pmatrix}^{(p-1)\times(p-1)}, \quad \boldsymbol{b} = \begin{pmatrix} 1 - \mathbf{1}'\boldsymbol{\eta}_1 \\ 1 - \mathbf{1}'\boldsymbol{\eta}_2 \\ \vdots \\ 1 - \mathbf{1}'\boldsymbol{\eta}_{p-1} \end{pmatrix}^{(p-1)\times 1},$$

$$\boldsymbol{\eta}_h = \begin{pmatrix} \eta_{h1} \\ \eta_{h2} \\ \vdots \\ \eta_{hp} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta}_h^{(1)} \\ \boldsymbol{\eta}_h^{(2)} \end{pmatrix}, \quad \boldsymbol{\eta}_h^{(1)} = \eta_{h1}, \quad \boldsymbol{\eta}_h^{(2)} = \begin{pmatrix} \eta_{h2} \\ \eta_{h3} \\ \vdots \\ \eta_{hp} \end{pmatrix}, \quad \boldsymbol{\delta}_{p-1} = \begin{pmatrix} \eta_{11} \\ \eta_{21} \\ \vdots \\ \eta_{(p-1)1} \end{pmatrix}, \quad \textit{and}$$

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix}.$$

Proof: Proof of this theorem is relegated to Appendix 2. ■

4. Pegram's MA(q) process

A stationary Pegram's operator-based MA(q) process proposed by Biswas and Song [10] is defined as

$$Y_t = (\epsilon_t, \theta_0) * (I(\epsilon_{t-1}), \theta_1) * \cdots * (I(\epsilon_{t-q}), \theta_q), \tag{11}$$

which implies that for every $t \in 0, \pm 1, \pm 2, \ldots$, the conditional probability function takes the form

$$P(Y_t = C_i \mid \epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-q}) = \theta_0 I(\epsilon_t = C_i) + \theta_1 I(\epsilon_{t-1} = C_i) + \ldots + \theta_q I(\epsilon_{t-q} = C_i),$$

where $\theta_i \geq 0$ for all i, and $\sum_{i=0}^{q} \theta_i = 1$. It is easy to see that marginally $Y_t \sim D\{(C_i, \pi_i),\ i = 0, 1, \ldots, k\}$ for all t. Note that the above model is equivalent to DMA(q) model of Jacobs and Lewis [4].

In particular, auto-covariance function (ACVF) of lag h if it is defined, can be written as follows:

$$\gamma(h) = \begin{cases} \sum_{r=0}^{q-h} \theta_r \theta_{r+h} & \text{if } 0 \leq h \leq q, \\ 0 & \text{if } h > q. \end{cases}$$

See, e.g. Jacobs and Lewis [4] for the derivation. But if the time series is categorical in nature, then the above function will no longer be useful to measure the serial auto association. However, to study the serial association for the categorical MA(q) process, we derive the following result based on the auto-odds ratio.

Theorem 4.1: *For a stationary MA(q) process defined in Equation* (11) *with* $(k+1)$ *categories, the AORF between* Y_t *and* Y_{t-h} *can be written as follows:*

$$\theta(h) = 1 + \frac{k(k+1)\left(\sum_{r=0}^{q-h} \theta_r \theta_{r+h}\right)^2 + 2k\left(\sum_{r=0}^{q-h} \theta_r \theta_{r+h}\right)\left(1 - \sum_{r=0}^{q-h} \theta_r \theta_{r+h}\right)}{\left(1 - \sum_{r=0}^{q-h} \theta_r \theta_{r+h}\right)^2 (1 - \boldsymbol{\pi}^T \boldsymbol{\pi})}$$
$$= f\left(\sum_{r=0}^{q-h} \theta_r \theta_{r+h}\right), \quad 0 \leq h \leq q,$$

where $f(\sum_{r=0}^{q-h} \theta_r \theta_{r+h})$ *is an increasing function of* $\sum_{r=0}^{q-h} \theta_r \theta_{r+h}$ *and* $\theta(h) = 0$ *for* $h > q$.

Proof: Proof of this theorem is presented in Appendix 3. ■

5. Parameter estimation and model selection

Note that the objective here is to fit a time series model for categorical data observed over time, like condition of a patient recorded over every hour as good, fair, serious, and severe. In such cases, the proposed measure bypasses the numerical scaling and analyses the data directly and effectively. Even for the ordinal categorical time series where one can assign some numerical values to the categories, since there is no unique way to assign numerical values, the value of the ACF changes drastically with the numerical scaling of the categories. For example, if $C_0, C_1, \ldots, C_k$ are $(k+1)$ categories of the process $\{Y_t\}$, then the ACF between Y_t and Y_{t-h} may vary for two sets of (nonlinearly related) scalings of the categories. As a result, parameter estimation using YW-type method may lead to spurious inference, or two people using two different numerical scalings may end up with two different YW estimates of the same parameter. As an alternative, in this section, we discuss the parameter estimation for the AR(p) process based on the AORF $\theta(h)$. This estimation method is applied in the subsequent simulation and data analysis sections and we compare its efficacy with the maximum likelihood estimation (MLE) and estimation based on Cohen's $\kappa(h)$ measure with respect to both bias and standard error (se).

Another important issue with the categorical time series is model selection which can be done by using the Akaike information criteria (AIC) and Bayesian information criteria (BIC). These criteria have been used extensively in [10,16] for studying model comparison in the context of the categorical time series analysis. However, in this article, we select the order of the AR(p) process based on the sample behaviour of the AORF over various lagged values.

Let $\{Y_t\}$ be a categorical time series process with $(k+1)$ categories, namely $\{C_0, C_1, \ldots, C_k\}$, and let $\{Y_1, Y_2, \ldots, Y_n\}$ be a sample of size n from the process Y_t. Define

$$\begin{aligned} N_i &= \text{the number of } \{Y_t = C_i\} \text{ in the sample}, \\ N_{ij}(h) &= \text{the number of } \{(Y_t, Y_{t-h}) = (C_i, C_j)\} \text{ in the sample}. \end{aligned}$$

Then, the sample proportionate estimators for the π_i and $p_{ij}(h)$ are given by

$$\hat{\pi}_i = \frac{1}{n} \cdot N_i \quad \text{and} \quad \hat{p}_{ij}(h) = \frac{1}{n-h} \cdot N_{ij}(h), \tag{12}$$

respectively. Plugging (12) in Equations (2) and (4), we get

$$\hat{\theta}_{ij}(h) = \frac{N_{ii}(h)N_{jj}(h)}{N_{ij}(h)N_{ji}(h)},$$

and

$$\hat{w}_{ij}(h) = \frac{N_{ij}(h)}{\sum\sum_{0\le r\neq s\le k} N_{rs}(h)}.$$

Hence, the estimator of the combined AORF $\theta(h)$ is given by

$$\hat{\theta}_n(h) = \sum\sum_{0\le i\neq j\le k} \hat{\theta}_{ij}(h)\hat{w}_{ij}(h). \tag{13}$$

For the AR(1) process, equating (13) with Equation (8), we can get $\hat{\phi}_{aor} = g^{-1}(\hat{\theta}_n(1))$. Although, for the AR(1) process, we can see that there is no identifiability problem in estimating the mixing parameter ϕ using the AORF method, however for the AR(p) process, it is not so straight forward. In order to estimate $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^{\mathrm{T}}$ for the AR(p) process, we need to solve p nonlinear equations which might suffer from identifiability crisis.

6. Large sample distribution of $\hat{\theta}_n(h)$

To derive the asymptotic distribution of the estimator $\hat{\theta}_n(h)$ of $\theta(h)$ for the general AR(p) process, we assume that the data are coming from the iid process. Note that, in the context of the categorical time series analysis, the large sample properties of serial measures, such as Cohen's κ, Cramer's ν, Goodman and Kruskal's τ, have been extensively studied by Weiß and Göb [2], Weiß [12,13]. Large sample asymptotic distributions of those measures under various set-ups (e.g. under serial independence for NDARMA process) with more than two categories have been derived in [12,13]. Under the iid set-up, here we derive the asymptotic normality of $\hat{\theta}_n(h)$ and the result can be given as follows.

Let $P(h)$ denote the transition probability matrix of order $(m+1) \times (m+1)$ and is defined as follows:

$$P(h) = ((p_{ij}(h)))_{i,j=0}^{m},$$

where $p_{ij}(h) = P(Y_t = C_i, Y_{t-h} = C_j)$ is the joint probability distribution of Y_t and Y_{t-h}. A natural estimator of $P(h)$ based on sample proportions is $\hat{P}_n(h) = ((\hat{p}_{ij}(h)))_{i,j=0}^{m}$, where the sample proportion estimators $\hat{p}_{ij}(h)$'s are given in Equation (12). Also let $vec(P(h))$, the vectorization of the matrix $P(h)$, denote the $(m+1)^2 \times 1$ column vector obtained by stacking the columns of the matrix $P(h)$ on top of one another which can be presented as follows:

$$\mathrm{vec}(P(h)) = [p_{00}(h), p_{10}(h), \ldots, p_{m0}(h), p_{01}(h), \ldots, p_{m1}(h), \ldots, p_{0m}(h), \ldots, p_{mm}(h)]^{\mathrm{T}}.$$

Similarly, we can define $\mathrm{vec}(\hat{P}_n(h))$, an estimator of $\mathrm{vec}(P(h))$ by replacing each $p_{ij}(h)$ by $\hat{p}_{ij}(h)$ in $\mathrm{vec}(P(h))$. It can be shown that $\mathrm{vec}(\hat{P}_n(h))$ converges in probability to $\mathrm{vec}(P(h))$ as n goes to ∞. This is because the variance–covariance matrix of $\mathrm{vec}(\hat{P}_n(h))$, denoted by $\Sigma_n(h)$, goes to $\mathbf{0}$ as n increases to ∞ (see the derivation given in Appendix 4). Since $\hat{\theta}_n(h)$ is a continuous function of $\mathrm{vec}(\hat{P}_n(h))$, using the delta-method, we can show that

$$E(\hat{\theta}_n(h)) = \theta(h) \quad \text{and} \quad V(\hat{\theta}_n(h)) = \frac{1}{n-h}(\nabla\theta(h))^{\mathrm{T}}\Sigma_n(h)(\nabla\theta(h)),$$

where $\nabla\theta(h)$ is $(m+1)^2 \times 1$ vector of partial derivatives with respect to $p_{ij}(h)$, and hence we have $\sqrt{n-h}(\hat{\theta}_n(h) - \theta(h)) \overset{a}{\sim} N(0, (\nabla\theta(h))^{\mathrm{T}}\Sigma_n(h)(\nabla\theta(h)))$.

7. Simulation study

In this section, we use some simulation experiments to evaluate the proposed measure numerically. To begin with, we generated samples from Pegram's AR(1) process with four categories and for various values of the mixing parameter ϕ lies between 0 and 1. The marginal probability distribution $\boldsymbol{\pi} = (0.2, 0.35, 0.3, 0.15)$ is chosen arbitrarily. However, one can select some other values of $\boldsymbol{\pi}$ and do the same exercise. For a fixed value of ϕ and above chosen $\boldsymbol{\pi}$, we simulated samples of size $n = 500$ and computed the empirical estimates of $\log\theta(h)$ for varying h. Based on 1000 Monte Carlo repetitions, we averaged the values of $\log\theta(h)$ for varying h and plotted in Figure 1. As we can see, the AORF decreases as h increases, which implies that the auto-association between Y_t and Y_{t-h} decreases as they go away from each other, which is quite expected.

To compare the finite sample performance of AORF-based estimation with MLE and Cohen's κ-based estimation, we generated samples of sizes $n = 100$, 500, 1000, 5000 and 10,000 from the above-mentioned process. Note that samples of sizes 100 and 500 are explored to study the small sample properties, whereas 5000 and 10,000 give some idea about large sample properties and a sample of size 1000 reflects the moderate sample properties. For a fixed sample size and fixed value of ϕ, we computed the estimated ϕ, its bias and standard error (se) using all the three methods; and averaged them over 1000 Monte Carlo repetitions. The results are presented in Table 1. Note that, maximum likelihood estimates are observed using the likelihood function given in Pegram [9], andCohen's κ-based estimates are obtained using the formula given in [2]. As we can see, for all the values of ϕ and sample sizes, MLE dominates the other two methods, especially for smaller sample sizes MLE performs better than the other twowith respect to both bias and se. Given the true likelihood function, such results (i.e. dominated nature of MLE) are quite expected. On the other hand, as we can see from Table 1, for smaller sample sizes Cohen's κ-based estimate performs better than our proposed measure with respect to both bias and se. However, as sample size increases, both methods appear to be equally effective. Same kind of scenario is observed from Figure 2 which displays the mean squared error (MSE) of $\hat{\phi}$ for the above three estimation methods.

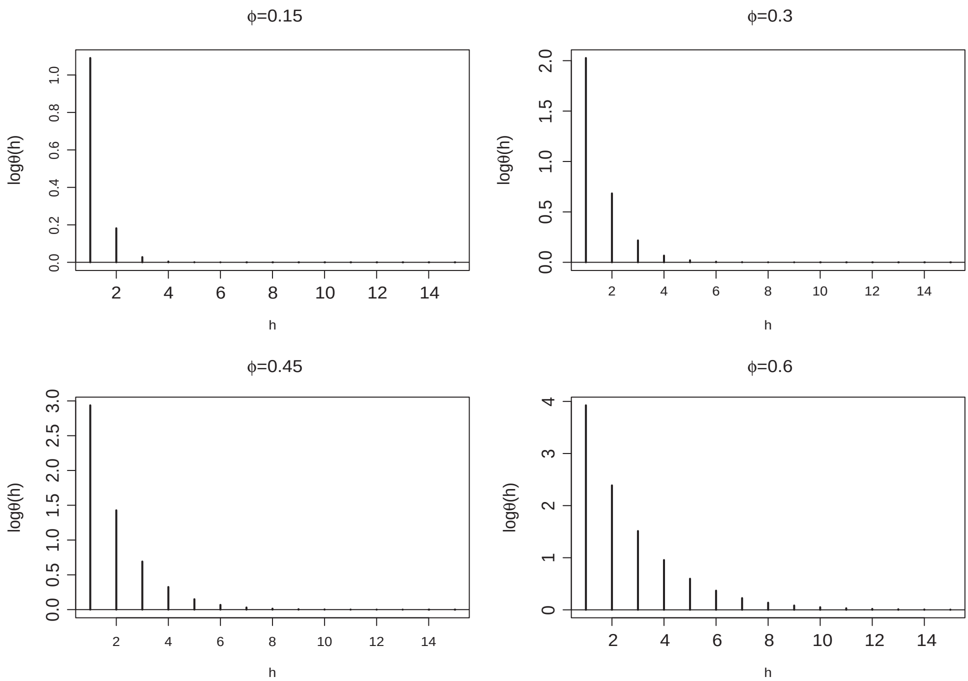

Figure 1. Plots of sample $\log\theta(h)$ for the AR(1) process with $\phi = 0.15, 0.3, 0.45, 0.6$.

Table 1. Sample estimates with its bias and standard error (se) by three different methods of estimation.

	$\widehat{\phi}\begin{pmatrix}\text{bias}\\\text{se}\end{pmatrix}$			$\widehat{\phi}\begin{pmatrix}\text{bias}\\\text{se}\end{pmatrix}$		
Sample size (n)	MLE	Cohen's κ	AOR	MLE	Cohen's κ	AOR
		$\phi = 0.15$			$\phi = 0.30$	
200	0.130 $\begin{pmatrix}0.00681\\0.03392\end{pmatrix}$	0.130 $\begin{pmatrix}0.00753\\0.03673\end{pmatrix}$	0.160 $\begin{pmatrix}0.01529\\0.03969\end{pmatrix}$	0.290 $\begin{pmatrix}0.00672\\0.04146\end{pmatrix}$	0.287 $\begin{pmatrix}0.00719\\0.04311\end{pmatrix}$	0.327 $\begin{pmatrix}0.01908\\0.05646\end{pmatrix}$
500	0.151 $\begin{pmatrix}0.00531\\0.02202\end{pmatrix}$	0.150 $\begin{pmatrix}0.00582\\0.02477\end{pmatrix}$	0.160 $\begin{pmatrix}0.01124\\0.02504\end{pmatrix}$	0.294 $\begin{pmatrix}0.00146\\0.02140\end{pmatrix}$	0.293 $\begin{pmatrix}0.00229\\0.02614\end{pmatrix}$	0.308 $\begin{pmatrix}0.00915\\0.02935\end{pmatrix}$
1000	0.145 $\begin{pmatrix}0.00416\\0.02362\end{pmatrix}$	0.145 $\begin{pmatrix}0.00436\\0.02390\end{pmatrix}$	0.150 $\begin{pmatrix}0.00457\\0.02423\end{pmatrix}$	0.300 $\begin{pmatrix}0.00101\\0.02490\end{pmatrix}$	0.300 $\begin{pmatrix}0.00105\\0.02639\end{pmatrix}$	0.308 $\begin{pmatrix}0.00145\\0.02753\end{pmatrix}$
5000	0.148 $\begin{pmatrix}0.00141\\0.00935\end{pmatrix}$	0.148 $\begin{pmatrix}0.00145\\0.00942\end{pmatrix}$	0.149 $\begin{pmatrix}0.00143\\0.00941\end{pmatrix}$	0.299 $\begin{pmatrix}0.00151\\0.00869\end{pmatrix}$	0.299 $\begin{pmatrix}0.00162\\0.00884\end{pmatrix}$	0.301 $\begin{pmatrix}0.00167\\0.00902\end{pmatrix}$
10,000	0.150 $\begin{pmatrix}0.00007\\0.00559\end{pmatrix}$	0.150 $\begin{pmatrix}0.00009\\0.00569\end{pmatrix}$	0.150 $\begin{pmatrix}0.00019\\0.00571\end{pmatrix}$	0.299 $\begin{pmatrix}0.00112\\0.00677\end{pmatrix}$	0.299 $\begin{pmatrix}0.00116\\0.00681\end{pmatrix}$	0.300 $\begin{pmatrix}0.00114\\0.00688\end{pmatrix}$
		$\phi = 0.45$			$\phi = 0.60$	
200	0.449 $\begin{pmatrix}0.00381\\0.04216\end{pmatrix}$	0.443 $\begin{pmatrix}0.00923\\0.04228\end{pmatrix}$	0.489 $\begin{pmatrix}0.03703\\0.05673\end{pmatrix}$	0.590 $\begin{pmatrix}0.00931\\0.04549\end{pmatrix}$	0.583 $\begin{pmatrix}0.01905\\0.04974\end{pmatrix}$	0.616 $\begin{pmatrix}0.02792\\0.05804\end{pmatrix}$
500	0.443 $\begin{pmatrix}0.00916\\0.03126\end{pmatrix}$	0.442 $\begin{pmatrix}0.01219\\0.03162\end{pmatrix}$	0.464 $\begin{pmatrix}0.02686\\0.03348\end{pmatrix}$	0.594 $\begin{pmatrix}0.00616\\0.02609\end{pmatrix}$	0.591 $\begin{pmatrix}0.00841\\0.02779\end{pmatrix}$	0.614 $\begin{pmatrix}0.01499\\0.03428\end{pmatrix}$
1000	0.451 $\begin{pmatrix}0.00625\\0.02017\end{pmatrix}$	0.450 $\begin{pmatrix}0.00766\\0.02185\end{pmatrix}$	0.461 $\begin{pmatrix}0.00641\\0.02364\end{pmatrix}$	0.599 $\begin{pmatrix}0.00245\\0.01552\end{pmatrix}$	0.598 $\begin{pmatrix}0.00413\\0.01794\end{pmatrix}$	0.609 $\begin{pmatrix}0.00583\\0.01937\end{pmatrix}$
5000	0.450 $\begin{pmatrix}0.00101\\0.00827\end{pmatrix}$	0.450 $\begin{pmatrix}0.00197\\0.00845\end{pmatrix}$	0.453 $\begin{pmatrix}0.00195\\0.00854\end{pmatrix}$	0.599 $\begin{pmatrix}0.00107\\0.00888\end{pmatrix}$	0.599 $\begin{pmatrix}0.00299\\0.00938\end{pmatrix}$	0.602 $\begin{pmatrix}0.00282\\0.00986\end{pmatrix}$
10,000	0.448 $\begin{pmatrix}0.00019\\0.00586\end{pmatrix}$	0.448 $\begin{pmatrix}0.00101\\0.00600\end{pmatrix}$	0.448 $\begin{pmatrix}0.00107\\0.00638\end{pmatrix}$	0.600 $\begin{pmatrix}0.00015\\0.00552\end{pmatrix}$	0.599 $\begin{pmatrix}0.00019\\0.00554\end{pmatrix}$	0.600 $\begin{pmatrix}0.00080\\0.00564\end{pmatrix}$

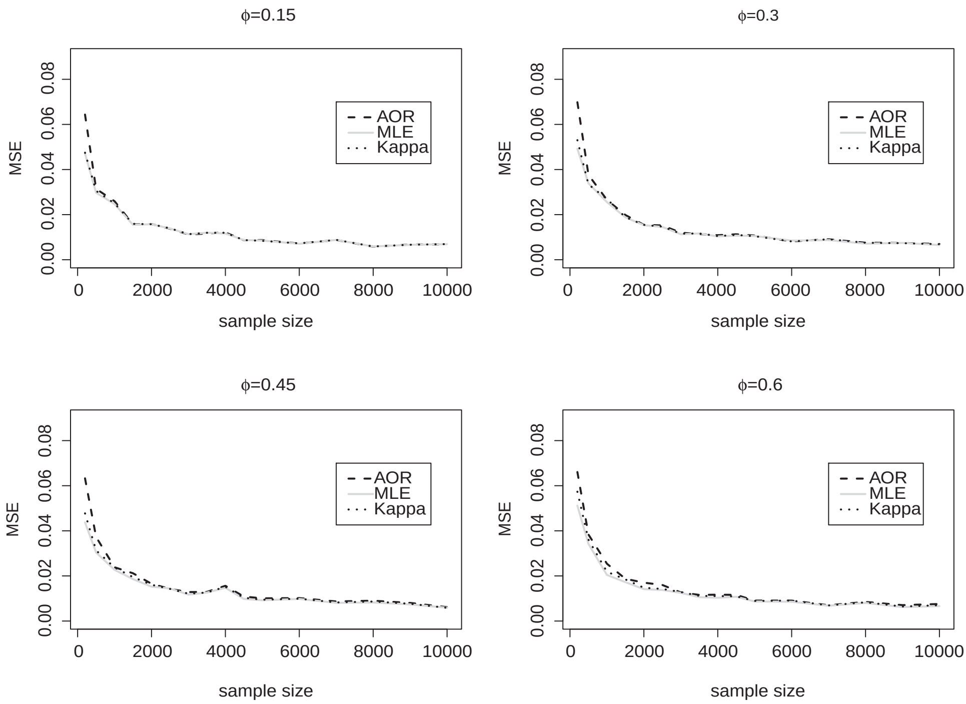

Figure 2. Comparison between three different methods of estimation using standard error of ϕ for $\phi = 0.15, 0.30, 0.45, 0.60$.

In our third simulation study, we plotted the 95% confidence interval of ϕ based on the AORF method. The whole range of ϕ, (0, 1), is partitioned into as many points as possible. Then for a fixed sample size and for each point of the partition, we obtained the estimated values of ϕ using the AORF method, and the 95% confidence interval under the approximate normality. Based on 1000 Monte Carlo replications, we averaged the estimated values and the confidence intervals and plotted them in Figure 3 along with the actual values of ϕ. As we can see from the figure, the solid line (–) represents the actual values and the dashed line (–) represents the estimated values of ϕ, whereas the dotted line (⋯) presents the upper and lower 95% confidence intervals. It can be seen from the above figure that for large sample sizes the actual and the estimated values overlap with the 95% confidence intervals. In other words, the length of the confidence intervals decreases as sample size increases, resulting in a consistent estimator of ϕ. Hence, the proposed measure can be an additional choice in the categorical time series analysis which can be used to study serial association as well as to estimate model parameters.

8. Data analysis

8.1. Infant sleep status data

Here, we consider the Infant sleep status data reported and analysed by Stoffer et al. [18]. The data consist of a collection of 24 categorical time series of infant sleep status in an EEG study. Each of these 24 time series is observed for 128 min. One such single time series is considered in this article. One important feature of the data is that it is categorical and categories are of ordinal nature.

During minute t, infant's sleep status was recorded in 6 categories, namely C_0 being 'quiet sleep' with trace alternate, C_1 being 'quiet sleep' with high voltage, C_2 being 'indeterminate sleep', C_3 being 'active sleep' with low voltage, C_4 being 'active sleep' with mixed voltage, and C_5 being 'awake'. We consider a particular series where the proportion of time spent by an infant in the sleep status

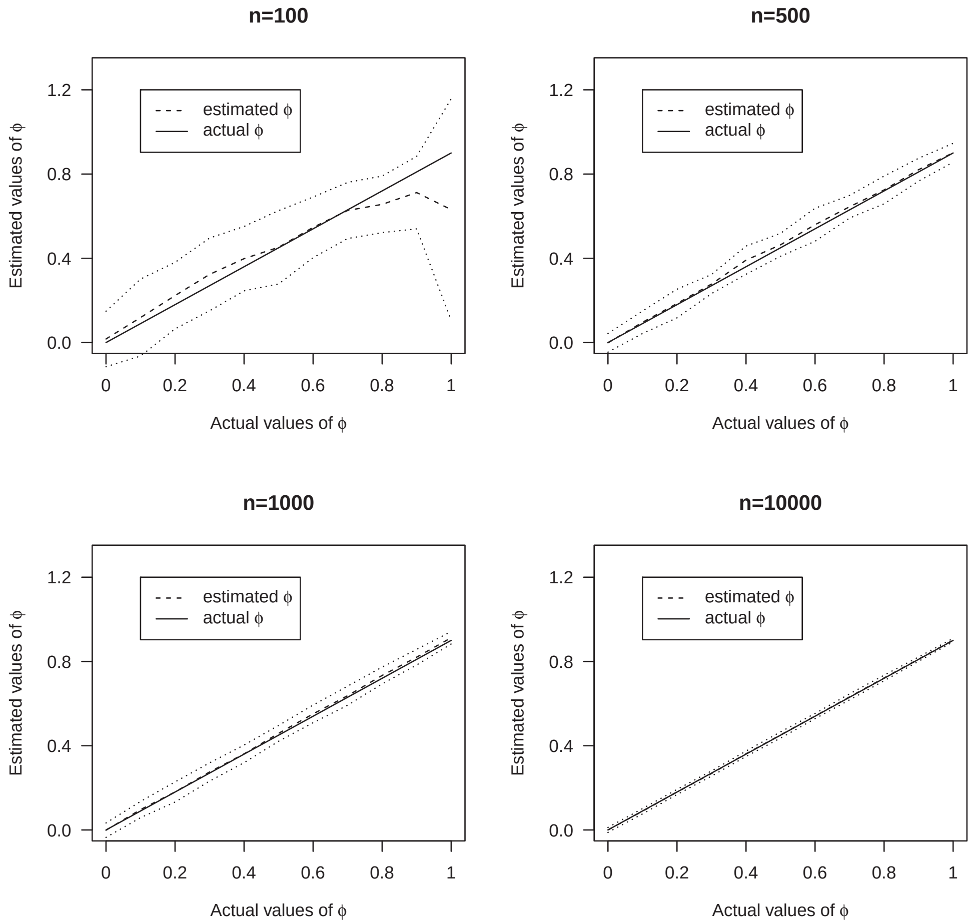

Figure 3. Estimates of AR(1) parameter ϕ through the AOR method with its approximate 95% CI.

$C_0, C_1, \ldots, C_5$ are 0.062, 0.352, 0.008, 0.109, 0.430 and 0.039, respectively. This indicates that infant spent maximum time in active sleep.

As we pointed out, the data are of ordinal in nature. The raw data and its $\log\theta(h)$ for various lagged values are presented in Figure 4. As we can see, the sample estimate of $\log\theta(h)$ decreases as the lagged value h increases. After fitting an AR(1) process to the data using AORF, the estimated value of the mixing parameter ϕ is computed as 0.270 with an approximate 95% confidence interval (0.162, 0.378) which does not include 0. It implies that the estimated value of ϕ is a positive significant number and therefore we can say that there is a strong serial association between the consecutive observations in the data and a AR(1) process seems to fit the data best. Also its estimated bias and standard error are reported as 0.012 and 0.055, respectively, comparing to the bias and standard error of Cohen's κ based estimate are 0.010 and 0.048, respectively. As we can see, Cohen's κ appears to be better than the AORF in terms of bias and standard error of $\hat{\phi}$, which agrees with our simulated result that for smaller sample sizes Cohen's κ performs better than the AORF $\theta(h)$.

8.2. Champions league data

In our second example, we consider the UEFA Champions League data as a non-ordered or nominal categorical to illustrate our method. Note that, the data give a list of champion teams in

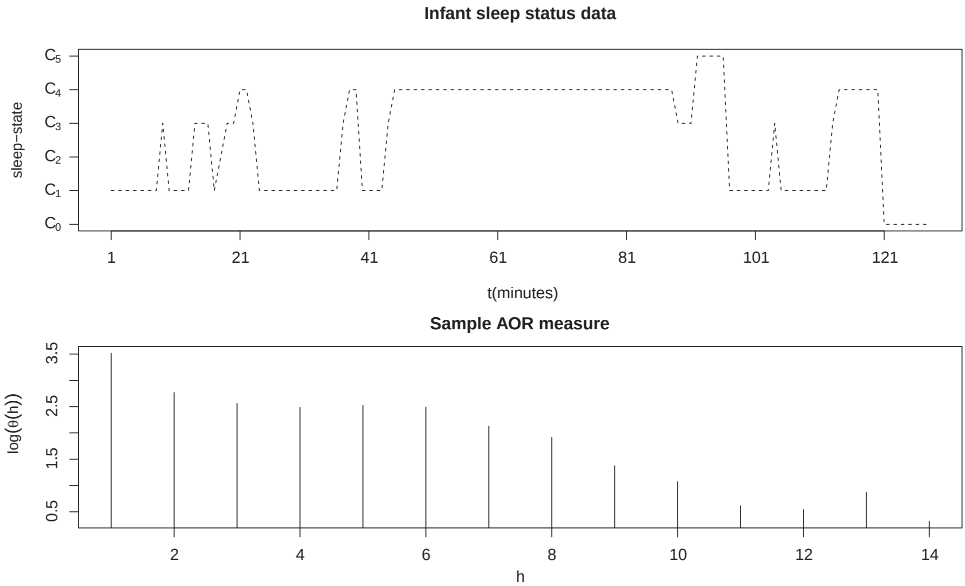

Figure 4. Infant sleep status data with its AORF.

UEFA champion league during the period from 1955 to 2012. As a general information, so far the competition has been won by 22 different clubs, 12 of which have won it more than once.

We categorize the champion clubs into five categories based on their success history, namely Real Madrid (RM), Barcelona (B), Bayern Munich (BM), Manchester United (MU) and Others (O). Therefore, the data give rise to a non-order or nominal categorical time series consisting of 58 observations having five classes starting from 1955 to 2012. The proportions of years won by the above five categories are 0.156, 0.052, 0.086, 0.069, 0.637, respectively. Since the data is a non-order in nature, one

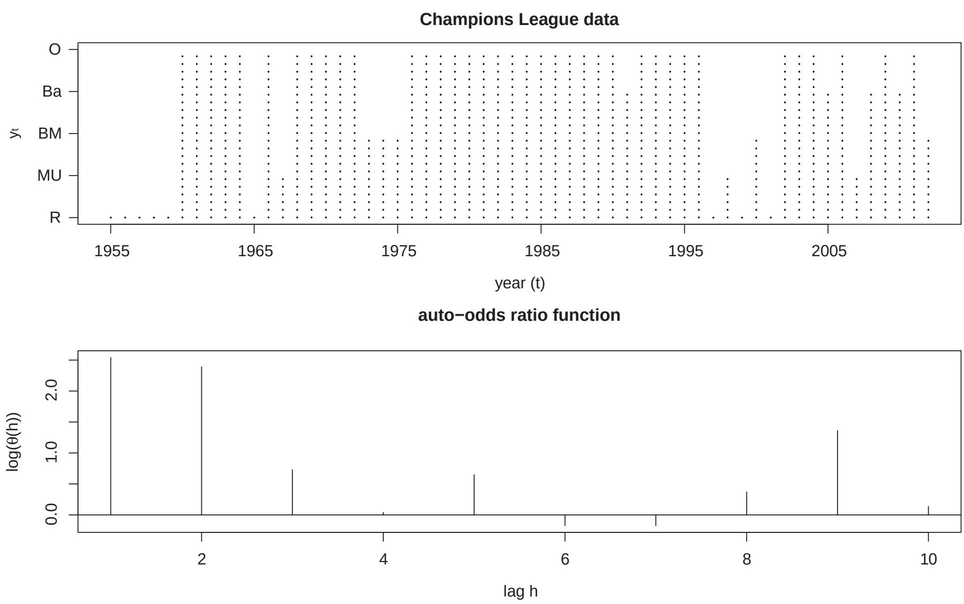

Figure 5. Champions league data with its AORF.

cannot obtain the ACF and partial ACF (PACF) for model selection. As an alternative, we computed the $\log\theta(h)$ for various lagged values h and presented in Figure 5 with the real data plot. As we can see, there is a decreasing pattern of $\log\theta(h)$ which suggests that an AR(1) process would possibly be a good fit to the data. After fitting an AR(1) process to the data, it is observed that the estimated value of the AR(1) mixing parameter ϕ using the AORF method is estimated as 0.29 with an approximate 95% confidence interval $(0.195, 0.485)$. Since the confidence interval of ϕ does not include 0, we can infer that there is a significant dependence in the data through Pegram's AR(1) process. Here also, we reported the bias and standard error of the AORF method and Cohen's κ method which are (0.033, 0.062) and (0.009, 0.041), respectively, with Cohen's κ-based estimate of ϕ is 0.175. As we can see Cohen's κ method has lesser bias and lesser standard error than that of the AORF $\theta(h)$. This is also in conformity with our simulation results which tells that for small sample sizes Cohen's κ method has lesser bias and lesser standard error compared to our method. However, the measure can still be an additional option to the users.

9. Conclusions

In this paper, we have considered the AORF to be a measure of serial dependence for categorical time series including both ordered and non-ordered categories. Earlier, Biswas and Song [10] studied this measure for AR(1) process with binary categories. Here, we have extended the measure to a more general set-up, i.e. it has been studied for stationary Pegram's AR(p) and MA(q) processes with more than two categories. Following Weiß [12], asymptotic normality of the measure has been established. Using some simulation experiments, consistency of the parameter estimates based on the proposed measure has been presented and compared its bias and standard error with the MLE and Cohen's κ methods. Results show that for large sample sizes, our proposed method of estimation appears to be equally efficient as the MLE method in terms of both bias and standard error. Thus, the extended measure would be an additional competitive choice in the field of the categorical time series analysis for studying serial dependency, model selection and parameter estimation.

Acknowledgments

The authors wish to thank the two anonymous reviewers and the associate editor for their careful reading and constructive suggestions which led to this improved version of the paper. Major part of the work was done when the first author was a PhD student in Indian Statistical Institute, Kolkata.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Agresti A. Categorical data analysis. Vol. 359. Hoboken (NJ): John Wiley & Sons; 2002.
[2] Weiß CH, Göb R. Measuring serial dependence in categorical time series. AStA Adv Stat Anal. 2008;92(1):71–89.
[3] Biswas A, Guha A. Time series analysis of categorical data using auto-mutual information. J Stat Plan Inference. 2009;139(9):3076–3087.
[4] Jacobs PA, Lewis PA. Discrete time series generated by mixtures. I: correlational and runs properties. J R Stat Soc Ser B (Methodol). 1978;40(1):94–105.
[5] Jacobs PA, Lewis PA. Discrete time series generated by mixtures II: asymptotic properties. J R Stat Soc Ser B (Methodol). 1978;40(2):222–228.
[6] Jacobs PA, Lewis PA, Discrete time series generated by mixtures III: autoregressive processes (DAR(p)). Monterey, CA: Naval Postgraduate School; 1978 (Technical report).
[7] Raftery AE. A model for high-order Markov chains. J R Stat Soc Ser B (Methodol). 1985;47(3):528–539.
[8] Berchtold A, Raftery AE. The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. Stat. Sci. 2002;17(3):328–356.

[9] Pegram GGS. An autoregressive model for multilag Markov chains. J Appl Probabil. 1980;17(2):350–362.
[10] Biswas A, Song PX-K. Discrete-valued ARMA processes. Stat Probabil Lett. 2009;79(17):1884–1889.
[11] Jacobs PA, Lewis PAW. Stationary discrete autoregressive-moving average time series generated by mixtures. J Time Ser Anal. 1983;4(1):19–36.
[12] Weiß CH. Empirical measures of signed serial dependence in categorical time series. J Stat Comput Simul. 2011;81(4):411–429.
[13] Weiß CH. Serial dependence of NDARMA process. Comput Stat Data Anal. 2013;68(1):213–238.
[14] Biswas A, del Carmen Pardo M., Guha A. Auto-association measures for stationary time series of categorical data. TEST. 2014;23(3):487–514.
[15] Fokianos K, Kedem B. Regression theory for categorical time series. Stat Sci. 2003;18(3):357–376.
[16] Maiti R, Biswas A, Coherent forecasting for stationary time series of discrete data. Adv Stat Anal. 2015; 99:337–365.doi:10.1007/s10182-014-0243-3
[17] Heagerty PJ, Zeger SL. Lorelogram: a regression approach to exploring dependence in longitudinal categorical responses. J Am Stat Assoc. 1998;93(441):150–162.
[18] Stoffer DS, Scher MS, Richardson GA, et al. A walsh-Fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. J Am Stat Assoc. 1988;83(404):954–963.

Appendices

Appendix 1. Proof of Theorem 3.1

Note that for a stationary Pegram's AR(1) process which is a particular case of Equation (6), the conditional distribution of Y_t given Y_{t-h} can be written as follows:

$$P(Y_t = C_i \mid Y_{t-h} = C_j) = \phi^h I(C_i = C_j) + (1-\phi^h)\pi_i; \quad i,j = 0,1,\ldots,k. \tag{A1}$$

Incorporating (A1) in Equation (2), we get the AORF between Y_t and Y_{t-h} for a pair of categories indexed by (C_i, C_j), $i \neq j$ as follows:

$$\theta_{i,j}(h) = 1 + \frac{\phi^{2h} + \phi^h(1-\phi^h)(\pi_i + \pi_j)}{(1-\phi^h)^2\pi_i\pi_j}, \quad i,j = 0,1,\ldots,k.$$

For intermediate steps, see Biswas and Song [10]. This describes the ratio of odds that $Y_t = C_i$ when $Y_{t-h} = C_i$ compared to when $Y_{t-h} = C_j$. Furthermore, the weight function $w_{ij}(h)$ for $i \neq j$ given in Equation (4) for the above AR(1) process can obtained as follows:

$$\begin{aligned} w_{ij}(h) &= \frac{P(Y_t = C_i, Y_{t-h} = C_j)}{\sum\sum_{0\le r\neq s\le k} P(Y_t = C_r, Y_{t-h} = C_s)} \\ &= \frac{\{\phi^h I(i=j) + (1-\phi^h)\pi_i\}\pi_j}{\sum\sum_{0\le r\neq s\le k}\{\phi^h I(r=s) + (1-\phi^h)\pi_r\}\pi_s} \\ &= \frac{\pi_i\pi_j}{1-\boldsymbol{\pi}^T\boldsymbol{\pi}}, \end{aligned} \tag{A2}$$

where $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_k)^{\mathrm{T}}$, unless define otherwise T refers to transpose of a matrix.

Plugging these weights in Equation (3), the AORF takes the form

$$\begin{aligned} \theta(h) &= 1 + \frac{k(k+1)\phi^{2h} + 2k\phi^h(1-\phi^h)}{(1-\phi^h)^2(1-\boldsymbol{\pi}^T\boldsymbol{\pi})} \\ &= g(\phi^h). \end{aligned}$$

Using some simple algebra, we get

$$\frac{\mathrm{d}}{\mathrm{d}\phi}\theta(1) = \frac{2k(1-\phi)^2(1-\boldsymbol{\pi}^T\boldsymbol{\pi})\{(k-1)\phi+1\} + 2k(1-\phi)(1-\boldsymbol{\pi}^T\boldsymbol{\pi})\{(k+1)\phi^2 + 2\phi(1-\phi)\}}{(1-\phi)^4(1-\boldsymbol{\pi}^T\boldsymbol{\pi})^2},$$

which is > 0 for $\phi \in (0,1)$ and $k \geq 1$. Therefore, for a stationary AR(1) process the AORF of order 1, $\theta(1)$ $(= g(\phi))$ is an increasing function of ϕ and hence it follows that the AORF of order h, $g(\phi^h)$, is also an increasing function of ϕ. This fact ensures the unique estimate of ϕ using the AORF method of estimation.

Appendix 2. Proof of Theorem 3.2

In order to derive the above result, we need to find the exact expressions of $\theta_{i,j}(h)$ and $w_{ij}(h)$ for the stationary AR(p) process given in Equation (6), which is equivalent to find the conditional distribution $P(Y_t = C_i \mid Y_{t-h} = C_j)$ for all $i \neq j$. Using the theorem of total probability, the required conditional probability can be written as follows:

$$
\begin{aligned}
p_{i\,|\,j}(h) &= P(Y_t = C_i \mid Y_{t-h} = C_j) \\
&= \eta_{h1} I(C_i = C_j) + \eta_{h2}\frac{\pi_i}{\pi_j} p_{j|i}(1) + \eta_{h3}\frac{\pi_i}{\pi_j} p_{j|i}(2) + \cdots + \eta_{hp}\frac{\pi_i}{\pi_j} p_{j|i}(p-1) \\
&\quad + (1 - \eta_{h1} - \eta_{h2} - \ldots - \eta_{hp})\pi_i, \quad h = 1, 2, \ldots.
\end{aligned} \tag{A3}
$$

After writing $\boldsymbol{p}_{i\,|\,j} = (p_{i|j}(1), \, p_{i|j}(2), \, \ldots, \, p_{i|j}(p-1))^{\mathrm{T}}$, we have for $i \neq j$,

$$
\begin{aligned}
\boldsymbol{p}_{i|j} &= \frac{\pi_i}{\pi_j} \boldsymbol{A} \boldsymbol{p}_{j|i} + \boldsymbol{b}\pi_i \\
&= \frac{\pi_i}{\pi_j} \boldsymbol{A} \left(\frac{\pi_j}{\pi_i} \boldsymbol{A} \boldsymbol{p}_{i|j} + \boldsymbol{b}\pi_j \right) + \boldsymbol{b}\pi_i \\
&= \boldsymbol{A}^2 \boldsymbol{p}_{i|j} + (\boldsymbol{I} + \boldsymbol{A})\boldsymbol{b}\pi_i,
\end{aligned}
$$

and hence

$$
\begin{aligned}
\boldsymbol{p}_{i|j} &= (\boldsymbol{I} - \boldsymbol{A}^2)^{-1}(\boldsymbol{I} + \boldsymbol{A})\boldsymbol{b}\pi_i \\
&= (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{b}\pi_i.
\end{aligned}
$$

For $i = j$,

$$\boldsymbol{p}_{i|i} = \boldsymbol{\delta}_{p-1} + \boldsymbol{A}\boldsymbol{p}_{i|i} + \boldsymbol{b}\pi_i$$

and hence

$$\boldsymbol{p}_{i|i} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\delta}_{p-1} + (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{b}\pi_i.$$

For $i \neq j$, the above conditional probability (A3) can explicitly be written as follows:

$$
\begin{aligned}
p_{i|j}(h) &= \eta_{h2}\frac{\pi_i}{\pi_j} p_{j|i}(1) + \eta_{h3}\frac{\pi_i}{\pi_j} p_{j|i}(2) + \cdots + \eta_{hp}\frac{\pi_i}{\pi_j} p_{j|i}(p-1) \\
&\quad + (1 - \eta_{h1} - \eta_{h2} - \cdots - \eta_{hp})\pi_i \\
&= \frac{\pi_i}{\pi_j} \boldsymbol{\eta}_h^{(2)'} \boldsymbol{p}_{j|i} + (1 - \mathbf{1}'\boldsymbol{\eta}_h)\pi_i \\
&= \frac{\pi_i}{\pi_j} \boldsymbol{\eta}_h^{(2)'} (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{b}\pi_j + (1 - \mathbf{1}'\boldsymbol{\eta}_h)\pi_i \\
&= \underbrace{\{\boldsymbol{\eta}_h^{(2)'} (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{b} + (1 - \mathbf{1}'\boldsymbol{\eta}_h)\}}_{f_1(\boldsymbol{\phi})}\pi_i \\
&= f_1(\boldsymbol{\phi})\pi_i.
\end{aligned}
$$

In similar way, for $i = j$, Equation (A3) can be written as follows:

$$
\begin{aligned}
p_{i|i}(h) &= \eta_{h1} + \boldsymbol{\eta}_h^{(2)'} \boldsymbol{p}_{i|i} + (1 - \mathbf{1}'\boldsymbol{\eta}_h)\pi_i \\
&= \eta_{h1} + \boldsymbol{\eta}_h^{(2)'} \{(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\delta}_{p-1} + (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{b}\pi_i\} + (1 - \mathbf{1}'\boldsymbol{\eta}_h)\pi_i \\
&= \underbrace{\{\eta_{h1} + \boldsymbol{\eta}_h^{(2)'} (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\delta}_{p-1}\}}_{f_2(\boldsymbol{\phi})} + \underbrace{\{\boldsymbol{\eta}_h^{(2)'} (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{b} + (1 - \mathbf{1}'\boldsymbol{\eta}_h)\}}_{f_1(\boldsymbol{\phi})}\pi_i \\
&= f_2(\boldsymbol{\phi}) + f_1(\boldsymbol{\phi})\pi_i.
\end{aligned}
$$

As defined in Equation (2), the AORF for $i \neq j$, is given by

$$\theta_{i,j}(h) = 1 + \frac{f_2^2(\boldsymbol{\phi}) + f_1(\boldsymbol{\phi})f_2(\boldsymbol{\phi})(\pi_i + \pi_j)}{f_1^2(\boldsymbol{\phi})\pi_i\pi_j}, \quad h = 1, 2, \ldots.$$

It is important to note that for a stationary AR(p) process, the normalizing weight $w_{ij}(h),\ i \neq j$ defined in Equation (4) is exactly the same with that of the AR(1) process derived in Equation (A2) and is given by

$$\begin{aligned} w_{ij}(h) &= \frac{P(Y_t = C_i, Y_{t-h} = C_j)}{\sum\sum_{0 \le r \neq s \le k} P(Y_t = C_r, Y_{t-h} = C_s)} \\ &= \frac{f_1(\boldsymbol{\phi})\,\pi_i\pi_j}{\sum\sum_{0 \le r \neq s \le k} f_1(\boldsymbol{\phi})\,\pi_r\pi_s} = \frac{\pi_i\pi_j}{1 - \boldsymbol{\pi}'\boldsymbol{\pi}}, \end{aligned}$$

and hence the AORF of lag h can be derived as follows:

$$\theta(h) = 1 + \frac{kf_2(\boldsymbol{\phi})}{f_1^2(\boldsymbol{\phi})(1 - \boldsymbol{\pi}'\boldsymbol{\pi})}\{(k+1)f_2(\boldsymbol{\phi}) + 2f_1(\boldsymbol{\phi})\}. \tag{A4}$$

Hence the result.

Appendix 3. Proof of Theorem 4.1

It is observed that for a stationary MA(q) process defined in Equation (11) and for $h > q$, $P(Y_t = C_i, Y_{t+h} = C_j) = \pi_i\pi_j$. Therefore by definition $\theta_{ij}(h) = 0$ for $h > q$, which implies that $\theta(h) = 0$ for $h > q$. Now for $0 \le h \le q$, it can be derived as

$$P(Y_t = C_i, Y_{t+h} = C_j) = \pi_i\pi_j + \left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)\pi_i\{I(i=j) - \pi_j\}, \tag{A5}$$

where $I(\cdot)$ is an indicator function (see [3] for the derivation). Hence using Equation (A5), we have

$$\begin{aligned} \theta_{ij}(h) &= \frac{\left\{\pi_i^2 + \left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)\pi_i(1-\pi_i)\right\}\left\{\pi_j^2 + \left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)\pi_j(1-\pi_j)\right\}}{\left\{\pi_i\pi_j - \left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)\pi_i\pi_j\right\}^2} \\ &= \frac{\pi_i\pi_j\left(1 - \sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)^2 + (\pi_i+\pi_j)\left(1 - \sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)\left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right) + \left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)^2}{\pi_i\pi_j\left(1 - \sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)^2} \\ &= 1 + \frac{\left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)^2 + \left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)\left(1 - \sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)(\pi_i+\pi_j)}{\left(1 - \sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)^2\pi_i\pi_j}. \end{aligned}$$

For a stationary MA(q) process, the suggested weight $w_{ij}(h)$ defined in Equation (4) is also equal with that for the AR(p) process and has the form $w_{ij}(h) = \pi_i\pi_j/(1-\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{\pi})$, which is free of h and hence the AORF of lag h, $\theta(h)$ is given by

$$\theta(h) = 1 + \frac{k(k+1)\left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)^2 + 2k\left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)\left(1-\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)}{\left(1-\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right)^2(1-\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{\pi})}$$
$$= f\left(\sum_{r=0}^{q-h}\theta_r\theta_{r+h}\right).$$

Hence the result.

Appendix 4.

Define $Y_{ti} = 1$ or 0 according as $Y_t = C_i$ or not. Hence, we can write

$$N_i = \sum_{t=1}^{n} Y_{ti} \quad \text{and} \quad N_{ij}(h) = \sum_{t=(h+1)}^{n} Y_{ti}Y_{(t-h)j} \quad \text{with } h \geq 1.$$

Note that

$$E(Y_{ti}) = \pi_i \quad \text{and} \quad E(Y_{ti}Y_{(t-h)j}) = p_{ij}(h).$$

Again

$$\mathrm{Var}(N_{ij}(h)) = \mathrm{Var}\left(\sum_{t=(h+1)}^{n} Y_{ti}Y_{(t-h)j}\right) = \sum_{t=(h+1)}^{n}\sum_{s=(h+1)}^{n}\mathrm{Cov}(Y_{ti}Y_{(t-h)j}, Y_{si}Y_{(s-h)j}). \tag{A6}$$

Now,

$$\mathrm{Cov}(Y_{ti}Y_{(t-h)j}, Y_{si}Y_{(s-h)j}) = E(Y_{ti}Y_{(t-h)j}Y_{si}Y_{(s-h)j}) - (\pi_i\pi_j)^2$$
$$= \begin{cases} \pi_i\pi_j(1-\pi_i\pi_j) & \text{if } s = t \\ \pi_i\pi_j\delta_{i,j}\pi_i - (\pi_i\pi_j)^2 & \text{if } s = t+h \\ \pi_i\pi_j\delta_{i,j}\pi_i - (\pi_i\pi_j)^2 & \text{if } s = t-h \\ 0 & \text{otherwise} \end{cases}$$

Combining the above cases, we can write

$$\mathrm{Cov}(Y_{ti}Y_{(t-h)j}, Y_{si}Y_{(s-h)j}) = \pi_i\pi_j(1-\pi_i\pi_j)\delta_{s,t} + \pi_i^2\pi_j(\delta_{i,j}-\pi_j)(\delta_{s,(t+h)} + \delta_{s,(t-h)}). \tag{A7}$$

After plugging (A7) in Equation (A6), we get

$$\mathrm{Var}(N_{ij}(h)) = \sum_{t=(h+1)}^{n}\sum_{s=(h+1)}^{n}[\pi_i\pi_j(1-\pi_i\pi_j)\delta_{s,t} + \pi_i^2\pi_j(\delta_{i,j}-\pi_j)(\delta_{s,(t+h)} + \delta_{s,(t-h)})]$$
$$= (n-h)\pi_i\pi_j(1-\pi_i\pi_j) + 2(n-2h)\pi_i^2\pi_j(\delta_{i,j}-\pi_j).$$

Hence,

$$\mathrm{Var}(\hat{p}_{ij}(h)) = \frac{1}{n-h}\pi_i\pi_j(1-\pi_i\pi_j) + \frac{2(n-2h)}{(n-h)^2}\pi_i^2\pi_j(\delta_{i,j}-\pi_j)$$
$$= \frac{1}{n-h}[\pi_i\pi_j(1-\pi_i\pi_j) + 2\pi_i^2\pi_j(\delta_{i,j}-\pi_j)] + O(n^{-2}),$$

which implies that as n goes to ∞, $\mathrm{Var}(\hat{p}_{ij}(h)) \to 0$. Therefore, $\hat{p}_{ij}(h)$ is a consistent estimator of $p_{ij}(h)$ for all $i,j = 0,1,\ldots,h$. Furthermore, to compute the asymptotic distribution of $\mathrm{vec}(\hat{P}_n(h))$, we find all combinations of covariance function which are given as follows:

(I)

$$\mathrm{Cov}(\hat{p}_{ii}(h), \hat{p}_{ij}(h)) = \frac{1}{(n-h)^2} \sum_{t=(h+1)}^{n} \sum_{s=(h+1)}^{n} Cov(Y_{ti}Y_{(t-h)i}, Y_{si}Y_{(s-h)j}).$$

$$\mathrm{Cov}(Y_{ti}Y_{(t-h)i}, Y_{si}Y_{(s-h)j}) = E(Y_{ti}Y_{(t-h)i}Y_{si}Y_{(s-h)j}) - \pi_i^3\pi_j$$

$$= \begin{cases} \pi_i\delta_{i,j}\pi_i - \pi_i^3\pi_j & \text{if } s = t, \\ \pi_i^2\delta_{i,j}\pi_i - \pi_i^3\pi_j & \text{if } s = t+h, \\ \pi_i^2\pi_j - \pi_i^3\pi_j & \text{if } s = t-h, \\ 0 & \text{otherwise.} \end{cases}$$

Combining the above equations, we have

$$\mathrm{Cov}(Y_{ti}Y_{(t-h)i}, Y_{si}Y_{(s-h)j}) = \delta_{s,t}\pi_i^2(\delta_{i,j} - \pi_i\pi_j) + \delta_{s,(t+h)}\pi_i^3(\delta_{i,j} - \pi_j) + \delta_{s,(t-h)}\pi_i^2\pi_j(1-\pi_i),$$

and hence

$$\begin{aligned} &\mathrm{Cov}(\hat{p}_{ii}(h), \hat{p}_{ij}(h)) \\ &= \frac{1}{(n-h)^2} \sum_{t=(h+1)}^{n} \sum_{s=(h+1)}^{n} [\delta_{s,t}\pi_i^2(\delta_{i,j} - \pi_i\pi_j) + \delta_{s,(t+h)}\pi_i^3(\delta_{i,j} - \pi_j) + \delta_{s,(t-h)}\pi_i^2\pi_j(1-\pi_i)] \\ &= \frac{1}{(n-h)^2}[(n-h)\pi_i^2(\delta_{i,j} - \pi_i\pi_j) + (n-2h)\pi_i^3(1-\pi_j) + (n-2h)\pi_i^2\pi_j(1-\pi_i)] \\ &= \frac{1}{n-h}[\pi_i^2(\delta_{i,j} - \pi_i\pi_j) + \pi_i^3(1-\pi_j) + \pi_i^2\pi_j(1-\pi_i)] + O(n^{-2}). \end{aligned}$$

(II)

$$\mathrm{Cov}(\hat{p}_{ii}(h), \hat{p}_{ji}(h)) = \frac{1}{(n-h)^2} \sum_{t=(h+1)}^{n} \sum_{s=(h+1)}^{n} Cov(Y_{ti}Y_{(t-h)i}, Y_{sj}Y_{(s-h)i}).$$

$$\mathrm{Cov}(Y_{ti}Y_{(t-h)i}, Y_{sj}Y_{(s-h)i}) = E(Y_{ti}Y_{(t-h)i}Y_{sj}Y_{(s-h)i}) - \pi_i^3\pi_j$$

$$= \begin{cases} \pi_i\delta_{ij}\pi_i - \pi_i^3\pi_j & \text{if } s = t, \\ \pi_i^2\pi_j - \pi_i^3\pi_j & \text{if } s = t+h, \\ \pi_i^2\delta_{i,j}\pi_i - \pi_i^3\pi_j & \text{if } s = t-h, \\ 0 & \text{otherwise.} \end{cases}$$

Combining the above equations, we get

$$\mathrm{Cov}(Y_{si}Y_{(s-h)i}, Y_{ti}Y_{(t-h)j}) = \delta_{s,t}\pi_i^2(\delta_{i,j} - \pi_i\pi_j) + \delta_{s,(t+h)}\pi_i^2\pi_j(1-\pi_i) + \delta_{s,(t-h)}\pi_i^3(\delta_{i,j} - \pi_j),$$

and hence

$$\begin{aligned} &\mathrm{Cov}(\hat{p}_{ii}(h), \hat{p}_{ij}(h)) \\ &= \frac{1}{(n-h)^2} \sum_{t=(h+1)}^{n} \sum_{s=(h+1)}^{n} [\delta_{s,t}\pi_i^2(\delta_{i,j} - \pi_i\pi_j) + \delta_{s,(t+h)}\pi_i^2\pi_j(1-\pi_i) + \delta_{s,(t-h)}\pi_i^3(\delta_{i,j} - \pi_j)] \\ &= \frac{1}{(n-h)^2}[(n-h)\pi_i^2(\delta_{i,j} - \pi_i\pi_j) + (n-2h)\pi_i^2\pi_j(1-\pi_i) + (n-2h)\pi_i^3(\delta_{i,j} - \pi_j)] \\ &= \frac{1}{n-h}[\pi_i^2(\delta_{i,j} - \pi_i\pi_j) + \pi_i^2\pi_j(1-\pi_i) + \pi_i^3(\delta_{i,j} - \pi_j)] + O(n^{-2}). \end{aligned}$$

(III)

$$\mathrm{Cov}(\hat{p}_{ji}(h), \hat{p}_{ij}(h)) = \frac{1}{(n-h)^2} \sum_{t=(h+1)}^{n} \sum_{s=(h+1)}^{n} \mathrm{Cov}(Y_{sj}Y_{(s-h)i}, Y_{ti}Y_{(t-h)j}),$$

where

$$\begin{aligned}\mathrm{Cov}(Y_{ti}Y_{(t-h)j}, Y_{sj}Y_{(s-h)i}) &= E(Y_{ti}Y_{(t-h)j}Y_{sj}Y_{(s-h)i}) - \pi_i^2\pi_j^2,\\ &= \begin{cases}\delta_{i,j}\pi_i^2 - \pi_i^2\pi_j^2 & \text{if } s = t,\\ \pi_j^2\pi_i - \pi_i^2\pi_j^2 & \text{if } s = t+h,\\ \pi_i^2\pi_j - \pi_i^2\pi_j^2 & \text{if } s = t-h,\\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

Combining the above cases, we can write

$$\mathrm{Cov}(Y_{ti}Y_{(t-h)j}, Y_{sj}Y_{(s-h)i}) = \delta_{s,t}\pi_i^2(\delta_{i,j} - \pi_j^2) + \delta_{s,t+h}\pi_i\pi_j^2(1-\pi_i) + \delta_{s,t-h}\pi_i^2\pi_j(1-\pi_i)$$

and hence

$$\begin{aligned}\mathrm{Cov}(\hat{p}_{ji}(h), \hat{p}_{ij}(h)) &= \frac{1}{(n-h)^2}[(n-h)\pi_i^2(\delta_{i,j} - \pi_j^2) + (n-2h)\pi_i\pi_j^2(1-\pi_i) + (n-2h)\pi_i^2\pi_j(1-\pi_i)]\\ &= \frac{1}{(n-h)}[\pi_i^2(\delta_{i,j} - \pi_j^2) + \pi_i\pi_j^2(1-\pi_i) + \pi_i^2\pi_j(1-\pi_i)] + O(n^{-2}).\end{aligned}$$

(IV) From [12], we have

$$\mathrm{Cov}(\hat{p}_{ii}(h), \hat{p}_{jj}(h)) = \frac{1}{n-h}\pi_i^2[\delta_{i,j}(2\pi_j + 1) - 3\pi_j^2] + O(n^{-2}).$$

(V) In general for any (i, j) and (i', j'), we can write

$$\begin{aligned}\mathrm{Cov}(Y_{ti}Y_{(t-h)j}, Y_{si'}Y_{(s-h)j'}) &= E(Y_{ti}Y_{(t-h)j}Y_{si'}Y_{(s-h)j'}) - \pi_i\pi_j\pi_{i'}\pi_{j'},\\ &= \begin{cases}\delta_{i,i'}\pi_i\delta_{j,j'} - \pi_i\pi_j\pi_{i'}\pi_{j'} & \text{if } s = t,\\ \pi_j\pi_{i'}\delta_{i,j'}\pi_i - \pi_i\pi_j\pi_{i'}\pi_{j'} & \text{if } s = t+h,\\ \pi_i\pi_{j'}\delta_{j,i'}\pi_j - \pi_i\pi_j\pi_{i'}\pi_{j'} & \text{if } s = t-h,\\ 0 & \text{otherwise,}\end{cases}\end{aligned}$$

and hence

$$\mathrm{Cov}(\hat{p}_{ij}(h), \hat{p}_{i'j'}(h)) = \frac{1}{n-h}[\pi_i\pi_j(\delta_{i,i'}\delta_{j,j'} - \pi_{i'}\pi_{j'}) + \pi_i\pi_{i'}\pi_j(\delta_{i,j'} - \pi_{j'}) + \pi_i\pi_j\pi_{j'}(\delta_{j,i'} - \pi_{i'})] + O(n^{-2}).$$

Therefore,

$$\mathrm{Var}(\mathrm{vec}(\hat{P}_n(h))) = \Sigma_n(h) = ((\mathrm{Cov}(\hat{p}_{ij}(h), \hat{p}_{i'j'}(h))))$$

and it goes to zero as n goes to ∞.