dortmund university

fakultät statistik

Model Diagnostics and Inference for Count Time Series

Dissertation

by

Maxime Faymonville

in partial fulfillment of
the requirements for the degree of
Doktor der Naturwissenschaften (Dr. rer. nat.)

submitted to the
Department of Statistics
TU Dortmund University

Dortmund, July 2025

Acting Dean:

Prof. Dr. Philipp Doebler

Referees:

Prof. Dr. Carsten Jentsch (TU Dortmund University)

Prof. Dr. Christian Weiß (Helmut Schmidt University Hamburg)

Acknowledgments

I would like to express my heartfelt gratitude to all those who have supported me throughout this journey.

First of all, I would like to thank my supervisor, Carsten, whose guidance, expertise and encouragement were invaluable in shaping this research. Your insightful feedback and unwavering support motivated me to push through every challenge. Thanks for always having an open ear. I would also like to thank Christian for his prompt responses to all my mails. Whenever I felt stuck or had even only a small question, you were always ready to help and inspired me with your motivation.

Many thanks also to Sheila, Jacob and Steffen who are not only my colleagues but also my friends with whom I have been on this journey since the beginning of our bachelor's degree. Without you and our mutual exchange of experiences, there were times when I might have buried my head in the sand. Your encouragement and camaraderie have been invaluable throughout this journey, helping me navigate challenges and celebrate successes together.

I would like to acknowledge the financial support received from the DFG which facilitated my research endeavors. I also thank all my co-authors and colleagues for their useful insights and our lunch discussions.

Lastly, I wish to express my deepest gratitude to my family for their unconditional love and encouragement. Although you had trouble in understanding what I actually do the whole day, your belief in me has been a constant source of motivation.

"Wit beyond measure is man's greatest treasure."

– Rowena Ravenclaw

Abstract

Count time series naturally arise when counting occurrences of things and events over time resulting in numerous applications in various fields. However, research on count time series and, more generally, discrete time series is not as advanced as research on classical continuous time series. This gap highlights the importance of developing appropriate methodologies for the discrete case in order to effectively address its unique characteristics.

This cumulative dissertation is based on five articles that collectively extend the research on count time series and discrete-valued time series in general. We provide an improved semi-parametric estimation procedure for the integer-valued autoregressive (INAR) model and develop an R package allowing for simulation, estimation and bootstrapping of INAR data. Furthermore, we present a semi-parametric INAR bootstrap procedure and prove its joint consistency for the estimation of the INAR coefficient and the innovation distribution. Finally, we propose a goodness-of-fit test on the whole INAR model class and provide methodology to conduct prediction in the setup of discrete time series in general. In addition to outlining these methodologies, we always validate our findings by extensive simulations and apply them on real-data examples. While three articles are published in peer-reviewed journals, the other two are on arXiv and attached in their current version.

Zusammenfassung

Zähldatenzeitreihen entstehen naturgemäß, wenn Ereignisse oder Objekte über die Zeit hinweg gezählt werden, was zu zahlreichen Anwendungen in unterschiedlichen Fachbereichen führt. Die Forschung zu Zähldatenzeitreihen und allgemeiner zu diskreten Zeitreihen ist jedoch bislang weniger weit fortgeschritten als die zu klassischen stetigen Zeitreihen. Dabei ist die Entwicklung geeigneter Methoden für den diskreten Fall von zentraler Bedeutung, um die zugehörigen Charakteristiken adäquat berücksichtigen zu können.

Diese kumulative Dissertation stützt sich auf fünf Artikel, die zusammen die Forschung über Zähldatenzeitreihen bzw. diskrete Zeitreihen im Allgemeinen erweitern. Wir schlagen ein verbessertes semiparametrisches Schätzverfahren für das INAR (Integer-valued Autoregressive) Modell vor und entwickeln ein R-Paket, das Simulation, Schätzung und Bootstrapping von INAR-Daten ermöglicht. Darüber hinaus präsentieren wir ein semiparametrisches INAR-Bootstrapverfahren und beweisen dessen Konsistenz für die gemeinsame Schätzung der INAR-Koeffizienten und der Innovationsverteilung. Weiter schlagen wir einen neuartigen Anpassungstest für die gesamte INAR-Modellklasse vor und entwickeln eine Methode für die Vorhersage diskreter Zeitreihen. Neben der Beschreibung dieser Methoden validieren wir unsere Ergebnisse durch umfangreiche Simulationen und wenden sie auf reale Datenbeispiele an. Drei der Artikel wurden bereits in Journalen mit Peer-Review veröffentlicht, die anderen beiden Artikel sind auf arXiv zu finden und wurden in ihrer aktuellen Version der vorliegenden Arbeit angehängt.

List of Publications

This cumulative dissertation is based on the following five papers and manuscripts which are all five not part of another dissertation.

Article 1: Faymonville, M., Jentsch, C., Weiß, C. H. and Aleksandrov, B. (2023). "Semiparametric estimation of INAR models using roughness penalization". *Statistical Methods & Applications* 32.2, pp. 365-400. DOI: 10.1007/s10260-022-00655-0.

Contribution:
The author of this thesis wrote the manuscript, implemented and evaluated the simulations and did the real-data applications. Christian Weiß shared his code for the (unpenalized) semiparametric estimation, came up with appropriate data sets and together with Carsten Jentsch, he provided his expertise in the research field while they both made suggestions for improvement of the contribution. The discussion and revision of the paper was done with all four (co-)authors.

Article 2: Faymonville, M., Riffo, J., Rieger, J. and Jentsch C. (2024). "spINAR: An R package for semiparametric and parametric estimation and bootstrapping of integer-valued autoregressive (INAR) models". *Journal of Open Source Software* 9.97, p. 5386. DOI: 10.21105/joss.05386.

Contribution:
The initial idea for the R package was given by Carsten Jentsch. The author of this thesis wrote the code of this package and provided her expertise in this research field. Jonas Rieger provided his expertise in publishing an R package and version control. Javiera Riffo implemented the functionality tests.

Article 3: Faymonville, M. and Jentsch, C. (2025). "Joint semi-parametric INAR bootstrap inference for model coefficients and innovation distribution". https://arxiv.org/abs/2507.11124.

Contribution:
The first idea came up during the Master's thesis of the author of the thesis which has been supervised by Carsten Jentsch. The author of this thesis implemented and evaluated the simulations, did the real-data applications and wrote a first draft of the manuscript. Together with Carsten Jentsch, both extensively worked on the underlying theory, constantly improving the manuscript.

Article 4: Faymonville, M., Jentsch, C., and Weiß, C. H. (2025b). "Semi-parametric goodness-of-fit testing for INAR models". *Bernoulli* 31.4, pp. 3213-3234. DOI: 10.3150/24-BEJ1844.

and the corresponding supplements

Faymonville, M., Jentsch, C., and Weiß, C. H. (2025c). "Supplement I to 'Semi-parametric goodness-of-fit testing for INAR models'". DOI: 10.3150/24-BEJ1844SUPPA.

containing three real-world data applications, additional tables and all the proofs and

Faymonville, M., Jentsch, C., and Weiß, C. H. (2025d). "Supplement II to 'Semi-parametric goodness-of-fit testing for INAR models'". DOI: 10.3150/24-BEJ1844SUPPB.

containing the MATLAB code for the real-world data applications.

Contribution:
The first idea of the test was developed together by Carsten Jentsch and Christian Weiß. The author of this thesis adapted and expanded it. She wrote the article and implemented and evaluated the simulations. Together with Carsten Jentsch, she did the theory, where Carsten Jentsch solely provided the power properties. Christian Weiß brought valuable expertise in this research field, identified relevant simulation scenarios and shared appropriate data sets for the real-data application.

Article 5: Faymonville, M., Jentsch, C. and Paparoditis, E. (2025a). "Predictive inference for discrete-valued time series". https://arxiv.org/abs/2507.16035.

Contribution:
The first idea came up during the Master's thesis of the author of the thesis which has been supervised by Carsten Jentsch. Together, they further developed the idea. After the author of this thesis wrote a first draft, implemented and evaluated some simulations and did some real-data applications, Efstathios Paparoditis joined and provided valuable input for expanding the scope of the work. The author of this thesis implemented them in the form of further simulations and rewriting the manuscript in terms of a different focus of the paper. All authors contributed equally to the derivation of the theoretical results.

Further publications:

- Aleksandrov, B., Weiß, C. H., Nik, S., Faymonville, M. and Jentsch, C. (2024). Modelling and diagnostic tests for Poisson and negative-binomial count time series. *Metrika*, 87, pp. 843-887. https://doi.org/10.1007/s00184-023-00934-0.
- Weiß, C. H., Aleksandrov, B., Faymonville, M. and Jentsch, C. (2023). Partial autocorrelation diagnostics for count time series. *Entropy*, 25.1, 105. https://doi.org/10.3390/e25010105.
- Aleksandrov, B., Weiß, C. H., Jentsch, C. and Faymonville, M. (2022). Novel goodness-of-fit tests for binomial count time series. *Statistics*, 56.5, pp. 957-990. https://doi.org/10.1080/02331888.2022.2134384.

Abbreviations

This list includes all abbreviations used in this thesis and some important, recurring notation.

$\rightsquigarrow$	weak convergence
$\mathbb{1}$	indicator function
ACF	autocorrelation function
AR	autoregressive
ARMA	autoregressive moving average
$\text{Bin}(n, p)$	binomial distribution with parameters $n \in \mathbb{N}_0$ and $p \in [0, 1]$
CLT	central limit theorem
CRAN	comprehensive R archive network
DAR	discrete autoregressive
DGP	data generating process
EDF	empirical distribution function
GARCH	generalized autoregressive conditional heteroscedasticity
$\text{Geo}(p)$	geometric distribution with parameter $p \in (0, 1]$ and support $\mathbb{N}_0$
ID	index of dispersion
INAR	integer-valued autoregressive
INARCH	integer-valued autoregressive conditional heteroscedasticity
INGARCH	integer-valued generalized autoregressive conditional heteroscedasticity
LASSO	least absolute shrinkage and selection operator
ML	maximum likelihood
MSE	mean squared error
$\mathbb{N}$	$\{1, 2, \ldots\}$
$\mathbb{N}_0$	$\{0, 1, 2, \ldots\}$
$\text{NB}(N, \pi)$	negative binomial distribution with parameters $N \in \mathbb{N}$ and $\pi \in [0, 1]$
NPMLE	non-parametric maximum likelihood estimator
PACF	partial autocorrelation function
pgf	probability generating function
pmf	probability mass function
$\text{Poi}(\lambda)$	Poisson distribution with parameter $\lambda > 0$
$\mathbb{R}$	set of all real numbers
spINAR	semi-parametric integer-valued autoregressive
$\mathbb{Z}$	set of all integers
$\text{ZIP}(\pi, \lambda)$	zero-inflated Poisson distribution with parameters $\pi \in (0, 1)$ and $\lambda > 0$

Contents

1

Introduction

Time series are a fundamental concept in statistics and mathematics. They consist of data points observed at different points or intervals over time and find application in many relevant fields such as environmental science, economics, health care and many others. Their analysis allows researchers from both theory and practice to discover underlying structures such as trends or seasonal components and to use them to understand past behavior or predict future values.

The literature on time series is broad; see, for example, Box and Jenkins (1970), Brockwell and Davis (1987) and Shumway and Stoffer (2000) as introductions. Most of these textbooks mainly cover the setup of continuous-valued time series, that is, time series with real numbers or vectors as outcome. Besides these continuous time series, there also exist so-called discrete time series, that is, time series with discrete range. Examples are numerous: There are count time series, where the range of the time series only allows for natural numbers. Another example are categorical time series, which include ordinal time series, where the data exhibits a qualitative range consisting of a finite number of categories. And as a last example, there are such apparently simple setups as binary time series, where the time series takes two possible values only. All these exemplary settings have in common that the classical methods for continuous time series are not able to account for the discrete range of these time series. Despite its noted relevance, McKenzie (2003) states in his survey article that modeling discrete variate time series is the most challenging and, at this time, the least well-developed one of all research areas in time series. Some textbooks briefly address the discrete case in their work, for example, Fahrmeir and Tutz (2001), Kedem and Fokianos (2002) and Cameron and Trivedi (2013) in the regression setup or Zucchini and MacDonald (2009) in their book about hidden Markov models for time series. In Turkman et al. (2014), one chapter deals with integer-valued time series and the book of Davis et al. (2011) provides some essays about discrete-valued time series. Finally, Weiß (2018) introduces the field of discrete-valued time series in general while covering common models for count time series and categorical time series and outlining their most important properties.

This cumulative dissertation mainly covers the topic of count time series, which arise naturally when counting things or events. As an example, consider Figure 1 containing the total number of climate-related disasters in France over the years and separately displaying the total numbers for each of the six considered disasters in this data set - drought, extreme temperature, flood, landslide, storm

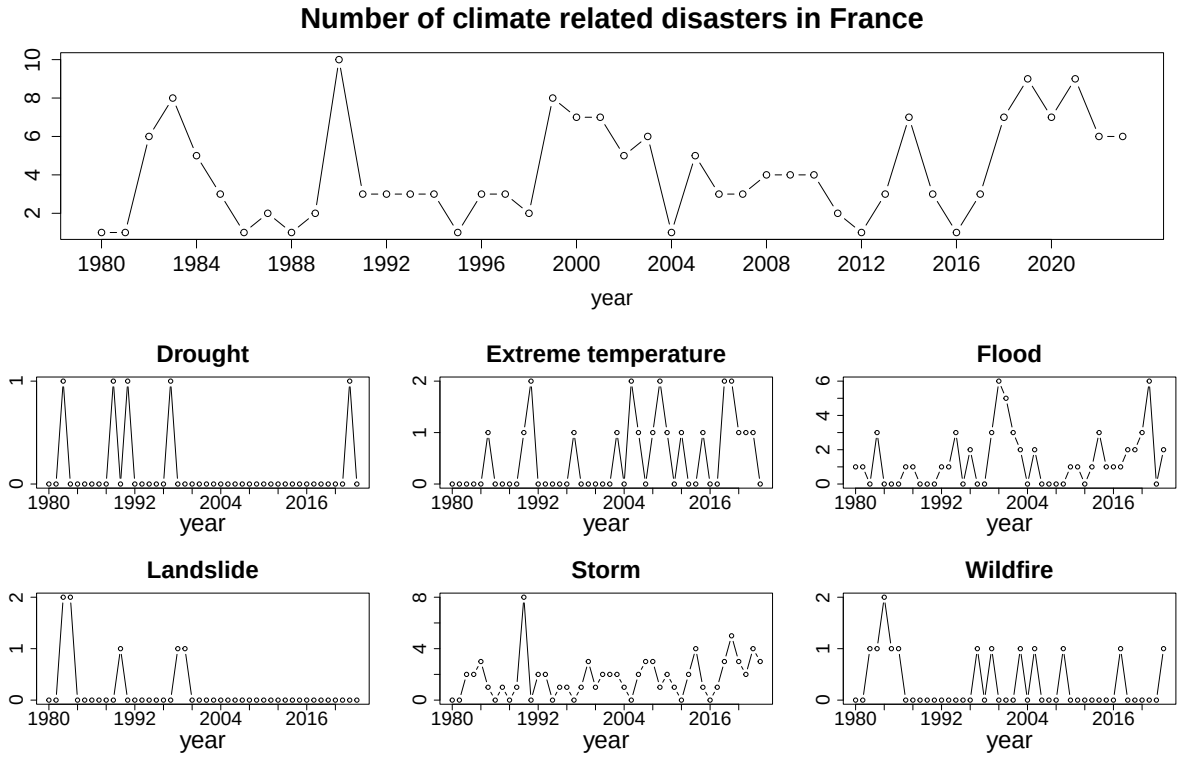

Figure 1: Number of total climate related disasters in France over the years (above) and (below) the separate numbers for each of the six considered disasters in the data set. Source: The Emergency Events Database (EM-DAT), Center for Research on the Epidemiology of Disasters (CRED)/Université catholique de Louvain (UCLouvain), Brussels, Belgium. `www.emdat.be`.

and wildfire. When it comes to the analysis of these time series, it becomes obvious that continuous models are not able to ensure the integer-valued range of the response. While they are sometimes used nevertheless, this procedure will only provide limited approximation, especially in the case of low counts as appearing in Figure 1. Additionally, they cannot account for specific characteristics often arising when dealing with count time series, such as asymmetry or excess zeros, also visible in Figure 1. This underlines the need for using appropriate count time series models.

The papers of this thesis contribute to the current state of research on count time series and partly also on discrete-valued time series in general. They address different challenges from various perspectives. The first paper (Faymonville et al., 2023) reveals that the performance of the semi-parametric estimator introduced in Section 2.2 may be inferior in case of small sample sizes. To this end, we propose a penalized version of the aforementioned estimator exploiting the smoothness of most innovation distributions of the INAR model introduced in Subsection 2.1.1. In this context, "smoothness of a distribution" means that two consecutive entries of its probability mass function differ only slightly from each other. Simulations display that our approach improves the estimation performance and that a combination of penalized and unpenalized approaches results in overall best INAR model fits.

The penalized and unpenalized semi-parametric estimation mentioned in the previous paragraph is implemented in the R package `spINAR` (Faymonville et al., 2024). This package also allows

for parametric estimation of the INAR model introduced in Subsection 2.1.1, (semi-)parametric bootstrapping (see Section 2.3) and flexible simulation of INAR data. In our peer-reviewed paper, we briefly describe the features of our package and set it into context with other related packages. It has been published via CRAN and includes tests for desired functionality. Besides the help pages in R, we provide extensive documentation on GitHub.

When introducing the semi-parametric INAR model estimator (see Section 2.2), Drost et al. (2009) not only prove efficiency and consistency of their estimator, they also derive its limiting distribution (see (2.8)), allowing for asymptotic inference. However, due to the infinite-dimensional parameter space, this asymptotic result is cumbersome to apply in practice, motivating the use of bootstrap inference. To this end, in Faymonville and Jentsch (2025), we propose an appropriate bootstrap procedure and prove that the bootstrap version of the semi-parametric estimator provides the same asymptotic distribution in probability. In addition to simulations that support this result, we also provide several methodological examples in which the result finds application, underlining its practical relevance.

One application of the result of Faymonville and Jentsch (2025) is the semi-parametric goodness-of-fit test for INAR models presented in Faymonville et al. (2025b). In this paper, we propose a test for the whole INAR model class where one does not have to specify a parametric family of innovation distributions. We prove consistency under fixed alternatives and analyze its asymptotic behavior under local alternatives. Our L_2-type test statistic relies on the joint probability generating function and its limiting distribution which we derive is not practicable. The result of Faymonville and Jentsch (2025) enables us to use bootstrap inference instead. In simulations, we illustrate the performance of our goodness-of-fit test and display that it can be improved by using test statistics of higher order.

In the fifth paper (Faymonville et al., 2025a), we address the general problem of predictive inference in the case of discrete time series. Due to the discrete range of the values, prediction *intervals* are not meaningful and prediction *sets* do generally not retain a desired coverage level. We therefore propose to reverse the construction principle and, in a nutshell, transform the prediction problem into a parameter estimation problem, where the parameter of interest is the probability for observing some values of interest in the future which arise from the respective application. We cover both parametric and non-parametric setups while providing asymptotic and bootstrap approaches. The latter have the advantage to circumvent possibly cumbersome limiting distributions and to imitate the distributions of interest also when it is not clear whether the model used for prediction holds true. While we first introduce the general procedure, we then illustrate our proposed methods by an application to both first-order INAR and INARCH models (see Section 2.1) using (conditional) maximum likelihood estimation. In simulations, we investigate the finite sample performance and we additionally discuss that extensions to higher model order and larger prediction horizon are straightforward.

To underline their practical relevance, we apply all the methods introduced in Faymonville et al. (2023), Faymonville and Jentsch (2025), Faymonville et al. (2025b) and Faymonville et al. (2025a) on real-world data sets and discuss our findings.

The remainder of this thesis is structured as follows. In Chapter 2, we provide the probably most popular count time series models in existing literature along with corresponding estimation methods and appropriate bootstrap techniques. Chapter 3 contains a summary of each of the five papers of this cumulative dissertation, emphasizing the innovative elements of each publication and their contribution to the respective research field. The thesis concludes with a discussion in Chapter 4, followed by all full-length papers.

2

Statistical Methods

This chapter provides a general methodological background of already existing methods in the context of count time series. They have been used and extended in the articles of this cumulative dissertation. Due to individual requests from reviewers of the journals we published in, the notation may sometimes slightly differ across the papers. In this chapter, we aim for a notation that is as consistent as possible.

2.1 Count Time Series Models

Count time series consist of non-negative integers observed over time, that is, the observed values have range $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. They arise naturally when counting things or events from all areas of life. Weiß (2018) analyzes, among others, the number of road accidents in an area of the Netherlands and the number of infections with, for example, Hantavirus and Legionella. Gouveia et al. (2018) investigate the number of rainy days of locations across Europe and Russia and Weiß and Kim (2014) deal with the number of countries in the EA17[1] with stable prices. Lately, many examples arose during the COVID-19 pandemic, where for each day the Robert-Koch institute (Robert-Koch institute, 2020) displayed different key numbers such as the number of infective people or hospital admissions to monitor the pandemic. Another example with increasing relevance is the number of extreme weather events partly already addressed in Figure 1.

2.1.1 INAR Model

When it comes to the modeling of such data sets, we have to take into account the integer nature of the data and cannot use well-known time series models as the AR(p) (**A**uto**R**egressive) model introduced by Yule (1927) and Slutsky (1937) as

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \varepsilon_t, t \in \mathbb{Z}. \tag{2.1}$$

[1] A group of 17 countries in the Euro area

Due to the *multiplication* of the coefficients α_i and the lags $X_{t-i}, i = 1, \ldots, p$, X_t would not be modeled as integer even if the innovations ε_t would only take integer values. Hence, one possibility to obtain an appropriate model is to replace the multiplication in (2.1) by a range-preserving operation. One approach of reducing $X_{t-i}, i = 1, \ldots, p$, in a way that preserves its integer values is the so-called binomial thinning operator denoted with "$\circ$". Steutel and Van Harn (1979) introduce it as

$$\alpha_i \circ X_{t-i} = \sum_{j=1}^{X_{t-i}} Z_j^{(t,i)}, \tag{2.2}$$

where $\left(Z_j^{(t,i)},\, j \in \mathbb{N},\, t \in \mathbb{Z}\right),\, i \in 1, \ldots, p$, are mutually independent Bernoulli distributed random variables $Z_j^{(t,i)} \sim \text{Bin}(1, \alpha_i)$. Thus, $\alpha_i \circ X_{t-i}$ can only take integer values between 0 and X_{t-i}. Using (2.2), Du and Li (1991) introduce the INAR(p) (**IN**teger-valued **A**uto**R**egressive) model as

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + \varepsilon_t, t \in \mathbb{Z} \tag{2.3}$$

where $\varepsilon_t \overset{\text{i.i.d.}}{\sim} G$ with G being a discrete distribution with range $\mathbb{N}_0$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ the vector of model coefficients fulfilling $\sum_{i=1}^{p} \alpha_i < 1$. Additionally, we have that $Z_j^{(t,i)}$ is independent of $(\varepsilon_t, t \in \mathbb{Z})$, the thinning operations are independent over time and of $(\varepsilon_t, t \in \mathbb{Z})$ and both the thinning operation at time t and ε_t are independent of X_s, $s < t$. The INAR(p) model of Du and Li (1991) is marked by these extensive assumptions, which contrast with those assumptions put forth by Alzaid and Al-Osh (1990), who introduce another version of the INAR(p) model while also using the recursion in (2.3). Since only the INAR(p) model proposed by Du and Li (1991) yields the traditional Yule-Walker equations (see Yule (1927) and Walker (1931)) for the autocorrelation function (ACF), it is generally favored in practice. Therefore, we will concentrate on this model specification for the rest of this chapter, just like in all of our papers on which this thesis is based. In most applications, the first-order INAR model ($p = 1$) finds application for which the two models coincide and simplify to the INAR(1) model first introduced by McKenzie (1985) and Al-Osh and Alzaid (1987).

The INAR(p) model is a p^{th} order Markov chain whose transition probabilities are given by

$$P^{\boldsymbol{\alpha},G}(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p}) = \left(\text{Bin}(x_{t-1}, \alpha_1) * \ldots * \text{Bin}(x_{t-p}, \alpha_p) * G\right)\{x_t\}, \tag{2.4}$$

where "$*$" denotes the convolution of distributions. In the special case of $p = 1$, they simplify to

$$P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1}) = \sum_{j=0}^{\min(x_t, x_{t-1})} \binom{x_{t-1}}{j} \alpha^j (1-\alpha)^{x_{t-1}-j} G(x_t - j),$$

where $(G(k),\, k \in \mathbb{N}_0)$ denotes the probability mass function of G. For the innovation distribution G, Al-Osh and Alzaid (1987) initially proposed the Poisson distribution, which can be regarded as the natural analog of the normal distribution in the case of continuous time series. This distribution is characterized by equidispersion, naturally leading to limitations. Therefore, several alternative innovation distributions for the INAR model have been proposed. For example, Savani and Zhigljavsky (2007) consider negative binomial innovations, Jazi et al. (2012b) geometric innovations, Jazi et al. (2012a) zero-inflated Poisson innovations and Qi et al. (2019) zero-and-one inflated Poisson innovations.

2.1.2 IN(G)ARCH Model

Besides the idea of the INAR model replacing the multiplication by an appropriate operation ensuring the integer range, another idea is to consider a regression model for the conditional mean $M_t := E(X_t|X_{t-1})$. This so-called INGARCH(p, q) (**IN**teger-valued **G**eneralized **A**uto**R**egressive **C**onditional **H**eteroscedasticity) model has been introduced by Rydberg and Shephard (2000) and Heinen (2003) (under different names) as

$$M_t = \beta_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j M_{t-j}, \tag{2.5}$$

with $X_t|X_{t-1}, \ldots \sim \text{Poi}(M_t)$ and $\beta_0 > 0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q \geq 0$. It has been further analyzed by Ferland et al. (2006) and Fokianos et al. (2009). For different choices of distributions, one obtains different INGARCH models. Assuming a Poisson distribution as conditional distribution leads to the basic model we always refer to when speaking of "the INGARCH model". Despite the equidispersion property of the Poisson distribution, the INGARCH model is suitable to model overdispersed observations; see Weiß (2018).

Setting $q = 0$ in (2.5), we naturally obtain the so-called INARCH(p) (**IN**teger-valued **A**uto**R**egressive **C**onditional **H**eteroscedasticity) model. Its transition probabilities are given by

$$P^{\boldsymbol{\alpha},\beta_0}(X_t = x_t|x_{t-1}, \ldots) = \exp\left(-\beta_0 - \sum_{i=1}^{p} \alpha_i x_{t-i}\right) \frac{(\beta_0 + \sum_{i=1}^{p} \alpha_i x_{t-i})^{x_t}}{x_t!}.$$

When additionally setting $p = 1$, we obtain the popular INARCH(1) model, particularly addressed by Weiß (2010).

Researchers who prefer the INARCH over the INAR model argue that a drawback of the latter is the commitment to an innovation distribution G coming with restrictions. But in a remarkable paper, Drost et al. (2009) come up with a semi-parametric INAR model which we address in the following Section 2.2, making the INAR model class very flexible. Additionally, the INAR model exhibits a nice and easy to interpret autoregressive structure. That is why in the remainder of this section and mostly in our papers, we focus on the INAR model.

2.2 Estimation Methods

The INAR model is completely specified by the model coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ and the innovation distribution G. When it comes to estimation of the model, we therefore need estimators for $\boldsymbol{\alpha}$ and G, where the latter might be tricky, especially in the case of unbounded counts. The literature mostly deals with parametric estimation assuming that G lies in some parametric class of distributions, that is,

$$G \in \{G_\gamma \mid \gamma \in \Gamma \subset \mathbb{R}^q\} \tag{2.6}$$

for some finite $q \in \mathbb{N}$. Since the INAR(p) model in (2.3) can be seen as the integer-valued counterpart of the AR(p) model in (2.1), it allows the use of classical (parametric) estimation approaches as e.g. Yule-Walker estimation, least squares estimation and maximum likelihood estimation; see for example Du and Li (1991), Silva and Silva (2006) and Bu et al. (2008) or Section 2.2 of Weiß (2018) for a summary of the most popular ones. While (2.6) might be too restrictive, for example, in terms of dispersion or possible zero-inflation, the semi-parametric model estimator of Drost et al. (2009) keeps the parametric binomial thinning operation, but the estimation of G is completely non-parametric. This leads to a very flexible estimation approach and for this reason, we mostly focus on this estimation procedure in our papers.

The semi-parametric estimation approach of Drost et al. (2009) treats the error distribution as (possibly infinite-dimensional) parameter. It jointly estimates the model coefficients and the probability mass function of the innovation distribution. To do so, they only need some mild assumptions. First of all, they require that

$$G \in \mathcal{G} = \{G \in \widetilde{\mathcal{G}} : 0 < G(0) < 1 : \mathrm{E}_G \varepsilon_t^{p+4} < \infty\},$$

where $\widetilde{\mathcal{G}}$ denotes the set of all probability measures on $\mathbb{N}_0$. This condition ensures that an observation can but does not always have to be equal to zero and requests the existence of some moments to ensure weak convergence of some empirical processes; see Drost et al. (2009) for details. Additionally, they impose that $\boldsymbol{\alpha} \in \Theta = \{\boldsymbol{\alpha} \in (0,1)^p : \sum_{i=1}^p \alpha_i < 1\}$ which ensures the stationarity of the time series and has already been included in our definition of the INAR model itself; see (2.3). Finally, their non-parametric maximum likelihood estimator is defined to maximize the conditional likelihood, that is,

$$\forall n \in \mathbb{N} : (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) \in \underset{(\boldsymbol{\alpha},G)\in[0,1]^p\times\widetilde{\mathcal{G}}}{\operatorname{argmax}} \left(\prod_{t=0}^{n} P^{\boldsymbol{\alpha},G}_{(X_{t-1},\ldots,X_{t-p}),X_t} \right), \tag{2.7}$$

where $P^{\boldsymbol{\alpha},G}_{(X_{t-1},\ldots,X_{t-p}),X_t} = P^{\boldsymbol{\alpha},G}(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p})$ corresponds to the transition probabilities of (2.4). Drost et al. (2009) prove the efficiency and the consistency of this estimator along with its limiting distribution given by

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1} \mathcal{S}^{\theta_0}, \tag{2.8}$$

where $\widehat{\theta}_n = (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$, $\theta_0 = (\boldsymbol{\alpha}_0, G_0)$, "$\rightsquigarrow$" denotes weak convergence, $\mathcal{S}^{\theta_0}$ is a tight, Borel measurable, Gaussian process and $\dot{\Psi}_{\theta_0}^{-1}$ is the continuous inverse of the Fréchet derivative of Ψ^{θ_0}. The latter denotes the population counterpart of the estimating equation(s) of $\widehat{\theta}_n$ which arise from its Z-estimator representation; see Drost et al. (2009) for details. The limiting distribution in (2.8) is a transformed Gaussian process. Because of the infinite-dimensional parameter space, this result is cumbersome to use for asymptotic inference in applications, motivating the use of bootstrap inference instead.

2.3 Bootstrap Methods

Although the INAR model in (2.3) can be seen as an integer-valued counterpart of the AR model in (2.1), they differ in one crucial point: In contrast to the linear AR model, the INAR model is non-linear, which is caused by the use of the (random) binomial thinning operation (2.2) instead of the (deterministic) multiplication. Particularly, Drost et al. (2009) highlight that having the estimated coefficients available does not directly provide the residuals. The latter is the case for the AR model, allowing to use the very popular AR bootstrap, which has been shown to be consistent under mild assumptions for a large class of statistics; see e.g. Kreiß (1992), Bühlmann (1997), Kreiß (1997) and Kreiß et al. (2011). Since the AR bootstrap is not only popular but also easy to implement, it would be desirable to transfer it to the INAR case. To do so, one would need appropriate residuals while correctly replicating the randomness of the binomial thinning operation. After exploring some naive approaches, Jentsch and Weiß (2019) conclude that this is not possible to a sufficient extent. Instead, they come up with another bootstrap scheme and distinguish between a parametric and a semi-parametric setting. In the following, we provide the algorithm for the semi-parametric case. By replacing the semi-parametric estimation by a suitable parametric estimation method, one gets the corresponding parametric approach.

Algorithm 2.3.1 (Semi-parametric INAR bootstrap of Jentsch and Weiß (2019))

1) Using the estimator of Drost et al. (2009) displayed in (2.7)*, semi-parametrically fit an INAR(p) process* $X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \varepsilon_t$ *to get estimated INAR coefficients* $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)$ *and the estimated probability mass function* $\widehat{G} = (\widehat{G}(k),\, k \in \mathbb{N}_0)$ *of* G*.*

2) Generate bootstrap observations $X_1^*, \ldots, X_n^*$ *according to*

$$X_t^* = \widehat{\alpha}_1 \circ^* X_{t-1}^* + \ldots + \widehat{\alpha}_p \circ^* X_{t-p}^* + \varepsilon_t^*,$$

where "$\circ^*$*" denotes (mutually independent) bootstrap binomial thinning operations and* ε_t^* *are i.i.d. random variables following* $\widehat{G}$*.*

Under a suitable set of meta assumptions, Jentsch and Weiß (2019) prove bootstrap consistency for the presented bootstrap scheme. Among others, the statistic of interest has to belong to the class of functions of generalized means, containing, for example, the sample mean, versions of sample autocovariances, sample (partial) autocorrelations and the Yule-Walker estimators. We refer to Jentsch and Weiß (2019) for more details.

3

Summary of the Articles

3.1 Article 1: Semiparametric Estimation of INAR Models using Roughness Penalization

The first article (Faymonville et al., 2023) provides an improved semi-parametric estimation method for the INAR model based on the one of Drost et al. (2009) introduced in Section 2.2. It uncovers that the estimation performance of the latter may be inferior in case of short time series, leading to estimated innovation distributions which are non-smooth. More concretely, they provide unnatural gaps in their estimated probability mass function (pmf) of the innovation distribution, although in practice most of the commonly used innovation distributions, such as the Poisson or the negative binomial distribution, are smooth, that is, consecutive entries of the pmf are close to each other. In this article, we leverage this prior knowledge and take advantage of the inherent qualitative (non-parametric) smoothness assumption by using a penalized approach. We highlight that we still do not assume a parametric class of innovation distributions.

The concept of penalization itself is not new in the context of count data and count time series; see e.g. Bui et al. (2022), Nardi and Rinaldo (2011), Fokianos (2010), Wang (2020) and Wang et al. (2020). While they mostly use the concept of penalization to shrink the model coefficients towards zero in order to perform variable selection, we come up with penalized estimation of the innovation distribution towards smoothness. To this end, we conduct a roughness penalization as introduced by Scott et al. (1980) and add a penalization term to the (log-)likelihood of Drost et al. (2009) (compare (2.7)), that is, we maximize the penalized log-likelihood

$$\log(\mathcal{L}_{\text{pen}}(\boldsymbol{\alpha}, G)) = \log(\mathcal{L}(\boldsymbol{\alpha}, G)) - \eta \cdot d_{G,m}.$$

The penalty term $\eta \cdot d_{G,m}$ is based on the idea of Tibshirani et al. (2005), who penalize differences of successive parameters. Hereby, η is the so-called smoothing or penalization parameter and $d_{G,m}$ denotes a suitable measure to quantify the roughness of G with m denoting the order of difference.

We consider two different roughness measures. The first one is inspired by Tibshirani et al. (2005) and is based on the L_1 distance (LASSO penalization). It is defined as

$$d_{G,m,1} = \sum_{i=m}^{\max(x_1,\ldots,x_n)} |\Delta^m G(i)|\,,$$

where $\Delta^m G(i) = \Delta^{m-1}(\Delta G(i))$ and $\Delta G(i) = G(i) - G(i-1)$. The second one uses the L_2 distance (Ridge penalization), that is,

$$d_{G,m,2} = \sum_{i=m}^{\max(x_1,\ldots,x_n)} (\Delta^m G(i))^2\,, \tag{3.1}$$

and is, in contrast to the first measure, characterized by not shrinking the difference of successive entries of the pmf to zero, but close to zero, which aligns better with the concept of a smooth distribution.

An important question that arises is how to select the penalization parameter η for a fixed roughness measure. To this end, we propose two different algorithms. The first one is an adaptation of the cross-validation procedure described in Adam et al. (2019). The second one is computationally more intensive and avoids potentially non-optimal selection due to an inappropriate choice of the initial value.

In simulations, we consider INAR models with different (non-)smooth innovation distributions and model order. We cover both first- and second-order differences. Overall, we see that the penalized estimation reduces both the variance and the bias of the estimated innovation distribution in case of smooth innovation distributions. Additionally, we find that L_2 penalization together with first-order differences ($m = 1$ in (3.1)) works best. One drawback of the penalized estimation, which we observe in the simulations, is that the joint estimation of the model coefficients and the innovation distribution is compromised. However, this issue can be solved by combining the unpenalized estimation of the model coefficients with the penalized estimation of the innovation distribution. Finally, we apply our procedure to intermittent demand time series and illustrate that the penalization approach is also beneficial for forecasting.

3.2 Article 2: spINAR: An R Package for Semiparametric and Parametric Estimation and Bootstrapping of Integer-valued Autoregressive (INAR) Models

The INAR model class is only marginally addressed in R (R Core Team, 2022) and other programming languages. In R, to the best of our knowledge, there are the packages `tscount` (Liboschik et al., 2017) and `ZINARp` (Medina Garay et al., 2022). The first one offers likelihood-based estimation for some count time series models, but not including the INAR model. Additionally, it is limited to

conditional Poisson or negative binomially distributed data. The second one allows for simulation and estimation of INAR data but is limited to parametric estimation methods (compare (2.6)) and only allows for Poisson or zero-inflated Poisson innovations. In Julia (Bezanson et al., 2017), the package `CountTimeSeries` (Stapper, 2022) addresses integer counterparts of ARMA and GARCH models, including the INAR model as a special case. While it supports parametric estimation methods, it also does not allow for non-parametric estimation of the innovation distribution. Moreover, none of these packages provides bootstrapping procedures for INAR models.

Our R package `spINAR` (Faymonville et al., 2024) closes this gap and allows for simulation, (semi-) parametric estimation and (semi-)parametric bootstrapping of INAR models. It covers models of first and second order, that is, $p \in \{1, 2\}$, which are mainly relevant in applications. It has been published via CRAN and, besides the provided help packages in R, we provide extensive documentation on GitHub.

3.3 Article 3: Joint Semiparametric INAR Bootstrap Inference for Model Coefficients and Innovation Distribution

In the literature, INAR models and in particular their innovation distributions are mostly estimated using parametric methods; see, for example, Franke and Seligmann (1993), Freeland and McCabe (2005), Brännäs and Hellström (2001) and Jung et al. (2005). That is, they assume that G belongs to a parametric family of distributions as displayed in (2.6). Following the discussion in Faymonville et al. (2025b), this makes the INAR model quite inflexible. Using the estimator of Drost et al. (2009), on the other hand, leads to much fewer restrictions and allows us to estimate the innovation distribution in a non-parametric way, as described in Section 2.2. Besides proving its consistency and efficiency, Drost et al. (2009) derive the limiting distribution of their INAR model estimator; see (2.8). Despite this apparently convenient closed form expression, the limiting distribution is cumbersome to use in applications, naturally motivating the use of bootstrap inference instead.

To generate the bootstrap data, we use the semi-parametric INAR bootstrap of Jentsch and Weiß (2019) for which the authors already established consistency. But this result only covers statistics of interest, which can be represented as functions of generalized means, which do in particular not include statistics depending on the estimated innovation distribution. In Faymonville and Jentsch (2025), we prove a more general result. First, we derive a Z-estimator representation of the bootstrap estimator $\widehat{\theta}_n^* = (\widehat{\boldsymbol{\alpha}}_n^*, \widehat{G}_n^*)$ which is the semi-parametric estimator of Drost et al. (2009) applied on the bootstrap data $X_1^*, \ldots, X_n^*$. Finally, under some technical and challenging to prove assumptions, we get that

$$\sqrt{n}\left(\widehat{\theta}_n^* - \widehat{\theta}_n\right) \rightsquigarrow^* -\dot{\Psi}_{\theta_0}^{-1}\mathcal{S}^{\theta_0} \text{ in probability,} \tag{3.2}$$

that is, the bootstrap estimator $\widehat{\theta}_n^*$ provides in probability the same limiting distribution as $\widehat{\theta}_n$; see (2.8). Furthermore, we state in a resulting corollary that by employing appropriate (bootstrap) delta

methods of Theorem 3.1 of Beutner and Zähle (2016) and Theorem 3.9.4 of Van der Vaart and Wellner (1996), we can extend the bootstrap consistency result to smooth functionals of the bootstrap estimator. That is, in probability, we have

$$d\left(\mathcal{L}\left(\sqrt{n}\left(\Xi(\widehat{\theta}_n)-\Xi\left(\theta_0\right)\right)\right),\mathcal{L}^*\left(\sqrt{n}\left(\Xi(\widehat{\theta}^*_n)-\Xi(\widehat{\theta}_n)\right)\right)\right)\underset{n\to\infty}{\to}0 \tag{3.3}$$

for $\Xi:[0,1]^p\times\tilde{\mathcal{G}}\to\mathbb{R}^q$ a sufficiently smooth functional such that the conditions in Theorem 3.1 of Beutner and Zähle (2016) are fulfilled and d an appropriate metric on (the distributions of) random elements in $\mathbb{R}^q$.

To validate the result in (3.2), we set up simulations in which we construct confidence intervals for the entries of the true model parameter $\theta_0=(\boldsymbol{\alpha}_0,G_0)$. We consider different INAR DGPs with different levels of dispersion and observations' mean. To illustrate the practical relevance of (3.2) and (3.3), we provide three different methodological applications. First of all, these results establish the theoretical framework of the semi-parametric goodness-of-fit test presented by Faymonville et al. (2025b). Second, we consider the prediction issue discussed in Faymonville et al. (2025a) and provide a semi-parametric version of their proposed procedure. Lastly, we present another application by constructing confidence intervals for both the dispersion index of the observations and the innovations. We especially highlight the novelty of the latter since only a non-parametric estimation of the innovation distribution allows us to address the innovations directly without imposing a parametric assumption (often) fixing the dispersion from the outset. Finally, we apply the semi-parametric prediction procedure on a real-data set from economics while also providing confidence intervals for the dispersion indices of the observations and the innovations.

3.4 Article 4: Semi-parametric Goodness-of-fit Testing for INAR Models

In this article, we propose a goodness-of-fit test on the INAR model class. We discuss the flexibility of the INAR model introduced in Subsection 2.1.1 and how an assumption of the form (2.6) will inevitably lead to limitations. That is why we should not test for restrictive parametric null hypotheses of the form

$$H_0^{\text{para}}:(X_t,\,t\in\mathbb{Z})\,\text{is INAR}(p)\,\text{with}\,G=G_\lambda\,\text{for some}\,\lambda\in\Lambda \tag{3.4}$$

for some parametric family of innovation distribution $\{G_\lambda,\lambda\in\Lambda\}$ with $\Lambda\subset\mathbb{R}^d$ and some (finite) d. Instead, we test for the null that the data follow an INAR(p) model with unspecified innovation distribution, that is,

$$H_0^{\text{semi}}:(X_t,\,t\in\mathbb{Z})\,\text{is INAR}(p). \tag{3.5}$$

Null hypotheses of the form (3.4) are usually considered in the literature since they notably simplify the estimation process and enable relatively straightforward testing strategies as can be seen in Meintanis and Karlis (2014), Hudecova et al. (2015), Schweer (2016), Aleksandrov et al. (2022), Aleksandrov et al. (2024) and a bivariate extension in Hudecova et al. (2021). However, this also

makes them particularly vulnerable to possible model misspecification, which is why null hypotheses of the form (3.5) are preferable.

To build a test statistic, we adopt the principal idea of Meintanis and Karlis (2014) and construct an L_2-type test statistic based on two estimators of the (joint) probability generating function (pgf). One estimator shall be consistent in general, the other one only under the null in (3.5). In contrast to Meintanis and Karlis (2014), where the pgf estimation under their null (3.4) simplifies a lot as a result of their parametric assumption, we have to deal with a general representation of the pgf, which we derive in our paper. By exploiting the INAR dependence structure of $X_t, \ldots, X_{t-p}$, we show that the joint pgf for INAR(p) models can be represented as

$$g_p(u_0, \ldots, u_p) = g_\varepsilon(u_0) \cdot E\left(\prod_{j=1}^{p} \left\{ u_j(1 + \alpha_j(u_0 - 1)) \right\}^{X_{t-j}} \right), \tag{3.6}$$

where $g_\varepsilon(u_0) = \sum_{k=0}^{\infty} P(\varepsilon_t = k)\, u_0^k = \sum_{k=0}^{\infty} G(k)\, u_0^k$. Using the estimator (2.7), we obtain a plug-in estimator of (3.6) under the null (3.5) given by

$$\widehat{g}_{p;H_0}(\mathbf{u}) := \widehat{g}_\varepsilon(u_0) \cdot \frac{1}{n-p} \sum_{t=p+1}^{n} \prod_{j=1}^{p} \left\{ u_j(1 + \widehat{\alpha}_{n,j}(u_0 - 1)) \right\}^{X_{t-j}}, \tag{3.7}$$

where $\mathbf{u} = (u_0, u_1, \ldots, u_p)$ and

$$\widehat{g}_\varepsilon(u_0) := \sum_{k=0}^{\infty} \widehat{P}(\varepsilon_t = k)\, u_0^k = \sum_{k=0}^{\infty} \widehat{G}_n(k)\, u_0^k = \sum_{k=0}^{\max(X_1,\ldots,X_n)} \widehat{G}_n(k)\, u_0^k.$$

A (non-parametric) estimator of the pgf of order p which is both consistent under the null (3.5) and the alternative is given by

$$\widehat{g}_p(\mathbf{u}) := \frac{1}{n-p} \sum_{t=p+1}^{n} \prod_{j=0}^{p} u_j^{X_{t-j}}. \tag{3.8}$$

By using the two estimators in (3.7) and (3.8), we naturally obtain the test statistic

$$T_n = n \int_0^1 \cdots \int_0^1 \left(\widehat{g}_{p;H_0}(\mathbf{u}) - \widehat{g}_p(\mathbf{u}) \right)^2 w(\mathbf{u}; a)\, d\mathbf{u}, \tag{3.9}$$

where $d\mathbf{u} := du_0 \cdots du_p$ and $w(\mathbf{u}; a) := (a+1)^{p+1} \prod_{j=0}^{p} u_j^a$ is a weighting function with weighting parameter $a \geq 0$. Intuitively, large values of (3.9) suggest a violation of the null in (3.5). In the article, we additionally derive a higher-order test statistic and a representation without numerical integration.

By showing that it can be represented as a degenerate V-statistic and applying asymptotic results of Leucht and Neumann (2013), we derive the cumbersome limiting distribution of the test statistic (3.9). Furthermore, we prove consistency of our test procedure under fixed alternatives and examine its asymptotic behavior under suitable local alternatives. As stated by Gürtler and Henze (2000), Meintanis and Swanepoel (2007), Leucht (2012), Leucht and Neumann (2013) and Meintanis and Karlis (2014) and proven in our paper for our specific test statistic (3.9), L_2-type test statistics as

in (3.9) do not have a conventional limiting distribution. To enable the application of the test, we propose an appropriate bootstrap procedure where Algorithm 2.3.1 finds application.

In simulations, we investigate the performance of our proposed goodness-of-fit test and compare the rejection rates with those from Meintanis and Karlis (2014). We elaborate on how we can increase the power by using higher-order test statistics. As illustration, we apply our test on three real-world data examples and provide all the MATLAB (The MathWorks Inc., 2022) code (Faymonville et al., 2025d). All the proofs of the paper and the real-data applications can be found in (Faymonville et al., 2025c), which is also attached to this dissertation.

3.5 Article 5: Predictive Inference for Discrete-valued Time Series

This article not only deals with count time series but also covers the topic of discrete-valued time series in general and is concerned with their prediction. When dealing with prediction in the non-discrete setup, one usually comes up with a point prediction and a corresponding prediction interval to quantify the uncertainty. A basic quality criterion for the latter is its asymptotic validity. Given a time series $X_1, \ldots, X_n$, according to Pan and Politis (2016), the interval $[l_n, u_n]$ is asymptotically valid for X_{n+1} if

$$P(l_n \leq X_{n+1} \leq u_n | X_1, \ldots, X_n) \to 1 - \beta \text{ for } n \to \infty, \tag{3.10}$$

where $1-\beta$, $\beta \in (0,1)$, denotes the true coverage level. However, such intervals cannot account for the discrete nature of a discrete-valued time series. To be in line with the notion of *coherent forecasting* introduced by Freeland and McCabe (2004), which expresses that predicted values of counts should also be counts themselves, we could instead consider prediction sets. In the article, we illustrate that they are, in general, not even asymptotically able to achieve validity in terms of (3.10). Instead, we propose to construct a confidence interval for the probability that X_{n+1} lies in an application-motivated set S given $X_1, \ldots, X_n$. Here, S can be any user-selected subset of possible values of the time series' range. We denote the aforementioned probability with $P_{S,x_n} := P(X_{n+1} \in S | X_n = x_n)$. While first focusing on the prediction of X_{n+1} and models of autoregressive order one, we later expand our approach to higher model order and larger prediction horizon.

We first cover asymptotic approaches and deal with the common situation in time series analysis where a parametric model is used to perform the prediction. Assuming some model class with parameter θ and a CLT for a corresponding estimator $\widehat{\theta}$, we obtain an asymptotically valid confidence interval for the parametric predictive probability of interest, where we denote the latter with $P_{S,x_n}^{(para)}(\theta_0)$. We emphasize that we nowhere assume that the observed time series really stems from the parametric model class. This is important as we address the practically relevant problem of model uncertainty/model misspecification. In this case, the parameter θ_0 does not denote the true parameter value but the best fit to the data under the assumed model class.

While in the parametric case we easily obtain a point estimator for the predictive probability by plugging in the estimated parameter value, in the non-parametric case, we make use of the relative frequencies and obtain

$$\widehat{P}_{S,x_n}^{(npara)} = \frac{\sum_{t=1}^{n-1} \mathbf{1}_{\{X_{t+1}\in S, X_t = x_n\}}}{\sum_{t=1}^{n-1} \mathbf{1}_{\{X_t = x_n\}}}$$

if $\sum_{t=1}^{n-1} \mathbf{1}_{\{X_t=x_n\}} \neq 0$ and 0 otherwise. By proving an appropriate CLT and using the delta method, we are able to derive a confidence interval for P_{S,x_n} in this non-parametric case.

Depending on the parametric assumption we rely on when asymptotically constructing the confidence interval, the estimation of the asymptotic variance can be complicated and bootstrapping might be beneficial. Furthermore, in order to correctly replicate the limiting distribution of the estimator, we have to take care which scenario we face: Either we use parametric or non-parametric prediction and either the data arise from the assumed model class or not. In the article, we provide a basic bootstrap algorithm, where the estimation and the generation of the bootstrap data must be specified according to which of the four cases we encounter. We establish bootstrap validity of the bootstrap procedures and stress that the bootstrap confidence interval for the parametric predictive probability $P_{S,x_n}^{(\text{para})}(\theta_0)$ retains (asymptotically) the desired level even if the data does not arise from the assumed model class.

As illustration, we apply our proposed prediction methodology on the INAR(1) and the INARCH(1) model (see Section 2.1) and focus on (conditional) maximum likelihood estimation of these models; see Section 2.2 of Weiß (2018) for details. Under some regularity conditions, we prove validity of our proposed procedures for these concrete settings. In addition to the basic bootstrap algorithm, we also consider exceptional cases, where the confidence intervals exceed the natural range of the probability and propose a small modification leading to range-preserving intervals.

In simulations, we cover both asymptotic and bootstrap-based approaches while considering parametric and non-parametric setups. For the asymptotic confidence intervals, we see there is a trade-off between robustness and efficiency. While the non-parametric approach is robust to model misspecification, the parametric approach is more efficient when the time series arises from the assumed model class. For the bootstrap confidence intervals, we observe that in the practically important case of model misspecification, the obtained confidence intervals inherit this uncertainty about the true model class and provide a slightly larger interval length. Lastly, we apply our methodology to two different data sets from economy, where we provide the possibility to quantify the uncertainty of the point prediction with a confidence interval. The sets of interest, S, are chosen to answer relevant questions of the respective field. We see that in case of a parametric approach, even the doubts about the validity of the model can be reflected in the confidence intervals.

4

Discussion and Outlook

Despite their widespread application, discrete time series are much less studied than their continuous counterparts. This discrepancy is particularly noteworthy in the context of statistical inference, where it is essential to consider the range of a time series. In this cumulative dissertation, we primarily focused on count time series, where the observed values have range $\mathbb{N}_0$ and further expanded the research in that area.

In a first paper (Faymonville et al., 2023), we proposed an improvement of the flexible semi-parametric estimation approach of the INAR model by proposing a penalized version. It exploited a qualitative smoothness assumption fulfilled by most common innovation distributions and added a roughness penalization to the existing likelihood. We then implemented these estimation methods together with several bootstrap procedures and the possibility of flexible simulation of INAR data in our own R package `spINAR` (Faymonville et al., 2024). It is available as open-source software on CRAN and extensively documented on GitHub and in several R help pages. Additionally, our peer-reviewed paper integrates our contribution into already existing software packages in different programming languages and describes its features. Our third paper (Faymonville and Jentsch, 2025) dealt with the cumbersome limiting distribution of the semi-parametric INAR model estimator. Although a closed-form expression exists, it is cumbersome to use in practice. We proposed to use an appropriate bootstrap procedure instead and proved its validity. This result allows for joint inference of the INAR model coefficients and the innovation distribution and we provided several methodological applications. One of these applications is the novel semi-parametric goodness-of-fit test we developed for the whole INAR model class (Faymonville et al., 2025b). Contrary to already existing literature, it does not rely on parametric assumptions about the nature of the innovations. The test statistic is of L_2-type and based on weighted integrals using probability generating functions, leading to a complex limiting distribution. To circumvent it, we specified a bootstrap procedure justified by the main result of Faymonville and Jentsch (2025). We proved consistency of our test under fixed alternatives and discussed its asymptotic behavior under local alternatives. In our last paper (Faymonville et al., 2025a), we dealt with the prediction issue of discrete-valued time series. Although prediction in general is a well-developed topic, the methods are usually limited to continuous time series. Instead of constructing prediction intervals or sets given a desired coverage, we proposed to construct confidence

intervals for the probability to observe certain values of interest in the future, where the latter arise from application. We covered both asymptotic and bootstrap approaches while considering parametric and non-parametric settings and proved their respective validity.

Naturally, there are limitations inherent in any research endeavor. While a semi-parametric approach offers considerably greater flexibility compared to its parametric counterparts, it is also more complex. Particularly with an increasing range of counts, the number of parameters that need to be estimated grows, which inevitably makes the estimation process more cumbersome. However, it is important to emphasize that the entire theory for discrete time series is especially relevant when dealing with low counts. In such cases, the dimensionality of the problem decreases. Furthermore, computational power plays a crucial role in this context. Setting up simulation studies requires substantial computational resources. But in application, analyses on a few samples can be done with little effort.

While our proposed prediction procedure is already applicable for discrete time series in general, future research could investigate whether the estimation procedures also find application on integer-valued models on $\mathbb{Z}$ such as those proposed by Kim and Park (2008) or Liu et al. (2021). Another promising research question is the joint modeling of real- and integer-valued autoregressive time series. While the literature on continuous autoregressive models is vast and an increasing amount of research is conducted for their integer counterparts, it is lacking a unified approach to model them jointly. Interesting questions include, for example, how to choose operations that respect the range of all included time series and whether classical estimation methods still find application. Also, the question about a suitable bootstrap procedure arises, as well as how to conduct impulse response analysis or inference in general.

References

Adam, T., Langrock, R., and Weiß, C. H. (2019). "Penalized estimation of flexible hidden Markov models for time series of counts". *METRON* 77.2, pp. 87–104. DOI: 10.1007/s40300-019-00153-6.

Al-Osh, M. A. and Alzaid, A. A. (1987). "First-order integer-valued autoregressive (INAR(1)) process". *Journal of Time Series Analysis* 8.3, pp. 261–275. DOI: 10.1111/j.1467-9892.1987.tb00438.x.

Aleksandrov, B., Weiß, C. H., and Jentsch, C. (2022). "Goodness-of-fit tests for Poisson count time series based on the Stein-Chen identity". *Statistica Neerlandica* 76.1, pp. 35–64. DOI: 10.1111/stan.12252.

Aleksandrov, B., Weiß, C. H., Nik, S., Faymonville, M., and Jentsch, C. (2024). "Modelling and diagnostic tests for Poisson and negative-binomial count time series". *Metrika* 87, pp. 843–887. DOI: 10.1007/s00184-023-00934-0.

Alzaid, A. A. and Al-Osh, M. A. (1990). "An integer-valued p*th*-order autoregressive structure (INAR(p)) process". *Journal of Applied Probability* 27.2, pp. 314–324. DOI: 10.2307/3214650.

Beutner, E. and Zähle, H. (2016). "Functional delta-method for the bootstrap of quasi-Hadamard differentiable functionals". *Electronic Journal of Statistics* 10.1, pp. 1181–1222. DOI: 10.1214/16-EJS1140.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. (2017). "Julia: A fresh approach to numerical computing". *SIAM Review* 59.1, pp. 65–98. DOI: 10.1137/141000671.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. 1st ed. Holden-Day. ISBN: 978-0816210947.

Brännäs, K. and Hellström, J. (2001). "Generalized integer-valued autoregression". *Econometric Reviews* 20.4, pp. 425–443. DOI: 10.1081/ETC-100106998.

Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods*. 1st ed. Springer. ISBN: 978-0387964065.

Bu, R., McCabe, B., and Hadri, K. (2008). "Maximum likelihood estimation of higher-order integer-valued autoregressive processes". *Journal of Time Series Analysis* 29.6, pp. 973–994. DOI: 10.1111/j.1467-9892.2008.00590.x.

Bühlmann, P. (1997). "Sieve bootstrap for time series". *Bernoulli* 3.2, pp. 123–148. DOI: 10.2307/3318584.

Bui, M. T., Potgieter, C. J., and Kamata, A. (2022). "Penalized likelihood methods for modeling count data". *Journal of Applied Statistics* 50.15, pp. 3157–3176. DOI: 10.1080/02664763.2022.2103101.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*. 2nd ed. Cambridge University Press. DOI: 10.1017/CBO9781139013567.

Davis, R. A., Holan, S. H., Lund, R., and Ravishanker, N. (2011). *Handbook of Discrete-Valued Time Series*. CRC Press. ISBN: 978-0367570392.

Drost, F., Van den Akker, R., and Werker, B. (2009). "Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models". *Journal of the Royal Statistical Society Series B* 71.2, pp. 467–485. DOI: 10.1111/j.1467-9868.2008.00687.x.

Du, J.-G. and Li, Y. (1991). "The integer valued autoregressive (INAR(p)) model". *Journal of Time Series Analysis* 12.2, pp. 129–142. DOI: 10.1111/j.1467-9892.1991.tb00073.x.

Fahrmeir, L and Tutz, G. (2001). *Multivariate Statistical Modeling based on Generalized Linear Models*. 2nd ed. Springer. DOI: 10.1007/978-1-4757-3454-6.

Faymonville, M. and Jentsch, C. (2025). *Joint semi-parametric INAR bootstrap inference for model coefficients and innovation distribution*. arXiv: 2507.11124 [stat.ME]. URL: https://arxiv.org/abs/2507.11124.

Faymonville, M., Jentsch, C., and Paparoditis, E. (2025a). *Predictive inference for discrete-valued time series*. arXiv: 2507.16035 [stat.ME]. URL: https://arxiv.org/abs/2507.16035.

Faymonville, M., Jentsch, C., and Weiß, C. H. (2025b). "Semi-parametric goodness-of-fit testing for INAR models". *Bernoulli* 31.4, pp. 3213–3234. DOI: 10.3150/24-BEJ1844.

— (2025c). "Supplement I to "Semi-parametric goodness-of-fit testing for INAR models"". DOI: 10.3150/24-BEJ1844SUPPA.

— (2025d). "Supplement II to "Semi-parametric goodness-of-fit testing for INAR models"". DOI: 10.3150/24-BEJ1844SUPPB.

Faymonville, M., Jentsch, C., Weiß, C. H., and Aleksandrov, B. (2023). "Semiparametric estimation of INAR models using roughness penalization". *Statistical Methods & Applications* 32.2, pp. 365–400. DOI: 10.1007/s10260-022-00655-0.

Faymonville, M., Riffo, J., Rieger, J., and Jentsch, C. (2024). "spINAR: An R package for semiparametric and parametric estimation and bootstrapping of integer-valued autoregressive (INAR) models". *Journal of Open Source Software* 9.97, p. 5386. DOI: 10.21105/joss.05386.

Ferland, R., Latour, A., and Oraichi, D. (2006). "Integer-valued GARCH processes". *Journal of Time Series Analysis* 27.6, pp. 923–942. DOI: 10.1111/j.1467-9892.2006.00496.x.

Fokianos, K. (2010). "Penalized estimation for integer autoregressive models". *Statistical modelling and regression structures*. Ed. by T. Kneib and G. Tutz. Springer, pp. 337–352. DOI: 10.1007/978-3-7908-2413-1_18.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). "Poisson autoregression". *Journal of the American Statistical Association* 104.488, pp. 1430–1439. DOI: 10.1198/jasa.2009.tm08270.

Franke, J. and Seligmann, T. (1993). "Conditional maximum-likelihood estimates for INAR(1) processes and their application to modelling epileptic seizure counts". *Developments in Time Series Analysis*. Ed. by T. S. Rao. London: Chapman & Hall, pp. 157–170. ISBN: 978-0412492600.

Freeland, R. K. and McCabe, B. P. M. (2004). "Forecasting discrete valued low count time series". *International Journal of Forecasting* 20.3, pp. 427–434. DOI: 10.1016/S0169-2070(03)00014-1.

— (2005). "Asymptotic properties of CLS estimators in the Poisson AR(1) model". *Statstics and Probability Letters* 73.2, pp. 147–153. DOI: 10.1016/j.spl.2005.03.006.

Gouveia, S., Möller, T., Weiß, C. H., and Scotto, M. (2018). "A full ARMA model for counts with bounded support and its application to rainy-days time series". *Stochastic Environmental Research and Risk Assessment* 32.9, pp. 2495–2514. DOI: 10.1007/s00477-018-1584-3.

Gürtler, N and Henze, N. (2000). "Recent and classical goodness-of-fit tests for the Poisson distribution". *Journal of Statistical Planning and Inference* 90.2, pp. 207–225. DOI: 10.1016/S0378-3758(00)00114-2.

Heinen, A. (2003). "Modelling time series count data: an autoregressive conditional Poisson model". *CORE Discussion Paper*. DOI: 10.2139/ssrn.1117187.

Hudecova, S., Huskova, M., and Meintanis, S. G. (2021). "Goodness-of-fit tests for bivariate time series of counts". *Econometrics* 9.1, p. 10. DOI: 10.3390/econometrics9010010.

Hudecova, S., Huskova, M., and Meintanis, S. G. (2015). "Tests for time series of counts based on the probability-generating function". *Statistics* 49.2, pp. 316–337. DOI: 10.1080/02331888.2014.979826.

Jazi, M., Jones, G., and Lai, C. (2012a). "First-order integer valued AR processes with zero inflated Poisson innovations". *Journal of Time Series Analysis* 33.6, pp. 954–963. DOI: 10.1111/j.1467-9892.2012.00809.x.

— (2012b). "Integer valued AR(1) with geometric innovations". *Journal of the Iranian Statistical Society* 11.2, pp. 173–190. URL: https://jirss.irstat.ir/article_253688.html.

Jentsch, C. and Weiß, C. H. (2019). "Bootstrapping INAR models". *Bernoulli* 25.3, pp. 2359–2408. DOI: 10.3150/18-BEJ1057.

Jung, R., Ronning, G., and Tremayne, A. (2005). "Estimation in conditional first order autoregression with discrete support". *Statistical Papers* 46.2, pp. 195–224. DOI: 10.1007/BF02762968.

Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*. Wiley. ISBN: 978-0471363552.

Kim, H. Y. and Park, Y. (2008). "A non-stationary integer-valued autoregressive model". *Statistical papers* 49.3, pp. 485–502. DOI: 10.1007/s00362-006-0028-1.

Kreiß, J.-P. (1992). "Bootstrap procedures for AR(∞) processes". *Bootstrapping and Related Techniques*. Ed. by K. Jöckel, G. Rothe, and W. Sendler. Vol. 376. Lecture Notes in Economics and Mathematical Systems. Heidelberg: Springer, pp. 107–113. DOI: 10.1007/978-3-642-48850-4_14.

— (1997). *Asymptotical properties of residual bootstrap for autoregressions*. Technical Report. Braunschweig, Germany: TU Braunschweig.

Kreiß, J.-P., Paparoditis, E., and Politis, D. N. (2011). "On the range of validity of the autoregressive sieve bootstrap". *The Annals of Statistics* 39.4, pp. 2103–2130. DOI: 10.1214/11-AOS900.

Leucht, A. (2012). "Characteristic function-based tests under weak dependence". *Journal of Multivariate Analysis* 108, pp. 67–89. DOI: 10.1016/j.jmva.2012.02.003.

Leucht, A. and Neumann, M. (2013). "Degenerate U- and V-statistics under ergodicity: asymptotics, bootstrap and applications in statistics". *Annals of the Institute of Statistical Mathematics* 65.2, pp. 349–386. DOI: 10.1007/s10463-012-0374-9.

Liboschik, T., Fokianos, K., and Fried, R. (2017). "tscount: An R package for analysis of count time series following generalized linear models". *Journal of Statistical Software* 82.5, pp. 1–51. DOI: 10.18637/jss.v082.i05.

Liu, Z., Li, Q., and Zhu, F. (2021). "Semiparametric integer-valued autoregresive models on $\mathbb{Z}$". *Canadian Journal of Statistics* 49.4, pp. 1317–1337. DOI: 10.1002/cjs.11621.

McKenzie, E. (1985). "Some simple models for discrete variate time series". *Water Resources Bulletin* 21.4, pp. 645–650. DOI: 10.1111/j.1752-1688.1985.tb05379.x.

— (2003). "Discrete variate time series". *Handbook of Statistics 21*. Ed. by D. Shanbhag and C. Rao. Elsevier, pp. 573–606. DOI: 10.1016/S0169-7161(03)21018-X.

Medina Garay, A. W., de Lima Medina, F., and Rossiter Araújo Monteiro, T. A. (2022). *ZINARp: Simulate INAR/ZINAR(p) models and estimate its parameters*. R package version 0.1.0. URL: https://CRAN.R-project.org/package=ZINARp.

Meintanis, S. G. and Karlis, D. (2014). "Validation tests for the innovation distribution in INAR time series models". *Computational Statistics* 29.5, pp. 1221–1241. DOI: 10.1007/s00180-014-0488-z.

Meintanis, S. G. and Swanepoel, J. (2007). "Bootstrap goodness-of-fit tests with estimated parameters based on empirical transforms". *Statistics & Probability Letters* 77.10, pp. 1004–1013. DOI: 10.1016/j.spl.2007.01.014.

Nardi, Y. and Rinaldo, A. (2011). "Autoregressive process modeling via the Lasso procedure". *Journal of Multivariate Analysis* 102.3, pp. 528–549. DOI: 10.1016/j.jmva.2010.10.012.

Pan, L. and Politis, D. (2016). "Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregression". *Journal of Statistical Planning and Inference* 177, pp. 1–27. DOI: 10.1016/j.jspi.2014.10.003.

Qi, X., Li, Q., and Zhu, F. (2019). "Modeling time series of count with excess zeros and ones based on INAR(1) model with zero-and-one inflated Poisson innovations". *Journal of Computational and Applied Mathematics* 346, pp. 572–590. DOI: 10.1016/j.cam.2018.07.043.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Robert-Koch institute (2020). *RKI COVID 19*. URL: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/nCoV.html.

Rydberg, T. H. and Shephard, N. (2000). *BIN models for trade-by-trade data. Modelling the number of trades in a fixed interval of time*. Econometric Society World Congress 2000 Contributed Papers 0740. Econometric Society. URL: https://ideas.repec.org/p/ecm/wc2000/0740.html.

Savani, V. and Zhigljavsky, A. (2007). "Efficient parameter estimation for independent and INAR(1) negative binomial samples". *Metrika* 65.2, pp. 207–225. DOI: 10.1007/s00184-006-0071-x.

Schweer, S. (2016). "A goodness-of-fit test for integer-valued autoregressive processes". *Journal of Time Series Analysis* 37.1, pp. 77–98. DOI: 10.1111/jtsa.12138.

Scott, D. W., Tapia, R. A., and Thompson, J. R. (1980). "Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria". *The Annals of Statistic* 8.4, pp. 820–832. DOI: 10.1214/aos/1176345074.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and its Applications*. 1st ed. Springer. ISBN: 978-0387989501.

Silva, I. and Silva, M. (2006). "Asymptotic distribution of the Yule-Walker estimator for INAR(p) processes". *Statstics and Probability Letters* 76.15, pp. 1655–1663. DOI: 10.1016/j.spl.2006.04.008.

Slutsky, E. (1937). "The summation of random causes as the source of cyclic processes". *Econometrica* 5.2, pp. 105–146. DOI: 10.2307/1907241.

Stapper, M. (2022). "ManuelStapper/CountTimeSeries.jl: v0.1.4". DOI: 10.5281/zenodo.7488440.

Steutel, F. W. and Van Harn, K. (1979). "Discrete analogues of self-decomposability and stability". *Annals of Probability* 7.5, pp. 893–899. DOI: 10.1214/aop/1176994950.

The MathWorks Inc. (2022). *MATLAB version: 9.13.0 (R2022b)*. URL: https://www.mathworks.com.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). "Sparsity and smoothness via the fused Lasso". *Journal of the Royal Statistical Society. Series B* 67.1, pp. 91–108. DOI: 10.1111/j.1467-9868.2005.00490.x.

Turkman, K. F., Scotto, M. G., and de Zea Bermudez, P. (2014). *Non-Linear Time Series: Extreme Events and Integer Value Problems*. Springer. ISBN: 978-3319070278.

Van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer. ISBN: 978-0387946405.

Walker, G. T. (1931). "On periodicity in series of related terms". *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 131.818, pp. 518–532. DOI: 10.1098/rspa.1931.0069.

Wang, X. (2020). "Variable selection for first-order Poisson integer-valued autoregressive model with covariables". *Australian and New Zealand Journal of Statistics* 62.2, pp. 278–295. DOI: 10.1111/anzs.12295.

Wang, X., Wang, D., and Yang, K. (2020). "Integer-valued time series model order shrinkage and selection via penalized quasi-likelihood approach". *Metrika* 84.5, pp. 713–750. DOI: 10.1007/s00184-020-00799-7.

Weiß, C. H. (2010). "The INARCH(1) model for overdispersed time series of counts". *Communications in Statistics - Simulation and Computation* 39.6, pp. 1269–1291. DOI: 10.1080/03610918.2010.490317.

— (2018). *An Introduction to Discrete-Valued Time Series*. 1st ed. Wiley. ISBN: 978-1119096962.

Weiß, C. H. and Kim, H.-Y. (2014). "Diagnosing and modeling extra-binomial variation for time-dependent counts". *Applied Stochastic Models in Business and Industry* 30.5, pp. 588–608. DOI: 10.1002/asmb.2005.

Yule, G. U. (1927). "On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers". *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 226, pp. 267–298. DOI: 10.1098/rsta.1927.0007.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. CRC Press. DOI: 10.1201/9781420010893.

Publications

Statistical Methods & Applications
https://doi.org/10.1007/s10260-022-00655-0

ORIGINAL PAPER

Semiparametric estimation of INAR models using roughness penalization

Maxime Faymonville[1] · Carsten Jentsch[1] · Christian H. Weiß[2] · Boris Aleksandrov[2]

Accepted: 17 August 2022

Abstract

Popular models for time series of count data are integer-valued autoregressive (INAR) models, for which the literature mainly deals with parametric estimation. In this regard, a semiparametric estimation approach is a remarkable exception which allows for estimation of the INAR models without any parametric assumption on the innovation distribution. However, for small sample sizes, the estimation performance of this semiparametric estimation approach may be inferior. Therefore, to improve the estimation accuracy, we propose a penalized version of the semiparametric estimation approach, which exploits the fact that the innovation distribution is often considered to be smooth, i.e. two consecutive entries of the PMF differ only slightly from each other. This is the case, for example, in the frequently used INAR models with Poisson, negative binomially or geometrically distributed innovations. For the data-driven selection of the penalization parameter, we propose two algorithms and evaluate their performance. In Monte Carlo simulations, we illustrate the superiority of the proposed penalized estimation approach and argue that a combination of penalized and unpenalized estimation approaches results in overall best INAR model fits.

Keywords Count data · Penalized estimation · Integer-valued autoregressions · Innovation distribution · Validation

Carsten Jentsch, Christian Weiß and Boris Aleksandrov contributed equally to this work.

Extended author information available on the last page of the article

Published online: 21 September 2022

1 Introduction

According to Du and Li (1991), the INAR(p) model is defined by the recursion

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \ldots + \alpha_p \circ X_{t-p} + \varepsilon_t,\ t \in \mathbb{Z}, \tag{1}$$

with innovation process $\varepsilon_t \overset{\text{i.i.d}}{\sim} G$, where the distribution G has range $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. Furthermore, let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)' \in (0,1)^p$ denote the vector of model coefficients with $\sum_{i=1}^{p} \alpha_i < 1$ and

$$\alpha_i \circ X_{t-i} = \sum_{j=1}^{X_{t-i}} Z_j^{(t,i)},$$

where "$\circ$" is the binomial thinning operator first introduced by Steutel and Van Harn (1979). Here, $\left(Z_j^{(t,i)}, j \in \mathbb{N}, t \in \mathbb{Z}\right)$, $i \in 1, \ldots, p$, are mutually independent Bernoulli-distributed random variables $Z_j^{(t,i)} \sim \text{Bin}(1, \alpha_i)$ with $P(Z_j^{(t,i)} = 1) = \alpha_i$ independent of $(\varepsilon_t, t \in \mathbb{Z})$. The special case $p = 1$ results in the INAR(1) model introduced by McKenzie (1985) and Al-Osh and Alzaid (1987). All the thinning operations "$\circ$" are independent of each other and of ε_t, $t \in \mathbb{Z}$. Furthermore, the thinning operation at time t and ε_t are independent of X_s, $s < t$.

Most researchers deal with parametric estimation of INAR models (see for example Franke and Seligmann (1993), Freeland and McCabe (2005), Brännäs and Hellström (2001) and Jung et al. (2005)), i.e. they assume G to lie in some parametric class of distributions $(G_\theta \mid \theta \in \Theta \subset \mathbb{R}^q)$ for some finite $q \in \mathbb{N}$. In contrast, Drost et al. (2009) introduced a semiparametric estimator, which on the one hand keeps the parametric assumption of the binomial thinning operation, but on the other hand allows to estimate the innovation distribution nonparametrically. Using empirical process theory, they derive asymptotic theory in terms of consistency and asymptotic normality results and proved efficiency. Consequently, their estimation approach does not require any parametric assumption regarding the innovation distribution, and avoids the risk of a falsely specified parametric assumption and its undesirable consequences. The approach estimates the coefficients of INAR models and the innovation distribution simultaneously. The resulting semiparametric maximum likelihood estimator

$$(\hat{\boldsymbol{\alpha}}_{sp}, \hat{G}_{sp}) = (\hat{\alpha}_{sp,1}, \ldots, \hat{\alpha}_{sp,p}, \hat{G}_{sp}(0), \hat{G}_{sp}(1), \hat{G}_{sp}(2), \ldots),$$

where $\hat{\boldsymbol{\alpha}}_{sp} = (\hat{\alpha}_{sp,1}, \ldots, \hat{\alpha}_{sp,p})$ denotes the vector of the estimated INAR coefficients and $\{\hat{G}_{sp}(k), k \in \mathbb{N}_0\}$ are the estimated entries of the probability mass function (PMF) of G, maximizes the conditional log-likelihood function $\log(\mathcal{L}(\boldsymbol{\alpha}, G))$, i.e.

$$\forall n \in \mathbb{Z}_+ : (\hat{\boldsymbol{\alpha}}_{sp}, \hat{G}_{sp}) \in \underset{(\boldsymbol{\alpha},G)\in[0,1]^p\times\tilde{\mathcal{G}}}{\arg\max} \left(\prod_{t=0}^{n} P^{\boldsymbol{\alpha},G}_{(X_{t-1},\ldots,X_{t-p}),X_t} \right). \tag{2}$$

Here, $\tilde{\mathcal{G}}$ is the set of all probability measures on $\mathbb{Z}_+$ and $P^{\boldsymbol{\alpha},G}_{(X_{t-1},\ldots,X_{t-p}),X_t}$ are the transition probabilities under the true model parameters $\boldsymbol{\alpha}$ and G, i.e.

$$\begin{aligned} P^{\boldsymbol{\alpha},G}_{(x_{t-1},\ldots,x_{t-p}),x_t} &= \mathbb{P}_{\boldsymbol{\alpha},G}\left(\sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \varepsilon_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p} \right) \\ &= (\mathrm{Bin}(x_{t-1}, \alpha_1) * \ldots * \mathrm{Bin}(x_{t-p}, \alpha_p) * G)\{x_t\}, \end{aligned}$$

with $\mathbb{P}$ the underlying probability measure and "$*$" denoting the convolution of distributions. In the special case of an INAR(1) model the transition probabilities are given by

$$\mathbb{P}_{\boldsymbol{\alpha},G}(X_t = x_t \mid X_{t-1} = x_{t-1}) = \sum_{j=0}^{\min(x_t, x_{t-1})} \binom{x_{t-1}}{j} \alpha^j (1-\alpha)^{x_{t-1}-j} \mathbb{P}_{\boldsymbol{\alpha},G}(\varepsilon_t = x_t - j),$$

where α is the coefficient of the INAR(1) model (McKenzie 1985; Al-Osh and Alzaid 1987). For $k < \min\{X_t - \sum_{i=1}^{p} X_{t-i} \mid t = p+1, \ldots, n\}$ or $k > \max\{X_t \mid t = 1, \ldots, n\}$, the values $\hat{G}_{sp}(k)$, $k \in \mathbb{N}_0$, are equal to 0. For further details, see Drost et al. (2009).

In practice, discrete probability distributions such as the Poisson, the negative binomial or the geometric distribution are often used as innovation distribution G, see Weiß (2018), Yang (2019), Al-Osh and Alzaid (1987), Al-Osh and Alzaid (1990). The common feature of all these distributions is their smoothness in the sense that consecutive entries of their PMFs differ only slightly from each other. However, for a small sample size n, the semiparametric estimation approach of Drost et al. (2009) may lead to rather non-smooth innovation distributions with unnatural gaps in their PMF. For illustration, we consider a time series containing counts of transactions of structured products (factor long certificates with leverage) from on-market and off-market trading per trading day between February 1, 2017 and July 31, 2018 (thus $n = 381$). These data, which are plotted in Fig. 1, have first been presented by Homburg et al. (2021), who derived them from the Cascade-Turnoverdata of the Deutsche Börse Group. In the upper right corner, we see the estimated innovation distribution using the semiparametric procedure of Drost et al. (2009) which turns out to be smooth. In the second row, we consider only the first 100 observations of the time series, where the first plot shows indeed a bimodal estimated innovation distribution. In the third row, we only considered the first 20 observations. The lower-left plot shows the resulting estimated PMF, which contains an unnatural gap with $\hat{G}_{sp}(3)$ being estimated exactly equal to zero while its neighbors $\hat{G}_{sp}(2)$ and $\hat{G}_{sp}(4)$ are estimated positive. Hence, the resulting estimation is not smooth contrary to the estimated innovation distribution on the whole time series. In general, such non-smooth innovation distributions are not common in practice and instead, smoothly estimated innovation distributions are often desired. In this paper, we want to use this

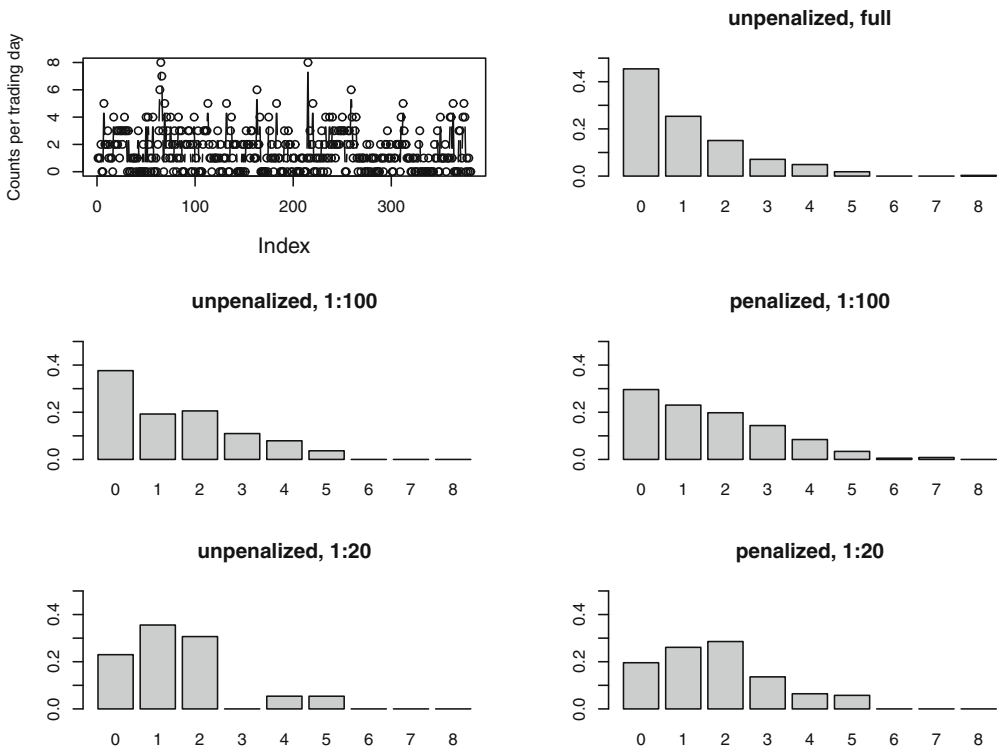

Fig. 1 From left to right and top to bottom: Plot of time series of counts of transactions of structured products per trading day, the unpenalized estimation of the corresponding innovation distribution based on the full data and the (un)penalized estimated innovation distribution for the first 100 and 20 observations, respectively

prior knowledge and take advantage of a natural qualitative smoothness assumption on the innovation distribution by proposing a version of the semiparametric estimation approach, which penalizes the roughness of the innovation distribution. The resulting estimated PMFs of this approach are contained in the right plots in the second and third row, respectively. In comparison, the penalized estimation now leads to a smoother estimation of the PMF without any gaps. We will have a closer look at additional real data examples in Sect. 4. For long time series, the smoothing caused by penalization is not of such great importance, because the distribution estimated without penalization will be sufficiently smooth by itself. But for short time series, estimation without smoothing will commonly lead to jagged estimated innovation distributions although the true distribution behind the data might be smooth. So the need for smoothing is of particular importance for short time series.

The paper is organized as follows. In Sect. 2, we introduce a penalized estimation approach using roughness penalization and propose two algorithms for the data-driven selection of the penalization parameter. Section 3 examines our estimation approach in a comprehensive simulation study, where we compare the estimation performance of the penalized and the unpenalized approach for different settings. In a real data application in Sect. 4, we analyze the monthly demand of car spare parts to illustrate our method and its practical relevance. In the conclusion in Sect. 5, we summarize the results and give an outlook on further research questions.

2 Penalized approach of fitting INAR models

Penalized estimation of count data is a modern topic in current statistical research. Bui et al. (2021) consider parameter estimation in count data models using penalized likelihood methods. In a time series context, Nardi and Rinaldo (2011) studied LASSO penalization for fitting autoregressive time series models to get sparse solutions, i.e. where some autoregressive coefficients are estimated exactly as zero. Fokianos (2010) proposed an alternative estimation scheme for the estimation of INAR models based on minimizing the least square criterion under ridge type of constraints. Wang (2020) proposed a variable selection procedure for INAR(1) models with Poisson distributed innovations including covariables by using penalized estimation and Wang et al. (2021) introduced an order selection procedure for INAR(p) and INARCH(p) models also by using penalized estimation. By contrast, in this paper, we propose a penalized estimation approach for INAR models which does not rely on a penalization of the INAR coefficients (towards zero), but on a penalization of the roughness of the innovation distribution (towards smoothness).

2.1 Penalized estimation approach using roughness penalty

The idea of our approach is to penalize the log-likelihood used in the semiparametric estimation of the INAR model according to Drost et al. (2009). Thus, we still do not assume a parametric class of distributions, we only use the assumed qualitative (i.e. nonparametric) property of smoothness. More precisely, this refers to a roughness penalization as introduced by Scott et al. (1980), which is e.g. used by Adam et al. (2019) for developing a nonparametric approach to fit hidden Markov models to time series of counts. We design the penalty term based on the idea of Tibshirani et al. (2005), where differences of successive parameters are penalized. In this regard, we allow for differences of order $m \in \mathbb{N}$. Applied to our setting, the estimation approach based on Drost et al. (2009) now maximizes the penalized log-likelihood (compare (2))

$$\log(\mathcal{L}_{\text{pen}}(\alpha, G)) = \log(\mathcal{L}(\alpha, G)) - \eta \cdot d_{G,m},$$

where $\eta > 0$ is the so-called smoothing or penalization parameter, $d_{G,m}$ denotes a suitable measure to quantify the roughness of G and m corresponds to the order of difference. According to Tibshirani et al. (2005), a first possible roughness measure for the penalization term is based on the L_1 distance (LASSO penalization), i.e.

$$d_{G,m,1} = \sum_{i=m}^{\max(x_1,\ldots,x_n)} | \Delta^m G(i) |, \tag{3}$$

where $\Delta^m G(i) = \Delta^{m-1}(\Delta G(i))$ and $\Delta G(i) = G(i) - G(i-1)$. In addition, we consider the squared L_2 distance (Ridge penalization) as second roughness measure, i.e.

$$d_{G,m,2} = \sum_{i=m}^{\max(x_1,\ldots,x_n)} (\Delta^m G(i))^2. \tag{4}$$

The idea behind choosing this second roughness measure is that it does not shrink the differences of the successive entries of the PMF exactly to 0 (contrary to the first roughness measure), but the differences become close to 0, which is more in line with the idea of a smooth distribution (note the analogy of penalized regression, where the L_1 penalization is used for variable selection because of this property, see Fahrmeir et al. (2013)). The order of the differences m is a tuning parameter. For $m = 1$, we penalize only the distance between two directly consecutive entries, for $m = 2$ the smoothness is extended to a triple of values, etc.

Remark 1 A possible extension would be to allow for different penalization weights (η_i) for the individual (higher-order) differences of the entries of the PMF. For instance, in the case of L_1 penalization, the goal could be to maximize

$$\log(\mathcal{L}(\alpha, G)) - \sum_{i=m}^{\max(x_1,\ldots,x_n)} \eta_i \mid \Delta^m G(i) \mid,$$

analogously for the case of L_2 penalization.

Figure 2 shows a first exemplary result on a sample of an INAR(1) process with $n = 25$ observations, order of difference $m = 1$ and smoothing parameter $\eta = 1$ roughly chosen by eye. In this example, the benefit of penalization already becomes clear. The penalized estimated innovation distributions are much closer to the true Poi(1) innovation distribution (which was truncated at value six for clarity) than the unpenalized estimated innovation distribution. Also, the difference between the L_1 and the L_2 penalization becomes visible. When using the L_2 penalization, the distances between the values of the PMF become small, when using the L_1 penalization they are shrinked to zero.

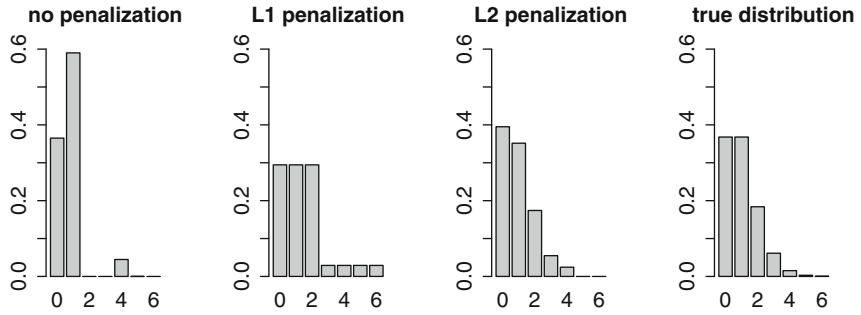

Fig. 2 Barplots of the (estimated) innovation distributions for one realization in the four cases (no penalization, L_1 penalization, L_2 penalization, true distribution)

2.2 Selection of the penalization parameter

Now, we propose two approaches to determine for a fixed roughness measure the optimal smoothing/penalization parameter η, which is a trade-off between fit to the data and the smoothness assumption. For this purpose, we adapt as a first approach the cross-validation procedure described in Adam et al. (2019) to our setting. Therefore, we split the data set into s blocks F_i, $i = 1, \ldots, s$, of roughly equal size. In each fold i, $F_{(-i)}$ denotes the in-sample data (data without F_i) and F_i the out-of-sample data. This replicates the correct dependence structure except for the "glue points", which only has a minor effect in practice when the data originate from an INAR model of small order. The greedy search algorithm is structured as follows:

Algorithm 1

(1) Choose an initial $\eta^{(0)} > 0$ and set $z = 0$.
(2) For each fold i and for each value on a specified grid

$$\{\ldots, \eta^{(z)} - 2c, \eta^{(z)} - c, \eta^{(z)}, \eta^{(z)} + c, \eta^{(z)} + 2c, \ldots\}$$

where $c \in \mathbb{R}$ is a small constant, estimate the model with penalization on $F_{(-i)}$ and compute the penalized log-likelihood on F_i.
(3) Average the resulting log-likelihood values across all folds i and choose $\eta^{(z+1)}$ as the penalization parameter on the grid that yields the maximum value.
(4) Repeat steps 2) and 3) until $\eta^{(z+1)} = \eta^{(z)}$ and define $\eta^{\text{opt}} := \eta^{(z+1)}$.

Furthermore, to avoid a potentially non-optimal selection of the penalization parameter η caused by an inappropriate choice of the initial value $\eta^{(0)}$, we propose a second optimization algorithm. How we split the data in each fold j, $j = 1, \ldots, \tilde{s}$, in in- and out-of-sample data is specified later in Sect. 3.

Algorithm 2

(1) For each fold j and each value η on a specified grid $\{0, \tilde{c}, 2\tilde{c}, 3\tilde{c}, \ldots, u\}$ on the interval $[0,u]$ for an appropriate upper bound u, estimate the model with penalization on the in-sample data and compute the penalized log-likelihood on the out-of-sample data.
(2) Average the resulting log-likelihood values across all folds j.
(3) Fit a polynomial of order r to the curve resulting from plotting the average out-of-sample log-likelihood against the grid.
(4) Choose η^{opt} as the value on the grid, where the curve takes its maximum value.

3 Simulation study

We investigate the performance of the proposed procedure in a simulation study with $K = 500$ Monte Carlo samples of size $n \in \{20, 50, 100, 250, 500, 1000\}$ generated from an INAR(1) process according to (1) for $p = 1$ with different coefficients

$\alpha \in \{0.2, 0.5, 0.8\}$ and innovation distributions $G \in \{\text{Poi}(1), \text{NB}\left(2, \frac{2}{3}\right), \text{Geo}\left(\frac{1}{2}\right), \text{ZIP}\left(\frac{1}{2}, 2\right)\}$, where ZIP denotes a zero-inflated Poisson distribution as in Jazi et al. (2012). The parameters of the negative binomial, geometric and zero-inflated Poisson distribution are hereby chosen to have the same expected value as the $\text{Poi}(1)$ distribution. But contrary to the $\text{Poi}(1)$ distribution which is equidispersed, i.e. the variance of the distribution equals its mean, they are overdispersed, i.e. their variances are larger than their mean values. Another difference between the considered innovation distributions is their (non-) smoothness, see also Fig. 12 in the appendix. The Poi(1), $\text{NB}\left(2, \frac{2}{3}\right)$ and $\text{Geo}\left(\frac{1}{2}\right)$ distributions are rather smooth, but the $\text{ZIP}\left(\frac{1}{2}, 2\right)$ distribution, which shows a pronounced zero probability, is not. The effect of this property on the roughness penalization is investigated in Subsect. 3.5. Moreover, in Subsect. 3.2, we also provide a small simulation setting for higher-order INAR processes and consider the case of an INAR(2) model. The implementation is straightforward but is a lot more demanding such that we restrict the considered setting to a rather small extent. To ensure the stationarity of the time series, we actually generate $n + 100$ observations and remove the first 100 observations. We consider first ($m = 1$) and second ($m = 2$) order differences in the penalization term (see Subsect. 3.4). As initialization for the smoothing parameter $\eta^{(0)}$, we set $\eta^{(0)} = 1$ as in the example in Fig. 2 for the sample sizes $n \in \{20, 50, 100, 250\}$ and for computing time reasons $\eta^{(0)} = 0.5$ for $n \in \{500, 1000\}$.[1] For the considered grid around the smoothing parameter (see Algorithm 1) we choose $c = 0.05$ resulting in $\{\eta^{(z)} - 0.1, \eta^{(z)} - 0.05, \eta^{(z)}, \eta^{(z)} + 0.05, \eta^{(z)} + 0.1\}$. Unless stated otherwise, we use $\alpha = 0.5$ as true INAR(1) coefficient and Algorithm 1 with 10-fold cross validation ($s = 10$) as optimization algorithm. For the realization of the simulation study, we use the statistical programming language R 4.1.2 (R Core Team 2021).

3.1 Roughness penalty for smooth innovations distributions and first order differences

Figure 3 shows the L_2 distances of the estimated innovation distributions to the true $\text{Poi}(1)$ innovation distribution,

$$d(\hat{G}, G) = \sum_{i=0}^{M} (\hat{G}(i) - G(i))^2,$$

for the different sample sizes and the respective estimation methods (unpenalized (up), L_1 penalization and L_2 penalization) for some large enough M. We use $M = 70$ as upper bound for the observations $x_1, \ldots, x_n$ since after this value the corresponding probabilities of occurrence are negligibly small. When the sample size n is small, the penalized estimation of the innovation distribution provides a large benefit compared to the unpenalized estimation: The L_2 distances of the penalized estimated to the true innovation distribution are much smaller than those of the unpenalized estimated to the true innovation distribution. Furthermore, the L_2 penalization performs better

[1] For large sample sizes n, η^{opt} will be close to zero, so a lower initial value decreases the number of iterations needed for Algorithm 1, which saves computing time.

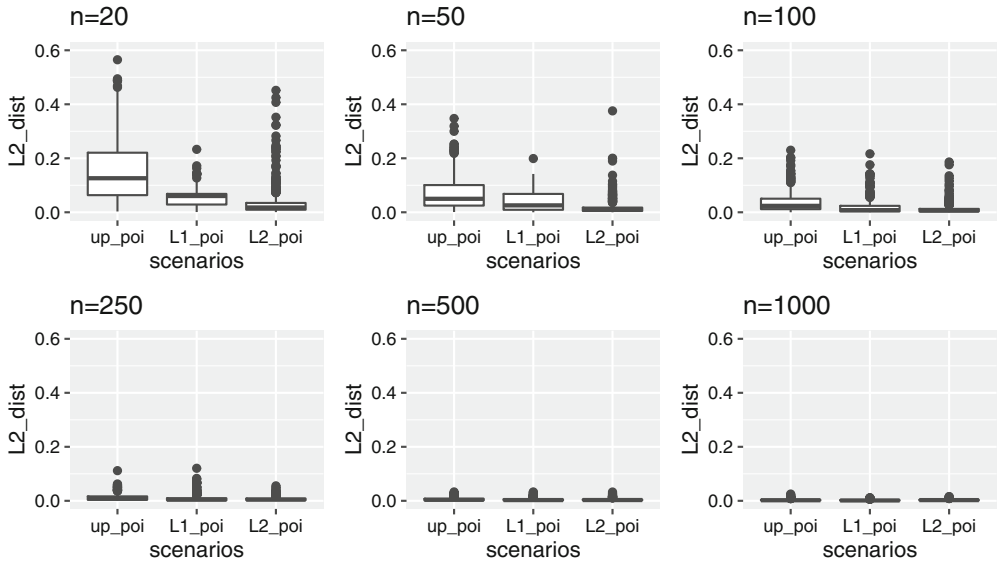

Fig. 3 Boxplots of the L_2 distances of the estimated innovation distribution to the true Poi(1) innovation distribution of an INAR(1) process for different sample sizes n. We report results for unpenalized (up), L_1 and L_2 penalized estimation

than the L_1 penalization. In Table 3 in the appendix, we also report the variance, the bias and the MSE of the first five estimated entries of the PMF resulting from the different procedures for the different sample sizes n. We see that the penalized estimation reduces both the variance, the absolute bias and consequently also the MSE of the estimated innovation distribution, especially for small n. Figures 17 and 18 and Tables 6 and 7 in the appendix show the analog results for a true $\mathrm{NB}\left(2, \frac{2}{3}\right)$ and $\mathrm{Geo}\left(\frac{1}{2}\right)$ distribution, respectively. In general, regardless of the distribution and up to a sample size of $n = 100$, we see a clear improvement concerning the estimation performance when using penalization. From a sample size of $n = 250$ on, this improvement can only be seen marginally with the different methods essentially coinciding for large n. In Fig. 13 and Table 4 in the appendix, we show the results for INAR coefficient $\alpha = 0.2$ and Poi(1) innovation distribution and, correspondingly, in Fig. 15 and Table 5 for $\alpha = 0.8$. In the latter case, the benefit of the penalized estimation compared to the unpenalized estimation is even larger than in the case $\alpha = 0.5$. This is plausible because it is in general more difficult to estimate the innovation distribution for a larger value of α as this also leads to a larger observations mean with innovations mean remaining constant. Therefore, more entries of the PMF have to be estimated with the same amount of data. Contrary, for $\alpha = 0.2$, we have (with analog arguments) less entries of the PMF which have to be estimated with the same amount of data, which simplifies the estimation of the PMF in general and the benefit of penalization decreases. Altogether, we can conclude that the benefit of penalization is more pronounced with larger α, that is, with larger serial dependency.

We get confirming conclusions, when we consider the values of the optimal smoothing parameter η, which approaches zero with increasing n, see Fig. 4 for the

case of a true Poi(1) innovation distribution, Figs. 19 and 20 in the appendix for the cases of a true NB$\left(2, \frac{2}{3}\right)$ and Geo$\left(\frac{1}{2}\right)$ innovation distribution and Figs. 14 and 16 in the appendix in case of a true Poi(1) innovation distribution with $\alpha = 0.2$ and $\alpha = 0.8$, respectively. Thus, for increasing n, the penalized and the unpenalized estimation coincide as intuitively expected: For large n, there are enough observations to learn the smoothness of the innovation distribution from the data even without imposing smoothness through penalization.

3.2 Higher-order INAR processes

To show that our proposed procedure is also applicable for higher-order INAR processes, we consider the case of a true INAR(2) process according to (1) for $p = 2$ with coefficients $\alpha_1 = 0.3$, $\alpha_2 = 0.2$ and $G = \text{Poi}(1)$. Due to the high computing time for the semiparametric estimation, we only consider a small simulation setup with $n = 50$ observations and $K = 100$ Monte Carlo samples. We consider L_1 and L_2 penalization with first order differences and compare the performance with the case of estimation without penalization. In Fig. 21 in the appendix, we see that also for higher-order INAR models, penalized estimation of the innovation distribution provides a clear benefit compared to unpenalized estimation. With penalization we are closer to the true innovation distribution than without and we are able to reduce the variance, the absolute bias and consequently the MSE of our estimation, see Table 1. Again, L_2 penalization works best.

3.3 Alternative selection of the penalization parameter

To investigate whether the results depend on the chosen initial parameter, we now determine the optimal penalization parameter alternatively using Algorithm 2 with $u = 5$, $\tilde{c} = 0.1$ and $r = 5$. In this context, we want to address a potential practical issue of Algorithm 1: the generation of the in- and out-of-sample data. For each of

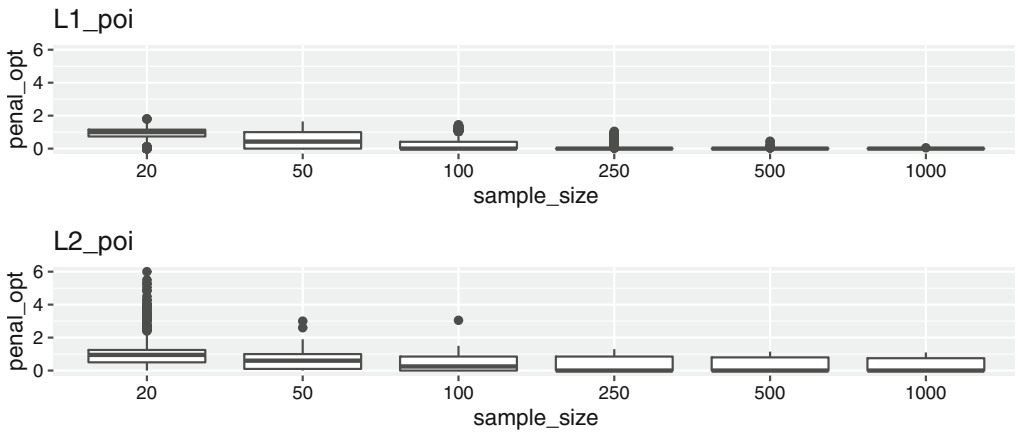

Fig. 4 Boxplots of the penalization parameter η selected by L_1 penalization (upper panel) and L_2 penalization (lower panel) for the different sample sizes n in the case of a true Poi(1) innovation distribution of an INAR(1) process

Table 1 Variance, bias and MSE of the first five estimated entries of the PMF for $n = 50$ in case of a true Poi(1) innovation distribution of an INAR(2) process. We report results for unpenalized (up), L_1 and L_2 penalized estimation

n		g0_up	g0_L1	g0_L2	g1_up	g1_L1	g1_L2
50	Variance	0.0269	0.0080	0.0070	0.0244	0.0037	0.0028
	Bias	−0.0499	−0.0571	−0.0016	−0.0317	−0.0786	−0.0393
	MSE	0.0294	0.0113	0.0070	0.0254	0.0099	0.0043

n		g2_up	g2_L1	g2_L2	g3_up	g3_L1	g3_L2
50	Variance	0.0193	0.0054	0.0045	0.0087	0.0073	0.0030
	Bias	0.0369	0.0266	0.0193	0.0227	0.0719	0.0202
	MSE	0.0207	0.0061	0.0048	0.0092	0.0124	0.0034

n		g4_up	g4_L1	g4_L2
50	Variance	0.0025	0.0043	0.0009
	Bias	0.0127	0.0511	0.0070
	MSE	0.0026	0.0069	0.0009

the 10 folds, 90% of the data becomes the in-sample data and the remaining 10% the out-sample data. For small n, 10% of the data is small. To avoid this, we now use an n-fold cross-validation ($\tilde{s} = n$) for sample sizes $n \in \{20, 50\}$ with Algorithm 2, where starting from each observation the following 50% of the data is in- and the other 50% is out-of-sample. When reaching the end of the time series, we start again from its beginning.

In Fig. 5, we see the results of this alternative procedure compared to the previous (iterative) procedure in Algorithm 1. It gives slightly better results than the iterative method, but overall the distances are very similar. The same can be concluded when considering Table 8. The alternative procedure leads to slightly lower MSE values, but altogether the values resemble each other. The 10-fold cross-validation also seems to be suitable and the resulting optimal parameters of the two procedures are close to each other (see Fig. 6). In conclusion, if we determine the optimal parameter from a sequence on a grid as in Algorithm 2, we tend to get slightly better results. However, the price to pay is a much higher computing time than with the iterative procedure. The iterative method needs a reasonably chosen starting value, but then it gives similarly good results in considerably less computing time. In addition, when using the alternative method, the question arises how to choose the upper limit of the interval adequately. In the following, we will continue to use the iterative method from Algorithm 1 but one should keep in mind that Algorithm 2 is also a practically useful procedure.

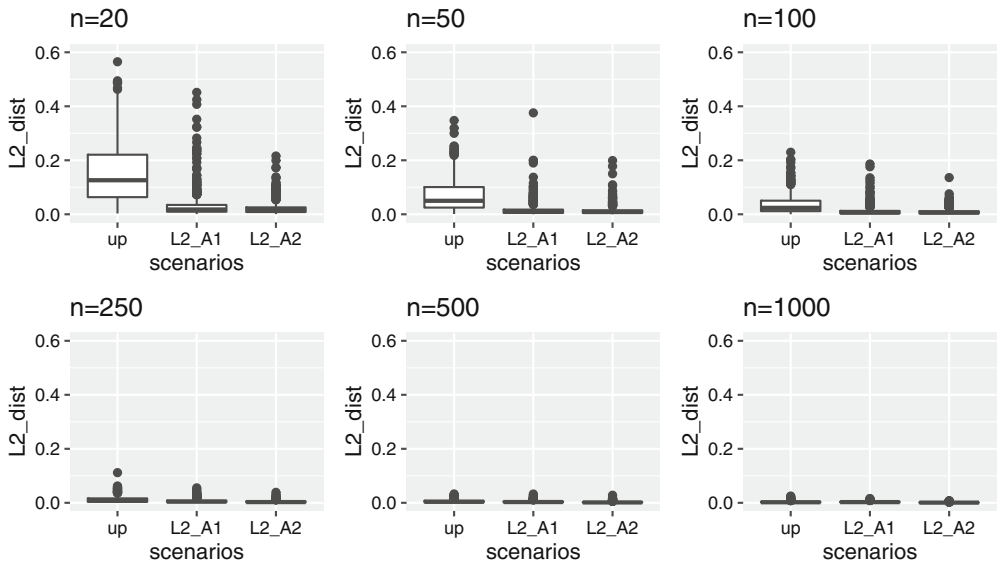

Fig. 5 Boxplots of the L_2 distances of the estimated innovation distribution to the true Poi(1) innovation distribution of an INAR(1) process for the different sample sizes n. We report results for unpenalized (up) and L_2 penalized estimation using either the iterated Algorithm 1 (A1) or the alternative Algorithm 2 (A2)

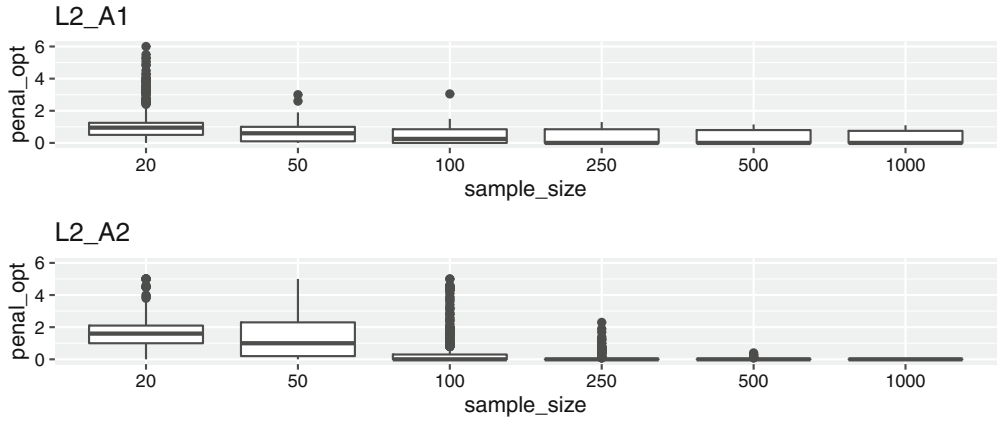

Fig. 6 Boxplots of the penalization parameter η selected by L_2 penalization using Algorithm 1 (A1, upper panel) and Algorithm 2 (A2, lower panel) for the different sample sizes n in the case of a true Poi(1) innovation distribution of an INAR(1) process

3.4 Higher-order differences in penalization term

So far we only considered first order differences ($m = 1$). Now we want to see if penalizing higher-order differences (e.g. $m = 2$) is able to improve the performance of our penalized estimation method. In Fig. 7 and Table 10 in the appendix it is visible for the case of a true Poi(1) innovation distribution and L_2 penalization that also the penalization of differences of higher order performs better than the unpenalized estimation in the cases of small sample sizes, and that it comes close to the penalization of first order differences. Similar results are in Fig. 22 and Table 9,

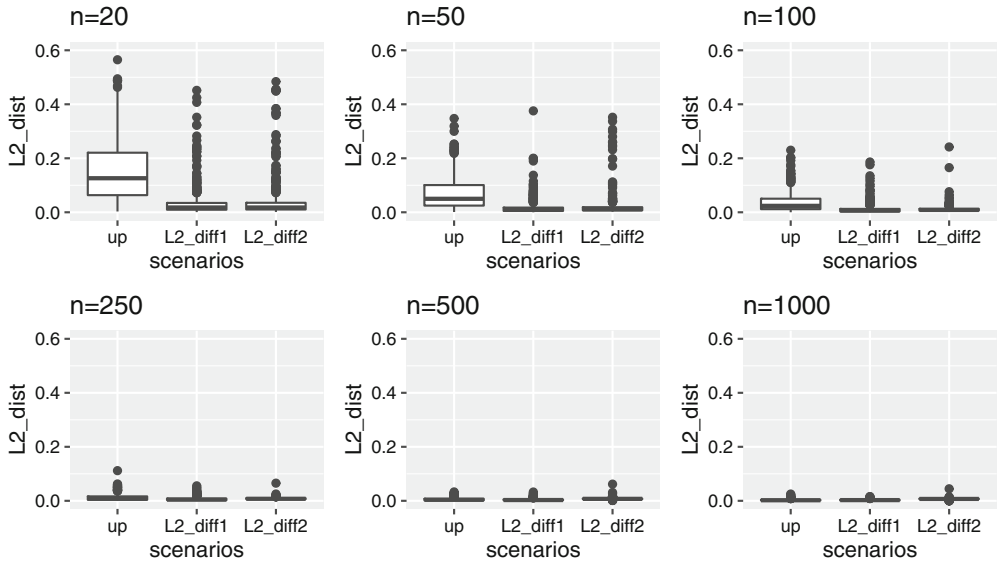

Fig. 7 Boxplots of the L_2 distances of the estimated innovation distribution to the true Poi(1) innovation distribution of an INAR(1) process for the different sample sizes n. We report results for unpenalized (up) and L_2 penalized estimation using either first order (diff1) or second order (diff2) differences

both in the appendix, where we see the results of first and second order differences for the L_1 penalization. In case of L_1 penalization, we would prefer second order differences for small sample sizes. Overall, however, the L_2 penalization of first-order differences performs best.

3.5 Non-smooth innovation distribution

Finally, let us consider the case of $\text{ZIP}\left(\frac{1}{2}, 2\right)$ distributed innovations and consider Fig. 8. The results are as expected. Since the ZIP distribution is not smooth (see Fig. 12 in the appendix), the smoothness assumption and hence the penalization is not suitable. The boxplots reflect this: Except for sample size $n = 20$, the penalized estimation procedure provides no benefit and for some n even leads to slightly higher L_2 distances from the true $\text{ZIP}\left(\frac{1}{2}, 2\right)$ distribution than the unpenalized procedure. As we can see in Table 11, the penalized estimation leads to a higher absolute bias when estimating the first (non-smooth) entry, $G(0)$, of the PMF. As sample size n increases, the penalization has less impact, as there is enough data to detect the incorrect assumption such that the unpenalized and the penalized procedures coincide.

For comparison, let's take a look at the results with a true $\text{ZIP}\left(\frac{1}{2}, 2\right)$ distribution when we exclude $G(0)$ from the penalization displayed in Fig. 9, i.e. when we consider

$$\tilde{d}_{G,m,1} = \sum_{i=m+1}^{\max(x_1,\ldots,x_n)} \mid \Delta^m G(i) \mid \quad \text{and} \quad \tilde{d}_{G,m,2} = \sum_{i=m+1}^{\max(x_1,\ldots,x_n)} (\Delta^m G(i))^2,$$

instead of $d_{G,m,1}$ and $d_{G,m,2}$ defined in (3) and (4). It becomes clear what we would

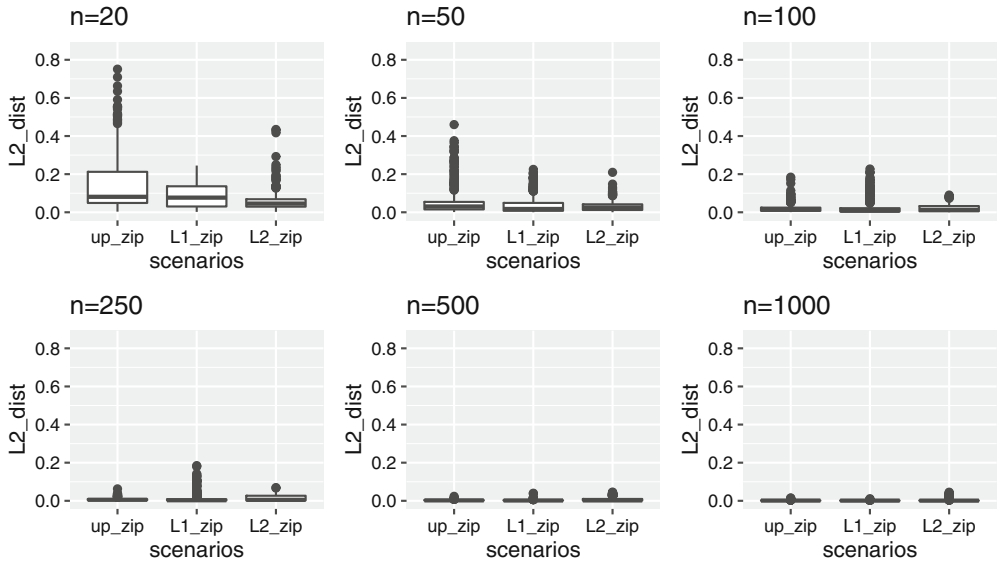

Fig. 8 Boxplots of the L_2 distances of the estimated innovation distribution to the true $\mathrm{ZIP}\left(\frac{1}{2}, 2\right)$ innovation distribution of an INAR(1) process for the different sample sizes n. We report results for unpenalized (up), L_1 and L_2 penalized estimation

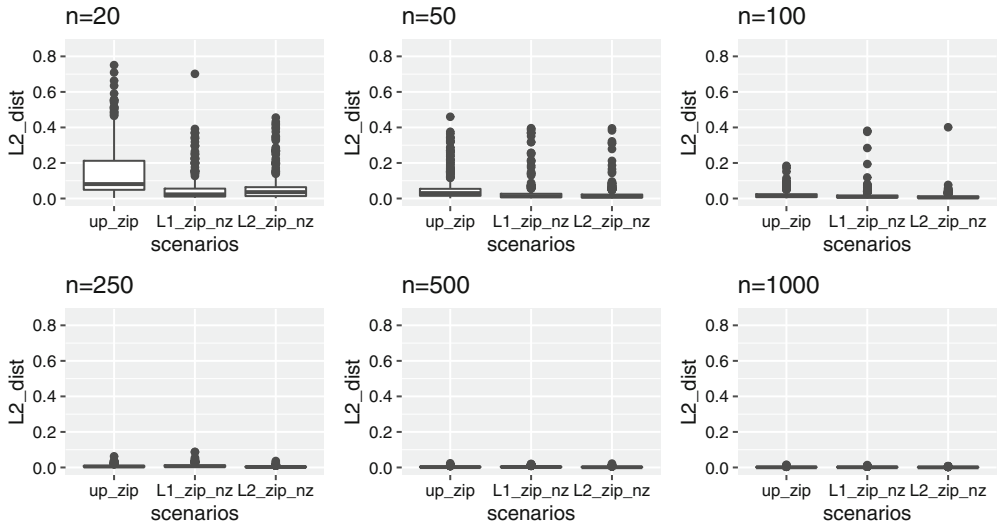

Fig. 9 Boxplots of the L_2 distances of the estimated innovation distribution to the true $\mathrm{ZIP}\left(\frac{1}{2}, 2\right)$ innovation distribution of an INAR(1) process for the different sample sizes n. We report results for unpenalized (up), L_1 and L_2 penalized estimation without smoothing of $G(0)$ (nz)

expect: By excluding the "non-smooth entry" $G(0)$ of the PMF of the innovation distribution from penalization, the penalized estimation works well again and provides a benefit for small n. In this case, the penalized estimation now results in a lower absolute bias of the estimated PMF's first entry compared to the unpenalized estimation (compare Table 12). However, this benefit is not as pronounced as in the cases of a true Poi(1), $\mathrm{NB}\left(2, \frac{2}{3}\right)$ and $\mathrm{Geo}\left(\frac{1}{2}\right)$ innovation distribution. This can probably be explained by the fact that the $\mathrm{ZIP}\left(\frac{1}{2}, 2\right)$ distribution has most of its mass

in zero and the corresponding entry of the PMF, $G(0)$, remains unaffected by the penalization. Consequently, the results from penalized and unpenalized estimation do not differ substantially from each other.

In summary, if the smoothness assumption of the innovation distribution is correctly imposed, it provides a large benefit for small sample size n. This holds whether the true underlying distribution is equidispersed or overdispersed. The best results are obtained for L_2 penalization and first-order differences.

3.6 Estimation of the INAR coefficient

A drawback of the penalized estimation is that the estimation of the INAR coefficient α no longer works well for small sample size n, see Fig. 23 in the appendix. A strength of the semiparametric estimation approach of Drost et al. (2009) is the accurate joint estimation of the INAR coefficient and the innovation distribution. This joint estimation accuracy is not maintained when penalization is used for small n. The L_2 distances of the penalized estimated INAR coefficient α to the true value are higher than for the unpenalized estimated coefficient. For increasing n, the estimation of α improves, but since the benefit of the penalized estimation lies in the cases where n is small, this is no comfort.

Instead, we can solve this problem by taking only the estimator for the innovation distribution from the penalized approach and estimating the INAR coefficient with the unpenalized (efficient) estimation approach of Drost et al. (2009). In Fig. 23, we see that it is indeed preferable to combine the unpenalized estimation of the INAR coefficient α and the penalized estimation of the innovation distribution G. Also when looking at the MSE, it is clear that this combination outperforms all other estimation approaches under consideration.

4 Real data example

For modeling intermittent demand, Syntetos and Boylan (2021) consider the equidispersed Poisson distribution on the one hand, and, as the demand variability may be severe when demand is intermittent, overdispersed distributions from the Compound-Poisson family (such as the negative binomial distribution) on the other hand. All these parametric distributions are smooth. With our novel penalized semiparametric estimation approach, we get smooth distributions without parametric assumptions, and as we saw in our simulations, our penalization procedure works well for both equi- and overdispersed distributions. By contrast, if using an unpenalized non-parametric estimation approach such as the empirical distribution function (EDF), Syntetos and Boylan (2021) criticize that demand values not observed in the past are automatically assigned zero probabilities for the future. Furthermore, they state that an EDF provides a perfect fit to the historical data, but it does not ensure the goodness of fit to the demand over the forecast horizon, especially with respect to higher percentiles. Again, these drawbacks are omitted with our penalized estimation approach. Finally, historical demand time series are often rather short, see the demand count time series provided by Snyder (2002) as

an example, such that smoothing approaches would be particularly welcome. For these reason, the forecasting of intermittent demand appears to be a promising application area for our proposed penalized semiparametric estimation procedure.

Therefore, we consider time series ($n = 51$) of the monthly demand of different car spare parts offered by an Australian subsidiary of a Japanese car company from January 1998 to March 2002 (Snyder 2002). Figure 10 contains an exemplary time series of car part 2404. The observations vary between 0 and 5 and the up and down movements indicate a moderate autocorrelation level. After inspecting the corresponding (P)ACF also included in Fig. 10, we conclude that an AR(1)-like model might be appropriate for describing the serial dependence of the time series. Moreover, L_2 penalization with first order differences leads to an estimated innovation distribution without any unnatural gaps, i.e. zero values, in the PMF.

Now consider the 1-step median prediction and the 90% quantile of the 1-step prediction of the demand for car spare part 2404. The latter serves here as a worst-case scenario for spare parts requirements. Therefore, we determine the median and the 90% quantile of the predictive distribution $P(\ldots \mid y)$, where $y \in 0, \ldots, 10$. Based on the results of the simulation study in Subsect. 3.6, we use the penalized estimated innovation distribution and the unpenalized estimated INAR coefficient to determine the conditional predictive distribution. Table 2 shows that the penalized estimation tends to lead to higher predicted values (more conservative prediction). Consequently, without penalizing the innovation distribution, the predictions for the demand for spare parts may be too low, which can lead to a lack of spare parts. Moreover, the penalization of the innovation distribution (especially for such short

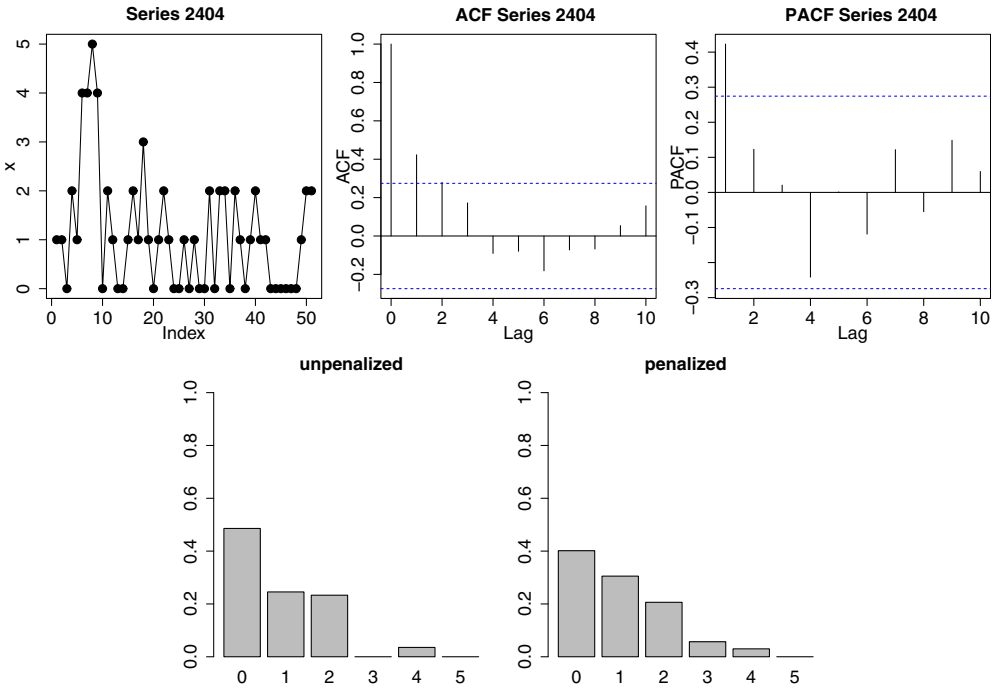

Fig. 10 From left to right and top to bottom: Plot of time series of monthly demand for car spare part 2404, its corresponding ACF and PACF and the unpenalized and the penalized estimated innovation distribution

Table 2 Unpenalized and penalized 1-step median prediction and 90% quantile of the 1-step prediction of the demand for car spare part 2404 when observing demand y

y	0	1	2	3	4	5	6	7	8	9	10
Median (unpenalized)	1	1	1	1	2	2	2	3	3	3	3
Median (penalized)	1	1	1	2	2	2	2	3	3	3	3
90% quantile (unpenalized)	2	2	3	3	4	4	4	5	5	5	6
90% quantile (penalized)	2	3	3	4	4	4	5	5	5	6	6

time series) can serve as a robustness analysis to identify possible uncertainties in the forecast at an early stage.

In addition, we consider car spare part 1971. Figure 11 again suggests an AR(1)-like model and a moderate autocorrelation level. The observations vary between 0 and 4 and there may be zero inflation in this time series. Therefore, in addition to the unpenalized and penalized estimates, we also consider the penalized estimate of the innovation distribution, where $G(0)$ is not smoothed (see Subsect. 3.5). It becomes clear that this last estimation procedure yields more plausible results than when $G(0)$ is smoothed. Again, the penalized estimation procedure yields a slightly smoother innovation distribution than the unpenalized estimation. In summary, if there is a reasonable suspicion of zero inflation, $G(0)$ should not be smoothed.

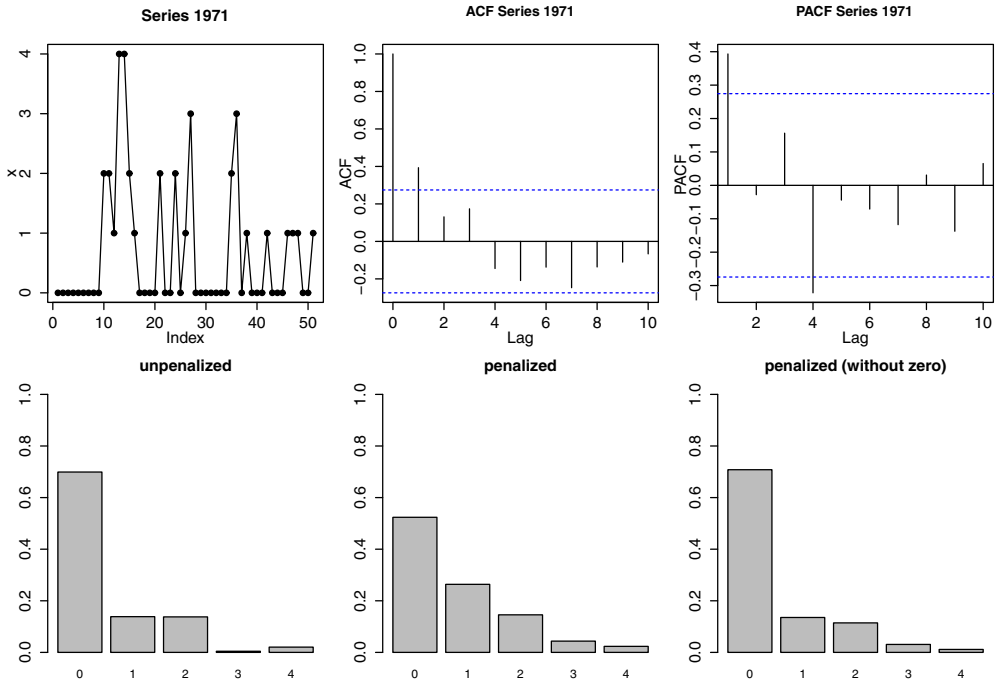

Fig. 11 Plot of time series of monthly demand for car spare part 1971, its corresponding ACF and PACF, the unpenalized and the penalized estimated innovation distribution and the penalized estimated innovation distribution excluding the first entry of the PMF (from left to right and from top to bottom)

5 Conclusion

Although semiparametric estimation yields a decent fit in INAR models, its performance is often not convincing for small sample sizes. Therefore, we proposed a penalization approach that exploits a qualitative smoothness assumption fulfilled by commonly used innovation distributions. A simulation study showed that our penalization approach provides a large benefit in estimating the innovation distribution, especially for small sample sizes. Additionally, we showed that the combination of unpenalized estimation of INAR coefficients and penalized estimation of the innovation distribution provided the best performance. Future research should investigate whether additional penalization of the INAR coefficients may result in further benefit. Furthermore, as the penalization approach proved to be beneficial for forecasting, one may also think of an application in statistical process control, i.e. for the design of control charts relying on a fitted INAR(1) model. Another interesting issue for future research is the application of our proposed method on integer-valued autoregressive models on $\mathbb{Z}$, such as those proposed by Kim and Park (2008) or Liu et al. (2021).

Appendix

See Figures 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22 and 23.

See Tables 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12.

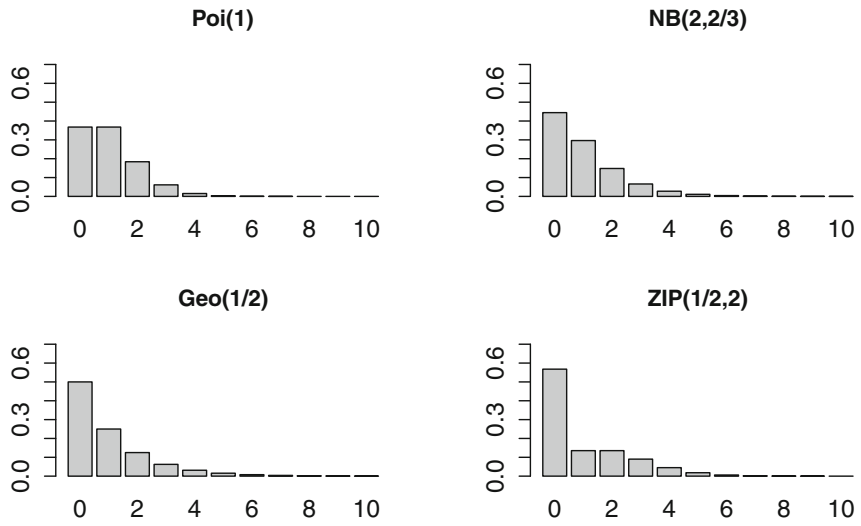

Fig. 12 Probability density functions of the Poi(1), $\mathrm{NB}\left(2,\frac{2}{3}\right)$, $\mathrm{Geo}\left(\frac{1}{2}\right)$ and $\mathrm{ZIP}\left(\frac{1}{2},2\right)$ distributions (truncated at value 10 for clarity)

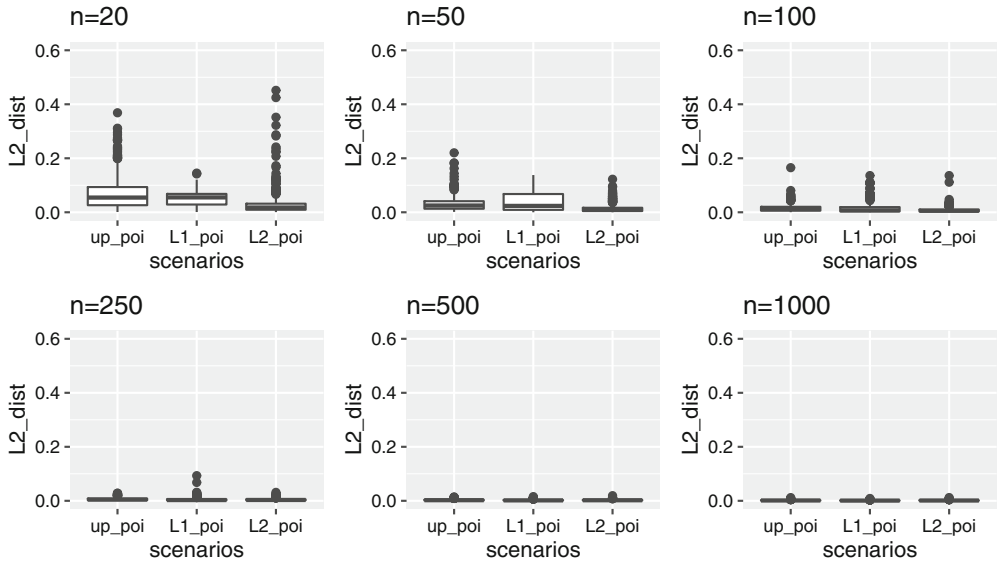

Fig. 13 Boxplots of the L_2 distances of the estimated innovation distribution to the true Poi(1) innovation distribution of an INAR(1) process for the different sample sizes n and $\alpha = 0.2$. We report results for unpenalized (up), L_1 and L_2 penalized estimation

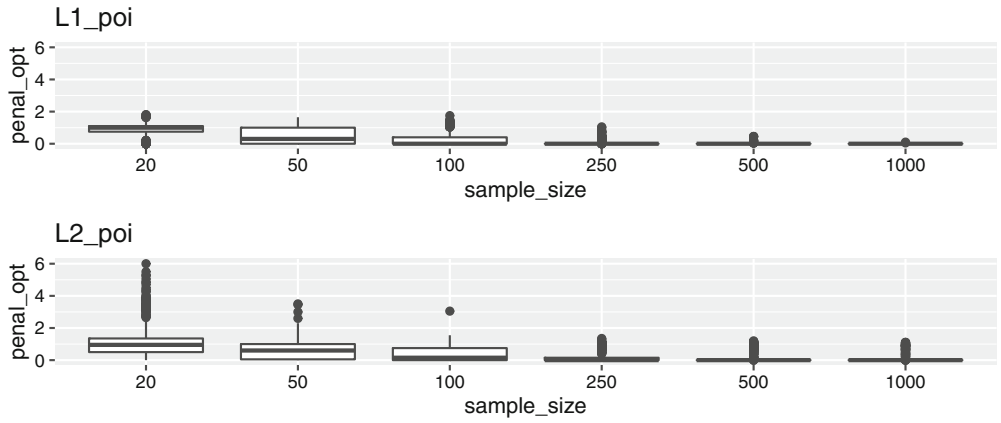

Fig. 14 Boxplots of the penalization parameter η selected by L_1 penalization (upper panel) and L_2 penalization (lower panel) for the different sample sizes n in the case of a true Poi(1) innovation distribution of an INAR(1) process and $\alpha = 0.2$

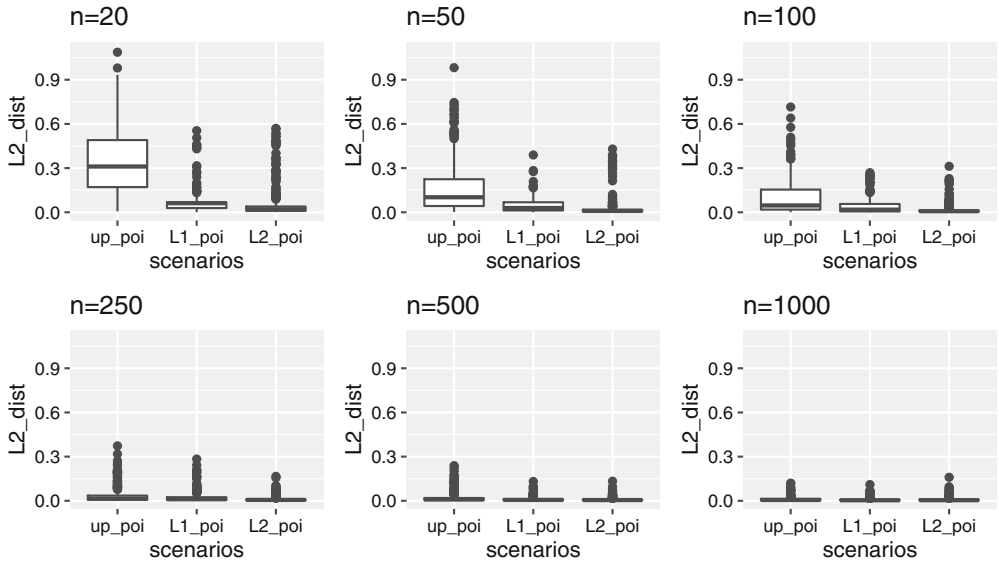

Fig. 15 Boxplots of the L_2 distances of the estimated innovation distribution to the true Poi(1) innovation distribution of an INAR(1) process for the different sample sizes n and $\alpha = 0.8$. We report results for unpenalized (up), L_1 and L_2 penalized estimation

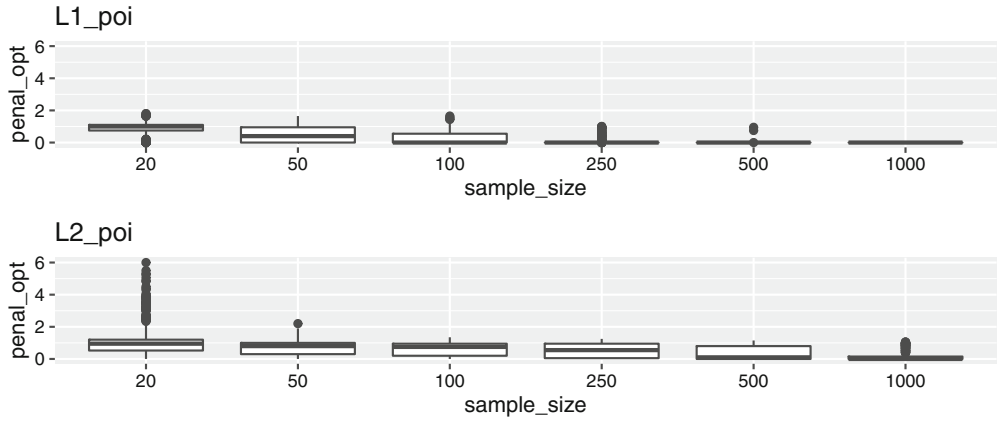

Fig. 16 Boxplots of the penalization parameter η selected by L_1 penalization (upper panel) and L_2 penalization (lower panel) for the different sample sizes n in the case of a true Poi(1) innovation distribution of an INAR(1) process and $\alpha = 0.8$

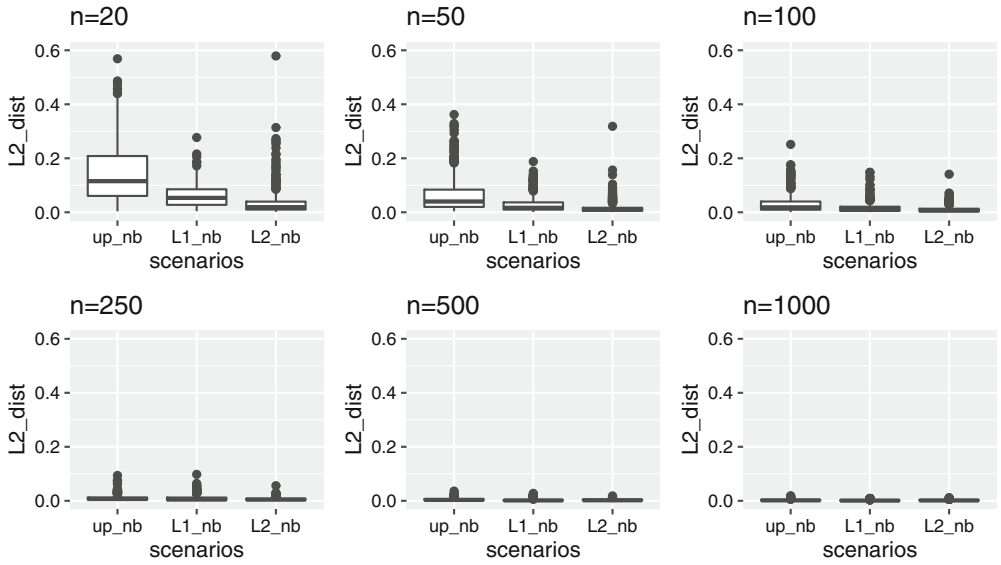

Fig. 17 Boxplots of the L_2 distances of the estimated innovation distribution to the true $\mathrm{NB}\left(2, \frac{2}{3}\right)$ innovation distribution of an INAR(1) process for the different sample sizes n. We report results for unpenalized (up), L_1 and L_2 penalized estimation

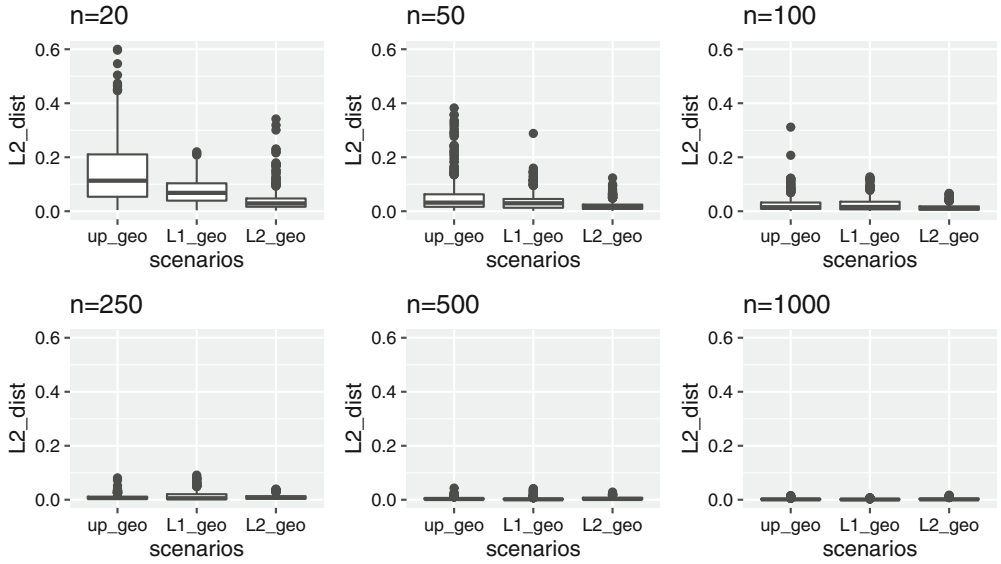

Fig. 18 Boxplots of the L_2 distances of the estimated innovation distribution to the true $\mathrm{Geo}\left(\frac{1}{2}\right)$ innovation distribution of an INAR(1) process for the different sample sizes n. We report results for unpenalized (up), L_1 and L_2 penalized estimation

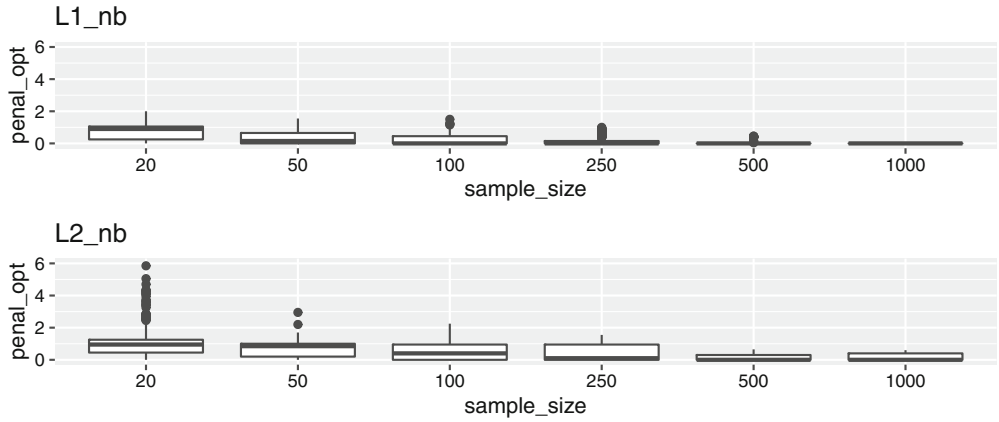

Fig. 19 Boxplots of the penalization parameter η selected by L_1 penalization (upper panel) and L_2 penalization (lower panel) for the different sample sizes n in the case of a true $\mathrm{NB}\left(2, \frac{2}{3}\right)$ innovation distribution of an INAR(1) process

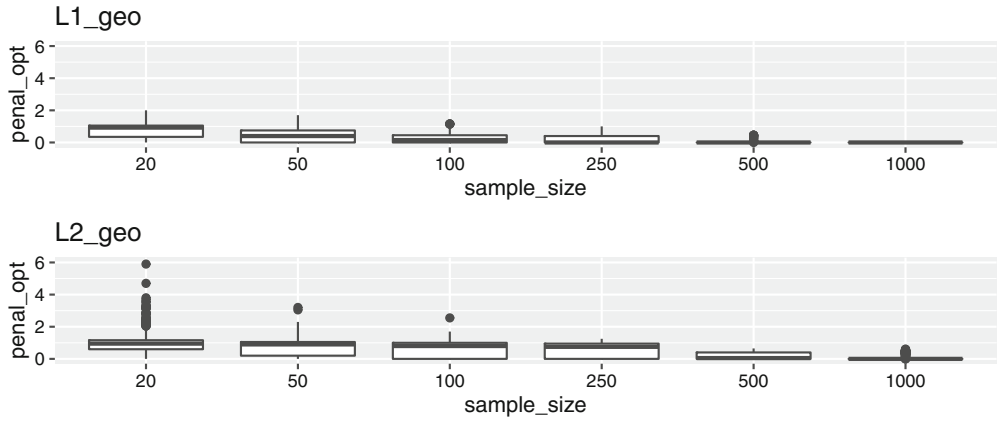

Fig. 20 Boxplots of the penalization parameter η selected by L_1 penalization (upper panel) and L_2 penalization (lower panel) for the different sample sizes n in the case of a true $\mathrm{Geo}\left(\frac{1}{2}\right)$ innovation distribution of an INAR(1) process

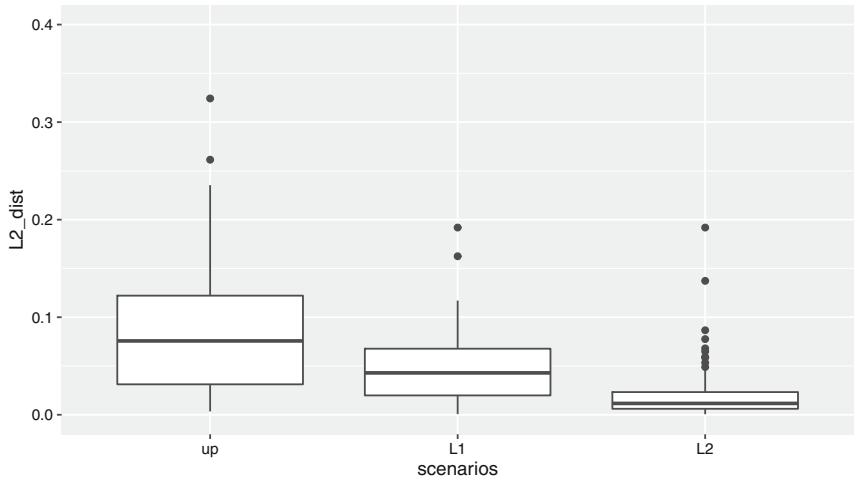

Fig. 21 Boxplots of the L_2 distances of the estimated innovations distribution to the true Poi(1) innovation distribution of an INAR(2) process and $n = 50$. We report results for unpenalized (up), L_1 and L_2 penalized estimation

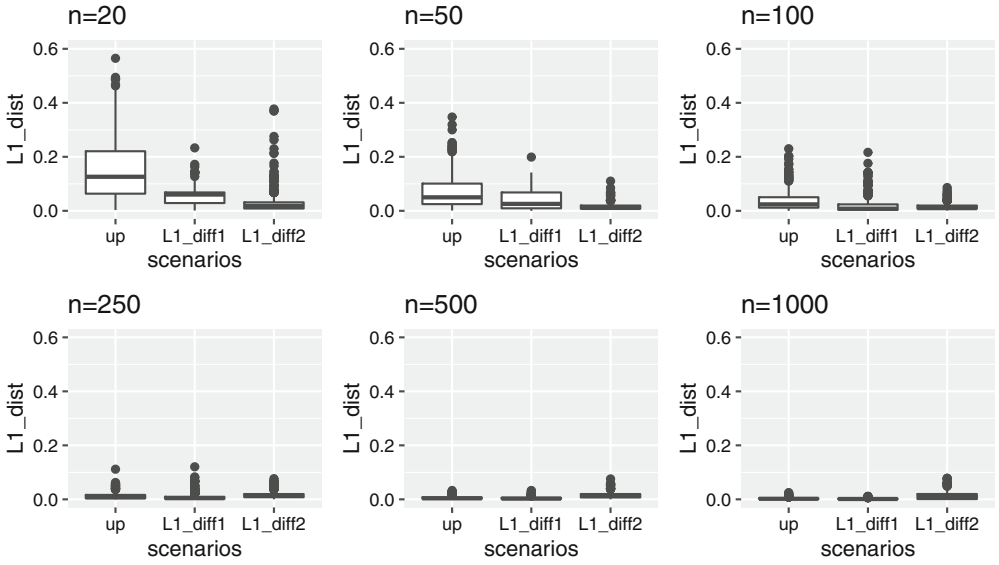

Fig. 22 Boxplots of the L_2 distances of the estimated innovation distribution to the true Poi(1) innovation distribution of an INAR(1) process for the different sample sizes n. We report results for unpenalized (up) and L_1 penalized estimation using either first order (diff1) or second order (diff2) differences

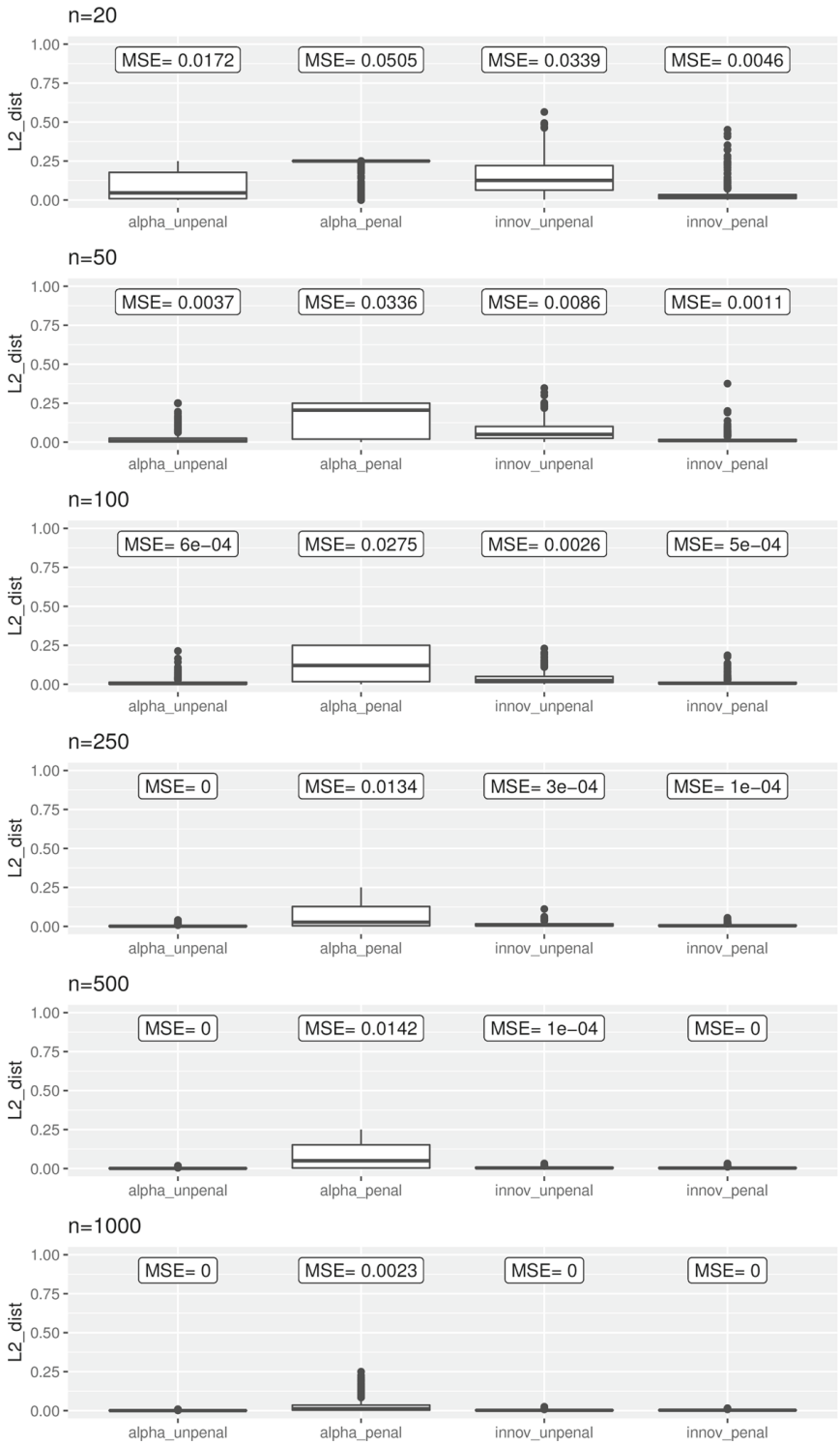

Fig. 23 Boxplots of the L_2 distances between (1) unpenalized estimated INAR coefficient α and the true INAR coefficient α, (2) L_2 penalized and true α, (3,4) their related estimated innovation distribution and the true Poi(1) innovation distribution of an INAR(1) process for different sample sizes n and their corresponding MSE (rounded to four digits)

Table 3 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true Poi(1) innovation distribution of an INAR(1) process. We report results for unpenalized (up), L_1 and L_2 penalized estimation

n		g0_up	g0_L1	g0_L2	g1_up	g1_L1	g1_L2	g2_up	g2_L1	g2_L2	g3_up	g3_L1	g3_L2	g4_up	g4_L1	g4_L2
20	Variance	0.0458	0.0069	0.0126	0.0382	0.0054	0.0066	0.0321	0.0059	0.0063	0.0156	0.0078	0.0053	0.0048	0.0042	0.0029
	Bias	−0.0921	−0.0700	−0.0114	−0.0217	−0.0751	−0.0371	0.0513	0.0456	0.0220	0.0376	0.1064	0.0330	0.0250	0.0890	0.0400
	MSE	0.0543	0.0118	0.0127	0.0387	0.0110	0.0080	0.0348	0.0080	0.0068	0.0170	0.0191	0.0064	0.0054	0.0122	0.0045
50	Variance	0.0246	0.0077	0.0060	0.0198	0.0068	0.0037	0.0155	0.0045	0.0034	0.0073	0.0079	0.0025	0.0012	0.0028	0.0006
	Bias	−0.0312	−0.0414	0.0003	−0.0017	−0.0468	−0.0251	0.0102	0.0149	0.0116	0.0165	0.0550	0.0118	0.0037	0.0307	0.0063
	MSE	0.0256	0.0094	0.0060	0.0198	0.0089	0.0043	0.0156	0.0047	0.0036	0.0076	0.0110	0.0026	0.0012	0.0037	0.0006
100	Variance	0.0125	0.0052	0.0035	0.0114	0.0052	0.0031	0.0080	0.0031	0.0027	0.0031	0.0042	0.0014	0.0006	0.0010	0.0003
	Bias	−0.0225	−0.0206	−0.0030	0.0032	−0.0276	−0.0216	0.0103	0.0120	0.0105	0.0067	0.0234	0.0136	0.0019	0.0096	0.0006
	MSE	0.0130	0.0056	0.0035	0.0114	0.0059	0.0035	0.0081	0.0032	0.0028	0.0032	0.0047	0.0016	0.0006	0.0011	0.0003
250	Variance	0.0043	0.0027	0.0019	0.0039	0.0023	0.0020	0.0025	0.0020	0.0012	0.0010	0.0012	0.0007	0.0002	0.0002	0.0002
	Bias	−0.0075	−0.0113	−0.0053	0.0006	−0.0086	−0.0164	0.0054	0.0106	0.0120	0.0026	0.0063	0.0084	−0.0014	0.0013	0.0011
	MSE	0.0044	0.0029	0.0019	0.0039	0.0024	0.0022	0.0025	0.0021	0.0014	0.0010	0.0012	0.0008	0.0002	0.0002	0.0002
500	Variance	0.0021	0.0015	0.0010	0.0019	0.0012	0.0013	0.0011	0.0009	0.0006	0.0004	0.0004	0.0004	0.0001	0.0001	0.0001
	Bias	−0.0044	−0.0029	−0.0050	0.0030	0.0007	−0.0142	0.0004	0.0017	0.0096	0.0009	0.0005	0.0081	0.0001	−0.0000	0.0011
	MSE	0.0022	0.0015	0.0010	0.0019	0.0012	0.0015	0.0011	0.0009	0.0007	0.0004	0.0004	0.0005	0.0001	0.0001	0.0001
1000	Variance	0.0012	0.0007	0.0006	0.0011	0.0006	0.0010	0.0006	0.0004	0.0004	0.0002	0.0002	0.0003	0.0000	0.0000	0.0000
	Bias	−0.0034	−0.0018	−0.0044	0.0014	0.0002	−0.0165	0.0007	0.0013	0.0108	0.0013	0.0002	0.0088	0.0002	0.0002	0.0015
	MSE	0.0012	0.0007	0.0006	0.0011	0.0006	0.0013	0.0006	0.0004	0.0005	0.0002	0.0002	0.0004	0.0000	0.0000	0.0000

Table 4 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true Poi(1) innovation distribution of an INAR(1) process and $\alpha = 0.2$. We report results for unpenalized (up), L_1 and L_2 penalized estimation

n		g0_up	g0_L1	g0_L2	g1_up	g1_L1	g1_L2	g2_up	g2_L1	g2_L2	g3_up	g3_L1	g3_L2	g4_up	g4_L1	g4_L2
20	Variance	0.0244	0.0063	0.0103	0.0225	0.0047	0.0066	0.0169	0.0059	0.0061	0.0054	0.0079	0.0045	0.0018	0.0041	0.0024
	Bias	−0.0126	−0.0662	−0.0073	−0.0064	−0.0737	−0.0317	0.0132	0.0473	0.0208	0.0060	0.1064	0.0284	0.0223	0.0851	0.0419
	MSE	0.0245	0.0107	0.0103	0.0225	0.0102	0.0076	0.0170	0.0081	0.0065	0.0055	0.0192	0.0053	0.0023	0.0114	0.0041
50	Variance	0.0118	0.0072	0.0046	0.0091	0.0060	0.0033	0.0075	0.0044	0.0029	0.0029	0.0078	0.0019	0.0006	0.0022	0.0005
	Bias	−0.0082	−0.0355	0.0027	−0.0021	−0.0457	−0.0210	0.0043	0.0160	0.0081	0.0012	0.0516	0.0084	0.0067	0.0263	0.0093
	MSE	0.0119	0.0085	0.0046	0.0091	0.0081	0.0038	0.0075	0.0047	0.0030	0.0029	0.0105	0.0020	0.0007	0.0029	0.0006
100	Variance	0.0058	0.0042	0.0027	0.0040	0.0039	0.0023	0.0034	0.0025	0.0017	0.0014	0.0043	0.0009	0.0004	0.0010	0.0003
	Bias	−0.0074	−0.0201	−0.0001	−0.0039	−0.0236	−0.0164	0.0054	0.0131	0.0075	0.0036	0.0219	0.0074	0.0013	0.0097	0.0027
	MSE	0.0059	0.0046	0.0027	0.0041	0.0045	0.0026	0.0035	0.0027	0.0018	0.0015	0.0048	0.0010	0.0004	0.0011	0.0003
250	Variance	0.0024	0.0016	0.0013	0.0016	0.0013	0.0013	0.0014	0.0012	0.0009	0.0005	0.0005	0.0004	0.0001	0.0001	0.0001
	Bias	−0.0057	−0.0037	−0.0022	−0.0009	−0.0060	−0.0079	0.0041	0.0072	0.0069	0.0014	0.0009	0.0025	0.0007	0.0012	0.0008
	MSE	0.0024	0.0016	0.0013	0.0016	0.0013	0.0014	0.0014	0.0013	0.0009	0.0005	0.0005	0.0004	0.0001	0.0001	0.0001
500	Variance	0.0011	0.0008	0.0007	0.0009	0.0006	0.0008	0.0006	0.0005	0.0005	0.0003	0.0002	0.0003	0.0001	0.0000	0.0001
	Bias	−0.0014	0.0005	−0.0012	0.0009	−0.0003	−0.0080	0.0005	0.0013	0.0048	0.0001	−0.0013	0.0038	−0.0003	−0.0002	0.0004
	MSE	0.0011	0.0008	0.0007	0.0009	0.0006	0.0009	0.0006	0.0006	0.0005	0.0003	0.0002	0.0003	0.0001	0.0000	0.0001
1000	Variance	0.0005	0.0004	0.0004	0.0005	0.0003	0.0005	0.0004	0.0003	0.0003	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000
	Bias	−0.0020	0.0005	−0.0009	0.0006	0.0006	−0.0056	0.0013	−0.0006	0.0029	0.0001	−0.0003	0.0029	0.0002	0.0001	0.0009
	MSE	0.0006	0.0004	0.0004	0.0005	0.0003	0.0005	0.0004	0.0003	0.0003	0.0001	0.0001	0.0002	0.0000	0.0000	0.0000

Table 5 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true Poi(1) innovation distribution of an INAR(1) process and $\alpha = 0.8$. We report results for unpenalized (up), L_1 and L_2 penalized estimation

n		g0_up	g0_L1	g0_L2	g1_up	g1_L1	g1_L2	g2_up	g2_L1	g2_L2	g3_up	g3_L1	g3_L2	g4_up	g4_L1	g4_L2
20	Variance	0.0575	0.0084	0.0162	0.0639	0.0063	0.0096	0.0491	0.0062	0.0071	0.0393	0.0080	0.0065	0.0237	0.0050	0.0049
	Bias	−0.2006	−0.0784	−0.0249	−0.1267	−0.0837	−0.0475	0.0533	0.0438	0.0172	0.1200	0.1056	0.0385	0.0842	0.0848	0.0442
	MSE	0.0977	0.0145	0.0168	0.0800	0.0133	0.0119	0.0519	0.0081	0.0074	0.0537	0.0191	0.0080	0.0308	0.0122	0.0068
50	Variance	0.0462	0.0096	0.0086	0.0377	0.0084	0.0042	0.0312	0.0052	0.0024	0.0174	0.0077	0.0036	0.0058	0.0031	0.0021
	Bias	−0.1184	−0.0450	−0.0112	−0.0111	−0.0533	−0.0424	0.0519	0.0202	0.0183	0.0559	0.0556	0.0235	0.0155	0.0261	0.0113
	MSE	0.0602	0.0117	0.0087	0.0378	0.0113	0.0060	0.0339	0.0056	0.0028	0.0205	0.0108	0.0041	0.0061	0.0038	0.0023
100	Variance	0.0351	0.0096	0.0044	0.0230	0.0086	0.0032	0.0194	0.0052	0.0024	0.0074	0.0054	0.0016	0.0013	0.0016	0.0004
	Bias	−0.0793	−0.0396	−0.0085	0.0020	−0.0324	−0.0374	0.0478	0.0223	0.0222	0.0237	0.0351	0.0230	0.0052	0.0119	0.0013
	MSE	0.0414	0.0112	0.0045	0.0230	0.0097	0.0046	0.0216	0.0057	0.0029	0.0080	0.0066	0.0021	0.0013	0.0017	0.0004
250	Variance	0.0150	0.0087	0.0031	0.0103	0.0066	0.0027	0.0045	0.0033	0.0015	0.0017	0.0019	0.0008	0.0002	0.0005	0.0002
	Bias	−0.0361	−0.0207	−0.0051	0.0217	−0.0053	−0.0323	0.0078	0.0147	0.0194	0.0059	0.0076	0.0165	0.0008	0.0032	0.0021
	MSE	0.0163	0.0091	0.0031	0.0108	0.0066	0.0037	0.0046	0.0035	0.0019	0.0017	0.0020	0.0011	0.0002	0.0005	0.0002
500	Variance	0.0087	0.0041	0.0027	0.0058	0.0034	0.0028	0.0023	0.0017	0.0010	0.0006	0.0006	0.0006	0.0001	0.0001	0.0001
	Bias	−0.0136	−0.0018	−0.0041	0.0083	0.0013	−0.0203	0.0049	0.0000	0.0125	−0.0003	0.0006	0.0107	0.0005	−0.0003	0.0015
	MSE	0.0089	0.0041	0.0027	0.0059	0.0034	0.0032	0.0024	0.0017	0.0012	0.0006	0.0006	0.0007	0.0001	0.0001	0.0001
1000	Variance	0.0053	0.0029	0.0034	0.0035	0.0022	0.0029	0.0009	0.0008	0.0009	0.0003	0.0003	0.0004	0.0001	0.0001	0.0001
	Bias	0.0013	−0.0031	−0.0074	−0.0003	0.0029	−0.0061	0.0002	−0.0001	0.0084	−0.0009	0.0001	0.0045	−0.0004	0.0002	0.0008
	MSE	0.0053	0.0029	0.0034	0.0035	0.0022	0.0030	0.0009	0.0008	0.0009	0.0003	0.0003	0.0004	0.0001	0.0001	0.0001

Table 6 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true NB$(2, \frac{2}{3})$ innovation distribution of an INAR(1) process. We report results for unpenalized (up), L_1 and L_2 penalized estimation

n		g0_up	g0_L1	g0_L2	g1_up	g1_L1	g1_L2	g2_up	g2_L1	g2_L2	g3_up	g3_L1	g3_L2	g4_up	g4_L1	g4_L2
20	Variance	0.0537	0.0153	0.0137	0.0338	0.0062	0.0054	0.0265	0.0068	0.0053	0.0135	0.0063	0.0041	0.0053	0.0037	0.0022
	Bias	−0.0720	−0.1058	−0.0340	−0.0186	−0.0201	−0.0048	0.0366	0.0377	0.0184	0.0277	0.0567	0.0152	0.0178	0.0580	0.0201
	MSE	0.0588	0.0265	0.0148	0.0341	0.0066	0.0054	0.0279	0.0083	0.0056	0.0143	0.0095	0.0044	0.0056	0.0070	0.0026
50	Variance	0.0236	0.0115	0.0057	0.0184	0.0043	0.0025	0.0115	0.0037	0.0024	0.0058	0.0035	0.0017	0.0019	0.0019	0.0009
	Bias	−0.0412	−0.0554	−0.0363	0.0129	−0.0069	0.0004	0.0092	0.0162	0.0168	0.0112	0.0222	0.0119	0.0038	0.0167	0.0073
	MSE	0.0253	0.0145	0.0070	0.0185	0.0044	0.0025	0.0116	0.0040	0.0026	0.0060	0.0040	0.0018	0.0019	0.0021	0.0009
100	Variance	0.0117	0.0068	0.0043	0.0093	0.0026	0.0016	0.0054	0.0026	0.0015	0.0025	0.0014	0.0009	0.0010	0.0005	0.0004
	Bias	−0.0279	−0.0408	−0.0343	0.0164	0.0053	0.0025	0.0050	0.0166	0.0183	0.0016	0.0087	0.0088	0.0037	0.0050	0.0034
	MSE	0.0125	0.0085	0.0054	0.0096	0.0026	0.0016	0.0054	0.0029	0.0019	0.0025	0.0015	0.0010	0.0010	0.0005	0.0004
250	Variance	0.0037	0.0038	0.0023	0.0033	0.0014	0.0009	0.0016	0.0014	0.0009	0.0008	0.0006	0.0005	0.0004	0.0003	0.0002
	Bias	−0.0069	−0.0290	−0.0308	0.0040	0.0081	0.0030	0.0011	0.0113	0.0160	0.0009	0.0054	0.0084	0.0006	0.0025	0.0030
	MSE	0.0038	0.0046	0.0032	0.0033	0.0015	0.0009	0.0016	0.0015	0.0011	0.0008	0.0006	0.0006	0.0004	0.0003	0.0002
500	Variance	0.0017	0.0013	0.0011	0.0016	0.0007	0.0007	0.0008	0.0004	0.0005	0.0004	0.0001	0.0001	0.0002	0.0001	0.0001
	Bias	−0.0029	−0.0077	−0.0167	0.0011	0.0046	0.0031	0.0007	−0.0027	0.0060	0.0006	0.0045	0.0068	0.0005	−0.0001	−0.0011
	MSE	0.0017	0.0013	0.0014	0.0016	0.0007	0.0007	0.0008	0.0004	0.0005	0.0004	0.0002	0.0002	0.0002	0.0001	0.0001
1000	Variance	0.0008	0.0004	0.0009	0.0007	0.0005	0.0003	0.0004	0.0002	0.0003	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001
	Bias	−0.0003	0.0031	−0.0170	0.0003	−0.0026	0.0026	−0.0008	−0.0002	0.0080	0.0002	−0.0002	0.0050	0.0006	−0.0015	−0.0007
	MSE	0.0008	0.0005	0.0012	0.0007	0.0005	0.0003	0.0004	0.0002	0.0004	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001

Table 7 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true Geo($\frac{1}{2}$) innovation distribution of an INAR(1) process. We report results for unpenalized (up), L_1 and L_2 penalized estimation

n		g0_up	g0_L1	g0_L2	g1_up	g1_L1	g1_L2	g2_up	g2_L1	g2_L2	g3_up	g3_L1	g3_L2	g4_up	g4_L1	g4_L2
20	Variance	0.0551	0.0189	0.0137	0.0364	0.0060	0.0050	0.0217	0.0074	0.0055	0.0118	0.0054	0.0033	0.0056	0.0030	0.0022
	Bias	−0.1071	−0.1490	−0.0743	0.0251	0.0128	0.0318	0.0279	0.0429	0.0230	0.0308	0.0553	0.0159	0.0181	0.0497	0.0205
	MSE	0.0666	0.0411	0.0192	0.0371	0.0062	0.0060	0.0225	0.0093	0.0060	0.0128	0.0084	0.0036	0.0060	0.0055	0.0026
50	Variance	0.0214	0.0143	0.0072	0.0140	0.0044	0.0025	0.0095	0.0037	0.0023	0.0046	0.0022	0.0014	0.0020	0.0013	0.0009
	Bias	−0.0296	−0.0868	−0.0618	0.0129	0.0188	0.0246	0.0062	0.0241	0.0219	0.0041	0.0152	0.0084	0.0044	0.0162	0.0064
	MSE	0.0223	0.0218	0.0110	0.0142	0.0048	0.0031	0.0096	0.0043	0.0028	0.0046	0.0024	0.0015	0.0020	0.0015	0.0009
100	Variance	0.0090	0.0094	0.0046	0.0078	0.0032	0.0016	0.0037	0.0023	0.0013	0.0019	0.0010	0.0007	0.0009	0.0005	0.0004
	Bias	−0.0162	−0.0737	−0.0561	0.0157	0.0263	0.0225	−0.0040	0.0214	0.0196	0.0023	0.0095	0.0080	0.0004	0.0071	0.0034
	MSE	0.0093	0.0148	0.0077	0.0080	0.0039	0.0021	0.0037	0.0027	0.0017	0.0019	0.0011	0.0007	0.0009	0.0005	0.0005
250	Variance	0.0030	0.0065	0.0034	0.0025	0.0020	0.0009	0.0014	0.0013	0.0006	0.0007	0.0005	0.0003	0.0004	0.0002	0.0002
	Bias	−0.0045	−0.0432	−0.0531	0.0060	0.0210	0.0215	−0.0006	0.0114	0.0199	−0.0024	0.0037	0.0062	0.0020	0.0035	0.0034
	MSE	0.0030	0.0084	0.0062	0.0026	0.0025	0.0014	0.0014	0.0014	0.0010	0.0007	0.0005	0.0004	0.0004	0.0002	0.0002
500	Variance	0.0014	0.0015	0.0020	0.0013	0.0009	0.0007	0.0007	0.0005	0.0004	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001
	Bias	−0.0022	−0.0044	−0.0316	0.0043	0.0036	0.0164	−0.0015	0.0004	0.0097	−0.0007	−0.0007	0.0023	0.0002	0.0009	0.0019
	MSE	0.0014	0.0015	0.0030	0.0013	0.0009	0.0010	0.0007	0.0005	0.0005	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001
1000	Variance	0.0007	0.0005	0.0013	0.0007	0.0004	0.0005	0.0003	0.0002	0.0003	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001
	Bias	−0.0007	0.0001	−0.0171	0.0018	0.0008	0.0094	−0.0013	−0.0017	0.0039	0.0000	0.0001	0.0018	0.0003	0.0007	0.0012
	MSE	0.0007	0.0005	0.0016	0.0007	0.0004	0.0006	0.0003	0.0002	0.0003	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001

Table 8 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true Poi(1) innovation distribution of an INAR(1) process. We report results for unpenalized (up), L_2 penalized estimation using either the iterated Algorithm 1 (A1) or the alternative Algorithm 2 (A2)

n		g0_up	g0_A1	g0_A2	g1_up	g1_A1	g1_A2	g2_up	g2_A1	g2_A2	g3_up	g3_A1	g3_A2	g4_up	g4_A1	g4_A2
20	Variance	0.0458	0.0126	0.0063	0.0382	0.0066	0.0041	0.0321	0.0063	0.0039	0.0156	0.0053	0.0037	0.0048	0.0029	0.0017
	Bias	−0.0921	−0.0114	−0.0127	−0.0217	−0.0371	−0.0433	0.0513	0.0220	0.0260	0.0376	0.0330	0.0428	0.0250	0.0400	0.0414
	MSE	0.0543	0.0127	0.0065	0.0387	0.0080	0.0060	0.0348	0.0068	0.0046	0.0170	0.0064	0.0055	0.0054	0.0045	0.0034
50	Variance	0.0246	0.0060	0.0043	0.0198	0.0037	0.0029	0.0155	0.0034	0.0020	0.0073	0.0025	0.0019	0.0012	0.0006	0.0007
	Bias	−0.0312	0.0003	−0.0007	−0.0017	−0.0251	−0.0345	0.0102	0.0116	0.0134	0.0165	0.0118	0.0172	0.0037	0.0063	0.0148
	MSE	0.0256	0.0060	0.0043	0.0198	0.0043	0.0041	0.0156	0.0036	0.0022	0.0076	0.0026	0.0022	0.0012	0.0006	0.0009
100	Variance	0.0125	0.0035	0.0030	0.0114	0.0031	0.0030	0.0080	0.0027	0.0020	0.0031	0.0014	0.0011	0.0006	0.0003	0.0003
	Bias	−0.0225	−0.0030	0.0044	0.0032	−0.0216	−0.0159	0.0103	0.0105	0.0056	0.0067	0.0136	0.0044	0.0019	0.0006	0.0033
	MSE	0.0130	0.0035	0.0030	0.0114	0.0035	0.0032	0.0081	0.0028	0.0020	0.0032	0.0016	0.0011	0.0006	0.0003	0.0003
250	Variance	0.0043	0.0019	0.0014	0.0039	0.0020	0.0013	0.0025	0.0012	0.0010	0.0010	0.0007	0.0004	0.0002	0.0002	0.0001
	Bias	−0.0075	−0.0053	0.0032	0.0006	−0.0164	−0.0013	0.0054	0.0120	−0.0002	0.0026	0.0084	−0.0012	−0.0014	0.0011	−0.0002
	MSE	0.0044	0.0019	0.0014	0.0039	0.0022	0.0013	0.0025	0.0014	0.0010	0.0010	0.0008	0.0004	0.0002	0.0002	0.0001
500	Variance	0.0021	0.0010	0.0007	0.0019	0.0013	0.0007	0.0011	0.0006	0.0004	0.0004	0.0004	0.0002	0.0001	0.0001	0.0000
	Bias	−0.0044	−0.0050	0.0035	0.0030	−0.0142	0.0006	0.0004	0.0096	−0.0017	0.0009	0.0081	−0.0015	0.0001	0.0011	−0.0006
	MSE	0.0022	0.0010	0.0008	0.0019	0.0015	0.0007	0.0011	0.0007	0.0004	0.0004	0.0005	0.0002	0.0001	0.0001	0.0000
1000	Variance	0.0012	0.0007	0.0006	0.0011	0.0006	0.0010	0.0006	0.0004	0.0004	0.0002	0.0002	0.0003	0.0000	0.0000	0.0000
	Bias	−0.0034	−0.0018	−0.0044	0.0014	0.0002	−0.0165	0.0007	0.0013	0.0108	0.0013	0.0002	0.0088	0.0002	0.0002	0.0015
	MSE	0.0012	0.0007	0.0006	0.0011	0.0006	0.0013	0.0006	0.0004	0.0005	0.0002	0.0002	0.0004	0.0000	0.0000	0.0000

Table 9 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true Poi(1) innovation distribution of an INAR(1) process. We report results for unpenalized (up) and L_1 penalized estimation using either first order (diff1) or second order (diff2) differences

n		g0_up	g0_d1	g0_d2	g1_up	g1_d1	g1_d2	g2_up	g2_d1	g2_d2	g3_up	g3_d1	g3_d2	g4_up	g4_d1	g4_d2
20	Variance	0.0458	0.0069	0.0105	0.0382	0.0054	0.0021	0.0321	0.0059	0.0023	0.0156	0.0078	0.0048	0.0048	0.0042	0.0032
	Bias	−0.0921	−0.0700	0.0154	−0.0217	−0.0751	−0.0760	0.0513	0.0456	0.0112	0.0376	0.1064	0.0467	0.0250	0.0890	0.0408
	MSE	0.0543	0.0118	0.0108	0.0387	0.0110	0.0079	0.0348	0.0080	0.0024	0.0170	0.0191	0.0069	0.0054	0.0122	0.0049
50	Variance	0.0246	0.0077	0.0029	0.0198	0.0068	0.0007	0.0155	0.0045	0.0002	0.0073	0.0079	0.0014	0.0012	0.0028	0.0012
	Bias	−0.0312	−0.0414	0.0074	−0.0017	−0.0468	−0.0813	0.0102	0.0149	0.0120	0.0165	0.0550	0.0447	0.0037	0.0307	0.0234
	MSE	0.0256	0.0094	0.0030	0.0198	0.0089	0.0073	0.0156	0.0047	0.0004	0.0076	0.0110	0.0034	0.0012	0.0037	0.0017
100	Variance	0.0125	0.0052	0.0020	0.0114	0.0052	0.0006	0.0080	0.0031	0.0001	0.0031	0.0042	0.0007	0.0006	0.0010	0.0010
	Bias	−0.0225	−0.0206	0.0008	0.0032	−0.0276	−0.0841	0.0103	0.0120	0.0124	0.0067	0.0234	0.0488	0.0019	0.0096	0.0208
	MSE	0.0130	0.0056	0.0020	0.0114	0.0059	0.0076	0.0081	0.0032	0.0002	0.0032	0.0047	0.0030	0.0006	0.0011	0.0015
250	Variance	0.0043	0.0027	0.0016	0.0039	0.0023	0.0011	0.0025	0.0020	0.0002	0.0010	0.0012	0.0006	0.0002	0.0002	0.0010
	Bias	−0.0075	−0.0113	−0.0117	0.0006	−0.0086	−0.0846	0.0054	0.0106	0.0121	0.0026	0.0063	0.0514	−0.0014	0.0013	0.0265
	MSE	0.0044	0.0029	0.0017	0.0039	0.0024	0.0083	0.0025	0.0021	0.0003	0.0010	0.0012	0.0033	0.0002	0.0002	0.0017
500	Variance	0.0021	0.0015	0.0013	0.0019	0.0012	0.0022	0.0011	0.0009	0.0002	0.0004	0.0004	0.0010	0.0001	0.0001	0.0009
	Bias	−0.0044	−0.0029	−0.0171	0.0030	0.0007	−0.0738	0.0004	0.0017	0.0083	0.0009	0.0005	0.0460	0.0001	−0.0000	0.0298
	MSE	0.0022	0.0015	0.0016	0.0019	0.0012	0.0077	0.0011	0.0009	0.0003	0.0004	0.0004	0.0032	0.0001	0.0001	0.0018
1000	Variance	0.0012	0.0007	0.0012	0.0011	0.0006	0.0031	0.0006	0.0004	0.0002	0.0002	0.0002	0.0012	0.0000	0.0000	0.0009
	Bias	−0.0034	−0.0018	−0.0194	0.0014	0.0002	−0.0586	0.0007	0.0013	0.0052	0.0013	0.0002	0.0369	0.0002	0.0002	0.0270
	MSE	0.0012	0.0007	0.0016	0.0011	0.0006	0.0065	0.0006	0.0004	0.0002	0.0002	0.0002	0.0026	0.0000	0.0000	0.0016

Table 10 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true Poi(1) innovation distribution of an INAR(1) process. We report results for unpenalized (up) and L_2 penalized estimation using either first order (diff1) or second order (diff2) differences

n		g0_up	g0_d1	g0_d2	g1_up	g1_d1	g1_d2	g2_up	g2_d1	g2_d2	g3_up	g3_d1	g3_d2	g4_up	g4_d1	g4_d2
20	Variance	0.0458	0.0126	0.0181	0.0382	0.0066	0.0053	0.0321	0.0063	0.0057	0.0156	0.0053	0.0058	0.0048	0.0029	0.0042
	Bias	−0.0921	−0.0114	0.0238	−0.0217	−0.0371	−0.0573	0.0513	0.0220	0.0089	0.0376	0.0330	0.0212	0.0250	0.0400	0.0294
	MSE	0.0543	0.0127	0.0186	0.0387	0.0080	0.0086	0.0348	0.0068	0.0058	0.0170	0.0064	0.0063	0.0054	0.0045	0.0050
50	Variance	0.0246	0.0060	0.0092	0.0198	0.0037	0.0011	0.0155	0.0034	0.0017	0.0073	0.0025	0.0028	0.0012	0.0006	0.0011
	Bias	−0.0312	0.0003	0.0433	−0.0017	−0.0251	−0.0595	0.0102	0.0116	0.0014	0.0165	0.0118	0.0122	0.0037	0.0063	0.0035
	MSE	0.0256	0.0060	0.0111	0.0198	0.0043	0.0046	0.0156	0.0036	0.0017	0.0076	0.0026	0.0029	0.0012	0.0006	0.0011
100	Variance	0.0125	0.0035	0.0025	0.0114	0.0031	0.0006	0.0080	0.0027	0.0008	0.0031	0.0014	0.0008	0.0006	0.0003	0.0002
	Bias	−0.0225	−0.0030	0.0517	0.0032	−0.0216	−0.0606	0.0103	0.0105	−0.0006	0.0067	0.0136	0.0118	0.0019	0.0006	−0.0017
	MSE	0.0130	0.0035	0.0052	0.0114	0.0035	0.0042	0.0081	0.0028	0.0008	0.0032	0.0016	0.0009	0.0006	0.0003	0.0002
250	variance	0.0043	0.0019	0.0010	0.0039	0.0020	0.0004	0.0025	0.0012	0.0003	0.0010	0.0007	0.0003	0.0002	0.0002	0.0001
	Bias	−0.0075	−0.0053	0.0503	0.0006	−0.0164	−0.0591	0.0054	0.0120	−0.0001	0.0026	0.0084	0.0110	−0.0014	0.0011	−0.0011
	MSE	0.0044	0.0019	0.0035	0.0039	0.0022	0.0038	0.0025	0.0014	0.0003	0.0010	0.0008	0.0004	0.0002	0.0002	0.0001
500	Variance	0.0021	0.0010	0.0008	0.0019	0.0013	0.0005	0.0011	0.0006	0.0002	0.0004	0.0004	0.0002	0.0001	0.0001	0.0001
	Bias	−0.0044	−0.0050	0.0495	0.0030	−0.0142	−0.0577	0.0004	0.0096	−0.0013	0.0009	0.0081	0.0110	0.0001	0.0011	−0.0005
	MSE	0.0022	0.0010	0.0032	0.0019	0.0015	0.0039	0.0011	0.0007	0.0002	0.0004	0.0005	0.0003	0.0001	0.0001	0.0001
1000	Variance	0.0012	0.0006	0.0009	0.0011	0.0010	0.0009	0.0006	0.0004	0.0002	0.0002	0.0003	0.0001	0.0000	0.0000	0.0000
	Bias	−0.0006	0.0010	0.0036	0.0003	−0.0005	0.0001	−0.0001	−0.0016	−0.0060	0.0005	0.0012	0.0012	−0.0001	0.0003	0.0014
	MSE	0.0005	0.0005	0.0005	0.0004	0.0004	0.0002	0.0003	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001

Table 11 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true ZIP($\frac{1}{2}$,2) innovation distribution of an INAR (1) process. We report results for unpenalized (up), L_1 and L_2 penalized estimation

n		g0_up	g0_L1	g0_L2	g1_up	g1_L1	g1_L2	g2_up	g2_L1	g2_L2	g3_up	g3_L1	g3_L2	g4_up	g4_L1	g4_L2
20	Variance	0.0581	0.0287	0.0174	0.0293	0.0060	0.0066	0.0174	0.0038	0.0045	0.0138	0.0042	0.0040	0.0072	0.0035	0.0026
	Bias	−0.1037	−0.1670	−0.1142	0.0424	0.0508	0.0752	0.0153	0.0289	0.0210	0.0211	0.0400	0.0130	0.0165	0.0556	0.0179
	MSE	0.0689	0.0566	0.0305	0.0310	0.0086	0.0123	0.0176	0.0046	0.0050	0.0142	0.0058	0.0042	0.0075	0.0066	0.0029
50	Variance	0.0205	0.0185	0.0104	0.0121	0.0044	0.0045	0.0087	0.0025	0.0022	0.0052	0.0020	0.0016	0.0026	0.0017	0.0010
	Bias	−0.0242	−0.0785	−0.0810	0.0118	0.0321	0.0565	0.0066	0.0098	0.0119	0.0043	0.0128	0.0077	−0.0022	0.0111	0.0025
	MSE	0.0211	0.0247	0.0169	0.0123	0.0054	0.0077	0.0087	0.0026	0.0023	0.0053	0.0022	0.0017	0.0026	0.0018	0.0010
100	Variance	0.0067	0.0119	0.0086	0.0053	0.0027	0.0039	0.0038	0.0016	0.0013	0.0025	0.0013	0.0009	0.0013	0.0007	0.0006
	Bias	−0.0132	−0.0496	−0.0598	0.0101	0.0194	0.0408	−0.0002	0.0090	0.0089	0.0028	0.0092	0.0059	0.0007	0.0045	0.0023
	MSE	0.0069	0.0144	0.0122	0.0054	0.0031	0.0055	0.0038	0.0017	0.0014	0.0025	0.0014	0.0010	0.0013	0.0007	0.0006
250	Variance	0.0023	0.0065	0.0063	0.0017	0.0019	0.0029	0.0013	0.0009	0.0006	0.0009	0.0006	0.0004	0.0005	0.0003	0.0002
	Bias	−0.0032	−0.0274	−0.0528	0.0013	0.0121	0.0353	0.0004	0.0063	0.0082	0.0018	0.0038	0.0045	−0.0002	0.0024	0.0029
	MSE	0.0023	0.0073	0.0091	0.0017	0.0021	0.0042	0.0013	0.0009	0.0007	0.0009	0.0006	0.0004	0.0005	0.0003	0.0002
500	Variance	0.0009	0.0010	0.0035	0.0007	0.0006	0.0019	0.0006	0.0004	0.0003	0.0004	0.0003	0.0002	0.0002	0.0002	0.0001
	Bias	0.0006	0.0010	−0.0298	0.0004	−0.0007	0.0230	−0.0012	−0.0008	0.0020	0.0008	0.0015	0.0030	0.0002	−0.0007	0.0012
	MSE	0.0009	0.0010	0.0044	0.0007	0.0006	0.0025	0.0006	0.0004	0.0003	0.0004	0.0003	0.0002	0.0002	0.0002	0.0001
1000	Variance	0.0005	0.0004	0.0016	0.0004	0.0003	0.0009	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
	Bias	−0.0006	−0.0011	−0.0100	0.0003	0.0007	0.0073	−0.0001	−0.0004	0.0008	0.0005	0.0009	0.0011	−0.0001	0.0001	0.0006
	MSE	0.0005	0.0004	0.0017	0.0004	0.0003	0.0009	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001

Table 12 Variance, bias and MSE of the first five estimated entries of the PMF for the different sample sizes n in case of a true ZIP($\frac{1}{2}$,2) innovation distribution of an INAR (1) process. We report results for unpenalized (up), L_1 and L_2 penalized estimation without smoothing of $G(0)$

n		g0_up	g0_L1	g0_L2	g1_up	g1_L1	g1_L2	g2_up	g2_L1	g2_L2	g3_up	g3_L1	g3_L2	g4_up	g4_L1	g4_L2
20	Variance	0.0581	0.0320	0.0338	0.0293	0.0031	0.0090	0.0174	0.0025	0.0055	0.0138	0.0022	0.0038	0.0072	0.0021	0.0022
	Bias	−0.1037	−0.0027	0.0034	0.0424	−0.0208	−0.0077	0.0153	−0.0217	−0.0021	0.0211	0.0077	0.0048	0.0165	0.0352	0.0100
	MSE	0.0689	0.0320	0.0338	0.0310	0.0036	0.0090	0.0176	0.0030	0.0055	0.0142	0.0022	0.0039	0.0075	0.0033	0.0023
50	Variance	0.0205	0.0183	0.0132	0.0121	0.0021	0.0033	0.0087	0.0017	0.0020	0.0052	0.0014	0.0015	0.0026	0.0016	0.0011
	Bias	−0.0242	0.0178	0.0117	0.0118	−0.0237	−0.0007	0.0066	−0.0282	−0.0101	0.0043	0.0002	−0.0022	−0.0022	0.0150	0.0004
	MSE	0.0211	0.0186	0.0134	0.0123	0.0027	0.0033	0.0087	0.0025	0.0021	0.0053	0.0014	0.0015	0.0026	0.0018	0.0011
100	Variance	0.0067	0.0088	0.0054	0.0053	0.0016	0.0017	0.0038	0.0013	0.0010	0.0025	0.0010	0.0007	0.0013	0.0010	0.0006
	Bias	−0.0132	0.0261	0.0080	0.0101	−0.0217	0.0002	−0.0002	−0.0229	−0.0081	0.0028	0.0024	0.0012	0.0007	0.0073	0.0002
	MSE	0.0069	0.0095	0.0055	0.0054	0.0021	0.0017	0.0038	0.0018	0.0011	0.0025	0.0010	0.0007	0.0013	0.0010	0.0006
250	Variance	0.0023	0.0042	0.0018	0.0017	0.0014	0.0006	0.0013	0.0012	0.0004	0.0009	0.0006	0.0003	0.0005	0.0005	0.0002
	Bias	−0.0032	0.0249	0.0092	0.0013	−0.0148	−0.0011	0.0004	−0.0166	−0.0096	0.0018	0.0030	−0.0001	−0.0002	0.0035	0.0024
	MSE	0.0023	0.0048	0.0019	0.0017	0.0017	0.0006	0.0013	0.0015	0.0005	0.0009	0.0006	0.0003	0.0005	0.0005	0.0002
500	Variance	0.0009	0.0013	0.0009	0.0007	0.0007	0.0004	0.0006	0.0007	0.0003	0.0004	0.0005	0.0002	0.0002	0.0003	0.0002
	Bias	0.0006	0.0090	0.0057	0.0004	−0.0043	−0.0002	−0.0012	−0.0061	−0.0065	0.0008	0.0020	0.0004	0.0002	−0.0001	0.0011
	MSE	0.0009	0.0013	0.0009	0.0007	0.0007	0.0004	0.0006	0.0007	0.0003	0.0004	0.0005	0.0002	0.0002	0.0003	0.0002
1000	Variance	0.0005	0.0005	0.0005	0.0004	0.0004	0.0002	0.0003	0.0003	0.0001	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001
	Bias	−0.0006	0.0010	0.0036	0.0003	−0.0005	0.0001	−0.0001	−0.0016	−0.0060	0.0005	0.0012	0.0012	−0.0001	0.0003	0.0014
	MSE	0.0005	0.0005	0.0005	0.0004	0.0004	0.0002	0.0003	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001

Acknowledgements The authors thank the two referees for their useful comments on an earlier draft of this article. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 437270842.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adam T, Langrock R, Weiß C (2019) Penalized estimation of flexible hidden Markov models for time series of counts. METRON 77:87–104

Al-Osh MA, Alzaid AA (1987) First-order integer-valued autoregressive (INAR(1)) process. J Time Ser Anal 8(3):261–275

Al-Osh MA, Alzaid AA (1990) An integer-valued pth order autoregressive structure (INAR(p)) process. J Appl Probab 27(2):314–324

Brännäs K, Hellström J (2001) Generalized integer-valued autoregression. Economet Rev 20:425–443

Bui MT, Potgieter CJ, Kamata A (2021) Penalized likelihood methods for modeling count data. arXiv preprint arXiv:2109.14010

Core Team R (2021): A language and environment for statistical computing. URL https://www.R-project.org/

Drost F, Van den Akker R, Werker B (2009) Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models. J Royal Stat Soc Ser B 71:467–485

Du JG, Li Y (1991) The integer valued autoregressive (INAR(p)) model. J Time Ser Anal 12(2):129–142

Fahrmeir L, Kneib T, Lang S (2013) Regression: models, methods and applications. Springer

Fokianos K (2010) Penalized estimation for integer autoregressive models. In: Statistical modelling and regression structures. Springer, pp 337–352

Franke J, Seligmann T (1993) Conditional maximum-likelihood estimates for INAR(1) processes and their applications to modelling epileptic seizure counts. Developments in Time Series, pp 310–330

Freeland R, McCabe B (2005) Asymptotic properties of cls estimators in the poisson AR(1) model. Stat Probab Lett 73:147–153

Homburg A, Weiß C, Frahm G et al (2021) Analysis and forecasting of risk in count processes. J Risk Finan Manag 14(4):182

Jazi M, Jones G, Lai C (2012) First-order integer valued AR processes with zero inflated Poisson innovations. J Time Ser Anal 33:954–963

Jung R, Ronning G, Tremayne A (2005) Estimation in conditional first order autoregression with discrete support. Stat Pap 46:195–224

Kim HY, Park Y (2008) A non-stationary integer-valued autoregressive model. Stat Pap 49(3):485–502

Liu Z, Li Q, Zhu F (2021) Semiparametric integer-valued autoregresive models on Z. Can J Stat 49:1317–1337

McKenzie E (1985) Some simple models for discrete variate time series. Water Resour Bull 21(4):645–650
Nardi Y, Rinaldo A (2011) Autoregressive process modeling via the lasso procedure. J Multivar Anal 102:528–549
Scott DW, Tapia RA, Thompson JR (1980) Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. Ann Stat 8(4):820–832
Snyder R (2002) Forecasting sales of slow and fast moving inventories. Eur J Oper Res 140:684–699
Steutel FW, Van Harn K (1979) Discrete analogues of self-decomposability and stability. Ann Probab 7 (5):893–899
Syntetos AA, Boylan JE (2021) Intermittent demand forecasting: Context, methods and applications. John Wiley & Sons, UK
Tibshirani R, Saunders M, Rosset S et al (2005) Sparsity and smoothness via the fused lasso. J Roy Stat Soc B 67(1):91–108
Wang X (2020) Variable selection for first-order poisson integer-valued autoregressive model with covariables. Aust N Z J Stat 62(2):278–295
Wang X, Wang D, Yang K (2021) Integer-valued time series model order shrinkage and selection via penalized quasi-likelihood approach. Metrika 84:713–750
Weiß C (2018) An Introduction to Discrete-Valued Time Series, 1st edn. Wiley, UK
Yang L (2019) The predictive distributions of thinning-based count processes. Scand J Stat 48(1):42–67

Authors and Affiliations

Maxime Faymonville[1] · Carsten Jentsch[1] · Christian H. Weiß[2] · Boris Aleksandrov[2]

✉ Maxime Faymonville
faymonville@statistik.tu-dortmund.de

Carsten Jentsch
jentsch@statistik.tu-dortmund.de

Christian H. Weiß
weissc@hsu-hh.de

Boris Aleksandrov
aleksanb@hsu-hh.de

[1] Department of Statistics, TU Dortmund University, D-44221 Dortmund, Germany

[2] Department of Mathematics and Statistics, Helmut-Schmidt-University, D-22008 Hamburg, Germany

The Journal of Open Source Software

spINAR: An R Package for Semiparametric and Parametric Estimation and Bootstrapping of Integer-Valued Autoregressive (INAR) Models

Maxime Faymonville[1], **Javiera Riffo**[1], **Jonas Rieger**[1], **and Carsten Jentsch**[1]

1 TU Dortmund University

DOI: 10.21105/joss.05386

Software

- Review
- Repository
- Archive

Editor: Claudia Solis-Lemus

Reviewers:

- @ManuelStapper
- @SaranjeetKaur
- @wittenberg

Submitted: 10 February 2023
Published: 08 May 2024

Summary

While the statistical literature on continuous-valued time series processes is vast and the toolbox for parametric, non-parametric and semiparametric approaches is methodologically sound, the literature on count data time series is considerably less developed. Such count data time series models are usually categorized in parameter-driven and observation-driven models. Among the observation-driven approaches, the integer-valued autoregressive (INAR) models that rely on the famous binomial thinning operation due to Steutel & Van Harn (1979) are arguably the most popular ones. They have a simple intuitive and easy interpretable structure and have been widely applied in practice (Weiß, 2009). In particular, the INAR(p) model can be seen as the discrete analogue of the well-known AR(p) model for continuous-valued time series. The INAR(1) model was first introduced by Al-Osh & Alzaid (1987) and McKenzie (1985), and its extension to the INAR(p) model by Du and Li (1991) is defined according to

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \ldots + \alpha_p \circ X_{t-p} + \varepsilon_t,$$

with $\varepsilon_t \overset{\text{i.i.d.}}{\sim} G$, where the innovation distribution G has range $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. The vector of INAR coefficients $\alpha = (\alpha_1, \ldots, \alpha_p)' \in (0,1)^p$ fulfills $\sum_{i=1}^{p} \alpha_i < 1$ and

$$\alpha_i \circ X_{t-i} = \sum_{j=1}^{X_{t-i}} Z_j^{(t,i)}, \; Z_j^{(t,i)} \sim \text{Bin}(1, \alpha_i),$$

where "$\circ$" denotes the binomial thinning operator first introduced by Steutel & Van Harn (1979). Although many contributions have been made during the last decades, most of the literature focuses on parametric INAR models and estimation techniques. We want to emphasize the efficient semiparametric estimation of INAR models (Drost, Van den Akker, & Werker, 2009).

Statement of need

INAR models find applications in a wide variety of fields such as medical sciences, environmentology and economics. For example, Franke & Seligmann (1993) model epileptic seizure counts using an INAR(1) model, Thyregod, Carstensen, Madsen, & Arnbjerg-Nielsen (1999) use integer-valued autoregressive models to model the dynamics of rainfall and McCabe & Martin (2005) to analyze wage loss claims data. They all have in common assuming that the innovation distribution belongs to a parametric class of distributions. Non- or semiparametric estimation of the INAR model was not considered until Drost et al. (2009) came up with their semiparametric estimation approach. A possible explanation is the complexity of the semiparametric setup since despite in the AR case the estimation in the INAR case cannot

The Journal of Open Source Software

be based on the residuals: Even if the autoregressive coefficients were known, observing the data does not imply observing the innovations (Drost et al., 2009). Nonetheless, one big advantage of semiparametric estimation is that we do not need to make a parametric distribution assumption on the innovations. The Poisson assumption is, for example, the most frequently used assumption for innovations and is characterized by equidispersion. In most cases, however, the data shows a higher variance than the mean value. The question arises when the distance between these two moments is large enough to not rather assume overdispersion, which would probably lead to assume negative binomially or geometrically distributed innovations. Furthermore, when dealing with low counts, we often observe many zeros in the data. This could be a sign for a zero-inflated innovation distribution such as the zero-inflated Poisson distribution (Jazi, Jones, & Lai, 2012). However, it is unclear at what point the zero is represented frequently enough in the data set to justify such an assumption. The mentioned points indicate that the assumption of an appropriate innovation distribution is often critical, bearing in mind that an incorrect assumption can lead to poor estimation performance. With semiparametric estimation, we do not have to commit to an innovation distribution, which makes this approach appealing.

To deal with count data time series, `R` (R Core Team, 2023) provides the package `tscount` (Liboschik, Fokianos, & Fried, 2017) which, a.o., includes likelihood-based estimation of parameter-driven count data time series models which do not include INAR models and exclusively allows for conditional Poisson or negative binomially distributed data. The R package `ZINARp` (Medina Garay, de Lima Medina, & Rossiter Araújo Monteiro, 2022) allows to simulate and estimate INAR data by using MCMC algorithms for estimation but the package is limited to parametric estimation of INAR models, that is, of the INAR coefficients and of a parametrically specified innovation distribution $\{G_\theta \,|\, \theta \in \mathbb{R}^q,\, q \in \mathbb{N}\}$ where they only cover the cases of Poisson or zero-inflated Poisson distributed innovations.The `Julia` (Bezanson, Edelman, Karpinski, & Shah, 2017) package `CountTimeSeries` (Stapper, 2022) deals with integer counterparts of ARMA and GARCH models and some generalizations including the INAR model. It covers the parametric estimation setup for INAR models but does also not allow for non-parametric estimation of the innovation distribution. Such a semiparametric estimation technique that still relies on the binomial thinning operation, but comes along without any parametric specification of the innovation distribution was proposed and proven to be efficient by Drost et al. (2009). Also neither of the three packages contains procedures for bootstrapping INAR models within these parametric and semiparametric setups. The `R` package `spINAR` fills this gap and combines simulation, estimation and bootstrapping of INAR models in a single package. Both, the estimation and the bootstrapping, are implemented semiparametrically and also parametrically. The package covers INAR models of order $p \in \{1, 2\}$, which are mainly used in applications.

Features

For the simulation of INAR data, our package allows for flexible innovation distributions that can be inserted in form of a parametric probability mass function (pmf) or by simply passing a user-defined vector as pmf argument. Regarding the estimation, it allows for moment- and maximum likelihood-based parametric estimation of INAR models with Poisson, geometrically or negative binomially distributed innovations (see for example Weiß (2018) for details), but the main contribution lies in the semiparametric maximum likelihood estimation of INAR models introduced by Drost et al. (2009) which they proved to be efficient. Additionally, a finite sample refinement for the semiparametric setup consisting of an estimation approach, that penalizes the roughness of the innovation distribution as well as a validation function for the penalization parameters is implemented (Faymonville, Jentsch, Weiß, & Aleksandrov, 2022). Furthermore, the package includes the possibility to bootstrap INAR data. Again, the user is able to choose the parametric or the more flexible semiparametric model specification and to perform the (semi)parametric INAR bootstrap described in Jentsch & Weiß (2017).

JOSS
The Journal of Open Source Software

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 437270842.

References

Al-Osh, M. A., & Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, *8(3)*, 261–275.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. doi:10.1137/141000671

Drost, F., Van den Akker, R., & Werker, B. (2009). Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models. *Journal of the Royal Statistical Society. Series B*, *71, Part 2*, 467–485.

Faymonville, M., Jentsch, C., Weiß, C. H., & Aleksandrov, B. (2022). Semiparametric estimation of INAR models using roughness penalization. *Statistical Methods and Applications*. doi:10.1007/s10260-022-00655-0

Franke, J., & Seligmann, T. (1993). Conditional maximum-likelihood estimates for INAR(1) processes and their applications to modeling epileptic seizure counts. *Developments in Time Series*, 310–330.

Jazi, M., Jones, G., & Lai, C. (2012). First-order integer valued AR processes with zero inflated poisson innovations. *Journal of Time Series Analysis*, *33*, 954–963. doi:10.1111/j.1467-9892.2012.00809.x

Jentsch, C., & Weiß, C. H. (2017). Bootstrapping INAR models. *Bernoulli*, *25(3)*, 2359–2408. doi:10.3150/18-BEJ1057

Liboschik, T., Fokianos, K., & Fried, R. (2017). tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, *82*(5), 1–51. doi:10.18637/jss.v082.i05

McCabe, B., & Martin, G. (2005). Bayesian predictions of low count time series. *International Journal of Forecasting*, *21(2)*, 315–330. doi:10.1016/j.ijforecast.2004.11.001

McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bulletin*, *21(4)*, 645–650.

Medina Garay, A. W., de Lima Medina, F., & Rossiter Araújo Monteiro, T. A. (2022). *ZINARp: Simulate INAR/ZINAR(p) models and estimate its parameters*. Retrieved from https://CRAN.R-project.org/package=ZINARp

R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Stapper, M. (2022). ManuelStapper/CountTimeSeries.jl: v0.1.4. doi:10.5281/zenodo.7488440

Steutel, F. W., & Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Annals of Probability*, *7(5)*, 893–899. doi:10.1214/aop/1176994950

Thyregod, P., Carstensen, J., Madsen, H., & Arnbjerg-Nielsen, K. (1999). Integer valued autoregressive models for tipping bucket rainfall measurements. *Environmetrics*, *10*, 395–411.

Weiß, C. H. (2009). *Categorical times series analysis and applications in statistical quality control*. dissertation.de.

Weiß, C. H. (2018). *An introduction to discrete-valued time series* (1st ed.). Wiley.

JOINT SEMI-PARAMETRIC INAR BOOTSTRAP INFERENCE FOR MODEL COEFFICIENTS AND INNOVATION DISTRIBUTION

MAXIME FAYMONVILLE

DEPARTMENT OF STATISTICS, TU DORTMUND UNIVERSITY, D-44221 DORTMUND, GERMANY;

FAYMONVILLE@STATISTIK.TU-DORTMUND.DE

AND

CARSTEN JENTSCH

DEPARTMENT OF STATISTICS, TU DORTMUND UNIVERSITY, D-44221 DORTMUND, GERMANY;

JENTSCH@STATISTIK.TU-DORTMUND.DE

ABSTRACT. For modeling the serial dependence in time series of counts, various approaches have been proposed in the literature. In particular, models based on a recursive, autoregressive-type structure such as the well-known integer-valued autoregressive (INAR) models are very popular in practice. The distribution of such INAR models is fully determined by a vector of autoregressive binomial thinning coefficients and the discrete innovation distribution. While fully parametric estimation techniques for these models are mostly covered in the literature, a semi-parametric approach allows for consistent and efficient joint estimation of the model coefficients and the innovation distribution without imposing any parametric assumptions. Although the limiting distribution of this estimator is known, which, in principle, enables asymptotic inference and INAR model diagnostics on the innovations, it is cumbersome to apply in practice.

In this paper, we consider a corresponding semi-parametric INAR bootstrap procedure and show its joint consistency for the estimation of the INAR coefficients and for the estimation of the innovation distribution. We discuss different application scenarios that include goodness-of-fit testing, predictive inference and joint dispersion index analysis for count time series. In simulations, we illustrate the finite sample performance of the semi-parametric INAR bootstrap using several innovation distributions and provide real-data applications.

Key words and phrases. Bootstrap inference, central limit theorem, count time series, dispersion index, semi-parametric estimation.

1. Introduction

Count time series consist of sequences of observations over time taking values in the non-negative integers $\mathbb{N}_0 := \mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$. They arise naturally when counting things or events over time and have therefore many relevant applications in various fields as, e.g., the number of infectious diseases, extreme weather events or phishing attacks. Unlike continuous time series, count data is inherently discrete-valued, which often leads to modeling challenges caused, e.g., by the presence of overdispersion (variance exceeding the mean) or zero inflation (excessive zeros in the data). One of the probably most used count time series models is the Integer-valued AutoRegressive (INAR) model of order $p \in \mathbb{N}$ introduced by Du and Li (1991). An INAR(p) process $(X_t, t \in \mathbb{Z})$ is defined by the recursion

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \ldots + \alpha_p \circ X_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{1.1}$$

where $(\varepsilon_t, t \in \mathbb{Z})$ denotes an i.i.d. innovation process with distribution G having range $\mathbb{N}_0$. We write $\varepsilon_t \sim G$ and identify the distribution G by its probability mass function (pmf), that is, $G = \{G(k), k \in \mathbb{N}_0\}$, where $G(k) = P(\varepsilon_t = k)$. Further, let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)' \in [0, 1)^p$ with $\sum_{i=1}^p \alpha_i < 1$ denote the vector of model coefficients and define

$$\alpha_i \circ X_{t-i} = \sum_{j=1}^{X_{t-i}} Z_j^{(t,i)},$$

where "$\circ$" is the binomial thinning operator first introduced by Steutel and Van Harn (1979). Here, $(Z_j^{(t,i)}, j \in \mathbb{N}, t \in \mathbb{N}_0)$, $i \in \{1, \ldots, p\}$, are mutually independent Bernoulli-distributed random variables with $Z_j^{(t,i)} \sim \text{Bin}(1, \alpha_i)$ such that $P(Z_j^{(t,i)} = 1) = \alpha_i = 1 - P(Z_j^{(t,i)} = 0)$ independent of $(\varepsilon_t, t \in \mathbb{N}_0)$. Note that, according to this construction, we have $\alpha_i \circ X_{t-i} | X_{t-i} \sim \text{Bin}(X_{t-i}, \alpha_i)$. All the thinning operations "$\circ$" are independent of each other and of $(\varepsilon_t, t \in \mathbb{Z})$. Furthermore, the thinning operation at time t and ε_t are both independent of X_s, $s < t$. The special case $p = 1$ results in the INAR(1) model introduced by McKenzie (1985) and Al-Osh and Alzaid (1987).

The existing literature mainly deals with *fully* parametric estimation of INAR models (see, e.g., Du and Li, 1991; Franke and Seligmann, 1993; Brännäs and Hellström, 2001; Freeland and McCabe, 2005; Jung et al., 2005; Silva and Silva, 2006; Bu et al., 2008), i.e., G is assumed to belong to a certain family of parametric distributions $\{G_\gamma \mid \gamma \in \Gamma \subset \mathbb{R}^q\}$ for some finite (and typically small) $q \in \mathbb{N}$. A summary of all the standard parametric estimation methods for INAR models as, e.g., moment estimation and (conditional) maximum likelihood estimation can be found in Section 2.2 of Weiß (2018). However, the use of such parametric assumptions considerably restricts the flexibility of the INAR model (1.1) - we refer to Faymonville et al. (2025b) for a discussion. A way more flexible estimator is the one presented by Drost et al.

(2009a). While keeping the *parametric* binomial thinning operation, it is able to estimate the innovation distribution in a completely *non-parametric* way. In the following Section 2, we introduce this semi-parametric estimator that allows to estimate *jointly* the INAR coefficients $\boldsymbol{\alpha}$ and the innovation distribution G and recap some of its theoretical (limiting) properties. An appropriate *semi-parametric* bootstrap procedure leading to corresponding bootstrap estimators is proposed in Section 3, where we also establish asymptotic theory and prove bootstrap consistency. Different application scenarios that include goodness-of-fit testing, predictive inference and joint dispersion index analysis for time series of counts are discussed in Section 4. We provide simulation results illustrating the finite sample performance of the semi-parametric INAR bootstrap procedure in Section 5 and discuss a real data application in Section 6. Section 7 concludes and the proofs of this paper are deferred to an appendix.

2. Preliminaries

By construction, the INAR(p) process defined by (1.1) is a pth order Markov chain. Under the model parameters $\boldsymbol{\alpha}$ and G and for $x_{t-i} \in \mathbb{N}_0$, $i = 0, 1, \ldots, p$, its transition probabilities are given by

$$P^{\boldsymbol{\alpha},G}_{(x_{t-1},\ldots,x_{t-p}),x_t} = \mathbb{P}_{\boldsymbol{\alpha},G}\left(\sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \varepsilon_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p}\right) \tag{2.1}$$
$$= \Big(\mathrm{Bin}(x_{t-1}, \alpha_1) * \cdots * \mathrm{Bin}(x_{t-p}, \alpha_p) * G\Big)\{x_t\},$$

where $\mathbb{P}$ is the underlying probability measure and "$*$" denotes the convolution of distributions. In the special case of an INAR(1) model, the transition probabilities can be written as

$$\mathbb{P}_{\alpha,G}(X_t = x_t \mid X_{t-1} = x_{t-1}) = \sum_{j=0}^{\min(x_t, x_{t-1})} \binom{x_{t-1}}{j} \alpha^j (1-\alpha)^{x_{t-1}-j} G(x_t - j), \tag{2.2}$$

where α is the coefficient of the INAR(1) model (McKenzie, 1985; Al-Osh and Alzaid, 1987). When estimating the INAR model as proposed in Drost et al. (2009a), that is, without using any parametric assumption on the family of the innovation distributions, one treats the error distribution as an (infinite-dimensional) parameter and jointly estimates the parameter sequence consisting of the p-dimensional vector of INAR model coefficients and the infinite-dimensional pmf of the innovations distribution. In the setup of generalized linear models, Huang (2014) considers a similar approach for the estimation of the error distribution. For the derivation of asymptotic theory, Drost et al. (2009a) impose the following (moment) assumptions on the innovation distribution and the model parameters.

Assumption 1 (Innovation distribution and model parameters; Drost et al. (2009a), Assumption 1)**.** *Let $\widetilde{\mathcal{G}}$ denote the set of all probability measures on $\mathbb{N}_0$. We assume that $G = \mathcal{L}(\varepsilon_t) \in \mathcal{G}$,*

where

$$\mathcal{G} = \left\{ G \in \widetilde{\mathcal{G}} : 0 < G(0) < 1, E_G(\varepsilon_t^{p+4}) < \infty \right\}$$

is the set of all probability measures on $\mathbb{N}_0$ *with* $0 < G(0) < 1$ *and finite* $(p+4)$*th moments. Furthermore, we assume that* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p) \in \Theta = \{\boldsymbol{\alpha} \in (0,1)^p : \sum_{i=1}^p \alpha_i < 1\}$.

For some of the results that Drost et al. (2009a) derive in their paper, weaker conditions than listed in Assumption 1 are sufficient. Nevertheless, the conditions included in Assumption 1 are not restrictive: The condition $0 < G(0) < 1$ makes sure that the innovations can be equal to zero, but that they are not always equal to zero, which is suitable for basically every practical application. The existence of the $(p+4)$th moment of the innovation distribution G is required to ensure the weak convergence of certain empirical processes; see Drost et al. (2009a) for details. The assumption $\boldsymbol{\alpha} \in (0,1)^p$ with $\sum_{i=1}^p \alpha_i < 1$ entails the standard assumption to ensure the stationarity of the INAR(p) process and avoids boundary issues by ruling out $\alpha_i = 0$ for all $i = 1, \ldots, p$.

Formally, Drost et al. (2009a) base their work on the experiments

$$\mathcal{E}^{(n)} = \left(\mathbb{N}_0^{n+1+p}, \mathcal{P}(\mathbb{N}_0^{n+1+p}), \mathbb{P}^{(n)}_{\nu_{\boldsymbol{\alpha},G},\boldsymbol{\alpha},G} | \boldsymbol{\alpha} \in \Theta, G \in \mathcal{G} \right), \quad n \in \mathbb{N}_0,$$

where $\mathcal{P}(\mathbb{N}_0^{n+1+p})$ denotes the power set of $\mathbb{N}_0^{n+1+p}$, $\mathbb{P}^{(n)}_{\nu_{\boldsymbol{\alpha},G},\boldsymbol{\alpha},G}$ denotes the law of $(X_{-p}, \ldots, X_n)$ under $\mathbb{P}_{\nu_{\boldsymbol{\alpha},G},\boldsymbol{\alpha},G}$, on the measurable space $(\mathbb{N}_0^{n+1+p}, \mathcal{P}(\mathbb{N}_0^{n+1+p}))$ and with stationary initial distribution $\nu_{\boldsymbol{\alpha},G}$ of $(X_{-p}, \ldots, X_0)$.[1] As one might expect, this semi-parametric model is more complicated to deal with than a parametric approach, but it is way more flexible with respect to the innovation distribution (see Drost et al., 2008, for the parametric counterparts).

The estimator is derived using a non-parametric (conditional) maximum likelihood approach. For any fixed $n \in \mathbb{N}_0$ and observations $(X_{-p}, \ldots, X_n)$ at hand, a non-parametric maximum likelihood estimator (NPMLE) $\widehat{\theta}_n := (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) = (\widehat{\alpha}_{n,1}, \ldots, \widehat{\alpha}_{n,p}, \widehat{G}_n(0), \widehat{G}_n(1), \ldots)$ of $(\boldsymbol{\alpha}, G)$ is defined to maximize the conditional likelihood, i.e.,

$$(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) \in \operatorname*{argmax}_{(\boldsymbol{\alpha},G) \in [0,1]^p \times \widetilde{\mathcal{G}}} \left(\prod_{t=0}^{n} P^{\boldsymbol{\alpha},G}_{(X_{t-1},\ldots,X_{t-p}),X_t} \right). \tag{2.3}$$

To guarantee its existence, note that they allow for values of $(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$ outside of $\Theta \times \mathcal{G}$. They further state that all the mass of $\widehat{G}_n$ is assigned to a subset $\{u_-, \ldots, u_+\} \subset \mathbb{N}_0$, where

$$u_- = \max\left\{0, \min_{t=0,\ldots,n} \left\{ X_t - \sum_{i=1}^{p} X_{t-i} \right\}\right\}, \quad u_+ = \max_{t=0,\ldots,n} \{X_t\}.$$

Then, $(\widehat{\theta}_n, \widehat{G}_n)$ maximizes the likelihood if and only if the following conditions hold:

[1] Although a sample $X_1, \ldots, X_n$ plus p pre-sample values $X_{-(p-1)}, \ldots, X_0$ would be sufficient, throughout this paper, we stick to the notation in Drost et al. (2009a), who suppose availability of $(X_{-p}, \ldots, X_{-1}), X_0, \ldots, X_n$, but nevertheless use a pre-factor of $1/n$ when taking sample averages.

i) $\widehat{G}_n(k) = 0$ for $k < u_-$ and $k > u_+$, and

ii) $(\widehat{\alpha}_{n,1}, \ldots, \widehat{\alpha}_{n,p}, \widehat{G}_n(u_-), \ldots, \widehat{G}_n(u_+))$ is a solution to the (constrained) polynomial optimization problem

$$\max_{\substack{x_1,\ldots,x_p \\ z_{u_-},\ldots,z_{u_+}}} \left\{ \prod_{t=0}^{n} \sum_{e=0 \vee X_t - \sum_{i=1}^{p} X_{t-i}}^{X_t} z_e \sum_{\substack{0 \le k_l \le X_{t-l}, l=1,\ldots,p \\ k_1+\ldots+k_p = X_t - e}} \prod_{l=1}^{p} \binom{X_{t-l}}{k_l} x_l^{k_l} (1-x_l)^{X_{t-l}-k_l} \right\} \tag{2.4}$$

subject to

$$0 \le x_k \le 1 \quad \text{for } k = 1, \ldots, p, \quad z_j \ge 0 \quad \text{for } j = u_-, \ldots, u_+, \quad z_{u_-} + \ldots + z_{u_+} = 1.$$

They stress that they do not impose the uniqueness of such a maximum location. A modification of the likelihood in (2.3) is proposed by Faymonville et al. (2023), who add a penalization term to account for the "smoothness" of most (discrete) innovation distributions, i.e., that $G(k+1) - G(k)$, $k \in \mathbb{N}_0$ is usually "small", which leads to an improved estimation performance for small sample sizes.

In their paper, Drost et al. (2009a) prove consistency, asymptotic normality and efficiency of their NPMLE under suitable regularity conditions. However, the practical use of their asymptotic theory is rather restricted as the derived limiting distribution follows a (transformed) Gaussian process, which is cumbersome to work with. Hence, with the goal of constructing a suitable and asymptotically valid semi-parametric INAR bootstrap procedure, in the following, we recap the consistency and asymptotic normality results established in Drost et al. (2009a).

Theorem 2.1 (Drost et al. (2009a), Theorem 1)**.** *Let Assumption 1 hold. For all* $(\boldsymbol{\alpha}_0, G_0) \in \Theta \times \mathcal{G}$ *and all initial probability measures* $\nu_{\boldsymbol{\alpha}_0, G_0}$ *on* $\mathbb{N}_0^p$*, any NPMLE* $(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$ *defined in* (2.3) *is consistent (as* $n \to \infty$*) in the following sense:*

$$\widehat{\boldsymbol{\alpha}}_n \xrightarrow{p} \boldsymbol{\alpha}_0 \quad \textit{and} \quad \sum_{k=0}^{\infty} \left| \widehat{G}_n(k) - G_0(k) \right| \xrightarrow{p} 0, \quad \textit{under } \mathbb{P}_{\nu_{\boldsymbol{\alpha}_0, G_0}, \boldsymbol{\alpha}_0, G_0}.$$

To derive the limiting distribution of the NPMLE, Drost et al. (2009a) show that their NPMLE is actually an infinite dimensional Z-estimator (see, e.g., Kosorok (2006) for a definition), i.e., that it solves an infinite number of moment conditions. We get these *infinitely many* moment conditions, because we allow for unbounded innovation distributions having support $\mathbb{N}_0$. This makes the setting flexible, but also complicates the derivation of the limiting distribution. Drost et al. (2009a) handle this by constructing (artificial) probability distributions on $\mathbb{N}_0$ bounded in direction $h : \mathbb{N}_0 \to \mathbb{R}$. They only use moment conditions arising from $h \in \mathcal{H}_1$ with $\mathcal{H}_1$ being the unit ball of $\ell^\infty(\mathbb{N}_0)$, where the latter denotes the Banach space of bounded sequences equipped with the supremum norm, i.e., they only consider functions

$h : \mathbb{N}_0 \to \mathbb{R}$ with $\sup_{e\in\mathbb{N}_0} |h(e)| \leq 1$. Finally, the estimating equations of the NPMLE are derived as $\Psi_n = (\Psi_{n1}, \Psi_{n2}) : (0,1)^p \times \widetilde{\mathcal{G}} \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ defined by

$$\Psi_{n1}(\boldsymbol{\alpha}, G) = \frac{1}{n}\sum_{t=0}^{n} \dot{l}_{\boldsymbol{\alpha}}(X_{t-p}, \ldots, X_t; \boldsymbol{\alpha}, G), \tag{2.5}$$

$$\Psi_{n2}(\boldsymbol{\alpha}, G)h = \frac{1}{n}\sum_{t=0}^{n} \left(A_{\boldsymbol{\alpha},G} h(X_{t-p}, \ldots, X_t) - \int h dG \right), \quad h \in \mathcal{H}_1, \tag{2.6}$$

where, for $x_{t-p}, \ldots, x_t \in \mathbb{N}_0$,

$$\dot{l}_{\boldsymbol{\alpha}}(x_{t-p}, \ldots, x_t; \boldsymbol{\alpha}, G) = \frac{\partial}{\partial \boldsymbol{\alpha}} \log\left(P^{\boldsymbol{\alpha},G}_{(x_{t-1},\ldots,x_{t-p}),x_t} \right) \tag{2.7}$$

with the convention that $\dot{l}_{\boldsymbol{\alpha}}(x_{t-p}, \ldots, x_t; \boldsymbol{\alpha}, G) = 0$ if $P^{\boldsymbol{\alpha},G}_{(x_{t-1},\ldots,x_{t-p}),x_t} = 0$ and

$$A_{\boldsymbol{\alpha},G} h(x_{t-p}, \ldots, x_t) = \mathrm{E}_{\boldsymbol{\alpha},G}\big(h(\varepsilon_t) | X_t = x_t, \ldots, X_{t-p} = x_{t-p}\big). \tag{2.8}$$

Additionally, for $(\boldsymbol{\alpha}_0, G_0)$, they introduce the population counterparts of these estimating equations, $\Psi^{\boldsymbol{\alpha}_0,G_0} = (\Psi_1^{\boldsymbol{\alpha}_0,G_0}, \Psi_2^{\boldsymbol{\alpha}_0,G_0}) : (0,1)^p \times \mathcal{G} \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$, where

$$\Psi_1^{\boldsymbol{\alpha}_0,G_0}(\boldsymbol{\alpha}, G) = \mathrm{E}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0} \left(\dot{l}_{\boldsymbol{\alpha}}(X_{t-p}, \ldots, X_t; \boldsymbol{\alpha}, G) \right), \tag{2.9}$$

$$\Psi_2^{\boldsymbol{\alpha}_0,G_0}(\boldsymbol{\alpha}, G)h = \mathrm{E}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0} \left(A_{\boldsymbol{\alpha},G}\, h(X_{-p}, \ldots, X_0) - \int h\, dG \right), \, h \in \mathcal{H}_1. \tag{2.10}$$

By exploiting that their NPMLE $(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$ provides a solution to the estimating equations, i.e., that $\Psi_n(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) = 0$ holds approximately with Ψ_n defined in equations (2.5) and (2.6), they derive a weak convergence result

$$\mathcal{S}_n^{\boldsymbol{\alpha}_0,G_0} = \sqrt{n}\left(\Psi_n(\boldsymbol{\alpha}_0, G_0) - \Psi^{\boldsymbol{\alpha}_0,G_0}(\boldsymbol{\alpha}_0, G_0) \right) \rightsquigarrow \mathcal{S}^{\boldsymbol{\alpha}_0,G_0} \tag{2.11}$$

in $\mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$, under $\mathbb{P}_{\nu_0,\boldsymbol{\alpha}_0,G_0}$, where "$\rightsquigarrow$" indicates weak convergence and $\mathcal{S}^{\boldsymbol{\alpha}_0,G_0}$ is a tight, Borel measurable, Gaussian process (see Drost et al., 2009a, Eq. (15)). Altogether, Drost et al. (2009a) get the following result for the limiting distribution of an NPMLE $\widehat{\theta}_n = (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$.

Theorem 2.2 (Drost et al. (2009a), Theorem 2)**.** *Suppose Assumption 1 holds. For $\theta_0 = (\boldsymbol{\alpha}_0, G_0) \in \Theta \times \mathcal{G}$, any NPMLE $\widehat{\theta}_n = (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$ satisfies*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0 \right) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1}\left(\mathcal{S}^{\theta_0} \right), \tag{2.12}$$

in $\mathbb{R}^p \times \ell^1(\mathbb{N}_0)$, under $\mathbb{P}_{\nu_{\theta_0},\theta_0}$, where $\dot{\Psi}_{\theta_0}^{-1}$ is the continuous inverse of the Fréchet derivative of Ψ at θ_0 and $\mathcal{S}^{\theta_0}$ is the tight, Borel measurable, Gaussian process determined by (2.11)*.*

According to Theorem 2.2, the limiting distribution of the NPMLE is a transformed Gaussian process. However, due to the infinite dimensional parameter space, this transformation is rather complicated and relies on the (inverted) Fréchet derivative $\dot{\Psi}_{\theta_0}^{-1}$ of Ψ^{θ_0} given in (2.9) and (2.10) and given in Section 3.2 of Drost et al. (2009a), which is cumbersome in practical applications (see also Huang, 2014, for a discussion). This generally motivates to use a suitable bootstrap

procedure for statistical inference instead of using asymptotic approximations that requires the cumbersome explicit estimation of many nuisance parameters. We propose and investigate a semi-parametric INAR bootstrap procedure in the following section.

3. Semi-parametric INAR Bootstrap

In this section, to enable suitable semi-parametric bootstrap inference in INAR models, we propose to use the semi-parametric version of the INAR bootstrap as proposed in Jentsch and Weiß (2019), which they proved to be consistent for statistics belonging to the large class of functions of generalized means of $(X_t, t \in \mathbb{Z})$ under mild assumptions. However, this class of statistics does not contain statistics depending on the estimated innovation distribution. For the purpose of extending the existing theory and to cover also such statistics, in Section 3.1, we introduce the semi-parametric INAR bootstrap scheme and provide the required notation. Further, in Section 3.2, we prove bootstrap consistency by showing that the bootstrap version of the NPMLE leads in probability (conditional on the data) to the same limiting distribution as obtained for the NPMLE given in Theorem 2.2.

3.1. The semi-parametric INAR Bootstrap Scheme.

For bootstrap inference of the NPMLE, the semi-parametric INAR bootstrap scheme, as proposed in Jentsch and Weiß (2019) for functions of generalized means and implemented in the R package *spINAR* by (Faymonville et al., 2024), is defined as follows:

Step 1.) Given a sample (including pre-sample values) of count data $(X_{-p}, \ldots, X_{-1}), X_0, \ldots, X_n$, fit semi-parametrically an INAR(p) process (1.1) using the NPMLE (2.3) as proposed by Drost et al. (2009a) to get estimators $\widehat{\boldsymbol{\alpha}}_n = (\widehat{\alpha}_{n,1}, \ldots, \widehat{\alpha}_{n,p})'$ and $\widehat{G}_n = (\widehat{G}_n(k),\, k \in \mathbb{N}_0)$ for the INAR coefficients and for the pmf of the innovation distribution, respectively.

Step 2.) Generate bootstrap observations $(X^*_{-p}, \ldots, X^*_{-1}), X^*_0, \ldots, X^*_n$ according to

$$X^*_t = \widehat{\alpha}_{n,1} \circ^* X^*_{t-1} + \ldots + \widehat{\alpha}_{n,p} \circ^* X^*_{t-p} + \varepsilon^*_t,$$

where "$\circ^*$" denotes (mutually independent) bootstrap binomial thinning operations and $(\varepsilon^*_t, t \in \mathbb{Z})$ denotes an i.i.d. bootstrap innovation process with $\varepsilon^*_t \sim \widehat{G}_n$ (conditionally on the data).

Step 3.) Compute the bootstrap NPMLE $\widehat{\theta}^*_n = (\widehat{\boldsymbol{\alpha}}^*_n, \widehat{G}^*_n)$, where $\widehat{\boldsymbol{\alpha}}^*_n = (\widehat{\alpha}^*_{n,1}, \ldots, \widehat{\alpha}^*_{n,p})$ and $\widehat{G}^*_n = (\widehat{G}^*_n(k),\, k \in \mathbb{N}_0)$, according to equation (2.3), but applied to the bootstrap sample $(X^*_{-p}, \ldots, X^*_{-1}), X^*_0, \ldots, X^*_n$. That is, for any fixed $n \in \mathbb{N}_0$, a bootstrap NPMLE (boot-NPMLE) $\widehat{\theta}^*_n := (\widehat{\boldsymbol{\alpha}}^*_n, \widehat{G}^*_n) = (\widehat{\alpha}^*_{n,1}, \ldots, \widehat{\alpha}^*_{n,p}, \widehat{G}^*_n(0), \widehat{G}^*_n(1), \ldots)$ is defined to maximize the bootstrap version of the conditional likelihood in (2.3), i.e.,

$$(\widehat{\boldsymbol{\alpha}}^*_n, \widehat{G}^*_n) \in \underset{(\boldsymbol{\alpha}, G) \in [0,1]^p \times \widetilde{\mathcal{G}}}{\operatorname{argmax}} \left(\prod_{t=0}^{n} P^{\boldsymbol{\alpha}, G}_{(X^*_{t-1}, \ldots, X^*_{t-p}), X^*_t} \right), \tag{3.1}$$

under analogous conditions as described in the restricted optimization problem (2.4).

Analogous to Drost et al. (2009a), for the bootstrap NPMLE (bootNPMLE) $\widehat{\theta}_n^* = (\widehat{\boldsymbol{\alpha}}_n^*, \widehat{G}_n^*)$, which fulfills a "bootstrap version" of Assumption 1 by Lemma B.1, we get a Z-estimator representation with corresponding estimating equations $\Psi_n^* = (\Psi_{n_1}^*, \Psi_{n_2}^*) : (0,1)^p \times \widetilde{\mathcal{G}} \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ defined by

$$\Psi_{n,1}^*(\boldsymbol{\alpha}, G) = \frac{1}{n}\sum_{t=0}^{n} \dot{l}_{\boldsymbol{\alpha}}(X_{t-p}^*, \ldots, X_t^*; \alpha, G), \tag{3.2}$$

$$\Psi_{n,2}^*(\boldsymbol{\alpha}, G)h = \frac{1}{n}\sum_{t=0}^{n} \left(A_{\boldsymbol{\alpha},G} h(X_{t-p}^*, \ldots, X_t^*) - \int h \, dG \right), \quad h \in \mathcal{H}_1. \tag{3.3}$$

Throughout this paper, as usual in bootstrap literature, we make use of the following notation. With $\mathrm{E}^*(\cdot)$, $\mathrm{Var}^*(\cdot)$ and $\mathrm{Cov}^*(\cdot)$, we denote the bootstrap expected value and (co)variance, respectively, given the original observations. That is, for $\mathbb{X} = (X_{-p}, \ldots, X_{-1}, X_0, \ldots, X_n)$, we have $\mathrm{E}^*(\cdot) = \mathrm{E}(\cdot|\mathbb{X})$, $\mathrm{Var}^*(\cdot) = \mathrm{Var}(\cdot|\mathbb{X})$ and $\mathrm{Cov}^*(\cdot,\cdot) = \mathrm{Cov}(\cdot,\cdot|\mathbb{X})$. To be in-line with the notation of Drost et al. (2009a), we clarify $\mathrm{E}(\cdot) \mathrel{\widehat{=}} \mathrm{E}_{\boldsymbol{\alpha}_0, G_0}(\cdot)$, $\mathrm{E}^*(\cdot) \mathrel{\widehat{=}} \mathrm{E}_{\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n}(\cdot|\mathbb{X})$ as well as $P \mathrel{\widehat{=}} \mathbb{P}_{\boldsymbol{\alpha},G}$ and $P^*(\cdot) \mathrel{\widehat{=}} \mathbb{P}_{\boldsymbol{\alpha},G}(\cdot|\mathbb{X})$.

3.2. **Bootstrap theory.** For the subsequent asymptotic theory for the semi-parametric INAR bootstrap, we impose the following regularity condition.

Assumption 2 (Innovation distribution regularity)**.** *Suppose that the pmf G is bounded away from zero by an exponentially decaying function. That is, let $G \in \mathcal{G}_u$, where*

$$\mathcal{G}_u = \Big\{ G \in \mathcal{G} : \exists c_1 \in (0,1],\, c_2 \in (0,\infty) \text{ such that } G(k) \geq c_1 e^{-c_2 k} \, \forall k \in \mathbb{N}_0 \Big\}. \tag{3.4}$$

The previous Assumption 2 is rather mild and not restrictive in practice. $\mathcal{G}_u$ contains all innovation distributions for which there exists a strictly positive and exponentially decaying function that always "lies below" the pmf of the innovation distribution. This property is fulfilled by several innovation distributions like, e.g., the Poisson, the negative binomial and the geometric distributions. In particular, it guarantees $G(k) > 0$ for all $k \in \mathbb{N}_0$ such that boundary issues are ruled out. Note that Drost et al. (2009a) do not make such an assumption. However, in view of the limiting distribution that they obtained in Theorem 2.1, asymptotic normality of $\widehat{G}(k)$ cannot hold due to $z_k \geq 0$ by construction in (2.4). Hence, it appears that Drost et al. (2009a) do require such an assumption as well.

Along the lines of the asymptotic theory established for the NPMLE in Drost et al. (2009a), we need to first derive "estimation consistency" of the bootNPMLE. Such a result constitutes the bootstrap version of Theorem 2.1 (i.e., of Theorem 1 in Drost et al. (2009a)) and is required to argue that also the bootNPMLE is away from the boundary of the parameter space $\Theta \times \mathcal{G}$. Under the assumptions introduced above, we get the following result.

Theorem 3.1 (Estimation consistency of the bootNPMLE)**.** *Suppose Assumptions 1 and 2 hold and we observe data $X_{-p},\dots,X_n$ from an INAR(p) process $(X_t, t\in\mathbb{Z})$ with INAR coefficients $\boldsymbol{\alpha}_0$ and innovation distribution G_0 such that $\theta_0=(\boldsymbol{\alpha}_0,G_0)\in\Theta\times\mathcal{G}$. Let $E(X_1^k)<\infty$ for some $k>2(p+4)$ and $\widehat{\theta}_n=(\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n)$ be an NPMLE of $(\boldsymbol{\alpha}_0,G_0)$. Suppose the bootstrap proposal from Section 3.1 is used to get a bootNPMLE $\widehat{\theta}_n^*=(\widehat{\boldsymbol{\alpha}}_n^*,\widehat{G}_n^*)$. Then, for all initial probability measures $\nu_{\boldsymbol{\alpha}_0,G_0}$ on $\mathbb{N}_0^p$, we have*

$$\widehat{\boldsymbol{\alpha}}_n^*-\widehat{\boldsymbol{\alpha}}_n\xrightarrow{p^*}0\quad\textit{and}\quad\sum_{k=0}^{\infty}|\widehat{G}_n^*(k)-\widehat{G}_n(k)|\xrightarrow{p^*}0$$

under $P^*(\cdot)=\mathbb{P}_{\nu_{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n},\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}(\cdot|\mathbb{X})$ in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability.

Now, we get to the main result of this paper, the limiting distribution of the bootNPMLE. As the resulting limiting distribution does coincide with the limiting distribution derived in Theorem 2.2 (i.e., in Theorem 2 in Drost et al., 2009a), the following Theorem 3.2 proves bootstrap consistency of the semi-parametric INAR bootstrap for the NPMLE.

Theorem 3.2 (Bootstrap consistency of bootNPMLE)**.** *Suppose Assumptions 1 and 2 hold and we observe data $X_{-p},\dots,X_n$ from an INAR(p) process $(X_t, t\in\mathbb{Z})$ with INAR coefficients $\boldsymbol{\alpha}_0$ and innovation distribution G_0 such that $\theta_0=(\boldsymbol{\alpha}_0,G_0)\in\Theta\times\mathcal{G}$. Let $E(X_1^k)<\infty$ for some $k>2(p+4)$, $E((X_t^3(1+\rho)^{X_t}))^{1+\delta}<\infty$ for some $\rho,\delta>0$ and $\widehat{\theta}_n=(\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n)$ be an NPMLE of $\theta_0=(\boldsymbol{\alpha}_0,G_0)$. Suppose the bootstrap proposal from Section 3.1 is used to get a bootNPMLE $\widehat{\theta}_n^*=(\widehat{\boldsymbol{\alpha}}_n^*,\widehat{G}_n^*)$. Then, for all initial probability measures $\nu_{\boldsymbol{\alpha}_0,G_0}$ on $\mathbb{N}_0^p$, we have*

$$\sqrt{n}\left(\widehat{\theta}_n^*-\widehat{\theta}_n\right)=\sqrt{n}\left((\widehat{\boldsymbol{\alpha}}_n^*,\widehat{G}_n^*)-(\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n)\right)\rightsquigarrow^*-\dot{\Psi}_{\theta_0}^{-1}\left(\mathcal{S}^{\theta_0}\right)\tag{3.5}$$

in $\mathbb{R}^p\times\ell^1(\mathbb{N}_0)$, under $P^*(\cdot)=\mathbb{P}_{\nu_{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n},\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}(\cdot|\mathbb{X})$ in $\mathbb{P}_{\nu_{\theta_0},\theta_0}$-probability.

Proof. Along the lines of the proof of Theorem 2 in Drost et al. (2009a), we make use of Theorem 3.3.1 in Van der Vaart and Wellner (1996) in the following to prove asymptotic normality in (3.5). By construction of the (boot)NPMLE as Z-estimator, we approximately have

$$\Psi_n(\widehat{\theta}_n)=0\quad\text{and}\quad\Psi_n^*(\widehat{\theta}_n^*)=0.\tag{3.6}$$

Using (3.6), we get

$$\sqrt{n}(\Psi_n(\widehat{\theta}_n^*)-\Psi_n(\widehat{\theta}_n))=\sqrt{n}(\Psi_n(\widehat{\theta}_n^*)-\Psi_n^*(\widehat{\theta}_n^*)).\tag{3.7}$$

Knowing from Lemma A.1 that Ψ_n is Fréchet-differentiable, according to (A.10), we can replace the left-hand side of (3.7) by

$$\sqrt{n}\left(\dot{\Psi}_n^{\xi_n}(\widehat{\theta}_n^*-\widehat{\theta}_n)\right),\tag{3.8}$$

where $||\xi_n - \widehat{\theta}_n|| \leq ||\widehat{\theta}_n^* - \widehat{\theta}_n||$. Using the uniform convergence of $\dot{\Psi}_n^{\xi_n}$ to $\dot{\Psi}^{\theta_0}$ established in Lemma A.2, (3.8) is asymptotically equivalent to

$$\sqrt{n}\left(\dot{\Psi}^{\theta_0}(\widehat{\theta}_n^* - \widehat{\theta}_n)\right). \tag{3.9}$$

Further, using the result of Lemma A.3, we can rewrite also the right-hand side of (3.7). Altogether, we have

$$\sqrt{n}\left(\dot{\Psi}^{\theta_0}(\widehat{\theta}_n^* - \widehat{\theta}_n)\right) = \sqrt{n}\left(\Psi_n(\widehat{\theta}_n) - \Psi_n^*(\widehat{\theta}_n)\right) + o_{p^*}(1). \tag{3.10}$$

Finally, using the asymptotic normality result from Lemma A.4 and the fact that $\dot{\Psi}_{\theta_0}$ is continuously invertible according to Lemma 2 in Drost et al. (2009a), making use of the continuous mapping theorem, the assertion follows.

□

Together with the bootstrap consistency for functions of generalized means derived in Jentsch and Weiß (2019), the asymptotic results established in Theorems 2.2 and 3.2 generalize the validity of the semi-parametric INAR bootstrap for a broader class of statistics. In particular, as the limiting distributions in Theorems 2.2 and 3.2 coincide, we obtain the following corollary.

Corollary 3.3 (First-order semi-parametric INAR bootstrap consistency)**.** *Let d be an appropriate metric on (the distributions of) random elements in $\mathbb{R}^p \times \ell^1(\mathbb{N}_0)$. Under the conditions of Theorem 2.2 and 3.2, we have*

$$d\Big(\mathcal{L}(\sqrt{n}(\widehat{\theta}_n - \theta_0)), \mathcal{L}^*(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))\Big) \underset{n\to\infty}{\to} 0$$

in $P_{\nu_{\theta_0},\theta_0}$-probability.

Furthermore, it is possible to make use of suitable delta methods to extend the bootstrap consistency result also to smooth functionals applied to the (boot)NPMLE. More precisely, for the bootNPMLE result (3.5), we can employ the delta method introduced in Theorem 3.1 of Beutner and Zähle (2016) which - in contrast to the conventional functional delta method as given by Theorems 3.9.11 and 3.9.13 of Van der Vaart and Wellner (1996) - does not make use of concepts of integrals and outer probabilities. Similarly, for the result of Theorem 2.2, Theorem 3.9.4. of Van der Vaart and Wellner (1996) can be applied which is valid for mappings between linear metric spaces. The paper of Beutner (2024) discusses several variants of functional delta methods. Consequently, we get the following result.

Corollary 3.4 (First-order bootstrap consistency for smooth functionals)**.** *Let $\Xi : [0,1]^p \times \tilde{\mathcal{G}} \to \mathbb{R}^q$ be a sufficiently smooth functional such that the conditions in Theorem 3.1 of Beutner and*

Zähle (2016) are fulfilled and let d be an appropriate metric on (the distributions of) random elements in $\mathbb{R}^q$. Under the conditions of Theorem 2.2 and 3.2, we have

$$d\left(\mathcal{L}\left(\sqrt{n}\left(\Xi(\widehat{\theta}_n)-\Xi\left(\theta_0\right)\right)\right),\mathcal{L}^*\left(\sqrt{n}\left(\Xi(\widehat{\theta}_n^*)-\Xi(\widehat{\theta}_n)\right)\right)\right)\underset{n\to\infty}{\to}0$$

in $P_{\nu_{\theta_0},\theta_0}$-probability.

The bootstrap consistency result from Corollary 3.4 finds application in diverse setups. For instance, a joint consistency result that combines the bootstrap consistency for functions of generalized means derived in Theorem 3.2 and Corollary 3.7 in Jentsch and Weiß (2019) and the bootstrap consistency for smooth functionals of $\widehat{\theta}_n$ from Corollary 3.4 above, can be applied for semi-parametric INAR goodness-of-fit testing discussed in Section 4.1. Further examples include, e.g., predictive inference in count time series setups covered in Section 4.2 or the analysis of the (joint) dispersion index covered in Section 4.3.

4. Methodological Applications

In the following subsections, wo discuss three methodological applications of Theorem 3.2 and Corollary 3.4 for different statistical tasks. They cover goodness-of-fit testing, predictive inference, and joint dispersion index analysis for (semi-parametrically estimated) INAR processes.

4.1. **Goodness-of-fit test on the whole INAR model class.** Together with the results from Jentsch and Weiß (2019), the result of our main Theorem 3.2 provides the theoretical foundation of the bootstrap-based semi-parametric goodness-of-fit test introduced in Faymonville et al. (2025b). They consider the null hypothesis "H_0: $(X_t, t \in \mathbb{Z})$ is an INAR(p) process" and their test is characterized by the lack of any parametric assumption on the innovation distribution which credits the flexibility of the INAR model class. The test statistic consists of an L_2-type distance of probability generating functions and can be represented as a degenerate V-statistic. This leads to a cumbersome χ^2-type limiting distribution requiring an appropriate bootstrap technique to make the testing procedure practicable. The semi-parametric INAR bootstrap described in Section 3.1 is used for simulations leading to good finite sample performance under the null and under the alternative. In this paper, we provide the theoretical justification the use of this bootstrap procedure for goodness-of-fit testing of semi-parametric INAR null hypotheses.

4.2. **Predictive inference for count time series.** In the setup of count time series and more general for discrete-valued time series, it is not straightforward to perform predictive inference. Faymonville et al. (2025a) discuss this topic and propose to solve this issue by transforming the prediction problem into a parameter estimation problem. That is, given $(X_{-p},\ldots,X_{-1}), X_0,\ldots,X_n$ and for the case of an underlying Markov process of order one, they

construct *confidence* intervals for the probability that X_{n+1} falls into a certain set S conditional on $X_n = x_n$ for some $x_n \in \mathbb{N}_0$, i.e., for $P_{S,x_n} = P(X_{n+1} \in S | X_n = x_n)$. They introduce asymptotic and bootstrap approaches and illustrate their proposed procedure on INAR and INARCH models. For this purpose, they consider parametric as well as non-parametric approaches and, moreover, they also discuss the practically important case of model misspecification, which we do not touch here.

In the setup of Faymonville et al. (2025a), also a semi-parametric INAR(1) variant *without* specifying the parametric family of the innovation distribution becomes suitable. Hence, assuming that the data is generated from an INAR(1) model, the procedure makes use of the semi-parametric INAR bootstrap from Section 3.1 as described in the following:

Step 1.) Given $X_1, \ldots, X_n$, calculate the NPMLE $\widehat{\theta}_n = (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$ as well as the semi-parametric estimator $\widehat{P}^{sp}_{S,x_n} := P_{S,x_n}(\widehat{\theta}_n)$ for the predictive probability P_{S,x_n}.

Step 2.) Use the semi-parametric INAR bootstrap from Section 3.1 to get bootstrap observations $X^*_1, \ldots, X^*_n$ and to calculate the bootNPMLE $\widehat{\theta}^*_n = (\widehat{\boldsymbol{\alpha}}^*_n, \widehat{G}^*_n)$ as well as the semi-parametric bootstrap estimator $\widehat{P}^{sp,*}_{S,x_n} := P_{S,x_n}(\widehat{\theta}^*_n)$.

Step 3.) Repeat Step 2.) B-times, where B is large, to get $L^{*,b}_n = \widehat{P}^{sp,*,b}_{S,x_n} - \widehat{P}^{sp}_{S,x_n}$, $b = 1, \ldots, B$, and construct the $(1-\delta)$-confidence interval for P_{S,x_n} as $[\widehat{P}_{S,x_n} - q^*_{1-\delta/2}, \widehat{P}_{S,x_n} - q^*_{\delta/2}]$, where q^*_α denotes the α-quantile of the empirical distribution of $L^{*,b}_n$, $b = 1, \ldots, B$.

In Section 5.2, we provide some simulations to numerically validate the above procedure.

4.3. **Joint dispersion index analysis for observations and innovations.** When it comes to the question of a suitable innovation distribution in the INAR model, dispersion plays a major role. In practice, one usually estimates the dispersion index ID_X by the estimator $\widehat{ID}_X = S^2/\overline{X}$, where S^2 is the sample variance and $\overline{X}$ is the sample mean based on the observations $X_1, \ldots, X_n$. Then, one makes use of the INAR model to derive also an estimator for the dispersion index ID_ϵ of the innovations denoted by $\widehat{ID}_\varepsilon$. For instance, in the case of an INAR(1) model with coefficient α, we have the relationship

$$ID_X = \frac{ID_\varepsilon + \alpha}{1 + \alpha}.$$

Hence, we can estimate the innovation dispersion ID_ε index by computing $\widehat{ID}_\varepsilon = \widehat{ID}_X(1 + \widehat{\alpha}) - \widehat{\alpha}$. However, the observations are over-/equi-/underdispersed if and only if the innovations are over-/equi-/underdispersed (see Section 2.2.1 in Weiß, 2018). Based on $\widehat{ID}_X$ one usually commits to a parametric family of innovation distributions and thus determines if the innovations are over-/equi-/underdispersed. However, for model diagnostics, the question remains whether a dispersion index deviates enough from 1 to suggest under- or overdispersion instead of equidispersion. While there are tests based on $\widehat{ID}_X$ proposed, e.g., by Schweer and Weiß

(2014), so far, there is no way to approach the innovations and their corresponding dispersion directly that is *without* imposing a parametric assumption on the innovation distribution. However, by assuming a certain family of innovations for testing purposes, the dispersion is (often) fixed from the outset. For example, if a Poi-INAR model is assumed, the innovations will be equidispersed, while in the case of an NB-INAR model, they will be overdispersed.

With the NPMLE estimation approach proposed of Drost et al. (2009a), the innovation distribution and thus also its dispersion can be estimated directly without a parametric assumption. As described in the following, we exploit this approach together with the semi-parametric INAR bootstrap from Section 3.1 to construct confidence intervals for the dispersion index of the innovations (i.e., for ID_ϵ) and of the observations (i.e., for ID_X) in one shot. The following algorithm is valid in case of an INAR(1) model:

Step 1.) Given $X_1, \dots, X_n$, calculate the NPMLE $\widehat{\theta}_n = (\widehat{\alpha}_n, \widehat{G}_n)$ and estimate (semi-parametrically) the dispersion indices of the innovations $\widehat{ID}_\varepsilon^{sp}$ and of the observations $\widehat{ID}_X^{sp}$ by computing

$$\widehat{ID}_\varepsilon^{sp} = \frac{\sum_{j\in\mathbb{N}_0} j^2 \widehat{G}_n(j) - \left(\sum_{j\in\mathbb{N}_0} j\widehat{G}_n(j)\right)^2}{\sum_{j\in\mathbb{N}_0} j\widehat{G}_n(j)} \quad \text{and} \quad \widehat{ID}_X^{sp} = \frac{\widehat{ID}_\varepsilon^{sp} + \widehat{\alpha}_n}{1+\widehat{\alpha}_n}.$$

Step 2.) Use the semi-parametric INAR bootstrap from Section 3.1 to get bootstrap observations $X_1^*, \dots, X_n^*$ and to calculate the bootNPMLE $\widehat{\theta}_n^* = (\widehat{\alpha}_n^*, \widehat{G}_n^*)$ and estimate (semi-parametrically) the bootstrap dispersion indices of the innovations $\widehat{ID}_\varepsilon^{sp,*}$ and of the observations $\widehat{ID}_X^{sp,*}$ by

$$\widehat{ID}_\varepsilon^{sp,*} = \frac{\sum_{j\in\mathbb{N}_0} j^2 \widehat{G}_n^*(j) - \left(\sum_{j\in\mathbb{N}_0} j\widehat{G}_n^*(j)\right)^2}{\sum_{j\in\mathbb{N}_0} j\widehat{G}_n^*(j)} \quad \text{and} \quad \widehat{ID}_X^{sp,*} = \frac{\widehat{ID}_\varepsilon^{sp,*} + \widehat{\alpha}_n^*}{1+\widehat{\alpha}_n^*}.$$

Step 3.) Repeat Step 2.) B-times, where B is large, to get $L_{\varepsilon,n}^{*,b} = \widehat{ID}_\varepsilon^{sp,*,b} - \widehat{ID}_\varepsilon^{sp}$ and $L_{X,n}^{*,b} = \widehat{ID}_X^{sp,*,b} - \widehat{ID}_X^{sp}$, $b = 1, \dots, B$, and construct the $(1-\delta)$-confidence intervals for ID_ε and ID_X as $\left[\widehat{ID}_\varepsilon^{sp} - q_{\varepsilon,1-\delta/2}^*, \widehat{ID}_\varepsilon^{sp} - q_{\varepsilon,\delta/2}^*\right]$ and $\left[\widehat{ID}_X^{sp} - q_{X,1-\delta/2}^*, \widehat{ID}_X^{sp} - q_{X,\delta/2}^*\right]$, respectively, where $q_{\varepsilon,\alpha}^*$ and $q_{X,\alpha}^*$ denote the α-quantiles of the empirical distributions of $L_{\varepsilon,n}^{*,b}$, $b = 1, \dots, B$ and of $L_{X,n}^{*,b}$, $b = 1, \dots, B$, respectively.

In Section 5.3, we provide some simulations to validate the above procedure.

5. Simulations

In this section, we investigate the finite sample performance of the semi-parametric INAR bootstrap from Subsection 3.1 and the methodological applications from Sections 4.2 and 4.3 by simulations.

5.1. **Bootstrap confidence intervals for $\theta_0 = (\boldsymbol{\alpha}_0, G_0)$.** First, we illustrate the bootstrap performance for the task of confidence interval construction for the true model parameter $\theta_0 = (\boldsymbol{\alpha}_0, G_0)$. That is, in a simulation study with $K = 500$ Monte Carlo samples and $B = 500$ bootstrap repetitions, we use the semi-parametric INAR bootstrap to construct Hall's confidence intervals for the entries of $\theta_0 = (\boldsymbol{\alpha}_0, G_0)$. Its validity is justified by Theorem 3.2 as it uses that $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$ provides in probability the same limiting distribution as $\sqrt{n}(\widehat{\theta}_n - \theta_0)$. Consequently, for significance level $\delta = 5\%$, this results in the confidence interval

$$\left[\widehat{\theta}_n - q^*_{1-\delta/2}, \widehat{\theta}_n - q^*_{\delta/2}\right], \tag{5.1}$$

where $q^*_{1-\delta/2}$ and $q^*_{\delta/2}$ are the corresponding quantiles of $\widehat{\theta}_n^* - \widehat{\theta}_n$. We consider sample sizes $n \in \{100, 500, 1000\}$ and INAR(1) DGPs (i.e., $p = 1$ in (1.1)) with different INAR coefficients $\alpha \in \{0.1, 0.3, 0.5, 0.9\}$ and with innovations following either a Poisson or a negative binomial distribution resulting in equi- and overdispersion, respectively. For implementing the simulation study, we use the R package *spINAR* (Faymonville et al., 2024).

At first, we consider INAR(1) DGPs with different α parameter and Poi(λ) distributed innovations with $\lambda \in \{1, 3\}$. Table 1 contains the coverage and the average length of the computed confidence intervals in case of $\alpha = 0.5$ and $\lambda = 1$. We see that the coverage increases for increasing n approaching the desired coverage level of 95%, while the average length decreases. In this parameter setup, where the mean of the observations equals two and the values of $\widehat{G}(k)$ for $k > 4$ become tiny (not larger than 0.0031), we only consider the first six entries of the parameter vector, that is, α and $G(k)$, $k \in \{0, 1, 2, 3, 4\}$. As $G(k)$ already becomes small for rather small k, this also explains the comparably low coverage for $G(4)$. These small entries of the parameter vector are difficult to estimate since the corresponding values are rarely observed. In this setup, where we have G=Poi(1), we exemplary get $G(4) = 0.015$. The results for the other parameterizations can be found in Tables 9, 11, 13, 15, 17 and 19 in the appendix. Overall, we get comparable results. However, it has to be noted that the semi-parametric INAR bootstrap approach slightly loses in terms of coverage performance, when the observations' mean increases (which is the case for larger α and λ). However, this could have been expected as the number of parameters, i.e., the entries of the pmf, to be estimated also increases, when the mean of the innovations increases. In addition, the considered first five entries of the pmf, i.e., $G(k)$, $k \in \{0, 1, 2, 3, 4\}$, then only cover a smaller portion of the whole probability mass in cases with large innovations' (and observations') mean.

Next, we consider an INAR(1) DGP with $\alpha = 0.5$ but now with an overdispered negative binomial innovation distribution NB(N, π) with parameters ($N = 10$, $\pi = 10/11$), ($N = 2$, $\pi = 2/3$) and ($N = 1$, $\pi = 1/2$). These parameter choices also lead to an observations' mean of two, but they cover different levels of overdispersion (increasing in mentioned order). Table

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.896	0.932	0.960	0.802	0.914	0.918	0.880	0.924	0.942
average length	0.369	0.147	0.102	0.385	0.177	0.123	0.391	0.171	0.119
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.850	0.928	0.950	0.698	0.946	0.950	0.446	0.730	0.852
average length	0.328	0.137	0.094	0.181	0.082	0.056	0.072	0.035	0.027

TABLE 1. Coverage and average length of the bootstrap confidence intervals based on the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(1)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.946	0.946	0.948	0.848	0.936	0.924	0.886	0.942	0.924
average length	0.337	0.130	0.090	0.380	0.164	0.115	0.361	0.157	0.110
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.846	0.934	0.934	0.734	0.958	0.948	0.568	0.814	0.924
average length	0.270	0.115	0.081	0.170	0.078	0.054	0.093	0.048	0.036

TABLE 2. Coverage and average length of the bootstrap confidence intervals based on the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a NB(2,2/3)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

2 displays the results for the case of moderate overdispersion ($N = 2$, $\pi = 2/3$). We see that for increasing n, we again approach the desired coverage level of 95%, whereas the average length of the confidence intervals decreases. In this setting of overdispersion, we have even better coverage and average length of the intervals compared to the results in Table 1, where the used innovation distribution is equidispersed. For the parameterizations ($N = 1$, $\pi = 1/2$) and ($N = 10$, $\pi = 10/11$), we get similar results shown in Tables 23 and 26 in the appendix.

In this paragraph, in Tables 3 and 4, we compare the results of the semi-parametric INAR bootstrap that does not rely on any parametric assumption on the innovation distribution with

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.926	0.948	0.964	0.906	0.928	0.936	0.984	0.978	0.978
average length	0.298	0.128	0.090	0.230	0.103	0.074	0.039	0.008	0.006
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.848	0.918	0.938	0.866	0.924	0.938	0.846	0.908	0.930
average length	0.108	0.051	0.037	0.078	0.035	0.025	0.033	0.013	0.010

TABLE 3. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(1)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.964	0.760	0.490	0.446	0.044	0	0.034	0	0
average length	0.320	0.137	0.096	0.223	0.101	0.071	0.046	0.013	0.008
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.422	0.050	0	0.884	0.848	0.782	0.668	0.512	0.330
average length	0.102	0.049	0.035	0.086	0.039	0.027	0.041	0.017	0.012

TABLE 4. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a NB(2,2/3)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

a fully parametric approach that assumes that the INAR(1) data at hand follows a Poisson distribution. In the latter case, we use the parametric maximum-likelihood method for parameter estimation (see, e.g., Weiß, 2018, for details) together with the *parametric* INAR bootstrap of Jentsch and Weiß (2019). Both are also implemented in the R package *spINAR* (Faymonville et al., 2024). In Tables 3 as well as in Tables 10, 12, 14, 16, 18 and 20 in the appendix, we see the results for the already considered Poisson DGPs, but now using the parametric approach. As one could expect, on the one hand, the procedure leads to (considerably) smaller confidence intervals while providing similar or even better coverage. This is not surprising since we use

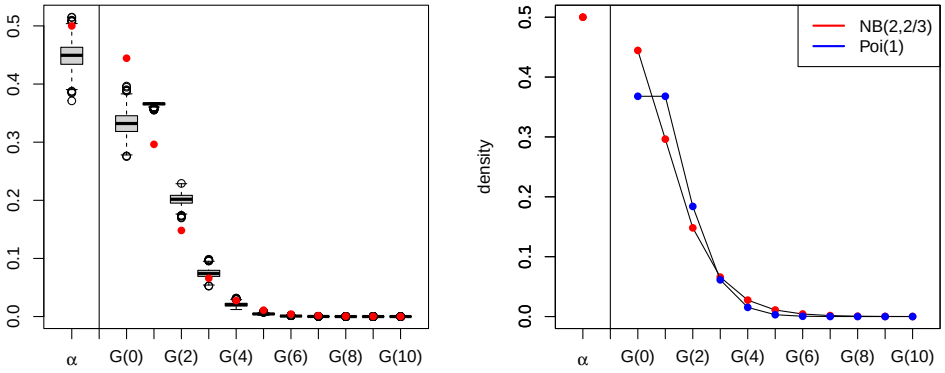

FIGURE 1. Left panel: Boxplots of the point estimators for the first twelve entries of the parameter vector θ for each Monte Carlo sample in case of $n = 1000$ with the true values being displayed in red. Right panel: pmfs of the NB(2,2/3) distribution (red) and the Poi(1) distribution (blue).

additional *true* information of the DGP. In contrast, on the other hand, Tables 4, 24 and 26 show what happens, when we use this additional distributional assumption in cases where it does not hold. The displayed results arise from applying parametric estimation and bootstrapping assuming a Poi(λ) innovation distribution in a case, where the innovation distribution is actually *not* Poisson, but negative binomial. Although the considered NB(N, π) innovation distribution has the same mean as before, the procedure breaks down. As expected, the confidence intervals do not hold the prescribed coverage rate. In particular, the coverage performance depends on the level of overdispersion. While the parameterization of $N = 10$ and $\pi = 10/11$ only leads to slight overdispersion, close to equidispersion, and not so inferior results (see Table 26), the results for the parameterizations with higher overdispersion become worse (see Table 24). Moreover, the coverage even decreases for increasing n, in some cases drastically. This again underlines the benefit of the semi-parametric approach. Without having to rely on any distribution assumptions that may not hold in practice, we achieve good results even compared to the parametric approach when a correct distribution assumption is used.

To get an intuition why the coverages may be extremely poor in some cases (even being equal to 0), consider the case of negative binomially distributed innovations with parameters $N = 2$ and $\pi = 2/3$, see Figure 1. In the left panel, we see the true parameter values ($\alpha = 0.5$ and the first eleven entries of the pmf in case of a NB(2,2/3) distribution) in red. The boxplots show the corresponding point estimates for each of the 500 Monte Carlo samples in case of $n = 1000$

DGP	n	100	500	1000	5000
Poi-INAR	coverage	0.918	0.938	0.944	0.946
	average length	0.210	0.089	0.066	0.028
NB-INAR	coverage	0.916	0.948	0.946	0.944
	average length	0.216	0.090	0.064	0.029
INARCH	coverage	0.916	0.752	0.558	0.372
	average length	0.215	0.094	0.065	0.029

TABLE 5. Coverage and average length of the confidence intervals for P_{S,x_n} for different sample sizes and different true DGPs.

when assuming an underlying Poisson distribution. It can be seen that some of these point estimators drastically over- or underestimate the true values. This is particularly the case for $G(0), G(1)$ and $G(2)$. This can be explained by the right panel, where, along with the true value of α, the pmf of a NB(2,2/3) distribution (red) and the one of a Poisson distribution providing the same mean (blue) are displayed. As can be seen, these two pmfs differ mainly (absolutely) in their first three entries.

Another finding is that although only the innovation distribution is misspecified, this misspecification also has a tremendous negative effect on the (point) estimation of the INAR coefficient α itself, as can be seen in Figure 1. Consequently, this also leads to a (too) low coverage of its confidence intervals. This is due to the fact that we use parametric (conditional) maximum likelihood estimation for implementing the parametric INAR bootstrap that estimates α and λ simultaneously. If we were to use moment estimators, i.e., least-squares or Yule-Walker estimation, to get an estimator for α in the first step and then to get an estimator for λ in the second step, this bias effect when estimating α would be avoided. However, in setups of a correctly specified Poi-INAR(1) process, the Yule-Walker estimation method is usually inferior to the maximum likelihood approach in finite samples such that the latter is mostly preferred in practice. In view of the results shown in Table 4, however, this should be done with care when doing inference.

5.2. **Predictive inference for count time series.** To validate the procedure described in Section 4.2, we set up simulations covering the same DGPs as in Faymonville et al. (2025a), namely Poi(1)-INAR(1) and NB(2,2/3)-INAR(1) with both $\alpha = 0.5$ and INARCH(1) with $\alpha = 0.5$ and $\beta = 1$. The results are displayed in Table 5 and are as expected. In the case, when the made assumption of an INAR model is true, the coverage of the confidence intervals increases towards 95% for increasing n, while the average interval length decreases. However,

		innovations				observations			
DGP	n	100	500	1000	5000	100	500	1000	5000
Poi-INAR	cov	0.838	0.944	0.942	0.942	0.876	0.950	0.940	0.942
	ave	0.778	0.365	0.260	0.118	0.545	0.246	0.175	0.079
NB-INAR	cov	0.854	0.910	0.928	0.940	0.878	0.918	0.932	0.938
	ave	1.139	0.550	0.393	0.179	0.771	0.361	0.257	0.117

TABLE 6. Coverage (cov) and average length (ave) of the confidence intervals for the dispersion indices for the innovations and the observations for different sample sizes and different true DGPs.

in case of the INARCH DGP, the INAR assumption is violated resulting in decreasing coverage for increasing n.

5.3. **Joint dispersion index analysis for observations and innovations.** To validate the performance of the bootstrap algorithm proposed in Section 4.3, we applied it again on a Poi(1)-INAR(1) DGP and an NB(2,2/3)-INAR(1) DGP both with $\alpha = 0.5$. Table 6 displays the results. We see that the coverage increases towards the desired coverage level of 95%, while the average length decreases. In the setup of the INAR model with negative binomial innovations, the intervals are systematically larger which can be explained by the underlying overdispersion of both the innovations and the observations compared to the equidispersion in case of the Poi-INAR model.

6. Real-Data Application

As a real-data application, based on the results of Theorem 3.2 and Corollary 3.4, we apply the procedures described in Sections 4.2 and 4.3 to semi-parametrically perform predictive inference and to construct confidence intervals for observations' and innovations' dispersion indices.

In one of their real-data applications, Faymonville et al. (2025a) considered a data set of the Deutsche Börse Group containing $n = 404$ transaction counts of structured products per trading day. The data along with the corresponding (P)ACF can be found in Figure 2. Based on the test result of Faymonville et al. (2025b) (see also Section 4.1) to not reject the semi-parametric null of an INAR(1) model at 5% level, they performed non-parametric as well as parametric predictive inference by assuming an INAR(1) model. For the parametric approaches, they separately imposed either a Poisson distribution or a geometric distribution for the innovations. They constructed 95% confidence intervals for $P_{S,x_n} := P(X_{n+1} \in S | X_n = x_n)$. As sets of interest S, they chose $S = \{0\}$ and $S = \{2\}$. We now apply the semi-parametric procedure

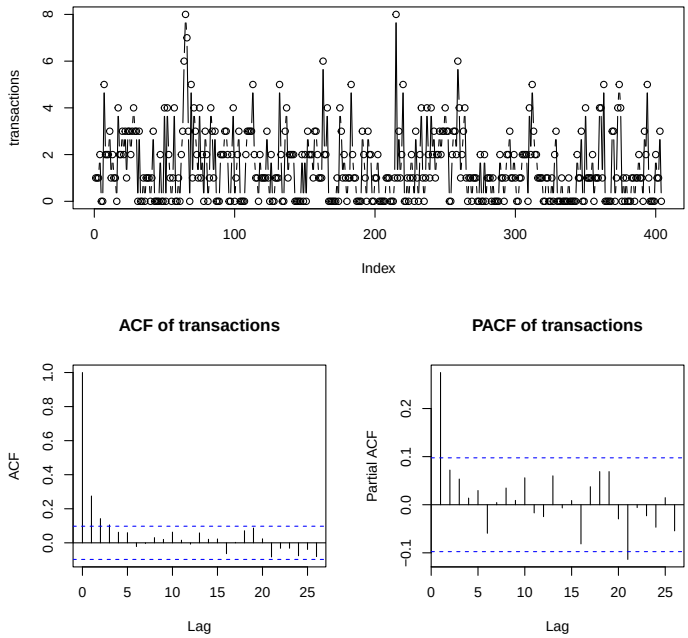

FIGURE 2. Plot of transaction counts and the corresponding (P)ACF (reproduced from Faymonville et al. (2025b).

described in Section 4.2 and give the point estimates along with confidence intervals for P_{S,x_n}. We display our results along with the ones of Faymonville et al. (2025a) (in italic) in Table 7. With their results, Faymonville et al. (2025a) could underline the results obtained in Faymonville et al. (2025b) that a Geo-INAR(1) model might be a good fit to the data. Our semi-parametric result fits well into this picture. The corresponding confidence intervals are nearly disjoint with the Poi-INAR ones and include the Geo-INAR ones, while being slightly shorter than the non-parametric ones.

To illustrate the procedure proposed in Section 4.3, we additionally construct confidence intervals for the dispersion indices of the innovations and the observations, see Table 8 for the result. Both confidence intervals do not include the 1. Hence, both intervals suggest that the data and also the innovations are overdispersed.

7. CONCLUSION

We proposed a semi-parametric INAR bootstrap procedure that allows for joint inference of the INAR model coefficients and the innovation distribution and circumvents the need for a cumbersome estimation of the limiting distribution. By using such a semi-parametric setting,

set	assumption	point estimation	CI
$S = \{0\}$	semi-param. INAR	0.4593	[0.3859, 0.5386]
	Poi-INAR	*0.3203*	*[0.2720,0.3676]*
	Geo-INAR	*0.4929*	*[0.4469,0.5290]*
	non-param.	*0.4884*	*[0.4023,0.5752]*
$S = \{2\}$	semi-param. INAR	0.1425	[0.0897, 0.1926]
	Poi-INAR	*0.2076*	*[0.1841, 0.2343]*
	Geo-INAR	*0.1267*	*[0.1189, 0.1365]*
	non-param.	*0.1318*	*[0.0676, 0.1824]*

TABLE 7. Point estimations and confidence intervals resulting from the semi-parametric procedure described in Section 4.2 for two different sets S along with the (non-)parametric results of Faymonville et al. (2025a) displayed in italic.

	point estimation	CI
innovations	1.6180	[1.3342, 1.8496]
observations	1.4832	[1.2660, 1.6593]

TABLE 8. Point estimation and confidence intervals for the dispersion indices of innovations and observations.

we avoid the use of any parametric assumptions on the innovation distribution. Indeed, such parametric assumptions may be too restrictive in practice and can have a large negative impact on the conducted inference, when they do not hold. Based on the semi-parametric INAR bootstrap proposed by Jentsch and Weiß (2019), which they showed to be consistent for functions of generalized means, we established bootstrap consistency results by proving a corresponding bootstrap central limit theorem. We illustrated the usefulness of our results by several methodological applications including goodness-of-fit testing, predictive inference and joint dispersion index analysis. In simulations, we illustrated our theoretical results by analyzing the coverage and the length of confidence intervals for different INAR model parameters in various setups. Concluding, the semi-parametric INAR bootstrap performs well in comparison to both non-parametric and parametric approaches. In particular, it turns out to be a practically relevant alternative that allows for robust inference in cases, where parametric assumptions are falsely imposed.

ACKNOWLEDGMENTS

Financial support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 437270842 (Model Diagnostics for Count Time Series) and 520388526 (TRR

391: Spatio-temporal Statistics for the Transition of Energy and Transport, Project A03) is gratefully acknowledged. Additionally, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

Al-Osh, M. A. and Alzaid, A. A. (1987). "First-order integer-valued autoregressive (INAR(1)) process". *J. Time Ser. Anal.* 8.3, pp. 261–275. DOI: 10.1111/j.1467-9892.1987.tb00438.x.

Beutner, E. (2024). "Functional delta method, asymptotic distribution". Working paper.

Beutner, E. and Zähle, H. (2016). "Functional delta-method for the bootstrap of quasi-Hadamard differentiable functionals". *Electron. J. Stat.* 10, pp. 1181–1222. DOI: 10.1214/16-EJS1140.

Bierens, H. (1982). "A uniform weak law of large numbers under ϕ-mixing with application to nonlinear least squares estimation". *Stat. Neerl.* 36.2, pp. 81–86. DOI: 10.1111/j.1467-9574.1982.tb00777.x.

Bradley, R. C. (2005). "Basic properties of strong mixing conditions. A survey and some open questions". *Probab. Surv.* 2. DOI: 10.1214/154957805100000104.

Brännäs, K. and Hellström, J. (2001). "Generalized integer-valued autoregression". *Econometric Rev.* 20.4, pp. 425–443. DOI: 10.1081/ETC-100106998].

Bu, R., McCabe, B., and Hadri, K. (2008). "Maximum likelihood estimation of higher-order integer-valued autoregressive processes". *J. Time Ser. Anal.* 29.6, pp. 973–994. DOI: 10.1111/j.1467-9892.2008.00590.x.

Davidson, J. (1994). *Stochastic Limit Theory. An Introduction for Econometricians.* Oxford University Press. ISBN: 978-0198774037.

Drost, F., Van den Akker, R., and Werker, B. (2008). "Local asymptotic normality and efficient estimation for INAR(p) models". *J. Time Ser. Anal.* 29.5, pp. 783–801. DOI: 10.1111/j.1467-9892.2008.00581.x.

— (2009a). "Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models". *J. R. Stat. Soc. B* 71.2, pp. 467–485. DOI: 10.1111/j.1467-9868.2008.00687.x.

— (2009b). *Technical Appendix to "Efficient Estimation of Auto-Regression Parameters and Innovation Distributions for Semiparametric Integer-Valued AR(p) Models".* URL: https://www.researchgate.net/publication/232767709_Technical_Appendix_to_Drost_Van_den_Akker_and_Werker_2009_-_JRSSB.

Du, J.-G. and Li, Y. (1991). "The integer valued autoregressive (INAR(p)) model". *J. Time Ser. Anal.* 12.2, pp. 129–142. DOI: 10.1111/j.1467-9892.1991.tb00073.x.

Faymonville, M., Jentsch, C., and Paparoditis, E. (2025a). "Predictive inference for discrete-valued time series". Working paper.

Faymonville, M., Jentsch, C., Weiß, C. H., and Aleksandrov, B. (2023). "Semiparametric estimation of INAR models using roughness penalization". *Stat. Meth. Appl.* 32.2, pp. 365–400. DOI: 10.1007/s10260-022-00655-0.

Faymonville, M., Jentsch, C., and Weiß, C. H. (2025b). "Semi-parametric goodness-of-fit testing for INAR models". *Forthcoming in: Bernoulli*. URL: https://www.e-publications.org/ims/submission/BEJ/user/submissionFile/63985?confirm=ca1b50bc.

Faymonville, M., Riffo, J., Rieger, J., and Jentsch, C. (2024). "spINAR: An R package for semi-parametric and parametric estimation and bootstrapping of integer-valued autoregressive (INAR) models". *J. Open Source Softw.* 9.97, p. 5386. DOI: 10.21105/joss.05386.

Franke, J. and Seligmann, T. (1993). "Conditional maximum-likelihood estimates for INAR(1) processes and their application to modelling epileptic seizure counts". *Developments in Time Series Analysis*. Ed. by T. S. Rao. London: Chapman & Hall, pp. 157–170. ISBN: 978-0412492600.

Freeland, R. and McCabe, B. (2005). "Asymptotic properties of CLS estimators in the Poisson AR(1) model". *Statist. Probab. Lett.* 73.2, pp. 147–153. DOI: 10.1016/j.spl.2005.03.006.

Huang, A. (2014). "Joint Estimation of the Mean and Error Distribution in Generalized Linear Models". *J. Am. Stat. Assoc.* 109.505, pp. 186–196. DOI: https://doi.org/10.1080/01621459.2013.824892.

Jentsch, C. and Weiß, C. (2019). "Bootstrapping INAR models". *Bernoulli* 25.3, pp. 2359–2408. DOI: 10.3150/18-BEJ1057.

Jung, R., Ronning, G., and Tremayne, A. (2005). "Estimation in conditional first order autoregression with discrete support". *Statist. Papers* 46.2, pp. 195–224. DOI: 10.1007/BF02762968.

Kosorok, M. (2006). *Introduction to Empirical Processes and Semiparametric Inference*. Springer. ISBN: 978-0387749778.

McKenzie, E. (1985). "Some simple models for discrete variate time series". *Water Resour. Bull.* 21.4, pp. 645–650. DOI: 10.1111/j.1752-1688.1985.tb05379.x.

Schweer, S. and Weiß, C. H. (2014). "Compound Poisson INAR(1) processes: Stochastic properties and testing for overdispersion". *Comput. Statist. Data Anal.* 77, pp. 267–284. DOI: 10.1016/j.csda.2014.03.005.

Silva, I. and Silva, M. (2006). "Asymptotic distribution of the Yule-Walker estimator for INAR(p) processes". *Statist. Probab. Lett.* 76.15, pp. 1655–1663. DOI: 10.1016/j.spl.2006.04.008.

Steutel, F. W. and Van Harn, K. (1979). "Discrete analogues of self-decomposability and stability". *Ann. Probab.* 7.5, pp. 893–899. DOI: 10.1214/aop/1176994950.

Van der Vaart, A. (2000). *Asymptotic Statistics*. 1st ed. Cambridge University Press. ISBN: 978-0521784504.

Van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer. ISBN: 978-0387946405.

Wald, A. (1949). "Note on the Consistency of the Maximum Likelihood Estimates". *Ann. Math. Stat.* 20.4, pp. 595–601. DOI: 10.1214/aoms/1177729952.

Weiß, C. H. (2018). *An Introduction to Discrete-Valued Time Series*. 1st ed. Wiley. ISBN: 978-1119096962.

Appendix A. Proofs of the Main Results

A.1. **Proof of Theorem 3.1.** For proving the estimation consistency of the bootstrap estimator $\widehat{\theta}_n^*$, we follow the proof idea of Theorem 2.1 that is given as Theorem 1 in Drost et al. (2009a). For this purpose, we adopt the notation introduced in Section B.2 of Drost et al. (2009b) and make use of similar arguments outlined in the following. Let $\widehat{\theta}_n = (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$ be an NPMLE of $(\boldsymbol{\alpha}_0, G_0)$ based on the sample $X_{-p}, \ldots, X_n$ and let $\widehat{\theta}_n^* = (\widehat{\boldsymbol{\alpha}}_n^*, \widehat{G}_n^*)$ be a bootNPMLE based on a bootstrap sample $X_{-p}^*, \ldots, X_n^*$ that is obtained from the bootstrap proposal from Section 3.1. Hence, the goal is to show that

$$\widehat{\boldsymbol{\alpha}}_n^* - \widehat{\boldsymbol{\alpha}}_n \xrightarrow{p^*} 0 \quad \text{and} \quad \sum_{k=0}^{\infty} |\widehat{G}_n^*(k) - \widehat{G}_n(k)| \xrightarrow{p^*} 0$$

holds in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability, where $P^*(\cdot) = \mathbb{P}_{\nu_{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n},\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}(\cdot|\mathbb{X})$. To prove this result, using that both $(\widehat{G}_n^*(k), k \in \mathbb{N}_0)$ and $(\widehat{G}_n(k), k \in \mathbb{N}_0)$ sum up to one by construction, i,e., $\sum_{k=0}^{\infty} \widehat{G}_n^*(k) = 1$ and $\sum_{k=0}^{\infty} \widehat{G}_n(k) = 1$ according to Drost et al. (2009b), it suffices to prove $\widehat{\boldsymbol{\alpha}}_n^* - \widehat{\boldsymbol{\alpha}}_n \xrightarrow{p^*} 0$ as well as $\widehat{G}_n^*(k) - \widehat{G}_n(k) \xrightarrow{p^*} 0$ for all $k \in \mathbb{N}_0$. We prove the latter by following the arguments of Wald's consistency theorem (see, e.g., the proof of Theorem 5.14 in Van der Vaart, 2000) and extend it to the bootstrap case. For this purpose, we have to consider the compactification of the parameter space analogously to Drost et al. (2009a). First, we introduce $\overline{\mathcal{G}}$, which denotes the class of all probability measures on $\mathbb{N}_0 \cup \infty$ and identify each $G \in \overline{\mathcal{G}}$ by its probability mass function (pmf) sequence $(G(k), k \in \mathbb{N}_0)$. This correspondence is 1-to-1 by the relationship $G(\infty) = 1 - \sum_{k \in \mathbb{N}_0} G(k)$. Hence, $\overline{\mathcal{G}}$ is a subset of $[0,1]^{\mathbb{N}_0}$ equipped with the norm $||a|| = \sum_{k=0}^{\infty} 2^{-k}|a(k)|$ for $a = (a(k), k \in \mathbb{N}_0)$. Following the arguments in Drost et al. (2009a) and Drost et al. (2009b), $\overline{\mathcal{G}}$ is a compact subset of $[0,1]^{\mathbb{N}_0}$. Further, for $G \in \overline{\mathcal{G}}$, we define $P_{x,\infty}^{\boldsymbol{\alpha},G} = 1 - \sum_{j \in \mathbb{N}_0} P_{x,j}^{\boldsymbol{\alpha},G} = G(\infty)$ for $x \in \mathbb{N}_0^p$ and $P_{x,\infty}^{\boldsymbol{\alpha},G} = 1$ if $\max_{i=1}^p x_i = \infty$. Second, by considering the compactification of the parameter space $\Theta = (0,1)^p$ of θ as well, i.e., $[0,1]^p$, we define $\overline{E} := [0,1]^p \times \overline{\mathcal{G}}$ equipped with the "sum distance" $d((\boldsymbol{\alpha}, G), (\boldsymbol{\alpha}', G')) = |\boldsymbol{\alpha} - \boldsymbol{\alpha}'| + ||(G(k), k \in \mathbb{N}_0) - (G'(k), k \in \mathbb{N}_0)||$. As $\overline{E}$ is the product of two compact spaces, it is itself compact.

Further, for $x_{-p}, \ldots, x_0 \in \mathbb{N}_0^{p+1}$, we define $m^{\boldsymbol{\alpha},G}(x_{-p}, \ldots, x_0) = \log P_{(x_{-1},\ldots,x_{-p}),x_0}^{\boldsymbol{\alpha},G}$, where $P_{(x_{-1},\ldots,x_{-p}),x_0}^{\boldsymbol{\alpha},G}$ is defined in (2.1), and introduce the (random) function $M_n : \overline{E} \to [-\infty, \infty)$ by

$$M_n(\boldsymbol{\alpha}, G) = \frac{1}{n} \sum_{t=0}^{n} m^{\boldsymbol{\alpha},G}(X_{t-p}, \ldots, X_t).$$

and its bootstrap analogue $M_n^* : \overline{E} \to [-\infty, \infty)$ by

$$M_n^*(\boldsymbol{\alpha}, G) = \frac{1}{n} \sum_{t=0}^{n} m^{\boldsymbol{\alpha},G}(X_{t-p}^*, \ldots, X_t^*).$$

We have to show that $M_n^*(\boldsymbol{\alpha}, G) - M_n(\boldsymbol{\alpha}, G) \xrightarrow{p^*} 0$ in probability for any $(\boldsymbol{\alpha}, G) \in \overline{E}$. Noting that

$$\mathrm{E}^*(M_n^*(\boldsymbol{\alpha}, G)) = M_n(\boldsymbol{\alpha}, G)$$

holds by construction (see, e.g., (A.6) in Faymonville et al., 2025a), we need to show that $M_n^*(\boldsymbol{\alpha}, G) - \mathrm{E}^*(M_n^*(\boldsymbol{\alpha}, G)) \xrightarrow{p^*} 0$ in probability. For this purpose, it remains to show that $\mathrm{Var}^*(M_n^*(\boldsymbol{\alpha}, G)) \xrightarrow{p^*} 0$ in probability as $n \to \infty$, where

$$\mathrm{Var}^*(M_n^*(\boldsymbol{\alpha}, G)) = \mathrm{Var}^*\left(\frac{1}{n}\sum_{t=0}^{n} m^{\boldsymbol{\alpha},G}(X_{t-p}^*, \ldots, X_t^*)\right) = \mathrm{Var}^*\left(\frac{1}{n}\sum_{t=0}^{n} \log P^{\boldsymbol{\alpha},G}_{(X_{t-1}^*,\ldots,X_{t-p}^*),X_t^*}\right).$$

For notational convenience, we focus on the case of an INAR(1) model, i.e., $p = 1$ exclusively, but the following can be extended analogously to the case of higher order INAR models. Using (2.2), we get

$$\begin{aligned}
&\mathrm{Var}^*(M_n^*(\alpha, G)) \\
&= \mathrm{Var}^*\left(\frac{1}{n}\sum_{t=0}^{n} \log\left(\sum_{j=0}^{\min(X_t^*, X_{t-1}^*)} \binom{X_{t-1}^*}{j} \alpha^j (1-\alpha)^{X_{t-1}^* - j} G(X_t^* - j)\right)\right) \\
&= \frac{1}{n^2}\sum_{t_1,t_2=0}^{n} \mathrm{Cov}^*\left(\log\left(\sum_{j_1=0}^{\min(X_{t_1}^*, X_{t_1-1}^*)} \binom{X_{t_1-1}^*}{j_1} \alpha^{j_1}(1-\alpha)^{X_{t_1-1}^* - j_1} G(X_{t_1}^* - j_1)\right),\right. \\
&\qquad \left.\log\left(\sum_{j_2=0}^{\min(X_{t_2}^*, X_{t_2-1}^*)} \binom{X_{t_2-1}^*}{j_2} \alpha^{j_2}(1-\alpha)^{X_{t_2-1}^* - j_2} G(X_{t_2}^* - j_2)\right)\right) \\
&= \frac{1}{n}\sum_{h=-(n-1)}^{n-1} \frac{1}{n}\sum_{t=\max(1,1-h)}^{\min(n,n-h)} \mathrm{Cov}^*\left(\log\left(\sum_{j_1=0}^{\min(X_t^*, X_{t-1}^*)} \binom{X_{t-1}^*}{j_1} \alpha^{j_1}(1-\alpha)^{X_{t-1}^* - j_1} G(X_t^* - j_1)\right),\right. \\
&\qquad \left.\log\left(\sum_{j_2=0}^{\min(X_{t+h}^*, X_{t+h-1}^*)} \binom{X_{t+h-1}^*}{j_2} \alpha^{j_2}(1-\alpha)^{X_{t+h-1}^* - j_2} G(X_{t+h}^* - j_2)\right)\right).
\end{aligned}$$

Due to the (strict) stationarity of $(X_t^*, t \in \mathbb{Z})$ (conditional on the data), the covariances on the last right-hand side only depend on h (and not on t). Further, due to $\min(n, n-h) - \max(1, 1-h) + 1 = n - |h| \leq n$, the last right-hand side can be bounded by

$$\begin{aligned}
&\frac{1}{n}\sum_{h=-(n-1)}^{n-1} \mathrm{Cov}^*\left(\log\left(\sum_{j_1=0}^{\min(X_1^*, X_0^*)} \binom{X_0^*}{j_1} \alpha^{j_1}(1-\alpha)^{X_0^* - j_1} G(X_1^* - j_1)\right),\right. \\
&\qquad \left.\log\left(\sum_{j_2=0}^{\min(X_{h+1}^*, X_h^*)} \binom{X_h^*}{j_2} \alpha^{j_2}(1-\alpha)^{X_h^* - j_2} G(X_{h+1}^* - j_2)\right)\right).
\end{aligned} \tag{A.1}$$

As argued by Drost et al. (2009a) in their Lemma 1(b), the INAR(p) process (here we focus on $p = 1$) is (geometrically) β-mixing under $P_{\nu_{\theta,G},\theta,G}$ for $(\theta, G) \in \Theta \times \mathcal{G}$, which is imposed by

Assumption 1. Consequently, according to Lemma B.1, with probability tending to one, the same holds for the bootstrap process $(X_t^*, t \in \mathbb{Z})$. Next, we define

$$Y_h^* := f(X_h^*, X_{h+1}^*) := \log\left(\sum_{j=0}^{\min(X_{h+1}^*, X_h^*)} \binom{X_h^*}{j} \alpha^j (1-\alpha)^{X_h^*-j} G(X_{h+1}^* - j)\right) \tag{A.2}$$

with an obvious notation for $f(\cdot,\cdot)$. As the argument of f consists of finitely many X_t^*'s, using Theorem 14.1 of Davidson (1994) and that β-mixing implies α-mixing (Bradley, 2005), we get that $(f(X_h^*, X_{h+1}^*, h \in \mathbb{Z})$ is also geometrically α-mixing with probability tending to one. Hence, using Corollary 14.3 in Davidson (1994), we get

$$|\mathrm{Cov}^*(Y_0^*, Y_h^*)| \leq 2\,(2^{1-1/q}+1)\,\alpha_{\mathrm{mix},n}(h)^{1-1/q-1/r}\,||Y_0^*||_q^*\,||Y_h^*||_r^* \tag{A.3}$$

with $q > 1, r > q/(q-1)$, where $||X_i^*||_s^* = (\mathrm{E}^*(|X_i^*|^s))^{1/s}$, $s \in \{q, r\}$ and $\alpha_{\mathrm{mix},n}(h)$ denotes the α-mixing coefficient of the bootstrap process $(X_t^*, t \in \mathbb{Z})$ at lag h, not to confuse with the coefficient estimator $\widehat{\alpha}_n$ of the INAR model. Then, as the bootstrap process is geometrically mixing with probability tending to one according to Lemma B.1 and the sequence $(\alpha_{\mathrm{mix},n}(h)^{1-1/q-1/r}, h \in \mathbb{N}_0)$ converges also exponentially fast to zero for $h \to \infty$, the covariances $\mathrm{Cov}^*(Y_0^*, Y_h^*)$ become absolutely summable with probability tending to one, whenever $||Y_0^*||_q^*$ and $||Y_h^*||_r^*$ are bounded in probability for some suitable q and r. The above term $2\,(2^{1-1/q}+1)$ is just a constant.

Hence, to conclude that the covariances in (A.3) are absolutely summable with probability tending to one, it remains to show that the moments of the function of the bootstrap data are bounded in probability, i.e.

$$\exists q > 1, r > q/(q-1) : ||Y_0^*||_q^* = O_p(1) \quad \text{and} \quad ||Y_h^*||_r^* = O_p(1). \tag{A.4}$$

Again, due to the strict stationarity of $(X_t^*, t \in \mathbb{Z})$ (conditional on the data), it suffices to show the previous for $h = 0$ and for some $q > 2$. If we want to prove that $\mathrm{E}^*(|Y_0^*|^q)^{1/q}$ is bounded in probability, it remains to prove that $\mathrm{E}^*(|Y_0^*|^q)$ is bounded in probability. Plugging in for Y_0^* as defined in (A.2) leads to

$$\mathrm{E}^*(|Y_0^*|^q) = \sum_{(x,y)\in\mathbb{N}_0^2} \left|\log\left(\sum_{j=0}^{\min(x,y)} \binom{x}{j} \alpha^j (1-\alpha)^{x-j} G(y-j)\right)\right|^q P^*\big((X_0^*, X_1^*) = (x,y)\big),$$

with $P^*(\cdot) = \mathbb{P}_{\nu_{\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n}, \widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n}(\cdot|\mathbb{X})$. Let $k > 0$ and consider the probability that the last right-hand side is larger than k, i.e.,

$$P\left(\sum_{(x,y)\in\mathbb{N}_0^2} \left|\log\left(\sum_{j=0}^{\min(x,y)} \binom{x}{j} \alpha^j (1-\alpha)^{x-j} G(y-j)\right)\right|^q P^*((X_0^*, X_1^*) = (x,y)) \geq k\right)$$

$$\leq P\left(\sum_{(x,y)\in\mathbb{N}_0^2, y\leq x} \left|\log\left(\sum_{j=0}^{y} \binom{x}{j} \alpha^j (1-\alpha)^{x-j} G(y-j)\right)\right|^q P^*((X_0^*, X_1^*) = (x,y)) \geq k/2\right)$$

$$+ P\left(\sum_{(x,y)\in\mathbb{N}_0^2, y>x} \left|\log\left(\sum_{j=0}^{x}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)\right)\right|^q P^*((X_0^*,X_1^*)=(x,y)) \geq k/2\right)$$

$$=: I + II.$$

Let us consider the first term I. As $\sum_{j=0}^{y}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j) = P_{x,y}^{\alpha,G} \in [0,1]$, we have $\log(\sum_{j=0}^{y}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)) \leq 0$. Hence, using that $-\log(\cdot)$ is a decreasing function and that all summands in $\sum_{j=0}^{y}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)$ are non-negative, we get an upper bound by only considering the last term of the sum, i.e., $-\log(\sum_{j=0}^{y}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)) \leq -\log(\binom{x}{y}\alpha^y(1-\alpha)^{x-y}G(0))$ and we have

$$\begin{aligned}
\left|\log\left(\sum_{j=0}^{y}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)\right)\right|^q &= \left(-\log\left(\sum_{j=0}^{y}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)\right)\right)^q \\
&\leq \left(-\log\left(\binom{x}{y}\alpha^y(1-\alpha)^{x-y}G(0)\right)\right)^q \\
&\leq \Big(-\log\left(\alpha^y(1-\alpha)^{x-y}G(0)\right)\Big)^q \\
&= \Big(-y\log(\alpha) - (x-y)\log(1-\alpha) - \log(G(0))\Big)^q,
\end{aligned}$$

where we used that $\binom{x}{y} \geq 1$ such that $-\log(\binom{x}{y}z) \leq -\log(z)$ holds for all $z \geq 0$. Finally, using that $\log(x) < 0$ for $x \in (0,1)$ and that $0 < G(0) < 1$ as well as $\alpha \in (0,1)$ by Assumption 1, we see that $-y\log(\alpha) - (x-y)\log(1-\alpha) - \log(G(0))$ is strictly positive and finite. Define $g(x) := x^q$ which is convex for $x \geq 0$. For a convex function $g : (0,\infty) \to (0,\infty)$, we know that $\forall x,\, y \in (0,\infty),\, \forall \theta \in [0,1]:\ g(\theta x + (1-\theta)y) \leq \theta g(x) + (1-\theta)g(y)$. Hence, with $\theta = \frac{1}{2}$ and $g(x) = x^q$, $x \geq 0$, we get

$$\begin{aligned}
&\Big(-y\log(\alpha) - (x-y)\log(1-\alpha) - \log(G(0))\Big)^q \\
&= \left(\theta\Big(2\left(-y\log(\alpha) - (x-y)\log(1-\alpha)\right)\Big) + (1-\theta)\left(2\left(-\log(G(0))\right)\right)\right)^q \\
&\leq \theta\Big(2(-y\log(\alpha) - (x-y)\log(1-\alpha))\Big)^q + (1-\theta)\left(2(-\log(G(0)))\right)^q \\
&= \frac{1}{2}\Big(2(-y\log(\alpha) - (x-y)\log(1-\alpha))\Big)^q + \frac{1}{2}\Big(2(-\log(G(0)))\Big)^q.
\end{aligned}$$

Similarly, we get

$$\Big(2(-y\log(\alpha) - (x-y)\log(1-\alpha))\Big)^q \leq \frac{1}{2}\Big(4(-y\log(\alpha))\Big)^q + \frac{1}{2}\Big(4(-(x-y)\log(1-\alpha))\Big)^q,$$

which altogether leads to

$$\begin{aligned}
&\Big(-y\log(\alpha) - (x-y)\log(1-\alpha) - \log(G(0))\Big)^q \\
&\leq \frac{1}{4}\Big(4(-y\log(\alpha))\Big)^q + \frac{1}{4}\Big(4(-(x-y)\log(1-\alpha))\Big)^q + \frac{1}{2}\Big(2(-\log(G(0)))\Big)^q
\end{aligned}$$

$$= 4^{q-1}\Big(- y\log(\alpha)\Big)^q + 4^{q-1}\Big(- (x-y)\log(1-\alpha)\Big)^q + 2^{q-1}\Big(- \log(G(0))\Big)^q.$$

In total, we have

$$\begin{aligned} I \leq & P\Bigg(\sum_{x,y\in\mathbb{N}_0, y\leq x} \Big[4^{q-1} y^q (-\log(\alpha))^q + 4^{q-1}(x-y)^q(-\log(1-\alpha))^q + 2^{q-1}(-\log(G(0)))^q \Big] \\ & \qquad P^*((X_0^*, X_1^*) = (x,y)) \geq k/2 \Bigg) \\ \leq & P\left(\sum_{x,y\in\mathbb{N}_0, y\leq x} 4^{q-1} y^q (-\log(\alpha))^q P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ & + P\left(\sum_{x,y\in\mathbb{N}_0, y\leq x} 4^{q-1} (x-y)^q (-\log(1-\alpha))^q P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ & + P\left(\sum_{x,y\in\mathbb{N}_0, y\leq x} 2^{q-1} (-\log(G(0)))^q P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ =: & I_1 + I_2 + I_3, \end{aligned}$$

where

$$\begin{aligned} I_1 &= P\left(4^{q-1}(-\log(\alpha))^q \sum_{x,y\in\mathbb{N}_0, y\leq x} y^q P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ &\leq P\left(4^{q-1}(-\log(\alpha))^q \sum_{x,y\in\mathbb{N}_0} y^q P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ &= P\Big(4^{q-1}(-\log(\alpha))^q \mathrm{E}^*((X_1^*)^q) \geq k/6 \Big), \\ I_2 &= P\left(4^{q-1}(-\log(1-\alpha))^q \sum_{x,y\in\mathbb{N}_0, y\leq x} (x-y)^q P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ &\leq P\left(4^{q-1}(-\log(1-\alpha))^q \sum_{x,y\in\mathbb{N}_0} |x-y|^q P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ &= P\Big(4^{q-1}(-\log(1-\alpha))^q \mathrm{E}^*(|X_0^* - X_1^*|^q) \geq k/6 \Big), \end{aligned}$$

and

$$\begin{aligned} I_3 &= P\left(2^{q-1}(-\log(G(0)))^q \sum_{x,y\in\mathbb{N}_0, y\leq x} P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ &\leq P\left(2^{q-1}(-\log(G(0)))^q \sum_{x,y\in\mathbb{N}_0} P^*((X_0^*, X_1^*) = (x,y)) \geq k/6 \right) \\ &= P\Big(2^{q-1}(-\log(G(0)))^q \geq k/6 \Big) = \mathbb{1}_{\{2^{q-1}(-\log(G(0)))^q \geq k/6\}}. \end{aligned}$$

Altogether, for the first term I, we have

$$I \leq P\left(4^{q-1}(-\log(\alpha))^q E^*((X_1^*)^q) \geq k/6\right) + P\left(4^{q-1}(-\log(1-\alpha))^q \mathrm{E}^*(|X_0^* - X_1^*|^q) \geq k/6\right) + \mathbb{1}_{\{2^{q-1}(-\log(G(0)))^q \geq k/6\}}.$$

Now, let us consider the second term II. We use similar arguments as employed above for deriving the bound for I. But instead of using the last summand (for $j = y$), we now use the first one (for $j = 0$) to construct an upper bound for II. Hence, we get

$$\begin{aligned}\left|\log\left(\sum_{j=0}^{x}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)\right)\right|^q &= \left(-\log\left(\sum_{j=0}^{x}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)\right)\right)^q \\ &\leq \Big(-\log((1-\alpha)^x G(y))\Big)^q \\ &= \Big(-x\log(1-\alpha) - \log(G(y))\Big)^q.\end{aligned}$$

Using Assumption 2, we have $G(y) \geq c_1 e^{-c_2 y}$ such that $-\log(G(y)) \leq -\log(c_1) + c_2 y$. Altogether, by using similar convexity arguments as used for term I, we get

$$\left|\log\left(\sum_{j=0}^{x}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)\right)\right|^q \leq 4^{q-1}\Big(-x\log(\alpha)\Big)^q + 4^{q-1}\Big(-\log(c_1)\Big)^q + 2^{q-1}\Big(c_2 y\Big)^q.$$

Analogously to the steps conducted for I, we get

$$\begin{aligned}II &\leq P\left(\sum_{x,y\in\mathbb{N}_0, y>x}\left[4^{q-1}(-x\log(\alpha))^q + 4^{q-1}(-\log(c_1))^q + 2^{q-1}(c_2 y)^q\right]P^*((X_0^*, X_1^*) = (x,y)) \geq k/2\right) \\ &\leq P\left(4^{q-1}(-\log(\alpha))^q \mathrm{E}^*((X_0^*)^q) \geq k/6\right) + \mathbb{1}_{\{(4^{q-1}(-\log(c_1))^q \geq k/6\}} + P\left(2^{q-1}c_2 \mathrm{E}^*((X_1^*)^q) \geq k/6\right).\end{aligned}$$

In summary, we showed

$$\begin{aligned}&P\left(\sum_{(x,y)\in\mathbb{N}_0^2}\left|\log\left(\sum_{j=0}^{\min(x,y)}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)\right)\right|^q P^*((X_0^*, X_1^*) = (x,y)) \geq k\right) \\ &\leq P\left(4^{q-1}(-\log(\alpha))^q \mathrm{E}^*((X_1^*)^q) \geq k/6\right) + P\left(4^{q-1}(-\log(1-\alpha))^q \mathrm{E}^*(|X_0^* - X_1^*|^q) \geq k/6\right) \\ &\quad + \mathbb{1}_{\{2^{q-1}(-\log(G(0)))^q \geq k/6\}} + P\left(4^{q-1}(-\log(\alpha))^q E^*((X_0^*)^q) \geq k/6\right) \\ &\quad + \mathbb{1}_{\{(4^{q-1}(-\log(c_1))^q \geq k/6\}} + P\left(2^{q-1}c_2 \mathrm{E}^*((X_1^*)^q) \geq k/6\right),\end{aligned}$$

where $c_1 \in (0,1]$ and $c_2 \in \mathbb{R}_+$. Due to the boundedness of the moments of the original observations, we conclude that the moments of the bootstrap innovations are also bounded with probability tending to one (see Lemma B.1 in Jentsch and Weiß, 2019). Using the Minkowski inequality, the same holds for their differences. With the previous arguments, the last right-hand side above becomes arbitrarily small for k chosen sufficiently large. Hence, (A.4) follows and, consequently, the covariances in (A.1) are summable. Hence, the right-hand side in (A.1)

is of order $O_P(\frac{1}{n})$ and vanishing in probability as $n \to \infty$. In summary, we have shown $\mathrm{Var}^*(M_n^*) \xrightarrow{p} 0$ as $n \to \infty$.

As shown by Drost et al. (2009a) in the proof of Theorem 2.1, for fixed $x_{-p}, \ldots, x_0 \in \mathbb{N}_0$, the map $\overline{E} \ni (\boldsymbol{\alpha}, G) \mapsto m^{\boldsymbol{\alpha},G}(x_{-p}, \ldots, x_0)$ is continuous because there appear only a finite number of $G(j)$'s in $P^{\boldsymbol{\alpha},G}_{(x_{-1},\ldots,x_{-p}),x_0}$. Moreover, for all $x_{-p}, \ldots, x_0 \in \mathbb{N}_0$, we have $m^{\boldsymbol{\alpha},G}(x_{-p}, \ldots, x_0) \leq \log(1) = 0$. Additionally, for the bootNPMLE, we have $M_n^*(\widehat{\boldsymbol{\alpha}}_n^*, \widehat{G}_n^*) \geq M_n^*(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$, because $(\widehat{\boldsymbol{\alpha}}_n^*, \widehat{G}_n^*)$ maximizes the likelihood by construction. Moreover, we need to show that the map $\overline{E} \ni (\boldsymbol{\alpha}, G) \mapsto M_n(\boldsymbol{\alpha}, G)$ has a unique maximum at $(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$ with probability tending to one. Since, in analogy to the identification argument used in Section B.2, part (C) of Drost et al. (2009b), with probability tending to one, we have the identification property (for general $p \in \mathbb{N}_0$), that is,

$$\begin{aligned}
&P^{\boldsymbol{\alpha},G}_{(X^*_{-1},\ldots,X^*_{-p}),X_0^*} = P^{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}_{(X^*_{-1},\ldots,X^*_{-p}),X_0^*} \quad \mathbb{P}_{\nu_{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n},\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}\text{-a.s. (in } \mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}\text{-probability)} \\
&\Rightarrow (\boldsymbol{\alpha}, G) = (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) \quad \mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}\text{-a.s.,} \qquad \text{(A.5)}
\end{aligned}$$

the assertion follows from

$$\begin{aligned}
&\mathrm{E}^*(M_n^*(\boldsymbol{\alpha}, G)) - \mathrm{E}^*(M_n^*(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) \\
&= \mathrm{E}^*\left(\frac{1}{n}\sum_{t=0}^{n} m^{\boldsymbol{\alpha},G}(X^*_{t-p}, \ldots, X^*_t)\right) - \mathrm{E}^*\left(\frac{1}{n}\sum_{t=0}^{n} m^{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}(X^*_{t-p}, \ldots, X^*_t)\right) \\
&= \frac{1}{n}\sum_{t=0}^{n} \mathrm{E}^*\left(\log P^{\boldsymbol{\alpha},G}_{(X^*_{t-1},\ldots,X^*_{t-p}),X^*_t}\right) - \frac{1}{n}\sum_{t=0}^{n} \mathrm{E}^*\left(\log P^{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}_{(X^*_{t-1},\ldots,X^*_{t-p}),X^*_t}\right) \\
&= \frac{1}{n}\, n\, \mathrm{E}^*\left(\log P^{\boldsymbol{\alpha},G}_{(X^*_{-1},\ldots,X^*_{-p}),X_0^*}\right) - \frac{1}{n}\, n\, \mathrm{E}^*\left(\log P^{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}_{(X^*_{-1},\ldots,X^*_{-p}),X_0^*}\right) \\
&\leq 2\, \mathrm{E}^*\left(\sqrt{\frac{P^{\boldsymbol{\alpha},G}_{(X^*_{-1},\ldots,X^*_{-p}),X_0^*}}{P^{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}_{(X^*_{-1},\ldots,X^*_{-p}),X_0^*}}} - 1\right) \\
&= 2\sum_{y\in\mathbb{N}_0^p} \nu_{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}}\{y\} \sum_{x_0=0}^{\infty} \sqrt{P^{\boldsymbol{\alpha},G}_{y,x_0} P^{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}_{y,x_0}} - 2 \\
&\leq -\sum_{y\in\mathbb{N}_0^p} \nu_{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}}\{y\} \sum_{x_0=0}^{\infty} \left(\sqrt{P^{\boldsymbol{\alpha},G}_{y,x_0}} - \sqrt{P^{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}_{y,x_0}}\right)^2 \\
&\leq 0 \quad \text{in } \mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}\text{-probability,}
\end{aligned}$$

where we used $\log(x) \leq 2(\sqrt{x} - 1)$ for $x \geq 0$ and the (conditional) stationarity of the bootstrap process. Note that the last inequality above holds in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability due to (A.5).

In summary, we showed that all conditions of Wald's consistency theorem (Wald, 1949) hold (in probability) and we obtain $d((\widehat{\boldsymbol{\alpha}}_n^*, \widehat{G}_n^*), (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)) \xrightarrow{p^*} 0$ in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability, which

immediately implies $\widehat{\boldsymbol{\alpha}}_n^* - \widehat{\boldsymbol{\alpha}}_n \xrightarrow{p^*} 0$ in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability as well as, for all $k \in \mathbb{N}_0$, $\widehat{G}_n^*(k) - \widehat{G}_n(k) \xrightarrow{p^*} 0$ in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability. □

Lemma A.1 (Fréchet derivative of Ψ_n)**.** *Let Assumption 1 hold true. For fixed n and for $\theta_0 = (\boldsymbol{\alpha}_0, G_0) \in \Theta \times \mathcal{G}$, the finite sample moment equations* (2.5) *and* (2.6)*, i.e., the maps $\Psi_n : (0,1)^p \times \widetilde{\mathcal{G}} \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$, are Fréchet differentiable with derivative $\dot{\Psi}_n^{\theta_0} : lin([0,1]^p \times \widetilde{\mathcal{G}}) \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ at θ_0 given by*

$$\dot{\Psi}_n^{\theta_0}(\theta - \theta_0) = \begin{pmatrix} \dot{\Psi}_{n11}^{\theta_0}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + \dot{\Psi}_{n12}^{\theta_0}(G - G_0) \\ \dot{\Psi}_{n21}^{\theta_0}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + \dot{\Psi}_{n22}^{\theta_0}(G - G_0) \end{pmatrix} \tag{A.6}$$

with $\dot{\Psi}_{n11}^{\theta_0} : \mathbb{R}^p \to \mathbb{R}^p, \dot{\Psi}_{n12}^{\theta_0} : lin(\mathcal{G}) \to \mathbb{R}^p, \dot{\Psi}_{n21}^{\theta_0} : \mathbb{R}^p \to \ell^\infty(\mathcal{H}_1)$ and $\dot{\Psi}_{n22}^{\theta_0} : lin(\mathcal{G}) \to \ell^\infty(\mathcal{H}_1)$ defined in (A.12) - (A.15)*, where $lin(\cdot)$ denotes the linear span. That is, for fixed n, we have*

$$\|\Psi_n(\theta) - \Psi_n(\theta_0) - \dot{\Psi}_n^{\theta_0}(\theta - \theta_0)\| = o_P\left(\|\theta - \theta_0\|\right) \tag{A.7}$$

and

$$\Psi_n(\theta) = \Psi_n(\theta_0) + \dot{\Psi}_n^{\xi}(\theta - \theta_0) \tag{A.8}$$

for some ξ, where $||\xi - \theta_0|| \leq ||\theta - \theta_0||$, where $\dot{\Psi}_n^{\theta_0} : lin([0,1]^p \times \widetilde{\mathcal{G}}) \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ is a continuous, linear mapping. Furthermore, as $\widehat{\theta}_n = (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) \in \Theta \times \mathcal{G}$ holds with P-probability tending to one as well as $\widehat{\theta}_n^* = (\widehat{\boldsymbol{\alpha}}_n^*, \widehat{G}_n^*) \in \Theta \times \mathcal{G}$ with P^*-probability tending to one (conditional on the data), by plugging-in $\widehat{\theta}_n$ for θ_0 and $\widehat{\theta}_n^*$ for θ in the above, we have

$$\|\Psi_n(\widehat{\theta}_n^*) - \Psi_n(\widehat{\theta}_n) - \dot{\Psi}_n^{\widehat{\theta}_n}(\widehat{\theta}_n^* - \widehat{\theta}_n)\| = o_{P^*}\left(\|\widehat{\theta}_n^* - \widehat{\theta}_n\|\right) \tag{A.9}$$

and

$$\Psi_n(\widehat{\theta}_n^*) = \Psi_n(\widehat{\theta}_n) + \dot{\Psi}_n^{\xi_n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \tag{A.10}$$

for some ξ_n, where $||\xi_n - \widehat{\theta}_n|| \leq ||\widehat{\theta}_n^* - \widehat{\theta}_n||$.

Proof. For notational convenience, we consider only the case of $p = 1$, but the following arguments can be extended to higher order $p > 1$. Let n be fixed, $\theta = (\alpha, G) = (\alpha, G(0), G(1), \ldots)$, $\theta_0 = (\alpha_0, G_0) = (\alpha_0, G_0(0), G_0(1), \ldots)$ and let $\Psi_n = (\Psi_{n1}, \Psi_{n2})$, where Ψ_{n1} and Ψ_{n2} are defined according to (2.5) and (2.6). Following Lemma B.2, its Fréchet derivative $\dot{\Psi}_n^{\theta_0}$ is defined as

$$\begin{aligned} \dot{\Psi}_n^{\theta_0}(\theta - \theta_0) &= \begin{pmatrix} \frac{\partial \Psi_{n1}(\alpha,G)}{\partial \alpha}|_{\theta=\theta_0}(\alpha - \alpha_0) + \sum_{k=0}^{\infty} \frac{\partial \Psi_{n1}(\alpha,G)}{\partial G(k)}|_{\theta=\theta_0}(G(k) - G_0(k)) \\ \frac{\partial \Psi_{n2}(\alpha,G)}{\partial \alpha}|_{\theta=\theta_0}(\alpha - \alpha_0) + \sum_{k=0}^{\infty} \frac{\partial \Psi_{n2}(\alpha,G)}{\partial G(k)}|_{\theta=\theta_0}(G(k) - G_0(k)) \end{pmatrix} \\ &=: \begin{pmatrix} \dot{\Psi}_{n11}^{\theta_0}(\alpha - \alpha_0) + \dot{\Psi}_{n12}^{\theta_0}(G - G_0) \\ \dot{\Psi}_{n21}^{\theta_0}(\alpha - \alpha_0) + \dot{\Psi}_{n22}^{\theta_0}(G - G_0) \end{pmatrix}, \end{aligned} \tag{A.11}$$

where

$$\dot{\Psi}^{\theta_0}_{n11}(\alpha-\alpha_0) = \left(\frac{1}{n}\sum_{t=0}^{n}\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \frac{1}{n}\sum_{t=0}^{n}\dot{l}^2_\alpha(X_{t-1},X_t;\alpha,G)\right)_{|_{\theta=\theta_0}}(\alpha-\alpha_0), \tag{A.12}$$

$$\begin{aligned}\dot{\Psi}^{\theta_0}_{n12}(G-G_0) = \sum_{k=0}^{\infty}\Bigg(&\frac{1}{n}\sum_{t=0}^{n}\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} \\ &- \frac{1}{n}\sum_{t=0}^{n}\dot{l}_\alpha(X_{t-1},X_t;\alpha,G)\frac{\frac{\partial}{\partial G(k)}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}}\Bigg)_{|_{\theta=\theta_0}}(G(k)-G_0(k))\end{aligned} \tag{A.13}$$

and, for $h \in \mathcal{H}_1$,

$$\begin{aligned}&\dot{\Psi}^{\theta_0}_{n21}(\alpha-\alpha_0)h \\ &= \Bigg(\frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty}h(j)\frac{\left(\frac{\partial}{\partial\alpha}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})\right)_{|_{(x_t,x_{t-1})=(X_t,X_{t-1})}}}{P^{\alpha,G}_{X_{t-1},X_t}} \\ &\quad - \frac{1}{n}\sum_{t=0}^{n}\dot{l}_\alpha(X_{t-1},X_t;\alpha,G)A_{\alpha,G}h(X_{t-1},X_t)\Bigg)_{|_{\theta=\theta_0}}(\alpha-\alpha_0),\end{aligned} \tag{A.14}$$

and

$$\begin{aligned}&\dot{\Psi}^{\theta_0}_{n22}(G-G_0)h \\ &= \sum_{k=0}^{\infty}\Bigg(\frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty}h(j)\frac{\left(\frac{\partial}{\partial G(k)}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})\right)_{|_{(x_t,x_{t-1})=(X_t,X_{t-1})}}}{P^{\alpha,G}_{X_{t-1},X_t}} - h(k) \\ &\quad - \frac{1}{n}\sum_{t=0}^{n}A_{\alpha,G}h(X_{t-1},X_t)\frac{\left(\frac{\partial}{\partial G(k)}P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})\right)_{|_{(x_t,x_{t-1})=(X_t,X_{t-1})}}}{P^{\alpha,G}_{X_{t-1},X_t})}\Bigg)_{|_{\theta=\theta_0}}(G(k)-\widehat{G}_n(k)),\end{aligned} \tag{A.15}$$

where $\dot{l}_\alpha$ and $A_{\alpha,G}h$ are defined in (2.7) and (2.8). Using similar arguments as in Drost et al. (2009a) to prove their Lemma 2(a) in Drost et al. (2009b), which makes also heavy use of inequalities derived in Drost et al. (2008), the assertion

$$\|\Psi_n(\theta) - \Psi_n(\theta_0) - \dot{\Psi}^{\theta_0}_n(\theta-\theta_0)\| = o_P\left(\|\theta-\theta_0\|\right).$$

follows.

Similarly, when plugging-in $\widehat{\theta}^*_n$ and $\widehat{\theta}_n$ for θ and θ_0, respectively, making use of Theorem 3.1, the same arguments lead to

$$\|\Psi_n(\widehat{\theta}^*_n) - \Psi_n(\widehat{\theta}_n) - \dot{\Psi}^{\widehat{\theta}_n}_n(\widehat{\theta}^*_n - \widehat{\theta}_n)\| = o_{P^*}\left(\|\widehat{\theta}^*_n - \widehat{\theta}_n\|\right).$$

□

Lemma A.2. *Suppose Assumptions 1 and 2 hold. Let $E(X_1^k) < \infty$ for some $k > 2(p+4)$ and $E((X_t^3(1+\rho)^{X_t}))^{1+\delta} < \infty$ for some $\rho, \delta > 0$. Then, we have*

$$\sqrt{n}\left(\dot{\Psi}_n^{\xi_n}(\widehat{\theta}_n^* - \widehat{\theta}_n)\right) - \sqrt{n}\left(\dot{\Psi}^{\theta_0}(\widehat{\theta}_n^* - \widehat{\theta}_n)\right) = o_{p^*}(1).$$

Recall that $\dot{\Psi}_n^{\xi_n} : lin([0,1]^p \times \widetilde{\mathcal{G}}) \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ and $\dot{\Psi}^{\theta_0} : lin([0,1]^p \times \widetilde{\mathcal{G}}) \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ are the Fréchet derivatives of Ψ_n at ξ_n and of Ψ at θ_0, respectively.

Proof. By adding and subtracting the Fréchet derivative of Ψ_n at θ_0, we get

$$\begin{aligned}
&||\sqrt{n}\Big(\dot{\Psi}_n^{\xi_n}(\widehat{\theta}_n^* - \widehat{\theta}_n)\Big) - \sqrt{n}\Big(\dot{\Psi}^{\theta_0}(\widehat{\theta}_n^* - \widehat{\theta}_n)\Big)|| \\
&= ||\dot{\Psi}_n^{\xi_n}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) \mp \dot{\Psi}_n^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) - \dot{\Psi}^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))|| \qquad \text{(A.16)} \\
&\leq ||\dot{\Psi}_n^{\xi_n}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) - \dot{\Psi}_n^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))|| + ||\dot{\Psi}_n^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) - \dot{\Psi}^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))||.
\end{aligned}$$

We examine both resulting differences separately. Let $\varepsilon, \delta > 0$ and consider the first term on the last right-hand side of (A.16). Then, we have

$$\begin{aligned}
&P(||\dot{\Psi}_n^{\xi_n}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) - \dot{\Psi}_n^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))|| > \varepsilon) \\
&= P(||\dot{\Psi}_n^{\xi_n}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) - \dot{\Psi}_n^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))|| > \varepsilon, ||\xi_n - \theta_0|| \leq \delta) \\
&\quad + P(||\dot{\Psi}_n^{\xi_n}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) - \dot{\Psi}_n^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))|| > \varepsilon, ||\xi_n - \theta_0|| > \delta) \\
&\leq P(||\dot{\Psi}_n^{\xi_n}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) - \dot{\Psi}_n^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))|| > \varepsilon, ||\xi_n - \theta_0|| \leq \delta) + P(||\xi_n - \theta_0|| > \delta),
\end{aligned}$$

where the last probability converges to zero for increasing n, because we have $||\xi_n - \widehat{\theta}_n|| \leq ||\widehat{\theta}_n^* - \widehat{\theta}_n|| \xrightarrow{p^*} 0$ in probability due to Theorem 3.1 as well as $\|\widehat{\theta}_n - \theta_0\| \xrightarrow{p} 0$ due to Theorem 2.1, which altogether gives $||\xi_n - \theta_0|| \xrightarrow{p^*} 0$ in probability. Hence, it remains to investigate the asymptotic behavior of

$$P(||\dot{\Psi}_n^{\xi_n}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)) - \dot{\Psi}_n^{\theta_0}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n))|| > \varepsilon, ||\xi_n - \theta_0|| \leq \delta).$$

For convenience, we will again consider only the case $p = 1$, but the same arguments can be used to prove the results for higher model order. Let us consider the first summand of the first entry of (A.11) and follow the calculations of the derivatives in the proof of Lemma A.1. Further, let $\xi_n = (\alpha_{\xi_n}, G_{\xi_n})$ and $\theta_0 = (\alpha_0, G_0)$. Then, we have

$$\begin{aligned}
&|\dot{\Psi}_{n11}^{\xi_n}(\sqrt{n}(\widehat{\alpha}_n^* - \widehat{\alpha}_n)) - \dot{\Psi}_{n11}^{\theta_0}(\sqrt{n}(\widehat{\alpha}_n^* - \widehat{\alpha}_n))| \\
&\leq \left| \frac{1}{n} \sum_{t=0}^{n} \left(\frac{\frac{\partial^2}{\partial \alpha^2} P_{X_{t-1},X_t|\alpha=\alpha_{\xi_n}}^{\alpha, G_{\xi_n}}}{P_{X_{t-1},X_t}^{\xi_n}} - \frac{\frac{\partial^2}{\partial \alpha^2} P_{X_{t-1},X_t|\alpha=\alpha_0}^{\alpha, G_0}}{P_{X_{t-1},X_t}^{\theta_0}} - \dot{l}_\alpha^2(X_{t-1}, X_t; \xi_n) + \dot{l}_\alpha^2(X_{t-1}, X_t; \theta_0) \right) \right| \\
&\qquad \times |\sqrt{n}(\widehat{\alpha}_n^* - \widehat{\alpha}_n)| \\
&\leq \left\{ \left| \frac{1}{n} \sum_{t=0}^{n} \left(\frac{\frac{\partial^2}{\partial \alpha^2} P_{X_{t-1},X_t|\alpha=\alpha_{\xi_n}}^{\alpha, G_{\xi_n}}}{P_{X_{t-1},X_t}^{\xi_n}} - \frac{\frac{\partial^2}{\partial \alpha^2} P_{X_{t-1},X_t|\alpha=\alpha_0}^{\alpha, G_0}}{P_{X_{t-1},X_t}^{\theta_0}} \right) \right|
\end{aligned}$$

$$+\left|\frac{1}{n}\sum_{t=0}^{n}\left(\dot{l}_\alpha^2(X_{t-1},X_t;\theta_0)-\dot{l}_\alpha^2(X_{t-1},X_t;\xi_n)\right)\right|\Bigg\}O_{P^*}(1)$$
$$=:(I+II)O_{P^*}(1). \tag{A.17}$$

Continuing with the second term II, using a binomial formula, it can be bounded by

$$\frac{1}{n}\sum_{t=0}^{n}\left|\dot{l}_\alpha(X_{t-1},X_t;\theta_0)+\dot{l}_\alpha(X_{t-1},X_t;\xi_n)\right|\left|\dot{l}_\alpha(X_{t-1},X_t;\theta_0)-\dot{l}_\alpha(X_{t-1},X_t;\xi_n)\right|. \tag{A.18}$$

Hence, using $||\xi_n-\theta_0||\leq\delta$, according to Lemma B.3, we get the bound

$$II\leq\frac{1}{n}\sum_{t=0}^{n}\left|\dot{l}_\alpha(X_{t-1},X_t;\theta_0)+\dot{l}_\alpha(X_{t-1},X_t;\xi_n)\right|$$
$$\delta\left(\frac{2X_{t-1}+CX_{t-1}^2\left(1+\rho\right)^{X_{t-1}-1}}{P^{\theta_0}_{X_{t-1},X_t}}+\frac{2X_{t-1}\left(CX_{t-1}\left(1+\rho\right)^{X_{t-1}-1}+1\right)}{P^{\theta_0}_{X_{t-1},X_t}P^{\xi_n}_{X_{t-1},X_t}}\right), \tag{A.19}$$

for some (generic) constant $C=C(\delta)$ and some $\rho=\rho(\delta)$, which becomes arbitrarily small for δ sufficiently small. Similarly, from Lemma B.4, we get

$$I\leq\frac{1}{n}\sum_{t=0}^{n}\delta\left(\frac{\widetilde{C}X_{t-1}^3(1+\widetilde{\rho})^{X_{t-1}-1}+\widetilde{C}X_{t-1}^2}{P^{\xi_n}_{X_{t-1},X_t}}+\frac{\widetilde{C}X_{t-1}^2\left(CX_{t-1}(1+\rho)^{X_{t-1}-1}+1\right)}{P^{\xi_n}_{X_{t-1},X_t}P^{\theta_0}_{X_{t-1},X_t}}\right). \tag{A.20}$$

With (A.19) and (A.20), we can now finally tackle (A.17) and get

$$P(|\dot{\Psi}^{\xi_n}_{n11}(\sqrt{n}(\widehat{\alpha}_n^*-\widehat{\alpha}_n))-\dot{\Psi}^{\theta_0}_{n11}(\sqrt{n}(\widehat{\alpha}_n^*-\widehat{\alpha}_n))|>\varepsilon,||\xi_n-\theta_0||\leq\delta)$$
$$\leq P\Bigg(\frac{1}{n}\sum_{t=0}^{n}\delta\left(\frac{\widetilde{C}X_{t-1}^3(1+\widetilde{\rho})^{X_{t-1}-1}+\widetilde{C}X_{t-1}^2}{P^{\xi_n}_{X_{t-1},X_t}}+\frac{\widetilde{C}X_{t-1}^2\left(CX_{t-1}(1+\rho)^{X_{t-1}-1}+1\right)}{P^{\xi_n}_{X_{t-1},X_t}P^{\theta_0}_{X_{t-1},X_t}}\right)$$
$$+\frac{1}{n}\sum_{t=0}^{n}|\dot{l}_\alpha(X_{t-1},X_t;\theta_0)+\dot{l}_\alpha(X_{t-1},X_t;\xi_n)|$$
$$\delta\left(\frac{2X_{t-1}+CX_{t-1}^2\left(1+\rho\right)^{X_{t-1}-1}}{P^{\theta_0}_{X_{t-1},X_t}}+\frac{2X_{t-1}\left(CX_{t-1}\left(1+\rho\right)^{X_{t-1}-1}+1\right)}{P^{\theta_0}_{X_{t-1},X_t}P^{\xi_n}_{X_{t-1},X_t}}\right)O_{P^*}(1)>\varepsilon\Bigg)$$
$$\leq P\left(\delta\left(\frac{1}{n}\sum_{t=0}^{n}w(X_t,X_{t-1};\rho)\right)O_{P^*}(1)>\varepsilon\right),$$

where we used similar techniques to deal with the factor $|\dot{l}_\alpha(X_{t-1},X_t;\theta_0)+\dot{l}_\alpha(X_{t-1},X_t;\xi_n)|$ and with $P^{\xi_n}_{X_{t-1},X_t}$ in the denominators. Using Assumption 2 and Theorem 14.1 of Davidson (1994), we can conclude that $w(X_t,X_{t-1};\rho)$ is also geometrically mixing. Bierens (1982) gives a weak law of law large numbers under geometric mixing and together with the moment assumptions we made we can conclude that $\forall\varepsilon>0\ \forall\gamma>0\ \exists\delta>0,\ \rho=\rho(\delta)>0:$ $P\left(\delta\frac{1}{n}\sum_{t=0}^{n}w(X_t,X_{t-1};\rho)>\varepsilon\right)<\gamma$.

So far, we showed the convergence result only for the first term of the first entry in (A.11) by proving Lipschitz-type continuity properties of $\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G}_{X_{t-1},X_t}$ and $\left(\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}\right)^2$. For the

second term of that first entry in (A.11), we can proceed analogously by showing with similar arguments the Lipschitz continuity of $\frac{\partial}{\partial G}\frac{\partial}{\partial \alpha}P^{\alpha,G}_{X_{t-1},X_t}$ and $\frac{\partial}{\partial \alpha}P^{\alpha,G}_{X_{t-1},X_t}\frac{\partial}{\partial G}P^{\alpha,G}_{X_{t-1},X_t}$ since

$$
\begin{aligned}
&|\dot{\Psi}^{\xi_n}_{n12}\big(\sqrt{n}(\widehat{G}^*_n-\widehat{G}_n)\big)-\dot{\Psi}^{\theta_0}_{n12}\big(\sqrt{n}(\widehat{G}^*_n-\widehat{G}_n)\big)|\\
=&\left|\sum_{k=0}^{\infty}\frac{1}{n}\sum_{t=0}^{n}\left(\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial \alpha}P^{\alpha,G}_{X_{t-1},X_t|(\alpha,G)=\xi_n}}{P^{\xi_n}_{X_{t-1},X_t}}-\dot{l}_\alpha(X_{t-1},X_t;\xi_n)\frac{\frac{\partial}{\partial G(k)}P^{\alpha_{\xi_n},G}_{X_{t-1},X_t|G=G_{\xi_n}}}{P^{\xi_n}_{X_{t-1},X_t}}\right)\big(\sqrt{n}(\widehat{G}^*_n(k)-\widehat{G}_n(k))\big)\right.\\
&\left.-\frac{1}{n}\sum_{t=0}^{n}\left(\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial \alpha}P^{\alpha,G}_{X_{t-1},X_t|(\alpha,G)=\theta_0}}{P^{\theta_0}_{X_{t-1},X_t}}-\dot{l}_\alpha(X_{t-1},X_t;\theta_0)\frac{\frac{\partial}{\partial G(k)}P^{\alpha_0,G}_{X_{t-1},X_t|G=G_0}}{P^{\theta_0}_{X_{t-1},X_t}}\right)\big(\sqrt{n}(\widehat{G}^*_n(k)-\widehat{G}_n(k))\big)\right|\\
\leq&\left(\sum_{k=0}^{\infty}\sqrt{n}|\widehat{G}^*_n(k)-\widehat{G}_n(k)|\right)\sum_{k=0}^{\infty}\left|\frac{1}{n}\sum_{t=0}^{n}\left(\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial \alpha}P^{\alpha,G}_{X_{t-1},X_t|(\alpha,G)=\xi_n}}{P^{\xi_n}_{X_{t-1},X_t}}-\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial \alpha}P^{\alpha,G}_{X_{t-1},X_t|(\alpha,G)=\theta_0}}{P^{\theta_0}_{X_{t-1},X_t}}\right.\right.\\
&\left.\left.+\dot{l}_\alpha(X_{t-1},X_t;\theta_0)\frac{\frac{\partial}{\partial G(k)}P^{\alpha_0,G}_{X_{t-1},X_t|G=G_0}}{P^{\theta_0}_{X_{t-1},X_t}}-\dot{l}_\alpha(X_{t-1},X_t;\xi_n)\frac{\frac{\partial}{\partial G(k)}P^{\alpha_{\xi_n},G}_{X_{t-1},X_t|G=G_{\xi_n}}}{P^{\xi_n}_{X_{t-1},X_t}}\right)\right|\\
\leq&O_{P^*}(1)\sum_{k=0}^{\infty}\left(\left|\frac{1}{n}\sum_{t=0}^{n}\left(\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial \alpha}P^{\alpha,G}_{X_{t-1},X_t|(\alpha,G)=\xi_n}}{P^{\xi_n}_{X_{t-1},X_t}}-\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial \alpha}P^{\alpha,G}_{X_{t-1},X_t|(\alpha,G)=\theta_0}}{P^{\theta_0}_{X_{t-1},X_t}}\right)\right|\right.\\
&\left.+\left|\frac{1}{n}\sum_{t=0}^{n}\left(\dot{l}_\alpha(X_{t-1},X_t;\theta_0)\frac{\frac{\partial}{\partial G(k)}P^{\alpha_0,G}_{X_{t-1},X_t|G=G_0}}{P^{\theta_0}_{X_{t-1},X_t}}-\dot{l}_\alpha(X_{t-1},X_t;\xi_n)\frac{\frac{\partial}{\partial G(k)}P^{\alpha_{\xi_n},G}_{X_{t-1},X_t|G=G_{\xi_n}}}{P^{\xi_n}_{X_{t-1},X_t}}\right)\right|\right).
\end{aligned}
$$

Then, using similar arguments as used for the first term of the first entry of (A.11), we get the claimed result. We omit the details.

Finally, for both terms of the second entry of $\Psi^{\theta_0}_n$, we require some Lipschitz continuity of $\sum_{j=0}^{\infty}\frac{\partial}{\partial \alpha}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})$, $\sum_{j=0}^{\infty}\frac{\partial}{\partial G}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})$ and $A_{\alpha,G}h(X_{t-1},X_t)$, respectively. Starting with $P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})$, we get the following closed form representation

$$
\begin{aligned}
&P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})=P^{\alpha,G}(\varepsilon_t=j,\alpha\circ X_{t-1}+\varepsilon_t=x_t|X_{t-1}=x_{t-1})\\
&=P^{\alpha,G}(\varepsilon_t=j,\alpha\circ X_{t-1}+j=x_t|X_{t-1}=x_{t-1})=P^{\alpha,G}(\varepsilon_t=j,\alpha\circ X_{t-1}=x_t-j|X_{t-1}=x_{t-1})\\
&=P^{\alpha,G}(\varepsilon_t=j|X_{t-1}=x_{t-1})P^{\alpha,G}(\alpha\circ X_{t-1}=x_t-j|X_{t-1}=x_{t-1})\\
&=P^{\alpha,G}(\varepsilon_t=j)P^{\alpha,G}(\alpha\circ x_{t-1}=x_t-j)\\
&=G(j)\binom{x_{t-1}}{x_t-j}\alpha^{x_t-j}(1-\alpha)^{x_{t-1}-(x_t-j)}\mathbb{1}_{j\in\{\max(0,x_t-x_{t-1}),\ldots,x_t\}},
\end{aligned}
\tag{A.21}
$$

where we used that ϵ_t and $\alpha\circ X_{t-1}$ are independent given X_{t-1}. Using the same arguments as above, we conclude Lipschitz continuity of both $\sum_{j=0}^{\infty}\frac{\partial}{\partial \alpha}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})$ and $\sum_{j=0}^{\infty}\frac{\partial}{\partial G}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})$. The same applies to argue Lipschitz continuity of

$A_{\alpha,G}h(X_{t-1},X_t)$ since, with (B.5) and (A.21), we have

$$\begin{aligned}
A_{\alpha,G}h(x_{t-1},x_t) &= \sum_{j=0}^{\infty} h(j)\frac{P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})}{P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})} \\
&= \sum_{j=0}^{\infty} h(j)\frac{G(j)\binom{x_{t-1}}{x_t-j}\alpha^{x_t-j}(1-\alpha)^{x_{t-1}-(x_t-j)}\mathbb{1}_{j\in\{\max(0,x_t-x_{t-1}),\dots,x_t\}}}{\sum_{s=0}^{\min(x_t,x_{t-1})}\binom{x_{t-1}}{s}\alpha^s(1-\alpha)^{x_{t-1}-s}G(x_t-s)} \\
&= \frac{\sum_{s=0}^{\min(x_t,x_{t-1})} h(x_t-s)\binom{x_{t-1}}{s}\alpha^s(1-\alpha)^{x_{t-1}-s}G(x_t-s)}{\sum_{s=0}^{\min(x_t,x_{t-1})}\binom{x_{t-1}}{s}\alpha^s(1-\alpha)^{x_{t-1}-s}G(x_t-s)},
\end{aligned} \tag{A.22}$$

which can be treated analogously.

Coming back to (A.16), it remains to investigate also the asymptotic behavior of the second difference, i.e., of

$$P(||\dot{\Psi}_n^{\theta_0}\big(\sqrt{n}(\widehat{\theta}_n^*-\widehat{\theta}_n)\big)-\dot{\Psi}^{\theta_0}\big(\sqrt{n}(\widehat{\theta}_n^*-\widehat{\theta}_n)\big)||>\widetilde{\varepsilon}). \tag{A.23}$$

Here again, we argue componentwise and apply the idea of the proof of weak of large numbers. For the first summand of the first entry of (A.11), we again have

$$P(|\dot{\Psi}_{n11}^{\theta_0}\big(\sqrt{n}(\widehat{\alpha}_n^*-\widehat{\alpha}_n)\big)-\dot{\Psi}_{11}^{\theta_0}\big(\sqrt{n}(\widehat{\alpha}_n^*-\widehat{\alpha}_n)\big)|>\widetilde{\varepsilon})\leq P(|\dot{\Psi}_{n11}^{\theta_0}-\dot{\Psi}_{11}^{\theta_0}|O_{P^*}(1)>\widetilde{\varepsilon})$$

and

$$\begin{aligned}
E_{\alpha_0,G_0}(\dot{\Psi}_{n11}^{\theta_0}) &= E_{\alpha_0,G_0}\left(\frac{1}{n}\sum_{t=0}^{n}\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_0}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t|\alpha=\alpha_0}}-\frac{1}{n}\sum_{t=0}^{n}\dot{l}_\alpha^2(X_{t-1},X_t;\alpha_0,G_0)\right) \\
&= E_{\alpha_0,G_0}\left(\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_0}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t|\alpha=\alpha_0}}\right)-E_{\alpha_0,G_0}\left(\dot{l}_\alpha^2(X_{t-1},X_t;\alpha_0,G_0)\right).
\end{aligned}$$

The second component corresponds to the term Drost et al. (2009a) derived for $\dot{\Psi}_{11}^{\theta_0}$ in case of $p=1$. For the first term, by dominated convergence, we have

$$\begin{aligned}
E_{\alpha_0,G_0}\left(\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_0}_{X_{-1},X_0|\alpha=\alpha_0}}{P^{\alpha_0,G_0}_{X_{-1},X_0}}\right) &= \sum_{m,l=0}^{\infty}\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_0}(X_0=m|X_{-1}=l)_{|\alpha=\alpha_0}}{P^{\alpha_0,G_0}(X_0=m|X_{-1}=l)}P^{\alpha_0,G_0}(X_0=m,X_{-1}=l) \\
&= \sum_{m,l=0}^{\infty}\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_0}(X_0=m|X_{-1}=l)_{|\alpha=\alpha_0}P^{\alpha_0,G_0}(X_{-1}=l) \\
&= \sum_{l=0}^{\infty}P^{\alpha_0,G_0}(X_{-1}=l)\sum_{m=0}^{\infty}\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_0}(X_0=m|X_{-1}=l)_{|\alpha=\alpha_0} \\
&= \sum_{l=0}^{\infty}P^{\alpha_0,G_0}(X_{-1}=l)\frac{\partial^2}{\partial\alpha^2}\sum_{m=0}^{\infty}P^{\alpha,G_0}(X_0=m|X_{-1}=l)_{|\alpha=\alpha_0} \\
&= \sum_{l=0}^{\infty}P^{\alpha_0,G_0}(X_{-1}=l)\frac{\partial^2}{\partial\alpha^2}1
\end{aligned}$$

$$= 0.$$

With the same arguments, we can treat the other three expressions to show that the corresponding first terms have mean zero. As a preliminary result, we proved that the expectations of the two terms in (A.23), i.e., of $\dot{\Psi}^{\theta_0}_{n11}$ and $\dot{\Psi}^{\theta_0}_{11}$, coincide. To conclude that the whole expression in (A.23) converges to zero, by Chebychev's inequality, it remains to show that

$$Var_{\alpha_0,G_0}(\dot{\Psi}^{\theta_0}_{n11}) \to 0 \quad \text{as } n \to \infty.$$

We proceed as in the proof of Theorem 3.1 and, analogously to (A.1), we get

$$\begin{aligned} &Var\left(\frac{\partial^2}{\partial\alpha^2}\frac{1}{n}\sum_{t=0}^{n} log\left(\sum_{j=0}^{\min\{X_{t-1},X_t\}}\binom{X_{t-1}}{j}\alpha^j(1-\alpha)^{X_{t-1}-j}G(X_t-j)\right)_{|\alpha=\alpha_0}\right) \\ &\leq \frac{1}{n}\sum_{h=-(n-1)}^{n-1}\text{Cov}\left(\frac{\partial^2}{\partial\alpha^2}\log\left(\sum_{j_1=0}^{\min(X_1,X_0)}\binom{X_0}{j_1}\alpha^{j_1}(1-\alpha)^{X_0-j_1}G(X_1-j_1)\right)_{|\alpha=\alpha_0},\right. \\ &\qquad\left.\frac{\partial^2}{\partial\alpha^2}\log\left(\sum_{j_2=0}^{\min(X_{h+1},X_h)}\binom{X_h}{j_2}\alpha^{j_2}(1-\alpha)^{X_h-j_2}G(X_{h+1}-j_2)\right)_{|\alpha=\alpha_0}\right). \quad \text{(A.24)} \end{aligned}$$

Consequently, by defining

$$Y_h := \frac{\partial^2}{\partial\alpha^2}\log\left(\sum_{j=0}^{\min(X_{h+1},X_h)}\binom{X_h}{j}\alpha^j(1-\alpha)^{X_h-j}G(X_{h+1}-j)\right)_{|\alpha=\alpha_0}, \quad \text{(A.25)}$$

and again using Assumption 2, Lemma B.1, Theorem 14.1 of Davidson (1994) and Corollary 14.3 of Davidson (1994), we get

$$|\text{Cov}(Y_0,Y_h)| \leq 2\left(2^{1-1/q}+1\right)\alpha_{\text{mix}}(h)^{1-1/q-1/r}\,||Y_0||_q||Y_h||_r. \quad \text{(A.26)}$$

Hence, it remains to show that

$$\exists q>1, r>q/(q-1) : ||Y_0||_q = O(1) \quad \text{and} \quad ||Y_h||_r = O(1), \quad \text{(A.27)}$$

where it again suffices to prove it for $h=0$ due to stationarity of $(X_t, t\in\mathbb{Z})$ and, for some $q>2$, we have to investigate

$$\sum_{(x,y)\in\mathbb{N}_0^2}\left|\frac{\partial^2}{\partial\alpha^2}\log\left(\sum_{j=0}^{\min(x,y)}\binom{x}{j}\alpha^j(1-\alpha)^{x-j}G(y-j)\right)_{|\alpha=\alpha_0}\right|^q P((X_0,X_1)=(x,y)). \quad \text{(A.28)}$$

Using that $q>1$ and $|a-b|\leq|a|+|b|$, we get an upper bound for (A.28) given by

$$\sum_{(x,y)\in\mathbb{N}_0^2}\left(\left|\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G}_{x,y|\alpha=\alpha_0}}{P^{\alpha,G}_{x,y}}\right| + \left|\frac{\frac{\partial}{\partial\alpha}P^{\alpha,G}_{x,y|\alpha=\alpha_0}}{P^{\alpha,G}_{x,y}}\right|^2\right)^q P((X_0,X_1)=(x,y)).$$

Using as in the Proof of Theorem 3.1 that x^q is a convex function for $x \geq 0$, that by (B.20) we have

$$\left| \frac{\partial^2}{\partial \alpha^2} P^{\alpha,G}_{x,y|\alpha=\alpha_0} \right| \leq 4x(x-1)(1-\alpha_0)^{-2}$$

and that by (B.16), we have

$$\left| \frac{\partial}{\partial \alpha} P^{\alpha,G}_{x,y|\alpha=\alpha_0} \right|^2 \leq 4x^2,$$

we get an upper bound for (A.28) given by

$$2^{3q-1}(1-\alpha_0)^{-2}\mathrm{E}((X_0(X_0-1))^q) + 2^{3q-1}\mathrm{E}((X_0^2)^q).$$

Using the boundedness of the moments of the observations together with Assumption 1, (A.27) follows and the covariance in (A.24) is summable. As in the proof of Theorem 3.1, the assertion that $Var_{\alpha_0,G_0}(\dot{\Psi}^{\theta_0}_{n11}) \to 0$ as $n \to \infty$ follows. For the other components, we proceed analogously. Thus, altogether, (A.23) converges to zero in probability. □

Lemma A.3. *Suppose Assumptions 1 and 2 hold true. Let $E(X_1^k) < \infty$ for some $k > 2(p+4)$ and $E((X_t^3(1+\rho)^{X_t}))^{1+\delta} < \infty$ for some $\rho, \delta > 0$. Then, we have*

$$\sqrt{n}(\Psi^*_n - \Psi_n)(\widehat{\theta}^*_n) - \sqrt{n}(\Psi^*_n - \Psi_n)(\widehat{\theta}_n) = o_{P^*}(1)$$

in probability.

Proof. Following the proof of (L4) in Drost et al. (2009b), we consider both components separately. That is, we have to show

$$\sqrt{n}(\Psi^*_{n1} - \Psi_{n1})(\widehat{\theta}^*_n) - \sqrt{n}(\Psi^*_{n1} - \Psi_{n1})(\widehat{\theta}_n) = o_{P^*}(1), \tag{A.29}$$

$$\sqrt{n}(\Psi^*_{n2} - \Psi_{n2})(\widehat{\theta}^*_n) - \sqrt{n}(\Psi^*_{n2} - \Psi_{n2})(\widehat{\theta}_n) = o_{P^*}(1) \tag{A.30}$$

in probability, respectively. We consider only (A.30) for $p = 1$ and omit the details for (A.29). Then, for $h \in \mathcal{H}_1$, we have

$$\begin{aligned}
&\sqrt{n}\big(\Psi^*_{n2}(\widehat{\theta}^*_n)h - \Psi_{n2}(\widehat{\theta}^*_n)h - \Psi^*_{n2}(\widehat{\theta}_n)h + \Psi_{n2}(\widehat{\theta}_n)h\big) \\
=&\sqrt{n}\left(\frac{1}{n}\sum_{t=0}^{n}\left(A_{\widehat{\alpha}^*,\widehat{G}^*}h(X^*_{t-1},X^*_t) - \int h\, d\widehat{G}^*\right) - \frac{1}{n}\sum_{t=0}^{n}\left(A_{\widehat{\alpha}^*,\widehat{G}^*}h(X_{t-1},X_t) - \int h d\widehat{G}^*\right)\right. \\
&\left. - \frac{1}{n}\sum_{t=0}^{n}\left(A_{\widehat{\alpha},\widehat{G}}h(X^*_{t-1},X^*_t) - \int h\, d\widehat{G}\right) + \frac{1}{n}\sum_{t=0}^{n}\left(A_{\widehat{\alpha},\widehat{G}}h(X_{t-1},X_t) - \int h d\widehat{G}\right)\right) \\
=&\frac{1}{\sqrt{n}}\sum_{t=0}^{n}\left(A_{\widehat{\alpha}^*,\widehat{G}^*}h(X^*_{t-1},X^*_t) - A_{\widehat{\alpha}^*,\widehat{G}^*}h(X_{t-1},X_t) - A_{\widehat{\alpha},\widehat{G}}h(X^*_{t-1},X^*_t) + A_{\widehat{\alpha},\widehat{G}}h(X_{t-1},X_t)\right)
\end{aligned}$$

where, according to (A.22) and for $x_{t-1}, x_t \in \mathbb{N}_0$ and $(\alpha, G) \in \Theta \times \mathcal{G}$, we have

$$A_{\alpha,G}h(x_{t-1}, x_t) = \frac{\sum_{s=0}^{\min(x_t,x_{t-1})} h(x_t - s)\binom{x_{t-1}}{s}\alpha^s(1-\alpha)^{x_{t-1}-s}G(x_t - s)}{\sum_{s=0}^{\min(x_t,x_{t-1})} \binom{x_{t-1}}{s}\alpha^s(1-\alpha)^{x_{t-1}-s}G(x_t - s)}.$$

Hence, we have

$$\begin{aligned}&\sqrt{n}\big(\Psi^*_{n2}(\widehat{\theta}^*_n) - \Psi_{n2}(\widehat{\theta}^*_n) - \Psi^*_{n2}(\widehat{\theta}_n) + \Psi_{n2}(\widehat{\theta}_n)\big)\\ =&\frac{1}{\sqrt{n}}\sum_{t=0}^{n}\left(\frac{d^*_{h,n}(\widehat{\theta}^*_n)}{d^*_n(\widehat{\theta}^*_n)} - \frac{d_{h,n}(\widehat{\theta}^*_n)}{d_n(\widehat{\theta}^*_n)} - \frac{d^*_{h,n}(\widehat{\theta}_n)}{d^*_n(\widehat{\theta}_n)} + \frac{d_{h,n}(\widehat{\theta}_n)}{d_n(\widehat{\theta}_n)}\right),\end{aligned} \tag{A.31}$$

where

$$\begin{aligned}d_{h,n}(\widehat{\theta}_n) &= \sum_{s=0}^{\min(X_t,X_{t-1})} h(X_t - s)\binom{X_{t-1}}{s}\widehat{\alpha}^s(1-\widehat{\alpha})^{X_{t-1}-s}\widehat{G}(X_t - s),\\ d_n(\widehat{\theta}_n) &= \sum_{s=0}^{\min(X_t,X_{t-1})} \binom{X_{t-1}}{s}\widehat{\alpha}^s(1-\widehat{\alpha})^{X_{t-1}-s}\widehat{G}(X_t - s),\\ d^*_{h,n}(\widehat{\theta}^*_n) &= \sum_{s=0}^{\min(X_t,X_{t-1})} h(X_t - s)\binom{X_{t-1}}{s}\widehat{\alpha}^{*s}(1-\widehat{\alpha}^*)^{X_{t-1}-s}\widehat{G}^*(X_t - s),\\ d^*_n(\widehat{\theta}^*_n) &= \sum_{s=0}^{\min(X_t,X_{t-1})} \binom{X_{t-1}}{s}\widehat{\alpha}^{*s}(1-\widehat{\alpha}^*)^{X_{t-1}-s}\widehat{G}^*(X_t - s)\end{aligned}$$

and $d_{h,n}(\widehat{\theta}^*_n), d_n(\widehat{\theta}^*_n), d^*_{h,n}(\widehat{\theta}_n)$ and $d^*_n(\widehat{\theta}_n)$ analog. Finally, by adding suitable zero and using the same technique as in the proof of Lemma B.3, we can bound the above expression (A.31) by $\sqrt{n}\|\widehat{\theta}^*_n - \widehat{\theta}_n\|$, which is $O_{P^*}(1)$ multiplied by a term that converges to zero in probability.

□

Lemma A.4. *Suppose Assumptions 1 and 2 hold. Let $E(X_1^k) < \infty$ for some $k > 2(p+4)$ and $E((X_t^3(1+\rho)^{X_t}))^{1+\delta} < \infty$ for some $\rho, \delta > 0$. Then, we have*

$$\mathcal{S}_n^{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n,*} = \sqrt{n}\left(\Psi^*_n(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) - \Psi_n(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)\right) \rightsquigarrow^* \mathcal{S}^{\boldsymbol{\alpha}_0,G_0} \tag{A.32}$$

in $\mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$, under $P^*(\cdot) = \mathbb{P}_{\nu_{\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n},\widehat{\boldsymbol{\alpha}}_n,\widehat{G}_n}(\cdot|\mathbb{X})$ in $\mathbb{P}_{\nu_{\theta_0},\theta_0}$-probability, where $\mathcal{S}^{\boldsymbol{\alpha}_0,G_0}$ is the tight, Borel measurable, Gaussian process obtained in (2.11).

Proof. Note that $\mathrm{E}^*(\Psi^*_n(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)) = \Psi_n(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)$ and by similar techniques as used in Lemma A.3, we get

$$\sqrt{n}\big(\Psi^*_n(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) - \Psi_n(\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n)\big) = \sqrt{n}\big(\Psi^*_n(\boldsymbol{\alpha}_0, G_0) - \mathrm{E}^*(\Psi^*_n(\boldsymbol{\alpha}_0, G_0))\big) + o_{P^*}(1).$$

Hence, recalling the definition of $\Psi^*_n(\boldsymbol{\alpha}, G)$ in (3.2) and (3.3), the leading term on the last right-hand side can be represented as a function of generalized means. That is, we have

$$\sqrt{n}\big(\Psi^*_n(\boldsymbol{\alpha}_0, G_0) - \mathrm{E}^*(\Psi^*_n(\boldsymbol{\alpha}_0, G_0))\big)$$

$$=\sqrt{n}\left(f\left(\frac{1}{n_m}\sum_{t=1}^{n_m} g(X_t^*,\ldots,X_{t+m-1}^*)\right) - f\left(\mathrm{E}^*\left(\frac{1}{n_m}\sum_{t=1}^{n_m} g(X_t^*,\ldots,X_{t+m-1}^*)\right)\right)\right),$$

where f is the identity function, $m = p+1$, $n_m = n-p-1$ and $g = (g_1, g_2)$ with

$$g_1(x_t,\ldots,x_{t-p}) = \dot{l}_{\boldsymbol{\alpha}}(x_{t-p},\ldots,x_t;\boldsymbol{\alpha}_0,G_0)$$

and

$$g_2(x_t,\ldots,x_{t-p})h = A_{\boldsymbol{\alpha}_0 G_0} h(x_{t-p},\ldots,x_t) - \int h dG_0, \quad h \in \mathcal{H}_1$$

for $x_t,\ldots,x_{t-p} \in \mathbb{N}_0$. This representation as a function of generalized means allows to use Corollary 4.2 in Jentsch and Weiß (2019). Finally, for allowing the application of this result, according to Assumption 1 in Jentsch and Weiß (2019), sufficient smoothness properties have to be fulfilled. As f is just the identity, all smoothness properties hold. It remains to argue that all partial derivatives of $g_1(x_t,\ldots,x_{t-p})$ and $g_2(x_t,\ldots,x_{t-p})$ with respect to $x_t,\ldots,x_{t-p}$ are Lipschitz continuous. However, as both function g_1 and g_2 ask for arguments from $\mathbb{N}_0^{p+1}$ only, they can be arbitrarily extended to functions on $\mathbb{R}^{p+1}$ such that all sufficient smoothness conditions will be fulfilled by construction. □

Appendix B. Auxiliary Results

Lemma B.1 (NPMLE fulfills $\widehat{\theta}_n = (\widehat{\boldsymbol{\alpha}}_n, \widehat{G}_n) \in \Theta \times \mathcal{G}$ in probability)**.** *Suppose Assumption 1 holds true. Let $E(X_1^k) < \infty$ for some $k > 2(p+4)$. Then, with $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability tending to one, it holds that $\widehat{G}_n \in \mathcal{G} = \{G \in \widetilde{\mathcal{G}} : 0 < G(0) < 1 : E_G(\varepsilon_t^{p+4}) < \infty\}$ and $\widehat{\boldsymbol{\alpha}}_n \in \Theta = \{\boldsymbol{\alpha} \in (0,1)^p : \sum_{i=1}^p \alpha_i < 1\}$, respectively. Note that $\widehat{G}_n = \mathcal{L}^*(\varepsilon_t^*)$ and $E^*({\varepsilon_t^*}^{p+4}) = \sum_{k=0}^\infty k^{p+4}\widehat{G}_n(k)$ for the bootstrap procedure described in Section 3.1*

Proof. We divide the proof into four parts. In the following, we will prove

(i) $0 < \widehat{G}_n(0) < 1$ in $P = \mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability, that is, with $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability tending to one,

(ii) $\sum_{k=0}^\infty k^{p+4}\widehat{G}_n(k) = O_P(1)$,

(iii) $\widehat{\boldsymbol{\alpha}}_n \in (0,1)^p$ in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability,

(iv) $\sum_{i=1}^p \widehat{\alpha}_{n,i} < 1$ in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability.

For showing part (i), we make use of the equivalence

$$P(0 < \widehat{G}_n(0) < 1) \underset{n\to\infty}{\to} 1 \quad \Leftrightarrow \quad P(\widehat{G}_n(0) \in \{0,1\}) \underset{n\to\infty}{\to} 0.$$

Further, we have

$$\begin{aligned} P(\widehat{G}_n(0) \in \{0,1\}) &= P\left(\widehat{G}_n(0) \in \{0,1\}, |\widehat{G}_n(0) - G_0(0)| < \frac{G_0(0)}{2}\right) \\ &\quad + P\left(\widehat{G}_n(0) \in \{0,1\}, |\widehat{G}_n(0) - G_0(0)| \geq \frac{G_0(0)}{2}\right) \\ &=: \mathrm{I} + \mathrm{II}, \end{aligned}$$

where

$$\mathrm{I} = P\left(\left\{\widehat{G}_n(0) \in \{0,1\}\right\} \cap \left\{\frac{-G_0(0)}{2} < \widehat{G}_n(0) - G_0(0) < \frac{G_0(0)}{2}\right\}\right) = P(\emptyset) = 0$$

and

$$\mathrm{II} \leq P\left(|\widehat{G}_n(0) - G_0(0)| \geq \frac{G_0(0)}{2}\right) \underset{n\to\infty}{\to} 0$$

according to Theorem 2.1. For proving part (ii), we have

$$\begin{aligned} &\sum_{k=0}^{\infty} k^{p+4}\widehat{G}_n(k) \\ &= \sum_{k=0}^{\max(X_1,\ldots,X_n)} k^{p+4}\widehat{G}_n(k) \\ &= \sum_{k=0}^{\max(X_1,\ldots,X_n)} k^{p+4}\big(\widehat{G}_n(k) - G_0(k)\big) + \sum_{k=0}^{\max(X_1,\ldots,X_n)} k^{p+4}G_0(k) \end{aligned}$$

$$= \sum_{k=0}^{\max(X_1,\ldots,X_n)} k^{p+4}\big(\widehat{G}_n(k) - G_0(k)\big) + \sum_{k=0}^{\infty} k^{p+4} G_0(k) - \sum_{k=\max(X_1,\ldots,X_n)+1}^{\infty} k^{p+4} G_0(k)$$

$$=: I + II + III.$$

For term I, we have

$$I = \sum_{k=0}^{\max(X_1,\ldots,X_n)} k^{p+4}\big(\widehat{G}_n(k) - G_0(k)\big) \leq \max(X_1,\ldots,X_n)^{p+4} \sum_{k=0}^{\max(X_1,\ldots,X_n)} \big(\widehat{G}_n(k) - G_0(k)\big).$$

With the use of $\sqrt{n}\sum_{k=0}^{\infty}(\widehat{G}_n(k) - G_0(k)) = O_p(1)$ (see Theorem 2.2), the latter becomes

$$\frac{\max(X_1,\ldots,X_n)^{p+4}}{\sqrt{n}} \left(\sqrt{n}\sum_{k=0}^{\infty}\big(\widehat{G}_n(k) - G_0(k)\big)\right) = O_p\left(\frac{\max(X_1,\ldots,X_n)^{p+4}}{\sqrt{n}}\right).$$

Further, for all $x > 0$ and all $k \geq 0$, from Markov's inequality, we get

$$P\left(\frac{\max(X_1,\ldots,X_n)^{p+4}}{\sqrt{n}} > x\right) = P\left(\max(X_1,\ldots,X_n) > x^{1/(p+4)} n^{1/(2(p+4))}\right)$$

$$= P\left(\bigcup_{i=1}^{n}\left\{X_i > x^{1/(p+4)} n^{1/(2(p+4))}\right\}\right) \leq \sum_{i=1}^{n} P\left(X_i > x^{1/(p+4)} n^{1/(2(p+4))}\right)$$

$$= n \cdot P\left(X_1 > x^{1/(p+4)} n^{1/(2(p+4))}\right) \leq \frac{n \cdot \mathrm{E}(X_1^k)}{x^{k/(p+4)} n^{k/(2(p+4))}}$$

$$= \frac{\mathrm{E}(X_1^k)}{x^{k/(p+4)}} \cdot n^{1-k/(2(p+4))},$$

which converges to zero for $k > 2(p+4)$ if $\mathrm{E}(X_1^k) < \infty$. This implies $\frac{\max(X_1,\ldots,X_n)^{p+4}}{\sqrt{n}} = o_p(1)$ such that $I = o_P(1)$ holds. As $II < \infty$, i.e., $II = O(1)$, by Assumption 1, the assertion follows from $III = o_P(1)$, which we show next. Here, we distinguish two cases. If the support of $(\varepsilon_t, t \in \mathbb{Z})$ is bounded, then $\sum_{k=\max(X_1,\ldots,X_n)+1}^{\infty} k^{p+4} G_0(k) = 0$ after one observation X_t attains a value greater or equal to the largest possible innovation which happens with probability tending to one. If the support is unbounded, we have $\max(X_1,\ldots,X_n) \to \infty$ such that $\sum_{k=\max(X_1,\ldots,X_n)+1}^{\infty} k^{p+4} G_0(k) = o_P(1)$ as again by Assumption 1, we have $\mathrm{E}_{G_0}(\varepsilon_t^{p+4}) < \infty$, i.e., $(k^{p+4}\, G_0(k), k \in \mathbb{N}_0)$ is summable. For part (iii), we have that $\widehat{\boldsymbol{\alpha}}_n \in (0,1)^p$ if and only if $\widehat{\alpha}_{n,i} \in (0,1)$ for all $i = 1,\ldots,p$. Hence, the proof is analogous to the proof of part (i).

Similarly, we can show that $\sum_{i=1}^{p} \widehat{\alpha}_{n,i} < 1$ holds in $\mathbb{P}_{\nu_{\boldsymbol{\alpha}_0,G_0},\boldsymbol{\alpha}_0,G_0}$-probability, because

$$\sum_{i=1}^{p} \widehat{\alpha}_{n,i} = \sum_{i=1}^{p} (\widehat{\alpha}_{n,i} - \alpha_{0,i}) + \sum_{i=1}^{p} \alpha_{0,i},$$

where the first term converges to zero in probability due to Theorem 2.1 and the second term is smaller than 1 according to Assumption 1. □

Lemma B.2 (Partial derivatives of Ψ_n)**.** *Under the assumptions of Lemma A.1, for $p = 1$ and for all $k \in \mathbb{N}_0$, we have*

$$\frac{\partial \Psi_{n1}(\alpha, G)}{\partial \alpha} = \frac{1}{n}\sum_{t=0}^{n} \frac{\frac{\partial^2}{\partial \alpha^2} P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \frac{1}{n}\sum_{t=0}^{n} \dot{l}^2_\alpha(X_{t-1}, X_t; \alpha, G), \tag{B.1}$$

$$\frac{\partial \Psi_{n1}(\alpha, G)}{\partial G(k)} = \frac{1}{n}\sum_{t=0}^{n} \frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial \alpha} P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \frac{1}{n}\sum_{t=0}^{n} \dot{l}_\alpha(X_{t-1}, X_t; \alpha, G)\frac{\frac{\partial}{\partial G(k)} P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} \tag{B.2}$$

and, for $h \in \mathcal{H}_1$,

$$\begin{aligned}\frac{\partial \Psi_{n2}(\alpha, G)}{\partial \alpha} h = {} & \frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty} h(j) \frac{\left(\frac{\partial}{\partial \alpha} P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})\right)_{|_{(x_t,x_{t-1})=(X_t,X_{t-1})}}}{P^{\alpha,G}_{X_{t-1},X_t}} \\ & - \frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty} h(j)\dot{l}_\alpha(X_{t-1}, X_t; \alpha, G) A_{\alpha,G} h(X_{t-1}, X_t),\end{aligned} \tag{B.3}$$

$$\begin{aligned}\frac{\partial \Psi_{n2}(\alpha, G)}{\partial G(k)} h = {} & \frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty} h(j) \frac{\left(\frac{\partial}{\partial G(k)} P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})\right)_{|_{(x_t,x_{t-1})=(X_t,X_{t-1})}}}{P^{\alpha,G}_{X_{t-1},X_t}} \\ & - \frac{1}{n}\sum_{t=0}^{n} A_{\alpha,G} h(X_{t-1}, X_t) \frac{\left(\frac{\partial}{\partial G(k)} P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})\right)_{|_{(x_t,x_{t-1})=(X_t,X_{t-1})}}}{P^{\alpha,G}_{X_{t-1},X_t}} \\ & - h(k).\end{aligned} \tag{B.4}$$

Proof. For the first two derivatives in (B.1) and (B.2), recall from (2.5) that

$$\Psi_{n1}(\alpha, G) = \frac{1}{n}\sum_{t=0}^{n} \dot{l}_\alpha(X_{t-1}, X_t; \alpha, G),$$

where

$$\dot{l}_\alpha(x_{t-1}, x_t; \alpha, G) = \frac{\partial}{\partial \alpha} \log\left(P^{\alpha,G}_{x_{t-1},x_t}\right).$$

Hence, for (B.1), we get

$$\begin{aligned}\frac{\partial \Psi_{n1}(\alpha, G)}{\partial \alpha} &= \frac{\partial}{\partial \alpha}\frac{1}{n}\sum_{t=0}^{n} \dot{l}_\alpha(X_{t-1}, X_t; \alpha, G) = \frac{\partial}{\partial \alpha}\frac{1}{n}\sum_{t=0}^{n} \frac{\partial}{\partial \alpha} \log\left(P^{\alpha,G}_{X_{t-1},X_t}\right) \\ &= \frac{1}{n}\sum_{t=0}^{n} \frac{\partial^2}{\partial \alpha^2} \log\left(P^{\alpha,G}_{X_{t-1},X_t}\right) = \frac{1}{n}\sum_{t=0}^{n} \frac{P^{\alpha,G}_{X_{t-1},X_t}\left(\frac{\partial^2}{\partial \alpha^2} P^{\alpha,G}_{X_{t-1},X_t}\right) - \left(\frac{\partial}{\partial \alpha} P^{\alpha,G}_{X_{t-1},X_t}\right)^2}{\left(P^{\alpha,G}_{X_{t-1},X_t}\right)^2} \\ &= \frac{1}{n}\sum_{t=0}^{n} \left(\frac{\frac{\partial^2}{\partial \alpha^2} P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \left(\frac{\frac{\partial}{\partial \alpha} P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}}\right)^2\right) \\ &= \frac{1}{n}\sum_{t=0}^{n} \left(\frac{\frac{\partial^2}{\partial \alpha^2} P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \dot{l}^2_\alpha(X_{t-1}, X_t; \alpha, G)\right)\end{aligned}$$

$$= \frac{1}{n}\sum_{t=0}^{n}\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \frac{1}{n}\sum_{t=0}^{n} i^2_\alpha(X_{t-1},X_t;\alpha,G).$$

Similarly, for (B.2) and for all $k \in \mathbb{N}_0$, we have

$$\begin{aligned}
\frac{\partial\Psi_{n1}(\alpha,G)}{\partial G(k)} &= \frac{\partial}{\partial G(k)}\frac{1}{n}\sum_{t=0}^{n} i_\alpha(X_{t-1},X_t;\alpha,G) = \frac{1}{n}\sum_{t=0}^{n}\frac{\partial}{\partial G(k)}\frac{\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} \\
&= \frac{1}{n}\sum_{t=0}^{n}\frac{\left(\frac{\partial}{\partial G(k)}\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}\right)P^{\alpha,G}_{X_{t-1},X_t} - \left(\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}\right)\left(\frac{\partial}{\partial G(k)}P^{\alpha,G}_{X_{t-1},X_t}\right)}{\left(P^{\alpha,G}_{X_{t-1},X_t}\right)^2} \\
&= \frac{1}{n}\sum_{t=0}^{n}\left(\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \frac{\left(\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}\right)\left(\frac{\partial}{\partial G(k)}P^{\alpha,G}_{X_{t-1},X_t}\right)}{\left(P^{\alpha,G}_{X_{t-1},X_t}\right)^2}\right) \\
&= \frac{1}{n}\sum_{t=0}^{n}\left(\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \dot{l}_\alpha(X_{t-1},X_t;\alpha,G)\frac{\frac{\partial}{\partial G(k)}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}}\right) \\
&= \frac{1}{n}\sum_{t=0}^{n}\frac{\frac{\partial}{\partial G(k)}\frac{\partial}{\partial\alpha}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}} - \frac{1}{n}\sum_{t=0}^{n}\dot{l}_\alpha(X_{t-1},X_t;\alpha,G)\frac{\frac{\partial}{\partial G(k)}P^{\alpha,G}_{X_{t-1},X_t}}{P^{\alpha,G}_{X_{t-1},X_t}}.
\end{aligned}$$

For the last two derivatives (B.3) and (B.4), recall from (2.6) that

$$\Psi_{n2}(\alpha,G)h = \frac{1}{n}\sum_{t=0}^{n}\left(A_{\alpha,G}h(X_{t-1},X_t) - \int h dG\right), \quad h \in \mathcal{H}_1,$$

holds, where

$$\begin{aligned}
A_{\alpha,G}h(x_{t-1},x_t) =& \mathrm{E}_{\alpha,G}(h(\varepsilon_t)|X_t = x_t, X_{t-1} = x_{t-1}) \\
=& \sum_{j=0}^{\infty} h(j)P^{\alpha,G}(\varepsilon_t = j|X_t = x_t, X_{t-1} = x_{t-1}) \\
=& \sum_{j=0}^{\infty} h(j)\frac{P^{\alpha,G}(\varepsilon_t = j, X_t = x_t, X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_t = x_t, X_{t-1} = x_{t-1})}\frac{P^{\alpha,G}(X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_{t-1} = x_{t-1})} \\
=& \sum_{j=0}^{\infty} h(j)\frac{P^{\alpha,G}(\varepsilon_t = j, X_t = x_t, X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_t = x_t|X_{t-1} = x_{t-1})P^{\alpha,G}(X_{t-1} = x_{t-1})} \\
=& \sum_{j=0}^{\infty} h(j)\frac{P^{\alpha,G}(\varepsilon_t = j, X_t = x_t|X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_t = x_t|X_{t-1} = x_{t-1})}. \qquad \text{(B.5)}
\end{aligned}$$

Hence, for (B.3), we get

$$\frac{\partial\Psi_{n2}(\alpha,G)}{\partial\alpha}h = \frac{\partial}{\partial\alpha}\frac{1}{n}\sum_{t=0}^{n}\left(A_{\alpha,G}h(X_{t-1},X_t) - \int h dG\right) = \frac{1}{n}\sum_{t=0}^{n}\frac{\partial}{\partial\alpha}A_{\alpha,G}h(X_{t-1},X_t),$$

where, for all $x_t, x_{t-1} \in \mathbb{N}_0$, using (B.5), we have

$$\frac{\partial}{\partial\alpha}A_{\alpha,G}h(x_{t-1},x_t)$$

$$
\begin{aligned}
&= \sum_{j=0}^{\infty} h(j) \frac{\partial}{\partial \alpha} \frac{P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})} \\
&= \sum_{j=0}^{\infty} h(j) \Bigg[\frac{\left(\frac{\partial}{\partial \alpha} P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})\right) P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})}{\left(P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})\right)^2} \\
&\quad - \frac{\left(\frac{\partial}{\partial \alpha} P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})\right) P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})}{\left(P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})\right)^2} \Bigg] \\
&= \sum_{j=0}^{\infty} h(j) \frac{\frac{\partial}{\partial \alpha} P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})} \\
&\quad - \sum_{j=0}^{\infty} h(j) \dot{l}_\alpha(x_{t-1}, x_t; \alpha, G) \frac{P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})} \\
&= \sum_{j=0}^{\infty} h(j) \frac{\frac{\partial}{\partial \alpha} P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})} - \dot{l}_\alpha(x_{t-1}, x_t; \alpha, G) A_{\alpha,G} h(x_{t-1}, x_t).
\end{aligned}
$$

Recalling that $P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1}) = P^{\alpha,G}_{x_{t-1},x_t}$, altogether, we have

$$
\begin{aligned}
\frac{\partial \Psi_{n2}(\alpha, G)}{\partial \alpha} h = \frac{1}{n} \sum_{t=0}^{n} \sum_{j=0}^{\infty} h(j) \frac{\left(\frac{\partial}{\partial \alpha} P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})\right)_{|(x_t, x_{t-1}) = (X_t, X_{t-1})}}{P^{\alpha,G}_{X_{t-1},X_t}} \\
- \frac{1}{n} \sum_{t=0}^{n} \dot{l}_\alpha(X_{t-1}, X_t; \alpha, G) A_{\alpha,G} h(X_{t-1}, X_t).
\end{aligned}
$$

Similarly, for (B.4) and for all $k \in \mathbb{N}_0$, we have

$$
\begin{aligned}
\frac{\partial \Psi_{n2}(\alpha, G))}{\partial G(k)} h &= \frac{\partial}{\partial G(k)} \frac{1}{n} \sum_{t=0}^{n} \left(A_{\alpha,G} h(X_{t-1}, X_t) - \int h dG \right), \\
&= \frac{1}{n} \sum_{t=0}^{n} \frac{\partial}{\partial G(k)} A_{\alpha,G} h(X_{t-1}, X_t) - \frac{1}{n} \sum_{t=0}^{n} \frac{\partial}{\partial G(k)} \int h dG,
\end{aligned}
$$

where

$$
\frac{1}{n} \sum_{t=0}^{n} \frac{\partial}{\partial G(k)} \int h dG = \frac{\partial}{\partial G(k)} \sum_{j=0}^{\infty} h(j) G(j) = \sum_{j=0}^{\infty} h(j) \frac{\partial}{\partial G(k)} G(j) = h(k) \tag{B.6}
$$

and, for all $x_t, x_{t-1} \in \mathbb{N}_0$, using (B.5), we have

$$
\begin{aligned}
&\frac{1}{n} \sum_{t=0}^{n} \frac{\partial}{\partial G(k)} A_{\alpha,G} h(x_{t-1}, x_t) \\
&= \frac{1}{n} \sum_{t=0}^{n} \sum_{j=0}^{\infty} h(j) \frac{\partial}{\partial G(k)} \left(\frac{P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})}{P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})} \right) \\
&= \frac{1}{n} \sum_{t=0}^{n} \sum_{j=0}^{\infty} h(j) \Bigg[\frac{\left(\frac{\partial}{\partial G(k)} P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})\right) P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})}{\left(P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})\right)^2} \\
&\quad - \frac{\left(\frac{\partial}{\partial G(k)} P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})\right) P^{\alpha,G}(\varepsilon_t = j, X_t = x_t | X_{t-1} = x_{t-1})}{\left(P^{\alpha,G}(X_t = x_t | X_{t-1} = x_{t-1})\right)^2} \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty}h(j)\frac{\frac{\partial}{\partial G(k)}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})}{P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})}\\
&-\frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty}h(j)\frac{\left(\frac{\partial}{\partial G(k)}P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})\right)P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})}{\left(P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})\right)^2}\\
=&\frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty}h(j)\frac{\frac{\partial}{\partial G(k)}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})}{P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})}\\
&-\frac{1}{n}\sum_{t=0}^{n}A_{\alpha,G}h(x_{t-1},x_t)\frac{\frac{\partial}{\partial G(k)}P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})}{P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})}.
\end{aligned}
$$

Consequently, altogether, we have

$$
\begin{aligned}
&\frac{\partial\Psi_{n2}(\alpha,G))}{\partial G(k)}h\\
&=\frac{1}{n}\sum_{t=0}^{n}\sum_{j=0}^{\infty}h(j)\frac{\left(\frac{\partial}{\partial G(k)}P^{\alpha,G}(\varepsilon_t=j,X_t=x_t|X_{t-1}=x_{t-1})\right)_{|_{(x_t,x_{t-1})=(X_t,X_{t-1})}}}{P^{\alpha,G}_{X_{t-1},X_t}}\\
&\quad-\frac{1}{n}\sum_{t=0}^{n}A_{\alpha,G}h(X_{t-1},X_t)\frac{\left(\frac{\partial}{\partial G(k)}P^{\alpha,G}(X_t=x_t|X_{t-1}=x_{t-1})\right)_{|_{(x_t,x_{t-1})=(X_t,X_{t-1})}}}{P^{\alpha,G}_{X_{t-1},X_t}}-h(k).
\end{aligned}
$$

□

Lemma B.3. *Suppose the Assumptions of Lemma A.2 hold. Then, for $||\xi_n-\theta_0||\leq\delta$, we have*

$$
\begin{aligned}
&\left|\dot{l}_\alpha(X_{t-1},X_t;\theta_0)-\dot{l}_\alpha(X_{t-1},X_t;\xi_n)\right|\\
\leq&\delta\left(\frac{2X_{t-1}+CX_{t-1}^2(1+\rho)^{X_{t-1}-1}}{P^{\theta_0}_{X_{t-1},X_t}}+\frac{2X_{t-1}\left(CX_{t-1}(1+\rho)^{X_{t-1}-1}+1\right)}{P^{\theta_0}_{X_{t-1},X_t}P^{\xi_n}_{X_{t-1},X_t}}\right),
\end{aligned}
\tag{B.7}
$$

for some (generic) constant $C=C(\delta)$ and some $\rho=\rho(\delta)$, which becomes arbitrarily small for δ sufficiently small.

Proof. By plugging-in, we get

$$
\begin{aligned}
&\left|\dot{l}_\alpha(X_{t-1},X_t;\theta_0)-\dot{l}_\alpha(X_{t-1},X_t;\xi_n)\right|\\
=&\left|\frac{\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}G_0(X_t-j)(j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j}-\alpha_0^j(X_{t-1}-j)(1-\alpha_0)^{X_{t-1}-j-1})}{\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}\alpha_0^j(1-\alpha_0)^{X_{t-1}-j}G_0(X_t-j)}\right.\\
&\left.-\frac{\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}G_{\xi_n}(X_t-j)(j\alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j}-\alpha_{\xi_n}^j(X_{t-1}-j)(1-\alpha_{\xi_n})^{X_{t-1}-j-1})}{\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}\alpha_{\xi_n}^j(1-\alpha_{\xi_n})^{X_{t-1}-j}G_{\xi_n}(X_t-j)}\right|
\end{aligned}
$$

$$=: \left|\frac{a}{b} - \frac{a_n}{b_n}\right| = \left|\frac{a-a_n}{b} + \frac{a_n(b_n-b)}{bb_n}\right| \leq \frac{|a-a_n|}{b} + \frac{|a_n||b_n-b|}{bb_n}, \tag{B.8}$$

where we used that both b and b_n are transition probabilities with $b, b_n \in [0,1]$. As by Assumption 1, $b > 0$ and $\alpha_0 \in (0,1)$ holds and $\alpha_{\xi_n} \to \alpha_0$ by Theorems 2.1 and 3.1, we also have $b_n > 0$ with probability tending to 1. Hence, it remains to consider the numerators $|a-a_n|$ and $|a_n| \cdot |b-b_n|$ separately in the following. First, for $|a-a_n|$, we have

$$\begin{aligned}
&|a-a_n| \\
\leq &\Bigg| \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_0(X_t-j)(j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} - \alpha_0^j(X_{t-1}-j)(1-\alpha_0)^{X_{t-1}-j-1}) \\
&\qquad - \binom{X_{t-1}}{j} G_{\xi_n}(X_t-j)(j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} - \alpha_0^j(X_{t-1}-j)(1-\alpha_0)^{X_{t-1}-j-1})\Bigg| \\
&+ \Bigg| \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t-j)(j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} - \alpha_0^j(X_{t-1}-j)(1-\alpha_0)^{X_{t-1}-j-1}) \\
&\qquad - \binom{X_{t-1}}{j} G_{\xi_n}(X_t-j)(j\alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j} - \alpha_{\xi_n}^j(X_{t-1}-j)(1-\alpha_{\xi_n})^{X_{t-1}-j-1})\Bigg| \\
=: &II_{a,1} + II_{a,2}.
\end{aligned} \tag{B.9}$$

For the first term $II_{a,1}$, we get

$$\begin{aligned}
II_{a,1} \leq &\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} |G_0(X_t-j) - G_{\xi_n}(X_t-j)| \\
&\qquad |j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} - \alpha_0^j(X_{t-1}-j)(1-\alpha_0)^{X_{t-1}-j-1}| \\
\leq &\sum_{m=0}^{\infty} |G_0(m) - G_{\xi_n}(m)| \left(\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} \right. \\
&\qquad \left. + \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} \alpha_0^j(X_{t-1}-j)(1-\alpha_0)^{X_{t-1}-j-1} \right) \\
\leq &2X_{t-1} \sum_{m=0}^{\infty} |G_0(m) - G_{\xi_n}(m)| \\
\leq &2X_{t-1}\delta,
\end{aligned}$$

where we used $\sum_{m=0}^{\infty} |G_0(m) - G_{\xi_n}(m)| \leq \|\xi_n - \theta_0\| \leq \delta$ for the last inequality and, for the second last inequality, we made use of the binomial theorem to get

$$\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j}$$

$$\leq \sum_{j=0}^{X_{t-1}} \binom{X_{t-1}}{j} j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} = \sum_{j=1}^{X_{t-1}} \binom{X_{t-1}}{j} j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j}$$

$$= \sum_{j=0}^{X_{t-1}-1} \binom{X_{t-1}}{j+1}(j+1)\alpha_0^{j}(1-\alpha_0)^{X_{t-1}-(j+1)} = X_{t-1} \sum_{j=0}^{X_{t-1}-1} \binom{X_{t-1}-1}{j}\alpha_0^{j}(1-\alpha_0)^{X_{t-1}-1-j}$$

$$= X_{t-1} \tag{B.10}$$

as well as

$$\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j}\alpha_0^{j}(X_{t-1}-j)(1-\alpha_0)^{X_{t-1}-j-1}$$

$$\leq \sum_{j=0}^{X_{t-1}} \binom{X_{t-1}}{j}(X_{t-1}-j)\alpha_0^{j}(1-\alpha_0)^{X_{t-1}-j-1} = \sum_{j=0}^{X_{t-1}-1} \binom{X_{t-1}}{j}(X_{t-1}-j)\alpha_0^{j}(1-\alpha_0)^{X_{t-1}-j-1}$$

$$= X_{t-1} \sum_{j=0}^{X_{t-1}-1} \binom{X_{t-1}-1}{j}\alpha_0^{j}(1-\alpha_0)^{X_{t-1}-1-j} = X_{t-1}. \tag{B.11}$$

When dealing with the second term $II_{a,2}$, we have

$$\begin{aligned}
&\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} - \alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j} \\
&= \alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} - \alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j} \mp \alpha_0^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j} \\
&= \alpha_0^{j-1}\Big((1-\alpha_0)^{X_{t-1}-j} - (1-\alpha_{\xi_n})^{X_{t-1}-j}\Big) + (1-\alpha_{\xi_n})^{X_{t-1}-j}\Big(\alpha_0^{j-1} - \alpha_{\xi_n}^{j-1}\Big).
\end{aligned} \tag{B.12}$$

Further, for the last expression in brackets, we have

$$\begin{aligned}
\alpha_0^{j-1} - \alpha_{\xi_n}^{j-1} &= \alpha_0^{j-1} - \alpha_{\xi_n}^{j-1} \mp \alpha_0^{j-2}\alpha_{\xi_n} \\
&= \alpha_0^{j-2}(\alpha_0 - \alpha_{\xi_n}) + \alpha_0^{j-2}\alpha_{\xi_n} - \alpha_{\xi_n}^{j-1} \mp \alpha_0^{j-3}\alpha_{\xi_n}^2 \\
&= \alpha_0^{j-2}(\alpha_0 - \alpha_{\xi_n}) + \alpha_0^{j-3}\alpha_{\xi_n}(\alpha_0 - \alpha_{\xi_n}) + \alpha_0^{j-3}\alpha_{\xi_n}^2 - \alpha_{\xi_n}^{j-1} \mp \alpha_0^{j-4}\alpha_{\xi_n}^3 \\
&= \dots \\
&= (\alpha_0 - \alpha_{\xi_n}) \sum_{l=0}^{j-2} \alpha_0^{j-2-l}\alpha_{\xi_n}^{l}
\end{aligned} \tag{B.13}$$

and, analogously,

$$(1-\alpha_0)^{X_{t-1}-j} - (1-\alpha_{\xi_n})^{X_{t-1}-j} = (\alpha_{\xi_n} - \alpha_0) \sum_{l=0}^{X_{t-1}-j-1} (1-\alpha_0)^{X_{t-1}-j-1-l}(1-\alpha_{\xi_n})^{l}. \tag{B.14}$$

Consequently, plugging-in (B.13) and (B.14) in (B.12) leads to

$$\begin{aligned}
&\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} - \alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j} \\
&= \alpha_0^{j-1}(\alpha_{\xi_n} - \alpha_0) \sum_{l=0}^{X_{t-1}-j-1} (1-\alpha_0)^{X_{t-1}-j-1-l}(1-\alpha_{\xi_n})^{l} + (1-\alpha_{\xi_n})^{X_{t-1}-j}(\alpha_0 - \alpha_{\xi_n}) \sum_{l=0}^{j-2} \alpha_0^{j-2-l}\alpha_{\xi_n}^{l}.
\end{aligned}$$

Following the same procedure, we get

$$\begin{aligned}
&\alpha_{\xi_n}^j(1-\alpha_{\xi_n})^{X_{t-1}-j-1} - \alpha_0^j(1-\alpha_0)^{X_{t-1}-j-1} \\
&= \alpha_{\xi_n}^j(\alpha_0-\alpha_{\xi_n}) \sum_{l=0}^{X_{t-1}-j-2} (1-\alpha_{\xi_n})^{X_{t-1}-j-2-l}(1-\alpha_0)^l + (1-\alpha_0)^{X_{t-1}-j-1}(\alpha_{\xi_n}-\alpha_0)\sum_{l=0}^{j-1}\alpha_{\xi_n}^{j-1-l}\alpha_0^l.
\end{aligned}$$

Altogether, we obtain

$$\begin{aligned}
II_{a,2} &= \Bigg| \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t-j)\Big(j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} - j\alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j} \\
&\qquad +\alpha_{\xi_n}^j(X_{t-1}-j)(1-\alpha_{\xi_n})^{X_{t-1}-j-1} - \alpha_0^j(X_{t-1}-j)(1-\alpha_0)^{X_{t-1}-j-1}\Big)\Bigg| \\
&\leq |\alpha_0-\alpha_{\xi_n}| \sum_{j=0}^{\min(X_{t-1},X_t)} \Bigg[\binom{X_{t-1}}{j} G_{\xi_n}(X_t-j)\Bigg(j\Bigg(\alpha_0^{j-1}\sum_{l=0}^{X_{t-1}-j-1}(1-\alpha_0)^{X_{t-1}-j-1-l}(1-\alpha_{\xi_n})^l \\
&\quad +(1-\alpha_{\xi_n})^{X_{t-1}-j}\sum_{l=0}^{j-2}\alpha_0^{j-2-l}\alpha_{\xi_n}^l\Bigg) + (X_{t-1}-j)\Bigg(\alpha_{\xi_n}^j\sum_{l=0}^{X_{t-1}-j-2}(1-\alpha_{\xi_n})^{X_{t-1}-j-2-l}(1-\alpha_0)^l \\
&\quad +(1-\alpha_0)^{X_{t-1}-j-1}\sum_{l=0}^{j-1}\alpha_{\xi_n}^{j-1-l}\alpha_0^l\Bigg)\Bigg)\Bigg] \\
&=: |\alpha_0-\alpha_{\xi_n}|\big(II_{a,2,1}+II_{a,2,2}+II_{a,2,3}+II_{a,2,4}\big)
\end{aligned}$$

with an obvious notation for $II_{a,2,1}, II_{a,2,2}, II_{a,2,3}$ and $II_{a,2,4}$ according to the four terms on the last right-hand side. Let us consider $II_{a,2,1}$ in more detail. Making use of $|\alpha_0-\alpha_{\xi_n}| \leq \|\xi_n-\theta_0\| \leq \delta$, we have

$$\begin{aligned}
&\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t-j) j\alpha_0^{j-1} \sum_{l=0}^{X_{t-1}-j-1}(1-\alpha_0)^{X_{t-1}-j-1-l}(1-\alpha_{\xi_n})^l \qquad \text{(B.15)} \\
=&\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t-j) j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j-1} \sum_{l=0}^{X_{t-1}-j-1}(1-\alpha_0)^{-l}(1-\alpha_{\xi_n})^l \\
=&\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t-j) j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j-1} \sum_{l=0}^{X_{t-1}-j-1}\left(\frac{1-\alpha_{\xi_n}}{1-\alpha_0}\right)^l \\
=&\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t-j) j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j-1} \sum_{l=0}^{X_{t-1}-j-1}\left(1+\frac{\alpha_0-\alpha_{\xi_n}}{1-\alpha_0}\right)^l \\
\leq&\frac{1}{1-\alpha_0}\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j} \sum_{l=0}^{X_{t-1}-j-1}\left(1+\frac{|\alpha_0-\alpha_{\xi_n}|}{1-\alpha_0}\right)^l \\
\leq&\frac{1}{1-\alpha_0}\left(\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} j\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j}\right)\left(\sum_{l=0}^{X_{t-1}-1}\left(1+\frac{\delta}{1-\alpha_0}\right)^l\right) \\
=&\frac{1}{1-\alpha_0}X_{t-1}^2\left(1+\frac{\delta}{1-\alpha_0}\right)^{X_{t-1}-1}.
\end{aligned}$$

Using the same steps, we get for the three other terms

$$II_{a,2,2} \leq \frac{1}{\alpha_{\xi_n}} X_{t-1}^2 \left(1 + \frac{\delta}{\alpha_{\xi_n}}\right)^{X_{t-1}-1},$$
$$II_{a,2,3} \leq \frac{1}{1-\alpha_{\xi_n}} X_{t-1}^2 \left(1 + \frac{\delta}{1-\alpha_{\xi_n}}\right)^{X_{t-1}-1} \quad \text{and}$$
$$II_{a,2,4} \leq \frac{1}{\alpha_0} X_{t-1}^2 \left(1 + \frac{\delta}{\alpha_0}\right)^{X_{t-1}-1},$$

where we used that, e.g., $\sum_{l=0}^{j-2} \alpha_0^{j-2-l}\alpha_{\xi_n}^l = \sum_{l=0}^{j-2} \alpha_{\xi_n}^{j-2-l}\alpha_0^l$ and $\min(X_t, X_{t-1}) \leq X_{t-1}$. Altogether, using again $|\alpha_0 - \alpha_{\xi_n}| \leq \|\xi_n - \theta_0\| \leq \delta$, this leads to

$$II_{a,2} \leq C|\alpha_0 - \alpha_{\xi_n}| X_{t-1}^2 (1+\rho)^{X_{t-1}-1} \leq C\delta X_{t-1}^2 (1+\rho)^{X_{t-1}-1}$$

for some (generic) constant $C = C(\delta)$ and some $\rho = \rho(\delta)$ which becomes arbitrarily small for δ sufficiently small. Now, consider the second term of (B.8). We have

$$\begin{aligned}
|a_n| = &\left| \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t - j) \right. \\
&\qquad \left. \left(j\alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j} - \alpha_{\xi_n}^j (X_{t-1}-j)(1-\alpha_{\xi_n})^{X_{t-1}-j-1}\right)\right| \\
\leq &\sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} j\alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j} + \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} \alpha_{\xi_n}^j (X_{t-1}-j)(1-\alpha_{\xi_n})^{X_{t-1}-j-1} \\
\leq &2X_{t-1}, \qquad \text{(B.16)}
\end{aligned}$$

where we used that $G_{\xi_n}(k) \leq 1$ for all $k \in \mathbb{N}_0$ and the bounds obtained in (B.10) and (B.11). Further, we have

$$\begin{aligned}
|b_n - b| \leq &\left| \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t - j) \left(\alpha_{\xi_n}^j (1-\alpha_{\xi_n})^{X_{t-1}-j} - \alpha_0^j (1-\alpha_0)^{X_{t-1}-j}\right)\right| \\
&+ \left| \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} \left(G_{\xi_n}(X_t - j) - G_0(X_t - j)\right) \left(\alpha_0^j (1-\alpha_0)^{X_{t-1}-j}\right)\right| \\
=: &II_{b,1} + II_{b,2}.
\end{aligned}$$

We can proceed analogously as we did for $II_{a,1}$ and $II_{a,2}$ to get

$$\begin{aligned}
II_{b,1} = &\left| \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} G_{\xi_n}(X_t - j) \left(\alpha_{\xi_n}^j (\alpha_0 - \alpha_{\xi_n}) \sum_{l=0}^{X_{t-1}-j-1} (1-\alpha_{\xi_n})^{X_{t-1}-j-1-l}(1-\alpha_0)^l \right.\right. \\
&\qquad \left.\left. + (1-\alpha_0)^{X_{t-1}-j}(\alpha_{\xi_n} - \alpha_0) \sum_{l=0}^{j-1} \alpha_{\xi_n}^{j-1-l}\alpha_0^l\right)\right| \\
\leq &|\alpha_0 - \alpha_{\xi_n}| \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} \left(\alpha_{\xi_n}^j \sum_{l=0}^{X_{t-1}-j-1} (1-\alpha_{\xi_n})^{X_{t-1}-j-1-l}(1-\alpha_0)^l \right.
\end{aligned}$$

$$+(1-\alpha_0)^{X_{t-1}-j}\sum_{l=0}^{j-1}\alpha_{\xi_n}^{j-1-l}\alpha_0^l\Bigg)$$

$$\leq C|\alpha_0-\alpha_{\xi_n}|X_{t-1}\,(1+\rho)^{X_{t-1}-1}$$

$$\leq C\delta X_{t-1}\,(1+\rho)^{X_{t-1}-1}$$

and

$$\begin{aligned}
II_{b,2} &\leq \sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}|G_{\xi_n}(X_t-j)-G_0(X_t-j)|(\alpha_0^j(1-\alpha_0)^{X_{t-1}-j}) \\
&\leq \sum_{m=0}^{\infty}|G_{\xi_n}(m)-G_0(m)|\sum_{j=0}^{X_{t-1}}\binom{X_{t-1}}{j}\alpha_0^j(1-\alpha_0)^{X_{t-1}-j} \\
&= \sum_{m=0}^{\infty}|G_{\xi_n}(m)-G_0(m)| \\
&\leq \delta,
\end{aligned}$$

where C and ρ are as above. Altogether, this completes the proof. □

Lemma B.4. *Suppose the Assumptions of Lemma A.2 hold. Then, for $||\xi_n-\theta_0||\leq\delta$, we have*

$$\begin{aligned}
&\left|\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_{\xi_n}}_{X_{t-1},X_t|\alpha=\alpha_{\xi_n}}}{P^{\xi_n}_{X_{t-1},X_t}}-\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_0}_{X_{t-1},X_t|\alpha=\alpha_0}}{P^{\theta_0}_{X_{t-1},X_t}}\right| \\
\leq&\delta\left(\frac{\widetilde{C}X_{t-1}^3(1+\widetilde{\rho})^{X_{t-1}-1}+\widetilde{C}X_{t-1}^2}{P^{\xi_n}_{X_{t-1},X_t}}+\frac{\widetilde{C}X_{t-1}^2\left(\widetilde{C}X_{t-1}(1+\widetilde{\rho})^{X_{t-1}-1}+1\right)}{P^{\xi_n}_{X_{t-1},X_t}P^{\theta_0}_{X_{t-1},X_t}}\right) \qquad \text{(B.17)}
\end{aligned}$$

for some (generic) constant $\widetilde{C}=\widetilde{C}(\delta)$ and some $\widetilde{\rho}=\widetilde{\rho}(\delta)$, which becomes arbitrarily small for δ sufficiently small.

Proof. We follow the proof technique of Lemma B.3, but have to deal with the second derivative. First, we get the bound

$$\left|\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_{\xi_n}}_{X_{t-1},X_t|\alpha=\alpha_{\xi_n}}}{P^{\xi_n}_{X_{t-1},X_t}}-\frac{\frac{\partial^2}{\partial\alpha^2}P^{\alpha,G_0}_{X_{t-1},X_t|\alpha=\alpha_0}}{P^{\theta_0}_{X_{t-1},X_t}}\right|=:\left|\frac{c_n}{b_n}-\frac{c}{b}\right|\leq\frac{|c_n-c|}{b_n}+\frac{|c||b-b_n|}{b_nb},$$

where b_n and b are defined as in (B.8) and

$$\begin{aligned}
c_n=&\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}G_{\xi_n}(X_t-j)\alpha_{\xi_n}^{j-2}(1-\alpha_{\xi_n})^{X_{t-1}-j-2} \\
&\qquad\Big(\alpha_{\xi_n}^2(X_{t-1}-1)X_{t-1}+j(-2\alpha_{\xi_n}(X_{t-1}-1)-1)+j^2\Big), \\
c=&\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}G_0(X_t-j)\alpha_0^{j-2}(1-\alpha_0)^{X_{t-1}-j-2}
\end{aligned}$$

$$\left(\alpha_0^2(X_{t-1}-1)X_{t-1}+j(-2\alpha_0(X_{t-1}-1)-1)+j^2\right).$$

Hence, it remains to investigate $|c_n-c|$ and $|c|$. First, we note that c (and analogously c_n) can equivalently be written as

$$\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}G_0(X_t-j)\left(\alpha_0^j(1-\alpha_0)^{X_{t-1}-j-2}(X_{t-1}-1)X_{t-1}\right.$$
$$\left.-\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j-2}2j(X_{t-1}-1)-\alpha_0^{j-2}(1-\alpha_0)^{X_{t-1}-j-2}(j-j^2)\right).$$

Adding and subtracting the mixed terms as we did in (B.9), we obtain

$$\begin{aligned}
|c_n-c|\leq&\left|\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}G_{\xi_n}(X_t-j)\left(\alpha_{\xi_n}^j(1-\alpha_{\xi_n})^{X_{t-1}-j-2}(X_{t-1}-1)X_{t-1}\right.\right.\\
&-\alpha_0^j(1-\alpha_0)^{X_{t-1}-j-2}(X_{t-1}-1)X_{t-1}+\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j-2}2j(X_{t-1}-1)\\
&-\alpha_{\xi_n}^{j-1}(1-\alpha_{\xi_n})^{X_{t-1}-j-2}2j(X_{t-1}-1)+\alpha_0^{j-2}(1-\alpha_0)^{X_{t-1}-j-2}(j-j^2)\\
&\left.\left.-\alpha_{\xi_n}^{j-2}(1-\alpha_{\xi_n})^{X_{t-1}-j-2}(j-j^2)\right)\right|\\
&+\left|\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}\left(G_{\xi_n}(X_t-j)-G_0(X_t-j)\right)\left(\alpha_0^j(1-\alpha_0)^{X_{t-1}-j-2}(X_{t-1}-1)X_{t-1}\right.\right.\\
&\left.\left.-\alpha_0^{j-1}(1-\alpha_0)^{X_{t-1}-j-2}2j(X_{t-1}-1)-\alpha_0^{j-2}(1-\alpha_0)^{X_{t-1}-j-2}(j-j^2)\right)\right|\\
:=&I_{c,1}+I_{c,2}.
\end{aligned}$$

For $I_{c,1}$, we rewrite the last factor of the sum analogously to (B.12), (B.13) and (B.14) to get

$$\begin{aligned}
&I_{c,1}\\
\leq&|\alpha_0-\alpha_{\xi_n}|\sum_{j=0}^{\min(X_{t-1},X_t)}\binom{X_{t-1}}{j}\left[(X_{t-1}-1)X_{t-1}\left(\alpha_{\xi_n}^j\sum_{l=0}^{X_{t-1}-j-3}(1-\alpha_{\xi_n})^{X_{t-1}-j-3-l}(1-\alpha_0)^l\right.\right.\\
&\left.+(1-\alpha_0)^{X_{t-1}-j-2}\sum_{l=0}^{j-1}\alpha_{\xi_n}^{j-1-l}\alpha_0^l\right)+2j(X_{t-1}-1)\left(\alpha_0^{j-1}\sum_{l=0}^{X_{t-1}-j-3}(1-\alpha_0)^{X_{t-1}-j-3-l}(1-\alpha_{\xi_n})^l\right.\\
&\left.+(1-\alpha_{\xi_n})^{X_{t-1}-j-2}\sum_{l=0}^{j-1}\alpha_0^{j-1-l}\alpha_{\xi_n}^l\right)+(j-j^2)\left(\alpha_0^{j-2}\sum_{l=0}^{X_{t-1}-j-3}(1-\alpha_0)^{X_{t-1}-j-3-l}(1-\alpha_{\xi_n})^l\right.\\
&\left.\left.+(1-\alpha_{\xi_n})^{X_{t-1}-j-2}\sum_{l=0}^{j-3}\alpha_0^{j-3-l}\alpha_{\xi_n}^l\right)\right]\\
=:&|\alpha_0-\alpha_{\xi_n}|\left(I_{c,1,1}+I_{c,1,2}+I_{c,1,3}+I_{c,1,4}+I_{c,1,5}+I_{c,1,6}\right)
\end{aligned}$$

with an obvious notation for $I_{c,1,1}, I_{c,1,2}, I_{c,1,3}, I_{c,1,4}, I_{c,1,5}$ and $I_{c,1,6}$ according to the six terms on the last right-hand side. With similar steps as in (B.15), we get

$$\begin{aligned}
I_{c,1,1} &\leq \frac{1}{(1-\alpha_{\xi_n})^3} X_{t-1}^2 (X_{t-1}-1) \left(1+\frac{\delta}{1-\alpha_{\xi_n}}\right)^{X_{t-1}-1}, \\
I_{c,1,2} &\leq \frac{1}{\alpha_0(1-\alpha_0)^2} X_{t-1}^2 (X_{t-1}-1) \left(1+\frac{\delta}{\alpha_0}\right)^{X_{t-1}-1}, \\
I_{c,1,3} &\leq \frac{1}{(1-\alpha_0)^3} 2X_{t-1}^2 (X_{t-1}-1) \left(1+\frac{\delta}{1-\alpha_0}\right)^{X_{t-1}-1}, \\
I_{c,1,4} &\leq \frac{1}{(1-\alpha_{\xi_n})^2} 2X_{t-1}^2 (X_{t-1}-1) \left(1+\frac{\delta}{\alpha_{\xi_n}}\right)^{X_{t-1}-1}, \\
I_{c,1,5} &\leq \frac{1}{(1-\alpha_0)^3} X_{t-1}^2 (X_{t-1}-1) \left(1+\frac{\delta}{1-\alpha_0}\right)^{X_{t-1}-1} \quad \text{and} \\
I_{c,1,6} &\leq \frac{1}{\alpha_{\xi_n}(1-\alpha_{\xi_n})^2} X_{t-1}^2 (X_{t-1}-1) \left(1+\frac{\delta}{\alpha_{\xi_n}}\right)^{X_{t-1}-1}.
\end{aligned}$$

Altogether, this leads to

$$I_{c,1} \leq \widetilde{C}|\alpha_0 - \alpha_{\xi_n}| X_{t-1}^2 (X_{t-1}-1)(1+\widetilde{\rho})^{X_{t-1}-1} \leq \widetilde{C}\delta X_{t-1}^3 (1+\widetilde{\rho})^{X_{t-1}-1}$$

for some (generic) constant $\widetilde{C} = \widetilde{C}(\delta)$ and some $\widetilde{\rho} = \widetilde{\rho}(\delta)$ which becomes arbitrary small for δ sufficiently small. For $I_{c,2}$, we re-use the binomial theorem as in (B.10) and (B.11) and get

$$\begin{aligned}
I_{c,2} \leq & \sum_{j=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{j} |G_{\xi_n}(X_t - j) - G_0(X_t - j)| (\alpha_0^j (1-\alpha_0)^{X_{t-1}-j-2} (X_{t-1}-1) X_{t-1} \\
& - \alpha_0^{j-1} (1-\alpha_0)^{X_{t-1}-j-2} 2j(X_{t-1}-1) - \alpha_0^{j-2} (1-\alpha_0)^{X_{t-1}-j-2} (j-j^2)) \\
\leq & \sum_{m=0}^{\infty} |G_{\xi_n}(m) - G_0(m)| 4X_{t-1}(X_{t-1}-1)(1-\alpha_0)^{-2} \qquad \text{(B.18)} \\
\leq & \widetilde{C}\delta X_{t-1}^2 \qquad \text{(B.19)}
\end{aligned}$$

As last term, we have to consider $|c|$ for which we get

$$|c| \leq 4X_{t-1}(X_{t-1}-1)(1-\alpha_0)^{-2} \leq \widetilde{C} X_{t-1}^2 \qquad \text{(B.20)}$$

with the same arguments as in (B.18). Altogether, using $||\xi_n - \theta_0|| \leq \delta$, this completes the proof. □

Appendix C. Additional Tables

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.520	0.806	0.914	0.886	0.918	0.946	0.960	0.948	0.962
average length	0.264	0.163	0.125	0.242	0.115	0.083	0.223	0.098	0.069
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.918	0.938	0.934	0.820	0.934	0.934	0.652	0.834	0.942
average length	0.191	0.087	0.062	0.113	0.055	0.039	0.047	0.027	0.019

Table 9. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(1)-INAR(1) DGP with $\alpha = 0.1$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.652	0.834	0.930	0.876	0.916	0.944	0.974	0.976	0.986
average length	0.288	0.164	0.124	0.195	0.093	0.067	0.026	0.006	0.003
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.848	0.908	0.948	0.864	0.898	0.948	0.846	0.872	0.944
average length	0.094	0.0046	0.033	0.061	0.031	0.022	0.024	0.012	0.008

Table 10. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(1)-INAR(1) DGP with $\alpha = 0.1$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.410	0.750	0.866	0.754	0.892	0.916	0.856	0.864	0.912
average length	0.224	0.153	0.123	0.110	0.052	0.038	0.199	0.094	0.069
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.898	0.944	0.942	0.918	0.942	0.942	0.878	0.958	0.940
average length	0.236	0.107	0.076	0.242	0.107	0.076	0.219	0.104	0.074

Table 11. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(3)-INAR(1) DGP with $\alpha = 0.1$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.596	0.842	0.936	0.774	0.896	0.948	0.768	0.910	0.942
average length	0.283	0.162	0.122	0.073	0.032	0.023	0.126	0.062	0.045
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.694	0.880	0.960	0.970	0.978	0.960	0.704	0.866	0.944
average length	0.080	0.045	0.034	0.028	0.007	0.004	0.071	0.034	0.026

Table 12. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(3)-INAR(1) DGP with $\alpha = 0.1$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.804	0.952	0.944	0.828	0.928	0.934	0.914	0.938	0.924
average length	0.379	0.171	0.120	0.305	0.141	0.100	0.287	0.129	0.091
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.878	0.928	0.944	0.732	0.936	0.944	0.486	0.782	0.898
average length	0.253	0.112	0.078	0.143	0.069	0.048	0.058	0.031	0.024

TABLE 13. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(1)-INAR(1) DGP with $\alpha = 0.3$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.890	0.952	0.954	0.926	0.932	0.954	0.984	0.978	0.982
average length	0.363	0.162	0.114	0.222	0.100	0.071	0.034	0.007	0.004
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.878	0.924	0.938	0.872	0.924	0.938	0.836	0.916	0.940
average length	0.105	0.050	0.036	0.074	0.034	0.024	0.031	0.013	0.009

TABLE 14. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(1)-INAR(1) DGP with $\alpha = 0.3$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.698	0.966	0.946	0.566	0.872	0.912	0.596	0.890	0.918
average length	0.352	0.188	0.128	0.123	0.079	0.058	0.258	0.168	0.122
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.688	0.884	0.932	0.750	0.902	0.932	0.686	0.836	0.932
average length	0.324	0.211	0.155	0.340	0.231	0.172	0.311	0.224	0.170

Table 15. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(3)-INAR(1) DGP with $\alpha = 0.3$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.904	0.946	0.958	0.842	0.926	0.952	0.872	0.934	0.946
average length	0.357	0.157	0.111	0.083	0.036	0.026	0.151	0.070	0.051
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.796	0.902	0.982	0.978	0.980	0.982	0.686	0.882	0.934
average length	0.110	0.053	0.038	0.045	0.010	0.007	0.075	0.038	0.029

Table 16. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(3)-INAR(1) DGP with $\alpha = 0.3$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.896	0.964	0.952	0.408	0.684	0.818	0.550	0.794	0.832
average length	0.422	0.213	0.172	0.200	0.165	0.168	0.325	0.254	0.209
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.592	0.812	0.704	0.692	0.764	0.704	0.664	0.734	0.724
average length	0.385	0.293	0.244	0.423	0.299	0.253	0.389	0.287	0.237

TABLE 17. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(3)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.952	0.962	0.960	0.832	0.924	0.930	0.872	0.932	0.952
average length	0.288	0.122	0.087	0.088	0.038	0.030	0.159	0.075	0.056
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.784	0.904	0.974	0.978	0.990	0.974	0.652	0.878	0.922
average length	0.117	0.055	0.044	0.052	0.011	0.010	0.078	0.041	0.034

TABLE 18. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(3)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.958	0.946	0.810	0.412	0.258	0.196	0.682	0.534	0.478
average length	0.205	0.096	0.086	0.408	0.222	0.152	0.684	0.432	0.399
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.808	0.864	0.864	0.728	0.880	0.864	0.488	0.786	0.844
average length	0.559	0.348	0.306	0.401	0.230	0.204	0.235	0.071	0.065

TABLE 19. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(1)-INAR(1) DGP with $\alpha = 0.9$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.940	0.956	0.952	0.936	0.964	0.952	0.982	0.970	0.980
average length	0.068	0.029	0.022	0.237	0.107	0.078	0.040	0.008	0.012
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.878	0.954	0.942	0.900	0.948	0.942	0.858	0.926	0.934
average length	0.111	0.053	0.043	0.080	0.036	0.029	0.035	0.014	0.011

TABLE 20. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(1)-INAR(1) DGP with $\alpha = 0.9$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.938	0.920	0.875	0	0	0	0.012	0.008	0.010
average length	0.229	0.082	0.055	0.006	0.001	0.001	0.005	0.001	0.001
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.020	0.008	0.006	0.032	0.008	0.006	.032	0.006	0.010
average length	0.005	0.001	0.001	0.005	0.001	0.001	0.005	0.001	0.001

TABLE 21. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(3)-INAR(1) DGP with $\alpha = 0.9$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.948	0.950	0.938	0.850	0.924	0.910	0.884	0.940	0.920
average length	0.061	0.026	0.020	0.094	0.040	0.028	0.166	0.078	0.054
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.766	0.910	0.964	0.984	0.968	0.964	0.698	0.874	0.880
average length	0.117	0.058	0.040	0.055	0.012	0.006	0.084	0.042	0.030

TABLE 22. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a Poi(3)-INAR(1) DGP with $\alpha = 0.9$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.940	0.960	0.950	0.866	0.950	0.952	0.918	0.952	0.958
average length	0.303	0.117	0.081	0.360	0.151	0.105	0.335	0.143	0.100
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.856	0.946	0.940	0..732	0.938	0.938	0.634	0.852	0.940
average length	0.235	0.101	0.071	0.153	0.072	0.050	0.096	0.050	0.037

Table 23. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a NB(1,1/2)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.926	0.446	0.104	0.150	0	0	0.018	0	0
average length	0.333	0.142	0.100	0.219	0.099	0.070	0.053	0.017	0.011
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.184	0	0	0.832	0.566	0.304	0.660	0.558	0.364
average length	0.097	0.047	0.034	0.089	0.041	0.029	0.046	0.019	0.013

Table 24. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a NB(1,1/2)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.932	0.952	0.940	0.826	0.928	0.926	0.898	0.918	0.930
average length	0.362	0.142	0.100	0.385	0.175	0.124	0.385	0.169	0.120
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.832	0.938	0.944	0.710	0.912	0.932	0.486	0.756	0.874
average length	0.314	0.131	0.092	0.177	0.081	0.057	0.079	0.039	0.030

TABLE 25. Coverage and average length of the semi-parametrically constructed bootstrap confidence intervals based on the semi-parametric estimation of Drost et al. (2009a) and the semi-parametric INAR bootstrap from Section 3.1 for $\alpha, G(0), \ldots, G(4)$ in case of a NB(10,10/11)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

	α			$G(0)$			$G(1)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.948	0.952	0.922	0.866	0.782	0.662	0.206	0.004	0
average length	0.304	0.131	0.091	0.229	0.103	0.073	0.040	0.008	0.004
	$G(2)$			$G(3)$			$G(4)$		
n	100	500	1000	100	500	1000	100	500	1000
coverage	0.806	0.770	0.660	0.890	0.932	0.936	0.794	0.794	0.762
average length	0.107	0.051	0.036	0.081	0.036	0.025	0.036	0.014	0.010

TABLE 26. Coverage and average length of the parametrically constructed bootstrap confidence intervals based on *parametric* ML estimation and a *parametric* Poi-INAR bootstrap for $\alpha, G(0), \ldots, G(4)$ in case of a NB(10,10/11)-INAR(1) DGP with $\alpha = 0.5$ for different sample sizes.

Bernoulli **31**(4), 2025, 3213–3234
https://doi.org/10.3150/24-BEJ1844

Semi-parametric goodness-of-fit testing for INAR models

MAXIME FAYMONVILLE[1,a], CARSTEN JENTSCH[1,b] and CHRISTIAN H. WEISS[2,c]

[1]*Department of Statistics, TU Dortmund University, Dortmund, Germany,*
[a]*faymonville@statistik.tu-dortmund.de,* [b]*jentsch@statistik.tu-dortmund.de*
[2]*Department of Mathematics and Statistics, Helmut-Schmidt-University Hamburg, Hamburg, Germany,*
[c]*weissc@hsu-hh.de*

Among the various models designed for dependent count data, integer-valued autoregressive (INAR) processes enjoy great popularity. Typically, statistical inference for INAR models uses asymptotic theory that relies on rather stringent (parametric) assumptions on the innovations such as Poisson or negative binomial distributions. In this paper, we present a novel semi-parametric goodness-of-fit test tailored for the INAR model class. Relying on the INAR-specific shape of the joint probability generating function, our approach allows for model validation of INAR models without specifying the (family of the) innovation distribution. We derive the limiting null distribution of our proposed test statistic, prove consistency under fixed alternatives and discuss its asymptotic behavior under local alternatives. By manifold Monte Carlo simulations, we illustrate the overall good performance of our testing procedure in terms of power and size properties. In particular, it turns out that the power can be considerably improved by using higher-order test statistics. In supplementary material, we provide an application to three real-world economic data sets.

Keywords: Bootstrap; count time series; goodness-of-fit; local power; probability generating function; semi-parametric estimation

1. Introduction

Integer-valued autoregressive (INAR) models represent a powerful model class for modeling count time series. They offer a flexible and versatile approach for dealing with autoregressive (AR) time series of non-negative integer values and are the natural analog of the well-known AR model for continuous-valued time series. Du and Li (1991) introduce the INAR model of order p to follow the recursion

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}, \tag{1.1}$$

where $\varepsilon_t \overset{\text{i.i.d.}}{\sim} G$, that is, the innovations $(\varepsilon_t, t \in \mathbb{Z})$ are independent and identically distributed and follow a discrete distribution G with range $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ and probability mass function (pmf) $(G(k),\, k \in \mathbb{N}_0)$. The vector of model coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p) \in (0,1)^p$ fulfills $\sum_{j=1}^{p} \alpha_j < 1$. To ensure the integer-valued modeling of the time series, the model uses the binomial thinning operator "$\circ$" introduced by Steutel and van Harn (1979) as

$$\alpha_j \circ X_{t-j} = \sum_{i=1}^{X_{t-j}} Z_i^{(t,j)}, \tag{1.2}$$

with $(Z_i^{(t,j)}, i \in \mathbb{N}, t \in \mathbb{Z})$, $j \in \{1, \ldots, p\}$ being mutually independent Bernoulli-distributed random variables $Z_i^{(t,j)} \sim \text{Bin}(1, \alpha_j)$ independent of $(\varepsilon_t,\, t \in \mathbb{Z})$. Hence, the thinning operations are independent over time and independent of $(\varepsilon_t,\, t \in \mathbb{Z})$. Additionally, the thinning operation at time t and ε_t are

both independent of X_s, $s < t$. These comprehensive independence assumptions are characteristic for the INAR(p) model formulation of Du and Li (1991) and they differ from the one of Alzaid and Al-Osh (1990). But as only the INAR(p) model of Du and Li (1991) leads to the traditional Yule–Walker equations for the autocorrelation function (ACF), this model is usually preferred in practice and we focus on this model specification for the remainder of this paper. However, for $p = 1$, both versions of INAR(p) models simplify to the INAR(1) model first introduced by McKenzie (1985) and Al-Osh and Alzaid (1987).

The aforementioned flexibility of INAR models gets lost if one imposes a (parametric) family of distributions for the innovations, which, however, mostly happens in the literature because it simplifies considerably the estimation and the inference for these models. Initially, Al-Osh and Alzaid (1987) suggested a Poisson distribution for the innovations, which can be considered as the natural analog of the normal distribution in the continuous case. However, a Poisson distribution may be too restrictive in applications as it only allows for equidispersion. In the following years, INAR models with several alternative innovation distributions have been considered. For instance, Savani and Zhigljavsky (2007) deal with negative binomial innovations, Jazi, Jones and Lai (2012a) with geometric innovations, Jazi, Jones and Lai (2012b) with a zero-inflated Poisson innovation distribution, and Qi, Li and Zhu (2019) with zero-and-one inflated Poisson innovations. But regardless of which innovation distribution we choose, we will always face restrictions and lose some of the flexibility of the INAR(p) model in (1.1). Hence, we should not test for restrictive (parametric) null hypotheses H_0^{para} with *pre-defined* innovation distributions, where

$$H_0^{\text{para}} : (X_t,\, t \in \mathbb{Z}) \text{ is INAR}(p) \text{ with } G = G_\lambda \text{ for some } \lambda \in \Lambda \tag{1.3}$$

for some parametric family of innovation distributions $\{G_\lambda,\, \lambda \in \Lambda\}$ with $\Lambda \subset \mathbb{R}^d$ for some (finite) d as it is usually considered in the literature, see e.g. Meintanis and Karlis (2014), Hudecová, Hušková and Meintanis (2015), Schweer (2016), Aleksandrov, Weiß and Jentsch (2022), Aleksandrov et al. (2024) and a bivariate extension in Hudecova, Huskova and Meintanis (2021). Such parametric assumptions considerably facilitate the estimation and allow for relatively simple testing strategies, but they also make the tests prone to possible model misspecification. Instead, we want to test the (semi-parametric) null hypothesis H_0^{semi} that the data at hand follow an INAR(p) model as in (1.1) with *unspecified* innovation distribution, i.e.

$$H_0^{\text{semi}} : (X_t,\, t \in \mathbb{Z}) \text{ is INAR}(p). \tag{1.4}$$

In practice, irrespective of the concrete underlying innovation distribution, it is very helpful to know whether an INAR(p) process (1.1) is suitable to adequately capture the dependence structure of the count time series. In this case, a semi-parametric estimation approach can be used. The general relevance of semi-parametric approaches for $\mathbb{Z}$-valued time series was recently demonstrated by Liu, Li and Zhu (2021), who consider a similar model setup. In a more general time series setup, Armillotta and Gorgi (2024) introduce the semi-parametric pseudo-variance quasi-maximum-likelihood estimation, which is based on a Gaussian quasi-likelihood function relying on the specification of the pseudo-variance. The latter transfers naturally to time series of bounded counts, which are not considered here. Instead, we use the semi-parametric estimator of Drost, van den Akker and Werker (2009), which does not impose any parametric assumption on the innovation distribution and estimates G non-parametrically. Hence, even without imposing a parametric assumption on the innovation distribution, such (semi-parametric) INAR processes are generally attractive in applications as they are very flexible and still easily interpretable due to their autoregressive nature.

The paper is organized as follows. We introduce a test statistic for the null hypothesis defined in (1.4), derive its limiting null distribution and derive its asymptotic behavior under fixed and local

alternatives in Section 2. As the limiting null distribution is cumbersome to estimate, we introduce an appropriate bootstrap procedure in Section 3 to get critical values. Corresponding simulation results are provided in Section 4. The paper concludes with Section 5, where we summarize the results and give an outlook on possible future research questions. All proofs are deferred to supplementary material (Faymonville, Jentsch and Weiß, 2025a), which also contains additional tables and three real-world data applications. The corresponding MATLAB code is provided in further supplementary material (Faymonville, Jentsch and Weiß, 2025b).

2. Semi-parametric goodness-of-fit test for INAR models

Suppose we observe a sample $X_1, \ldots, X_n$ of time-series count data and we want to construct a test statistic for the null hypothesis H_0^{semi} in (1.4). While Meintanis and Karlis (2014) exclusively consider parametric null hypotheses H_0^{para} of the form (1.3) without providing asymptotic theory, we adopt their idea of constructing a suitable L_2-type test statistic based on two estimators of the (joint) probability generating function (pgf). The first pgf estimator shall be consistent in general, and the second one *only* under the null H_0^{semi}. In Meintanis and Karlis (2014), the pgf estimation under their (parametric) null H_0^{para} facilitates a lot due to their parametric assumption on the innovations, which results in closed form expressions for the pgf, depending on a *finite* number of estimated parameters determining the innovation distribution. In what follows, by contrast, we deal with more general expressions under the semi-parametric setup in (1.1).

2.1. Joint probability generating function of INAR models

As we want to test for INAR-type dependence structure of order p, we consider the joint pgf of $p+1$ consecutive random variables $X_t, \ldots, X_{t-p}$, which uniquely determines the full dependence structure of a (stationary) Markov process of order p. Then, for $u_0, \ldots, u_p \in [0,1]$, the joint pgf of $X_t, \ldots, X_{t-p}$ is defined as

$$g_p(u_0, \ldots, u_p) := g_{X_t,\ldots,X_{t-p}}(u_0, \ldots, u_p) := E\left(u_0^{X_t} \cdots u_p^{X_{t-p}}\right). \tag{2.1}$$

For the construction of a suitable pgf-based goodness-of-fit test statistic, we exploit the INAR dependence structure of $X_t, \ldots, X_{t-p}$ and derive an explicit representation of the joint pgf g_p for INAR(p) models.

Lemma 2.1 (Joint pgf of INAR(p) processes). *For $X_t, \ldots, X_{t-p}$ following an INAR(p) process (1.1), the pgf g_p defined in (2.1) can be represented by*

$$g_p(u_0 \ldots, u_p) = g_\varepsilon(u_0) \cdot E\left(\prod_{j=1}^{p} \left\{u_j\big(1+\alpha_j(u_0-1)\big)\right\}^{X_{t-j}}\right), \tag{2.2}$$

where $g_\varepsilon(u_0) = \sum_{k=0}^{\infty} P(\varepsilon_t = k)\, u_0^k = \sum_{k=0}^{\infty} G(k)\, u_0^k$.

The proof is contained in Subsection C.1 in the Supplement (Faymonville, Jentsch and Weiß, 2025a). Taking a closer look at (2.2), we see that the pgf g_p can be represented as a product of two factors. While the first factor *exclusively* depends on the pmf of the innovation distribution, $(G(k),\ k \in \mathbb{N}_0)$, the second factor is the joint pgf of (only) p consecutive random variables $X_{t-1}, \ldots, X_{t-p}$, whose arguments $u_j(1+\alpha_j(u_0-1))$, $j = 1, \ldots, p$ also depend on the autoregressive model coefficients $\alpha_1, \ldots, \alpha_p$. Note that the representation (2.2) does *not* require any further (parametric) assumptions on G.

2.2. Goodness-of-fit test statistic

When we renounce all parametric assumptions on the innovation distribution, the main challenge is the estimation of the INAR model, determined by the model coefficients $\alpha_1, \ldots, \alpha_p$ and the innovation distribution G. We cannot resort to (parametric) estimation methods such as moment or (conditional) maximum-likelihood estimation (see e.g. Weiß, 2018) that are usually employed to consistently estimate the (low-dimensional) parameter vector determining the innovation distribution. Instead, we use the semi-parametric estimator introduced in Drost, van den Akker and Werker (2009), which maintains the parametric binomial thinning, while simultaneously enabling the *non-parametric* estimation of the innovation distribution. For a small-sample refinement of this semi-parametric estimator using penalization techniques, we refer to Faymonville et al. (2023).

The estimation procedure proposed by Drost, van den Akker and Werker (2009) allows for joint estimation of the INAR coefficients and the pmf of the innovation distribution, $(G(k),\, k \in \mathbb{N}_0)$. Given $X_1, \ldots, X_n$, their semi-parametric maximum-likelihood estimator

$$(\widehat{\boldsymbol{\alpha}}_{\mathrm{sp}}, \widehat{G}_{\mathrm{sp}}) = (\widehat{\alpha}_{\mathrm{sp},1}, \ldots, \widehat{\alpha}_{\mathrm{sp},p}, \widehat{G}_{\mathrm{sp}}(0), \widehat{G}_{\mathrm{sp}}(1), \ldots), \tag{2.3}$$

which they prove to be consistent and efficient, is defined to maximize the conditional likelihood, i.e.

$$(\widehat{\boldsymbol{\alpha}}_{\mathrm{sp}}, \widehat{G}_{\mathrm{sp}}) \in \underset{(\boldsymbol{\alpha},G)\in[0,1]^p\times\widetilde{\mathcal{G}}}{\arg\max} \left(\prod_{t=p+1}^{n} P^{\boldsymbol{\alpha},G}_{(X_{t-1},\ldots,X_{t-p}),X_t} \right), \tag{2.4}$$

where $\widetilde{\mathcal{G}}$ contains all probability measures on $\mathbb{N}_0$ and $P^{\boldsymbol{\alpha},G}_{(X_{t-1},\ldots,X_{t-p}),X_t}$ denotes the transition probabilities

$$\begin{aligned} P^{\boldsymbol{\alpha},G}_{(x_{t-1},\ldots,x_{t-p}),x_t} &= \mathbb{P}_{\boldsymbol{\alpha},G}\left(\sum_{j=1}^{p} \alpha_i \circ X_{t-j} + \varepsilon_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p} \right) \\ &= (\mathrm{Bin}(x_{t-1},\alpha_1) * \ldots * \mathrm{Bin}(x_{t-p},\alpha_p) * G)\{x_t\}. \end{aligned}$$

Here, $\mathbb{P}_{\boldsymbol{\alpha},G}$ denotes the underlying probability measure induced by an INAR(p) process with coefficients $\boldsymbol{\alpha}$ and innovation distribution G, $\mathrm{Bin}(x_{t-j},\alpha_j)$ is the binomial distribution with parameters x_{t-j} and α_j, $j = 1, \ldots, p$ and "$*$" denotes the convolution of distributions. The estimator in (2.3) allows to estimate the pgf in (2.2) under the semi-parametric null H_0^{semi} in (1.4) *without* using any parametric assumption on the innovation distribution. Naturally, we use the plug-in estimator

$$\widehat{g}_{p;H_0}(\mathbf{u}) := \widehat{g}_{\varepsilon}(u_0) \cdot \frac{1}{n-p} \sum_{t=p+1}^{n} \prod_{j=1}^{p} \left\{ u_j\big(1 + \widehat{\alpha}_{\mathrm{sp},j}(u_0 - 1)\big) \right\}^{X_{t-j}}, \tag{2.5}$$

where $\mathbf{u} = (u_0, u_1, \ldots, u_p)$ and

$$\widehat{g}_{\varepsilon}(u_0) := \sum_{k=0}^{\infty} \widehat{P}(\varepsilon_t = k)\, u_0^k = \sum_{k=0}^{\infty} \widehat{G}_{\mathrm{sp}}(k)\, u_0^k = \sum_{k=0}^{\max(X_1,\ldots,X_n)} \widehat{G}_{\mathrm{sp}}(k)\, u_0^k \tag{2.6}$$

with the semi-parametric estimators $\widehat{\alpha}_{\mathrm{sp},j}$, $j \in \{1, \ldots, p\}$ and $(\widehat{G}_{\mathrm{sp}}(k),\, k \in \mathbb{N}_0)$ from (2.3). The last equality in (2.6) holds, because, for fixed $n \in \mathbb{N}$, we have $\widehat{G}_{\mathrm{sp}}(k) = 0\ \forall k > \max(X_1, \ldots, X_n)$; see Drost, van den Akker and Werker (2009) for details.

While the estimator $\widehat{g}_{p;H_0}(\mathbf{u})$ explicitly makes use of the INAR structure, the (non-parametric) estimator $\widehat{g}_p(\mathbf{u})$, defined by

$$\widehat{g}_p(\mathbf{u}) := \frac{1}{n-p} \sum_{t=p+1}^{n} \prod_{j=0}^{p} u_j^{X_{t-j}} \tag{2.7}$$

is consistent in general, that is, under the null *and* under the alternative. Hence, under the null H_0^{semi} in (1.4) of an underlying INAR(p) process, both $\widehat{g}_{p;H_0}(\mathbf{u})$ and $\widehat{g}_p(\mathbf{u})$ estimate the same quantity. This allows to construct the L_2-type test statistic T_n defined by

$$T_n = n \int_0^1 \cdots \int_0^1 \left(\widehat{g}_{p;H_0}(\mathbf{u}) - \widehat{g}_p(\mathbf{u})\right)^2 w(\mathbf{u}; a)\, d\mathbf{u}, \tag{2.8}$$

where $d\mathbf{u} := du_0 \cdots du_p$ and $w(\mathbf{u}; a) := (a+1)^{p+1} \prod_{j=0}^{p} u_j^a$ is a weighting function with weighting parameter $a \geq 0$. The weighting function is constructed to integrate to one such that w becomes a probability density function (pdf) on $[0,1]^{p+1}$. Choosing $a = 0$ corresponds to no weighting, whereas larger values for $a > 0$ (common choices are e.g. $a = 2$ and $a = 5$) put more weight close to the right boundaries of the integration intervals $[0,1]$, see Gürtler and Henze (2000). In what follows, we shall use a more precise notation of T_n than introduced in (2.8). Note that T_n is naturally a function of $X_1, \ldots, X_n$, but also that the definition of T_n relies on (nuisance) parameter estimators $\widehat{\theta}_{\text{sp}} := (\widehat{\boldsymbol{\alpha}}_{\text{sp}}, \widehat{G}_{\text{sp}})$, which are also functions of $X_1, \ldots, X_n$ themselves. This justifies the notations

$$T_n = T_n(X_1, \ldots, X_n) = T_n(\widehat{\theta}_{\text{sp}}; X_1, \ldots, X_n) = T_n(\widehat{\theta}_{\text{sp}}). \tag{2.9}$$

The test statistic T_n in (2.8) is of a similar structure as in Meintanis and Karlis (2014). However, by contrast to their parametric approach, we are using the semi-parametric estimator $\widehat{\theta}_{\text{sp}} = (\widehat{\boldsymbol{\alpha}}_{\text{sp}}, \widehat{G}_{\text{sp}})$ from (2.3). We should reject the null hypothesis H_0^{semi} in (1.4) for large values of T_n in (2.8).

Remark 2.2 (Higher-order test statistics). For the construction of test statistics in the spirit of (2.8) to test for the null H_0^{semi} in (1.4), we could also consider pgfs of higher order $s \geq p$, which may be beneficial to better detect Markov chain alternatives of higher order than p. That is, for any $s \geq p$, we can define the test statistic

$$T_n^{(s)} = n \int_0^1 \cdots \int_0^1 \left(\widehat{g}_{s;H_0}(u_0, \ldots, u_s) - \widehat{g}_s(u_0, \ldots, u_s)\right)^2 w(u_0, \ldots, u_s; a)\, du_0 \cdots du_s, \tag{2.10}$$

where $\widehat{g}_s$ is defined as in (2.7) with p replaced by s, and

$$\widehat{g}_{s;H_0}(u_0, \ldots, u_s) := \widehat{g}_\varepsilon(u_0) \cdot \frac{1}{n-s} \sum_{t=s+1}^{n} \prod_{j=1}^{s} \left\{u_j\left(1 + \widehat{\alpha}_{\text{sp},j}(u_0 - 1)\right)\right\}^{X_{t-j}}$$

with $\widehat{\alpha}_{\text{sp},j} := 0$ for $j = p+1, \ldots, s$. Note that $T_n^{(p)} = T_n$ holds.

While the test statistic T_n as proposed in (2.8) requires (numerical) integration, making use of (2.5) and (2.7), it can also be expressed without any integrals.

Lemma 2.3 (Calculation of T_n without integrals). *The test statistic T_n given by (2.8) can be equivalently written as*

$$T_n = \frac{n}{(n-p)^2}(a+1)^{p+1}\left(\prod_{j=1}^{p}\frac{1}{1+X_{t-j}+X_{s-j}+a}\right)\Bigg[\frac{1}{1+X_t+X_s+a}$$
$$+\sum_{k_1,k_2=0}^{\max(X_1,\ldots,X_n)}\widehat{G}_{sp}(k_1)\widehat{G}_{sp}(k_2)\sum_{i_1=0}^{X_{t-1}+X_{s-1}}\sum_{h_1=0}^{i_1}\cdots\sum_{i_p=0}^{X_{t-p}+X_{s-p}}\sum_{h_p=0}^{i_p}\frac{1}{1+k_1+k_2+a+\sum_{m=1}^{p}h_m}$$
$$\prod_{j=1}^{p}\binom{X_{t-j}+X_{s-j}}{i_j}\widehat{\alpha}_{sp,j}^{i_j}(-1)^{i_j-h_j}\binom{i_j}{h_j}-2\sum_{k=0}^{\max(X_1,\ldots,X_n)}\widehat{G}_{sp}(k)$$
$$\sum_{i_1=0}^{X_{t-1}+X_{s-1}}\sum_{h_1=0}^{i_1}\cdots\sum_{i_p=0}^{X_{t-p}+X_{s-p}}\sum_{h_p=0}^{i_p}\frac{1}{1+k+X_s+a+\sum_{m=1}^{p}h_m}\prod_{j=1}^{p}\binom{X_{t-j}}{i_j}\widehat{\alpha}_{sp,j}^{i_j}(-1)^{i_j-h_j}\binom{i_j}{h_j}\Bigg].$$

The proof is contained in Subsection C.2 in the Supplement (Faymonville, Jentsch and Weiß, 2025a).

2.3. Asymptotic theory

To derive the limiting distribution of our test statistic T_n, recall that, according to (2.9), we are actually confronted with $T_n(\widehat{\theta}_{\mathrm{sp}})$. Before addressing this case in Section 2.3.2 below, let us consider first the somewhat simpler case of $T_n(\theta_0)$ in Section 2.3.1, where the semi-parametric estimator $\widehat{\theta}_{\mathrm{sp}}$ is replaced by some $\theta_0 = (\boldsymbol{\alpha}_0, G_0) \in [0,1]^p \times \widetilde{\mathcal{G}}$.

2.3.1. Limiting distribution of $T_n(\theta_0)$

Let us consider the test statistic $T_n(\theta_0) = T_n(\theta_0; X_1, \ldots, X_n)$ in more detail. It is defined similarly to $T_n = T_n(\widehat{\theta}_{\mathrm{sp}})$ in (2.8), but with $\widehat{g}_{p;H_0}$ in (2.5) replaced by $g_{p;H_0}$, where

$$g_{p;H_0}(\mathbf{u}) := g_{0,\varepsilon}(u_0)\cdot\frac{1}{n-p}\sum_{t=p+1}^{n}\prod_{j=1}^{p}\left\{u_j\big(1+\alpha_{0,j}(u_0-1)\big)\right\}^{X_{t-j}},$$

with $g_{0,\varepsilon}(u_0) := \sum_{k=0}^{\infty} G_0(k)\, u_0^k$. While the original test statistic $T_n(\widehat{\theta}_{\mathrm{sp}})$ introduced in (2.8) allows to test for H_0^{semi} in (1.4), that is, for the *whole* INAR(p) model class, $T_n(\theta_0)$ is useful to test for null hypotheses of the form

$$H_0^{\mathrm{semi}}(\theta_0) : (X_t,\, t\in\mathbb{Z}) \text{ is INAR}(p) \text{ with } \theta = \theta_0 \tag{2.11}$$

for $\theta_0 = (\boldsymbol{\alpha}_0, G_0)$ with some pre-specified $\boldsymbol{\alpha}_0 = (\alpha_{0,1}, \ldots, \alpha_{0,p})$ and $G_0 = (G_0(k),\, k\in\mathbb{N}_0)$. Note that $H_0^{\mathrm{semi}}(\theta_0) \subset H_0^{\mathrm{semi}}$ for all $\theta_0 \in [0,1]^p \times \widetilde{\mathcal{G}}$.

The following proposition shows that, under the null $H_0^{\mathrm{semi}}(\theta_0)$, the test statistic $T_n(\theta_0)$ can be represented as a degenerate V-statistic, which enables the derivation of its limiting distribution.

Proposition 2.4 ($T_n(\theta_0)$ as a V-statistic). *Let $\theta_0 \in [0,1]^p \times \widetilde{\mathcal{G}}$. Suppose the null hypothesis $H_0^{\mathrm{semi}}(\theta_0)$ in (2.11) holds, that is, $X_1, \ldots, X_n$ follow an INAR(p) process with coefficients $\boldsymbol{\alpha}_0 = (\alpha_{0,1}, \ldots, \alpha_{0,p})$*

and innovation distribution $G_0 = (G_0(k),\, k \in \mathbb{N}_0)$*. Then, the test statistic* $T_n(\theta_0)$ *is a degenerate V-statistic. That is,* $T_n(\theta_0)$ *can be represented as*

$$T_n(\theta_0) = \frac{n}{(n-p)^2} \sum_{t=p+1}^{n} \sum_{s=p+1}^{n} h(Y_t, Y_s; \theta_0),$$

where $Y_t = (X_t, \ldots, X_{t-p})$*,* $Y_s = (X_s, \ldots, X_{s-p})$*,* $\theta_0 = (\boldsymbol{\alpha}_0, G_0)$ *and* $h : \mathbb{R}^{p+1} \times \mathbb{R}^{p+1} \to \mathbb{R}$ *with*

$$\begin{aligned} h(Y_t, Y_s; \theta_0) = \int_0^1 \ldots \int_0^1 \Bigg(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1 + \alpha_{0,j}(u_0 - 1)))^{X_{t-j}} - \prod_{j=0}^{p} u_j^{X_{t-j}} \Bigg) \\ \times \Bigg(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1 + \alpha_{0,j}(u_0 - 1)))^{X_{s-j}} - \prod_{j=0}^{p} u_j^{X_{s-j}} \Bigg) w(u_0, \ldots, u_p; a)\, du_0 \ldots du_p \end{aligned} \tag{2.12}$$

the symmetric and continuous kernel and $E(h(y_t, Y_s; \theta_0)) = 0$ *for all* $y_t = (x_t, \ldots, x_{t-p}) \in \mathbb{R}^{p+1}$*.*

The proof is contained in Subsection C.3 in the Supplement (Faymonville, Jentsch and Weiß, 2025a). This finding allows to use the asymptotic results established for degenerate V-statistics by Leucht and Neumann (2013), which leads to the following theorem.

Theorem 2.5 (Limiting distribution of $T_n(\theta_0)$ **under** $H_0^{\mathbf{semi}}(\theta_0)$**).** *Let* $\theta_0 \in [0,1]^p \times \widetilde{\mathcal{G}}$*. Suppose the null hypothesis* $H_0^{semi}(\theta_0)$ *in* (2.11) *holds, that is,* $X_1, \ldots, X_n$ *follow an INAR(p) process with coefficients* $\boldsymbol{\alpha}_0 = (\alpha_{0,1}, \ldots, \alpha_{0,p})$ *and innovation distribution* $G_0 = (G_0(k),\, k \in \mathbb{N}_0)$*. Then, for* $n \to \infty$*, we have*

$$T_n(\theta_0) \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k Z_k^2,$$

where $(Z_k)_k$ *is a sequence of independent standard normal random variables and* $(\lambda_k)_k$ *the sequence of nonzero eigenvalues of the equation*

$$E(h(y, Y_0; \theta_0)\Phi(Y_0)) = \lambda \Phi(y) \tag{2.13}$$

enumerated according their multiplicity, where $(\Phi_k)_k$ *are the associated orthonormal eigenfunctions and the kernel* h *is defined in* (2.12)*.*

The proof is contained in Subsection C.4 in the Supplement (Faymonville, Jentsch and Weiß, 2025a). Based on the asymptotic results from Theorem 2.5, we can define the (unconditional) test

$$\varphi_{n,\theta_0} := \mathbf{1}_{(q_{1-\gamma}, \infty)}(T_n(\theta_0)),$$

where $q_{1-\gamma}$ denotes the $(1-\gamma)$-quantile of the χ^2-type limiting distribution of $\sum_{k=1}^{\infty} \lambda_k Z_k^2$.

2.3.2. *Limiting distribution of* $T_n(\widehat{\theta}_{sp})$

Now, let us consider again the originally proposed test statistic $T_n := T_n(\widehat{\theta}_{\text{sp}})$ as defined in (2.8). To derive its limiting distribution, we impose the following condition which ensures the consistency of the estimator (2.3), see Drost, van den Akker and Werker (2009) for details.

Assumption 1. We assume that the true parameter $\theta_0 = (\boldsymbol{\alpha}_0, G_0) \in A \times \mathcal{G}$, where $\mathcal{G} = \{G \in \widetilde{\mathcal{G}} : 0 < G(0) < 1; E_G \varepsilon_t^{p+4} < \infty\}$ and $A = \{\boldsymbol{\alpha} \in (0,1)^p : \sum_{j=1}^p \alpha_j < 1\}$.

The limiting distribution under the null H_0^{semi} in (1.4) is given in the following theorem. For the proof, we make use of the fact that

$$T_n(\widehat{\theta}_{\text{sp}}) = \widetilde{T}_n(\theta_0) + o_P(1), \tag{2.14}$$

where $\widetilde{T}_n$ can be defined as T_n in (2.8), but with kernel h replaced by $\widetilde{h}$ defined in (C.8) in Faymonville, Jentsch and Weiß (2025a), and $\theta_0 = (\boldsymbol{\alpha}_0, G_0) \in A \times \mathcal{G}$. As $\widetilde{T}_n(\theta_0)$ is also a degenerate V-statistic under H_0^{semi}, we can use again the theory derived in Leucht and Neumann (2013) to establish the limiting distribution of $\widetilde{T}_n(\theta_0)$ and, together with (2.14), also of $T_n(\widehat{\theta}_{\text{sp}})$ under the null H_0^{semi}.

Theorem 2.6 (Limiting distribution of $T_n = T_n(\widehat{\theta}_{\text{sp}})$ under H_0^{semi}). *Suppose the null hypothesis H_0^{semi} in (1.4) holds, that is, $X_1, \ldots, X_n$ follow an INAR(p) process. Let Assumption 1 be satisfied. Then, for $n \to \infty$, we have*

$$T_n \xrightarrow{d} \sum_{k=1}^{\infty} \widetilde{\lambda}_k Z_k^2, \tag{2.15}$$

where $(Z_k)_k$ is a sequence of independent standard normal random variables and $(\widetilde{\lambda}_k)_k$ the sequence of nonzero eigenvalues of the equation

$$E(\widetilde{h}(y, Y_0; \theta_0)\widetilde{\Phi}(Y_0)) = \widetilde{\lambda}\widetilde{\Phi}(y) \tag{2.16}$$

enumerated according their multiplicity, where $(\widetilde{\Phi}_k)_k$ are the associated orthonormal eigenfunctions and the kernel $\widetilde{h}$ is defined in (C.8).

The proof is contained in Subsection C.5 in the Supplement (Faymonville, Jentsch and Weiß, 2025a). In the latter, we see that the effect of the estimator $\widehat{\theta}_{\text{sp}}$ on the limiting distribution is captured by the Fréchet derivative of $E_{\theta_0}(h_1(Y_1, u, \theta))$ with respect to θ and evaluated at $\theta = \theta_0$ with h_1 defined in (C.3). If this Fréchet derivative would be equal to zero, the substitution of θ_0 by $\widehat{\theta}_{\text{sp}}$ would not affect the limiting distribution at all. But according to (C.4) and (C.5), we see that it is generally not equal to zero and, consequently, the estimator $\widehat{\theta}_{\text{sp}}$ does affect the limiting distribution of the test statistic.

Now, based on the asymptotic results from Theorem 2.6, we are ready to define the (unconditional) test

$$\varphi_n := \mathbf{1}_{(\widetilde{q}_{1-\gamma}, \infty)}(T_n), \tag{2.17}$$

where $\widetilde{q}_{1-\gamma}$ denotes the $(1-\gamma)$-quantile of the χ^2-type limiting distribution of $\sum_{k=1}^{\infty} \widetilde{\lambda}_k Z_k^2$. As this distribution is not pivotal and cumbersome to estimate, we recommend to use a suitable bootstrap procedure in Section 3 that was proposed by Jentsch and Weiss (2019).

Remark 2.7 (Testing parametric null hypotheses H_0^{para}). Under the parametric null hypothesis H_0^{para} in (1.3), a test statistic of the form T_n in (2.8) can be used. Then, in (2.5), we have to replace $\widehat{g}_\epsilon(u_0)$ by $g_{\widehat{\lambda},\varepsilon}(u_0) := \sum_{k=0}^{\infty} G_{\widehat{\lambda}}(k)\, u_0^k$, where $\widehat{\lambda}$ is some $\sqrt{n}$-consistent estimator for λ, G is sufficiently smooth in λ and some arbitrary $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha}$ has to be used. Then, its limiting distribution can be derived by similar arguments.

2.4. Power properties

In this section, we prove consistency of our proposed test under fixed alternatives in Section 2.4.1 and discuss its asymptotic behavior under suitable local alternatives in Section 2.4.2.

2.4.1. Power under fixed alternatives

Suppose we observe data $X_1, \ldots, X_n$ from some (strictly) stationary count time series process $(X_t, t \in \mathbb{Z})$. Then, we want to test the (semi-parametric) null hypothesis H_0^{semi} in (1.4) against the (natural) alternative

$$H_1^{\text{semi}} : (X_t,\, t \in \mathbb{Z}) \text{ is } \textit{not} \text{ an INAR}(p). \tag{2.18}$$

Consequently, H_1^{semi} consists of all stationary count time series processes that are not INAR processes of some (fixed) order p (with *unspecified* innovation distribution).

Furthermore, as the test statistic T_n in (2.8) relies on the semi-parametric estimator $\widehat{\theta}_{\text{sp}}$ of Drost, van den Akker and Werker (2009), who presume an underlying INAR(p) process, we have to make sure that $\widehat{\theta}_{\text{sp}}$ still behaves well if the INAR model is misspecified. That is, in analogy to Theorem 2 in Drost, van den Akker and Werker (2009), we assume that a $\theta_{\text{mis}} = (\boldsymbol{\alpha}_{\text{mis}}, G_{\text{mis}}) \in (0,1)^p \times \mathcal{G}$ exists such that a weak convergence result

$$\sqrt{n}\left(\widehat{\theta}_{\text{sp}} - \theta_{\text{mis}}\right) \rightsquigarrow -\dot{\Psi}_{\theta_{\text{mis}}}^{\text{mis},-1} \mathbb{S}^{\theta_{mis}} \tag{2.19}$$

holds, where $\mathbb{S}^{\theta_{mis}}$ is some tight, Borel measurable Gaussian process obtained as in equation (15) in Drost, van den Akker and Werker (2009), but under the alternative, i.e., when the fitted INAR model is misspecified. Similarly, $\dot{\Psi}_{\theta_{\text{mis}}}^{\text{mis},-1}$ denotes here the inverse Fréchet derivative of the 'limiting' estimating equations under the alternative of a misspecifed INAR model, which can be obtained as in equations (8) and (9) in Drost, van den Akker and Werker (2009).

Hence, for studying the behavior of the test φ_n under *fixed* alternatives, let $(X_t, t \in \mathbb{Z})$ be some (strictly) stationary count time series process such that (2.19) holds. Furthermore, we assume that the joint pgf of $X_t, \ldots, X_{t-p}$, which is denoted by g_p, *differs* from $g_{p;H_0}$, which is the joint pgf of the theoretical best INAR(p) fit to the process $(X_t, t \in \mathbb{Z})$ under the alternative. More precisely, we suppose that

$$g_{p;H_0}(\mathbf{u}) - g_p(\mathbf{u}) = C(\mathbf{u}) \tag{2.20}$$

holds for some (bounded) function $C : [0,1]^{p+1} \to \mathbb{R}$ that is non-zero on some (sub)set $S \subseteq [0,1]^{p+1}$ having strictly positive Lebesgue measure $\lambda(S) > 0$. Additionally, we assume that the weak convergence

$$\left\{\sqrt{n}\left(\widehat{g}_{p;H_0}(\mathbf{u}) - \widehat{g}_p(\mathbf{u}) - \left(g_{p;H_0}(\mathbf{u}) - g_p(\mathbf{u})\right)\right), \mathbf{u} \in U\right\} \Rightarrow \left\{G_{mis}(\mathbf{u}), \mathbf{u} \in U\right\} \tag{2.21}$$

holds for some centered Gaussian process $\{G_{mis}(\mathbf{u}), \mathbf{u} \in U\}$ with covariance kernel $K_{mis}(\mathbf{u}_1, \mathbf{u}_2)$, $\mathbf{u}_1, \mathbf{u}_2 \in U$, where $U := [0,1]^{p+1}$. Then, we have the following result on the asymptotic behavior of T_n under fixed alternatives.

Theorem 2.8 (Consistency of T_n under fixed alternatives). *Suppose $(X_t, t \in \mathbb{Z})$ is a strictly stationary count time series process generated under the alternative H_1^{semi} such that (2.19), (2.20) and (2.21) hold. Then, for all $\gamma \in (0,1)$, the test φ_n from (2.17) is consistent for testing H_0^{semi} against alternatives H_1^{semi}, that is, $E(\varphi_n) \to 1$ as $n \to \infty$.*

The proof is contained in Subsection C.6 in the Supplement (Faymonville, Jentsch and Weiß, 2025a). Before we consider the case of local alternatives in the next subsection, Remark 2.9 gives a sufficient Markov condition for (2.20) to hold.

Remark 2.9 (Consistency of T_n for Markov processes of order p). The class of processes defined by H_1^{semi} in (2.18) is too rich to achieve consistency for T_n under general alternatives in H_1^{semi}. For instance, this is because the stationary distribution of a Markov process of some order $p' > p$ is *not* completely determined by the joint pgf g_p of $X_t, \ldots, X_{t-p}$. In fact, we would require a test that makes use of the joint pgf $g_{p'}$ of $X_t, \ldots, X_{t-p'}$, that is, a test of higher order in the sense of Remark 2.2. More precisely, by using a higher-order test $T_n^{(s)}$ with $s \geq p'$, consistency is guaranteed in the case of Markovian alternatives of order $p' > p$.

Hence, when using $T_n = T_n^{(p)}$ for testing the null H_0^{semi} against $H_1^{\text{semi,Markov}(p)}$, where

$$H_1^{\text{semi,Markov}(p)} : (X_t,\, t \in \mathbb{Z}) \text{ is a Markov process of order } p, \text{ but } \textit{not} \text{ an INAR}(p), \tag{2.22}$$

due to continuity of g_p and $g_{p;H_0}$, property (2.20) will generally hold. Consequently, we get consistency of T_n against *all* alternatives in $H_1^{\text{semi,Markov(p)}}$ that also fulfill (2.21).

In the following example, we illustrate the fixed-alternative setup covered by Theorem 2.8.

Example 2.10 (Fixed INAR(2) alternative). Suppose we want to use T_n in (2.8) (with $p = 1$) for testing the null H_0^{semi} of an INAR(1) process and the data $X_1, \ldots, X_n$ is generated from an INAR(2) process with $\theta = ((\alpha_1, \alpha_2), G)$, that is, with coefficients $\alpha_1, \alpha_2 \in (0,1)$, $\alpha_1 + \alpha_2 < 1$, where $\alpha_2 \neq 0$, and some innovation distribution $G \in \mathcal{G}$. Then, on the one hand, we have

$$g_{1;H_0}(u_0, u_1) = g_{\varepsilon,H_0}(u_0) \cdot E\left(\left\{u_1\big(1 + \alpha_{1,H_0}(u_0 - 1)\big)\right\}^{X_{t-1}}\right),$$

where $g_{\varepsilon,H_0}(u_0) = \sum_{k=0}^{\infty} G_{H_0}(k)\, u_0^k$ and $\theta_{H_0} = (\alpha_{1,H_0}, G_{H_0})$ is the solution of the population analogue of (2.4) (i.e. the argmax of the expectation of the log-likelihood), when fitting an INAR(1) to $X_1, \ldots, X_n$, when the true DGP $(X_t, t \in \mathbb{Z})$ is the INAR(2) specified above. On the other hand, using $g_1(u_0, u_1) = E(u_0^{X_t} u_1^{X_{t-1}}) = E(u_0^{X_t} u_1^{X_{t-1}} 1^{X_{t-2}}) = g_2(u_0, u_1, 1)$, we have

$$g_1(u_0, u_1) = g_{\varepsilon}(u_0) \cdot E\left(\left\{u_1\big(1 + \alpha_1(u_0 - 1)\big)\right\}^{X_{t-1}} \left\{\big(1 + \alpha_2(u_0 - 1)\big)\right\}^{X_{t-2}}\right).$$

Furthermore, writing $g_1(u_0, u_1) = g_1(u_0, u_1; \theta)$ and using a Taylor series argument, we have

$$g_1(u_0, u_1; \theta) = g_1(u_0, u_1; \widetilde{\theta}_{H_0}) + \dot{g}_1(u_0, u_1; \widetilde{\theta}_{H_0})(\theta - \widetilde{\theta}_{H_0}) + o\left(\frac{1}{\sqrt{n}}\right),$$

where $\widetilde{\theta}_{H_0} = ((\alpha_{1,H_0}, 0), G_{H_0})$. Altogether, due to $g_{1;H_0}(u_0, u_1) = g_{1;H_0}(u_0, u_1; \widetilde{\theta}_{H_0}) = g_1(u_0, u_1; \widetilde{\theta}_{H_0})$, this leads to

$$g_{1;H_0}(u_0, u_1) - g_1(u_0, u_1) = -\dot{g}_1(u_0, u_1; \widetilde{\theta}_{H_0})(\theta - \widetilde{\theta}_{H_0}) + o\left(\frac{1}{\sqrt{n}}\right)$$

uniformly in $(u_0, u_1) \in [0,1]^2$. Furthermore, as $\alpha_2 \neq 0$, we also have $\theta - \widetilde{\theta}_{H_0} \neq 0$ in $[0,1]^2 \times \mathcal{G}$ such that

$$n \int_0^1 \int_0^1 \left(g_{1;H_0}(u_0,u_1) - g_1(u_0,u_1)\right)^2 w(u_0,u_1;a) du_0 du_1 \to +\infty \tag{2.23}$$

as $n \to \infty$, because $\dot{g}_1(u_0,u_1;\widetilde{\theta}_{H_0})(\theta - \widetilde{\theta}_{H_0})$ is non-zero on a set with strictly positive Lebesgue measure such that (2.20) holds.

2.4.2. *Power under local alternatives*

For studying the behavior of the test φ_n under *local* alternatives, we have to consider observations from a triangular array of count time series $(X_{n,t}, t = 1,\dots,n,\ n \in \mathbb{N})$. For each fixed n, suppose that $X_{n,1},\dots,X_{n,n}$ are generated from a (strictly) stationary count time series process under H_1^{semi} such that the DGP under the alternative depends on n and converges to a DGP under H_0^{semi} as $n \to \infty$. We denote the limiting process under H_0^{semi} by $(X_{0,t}, t \in \mathbb{Z})$. For all $n \in \mathbb{N}$, let $g_{p,n}$ with $g_{p,n}(\mathbf{u}) = E(u_0^{X_{n,t}} \cdots u_p^{X_{n,t-p}})$ denote the joint pgf of $X_{n,t},\dots,X_{n,t-p}$ that *differs* from $g_{p;H_0,n}$, which is the joint pgf of the theoretically best INAR(p) fit to $(X_{n,t}, t \in \mathbb{Z})$. Furthermore, let $\widehat{g}_{p;H_0,n}(\mathbf{u})$ and $\widehat{g}_{p,n}(\mathbf{u})$ be the corresponding estimators as defined in (2.5) and (2.7) based on $X_{n,1},\dots,X_{n,n}$. Note that both quantities are now equipped with an n to match the notation of $g_{p,n}(\mathbf{u})$ and $g_{p;H_0,n}(\mathbf{u})$, but $\widehat{g}_p(\mathbf{u})$ and $\widehat{g}_{p;H_0}(\mathbf{u})$ defined in (2.5) and (2.7) depend of course already on n. Furthermore, suppose that $(X_{n,t}, t = 1,\dots,n,\ n \in \mathbb{N})$ is constructed such that

$$g_{p;H_0,n}(\mathbf{u}) - g_{p,n}(\mathbf{u}) = a_n C(\mathbf{u}) + o(a_n) \tag{2.24}$$

holds uniformly over $\mathbf{u} \in U$ for some $a_n \to 0$ as $n \to \infty$ and some (bounded) function $C : [0,1]^{p+1} \to \mathbb{R}$ that is non-zero on some (sub)set $S \subseteq [0,1]^{p+1}$ having positive Lebesgue measure $\lambda(S) > 0$. Additionally, we assume that the weak convergence

$$\left\{\sqrt{n}\left(\widehat{g}_{p;H_0,n}(\mathbf{u}) - \widehat{g}_{p,n}(\mathbf{u}) - \left(g_{p;H_0,n}(\mathbf{u}) - g_{p,n}(\mathbf{u})\right)\right), \mathbf{u} \in U\right\} \Rightarrow \left\{G_{loc}(\mathbf{u}), \mathbf{u} \in U\right\} \tag{2.25}$$

holds for some centered Gaussian process $\{G_{loc}(\mathbf{u}), \mathbf{u} \in U\}$ with covariance kernel $K_{loc}(\mathbf{u}_1, \mathbf{u}_2)$, $\mathbf{u}_1, \mathbf{u}_2 \in U$. Then, we have the following result on the asymptotic behavior of T_n under local alternatives of the form described above.

Theorem 2.11 (Power of T_n under local alternatives). *Suppose $(X_{n,t}, t = 1,\dots,n,\ n \in \mathbb{N})$ forms a triangular array of count time series and, for each fixed n, $X_{n,1},\dots,X_{n,n}$ is generated from a strictly stationary count time series process under the alternative H_1^{semi} such that (2.24) with $a_n = n^{-1/2}$ and (2.25) hold. Then, for all $\gamma \in (0,1)$, the test φ_n from (2.17) fulfills $\lim_{n\to\infty} E(\varphi_n) = 1 - F_{loc}(\widetilde{q}_{1-\gamma})$, where F_{loc} denotes the cumulative distribution function of the limiting distribution of T_n under local alternatives, that is, of*

$$\int_0^1 \cdots \int_0^1 \Big(G_{loc}(\mathbf{u}) + C(\mathbf{u})\Big)^2 w(\mathbf{u};a) d\mathbf{u}. \tag{2.26}$$

If $a_n \to 0$ such that $\sqrt{n}\, a_n \to \infty$, the test φ_n remains consistent, that is, we have $E(\varphi_n) \to 1$ as $n \to \infty$. If $a_n = o(n^{-1/2})$, the test φ_n has no asymptotic power, that is, we have $E(\varphi_n) \to \gamma$ as $n \to \infty$.

The proof of Theorem 2.11 is contained in Subsection C.7 in the Supplement (Faymonville, Jentsch and Weiß, 2025a). Taking a closer look at the limiting distribution under local alternatives in (2.26), we see that it can be decomposed in three additive terms A_1, A_2 and A_3, where

$$A_1 = \int_0^1 \cdots \int_0^1 G_{loc}^2(\mathbf{u}) w(\mathbf{u}; a) d\mathbf{u}, \qquad A_2 = 2 \int_0^1 \cdots \int_0^1 G_{loc}(\mathbf{u}) C(\mathbf{u}) w(\mathbf{u}; a) d\mathbf{u},$$
$$A_3 = \int_0^1 \cdots \int_0^1 C^2(\mathbf{u}) w(\mathbf{u}; a) d\mathbf{u}.$$

While A_1 corresponds to the χ^2-type limiting distribution of $T_n = T_n(\widehat{\theta}_{\text{sp}})$ derived in (2.15) for the underlying (limiting) process $(X_{0,t}, t \in \mathbb{Z})$ under the null H_0^{semi} and to the first term of T_n discussed in the proof of Theorem 2.8 for fixed alternatives, the second term A_2 has a centered normal distribution with variance determined by the covariance kernel K_{loc} of the Gaussian process $G_{loc} = \{G_{loc}(\mathbf{u}), \mathbf{u} \in [0,1]^{p+1}\}$ and by the function $C(\mathbf{u})$, $\mathbf{u} \in [0,1]^{p+1}$. Finally, the third term A_3 is deterministic and strictly positive as the function C is assumed to be non-zero on some set with positive Lebesgue measure. Consequently, together, $A_2 + A_3$ has a *non-centered* normal distribution with mean A_3 determined by C and variance determined by G_{loc} and C. As A_1 and A_2 are driven by the same Gaussian process G_{loc}, A_1 and A_2 will be typically dependent. Altogether, we see that the limiting distribution of T_n under local alternatives consists of a χ^2-type limiting distribution that is *shifted* by a non-centered normal distribution.

In continuation of Example 2.10 dealing with a fixed-alternative setup, we illustrate local alternatives covered by Theorem 2.11 in the following example.

Example 2.12 (Local INAR(2) alternatives). Suppose we want to use T_n with $p = 1$ for testing the null H_0^{semi} of an INAR(1) process and the data $X_{n,1}, \ldots, X_{n,n}$ is generated from an INAR(2) process with coefficients $\alpha_1, \alpha_{2,n} \in (0,1)$, where $\alpha_{2,n} = c/\sqrt{n}$ for some $c \in \mathbb{R}$ such that $\alpha_1, \alpha_{2,n} \in (0,1)$, $\alpha_1 + \alpha_{2,n} < 1$ for all $n \in \mathbb{N}$ and some innovation distribution $G \in \mathcal{G}$. Then, on the one hand, we have

$$g_{1;H_0,n}(u_0, u_1) = g_{\varepsilon,H_0,n}(u_0) \cdot E\left(\left\{u_1\big(1 + \alpha_{1,H_0,n}(u_0 - 1)\big)\right\}^{X_{t-1}}\right),$$

where $g_{\varepsilon,H_0,n}(u_0) = \sum_{k=0}^{\infty} G_{H_0,n}(k)\, u_0^k$ and $\theta_{H_0,n} = (\alpha_{1,H_0,n}, G_{H_0,n})$ is the solution of the population analogue of (2.4) (i.e. the argmax of the expectation of the log-likelihood), if fitting an INAR(1) to $X_{n,1}, \ldots, X_{n,n}$ when the true DGPs of $(X_{n,t}, t = 1, \ldots, n,\ n \in \mathbb{N})$ are INAR(2) processes. On the other hand, using

$$g_{1,n}(u_0, u_1) = E(u_0^{X_{t,n}} u_1^{X_{t-1,n}}) = E(u_0^{X_{t,n}} u_1^{X_{t-1,n}} 1^{X_{t-2,n}}) = g_{2,n}(u_0, u_1, 1),$$

we have

$$g_{1,n}(u_0, u_1) = g_{\varepsilon}(u_0) \cdot E\left(\left\{u_1\big(1 + \alpha_1(u_0 - 1)\big)\right\}^{X_{t-1,n}} \left\{\big(1 + \alpha_{2,n}(u_0 - 1)\big)\right\}^{X_{t-2,n}}\right),$$

where we used that $g_{\varepsilon,n}(u_0) = g_{\varepsilon}(u_0)$ holds as G does not depend on n by construction. Then, writing $g_{1;H_0,n}(u_0, u_1) = g_{1;H_0,n}(u_0, u_1; \widetilde{\theta}_{H_0,n})$, where $\widetilde{\theta}_{H_0,n} = ((\alpha_{1,H_0,n}, 0), G_{H_0,n}) \to \widetilde{\theta}_0$, and using a Taylor-series argument, we have

$$g_{1;H_0,n}(u_0, u_1; \widetilde{\theta}_{H_0,n}) = g_{1;H_0,n}(u_0, u_1; \widetilde{\theta}_0) + \dot{g}_{1;H_0,n}(u_0, u_1; \widetilde{\theta}_0)(\widetilde{\theta}_{H_0,n} - \widetilde{\theta}_0) + o\left(\frac{1}{\sqrt{n}}\right),$$

where $\dot{g}_{1;H_0,n}(u_0,u_1;\widetilde{\theta}_0)(\cdot)$ denotes the Fréchet derivative of $g_{1;H_0,n}(u_0,u_1;\theta)$ (with respect to θ; see also (C.4) and (C.5) in the Supplement (Faymonville, Jentsch and Weiß, 2025a)) evaluated in $\widetilde{\theta}_0$. Similarly, writing $g_{1,n}(u_0,u_1) = g_{1,n}(u_0,u_1;\theta_n)$, where $\theta_n = ((\alpha_1,\alpha_{2,n}),G) \to ((\alpha_1,0),G) = \widetilde{\theta}_0$ as $n \to \infty$, we get

$$g_{1,n}(u_0,u_1;\theta_n) = g_{1,n}(u_0,u_1;\widetilde{\theta}_0) + \dot{g}_{1,n}(u_0,u_1;\widetilde{\theta}_0)(\theta_n - \widetilde{\theta}_0) + o\left(\frac{1}{\sqrt{n}}\right).$$

Altogether, due to $g_{1;H_0,n}(u_0,u_1;\widetilde{\theta}_0) = g_{1,n}(u_0,u_1;\bar{\theta}_0)$, this leads to

$$g_{1;H_0,n}(u_0,u_1) - g_{1,n}(u_0,u_1) = \dot{g}_{1;H_0,n}(u_0,u_1;\widetilde{\theta}_0)(\widetilde{\theta}_{H_0,n} - \widetilde{\theta}_0) - \dot{g}_{1,n}(u_0,u_1;\widetilde{\theta}_0)(\theta_n - \widetilde{\theta}_0) + o\left(\frac{1}{\sqrt{n}}\right).$$

Furthermore, we have $\sqrt{n}(\widetilde{\theta}_{H_0,n} - \widetilde{\theta}_0) = ((0,c),0_{\mathcal{G}})$ and $\sqrt{n}(\theta_n - \bar{\theta}_0) \to ((d,0),0_{\mathcal{G}})$ for some $d \in \mathbb{R}$ by construction, where $0_{\mathcal{G}}$ denotes the zero-sequence $0_{\mathcal{G}} = (0,0,0,\ldots) \in \mathbb{R}^{\dim(G)}$. Note that the d is in the first entry of the limit, while the c above is in the second entry. Finally, using uniform convergence of *both* $\dot{g}_{1;H_0,n}(u_0,u_1;\widetilde{\theta}_0)$ and of $\dot{g}_{1;n}(u_0,u_1;\widetilde{\theta}_0)$ to $\dot{g}_1(u_0,u_1;\widetilde{\theta}_0)$, we get

$$\begin{aligned} & n\int_0^1\int_0^1 \big(g_{1;H_0,n}(u_0,u_1) - g_{1,n}(u_0,u_1)\big)^2 w(u_0,u_1;a)\,du_0\,du_1 \\ \to & \int_0^1\int_0^1 \Big(\dot{g}_1(u_0,u_1;\widetilde{\theta}_0)(((0,c),0_{\mathcal{G}})) - \dot{g}_1(u_0,u_1;\bar{\theta}_0)(((d,0),0_{\mathcal{G}}))\Big)^2 w(u_0,u_1;a)\,du_0\,du_1 \end{aligned} \quad (2.27)$$

as $n \to \infty$. The last right-hand side is strictly positive, because the expression that is squared and weighted before integration, that is, $\dot{g}_{1;H_0}(u_0,u_1;\widetilde{\theta}_0)(((0,c),0_{\mathcal{G}})) - \dot{g}_1(u_0,u_1;\bar{\theta}_0)(((d,0),0_{\mathcal{G}}))$ is non-zero on a set with strictly positive Lebesgue measure such that (2.24) holds with $a_n = 1/\sqrt{n}$. If $a_n \to 0$ as a slower rate such that $\sqrt{n}\,a_n \to \infty$, we get divergence to $+\infty$ as for fixed alternatives such that the corresponding test remains also consistent. If $a_n = o(n^{-1/2})$, (c,d) has to be replaced by $(0,0)$ such that the corresponding test has no asymptotic power as the function to be squared and integrated is exactly zero over $[0,1]^2$.

3. Bootstrap inference

As we have seen in the previous section and as already stated by Meintanis and Karlis (2014) and beforehand by e.g. Gürtler and Henze (2000), Meintanis and Swanepoel (2007), Leucht (2012), and Leucht and Neumann (2013), L_2-type test statistics as proposed in (2.8) do not exhibit a conventional (Gaussian) limiting distribution under the null. Although Drost, van den Akker and Werker (2009) derive a CLT for their semi-parametric estimator $(\widehat{\alpha}_{\mathrm{sp}},\widehat{G}_{\mathrm{sp}})$, this does not lead to a simple limiting distribution of T_n, see Theorem 2.6. Therefore, we propose a tailored bootstrap technique to make the testing procedure practicable. On the one hand, the bootstrap has to replicate correctly the binomial thinning operations (1.2) in the INAR recursion (1.1) and, on the other hand, we have to use appropriate bootstrap innovations that capture the correct, but unspecified innovation distribution.

Hence, a bootstrap procedure that fulfills these requirements is the semi-parametric INAR bootstrap proposed by Jentsch and Weiss (2019), which we will outline in the following:

1.) Fit semi-parametrically an INAR(p) process (1.1) using the estimator (2.3) proposed by Drost, van den Akker and Werker (2009) to get estimates $\widehat{\boldsymbol{\alpha}}_{\text{sp}} = (\widehat{\alpha}_{\text{sp},1}, \ldots, \widehat{\alpha}_{\text{sp},p})$ and $\widehat{G}_{\text{sp}} = \big(\widehat{G}_{\text{sp}}(k),\, k \in \mathbb{N}_0\big)$ for the INAR coefficients and for the pmf of the innovation distribution, respectively.

2.) Compute the test statistic $T_n = T_n(\widehat{\theta}_{\text{sp}}; X_1, \ldots, X_n)$, where T_n is defined in (2.8) and $\widehat{\theta}_{\text{sp}} = \widehat{\theta}_{\text{sp}}(X_1, \ldots, X_n) = (\widehat{\boldsymbol{\alpha}}_{\text{sp}}, \widehat{G}_{\text{sp}})$ is defined in (2.3).

3.) Generate bootstrap observations $X_1^*, \ldots, X_n^*$ according to

$$X_t^* = \widehat{\alpha}_{\text{sp},1} \circ^* X_{t-1}^* + \ldots + \widehat{\alpha}_{\text{sp},p} \circ^* X_{t-p}^* + \varepsilon_t^*,$$

where "$\circ^*$" denotes (mutually independent) bootstrap binomial thinning operations and $\varepsilon_t^* \overset{\text{i.i.d.}}{\sim} \widehat{G}_{\text{sp}}$ (conditionally on $X_1, \ldots, X_n$).

4.) Compute the bootstrap test statistic $T_n^* := T_n(\widehat{\theta}_{\text{sp}}^*; X_1^*, \ldots, X_n^*)$, where T_n is defined in (2.8) and $\widehat{\theta}_{\text{sp}}^* = \widehat{\theta}_{\text{sp}}(X_1^*, \ldots, X_n^*) = (\widehat{\boldsymbol{\alpha}}_{\text{sp}}^*, \widehat{G}_{\text{sp}}^*)$ is the bootstrap analog of $\widehat{\theta}_{\text{sp}}$ based on $X_1^*, \ldots, X_n^*$.

5.) Repeat the Steps 3.) and 4.) B times, with B sufficiently large, to get bootstrap test statistics $T_n^{*,b}$, $b \in \{1, \ldots, B\}$.

6.) Reject the null hypothesis (1.4) at significance level γ if $T_n = T_n(\widehat{\theta}_{\text{sp}}; X_1, \ldots, X_n)$ exceeds the $(1-\gamma)$-quantile of the empirical distribution of $T_n^{*,b}$, $b \in \{1, \ldots, B\}$.

To ensure the (approximate) stationarity of the bootstrap time series in Step 3 of the above algorithm, we use a burn-in period of r observations, which we will then cut off, i.e., we generate $X_1^*, \ldots, X_{n+r}^*$ and cut the first r values. In the simulation study in Section 4, we use $r = 100$. To initialize this burn-in period, we use the rounded mean value of the original observations.

The semi-parametric INAR bootstrap procedure is proved to be (first-order) consistent for the large class of *functions of generalized means* in Jentsch and Weiss (2019) under mild conditions, and it relies on the semi-parametric estimator proposed by Drost, van den Akker and Werker (2009), who proved its (estimation) efficiency. When it comes to formal statements about the efficiency of the bootstrap procedure itself in the sense of higher-order refinements that make typically use of Edgeworth expansions, this would require a lot (more) technical details and is beyond the scope of this paper.

Remark 3.1 (Bootstrap for testing parametric null hypotheses H_0^{para}). In the situation of a parametric null hypothesis H_0^{para} (1.3) discussed in Remark 2.7, the *parametric* INAR bootstrap of Jentsch and Weiss (2019) finds application. There, the semi-parametric estimators $\widehat{\theta}_{\text{sp}} = (\widehat{\boldsymbol{\alpha}}_{\text{sp}}, \widehat{G}_{\text{sp}})$ and $\widehat{\theta}_{\text{sp}}^* = (\widehat{\boldsymbol{\alpha}}_{\text{sp}}^*, \widehat{G}_{\text{sp}}^*)$ of Steps 1.) and 4.) are replaced by suitable parametric estimators.

4. Simulations

We investigate the performance of the proposed goodness-of-fit test through a simulation study, where we simulate data from different data generating processes (DGPs) under the null and under the alternative, and where we compute the resulting size and power, respectively. Additionally, we propose a way to better detect violations of the nulls in terms of model order and highlight a big advantage of our semi-parametric test being able to test for deviations from the null in terms of the INAR *structure*. We mainly focus on the null hypothesis H_0^{semi} in (1.4) with $p = 1$ and consider sample sizes $n \in \{100, 500\}$. To ensure stationarity of the generated data, we include a prerun of 100 observations which will be omitted afterwards. The significance level γ equals 5%. We compare our rejection rates with those from the simulation study performed in Meintanis and Karlis (2014), where the authors considered the parametric null hypothesis "$H_0 : (X_t,\, t \in \mathbb{Z})$ is a Poi-INAR(1) process", which is of the form H_0^{para} in (1.3). They

considered four different DGPs with different parameterizations: INAR(1) with Poi(λ) innovations, INAR(1) with NB(N, π) innovations, Poi-INAR(2) with Poi(λ) innovations, and Poi-INGARCH(1, 1). For the latter process, we have $X_t|X_{t-1}, X_{t-2}, \ldots \sim \text{Poi}(M_t)$, where $M_t = \beta_0 + \beta_1 M_{t-1} + \alpha_1 X_{t-1}$. In addition to the DGPs considered in Meintanis and Karlis (2014), we study further data and test scenarios later on. We consider different weighting parameters $a \in \{0, 2, 5\}$. For convenience of comparison, we state the rejection rates of Meintanis and Karlis (2014, Table 1) in italic numbers (where available). They used $a = 2$ in their simulations, so we only get direct comparability for this moderate weighting. Since we are concerned with a large computational burden due to the semi-parametric estimation and multiple integration, for conducting bootstrap simulation studies, we use the warp-speed approach with $M = 10^4$ Monte–Carlo samples (see Giacomini, Politis and White (2013) for details). While R (R Core Team, 2022) is sufficient for moderate sample sizes, also see the *spINAR* package (Faymonville et al., 2024), we recommend MATLAB (The MathWorks Inc., 2022) for larger n. To get an idea of the computing time, see Tables 1 and 2 in the supplementary material (Faymonville, Jentsch and Weiß, 2025a), which for different DGPs contain the computing time in seconds for one Monte Carlo sample using the warp speed method and MATLAB.

4.1. Performance under the null

First, we investigate how our test performs under the null H_0^{semi} in (1.4). In Table 1, we see the rejection rates for a Poi(λ)-INAR(1) DGP with model coefficient α and different weighting parameters a. For all weightings, we are rather conservative but we keep the level of 5%. Meintanis and Karlis (2014) better exploit the level which could be expected due to their additional (true) information about the innovation distribution. Additionally, to the parameterizations set by Meintanis and Karlis (2014), we also consider values of α lying closer to the boundaries of its parameter range, i.e., $\alpha \in \{0.1, 0.9\}$, see Table 3 in the Supplement (Faymonville, Jentsch and Weiß, 2025a). These parameterizations lead to similar results.

Next, we consider the case of an NB(N, π)-INAR(1) DGP. Table 2 shows that in this case of overdispersion, we are now even closer to the desired size of 5%. The (parametric) test of Meintanis and Karlis (2014) will generally reject the INAR model class since this DGP represents a scenario under the alternative for their null. We also test for the null H_0^{semi} in (1.4) with $p = 2$. Table 3 displays the rejection rates for the different Poi-INAR(2) DGPs. We see that we also keep the level when testing for the null of an INAR process of order 2. Note that Meintanis and Karlis (2014) solely applied their test to the first-order null.

In all simulation setups, we keep the level of 5%, but the results are rather conservative. Although Drost, van den Akker and Werker (2009) proved efficiency of their *semi-parametric* INAR model estimator, a large number of (innovations) parameters has to be estimated. In contrast, for the *parametric*

Table 1. Actual sizes in case of a Poi(λ)-INAR(1) DGP when testing for H_0^{semi} in (1.4) with $p = 1$. Numbers in italic are taken from Table 1 in Meintanis and Karlis (2014) who test for H_0^{para} in (1.3) with $p = 1$ and G_λ =Poi(λ).

		$a = 0$		$a = 5$		$a = 2$		*MK*, $a = 2$	
λ	α	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.3	0.036	0.033	0.037	0.040	0.036	0.035	*0.049*	*0.055*
1	0.5	0.040	0.041	0.044	0.046	0.043	0.042	*0.051*	*0.053*
3	0.3	0.032	0.034	0.043	0.022	0.038	0.027	*0.056*	*0.047*
3	0.5	0.032	0.032	0.037	0.029	0.039	0.034	*0.055*	*0.059*

Table 2. Actual sizes in case of a NB(N,π)-INAR(1) DGP when testing for H_0^{semi} in (1.4) with $p=1$. Numbers in italic display the power values taken from Table 1 in Meintanis and Karlis (2014) who test for H_0^{para} in (1.3) with $p=1$ and G_λ =Poi(λ).

			$a=0$		$a=5$		$a=2$		*MK*, $a=2$	
N	π	α	$n=100$	$n=500$	$n=100$	$n=500$	$n=100$	$n=500$	$n=100$	$n=500$
1	1/2	0.5	0.048	0.050	0.050	0.053	0.052	0.052	*0.686*	*1.000*
2	2/3	0.5	0.048	0.049	0.048	0.053	0.045	0.049	*0.327*	*0.897*
10	10/11	0.5	0.047	0.046	0.049	0.046	0.046	0.045	*0.120*	*0.149*

INAR model, there will be typically no more than $p+2$ parameters (p INAR coefficients plus one or two parameters for the innovations' mean and variance), which is considerably lower. As these estimators can also leverage the parametric family of innovation distributions, this explains why the semi-parametric test may not hold the level as good as parametric test procedures in finite samples.

4.2. Performance under the alternative

Now, we assess how our test performs when the DGP at hand deviates from the null in terms of model order or model structure. First, we consider the scenario of an INAR(2) with Poi(λ) innovations with the same parameterizations as in Table 3. Table 4 shows that we perform similar to Meintanis and Karlis (2014) though slightly less powerful due to their correctly imposed assumption on the innovations' distribution under the null. With higher weight, however, the power increases, partly surpassing the parametric approach of Meintanis and Karlis (2014). The same holds in case of an INGARCH$(1,1)$ DGP as demonstrated in Table 5. Also in the setup of a Poi-INAR(2) DGP, in addition to the parameterizations considered in Meintanis and Karlis (2014), we examine settings where $\alpha_1+\alpha_2$ are close to 0 or 1, see Table 4 in the Supplement (Faymonville, Jentsch and Weiß, 2025a). When both $\alpha_1=\alpha_2=0.05$, the test has low power which has been expected since violations of the hypothetical dependence structure are hard to recognize for such a low level of dependence. In particular, the small value of α_2 implies that the corresponding INAR(2) process does not differ much from an INAR(1) process. In case of $\alpha_1=0.4$ and $\alpha_2=0.5$ however, we achieve even higher power results than for $\alpha_1=0.5$ and $\alpha_2=0.3$ in Table 4, explainable through the higher value of α_2. Again, the power increases in the weighting parameter a. For a better understanding of the weighting concept, consider Figure 1, which contains two heatmaps of $(\widehat{g}_{1;H_0}(u_0,u_1)-\widehat{g}_1(u_0,u_1))^2 w(u_0,u_1;a)$ for a large sample of an INGARCH$(1,1)$ DGP with $\beta_0=1$, $\beta_1=0.1$ and $\alpha_1=0.5$ for $n=5000$. The left heatmap corresponds to $a=0$, i.e., no weighting, the right one to $a=5$. We see that the weighting shifts the differences to the upper right corner marking the endpoint of the integration intervals $[0,1]$. In this area, we get darker color, i.e., the pgfs differ more.

Table 3. Actual sizes in case of a Poi(λ)-INAR(2) DGP when testing for H_0^{semi} in (1.4) with $p=2$.

			$a=0$		$a=5$		$a=2$	
λ	α_1	α_2	$n=100$	$n=500$	$n=100$	$n=500$	$n=100$	$n=500$
1	0.3	0.1	0.037	0.040	0.036	0.024	0.035	0.024
1	0.5	0.1	0.039	0.043	0.039	0.037	0.037	0.036
1	0.5	0.3	0.034	0.041	0.042	0.033	0.040	0.035

Table 4. Power in case of a Poi(λ)-INAR(2) DGP when testing for H_0^{semi} in (1.4) with $p = 1$. Numbers in italic are taken from Table 1 in Meintanis and Karlis (2014) who test for H_0^{para} in (1.3) with $p = 1$ and G_λ =Poi(λ).

			$a = 0$		$a = 5$		$a = 2$		*MK*, $a = 2$	
λ	α_1	α_2	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.3	0.1	0.039	0.046	0.045	0.050	0.040	0.043	*0.088*	*0.035*
1	0.5	0.1	0.044	0.063	0.056	0.104	0.050	0.086	*0.089*	*0.127*
1	0.5	0.3	0.097	0.206	0.178	0.525	0.144	0.415	*0.191*	*0.487*

Conspicuous at first glance may be the partially poor power results for both the Poi-INAR(2) and the INGARCH(1, 1) DGP. In the Poi-INAR(2) case, see Table 4, this can be explained by the similarity of an INAR(1) process and an INAR(2) process with small α_2. For the INGARCH(1, 1) DGPs, Meintanis and Karlis (2014) provide the explanation that such DGPs do not differ much from INAR(1) processes if the ACF at lag 1, i.e., $\rho(1)$, is small. As we can see in Table 5, the power distortions exactly occur in such cases of small autocorrelation.

An additional explanation for the generally rather low power values is that when testing the null of an INAR(1) process, we consider the pgf of order 1. However, this does not explain the *entire* dependence structure of an alternative Poi-INAR(2) or INGARCH(1, 1) process, since both of them are no first-order Markov chain. To address this, we fit an INAR(1) model to the DGP but now use the second-order test statistic $T_n^{(2)}$ ($s = 2$ in (2.10)) by setting $\widehat{\alpha}_{\text{sp},2} := 0$ as suggested in Remark 2.2. The size results are presented in Tables 5 and 6 in the Supplement (Faymonville, Jentsch and Weiß, 2025a), where we see that we still keep the level of 5%. Tables 6 and 7 in this file and Table 7 in the Supplement contain the resulting power values. For comparison, for the two first DGPs, we again included the power values of Meintanis and Karlis (2014) in italics, which still result from using the first-order ($p = 1$) test statistic T_n in (2.8). As anticipated, when using the second-order test statistic, the power increased, in some settings substantially. Again, we tend to achieve higher power values with higher weighting.

The parametric testing approaches as in Meintanis and Karlis (2014) allow to test for deviations from the null in terms of model order, which can also be done with our semi-parametric goodness-of-fit test. In addition, due to the flexible and non-restrictive nature of the null hypothesis H_0^{semi} in (1.4), we are able to test for deviations from the INAR structure in general. For this purpose, we consider INARCH(1) and Poi-DAR(1) DGPs with different parameterizations, see Weiß (2018).

Table 5. Power in case of an INGARCH(1, 1) DGP when testing for H_0^{semi} in (1.4) with $p = 1$. Numbers in italic are taken from Table 1 in Meintanis and Karlis (2014) who test for H_0^{para} in (1.3) with $p = 1$ and G_λ =Poi(λ).

				$a = 0$		$a = 5$		$a = 2$		*MK*, $a = 2$	
β_0	β_1	α_1	$\rho(1)$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
0.2	0.4	0.10	0.09	0.026	0.018	0.026	0.030	0.023	0.021	*0.012*	*0.050*
0.2	0.4	0.20	0.21	0.031	0.087	0.054	0.131	0.044	0.117	*0.042*	*0.108*
1.0	0.1	0.50	0.52	0.068	0.205	0.159	0.645	0.113	0.501	*0.152*	*0.512*
0.5	0.1	0.50	0.52	0.177	0.707	0.256	0.865	0.237	0.833	*0.248*	*0.724*
0.1	0.4	0.40	0.53	0.264	0.829	0.255	0.816	0.264	0.825	*0.296*	*0.902*
0.6	0.1	0.60	0.66	0.271	0.881	0.446	0.982	0.399	0.969	*0.376*	*0.988*
0.1	0.2	0.60	0.70	0.497	0.985	0.443	0.973	0.468	0.980	*0.534*	*0.998*
0.1	0.5	0.45	0.78	0.526	0.997	0.542	0.996	0.562	0.997	*0.669*	*0.986*

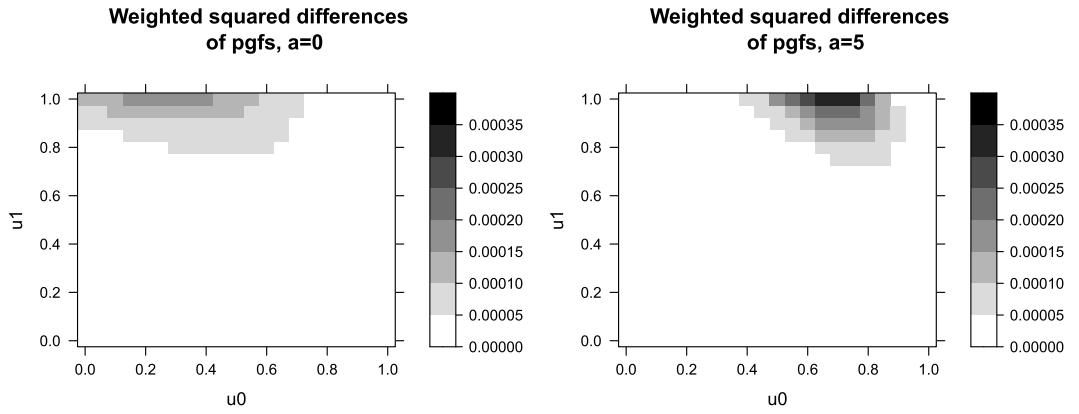

Figure 1. Heatmaps of the weighted squared difference of the two estimated pgfs for an INGARCH(1, 1) DGP (left: $a = 0$, right: $a = 5$).

Both models exhibit an autoregressive structure of order 1, but are distinct from an INAR model. The INARCH(1) process is a special case of the INGARCH(1, 1) process, i.e. $X_t|X_{t-1}, X_{t-2}, \ldots \sim \text{Poi}(M_t)$, where $M_t = \beta + \alpha X_{t-1}$. A Poi-DAR(1) process is characterized by $X_t = a_t X_{t-1} + b_t \varepsilon_t$, where $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \text{Poi}(\lambda)$ and $(a_t, b_t) \overset{\text{i.i.d.}}{\sim} \text{Mult}(1, \alpha, 1-\alpha)$. In this latter model class, each observation either chooses the previous observation or the innovation, so the stationary marginal distribution of the innovations equals that of the observations. We choose such values for λ and β to obtain similar observation means as for the other DGPs. For both DGPs, INARCH(1) and Poi-DAR(1), the power is larger for higher autocorrelation levels (provided that $a > 0$), see Tables 8 and 9. This is plausible since for high autocorrelation (and small innovation mean), an INAR(1) process tends to produce "runs", i.e., the same value is realized in consecutive time points. In contrast, the INARCH(1) model shows more erratic behavior. The Poi-DAR(1) process, on the other hand, exhibits even more extreme runs with higher autocorrelation combined with further "jumps" in-between these runs.

While we have high power in most scenarios, we also encounter some parameterizations with low power, e.g., for INARCH(1) processes with low autocorrelation α and high intercept β. In general, we seem to lose power for increasing mean of observations (due to additive terms rather than increasing autocorrelation). A higher observation mean leads to a wider range of observa-

Table 6. Power in case of a Poi(λ)-INAR(2) DGP when testing for H_0^{semi} in (1.4) with $p = 1$ using test statistic (2.10) with $s = 2$. Numbers in italic display the power values taken from Table 1 in Meintanis and Karlis (2014) who test for H_0^{para} in (1.3) with $p = 1$ and G_λ =Poi(λ) not using a higher-order test statistic analogously to (2.10).

			$a = 0$		$a = 5$		$a = 2$		*MK*, $a = 2$	
λ	α_1	α_2	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.3	0.1	0.068	0.205	0.079	0.391	0.082	0.354	*0.088*	*0.035*
1	0.5	0.1	0.060	0.141	0.080	0.311	0.072	0.251	*0.089*	*0.127*
1	0.5	0.3	0.114	0.401	0.328	0.958	0.240	0.848	*0.191*	*0.487*

Table 7. Power in case of an INGARCH(1, 1) DGP when testing for H_0^{semi} in (1.4) with $p = 1$ using test statistic (2.10) with $s = 2$. Numbers in italic display the power values taken from Table 1 in Meintanis and Karlis (2014) who test for H_0^{para} in (1.3) with $p = 1$ and G_λ =Poi(λ) not using a higher-order test statistic analogously to (2.10).

				$a = 0$		$a = 5$		$a = 2$		MK, $a = 2$	
β_0	β_1	α_1	$\rho(1)$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
0.2	0.4	0.10	0.09	0.045	0.101	0.052	0.119	0.048	0.116	*0.012*	*0.050*
0.2	0.4	0.20	0.21	0.090	0.403	0.099	0.449	0.100	0.467	*0.042*	*0.108*
1.0	0.1	0.50	0.52	0.068	0.178	0.166	0.691	0.123	0.528	*0.152*	*0.512*
0.5	0.1	0.50	0.52	0.141	0.628	0.249	0.884	0.222	0.846	*0.248*	*0.724*
0.1	0.4	0.40	0.53	0.449	0.989	0.366	0.970	0.416	0.983	*0.296*	*0.902*
0.6	0.1	0.60	0.66	0.216	0.812	0.440	0.987	0.376	0.973	*0.376*	*0.988*
0.1	0.2	0.60	0.70	0.566	0.996	0.489	0.989	0.534	0.994	*0.534*	*0.998*
0.1	0.5	0.45	0.78	0.697	1.000	0.717	1.000	0.749	1.000	*0.669*	*0.986*

tions, potentially affecting both the semi-parametric and the non-parametric estimation of the pgf, where the latter is also included in the parametric testing approaches. Besides the challenges related to semi- and non-parametric estimation, the considered DGPs themselves may contribute to low power results. As mentioned at the beginning of our paper, without restraining to a certain parametric family of innovations, the INAR model class is very flexible, the unspecified innovation distribution presents a high degree of freedom. Consider for example the INARCH(1) DGP with $\beta = 3$ and $\alpha = 0.3$ which leads to low power results as displayed in Table 8. Additionally, consider an INGARCH(1, 1) DGP with $\beta_0 = 0.6$, $\beta_1 = 0.1$ and $\alpha_1 = 0.6$ which leads to good power results as displayed in Table 5. For both DGPs, we simulated a sample of $n = 5000$ observations and computed the integrand of (2.8) with $a = 0$, i.e., $(\widehat{g}_{1;H_0}(u_0, u_1) - \widehat{g}_1(u_0, u_1))^2$, for $u_0, u_1 \in \{0, 0.05, 0.1, \ldots, 1\}$. For the sake of comparison, we also considered an INAR(1) DGP with $\lambda = 3$ and $\alpha = 0.3$. Figure 2 shows the boxplots of the resulting values for the three different DGPs. While the two-dimensional pgf of order 1 of the considered INGARCH(1, 1) DGP differs much from the one of a semi-parametrically estimated INAR(1) process, the first-order pgf of the considered INARCH(1) DGP is very close to the latter. Actually, $\widehat{g}_{1;H_0}(u_0, u_1)$ and $\widehat{g}_1(u_0, u_1)$ do not differ much, the differences are even as small as for the considered INAR(1) DGP. This explains the low power results and underlines the flexibility of the INAR model with unspecified innovation distribution.

Table 8. Power in case of an INARCH(1) DGP when testing for H_0^{semi} in (1.4) with $p = 1$.

		$a = 0$		$a = 5$		$a = 2$	
β	α	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.30	0.032	0.044	0.048	0.129	0.039	0.094
1	0.50	0.071	0.264	0.159	0.658	0.119	0.540
1	0.75	0.271	0.914	0.604	0.999	0.506	0.997
3	0.30	0.035	0.030	0.046	0.022	0.038	0.025
3	0.50	0.025	0.028	0.052	0.054	0.037	0.034
3	0.75	0.074	0.184	0.185	0.634	0.133	0.460

Table 9. Power in case of a Poi(λ)-DAR DGP when testing for H_0^{semi} in (1.4) with $p = 1$.

		$a = 0$		$a = 5$		$a = 2$	
λ	α	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
2	0.25	0.144	0.437	0.115	0.226	0.117	0.212
2	0.50	0.447	0.992	0.535	0.995	0.538	0.994
2	0.75	0.326	0.982	0.400	0.997	0.401	0.996
6	0.25	0.142	0.279	0.160	0.159	0.163	0.230
6	0.50	0.117	0.371	0.234	0.707	0.190	0.643
6	0.75	0.016	0.041	0.277	0.980	0.121	0.920

In summary, under the null, it becomes clear that we achieve better results when there is no weighting of the test statistic. However, since a higher weighting of the test statistic leads to substantially higher power values under the alternative and we still keep the level under the null using higher weighting, we recommend using the test with a comparatively high weighting, i.e. $a = 5$.

5. Conclusion

In existing literature, goodness-of-fit tests for the INAR model class are restricted to specific parametric families of innovation distributions. In this paper, we introduced a novel goodness-of-fit test for INAR(p) processes that does not rely on parametric assumptions about the nature of the innovations. We derived the limiting null distribution of our L_2-type test statistic based on weighted integrals using probability generating functions. Additionally, we proved its consistency under fixed alternatives, discussed the asymptotic behavior under local alternatives, and specified a bootstrap procedure required to circumvent the complex limiting distribution. In an extensive simulation study, we compared our proposed procedure with the parametric competitor of Meintanis and Karlis (2014). Overall, we got similar results, but we were able to increase the power against Markov chain alternatives of higher order by using also higher-order test statistics. Moreover, unlike the parametric approaches of Meintanis and Karlis (2014), we are able to test for general deviations from the INAR structure and got good power results for the considered DGPs. We noticed that, when testing with higher weight parameter a, the test exhibited higher power. Finally, we applied our method to three real data examples from economics (Faymonville, Jentsch and Weiß, 2025a).

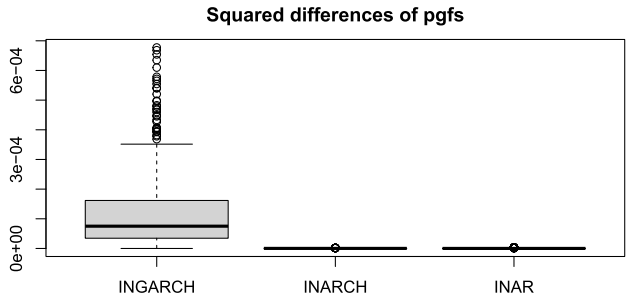

Figure 2. Boxplots of the squared differences of the two estimated pgfs for the considered INGARCH (left), INARCH (middle) and INAR (right) DGPs.

Acknowledgments

The authors thank the editor and the two referees for their useful comments on an earlier draft of this article. The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projeknummer 437270842.

Supplementary Material

Supplement I (DOI: 10.3150/24-BEJ1844SUPPA; .pdf). We provide three real-world data applications, additional tables and all the proofs of this paper.

Supplement II (DOI: 10.3150/24-BEJ1844SUPPB; .zip). We provide the MATLAB code for the real-world data applications.

References

Al-Osh, M.A. and Alzaid, A.A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *J. Time Series Anal.* **8** 261–275. MR0903755 https://doi.org/10.1111/j.1467-9892.1987.tb00438.x

Aleksandrov, B., Weiß, C.H. and Jentsch, C. (2022). Goodness-of-fit tests for Poisson count time series based on the Stein-Chen identity. *Stat. Neerl.* **76** 35–64. MR4374078 https://doi.org/10.1111/stan.12252

Aleksandrov, B., Weiß, C.H., Nik, S., Faymonville, M. and Jentsch, C. (2024). Modelling and diagnostic tests for Poisson and negative-binomial count time series. *Metrika* **87** 843–887. MR4795098 https://doi.org/10.1007/s00184-023-00934-0

Alzaid, A.A. and Al-Osh, M. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *J. Appl. Probab.* **27** 314–324. MR1052303 https://doi.org/10.2307/3214650

Armillotta, M. and Gorgi, P. (2024). Pseudo-variance quasi-maximum likelihood estimation of semi-parametric time series models. *J. Econometrics* **246** Paper No. 105894, 24. MR4832091 https://doi.org/10.1016/j.jeconom.2024.105894

Drost, F.C., van den Akker, R. and Werker, B.J.M. (2009). Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 467–485. MR2649605 https://doi.org/10.1111/j.1467-9868.2008.00687.x

Du, J.G. and Li, Y. (1991). The integer-valued autoregressive (INAR(p)) model. *J. Time Series Anal.* **12** 129–142. MR1108796 https://doi.org/10.1111/j.1467-9892.1991.tb00073.x

Faymonville, M., Jentsch, C. and Weiß, C.H. (2025a). Supplement I to "Semi-parametric goodness-of-fit testing for INAR models": Data examples, tables and proofs. https://doi.org/10.3150/24-BEJ1844SUPPB

Faymonville, M., Jentsch, C. and Weiß, C.H. (2025b). Supplement II to "Semi-parametric goodness-of-fit testing for INAR models": MATLAB code. https://doi.org/10.3150/24-BEJ1844SUPPA

Faymonville, M., Jentsch, C., Weiß, C.H. and Aleksandrov, B. (2023). Semiparametric estimation of INAR models using roughness penalization. *Stat. Methods Appl.* **32** 365–400. MR4606278 https://doi.org/10.1007/s10260-022-00655-0

Faymonville, M., Riffo, J., Rieger, J. and Jentsch, C. (2024). spINAR: An R package for semiparametric and parametric estimation and bootstrapping of integer-valued autoregressive (INAR) models. *J. Open Sour. Softw.* **9** 5386. https://doi.org/10.21105/joss.05386

Giacomini, R., Politis, D.N. and White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory* **29** 567–589. MR3064050 https://doi.org/10.1017/S0266466612000655

Gürtler, N. and Henze, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution. *J. Statist. Plann. Inference* **90** 207–225. MR1795597 https://doi.org/10.1016/S0378-3758(00)00114-2

Hudecová, Š., Hušková, M. and Meintanis, S.G. (2015). Tests for time series of counts based on the probability-generating function. *Statistics* **49** 316–337. MR3325362 https://doi.org/10.1080/02331888.2014.979826

Hudecova, S., Huskova, M. and Meintanis, S.G. (2021). Goodness-of-fit tests for bivariate time series of counts. *Econometrics* **9** 10. https://doi.org/10.3390/econometrics9010010

Jazi, M.A., Jones, G. and Lai, C.-D. (2012a). Integer valued AR(1) with geometric innovations. *J. Iran. Stat. Soc. (JIRSS)* **11** 173–190. MR3010343

Jazi, M.A., Jones, G. and Lai, C.-D. (2012b). First-order integer valued AR processes with zero inflated Poisson innovations. *J. Time Series Anal.* **33** 954–963. MR2991911 https://doi.org/10.1111/j.1467-9892.2012.00809.x

Jentsch, C. and Weiss, C.H. (2019). Bootstrapping INAR models. *Bernoulli* **25** 2359–2408. MR3961251 https://doi.org/10.3150/18-BEJ1057

Leucht, A. (2012). Characteristic function-based hypothesis tests under weak dependence. *J. Multivariate Anal.* **108** 67–89. MR2903134 https://doi.org/10.1016/j.jmva.2012.02.003

Leucht, A. and Neumann, M.H. (2013). Degenerate U- and V-statistics under ergodicity: Asymptotics, bootstrap and applications in statistics. *Ann. Inst. Statist. Math.* **65** 349–386. MR3011626 https://doi.org/10.1007/s10463-012-0374-9

Liu, Z., Li, Q. and Zhu, F. (2021). Semiparametric integer-valued autoregressive models on $\mathbb{Z}$. *Canad. J. Statist.* **49** 1317–1337. MR4349647 https://doi.org/10.1002/cjs.11621

McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resour. Bull.* **21** 645–650. https://doi.org/10.1111/j.1752-1688.1985.tb05379.x

Meintanis, S.G. and Karlis, D. (2014). Validation tests for the innovation distribution in INAR time series models. *Comput. Statist.* **29** 1221–1241. MR3266056 https://doi.org/10.1007/s00180-014-0488-z

Meintanis, S. and Swanepoel, J. (2007). Bootstrap goodness-of-fit tests with estimated parameters based on empirical transforms. *Statist. Probab. Lett.* **77** 1004–1013. MR2380538 https://doi.org/10.1016/j.spl.2007.01.014

Qi, X., Li, Q. and Zhu, F. (2019). Modeling time series of count with excess zeros and ones based on INAR(1) model with zero-and-one inflated Poisson innovations. *J. Comput. Appl. Math.* **346** 572–590. MR3864182 https://doi.org/10.1016/j.cam.2018.07.043

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Savani, V. and Zhigljavsky, A.A. (2007). Efficient parameter estimation for independent and INAR(1) negative binomial samples. *Metrika* **65** 207–225. MR2288059 https://doi.org/10.1007/s00184-006-0071-x

Schweer, S. (2016). A goodness-of-fit test for integer-valued autoregressive processes. *J. Time Series Anal.* **37** 77–98. MR3439533 https://doi.org/10.1111/jtsa.12138

Steutel, F.W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Ann. Probab.* **7** 893–899. MR0542141

The MathWorks Inc. (2022). MATLAB version: 9.13.0.

Weiß, C.H. (2018). *An Introduction to Discrete-Valued Time Series*, 1st ed. Wiley. https://doi.org/10.1002/9781119097013

Received February 2024 and revised December 2024

Submitted to Bernoulli

Supplement to "Semi-parametric goodness-of-fit testing for INAR models"

MAXIME FAYMONVILLE* [1,a], CARSTEN JENTSCH [1,b] and CHRISTIAN H. WEISS [2,c]

[1]*Department of Statistics, TU Dortmund University,* [a]*faymonville@statistik.tu-dortmund.de,* [b]*jentsch@statistik.tu-dortmund.de*
[2]*Department of Mathematics and Statistics, Helmut-Schmidt-University Hamburg,* [c]*weissc@hsu-hh.de*

Appendix A: Real-world data applications

To illustrate the application of our proposed goodness-of-fit test, we apply it on three economic real-world data examples using $B = 1000$ bootstrap repetitions. The corresponding MATLAB code is provided in further supplementary material (Faymonville, Jentsch and Weiß, 2024).

The first data set is sourced from Baker Hughes[1] containing weekly counts of active rotary drilling rigs. These counts serve as indicator for the demand of products used in drilling, well completion, oil production, and hydrocarbon processing and have been published since 1944. We specifically focus on the number of drilling rigs in Alaska from 1991 to 1997 ($n = 417$). This data set has been addressed before in Weiß (2018). Figure 1 shows a plot of the time series and the corresponding ACF and PACF. The characteristic INAR runs suggest a high serial dependence with small innovation mean, which are confirmed by the high and slowly decreasing autocorrelation level. Looking at the partial autocorrelation function (PACF), an AR(1)-like model seems to be an appropriate fit. Indeed, when applying our test on the data using $a = 5$, we do not reject the null of an INAR(1) process at 5% level.

The second data set was provided by the Deutsche Börse Group and has been discussed before in Homburg et al. (2021). It contains counts per trading day of transactions of structured products between February 2017 and August 2019 ($n = 404$). The data are displayed in Figure 2. As in the previous example of rig counts, the ACF and PACF suggest that an INAR(1) model might be an appropriate fit for the data. But applying the test of Meintanis and Karlis (2014), it rejects the null of a Poi-INAR(1) model at 5% level ($a = 2$). Our test, by contrast, also using $a = 2$, does not reject the INAR(1) null at 5% level. These different results may be explained by the dispersion of the data. With $\bar{x} \approx 1.47$ and $s^2 \approx 2.23$, the index of dispersion is approximately 1.51 suggesting overdispersed counts. This is additionally stressed out by Figure 3. It displays the semi-parametric (left plot) and parametric estimation of the innovation distribution, where for the latter we used a Poisson distribution (in the middle) and a geometric distribution (on the right), respectively. We see that the parametrically estimated (equidispersed) Poisson distribution differs much from the semi-parametrically estimated innovation distribution whereas the (overdispersed) geometric distribution seems much more appropriate as innovation distribution. Hence, while the parametric test of Meintanis and Karlis (2014) fails to detect the INAR model structure due to their too restrictive equidispersion property of the data under the null, our more flexible test does not reject the INAR structure.

In our third application, we consider a data set first published by Brännäs and Quoreshi (2010). It records the number of transaction of the Ericsson B stock per minute between 9:35 and 17:44. Originally, it provides data for the days between July 2 and 22, 2002. Fokianos, Rahbek and Tjøstheim

[1] phx.corporate-ir.net/phoenix.zhtml?c=79687&p=irol-rigcountsoverview.

2

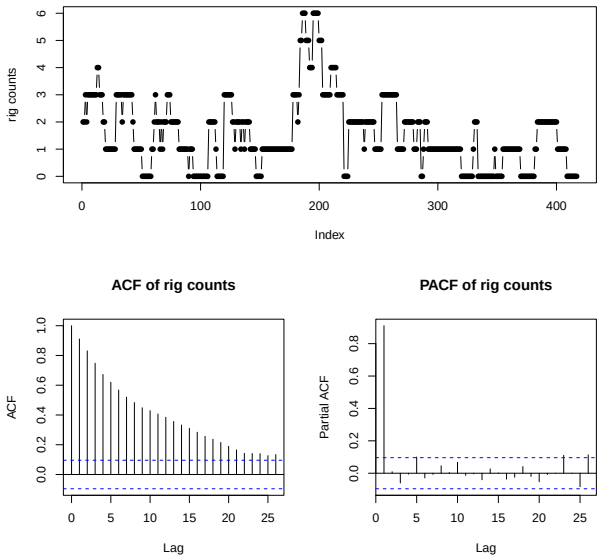

Figure 1. Plot of rig counts and the corresponding (P)ACF.

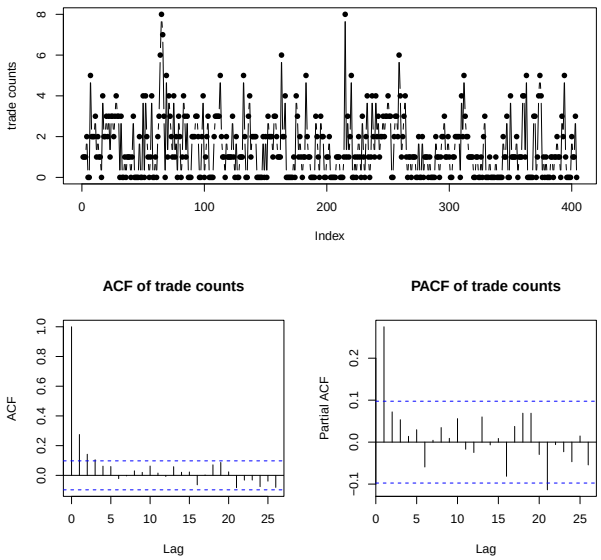

Figure 2. Plot of daily trade counts and the corresponding (P)ACF.

(2009), Zhu (2012), Christou and Fokianos (2015), Davis and Liu (2016),Weiß (2018) and Su and Zhu (2022) exclusively consider the data of July 2 and model the data by an INGARCH$(1,1)$ process. The resulting time series is of length $n = 460$. Figure 4 shows a plot of this time series along with the corre-

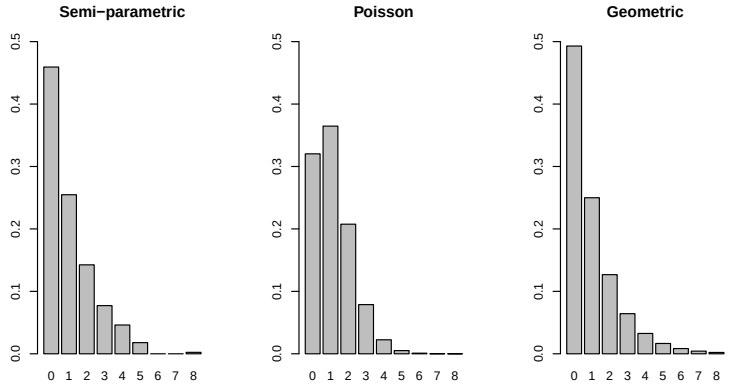

Figure 3. Plot of the semi-parametrically (left) and parametrically (Poisson in the middle, Geometric on the right) estimated innovation distribution.

sponding (P)ACF. By contrast to the first two examples, this data set exhibits a more erratic structure. The ACF is slowly decaying and the PACF suggests dependencies of higher order than 1. Applying our goodness-of-fit test to the null H_0^{semi} in (1.4) with $p = 1$, we are initially not able to reject the null at level 5%. However, when using the second-order test statistic $T_n^{(2)}$ ($s = 2$ in (2.10)), we ultimately reject the null (both with $a = 5$). This is plausible since we capture dependencies of higher order by considering the three-dimensional pgf of order 2, i.e., g_2, instead of the two-dimensional pgf of order 1, i.e., g_1.

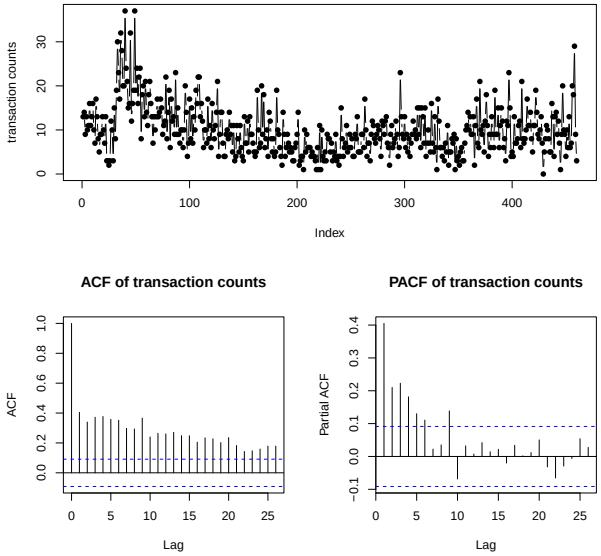

Figure 4. Plot of transaction counts and the corresponding (P)ACF.

4

Appendix B: Additional tables

Table 1. Computing time (in seconds) in case of a Poi-INAR(1) DGP, testing for H_0^{semi} in (1.4) with $p = 1$.

		$a = 0$		$a = 5$		$a = 2$	
λ	α	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.3	0.109	0.176	0.071	0.168	0.059	0.184
1	0.5	0.060	0.214	0.050	0.166	0.058	0.164
3	0.3	0.103	0.254	0.091	0.256	0.076	0.275
3	0.5	0.083	0.281	0.100	0.230	0.150	0.312

Table 2. Computing time (in seconds) in case of a Poi-INAR(2) DGP, testing for H_0^{semi} in (1.4) with $p = 2$.

			$a = 0$		$a = 5$		$a = 2$	
λ	α_1	α_2	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.3	0.1	5.450	11.910	2.970	6.344	4.424	11.102
1	0.5	0.1	5.611	18.004	3.630	9.614	4.545	16.480
1	0.5	0.3	8.123	32.553	4.396	17.867	6.796	19.555

Table 3. Actual sizes in case of a Poi(λ)-INAR(1) DGP when testing for H_0^{semi} in (1.4) with $p = 1$.

		$a = 0$		$a = 5$		$a = 2$	
λ	α	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.1	0.027	0.031	0.024	0.015	0.025	0.022
1	0.9	0.042	0.050	0.051	0.047	0.046	0.047
3	0.1	0.037	0.040	0.034	0.030	0.035	0.032
3	0.9	0.031	0.027	0.033	0.026	0.031	0.027

Table 4. Power in case of a Poi(λ)-INAR(2) DGP when testing for H_0^{semi} in (1.4) with $p = 1$.

			$a = 0$		$a = 5$		$a = 2$	
λ	α_1	α_2	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.05	0.05	0.039	0.034	0.031	0.020	0.033	0.022
1	0.4	0.5	0.103	0.311	0.198	0.657	0.164	0.538

Table 5. Actual sizes in case of a Poi(λ)-INAR(1) DGP when testing for H_0^{semi} in (1.4) with $p = 1$ using test statistic (2.10) with $s = 2$. Numbers in italic display the power values taken from Table 1 in Meintanis and Karlis (2014) who test for H_0^{para} in (1.3) with $p = 1$ and G_λ=Poi(λ) not using a higher-order test statistic analogously to (2.10).

		$a = 0$		$a = 5$		$a = 2$		*MK*, $a = 2$	
λ	α	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.3	0.047	0.047	0.047	0.041	0.044	0.044	*0.049*	*0.055*
1	0.5	0.039	0.047	0.042	0.051	0.038	0.050	*0.051*	*0.053*
3	0.3	0.027	0.035	0.048	0.036	0.042	0.037	*0.056*	*0.047*
3	0.5	0.027	0.030	0.043	0.037	0.038	0.037	*0.055*	*0.059*

Table 6. Actual sizes in case of a NB(N, π)-INAR(1) DGP when testing for H_0^{semi} in (1.4) with $p = 1$ using test statistic (2.10) with $s = 2$. Numbers in italic display the power values taken from Table 1 in Meintanis and Karlis (2014) who test for H_0^{para} in (1.3) with $p = 1$ and G_λ=Poi(λ) not using a higher-order test statistic analogously to (2.10).

			$a = 0$		$a = 5$		$a = 2$		*MK*, $a = 2$	
N	π	α	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	1/2	0.5	0.053	0.053	0.052	0.051	0.053	0.052	*0.686*	*1.000*
2	2/3	0.5	0.046	0.050	0.048	0.050	0.049	0.050	*0.327*	*0.897*
10	10/11	0.5	0.045	0.044	0.047	0.047	0.045	0.041	*0.120*	*0.149*

Table 7. Power in case of a Poi(λ)-INAR(2) DGP when testing for H_0^{semi} in (1.4) with $p = 1$ using test statistic (2.10) with $s = 2$.

			$a = 0$		$a = 5$		$a = 2$	
λ	α_1	α_2	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
1	0.05	0.05	0.042	0.087	0.041	0.114	0.040	0.111
1	0.4	0.5	0.131	0.515	0.384	0.983	0.268	0.900

Appendix C: Proofs

C.1. Proof of Lemma 2.1

Exploiting the INAR(p) model structure of $X_t, \ldots, X_{t-p}$, we get

$$E\left(\prod_{j=0}^{p} u_j^{X_{t-j}}\right) = E\left(E\left(\prod_{j=0}^{p} u_j^{X_{t-j}} | X_{t-1}, \ldots, X_{t-p}\right)\right)$$
$$= E\left(\prod_{j=1}^{p} u_j^{X_{t-j}} E\left(u_0^{X_t} | X_{t-1}, \ldots, X_{t-p}\right)\right).$$

Taking a closer look at the interior conditional expectation and inserting the model equation, we get

$$E\left(u_0^{X_t} | X_{t-1}, \ldots, X_{t-p}\right) = E\left(u_0^{\alpha_1 \circ X_{t-1} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t} | X_{t-1}, \ldots, X_{t-p}\right)$$

6

$$
\begin{aligned}
&= \left\{ \prod_{i=1}^{p} E\left(u_0^{\alpha_i \circ X_{t-i}} | X_{t-1}, \ldots, X_{t-p} \right) \right\} \cdot E(u_0^{\varepsilon_t}) \\
&= \left\{ \prod_{i=1}^{p} E\left(u_0^{\alpha_i \circ X_{t-i}} | X_{t-i} \right) \right\} \cdot E(u_0^{\varepsilon_t}) \\
&= \left\{ \prod_{i=1}^{p} \left(1 + \alpha_i (u_0 - 1) \right)^{X_{t-i}} \right\} \cdot g_\varepsilon(u_0),
\end{aligned}
$$

where we have used that given $X_{t-1}, \ldots, X_{t-p}$, the thinning operations and the innovation ε_t are independent. The thinning operation $\alpha_i \circ$ only depends on X_{t-i} and $\alpha_i \circ X_{t-i} | X_{t-i} \sim \text{Bin}(X_{t-i}, \alpha_i)$ together with the well-known formula for the pgf of a binomial distribution. Furthermore, we define $g_\varepsilon(u_0) := E(u_0^{\varepsilon_t})$ as the marginal pgf of the innovations. Hence, we get the following representation for the joint pgf of $X_t, \ldots, X_{t-p}$:

$$
\begin{aligned}
g_p(u_0, \ldots, u_p) &:= E\left(\prod_{j=0}^{p} u_j^{X_{t-j}} \right) \\
&= E\left(\prod_{j=1}^{p} u_j^{X_{t-j}} \left\{ \prod_{i=1}^{p} (1 + \alpha_i (u_0 - 1))^{X_{t-i}} \right\} \cdot g_\varepsilon(u_0) \right) \\
&= g_\varepsilon(u_0) \cdot E\left(\prod_{j=1}^{p} u_j^{X_{t-j}} \left\{ \prod_{i=1}^{p} (1 + \alpha_i (u_0 - 1))^{X_{t-i}} \right\} \right) \\
&= g_\varepsilon(u_0) \cdot E\left(\prod_{j=1}^{p} \left\{ u_j \left(1 + \alpha_j (u_0 - 1) \right) \right\}^{X_{t-j}} \right).
\end{aligned}
$$

□

C.2. Proof of Lemma 2.3

Plugging-in (2.5) and (2.7) and rearranging terms, we get

$$
\begin{aligned}
T_n =& n \int_0^1 \cdots \int_0^1 \left(\widehat{g}_{p;H_0}(u_0, \ldots, u_p) - \widehat{g}_p(u_0, \ldots, u_p) \right)^2 w(u_0, \ldots, u_p; a) \, du_0 \cdots du_p \\
=& \frac{n}{(n-p)^2} (a+1)^{p+1} \sum_{t,s=p+1}^{n} \int_0^1 \cdots \int_0^1 \left(\prod_{j=1}^{p} u_j^{X_{t-j}} \right) \left(\prod_{j=1}^{p} u_j^{X_{s-j}} \right) \left(\prod_{j=1}^{p} u_j^{a} \right) du_1 \cdots du_p \\
& \times \int_0^1 \left(u_0^{X_t} - \widehat{g}_\varepsilon(u_0) \prod_{j=1}^{p} \left(1 + \widehat{\alpha}_{\text{sp},j} (u_0 - 1) \right)^{X_{t-j}} \right) \\
& \left(u_0^{X_s} - \widehat{g}_\varepsilon(u_0) \prod_{j=1}^{p} \left(1 + \widehat{\alpha}_{\text{sp},j} (u_0 - 1) \right)^{X_{s-j}} \right) u_0^a \, du_0
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{n}{(n-p)^2}(a+1)^{p+1}\sum_{t,s=p+1}^{n}\left(\prod_{j=1}^{p}\int_0^1 u_j^{X_{t-j}+X_{s-j}+a}\,du_j\right)\\
&\times\int_0^1\Bigg(u_0^{X_t+X_s+a}+\widehat{g}_\varepsilon(u_0)^2\,u_0^a\prod_{j=1}^{p}\left(1+\widehat{\alpha}_{\mathrm{sp},j}(u_0-1)\right)^{X_{t-j}+X_{s-j}}\\
&-\widehat{g}_\varepsilon(u_0)\,u_0^{X_s+a}\prod_{j=1}^{p}\left(1+\widehat{\alpha}_{\mathrm{sp},j}(u_0-1)\right)^{X_{t-j}}-\widehat{g}_\varepsilon(u_0)\,u_0^{X_t+a}\prod_{j=1}^{p}\left(1+\widehat{\alpha}_{\mathrm{sp},j}(u_0-1)\right)^{X_{s-j}}\Bigg)du_0.
\end{aligned}
$$

Furthermore, by using that

$$
\int_0^1 u^x\,du=1/(1+x) \tag{C.1}
$$

holds for $x\neq -1$, T_n can be simplified to get

$$
\begin{aligned}
T_n=&\frac{n}{(n-p)^2}(a+1)^{p+1}\sum_{t,s=p+1}^{n}\left(\prod_{j=1}^{p}\frac{1}{1+X_{t-j}+X_{s-j}+a}\right)\\
&\times\Bigg(\frac{1}{1+X_t+X_s+a}+\int_0^1\widehat{g}_\varepsilon(u_0)^2u_0^a\prod_{j=1}^{p}\left(1+\widehat{\alpha}_{\mathrm{sp},j}(u_0-1)\right)^{X_{t-j}+X_{s-j}}du_0\\
&-2\int_0^1\widehat{g}_\varepsilon(u_0)\,u_0^{X_s+a}\prod_{j=1}^{p}\left(1+\widehat{\alpha}_{\mathrm{sp},j}(u_0-1)\right)^{X_{t-j}}du_0\Bigg).
\end{aligned}
$$

To be able to use the same integration rule (C.1) also for the two remaining integrals above, we need to isolate the terms in u_0. On the one hand, we have

$$
\widehat{g}_\varepsilon(u_0)^2=\sum_{k_1,k_2=0}^{\max(X_1,\ldots,X_n)}\widehat{G}_{\mathrm{sp}}(k_1)\widehat{G}_{\mathrm{sp}}(k_2)u_0^{k_1+k_2}.
$$

By using twice the binomial theorem in each case, on the other hand, we get

$$
\left(1+\widehat{\alpha}_{\mathrm{sp},j}(u_0-1)\right)^{X_{t-j}+X_{s-j}}=\sum_{i_j=0}^{X_{t-j}+X_{s-j}}\binom{X_{t-j}+X_{s-j}}{i_j}\widehat{\alpha}_{\mathrm{sp},j}^{i_j}\sum_{h_j=0}^{i_j}\binom{i_j}{h_j}u_0^{h_j}(-1)^{i_j-h_j}
$$

and

$$
\left(1+\widehat{\alpha}_{\mathrm{sp},j}(u_0-1)\right)^{X_{t-j}}=\sum_{i_j=0}^{X_{t-j}}\binom{X_{t-j}}{i_j}\widehat{\alpha}_{\mathrm{sp},j}^{i_j}\sum_{h_j=0}^{i_j}\binom{i_j}{h_j}u_0^{h_j}(-1)^{i_j-h_j},
$$

respectively. Altogether, this leads to

$$
T_n=\frac{n}{(n-p)^2}(a+1)^{p+1}\left(\prod_{j=1}^{p}\frac{1}{1+X_{t-j}+X_{s-j}+a}\right)\Bigg[\frac{1}{1+X_t+X_s+a}
$$

8

$$+ \sum_{k_1,k_2=0}^{\max(X_1,\ldots,X_n)} \widehat{G}_{\mathrm{sp}}(k_1)\widehat{G}_{\mathrm{sp}}(k_2) \sum_{i_1=0}^{X_{t-1}+X_{s-1}} \sum_{h_1=0}^{i_1} \cdots \sum_{i_p=0}^{X_{t-p}+X_{s-p}} \sum_{h_p=0}^{i_p}$$

$$\prod_{j=1}^{p} \binom{X_{t-j}+X_{s-j}}{i_j} \widehat{\alpha}_{\mathrm{sp},j}^{i_j} (-1)^{i_j-h_j} \binom{i_j}{h_j} \int_0^1 u_0^{1+k_1+k_2+a+\sum_{m=1}^{p} h_m} du_0 - 2 \sum_{k=0}^{\max(X_1,\ldots,X_n)} \widehat{G}_{\mathrm{sp}}(k)$$

$$\sum_{i_1=0}^{X_{t-1}+X_{s-1}} \sum_{h_1=0}^{i_1} \cdots \sum_{i_p=0}^{X_{t-p}+X_{s-p}} \sum_{h_p=0}^{i_p} \prod_{j=1}^{p} \binom{X_{t-j}}{i_j} \widehat{\alpha}_{\mathrm{sp},j}^{i_j} (-1)^{i_j-h_j} \binom{i_j}{h_j} \int_0^1 u_0^{1+k+X_s+a+\sum_{m=1}^{p} h_m} du_0 \Bigg].$$

Applying (C.1) again, the assertion follows. □

C.3. Proof of Proposition 2.4

The only aspect that remains to show is the degeneracy. Let h be the kernel defined in (2.12). Using $y_t = (x_t, \ldots, x_{t-p}) \in \mathbb{N}_0^{p+1}$, under the null $H_0^{semi}(\theta_0)$ in (2.11), we have

$$\begin{aligned}
E(h(y_t, Y_s; \theta_0)) = &\int_0^1 \cdots \int_0^1 \left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{x_{t-j}} - \prod_{j=0}^{p} u_j^{x_{t-j}} \right) \\
&\times E\left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{X_{s-j}} - \prod_{j=0}^{p} u_j^{X_{s-j}} \right) \\
&w(u_0, \ldots, u_p; a)\, du_0 \ldots du_p \\
= &\int_0^1 \cdots \int_0^1 \left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{x_{t-j}} - \prod_{j=0}^{p} u_j^{x_{t-j}} \right) \\
&\times \left(pgf_{H_0}(X_s, \ldots, X_{s-p}) - pgf(X_s, \ldots, X_{s-p}) \right) \\
&w(u_0, \ldots, u_p; a)\, du_0 \ldots du_p \\
= &\, 0,
\end{aligned}$$

because $pgf_{H_0}(X_s, \ldots, X_{s-p}) = pgf(X_s, \ldots, X_{s-p})$ under $H_0^{semi}(\theta_0)$. That is, we are in the case of a degenerate kernel. □

C.4. Proof of Theorem 2.5

Using Proposition 2.4, we already have the degeneracy of the kernel (2.12). In view of Theorem 1 in Leucht and Neumann (2013), under $H_0^{semi}(\theta_0)$, the kernel h even fulfills the stronger degeneracy condition

$$E(h(y_t, Y_s; \theta_0) | Y_1, \ldots, Y_{s-1})$$

$$= \int_0^1 \dots \int_0^1 \left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{x_{t-j}} - \prod_{j=0}^{p} u_j^{x_{t-j}} \right) \tag{C.2}$$

$$\times E\left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{X_{s-j}} - \prod_{j=0}^{p} u_j^{X_{s-j}} | Y_1, \dots, Y_{s-1} \right)$$

$$w(u_0, \dots, u_p; a)\, du_0 \dots du_p$$

$$= 0.$$

By using the Markov property of an INAR process, this is the case, because

$$E\left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{X_{s-j}} - \prod_{j=0}^{p} u_j^{X_{s-j}} | Y_1, \dots, Y_{s-1} \right)$$

$$= E\left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{X_{s-j}} - \prod_{j=0}^{p} u_j^{X_{s-j}} | X_{1-p}, \dots, X_{s-1} \right)$$

$$= E\left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{X_{s-j}} | X_{1-p}, \dots, X_{s-1} \right) - E\left(\prod_{j=0}^{p} u_j^{X_{s-j}} | X_{1-p}, \dots, X_{s-1} \right)$$

$$= g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{X_{s-j}} - \prod_{j=1}^{p} u_j^{X_{s-j}} E(u_0^{X_s} | X_{1-p}, \dots, X_{s-1})$$

$$= g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{X_{s-j}} - \prod_{j=1}^{p} u_j^{X_{s-j}} g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (1+\alpha_{0,j}(u_0-1))^{X_{s-j}}$$

$$= 0,$$

where we used the null $H_0^{semi}(\theta_0)$ to get $E(u_0^{X_s} | X_{1-p}, \dots, X_{s-1}) = g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (1+\alpha_{0,j}(u_0-1))^{X_{s-j}}$.

Furthermore, because the kernel h is of a quadratic form, it is positive semidefinite. That is, for all $m \in \mathbb{N}$ and for all $c_1, \dots, c_m \in \mathbb{R}$, $y_1, \dots, y_m \in \mathbb{N}_0^{p+1}$, it holds $\sum_{t,s=1}^{m} c_t c_s h(y_t, y_s) \geq 0$. Moreover, because all the terms of the integrand, i.e., $g_\varepsilon(u_0), u_0, \dots, u_p, \alpha_1, \dots, \alpha_p$, are bounded by 0 from below and by 1 from above, we have

$$E(h(Y_0, Y_0; \theta_0)) = \int_0^1 \dots \int_0^1 E\left(\left(g_{0,\varepsilon}(u_0) \prod_{j=1}^{p} (u_j(1+\alpha_{0,j}(u_0-1)))^{X_{t-j}} - \prod_{j=0}^{p} u_j^{X_{t-j}} \right)^2 \right)$$

$$w(u_0, \dots, u_p; a)\, du_0 \dots du_p$$

$$< \infty.$$

10

With $(X_t)_{t\in\mathbb{Z}}$ being a strictly stationary and ergodic process (Du and Li, 1991), we can apply Theorem 1 of Leucht and Neumann (2013). Precisely, under the null $H_0^{semi}(\theta_0)$ and for $n \to \infty$, this leads to

$$T_n(\theta_0) \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k Z_k^2,$$

where $(Z_k)_k$ is a sequence of independent standard normal random variables and $(\lambda_k)_k$ the sequence of nonzero eigenvalues of (2.13) enumerated according their multiplicity with $(\Phi_k)_k$ the associated orthonormal eigenfunctions. □

C.5. Proof of Theorem 2.6

Let $\theta_0 = (\boldsymbol{\alpha}_0, G_0) \in A \times \mathcal{G}$ denote the true parameter. For any $\theta \in \Theta$, we see that the kernel h can be represented as

$$h(x, y; \theta) = \int_{[0,1]^{p+1}} h_1(x, \mathbf{u}; \theta) h_1(y, \mathbf{u}; \theta) Q(d\mathbf{u}),$$

where $\mathbf{u} = (u_0, \ldots, u_p)$, $h_1 : \mathbb{R}^{p+1} \times [0,1]^{p+1} \times \Theta \to \mathbb{R}$, $\Theta = (0,1)^p \times \mathcal{G}$ with

$$h_1(y_t, \mathbf{u}; \theta) = g_\varepsilon(u_0) \prod_{j=1}^{p} (u_j(1 + \alpha_j(u_0 - 1)))^{x_{t-j}} - \prod_{j=0}^{p} u_j^{x_{t-j}} \tag{C.3}$$

with $y_t = (x_t, \ldots, x_{t-p})$ and the probability measure Q has probability density function w, that is, $dQ/d\mathbf{u} = w(\mathbf{u})$. Using that both terms of the difference in (C.3) only take values in $[0,1]$, $\boldsymbol{\alpha} \in (0,1)^p$, $G \in \mathcal{G}$ and that the integration limits of all the following integrals are 0 and 1, we see that (C.3) fulfills

$$\int_{[0,1]^{p+1}} h_1(y, \mathbf{u}; \theta_0)^2 Q(d\mathbf{u}) < \infty, \qquad \int_{[0,1]^{p+1}} E_{\theta_0}\left(h_1(Y_0, \mathbf{u}; \theta_0)^2\right) Q(d\mathbf{u}) < \infty$$

as well as the continuity condition

$$\int_{[0,1]^{p+1}} \left(h_1(y, \mathbf{u}; \theta_0) - h_1(\widetilde{y}, \mathbf{u}; \theta_0)\right)^2 Q(d\mathbf{u}) \to 0 \quad \text{for } \widetilde{y} - y \to 0.$$

Due to (C.2), we can conclude that $E_{\theta_0}(h_1(Y_t, \mathbf{u}; \theta_0) | Y_{t-1}, Y_{t-2}, \ldots) = 0$ holds as well.[2] Furthermore, the function h_1 in (C.3) is continuously differentiable with respect to θ. For the derivatives with respect to the model coefficients α_l, $l \in \{1, \ldots, p\}$, we have

$$\begin{aligned} \frac{\partial}{\partial \alpha_l} h_1(y_t, \mathbf{u}; \theta) &= \frac{\partial}{\partial \alpha_l} g_\varepsilon(u_0) \prod_{j=1}^{p} (u_j(1 + \alpha_j(u_0 - 1)))^{x_{t-j}} \\ &= g_\varepsilon(u_0) \left(\prod_{j=1,\, j\neq l}^{p} (u_j(1 + \alpha_j(u_0 - 1)))^{x_{t-j}} \right) \frac{\partial}{\partial \alpha_l} (u_l(1 + \alpha_l(u_0 - 1)))^{x_{t-l}} \end{aligned} \tag{C.4}$$

[2] We point out that in the original paper of Leucht and Neumann (2013), this condition contains a small typo.

$$= g_\varepsilon(u_0)\left(\prod_{j=1,\, j\neq l}^{p}(u_j(1+\alpha_j(u_0-1)))^{x_{t-j}}\right)x_{t-l}(u_l(1+\alpha_l(u_0-1)))^{x_{t-l}-1}u_l(u_0-1)$$

$$= \frac{g_\varepsilon(u_0)(u_0-1)}{1+\alpha_l(u_0-1)}x_{t-l}\prod_{j=1}^{p}(u_j(1+\alpha_j(u_0-1)))^{x_{t-j}}.$$

Recalling that $g_\varepsilon(u_0) = \sum_{k=0}^{\infty} G(k)u_0^k$, for the partial derivatives with respect to the entries of the pmf of the innovation distribution G, that is, $(G(k),\, k \in \mathbb{N}_0)$, we get

$$\frac{\partial}{\partial G(k)}h_1(y_t, \mathbf{u}, \theta) = \frac{\partial}{\partial G(k)}g_\varepsilon(u_0)\prod_{j=1}^{p}(u_j(1+\alpha_j(u_0-1)))^{x_{t-j}} \tag{C.5}$$

$$= u_0^k\prod_{j=1}^{p}(u_j(1+\alpha_j(u_0-1)))^{x_{t-j}},$$

which does not depend anymore on $(G(k),\, k \in \mathbb{N}_0)$. With the same arguments as used for (C.3), also $\dot{h}_1$, the Fréchet derivative of h_1 with respect to θ, fulfills

$$E_{\theta_0}\left(\int_{[0,1]^{p+1}}||\dot{h}_1(Y_0, \mathbf{u}; \theta_0)||_2^2 Q(d\mathbf{u})\right) < \infty,$$

$$\int_{[0,1]^{p+1}}||\dot{h}_1(y, \mathbf{u}; \theta_0) - \dot{h}_1(\widetilde{y}, \mathbf{u}; \theta_0)||_2^2 Q(d\mathbf{u}) \to 0 \quad \text{for } \widetilde{y} - y \to 0.$$

Moreover, $\dot{h}_1$ fulfills a Lipschitz-type condition in θ which we outline in the following. For simplicity, we set $p = 1$ but the subsequent arguments can be extended to higher order $p > 1$. Denote $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{G})$. Then, we get

$$\dot{h}_1(\cdot,\cdot,\theta) - \dot{h}_1(\cdot,\cdot,\tilde{\theta}) = \ddot{h}_1(\cdot,\cdot,\breve{\theta})(\theta - \tilde{\theta}),$$

where $\breve{\theta}$ is between θ and $\widetilde{\theta}$ and

$$\ddot{h}_1(\cdot,\cdot,\breve{\theta})(\theta - \tilde{\theta}) = \frac{\partial \dot{h}_1(\cdot,\cdot,\theta)}{\partial \alpha}_{|\theta=\breve{\theta}}(\alpha - \tilde{\alpha}) + \sum_{k=0}^{\infty}\frac{\partial \dot{h}_1(\cdot,\cdot,\theta)}{\partial G(k)}_{|\theta=\breve{\theta}}(G(k) - \tilde{G}(k))$$

with

$$\frac{\partial \dot{h}_1(Y_t, \mathbf{u}, \theta)}{\partial \alpha}_{|\theta=\breve{\theta}} = \begin{pmatrix} \breve{g}_\varepsilon(u_0)(u_0-1)^2 X_{t-1}u_1^{X_{t-1}}(X_{t-1}-1)(1+\breve{\alpha}(u_0-1))^{X_{t-1}-2} \\ u_0^0 u_1^{X_{t-1}} X_{t-1}(1+\breve{\alpha}(u_0-1))^{X_{t-1}-1}(u_0-1) \\ u_0^1 u_1^{X_{t-1}} X_{t-1}(1+\breve{\alpha}(u_0-1))^{X_{t-1}-1}(u_0-1) \\ \vdots \end{pmatrix} \tag{C.6}$$

12

and

$$\frac{\partial \dot{h}_1(Y_t,\mathbf{u},\theta)}{\partial G(k)}_{|\theta=\breve{\theta}} = \begin{pmatrix} u_0^k(u_0-1)X_{t-1}u_1^{X_{t-1}}(1+\breve{\alpha}(u_0-1))^{X_{t-1}-1} \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \quad k \in \mathbb{N}_0, \tag{C.7}$$

where we used the previous calculations of (C.4) and (C.5). For the first entry of (C.6), we have

$$\breve{g}_\varepsilon(u_0)(u_0-1)^2X_{t-1}u_1^{X_{t-1}}(X_{t-1}-1)(1+\breve{\alpha}(u_0-1))^{X_{t-1}-2} \leq X_{t-1}(X_{t-1}-1) \leq X_{t-1}^2$$

and for all other entries (which are equal to the first entry of (C.7), that is, for all $j \in \mathbb{N}$, we have

$$u_0^j(u_0-1)X_{t-1}u_1^{X_{t-1}}(1+\breve{\alpha}(u_0-1))^{X_{t-1}-1} \leq X_{t-1} \leq X_{t-1}^2.$$

Hence, using the notation $\|(a_n)_{n\in\mathbb{N}}\|_1 = \sum_{n=1}^\infty |a_n|$, we get

$$\begin{aligned} ||\dot{h}_1(Y_t,\mathbf{u},\theta) - \dot{h}_1(Y_t,\mathbf{u},\widetilde{\theta})||_1 &= ||\ddot{h}_1(Y_t,\mathbf{u},\widetilde{\theta})(\theta-\widetilde{\theta})||_1 \\ &\leq \left\| \begin{pmatrix} X_{t-1}^2(\alpha-\widetilde{\alpha}) + \sum_{k=0}^\infty X_{t-1}^2(G(k)-\widetilde{G}(k)) \\ X_{t-1}^2(G(0)-\widetilde{G}(0)) \\ X_{t-1}^2(G(1)-\widetilde{G}(1)) \\ \vdots \end{pmatrix} \right\|_1 \\ &\leq X_{t-1}^2 \left\| \begin{pmatrix} ||\theta-\widetilde{\theta}||_1 \\ ||G-\widetilde{G}||_1 \end{pmatrix} \right\|_1 \\ &\leq 2X_{t-1}^2||\theta-\widetilde{\theta}||_1, \end{aligned}$$

as $||G-\widetilde{G}||_1 \leq ||\theta-\widetilde{\theta}||_1$. That is, the last part of Assumption (A2) (iv) in Leucht and Neumann (2013) holds for existing second moments of X_t ensured by Assumption 1. Additionally, Drost, Van den Akker and Werker (2009) show that their proposed estimator (2.3) is regular and consistent and consequently exhibits the Bahadur-type expansion

$$\widehat{\theta}_{\text{sp}} = \theta_0 + \frac{1}{n}\sum_{t=1}^n l_t + o_p(n^{-1/2}),$$

with $l_t = L(Y_t, Y_{t-1}, \ldots)$ for some measurable function L, $E_{\theta_0}(l_t|Y_{t-1}, Y_{t-2}, \ldots) = 0$ and $E_{\theta_0}(||l_t||_2^2) < \infty$, where we refer to Section 5.3 of Van der Vaart (2000) proving this result for the class of M-estimators (including ML-estimators) under mild regularity conditions. Hence, all assumptions of Proposition 1 in Leucht and Neumann (2013), which is based on De Wet and Rangles (1987), are fulfilled and we get that $T_n(\widehat{\theta}_{\text{sp}})$ has the same limiting distribution as $\frac{1}{n}\sum_{t=1}^n\sum_{s=1}^n \widetilde{h}(\widetilde{Y}_t, \widetilde{Y}_s; \theta_0)$, where $\widetilde{Y}_t = (Y_t', l_t')'$ and

$$\widetilde{h}(x,y;\theta_0) = \int_{[0,1]^{p+1}} \big(h_1(x_1,\mathbf{u};\theta_0) + E_{\theta_0}(\dot{h}_1(Y_1,\mathbf{u};\theta_0)x_2)\big)\big(h_1(y_1,\mathbf{u};\theta_0) + E_{\theta_0}(\dot{h}_1(Y_1,\mathbf{u};\theta_0)y_2)\big)Q(d\mathbf{u}). \tag{C.8}$$

Altogether, using Theorem 1 of Leucht and Neumann (2013), we get

$$T_n = T_n(\widehat{\theta}_{\text{sp}}) \xrightarrow{d} \sum_{k=1}^{\infty} \widetilde{\lambda}_k Z_k^2,$$

where $(Z_k)_k$ is as before and $(\widetilde{\lambda}_k)_k$ denotes the sequence of nonzero eigenvalues of the equation (2.16) enumerated according their multiplicity with $(\widetilde{\Phi}_k)_k$ the associated orthonormal eigenfunctions. □

C.6. Proof of Theorem 2.8

Let $p \in \mathbb{N}$ and suppose we observe data $X_1, \ldots, X_n$ from some (strictly) stationary count time series process $(X_t, t \in \mathbb{Z})$ under the alternative H_1^{semi} in (2.18) such that (2.19), (2.20) and (2.21) hold. Then, by adding suitable zeros $g_{p;H_0}(\mathbf{u}) - g_{p;H_0}(\mathbf{u})$ and $g_p(\mathbf{u}) - g_p(\mathbf{u})$, where $\mathbf{u} = (u_0, \ldots, u_p)$, to the integrand of the test statistic T_n from (2.8) and expanding the squared term in brackets, we get

$$\begin{aligned}
T_n =& n \int_0^1 \cdots \int_0^1 \left(\widehat{g}_{p;H_0}(\mathbf{u}) - \widehat{g}_p(\mathbf{u})\right)^2 w(\mathbf{u}; a)\, du_0 \cdots du_p \\
=& n \int_0^1 \cdots \int_0^1 \left(\widehat{g}_{p;H_0}(\mathbf{u}) + \left(g_{p;H_0}(\mathbf{u}) - g_{p;H_0}(\mathbf{u})\right) - \widehat{g}_p(\mathbf{u}) + \left(g_p(\mathbf{u}) - g_p(\mathbf{u})\right)\right)^2 \\
& \qquad w(\mathbf{u}; a)\, du_0 \cdots du_p \\
=& n \int_0^1 \cdots \int_0^1 \left(\widehat{g}_{p;H_0}(\mathbf{u}) - g_{p;H_0}(\mathbf{u}) - \widehat{g}_p(\mathbf{u}) + g_p(\mathbf{u})\right)^2 w(\mathbf{u}; a)\, du_0 \cdots du_p \\
& + n \int_0^1 \cdots \int_0^1 \left(g_{p;H_0}(\mathbf{u}) - g_p(\mathbf{u})\right)^2 w(\mathbf{u}; a)\, du_0 \cdots du_p \\
& + n \int_0^1 \cdots \int_0^1 \left(\widehat{g}_{p;H_0}(\mathbf{u}) - g_{p;H_0}(\mathbf{u}) - \widehat{g}_p(\mathbf{u}) + g_p(\mathbf{u})\right)\left(g_{p;H_0}(\mathbf{u}) - g_p(\mathbf{u})\right) w(\mathbf{u}; a)\, du_0 \cdots du_p \\
=:& A_1 + A_2 + A_3
\end{aligned}$$

with an obvious notation for A_i, $i = 1, 2, 3$. When discussing these three terms separately, we see that the integrand of the first term A_1 is appropriately centered such that A_1 represents again a degenerate V-statistic (just with a different kernel) that converges to a (non-degenerate) χ^2-type distribution by making use of (2.21). The second term A_2 diverges to $+\infty$ as, for all $\mathbf{u} \in [0,1]^{p+1}$, $g_{p;H_0}(\mathbf{u}) - g_p(\mathbf{u})$ converges to a non-zero limit such that its square becomes a function that will be strictly positive on some subset of $[0,1]^{p+1}$ with strictly positive Lebesgue measure according to (2.20). This leads to an integral that is also strictly positive and as it is inflated with n, this second term diverges to $+\infty$ with rate n. Finally, the third term is a mixed term, where the difference in the first brackets, that is, $\widehat{g}_{p;H_0}(\mathbf{u}) - g_{p;H_0}(\mathbf{u}) - \widehat{g}_p(\mathbf{u}) + g_p(\mathbf{u})$, is of order $O(1/\sqrt{n})$. This is multiplied with $g_{p;H_0}(\mathbf{u}) - g_p(\mathbf{u})$, which is a function that is strictly positive on some set with positive Lebesgue measure. In total, the integral behaves like $O_P(1/\sqrt{n})$ such that the whole third term, when inflated with n behaves like $O_P(\sqrt{n})$. As this is slower than the rate $O(n)$ obtained for the second term, altogether, we have $T_n = O_P(n)$. That is, T_n diverges to $+\infty$ in probability such that, for all $\gamma \in (0,1)$, we have $E(\varphi_n) = P(T_n > q_{1-\gamma}) \longrightarrow 1$. Hence, this proves consistency of the test against fixed alternatives. □

14

C.7. Proof of Theorem 2.11

Let $p \in \mathbb{N}$ and suppose we observe data $X_{n,1}, \ldots, X_{n,n}$ from a triangular array $(X_{n,t}, t = 1, \ldots, n,\ n \in \mathbb{N})$ of count time series and, for each fixed n, $X_{n,1}, \ldots, X_{n,n}$ is generated from a stationary Markov chain of order p with state space $\mathcal{S} \subseteq \mathbb{N}_0$ generated under the alternative H_1^{semi} such that (2.24) with $a_n = n^{-1/2}$ as well as (2.25) hold. Then, similar to the proof of Theorem 2.8, by adding suitable zeros $g_{p;H_0,n}(\mathbf{u}) - g_{p;H_0,n}(\mathbf{u})$ and $g_{p,n}(\mathbf{u}) - g_{p,n}(\mathbf{u})$ to the integrand of the test statistic T_n from (2.8) and expanding the squared term in brackets, we get

$$
\begin{aligned}
T_n =& n \int_0^1 \cdots \int_0^1 \Big(\widehat{g}_{p;H_0,n}(\mathbf{u}) - \widehat{g}_{p,n}(\mathbf{u}) \Big)^2 w(\mathbf{u}; a)\, du_0 \cdots du_p \\
=& n \int_0^1 \cdots \int_0^1 \Big(\widehat{g}_{p;H_0,n}(\mathbf{u}) + \big(g_{p;H_0,n}(\mathbf{u}) - g_{p;H_0,n}(\mathbf{u}) \big) - \widehat{g}_{p,n}(\mathbf{u}) + \big(g_{p,n}(\mathbf{u}) - g_{p,n}(\mathbf{u}) \big) \Big)^2 \\
& \qquad w(\mathbf{u}; a)\, du_0 \cdots du_p \\
=& \int_0^1 \cdots \int_0^1 \Big(\sqrt{n} \big(\widehat{g}_{p;H_0,n}(\mathbf{u}) - \widehat{g}_{p,n}(\mathbf{u}) - \big(g_{p;H_0,n}(\mathbf{u}) - g_{p,n}(\mathbf{u}) \big) \big) + \sqrt{n} \big(g_{p;H_0,n}(\mathbf{u}) - g_{p,n}(\mathbf{u}) \big) \Big)^2 \\
& \qquad w(\mathbf{u}; a)\, du_0 \cdots du_p \\
\xrightarrow{d}& \int_0^1 \cdots \int_0^1 \Big(G(\mathbf{u}) + C(\mathbf{u}) \Big)^2 w(\mathbf{u}; a)\, du_0 \cdots du_p,
\end{aligned}
$$

where we used (2.24) with $a_n = 1/\sqrt{n}$ and (2.25). Otherwise, if $a_n \to 0$ such that $\sqrt{n}\, a_n \to \infty$, the test φ_n remains consistent, that is, we have $E(\varphi_n) \to 1$ as $n \to \infty$. If $a_n = o(n^{-1/2})$, the test φ_n has no asymptotic power, that is, we have $E(\varphi_n) \to \gamma$ as $n \to \infty$.

□

References

BRÄNNÄS, K. and QUORESHI, A. M. M. S. (2010). Integer-valued moving average modelling of the number of transactions in stocks. *Appl. Financ. Econ.* **20** 1429–1440. https://doi.org/10.1080/09603107.2010.498343

CHRISTOU, V. and FOKIANOS, K. (2015). Quasi-likelihood inference for negative binomial time series models. *J. Time Series Anal.* **35** 55–78. https://doi.org/10.1111/jtsa.12050

DAVIS, R. A. and LIU, H. (2016). Theory and inference for a class of nonlinear models with application to time series of counts. *Statist. Sinica* **26** 1673–1707. https://doi.org/10.5705/SS.2014.145T

DE WET, T. and RANGLES, R. (1987). On the effect of substituting parameter estimators in limiting χ^2 U and V statistics. *Ann. Statist.* **15** 398–412. https://doi.org/10.1214/aos/1176350274

DROST, F., VAN DEN AKKER, R. and WERKER, B. (2009). Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 467–485. https://doi.org/10.1111/j.1467-9868.2008.00687.x

DU, J. G. and LI, Y. (1991). The integer valued autoregressive (INAR(p)) model. *J. Time Series Anal.* **12** 129–142. https://doi.org/10.1111/j.1467-9892.1991.tb00073.x

FAYMONVILLE, M., JENTSCH, C. and WEISS, C. H. (2024). Supplement II to "Semi-parametric goodness-of-fit testing for INAR models": MATLAB code.

FOKIANOS, K., RAHBEK, A. and TJØSTHEIM, D. (2009). Poisson autoregression. *J. Amer. Statist. Assoc.* **104** 1430–1439. https://doi.org/10.1198/jasa.2009.tm08270

HOMBURG, A., WEISS, C., FRAHM, G., ALWAN, L. C. and GÖB, R. (2021). Analysis and forecasting of risk in count processes. *J. Risk Financial Manag.* **14** 182. https://doi.org/10.3390/jrfm14040182

LEUCHT, A. and NEUMANN, M. (2013). Degenerate U- and V-statistics under ergodicity: asymptotics, bootstrap and applications in statistics. *Ann. Inst. Statist. Math.* **65** 349–386. https://doi.org/10.1007/s10463-012-0374-9

MEINTANIS, S. G. and KARLIS, D. (2014). Validation tests for the innovation distribution in INAR time series models. *Comput. Statist.* **29** 1221–1241. https://doi.org/10.1007/s00180-014-0488-z

SU, B. and ZHU, F. (2022). Temporal aggregation and systematic sampling for INGARCH processes. *J. Statist. Plann. Inference* **219** 120–133. https://doi.org/10.1016/j.jspi.2021.12.002

VAN DER VAART, A. (2000). *Asymptotic Statistics*, 1 ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511802256

WEISS, C. H. (2018). *An Introduction to Discrete-Valued Time Series*, 1 ed. Wiley. https://doi.org/10.1002/9781119097013

ZHU, F. (2012). Modeling time series of counts with COM-Poisson INGARCH models. *Math Comput. Model.* **56** 191–203. https://doi.org/10.1016/j.mcm.2011.11.069

Predictive inference for discrete-valued time series

Maxime Faymonville[1] Carsten Jentsch[2] Efstathios Paparoditis[3]

Abstract

For discrete-valued time series, predictive inference cannot be implemented through the construction of prediction intervals to some predetermined coverage level, as this is the case for real-valued time series. To address this problem, we propose to reverse the construction principle by considering preselected sets of interest and estimating the probability that a future observation of the process falls into these sets. The accuracy of the prediction is then evaluated by quantifying the uncertainty associated with estimation of these predictive probabilities. We consider parametric and non-parametric approaches and derive asymptotic theory for the estimators involved. Suitable bootstrap approaches to evaluate the distribution of the estimators considered also are introduced. They have the advantage to imitate the distributions of interest under different possible settings, including the practical important case where uncertainty holds true about the correctness of a parametric model used for prediction. Theoretical justification of the bootstrap is given, which also requires investigation of asymptotic properties of parameter estimators under model misspecification. We elaborate on bootstrap implementations under different scenarios and focus on parametric prediction using INAR and INARCH models and (conditional) maximum likelihood estimators. Simulations investigate the finite sample performance of the predictive method developed and applications to real life data sets are presented.

Keywords: Bootstrap, discrete-valued time series, INAR model, INARCH model, inference, prediction

[1]Department of Statistics, TU Dortmund University, D-44221 Dortmund, Germany; faymonville@statistik.tu-dortmund.de; corresponding author

[2]Department of Statistics, TU Dortmund University, D-44221 Dortmund, Germany; jentsch@statistik.tu-dortmund.de

[3]Cyprus Academy of Sciences, Letters, and Arts, CY-1011 Nikosia, Cyprus; stathisp@ucy.ac.cy

2

1. Introduction

The class of discrete-valued time series models is vast; see Weiß (2018) and Davis et al. (2016) for an overview. A first family of such time series are so-called count time series which include, e.g., the INAR and the INARCH models that will be further discussed in Section 3. A second family are categorical time series (which also include ordinal time series), when the data exhibit a qualitative range consisting of a finite number of categories; see e.g., Pruscha (1993), Fokianos and Kedem (2003), Weiß (2020) and Liu et al. (2022). A special case of categorical time series are binary time series, where the observations take two possible values only; see e.g. Kedem and Fokianos (2002) and Jentsch and Reichmann (2021).

In the continuous time series setting, predictive inference commonly deals with point prediction which is accompanied by a prediction interval, where the goal of the latter is to take into account the uncertainty associated with the point prediction. Given time series data up to some time point n, that is, given $X_1, \ldots, X_n$, such prediction intervals are designed to cover future observations, say X_{n+h}, for some $h \in \mathbb{N}$, with a desired (high) probability. The main criterion to judge the quality of such prediction intervals is its asymptotic validity, which is specified, e.g., in Pan and Politis (2016) as follows. Given time series $X_1, \ldots, X_n$, suppose that the task is to predict one step ahead, that is, to predict X_{n+1}. Then, for $\beta \in (0, 1)$, the interval $[l_n, u_n]$ is called an asymptotically valid $(1 - \beta)$ prediction interval for X_{n+1} if

$$P(l_n \leq X_{n+1} \leq u_n | X_1, \ldots, X_n) \to 1 - \beta \quad \text{for } n \to \infty. \tag{1.1}$$

In the formulation above, $l_n = l_n(X_1, \ldots, X_n)$ and $u_n = u_n(X_1, \ldots, X_n)$ typically are functions of the observed sample $X_1, \ldots, X_n$.

In the context of discrete-valued time series, the above concept of an (asymptotically) valid prediction interval is not applicable since *intervals* do not account for the discrete nature of the data. In fact and for general choices of $\beta \in (0, 1)$, it is typically not possible to construct an interval which guarantees an exact (or asymptotic) level $1 - \beta$. To address this problem, in the special case of count time series, Freeland and McCabe (2004) introduced the notion of *coherent forecasting* according to which predicted values of count processes should also be counts themselves. In compliance with this notion, we could rather consider prediction *sets* than intervals. These can be obtained by looking at the set $[l_n, u_n] \cap range(X_{n+1})$. However, also for such prediction sets, it is in general not even asymptotically possible to achieve validity in the sense of (1.1). This is illustrated by the following example.

Example 1.1. *Let $X_1, \ldots, X_n$ be an ordinal stationary time series stemming from a first-order Markov process with $range(X_t) = \{0, 1, \ldots, 5\}$. Think of X_t as measuring the daily air quality level in different cities; see the data example of Liu et al. (2022) and the analysis by Jahn and*

3

Weiß (2024). Suppose that for some $x_n \in \{0, 1, \ldots, 5\}$*, e.g., for* $x_n = 2$*, the process has the following one step transition probabilities:* $P(X_{n+1} = 0|X_n = 2) = 0.58$*,* $P(X_{n+1} = 1|X_n = 2) = 0.2$*,* $P(X_{n+1} = 2|X_n = 2) = 0.11$*,* $P(X_{n+1} = 3|X_n = 2) = 0.09$ *and* $P(X_{n+1} = 4|X_n = 2) = P(X_{n+1} = 5|X_n = 2) = 0.01$*. Then, it is not possible to obtain a prediction set for* X_{n+1} *given that* $X_n = 2$ *with an exact or asymptotic coverage of* 95%*; see Table 1 for some exemplary prediction sets.*

$X_{n+1} \in$	Set	Cov	Set	Cov	Set	Cov	Set	Cov
	$\{1, 2, 3\}$	0.40	$\{0, 1, 2\}$	0.89	$\{0, 1, 2, 4, 5\}$	0.91	$\{0, 1, 2, 3\}$	0.98

TABLE 1. Coverages (Cov) for different exemplary prediction sets.

A way to achieve coherent forecasting for the setup of count processes is to return the full predictive probability mass function of X_{n+h} given $X_1, \ldots, X_n$, as proposed by Homburg et al. (2023). However, this does not account for the estimation uncertainty that comes with it. In this paper, we propose an alternative strategy to solve the problem of coherent forecasting by transforming the predictive inference problem into a parameter estimation problem and by constructing a corresponding *confidence* interval for the underlying parameter at some desired level $1-\delta$, where $\delta \in (0, 1)$. To be more specific, let $S \subset \mathbb{R}$ be any user-selected subset of possible values of interest that X_{n+h} can take. We propose to estimate the predictive probability of this set, that is, the probability that X_{n+h} takes a value in S given the observed stretch $X_1, \ldots, X_n$ of the process. The uncertainty associated with such a prediction can then be accounted for by constructing a $(1 - \delta)$-confidence interval for the corresponding predictive probability. Clearly, if $range(X_{n+h})$ is known, $S \subset range(X_{n+h})$ is a typical choice, which we assume to be the case for simplicity in what follows. Note that $|S| = \infty$ is also allowed.

In Section 2, we elaborate on this general idea. In Section 2.2, we derive some asymptotic theory for the corresponding prediction problem, where we concentrate on parametric and non-parametric implementations of this kind of predictive inference. As we will see, using the asymptotic distributions derived for the calculation of the described confidence intervals turns out to be cumbersome in practice. To circumvent this problem, we resort to bootstrap techniques. In particular, we use bootstrapping to construct *confidence* intervals of the predictive probabilities of interest under a variety of relevant settings. In Section 2.3, we propose different bootstrap algorithms and we also address the important problem of model uncertainty (in case the decision of the user is in favor of a parametric model) by providing a bootstrap procedure which takes this problem into account in constructing the confidence interval of interest. To derive theoretical results in this context, the investigation of the consistency and the distributional

4

properties of estimators under model misspecification is required, a topic which is of interest on its own. In Section 3, we concentrate on asymptotic and on bootstrap-based approaches for the case of (discrete-valued) count time series, where predictive inference is implemented using (conditional) maximum likelihood estimation and (first-order) INAR and INARCH models. Section 4 discusses some practical issues while Section 5 presents simulation results demonstrating the capabilities of the different asymptotic and bootstrap methods proposed. Section 6 gives applications to real-life data sets and Section 7 concludes the findings of our paper. All technical proofs are deferred to the Appendix.

2. Prediction for Discrete-Valued Time Series

2.1. **Predictive Probabilities.** Suppose we observe a sample $X_1, \ldots, X_n$ from a strictly stationary, discrete-valued process $(X_t, t \in \mathbb{Z})$, that is $range(X_0) \subset \mathbb{R}$ is a countable set. Without loss of generality, let $range(X_0) = \{y_k, k \in \mathcal{N}\}$, where $\mathcal{N} \subseteq \mathbb{N}$. Our goal is to predict the future value X_{n+h} of the process for some $h \in \mathbb{N}$. In this setup and for any $S \subset \mathbb{R}$, we denote by

$$P_{S,\underline{x}_n}^{(h)} = P(X_{n+h} \in S | \underline{X}_n = \underline{x}_n)$$

the *h-step predictive probability* of the subset S given that

$$\underline{X}_n := (X_1, \ldots, X_n) = (x_1, \ldots, x_n) =: \underline{x}_n.$$

$P_{S,\underline{x}_n}^{(h)}$ is the conditional probability that the future random variable X_{n+h} takes a value in the subset S given the realization $\underline{x}_n$ of $\underline{X}_n$. Let $\delta \in (0,1)$ be small. A $(1-\delta)$-confidence interval for $P_{S,\underline{x}_n}^{(h)}$, that is, an interval $[L_n^{(h)}, U_n^{(h)}]$ with the property

$$P\big(L_n^{(h)} \leq P_{S,\underline{x}_n}^{(h)} \leq U_n^{(h)}\big) = 1 - \delta,$$

is called a $(1-\delta)$ *confidence interval for the h-step predictive probability* of the subset S. If the above statement holds true for $n \to \infty$, that is, if

$$\lim_{n\to\infty} P\big(L_n^{(h)} \leq P_{S,\underline{x}_n}^{(h)} \leq U_n^{(h)}\big) = 1 - \delta,$$

the corresponding interval is called an *asymptotic* $(1-\delta)$ *confidence interval for the h-step predictive probability* of the subset S.

In the following, we concentrate on the case where the underlying process is a first-order, discrete-valued Markov process and $h = 1$. In this case $P_{S,\underline{x}_n}^{(h)} = P_{S,x_n}^{(h)}$ and we write for simplicity, since $h = 1$,

$$P_{S,x_n} := P(X_{n+1} \in S | X_n = x_n) = \sum_{k \in \mathcal{N}: y_k \in S} P(X_{n+1} = y_k | X_n = x_n).$$

5

We assume throughout this paper that $x_n \in range(X_0)$ is a given value which coincides with the value of the last observation of the time series $X_1, \ldots, X_n$ at hand. Furthermore and in order to simplify notation, we sometimes also write $P_{X_{t+1}=y_k|X_t=x_n}$ for the one step transition probability $P(X_{t+1} = y_k|X_t = x_n)$. Note that our derivations can easily be extended to Markov models of higher order and to larger prediction horizons, that is, $h > 1$; we refer to Section 4.2 for more details. We start by imposing the following assumption.

Assumption 1 (Markov process)**.** $(X_t, t \in \mathbb{Z})$ *is a discrete-valued, homogeneous, aperiodic, irreducible, positive recurrent and geometrically ergodic first-order Markov process.*

Remark 2.1. *Assumption 1 covers a wide range of discrete time series. It even includes appropriate nominal time series since we do not require an order of the potentially involved categories of the data. An example for the latter time series are DAR processes; see Jacobs and Lewis (1983) for details.*

As already mentioned, our aim is to construct a confidence interval for the (one step ahead) predictive probability of a set $S \subset \mathbb{R}$ of interest, which has an (asymptotic) coverage of $1 - \delta$ for some given $\delta \in (0, 1)$. In the next two subsections, we will consider parametric and non-parametric approaches towards this goal.

2.2. **Asymptotic Confidence Intervals.**

2.2.1. *Parametric Prediction.* A common situation in time series analysis occurs when a parametric model is used to perform prediction. In our context, the predictive probability P_{S,x_n} is estimated using a parametric, first-order Markov model, the properties of which will be specified in Assumption 2 below. In order to emphasize the dependence of the prediction on the used model and on the associated parameter vector denoted by θ, we write $(X_t(\theta), t \in \mathbb{Z})$ for the parametric model used for prediction. Note that we do not want to assume that the discrete-valued time series observed necessarily stems from the parametric process $(X_t(\theta), t \in \mathbb{Z})$ which will be used for prediction. This seems important in order to address the practically important situation of model misspecification in implementing the prediction and in constructing the confidence interval for the predictive probability. This situation occurs when the parametric model class used for prediction does not necessarily coincide with the data generating process. The following (high-level) assumption specifies our requirements including those needed to obtain an asymptotically valid confidence interval for the predictive probability of interest.

Assumption 2 (Parametric family of Markov processes)**.**

(i) $\mathcal{M}_\theta := \{(X_t(\theta), t \in \mathbb{Z}), \theta \in \Theta\}$ *is a parametric family of stationary, discrete-valued, aperiodic, positive recurrent and geometrically ergodic, first-order Markov process with*

6

parameter vector $\theta \in \Theta \subset \mathbb{R}^d$ for some finite $d \in \mathbb{N}$ and Θ a compact set. Furthermore, $\theta_1 \neq \theta_2$ implies $(X_t(\theta_1), t \in \mathbb{Z}) \neq (X_t(\theta_2), t \in \mathbb{Z})$.

(ii) There exists a unique $\theta_0 \in \Theta$ such that, when fitting a model from the family $\mathcal{M}_\theta$ to the process $(X_t, t \in \mathbb{Z})$, the estimator $\widehat{\theta}$ used fulfills

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_{\theta_0}),$$

where V_{θ_0} is a positive-definite covariance matrix.

Under the above assumption, we write $P^{(para)}_{S,x_n}(\theta_0) := P_{S,x_n}$ for the (parametric) predictive probability of the set S when model class $\mathcal{M}_\theta$ is used for prediction. Then,

$$P^{(para)}_{S,x_n}(\widehat{\theta}) = \sum_{y_k \in S} P^{(para)}_{X_{t+1}=y_k|X_t=x_n}(\widehat{\theta}),$$

is the estimator of $P^{(para)}_{S,x_n}(\theta_0)$ obtained by replacing θ_0 by $\widehat{\theta}$. Here, $P^{(para)}_{X_{t+1}=y_k|X_t=x_n}(\theta_0)$ denotes the one step transition probability associated with the parametric Markov process $(X_t(\theta_0), t \in \mathbb{Z}) \in \mathcal{M}_\theta$. Now assuming that $P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\theta)$ is continuous differentiable with respect to θ, we get

$$\sqrt{n}\big(P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\widehat{\theta}) - P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\theta_0)\big) = \nabla_\theta P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\theta')\sqrt{n}(\widehat{\theta} - \theta_0)^\top$$

for some θ' such that $\|\theta' - \theta_0\| \leq \|\widehat{\theta} - \theta_0\|$. Here and for any $\theta_1 \in \Theta$, $\nabla_\theta P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\theta_1)$ denotes the d-dimensional vector of partial derivatives of $P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\theta)$ with respect to θ and evaluated at θ_1. We then get by Assumption 2 and assuming continuity of the partial derivatives, that as $n \to \infty$,

$$\sqrt{n}\big(P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\widehat{\theta}) - P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\theta_0)\big) \xrightarrow{d} \mathcal{N}\Big(0, \sigma^2_S(\theta_0)\Big), \tag{2.1}$$

where

$$\sigma^2_S(\theta_0) := \nabla^\top_\theta P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\theta_0)\, V_{\theta_0}\, \nabla_\theta P^{(para)}_{X_{t+1}\in S|X_t=x_n}(\theta_0).$$

The following result easily follows.

Proposition 2.2 (Asymptotic parametric confidence interval for $P^{(para)}_{S,x_n}(\theta_0)$)**.** *Suppose Assumption 2 holds true and let $V_{\widehat{\theta}}$ be a consistent estimator of V_{θ_0} obtained by replacing θ_0 by $\widehat{\theta}$. Suppose that $P^{(para)}_{S,x_n}(\theta)$ is continuously differentiable around θ_0 and that $\nabla_\theta P^{(para)}_{S,x_n}(\theta_0) \neq 0$. Then,*

$$\left[P^{(para)}_{S,x_n}(\widehat{\theta}) - \frac{z_{\delta/2}\sigma_S(\widehat{\theta})}{\sqrt{n}}, P^{(para)}_{S,x_n}(\widehat{\theta}) + \frac{z_{\delta/2}\sigma_S(\widehat{\theta})}{\sqrt{n}}\right]$$

is an asymptotically valid confidence interval for the predictive probability $P^{(para)}_{S,x_n}(\theta_0)$ with confidence level $1-\delta$, where $\sigma_S(\widehat{\theta}) = \sqrt{\sigma^2_S(\widehat{\theta})}$ and $\sigma^2_S(\widehat{\theta})$ is the estimator of $\sigma^2_S(\theta_0)$ obtained by replacing θ_0 by $\widehat{\theta}$. Furthermore, $z_{\delta/2}$ denotes the upper $\delta/2$-quantile of the standard Gaussian distribution.

Remark 2.3.

(i) Note that in the case of model misspecification, which is allowed by Assumption 2, θ_0 is the parameter vector from the parameter space Θ which best fits the underlying process $(X_t, t \in \mathbb{Z})$ and that the particular value of θ_0 also depends on the estimation method used to obtain $\widehat{\theta}$. Furthermore, the same will be true for the (limiting) distribution of the estimator $\widehat{\theta}$ and in particular for the variance V_{θ_0} of this limiting distribution. We elaborate in Section 3 on the important case where $\widehat{\theta}$ is a (conditional) maximum likelihood estimator.

(ii) The construction of the asymptotic confidence interval for the predictive probability $P_{S,x_n}^{(para)}(\theta_0)$ relies on the estimator $\sigma_S^2(\widehat{\theta})$, which is typically difficult to obtain in practice due to the need to calculate the unknown quantities $\nabla_\theta P_{S,x_n}^{(para)}(\widehat{\theta})$ and $V_{\widehat{\theta}}$. In Section 2.3, we will introduce bootstrap procedures to estimate these quantities as well as the distribution of $P_{S,x_n}^{(para)}(\widehat{\theta})$.

(iii) Assumption 2 (ii) can be generalized to allow for an infinite dimensional parameter space Θ. For example, this is relevant for the semi-parametric INAR model investigated by Drost et al. (2009) and Faymonville and Jentsch (2025).

2.2.2. *Non-parametric Prediction.* Alternatively to the situation discussed in Section 2.2.1, we may consider a fully non-parametric approach to estimate the predictive probability and to obtain the corresponding confidence interval. In the absence of parametric assumptions, we can use the appropriate relative frequencies in order to estimate P_{S,x_n}. In the following, we denote the predictive probability P_{S,x_n} by $P_{S,x_n}^{(npara)}$ in order to distinguish it from the parametric case discussed in the previous section. A natural estimator of $P_{S,x_n}^{(npara)}$ is then obtained as follows:

$$\widehat{P}_{X_{t+1}\in S|X_t=x_n}^{(npara)} = \begin{cases} \frac{\sum_{t=1}^{n-1} \mathbf{1}_{\{X_{t+1}\in S, X_t=x_n\}}}{\sum_{t=1}^{n-1} \mathbf{1}_{\{X_t=x_n\}}} & \text{if } \sum_{t=1}^{n-1} \mathbf{1}_{\{X_t=x_n\}} \neq 0, \\ 0 & \text{if } \sum_{t=1}^{n-1} \mathbf{1}_{\{X_t=x_n\}} = 0. \end{cases} \tag{2.2}$$

Notice that by Assumption 1, the estimator $\widehat{P}_{X_{t+1}\in S|X_t=x_n}^{(npara)}$ converges in probability to $P_{X_{t+1}\in S|X_t=x_n}$ as $n \to \infty$; see Derman (1956), where also a limiting Gaussian distribution for these estimators has been established. An asymptotically valid $(1-\delta)$-confidence interval for $P_{S,x_n}^{(npara)}$ is given in the following proposition.

Proposition 2.4 (Asymptotic non-parametric confidence interval for $P_{S,x_n} = P_{S,x_n}^{(npara)}$)**.** *Suppose that Assumption 1 holds true. Then, an asymptotically valid confidence interval for $P_{S,x_n}^{(npara)}$*

8

Prediction \ Data	$X_1, \ldots, X_n$ stems from the class $\mathcal{M}_\theta$	$X_1, \ldots, X_n$ doesn't stem from the class $\mathcal{M}_\theta$
Non-parametric $\widehat{P}_{S,x_n}^{(npara)}$	Model is ignored	"Correct" prediction
Parametric $P_{S,x_n}^{(para)}(\widehat{\theta})$	"Correct" prediction	Model is misspecified

TABLE 2. Possible constellations that may occur when using a parametric or non-parametric estimator of predictive probabilities.

with confidence level $1-\delta$ *is given by*

$$\left[\widehat{P}_{S,x_n}^{(npara)} - \frac{z_{\delta/2}\widehat{\sigma}_S}{\sqrt{n-1}}, \widehat{P}_{S,x_n}^{(npara)} + \frac{z_{\delta/2}\widehat{\sigma}_S}{\sqrt{n-1}}\right],$$

where $\widehat{\sigma}_S = \sqrt{\widehat{\sigma}_S^2}$ *with* $\widehat{\sigma}_S^2 = \widehat{a}_S^T \widehat{\Sigma}_S \widehat{a}_S$. *Here, using the notation* $\widehat{Q}_{S,x_n} = \frac{1}{n-1}\sum_{t=1}^{n-1} \mathbf{1}_{\{X_{t+1}\in S, X_t = x_n\}}$ *and* $\widehat{Q}_{x_n} = \frac{1}{n-1}\sum_{t=1}^{n-1}\mathbf{1}_{\{X_t=x_n\}}$, *we have*

$$\widehat{a}_S = \left(\frac{1}{\widehat{Q}_{x_n}}, \frac{-\widehat{Q}_{S,x_n}}{\widehat{Q}_{x_n}^2}\right)^\top \quad \textit{and} \quad \widehat{\Sigma}_S = \begin{pmatrix} \widehat{\Sigma}_S^{(1,1)} & \widehat{\Sigma}_S^{(1,2)} \\ \widehat{\Sigma}_S^{(2,1)} & \widehat{\Sigma}_S^{(2,2)} \end{pmatrix},$$

where $\widehat{\Sigma}_S$ *is a consistent estimator for* Σ_S *defined in* (A.2). *The estimator* $\widehat{\Sigma}_S$ *is obtained by replacing the corresponding probabilities in* Σ_S *by their sample estimators and suitably truncating the corresponding infinite sums.*

2.3. **Bootstrap Confidence Intervals.** Despite the fact whether a parametric or a non-parametric approach is used to estimate the predictive probability, when it comes to the construction of the corresponding confidence interval, some care is needed. To elaborate, recall that parametric assumptions that have been imposed on the underlying process in order to obtain estimators of the predictive probabilities, may not necessarily hold true in reality. Depending on which constellation holds true, this (may) affect the (limiting) distribution of the estimators used and therefore, it has to be taken into account in order to properly construct the confidence interval of interest. In Table 2, we show the different constellations that may occur when it comes to prediction. This table summarizes different situations depending on whether the time series $X_1, \ldots, X_n$ observed stems from a particular parametric class of Markov processes denoted by $\mathcal{M}_\theta$ or not and whether a parametric or a non-parametric approach is used to perform prediction, that is, to estimate the predictive probability P_{S,x_n}.

As it can be seen from this table, there are two constellations where the particular method used to calculate the predictive probability is correct and which are termed as "Correct" Prediction. The first refers to the situation where the time series stems from the parametric class

9

$\mathcal{M}_\theta$ and the same class also is used to calculate the predictive probability. The second concerns the case where the time series does not stem from the parametric class $\mathcal{M}_\theta$ and the prediction is implement in a non-parametric way. In the other two constellations, either an error is made in performing the prediction or the approach is inefficient. To elaborate, an error occurs in the case where the parametric model class $\mathcal{M}_\theta$ is wrongly used (model misspecification). Analogously, the predictive approach becomes inefficient when the parametric model class $\mathcal{M}_\theta$ is erroneously not used and the prediction is implemented in a non-parametric way. Now, in order to construct a proper confidence interval for the predictive probability of interest, these different constellations have to be taken into account.

The bootstrap procedures developed in this section for the construction of confidence intervals are able to fully take into account the different situations that may occur as described in Table 2. Additionally to this important fact, the bootstrap also has the advantage to circumvent the possibly cumbersome calculations needed to implement in practice the limiting distributions derived in Sections 2.2.1 and 2.2.2 in order to construct the confidence intervals of interest.

The main structure of the bootstrap procedure proposed consists of four steps, where the first two have to be specified differently according to which setup from the four possible ones shown in Table 2 should be imitated in the bootstrap world. We first state this basic bootstrap algorithm, while the different specifications of the first two steps that should be made will be discussed later on in more detail. Notice that in the following description, $\widehat{P}_{S,x_n}$, respectively, $\widehat{P}^*_{S,x_n}$, are used to denote any estimator, respectively, bootstrap estimator, of the predictive probability P_{S,x_n}, where the specific form these estimators take will be clarified in the discussion following the basic bootstrap algorithm.

Algorithm 2.5 (Basic bootstrap algorithm for the construction of confidence intervals of predictive probabilities)**.**

Step 1: Given $X_1, \ldots, X_n$, calculate the estimator $\widehat{P}_{S,x_n} := \widehat{P}_{S,x_n}(X_1, \ldots, X_n)$ of the predictive probability P_{S,x_n}.

Step 2: Generate a pseudo time series $X_1^*, \ldots, X_n^*$ and calculate the same estimator as in Step 1 but based on $X_1^*, \ldots, X_n^*$ to get $\widehat{P}^*_{S,x_n} := \widehat{P}_{S,x_n}(X_1^*, \ldots, X_n^*)$.

Step 3: Repeat Step 2 a large number of times, say B times, and calculate

$$L_n^{*,(b)} = \widehat{P}^{*,(b)}_{S,x_n} - \widehat{P}_{S,x_n},\ b = 1, \ldots, B,$$

where $\widehat{P}^{*,(b)}_{S,x_n}$ denotes the estimator $\widehat{P}^*_{S,x_n}$ obtained in the bth bootstrap repetition.

Step 4: Compute the $(1-\delta)$-confidence interval as

$$\left[\widehat{P}_{S,x_n} - q^*_{1-\delta/2}, \widehat{P}_{S,x_n} - q^*_{\delta/2}\right],$$

10

Prediction \ Data	$X_1, \ldots, X_n$ stems from the class $\mathcal{M}_\theta$	$X_1, \ldots, X_n$ doesn't stem from the class $\mathcal{M}_\theta$
Step 1: Non-parametric $\widehat{P}_{S,x_n}^{(npara)}$	*Step 2:* $X_1^*, \ldots, X_n^*$ is generated using the estimated, model-based one step transition probabilities and $\widehat{P}_{S,x_n}^* = \widehat{P}_{S,x_n}^{*,(npara)}$	*Step 2:* $X_1^*, \ldots, X_n^*$ is generated using the non-parametrically estimated one step transition probabilities and $\widehat{P}_{S,x_n}^* = \widehat{P}_{S,x_n}^{*,(npara)}$
Step 1: Parametric $P_{S,x_n}^{(para)}(\widehat{\theta})$	*Step 2:* $X_1^*, \ldots, X_n^*$ is generated using the estimated, model-based one step transition probabilities and $\widehat{P}_{S,x_n}^* = P_{S,x_n}^{(para)}(\widehat{\theta}^*)$	*Step 2:* $X_1^*, \ldots, X_n^*$ is generated using the non-parametrically estimated one step transition probabilities and $\widehat{P}_{S,x_n}^* = P_{S,x_n}^{(para)}(\widehat{\theta}^*)$

TABLE 3. Different possible specifications of Step 1 and Step 2 of the basic bootstrap algorithm depending on which one of the four possible scenarios in Table 2 should be imitated.

where q_α^* denotes the α-quantile of the empirical distribution of L_n^*, that is, $P^*(L_n^* \leq q_\alpha^*) = \alpha$.

Recall that we aim to imitate the distribution of the estimator $\widehat{P}_{S,x_n}$ of the predictive probability under the different possible scenarios. So far, we have considered two types of estimators: the non-parametric estimator, that is, $\widehat{P}_{S,x_n} = \widehat{P}_{S,x_n}^{(npara)}$ and the parametric estimator, that is, $\widehat{P}_{S,x_n} = P_{S,x_n}^{(para)}(\widehat{\theta})$. Now, depending on whether the time series $X_1, \ldots, X_n$ stems from the parametric model class $\mathcal{M}_\theta$ or not, the distribution of the estimators $\widehat{P}_{S,x_n}^{(npara)}$, respectively, $P_{S,x_n}^{(para)}(\widehat{\theta})$ may be different. This requires a proper specification of Step 1 and Step 2 of the basic bootstrap algorithm, which is clarified in Table 3.

Remark 2.6 (Unconditional confidence interval vs. conditional prediction interval)**.** *In contrast to bootstrap procedures that aim to construct (conditionally) valid prediction intervals in the continuous case, we do explicitly not require that the pseudo time series $X_1^*, \ldots, X_n^*$ (in all four cases) has the property that $X_n^* = x_n$ holds true. This is because in the discrete setting considered in this paper, we propose to reverse the approach and aim for the construction of a* confidence *interval for the predictive probability P_{S,x_n} instead of constructing a* prediction *interval for X_{n+1} (conditional on $X_n = x_n$).*

In Sections 2.3.1 and 2.3.2, we will discuss in more detail the implementation and the properties of the different specifications of the basic bootstrap algorithm described in Table 3. Notice that investigations of the asymptotic validity of the constructed bootstrap prediction intervals

essentially need to show that the prediction error $L_n := \widehat{P}_{S,x_n} - P_{S,x_n}$ and the bootstrap prediction error $L_n^* := \widehat{P}^*_{S,x_n} - \widehat{P}_{S,x_n}$ converge, in a proper way, to the same limiting distribution, where the latter is assumed to be continuous. Denoting by q_α the α-quantile of the distribution of L_n, this convergence would justify the use of the bootstrap for constructing the prediction intervals of interest due to the following approximation

$$\begin{aligned} 1-\delta &= P(q_{\delta/2} \leq L_n \leq q_{1-\delta/2}) \\ &= P(\widehat{P}_{S,x_n} - q_{1-\delta/2} \leq P_{S,x_n} \leq \widehat{P}_{S,x_n} - q_{\delta/2}) \\ &\approx P(\widehat{P}_{S,x_n} - q^*_{1-\delta/2} \leq P_{S,x_n} \leq \widehat{P}_{S,x_n} - q^*_{\delta/2}). \end{aligned}$$

Here, q^*_α denotes the α-quantile of the distribution of the bootstrap error L_n^*, which, by the assumed convergence in distribution and the continuity of the limiting distribution, satisfies $|q^*_{\delta/2} - q_{\delta/2}| \to 0$, in probability, as $n \to \infty$.

2.3.1. *Parametric Prediction.* In the case of parametric prediction, that is, in the case where the estimator $P^{(para)}_{S,x_n}(\widehat{\theta})$ is used, the following two different constellations may occur as shown in the second row of Table 3:

(a) $X_1, \ldots, X_n$ stems from the parametric class used to perform the prediction.

(b) $X_1, \ldots, X_n$ does not stem from the parametric class used to perform the prediction.

We first elaborate on case (a). Under this scenario, the matrix of transition probabilities used to generate $X_1^*, \ldots, X_n^*$, is obtained from the parametric model and depends on the estimated parameter vector $\widehat{\theta}$. Its (i,j)th element is given by $P^{(para)}_{X_{t+1}=x_j|X_t=x_i}(\widehat{\theta})$. The bootstrap pseudo time series $X_1^*, \ldots, X_n^*$ is then generated using these parametric one step transition probabilities.

In case (b), the bootstrap time series $X_1^*, \ldots, X_n^*$ is generated using non-parametric estimators of the one step transition probabilities. The (i,j)th element of the corresponding transition probability matrix is $\widehat{P}^{(npara)}_{X_{t+1}=x_j|X_t=x_i}$ and it is given in (2.2) for $S = \{x_j\}$ and $x_n = x_i$.

However, in any one of the two scenarios discussed, the bootstrap estimator $\widehat{\theta}^*$ based on the pseudo time series $X_1^*, \ldots, X_n^*$, has to imitate the distribution of the estimator $\widehat{\theta}$ as stated in the following assumption. This high-level assumption has to be validated in a particular setting of interest, as this will be discussed in more detail in Section 3.

Assumption 3 (Bootstrap CLT for parameter estimator)**.** *The bootstrap estimator $\widehat{\theta}^*$ based on the pseudo time series $X_1^*, \ldots, X_n^*$ generated under the two different constellations (a) or (b), satisfies*

$$\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) \xrightarrow{d} \mathcal{N}(0, V_{\theta_0}) \text{ in probability,}$$

where V_{θ_0} is given in Assumption 2(ii).

12

Under this assumption, we can establish validity of the bootstrap procedure used to construct confidence intervals for the predictive probability in case the parametric estimator $P_{S,x_n}^{(para)}(\widehat{\theta})$ is used.

Theorem 2.7 (Parametric bootstrap confidence interval for $P_{S,x_n}^{(para)}(\theta_0)$)**.** *Suppose Assumptions 1, 2 and 3 hold true. Let $P_{S,x_n}^{(para)}(\theta)$ be continuously differentiable around θ_0 with gradient $\nabla_\theta P_{S,x_n}^{(para)}(\theta_0) \neq 0$. Then, the bootstrap confidence interval for $P_{S,x_n}^{(para)}(\theta_0)$ given by*

$$\big[P_{S,x_n}^{(para)}(\widehat{\theta}) - q_{1-\delta/2}^*, \quad P_{S,x_n}^{(para)}(\widehat{\theta}) - q_{\delta/2}^*\big],$$

is asymptotically of level $1-\delta$. Here, q_α^* denotes the α-quantile of the distribution of $L_n^* := P_{S,x_n}^{(para)}(\widehat{\theta}^*) - P_{S,x_n}^{(para)}(\widehat{\theta})$.

As the above theorem shows, the bootstrap confidence interval for the parametric predictive probability $P_{S,x_n}^{(para)}(\theta_0)$ retains (asymptotically) the desired level $1-\delta$ even if the time series observed does not stem from the model class used to perform the prediction. As we will see in the simulations in Section 5, in this case of model misspecification, the confidence intervals obtained inherit this uncertainty, which manifests itself in larger but more honest confidence intervals compared to those obtained for the case where $X_1, \ldots, X_n$ stems from the model class $\mathcal{M}_\theta$ and this class also is used to implement the bootstrap.

2.3.2. *Non-parametric Prediction.* Consider next the case where the non-parametric estimator of the predictive probability $\widehat{P}_{S,x_n}^{(npara)}$ is used to perform the prediction; see the first row of Table 3. In this case, the time series $X_1, \ldots, X_n$ may or may not stem from the parametric family $\mathcal{M}_\theta$. According to which one of the two possible scenarios should be imitated in the bootstrap world, one can again use the estimated parametric one step transition probabilities, $P_{X_{t+1}=x_j|X_t=x_i}^{(para)}(\widehat{\theta})$, or the non-parametric analogue, $\widehat{P}_{X_{t+1}=x_j|X_t=x_i}^{(npara)}$, to generate the bootstrap pseudo time series $X_1^*, \ldots, X_n^*$. The following theorem can then be established.

Theorem 2.8 (Non-parametric bootstrap confidence interval for $P_{S,x_n} = P_{S,x_n}^{(npara)}$)**.**

(i) Suppose Assumptions 1 holds true and that $X_1^*, \ldots, X_n^*$, is generated using the non-parametrically estimated one step transition probabilities $\widehat{P}_{X_{t+1}=x_j|X_t=x_i}^{(npara)}$. Then, the bootstrap confidence interval for P_{S,x_n} given by

$$\big[\widehat{P}_{S,x_n}^{(npara)} - q_{1-\delta/2}^*, \quad \widehat{P}_{S,x_n}^{(npara)} - q_{\delta/2}^*\big],$$

is asymptotically of level $1-\delta$. Here, q_α^* denotes the α-quantile of the distribution of $L_n^* := \widehat{P}_{S,x_n}^{*,(npara)} - \widehat{P}_{S,x_n}^{(npara)}$.

(ii) Suppose Assumptions 1, 2 and 3 hold true and that $X_1^*, \ldots, X_n^*$ is generated using the estimated, model-based one step transition probabilities $P_{X_{t+1}=x_j|X_t=x_i}^{(para)}(\widehat{\theta})$. Then, the confidence interval constructed as in (i) is asymptotically of level $1-\delta$.

3. Parametric Prediction Using Specific Estimators and Models

We illustrate our proposed prediction methodology by an application to the case where the popular INAR(1) and INARCH(1) models for count time series are used for prediction. We also concentrate on (conditional) maximum likelihood estimators of the parameters of these models. We first elaborate on properties of conditional maximum likelihood estimators under a setting which allows for model misspecification.

3.1. **Conditional Maximum Likelihood Estimators.** Let

$$l_n(\theta|X_1) = \sum_{t=2}^{n} \log\big(P^{(para)}_{X_t|X_{t-1}}(\theta)\big)$$

be the (conditional on X_1) log-likelihood function, where $P^{(para)}_{X_t|X_{t-1}}(\theta)$ denotes the one step transition probabilities of the process $(X_t(\theta), t \in \mathbb{Z})$ belonging to the class $\mathcal{M}_\theta$. We denote by $\widehat{\theta}_{ML} = \text{argmax}_{\theta\in\Theta} l_n(\theta|X_1)$, the (conditional) maximum likelihood estimator of θ based on the time series $X_1, \ldots, X_n$ stemming from a process $(X_t, t \in \mathbb{Z})$ satisfying Assumption 1.

Since we are in the setting of a parametric prediction and we will use the estimator $P^{(para)}_{S,x_n}(\widehat{\theta}_{\text{ML}})$ of the predictive probability, in order to construct a confidence interval for $P^{(para)}_{S,x_n}(\theta)$, a parametric or a non-parametric bootstrap approach to generate the bootstrap time series $X^*_1, \ldots, X^*_n$ can be used; see the second row of Table 3. Recall that using a non-parametric approach in this context to generate the pseudo time series, enables us to imitate the distribution of $P^{(para)}_{S,x_n}(\widehat{\theta}_{ML})$ under model misspecification. To establish validity of the bootstrap in this setting, the key step is to show that the estimator $\widehat{\theta}^*_{ML}$ satisfies Assumption 3. Towards this goal, some additional assumptions regarding the properties of the underlying process have to be imposed. We begin with the following assumption.

Assumption 4 (Data generating process under potential model misspecification)**.**

(i) $\mathrm{E}\big(\log P^{(para)}_{X_{t+1}|X_t}(\theta)\big)$ *has a unique maximum at* θ_0 *in the interior of* Θ*.*

(ii) $P^{(para)}_{X_{t+1}|X_t}(\theta)$ *is two times differentiable around* θ_0 *with respect to* θ *with Lipschitz continuous second-order partial derivatives.*

(iii) $\mathrm{E}\|\nabla_\theta \log P^{(para)}_{X_{t+1}|X_t}(\theta)\|^2 < \infty$ *for every* $\theta \in \Theta$ *and the matrix* $\mathrm{E}\big(\nabla^2_\theta \log P^{(para)}_{X_{t+1}|X_t}(\theta_0)\big)$ *is positive definite.*

(iv) $\mathrm{E}\big(\nabla_\theta \log P^{(para)}_{X_{t+1}|X_t}(\theta)\big) = \nabla_\theta \mathrm{E}\big(\log P^{(para)}_{X_{t+1}|X_t}(\theta)\big)$ *for every* $\theta \in \Theta$*.*

Part (i) of this assumption guarantees the uniqueness of the limit of the maximum likelihood estimator which is important when $(X_t, t \in \mathbb{Z})$ does not necessarily belong to $\mathcal{M}_\theta$. Assumption 4(ii) imposes smoothness conditions on the parametric one step transition probabilities, Assumption 4(iii) ensures the existence of second-order moments while Assumption 4(iv) allows for the interchangeability of expectation and differentiation. Such assumptions are common for

14

investigating the consistency and distributional properties of maximum likelihood estimators; see for instance Condition 1.1 in Billingsley (1961).

Using Assumption 4(ii), a Taylor series expansion of $l_n'(\theta|X_1) = \sum_{t=2}^{n} \nabla_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta))$ around θ_0 and the fact that $l_n'(\widehat{\theta}_{ML}|X_1) = 0$, we get the basic expression

$$\frac{1}{n-1} l_n''(\widetilde{\theta}_n|X_1)\sqrt{n-1}\big(\widehat{\theta}_{ML} - \theta_0)\big) = -\frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} \nabla_\theta \log P^{(para)}_{X_t|X_{t-1}}(\theta)\big|_{\theta=\theta_0}, \tag{3.1}$$

for some $\widetilde{\theta}_n \in \Theta$ such that $\|\widetilde{\theta}_n - \theta_0\| \leq \|\widehat{\theta}_{ML} - \theta_0\|$. We first establish the following result which generalizes consistency of the (conditional) maximum likelihood estimator also for case where the time series at hand does not necessarily stem from the model class $\mathcal{M}_\theta$.

Theorem 3.1 (Convergence of $\widehat{\theta}_{ML}$)**.** *Assume that $X_1, \ldots, X_n$ stems from a process satisfying Assumption 1. Let $\mathcal{M}_\theta$ be a model class satisfying Assumption 2(i) and suppose that Assumption 4(i) is fulfilled. Then, as $n \to \infty$,*

$$\widehat{\theta}_{ML} \xrightarrow{P} \theta_0.$$

Note that analogous to (3.1), a similar expression can also be obtained in the bootstrap world based on the pseudo time series $X_1^*, \ldots, X_n^*$, that is, we have,

$$\frac{1}{n-1} l_n''(\widetilde{\theta}_n^*|X_1^*)\sqrt{n-1}\big(\widehat{\theta}_{ML}^* - \widehat{\theta}_{ML}\big) = -\frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} \nabla_\theta \log(P^{(para)}_{X_t^*|X_{t-1}^*}(\theta))\big|_{\theta=\widehat{\theta}_{ML}}, \tag{3.2}$$

where $\|\widetilde{\theta}_n^* - \widehat{\theta}_{ML}\| \leq \|\widehat{\theta}_{ML}^* - \widehat{\theta}_{ML}\|$.

Now, when a parametric model from the class $\mathcal{M}_\theta$ is fitted using (conditional) maximum likelihood estimation and this model is used in the bootstrap procedure to generate the pseudo time series $X_1^*, \ldots, X_n^*$, then the generated bootstrap data stems from a parametric model with parameter vector $\widehat{\theta}_{ML}$ irrespective of whether this also holds true or not for the observed time series $X_1, \ldots, X_n$. In this case, and as we will see in the following, asymptotic normality of the bootstrap sequence

$$\frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} \nabla_\theta \log(P^{(para)}_{X_t^*|X_{t-1}^*}(\theta))\Big|_{\theta=\widehat{\theta}_{ML}} \tag{3.3}$$

can typically be established using a central limit theorem for triangular arrays of martingale differences; see expression (A.9) in the Appendix. However, in the case where $X_1^*, \ldots, X_n^*$ is generated non-parametrically using the estimated one step transition probabilities $\widehat{P}^{(npara)}_{X_{t+1}=x_j|X_t=x_i}$, such a property for the corresponding bootstrap sequence does not necessarily hold true. In this case, the asymptotic normality of the bootstrap sequence (3.3) relies on mixing properties which are related to corresponding properties of the underlying process $(X_t, t \in \mathbb{Z})$. To

elaborate, consider the sequence

$$\frac{1}{\sqrt{n-1}}\sum_{t=2}^{n}\nabla_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta_0)\big) \\ = \frac{1}{\sqrt{n-1}}\sum_{t=2}^{n}\Big(\nabla_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta_0)\big) - \mathrm{E}\nabla_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta_0))\Big), \tag{3.4}$$

appearing in (3.1) and where $\mathrm{E}(\nabla_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta_0))) = 0$ by Assumption 4(iv). For instance, it is well known that countable Markov chains satisfying Assumption 1 are β-mixing with $\beta(k) \to 0$ as $k \to \infty$; see Bradley (2007), Theorem 7.7. The maximal correlation coefficient (ρ-mixing), to which we focus in the following, is an alternative way to control the dependence structure of a Markov process and to establish the weak convergence result

$$\frac{1}{\sqrt{n-1}}\sum_{t=2}^{n}\nabla_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta)\Big|_{\theta=\theta_0} \xrightarrow{d} \mathcal{N}\Big(0, \mathrm{E}\big(\nabla^2_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta_0))\big)\Big).$$

From this, Assumption 4, Theorem 3.1 and expression (3.1), we get that

$$\sqrt{n}(\widehat{\theta}_{ML} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_{\theta_0}), \tag{3.5}$$

where $V^{-1}_{\theta_0} = \mathrm{E}\big(\nabla^2_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta_0))\big)$. We can now establish the following important result for bootstrap consistency, which shows that the (conditional) maximum likelihood estimator fulfills the requirements of Assumption 3.

Theorem 3.2 (Asymptotic normality of $\widehat{\theta}_{ML}$)**.** *Suppose Assumptions 1 and 2(i) and 4 hold true.*

(i) Let the pseudo time series $X^*_1, \ldots, X^*_n$ be generated using the one step transition probabilities $P^{(para)}_{X_{t+1}=x_j|X_t=x_i}(\widehat{\theta}_{ML})$ for $x_i, x_j \in range(X_0)$. Then, as $n \to \infty$,

$$\sqrt{n}(\widehat{\theta}^*_{ML} - \widehat{\theta}_{ML}) \xrightarrow{d} \mathcal{N}(0, V_{\theta_0}), \tag{3.6}$$

in probability, where $V^{-1}_{\theta_0} = \mathrm{E}_{\theta_0}\big(\nabla^2_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big)$.

(ii) Let the pseudo time series $X^*_1, \ldots, X^*_n$ be generated using the non-parametrically estimated one step transition probabilities $\widehat{P}^{(npara)}_{X_{t+1}=x_j|X_t=x_i}$. Assume that $(X_t, t \in \mathbb{Z})$ is ρ-mixing with ρ-mixing coefficient $\rho_1 < 1$ and $E(|X_0|^{12+\delta}) < \infty$ for some $\delta > 0$ and

$$\max_{1\le t\le n} \nabla_\theta \log(P^{(para)}_{X^*_t|X^*_{t-1}}(\theta))\big|_{\theta=\widehat{\theta}_{ML}} = o_{P^*}(\sqrt{n}). \tag{3.7}$$

Then, as $n \to \infty$, assertion (3.6) holds true with $V^{-1}_{\theta_0} = \mathrm{E}\big(\nabla^2_\theta \log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big)$.

Remark 3.3. *Note that in assertion (i) of the above theorem, the expectation denoted by $\mathrm{E}_{\theta_0}(\cdot)$ in the definition of V_{θ_0} is taken with respect to the probability measure of the parametric model with parameter θ_0. Furthermore, in assertion (ii) of the same theorem, the expectations appearing in the definition of V_{θ_0} and denoted by $\mathrm{E}(\cdot)$ is taken with respect to the probability measure*

of the underlying Markov process, which does not necessarily belong to the parametric class $\mathcal{M}_\theta$. Clearly, $\mathrm{E}(\cdot) = \mathrm{E}_{\theta_0}(\cdot)$ if $(X_t, t \in \mathbb{Z})$ belongs to $\mathcal{M}_\theta$.

Remark 3.4. *According to Roberts and Rosenthal (1997), any stationary Markov chain that is geometrically ergodic and reversible is ρ-mixing with $\rho_1 < 1$, where*

$$\rho_k = \rho\big(\sigma(X_k), \sigma(X_0)\big), \quad \rho_k \leq \rho_1^k. \tag{3.8}$$

The maximal correlation coefficient $\rho(\cdot,\cdot)$ is defined by

$$\rho(\mathcal{A}, \mathcal{B}) = \sup_{f \in L_2(\mathcal{A}),\, g \in L_2(\mathcal{B})} |Corr(f, g)|,$$

where $L_2(\mathcal{A})$ is the space of all random variables that are $\mathcal{A}$-measurable and square integrable. By Theorem 4.4(b1) of Bradley (2007), for a process $(X_t, t \in \mathbb{Z})$, we have

$$\rho_1 = \sup_{f,g} \left\{ \frac{\big|E(f(X_i)g(X_{i-1})) - E(f(X_i))E(g(X_{i-1}))\big|}{\sqrt{E(f^2(X_i))}\sqrt{E(g^2(X_{i-1}))}}; \|f(X_i)\|_2 < \infty,\ \|g(X_{i-1})\|_2 < \infty \right\},$$

where we used the notation $\|X\|_p = (E(X^p))^{1/p}$ for $p > 1$.

3.2. **The INAR(1) Model.** The INAR(1) model was first introduced by McKenzie (1985) and Al-Osh and Alzaid (1987) and extended to order p by Du and Li (1991). Recall that according to the INAR(p) model, X_t is generated as

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{3.9}$$

where $\varepsilon_t \overset{\text{i.i.d.}}{\sim} G$ with probability mass function $G(k)$, $k \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and the vector of coefficients $\alpha = (\alpha_1, \ldots, \alpha_p)^\top \in (0,1)^p$ fulfills $\sum_{i=1}^p \alpha_i < 1$. Recall that in order to account for the integer nature of the data, model (3.9) uses the binomial thinning operator "$\circ$" first introduced by Steutel and Van Harn (1979) and defined as

$$\alpha_i \circ X_{t-i} = \sum_{j=1}^{X_{t-i}} Z_j^{(t,i)}, \tag{3.10}$$

with $\big(Z_j^{(t,i)}, j \in \mathbb{N}, t \in \mathbb{Z}\big), i \in \{1, \ldots, p\}$ being mutually independent random variables with $Z_j^{(t,i)} \sim \mathrm{Bin}(1, \alpha_i)$ and independent of the innovation process $(\varepsilon_t,\, t \in \mathbb{Z})$. Here, $\mathrm{Bin}(L, p)$ denotes the binomial distribution with parameters $L \in \mathbb{N}$ and $p \in [0,1]$. Note that according to this construction, $\alpha_i \circ X_{t-i} | X_{t-i} \sim \mathrm{Bin}(X_{t-i}, \alpha_i)$ holds.

In the case of an INAR(1) model, the one step transition probabilities are given by

$$P^{(para)}_{X_t = x_t | X_{t-1} = x_{t-1}}(\theta) = \sum_{j=0}^{\min(x_t, x_{t-1})} \binom{x_{t-1}}{j} \alpha_1^j (1-\alpha_1)^{x_{t-1}-j} G(x_t - j) \tag{3.11}$$

for $x_t, x_{t-1} \in \mathbb{N}_0$. Assume now that an INAR(1) model with specified innovation distribution G_γ depending on a parameter γ is used for prediction. Letting $\theta = (\alpha, \gamma)$, the predictive

probability of this model is denoted by $P^{(para)}_{S,x_n}(\theta) =: P^{(inar)}_{S,x_n}(\theta)$. By (3.11) and for the (conditional) maximum likelihood estimator $\widehat{\theta}_{\mathrm{ML}} = (\widehat{\alpha}_{\mathrm{ML}}, \widehat{\gamma}_{\mathrm{ML}})$ of θ, the predictive probability can be estimated by

$$P^{(inar)}_{S,x_n}(\widehat{\theta}_{\mathrm{ML}}) = \sum_{j\in S} \sum_{k=0}^{\min(x_n,j)} \binom{x_n}{k} \widehat{\alpha}^k_{\mathrm{ML}} (1-\widehat{\alpha}_{\mathrm{ML}})^{x_n-k} G_{\widehat{\gamma}_{\mathrm{ML}}}(j-k). \tag{3.12}$$

To obtain a confidence interval for $P^{(inar)}_{S,x_n}(\theta_0)$ one can use the limiting distribution of $\widehat{\theta}_{ML}$. To elaborate, it is well known that under appropriate conditions including the assumption that $X_1, \ldots, X_n$ stems from a INAR(1) model, it holds true that, as $n \to \infty$,

$$\sqrt{n}\,(\widehat{\theta}_{\mathrm{ML}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0)), \tag{3.13}$$

where $\theta_0 = (\alpha_0, \gamma_0)$ is the true parameter and $I(\theta_0) = \mathrm{E}(J_t(\theta_0))$, where $J_t(\theta_0)$ is the Hessian of $-\log P^{(inar)}_{X_t|X_{t-1}}(\theta))$ evaluated at θ_0; see for instance Billingsley (1961), Freeland (1998) and Weiß (2018). Note that $\mathrm{E}(J_t(\theta_0))$ can be estimated by $(n-1)^{-1}\sum_{t=2}^{n} J_t(\widehat{\theta}_{ML})$; also see Remark B.2.1.2 in Weiß (2018).

Remark 3.5. *The asymptotic result* (3.13) *holds true for a broad variety of INAR processes including the prominent examples of a Poisson INAR(1), (Poi-INAR(1)), and a negative binomial INAR(1), (NB-INAR(1)), process.*

From (3.13), we immediately get

$$\sqrt{n-1}\,(P^{(inar)}_{S,x_n}(\widehat{\theta}_{\mathrm{ML}}) - P^{(inar)}_{S,x_n}(\theta_0)) \xrightarrow{d} \mathcal{N}(0, s^2_{\theta_0}), \tag{3.14}$$

where $s^2_{\theta_0} = \nabla_\theta g(\theta_0)^\top I^{-1}(\theta_0) \nabla_\theta g(\theta_0)$ and $\nabla_\theta g(\theta_0) = \sum_{j\in S} \nabla_\theta P^{(inar)}_{X_{t+1=j}|X_t=x_n}(\theta)\big|_{\theta=\theta_0}$. This leads to the asymptotically valid $(1-\delta)$-confidence interval

$$\left[P^{(inar)}_{S,x_n}(\widehat{\theta}_{ML}) - \frac{z_{\delta/2}\widehat{s}_{\theta_0}}{\sqrt{n-1}},\ P^{(inar)}_{S,x_n}(\widehat{\theta}_{ML}) + \frac{z_{\delta/2}\widehat{s}_{\theta_0}}{\sqrt{n-1}}\right]$$

for $P^{(inar)}_{S,x_n}(\theta_0)$, where $\widehat{s}^2_{\theta_0} = \nabla_\theta g(\widehat{\theta}_{\mathrm{ML}})^\top I^{-1}(\widehat{\theta}_{\mathrm{ML}}) \nabla_\theta g(\widehat{\theta}_{\mathrm{ML}})$ and $z_{\delta/2}$ the $\delta/2$-quantile of the standard Gaussian distribution.

Remark 3.6. *Note that in the case of a Poi-INAR(1) process, we have for* $\theta = (\alpha, \lambda)$,

$$g(\alpha,\lambda) := P^{(inar)}_{S,x_n}(\theta) = \sum_{j\in S} \sum_{k=0}^{\min(x_n,j)} \binom{x_n}{k} \alpha^k (1-\alpha)^{x_n-k} \frac{\lambda^{j-k} e^{-\lambda}}{(j-k)!}.$$

In the case of a NB-INAR(1) process, we have for $\theta = (\alpha, N, \pi)$,

$$g(\alpha,N,\pi) := P^{(inar)}_{S,x_n}(\theta) = \sum_{j\in S} \sum_{k=0}^{\min(x_n,j)} \binom{x_n}{k} \alpha^k (1-\alpha)^{x_n-k} \frac{\Gamma(j-k+N)}{\Gamma(N)(j-k)!} \pi^N (1-\pi)^{j-k},$$

18

where $\Gamma(\cdot)$ *denotes the gamma function. For the Poi-INAR(1) case, we exemplarily get*

$$\nabla g(\alpha,\lambda)) = \begin{pmatrix} \sum\limits_{j\in S}\sum\limits_{k=0}^{\min(x_n,j)} \binom{x_n}{k} \frac{\lambda^{j-k}}{(j-k)!} e^{-\lambda}\alpha^{k-1}(k-x_n\alpha)(1-\alpha)^{x_n-k-1} \\ \sum\limits_{j\in S}\sum\limits_{k=0}^{\min(x_n,j)} \binom{x_n}{k}\alpha^k(1-\alpha)^{x_n-k}\frac{1}{(j-k)!}(-e^{-\lambda})\lambda^{j-k-1}(-j+k+\lambda). \end{pmatrix} \tag{3.15}$$

These expressions have also been used for the numerical calculations in Section 5.

However, if $X_1,\ldots,X_n$ does *not* stem from an INAR(1) model, but this model is fitted to the data at hand, then one can take advantage of the convergence (3.5) in order to construct a confidence interval for the predictive probability $P^{(inar)}_{S,x_n}(\theta_0)$ and proceed analogous to the case where (3.13) holds true. The main difference in this case is that $V_{\theta_0} \neq I^{-1}(\theta_0)$ and $\theta_0 = (\alpha_0,\gamma_0)$ is the unique parameter from the parameter space Θ which minimizes

$$\mathrm{E}\big(\log P^{(inar)}_{X_t|X_{t-1}}(\theta)\big) = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\log(P^{(inar)}_{X_t=r|X_{t-1}=s}(\theta))P(X_t=r,X_{t-1}=s);$$

see Assumption 4(i).

Instead of using the asymptotic Gaussian distributions (3.13), respectively, (3.5) and in order to avoid estimation of the quantities $I^{-1}(\theta_0)$, respectively, V_{θ_0}, one can use the bootstrap in both versions described in the second row of Table 3 in order to obtain a confidence interval for $P^{(inar)}_{S,x_n}(\theta_0)$. In view of the results obtained in Section 3.1 and, in particular, Theorem 3.2, the following bootstrap confidence intervals can be constructed depending on whether one wants to imitate the situation of a time series stemming from the fitted model class or not.

Suppose that $X_1,\ldots,X_n$ does not necessarily stem from an INAR(1) model, but that solely Assumption 1 holds true. Let the bootstrap pseudo time series $X_1^*,\ldots,X_n^*$ be generated using the one step transition probabilities of the estimated INAR(1) model given by

$$P^{(inar)}_{X_t=x_t|X_{t-1}=x_{t-1}}(\widehat{\theta}_{ML}) = \sum_{j=0}^{\min(x_t,x_{t-1})}\binom{x_{t-1}}{j}\widehat{\alpha}^j_{ML}(1-\widehat{\alpha}_{ML})^{x_{t-1}-j}f_{\varepsilon,\widehat{\gamma}_{ML}}(x_t-j). \tag{3.16}$$

Then, the bootstrap confidence interval,

$$\big[P^{(inar)}_{S,x_n}(\widehat{\theta}_{ML}) - q^*_{1-\delta/2},\ \ P^{(inar)}_{S,x_n}(\widehat{\theta}_{ML}) - q^*_{\delta/2}\big], \tag{3.17}$$

is asymptotically of level $1-\delta$ for $P_{S,x_n}(\theta_0)$, where q^*_α denotes the α-quantile of the distribution of $L^*_n = P^{*,(inar)}_{S,x_n}(\widehat{\theta}^*_{ML}) - P^{(inar)}_{S,x_n}(\widehat{\theta}_{ML})$.

Alternatively and if the conditions stated in Theorem 3.2(ii) hold true and the bootstrap pseudo time series $X_1^*,\ldots,X_n^*$ is generated using the non-parametrically estimated one step transition probabilities $\widehat{P}^{(npara)}_{X_{t+1}=x_j|X_t=x_i}$, then, using the corresponding bootstrap distribution, a confidence interval analogue to (3.17) can be constructed.

3.3. **The INARCH(1) Model.** The INARCH model belongs to the class of more general INGARCH models, which have been introduced by Heinen (2003) (under a different name) and further analyzed by, e.g., Ferland et al. (2006) and Fokianos et al. (2009). The INARCH(p) model is described as

$$X_t|X_{t-1}, X_{t-2}, \ldots \sim Poi(\beta + \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p}), \tag{3.18}$$

where $\beta > 0$, $\alpha_1, \ldots, \alpha_p \geq 0$ and $\sum_{i=1}^{p} \alpha_i < 1$. The one step transition probabilities of model (3.18) are given by

$$P^{(para)}_{X_t=x_t|X_{t-1}=x_{t-1},\ldots,X_{t-p}=x_{t-p}} = \exp\left(-\beta - \sum_{i=1}^{p} \alpha_i x_{t-i}\right) \frac{\left(\beta + \sum\limits_{i=1}^{p} \alpha_i x_{t-i}\right)^{x_t}}{x_t!}. \tag{3.19}$$

For $p = 1$, we get the special case of an INARCH(1) model, which has been discussed in Weiß (2010) and the corresponding transition probabilities reduce to

$$P^{(para)}_{X_t=x_t|X_{t-1}=x_{t-1}} = \exp(-\beta - \alpha x_{t-1}) \frac{(\beta + \alpha x_{t-1})^{x_t}}{x_t!}. \tag{3.20}$$

Assuming that an INARCH(1) model with parameter $\theta = (\beta, \alpha)$ is used for prediction and using the maximum likelihood estimator $\widehat{\theta}_{\mathrm{ML}} = (\widehat{\beta}_{\mathrm{ML}}, \widehat{\alpha}_{\mathrm{ML}})$, the estimated predictive probability $P^{(para)}_{S,x_n}(\theta_0) =: P^{(inarch)}_{S,x_n}(\theta_0)$ is given by

$$P^{(inarch)}_{S,x_n}(\widehat{\theta}_{\mathrm{ML}}) = \sum_{j \in S} \exp(-\widehat{\beta}_{\mathrm{ML}} - \widehat{\alpha}_{\mathrm{ML}} x_n) \frac{(\widehat{\beta}_{\mathrm{ML}} + \widehat{\alpha}_{\mathrm{ML}} x_n)^j}{j!}. \tag{3.21}$$

To construct an asymptotic or a bootstrap confidence interval for $P^{(inarch)}_{S,x_n}(\theta_0)$, one can proceed as discussed in Section 3.2 for the case of an INAR(1) model.

Notice that under the assumption that $(X_t, t \in \mathbb{Z})$ is an INARCH(1) model, Zhu and Wang (2009) proved that $\sqrt{n-1}(\widehat{\theta}_{\mathrm{ML}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$, where $I(\theta_0)$ can be estimated by $(n-1)^{-1} \sum\limits_{t=2}^{n} J_t(\theta_0)$ and

$$J_t(\theta_0) = \begin{pmatrix} \frac{X_t}{(\beta+\alpha X_{t-1})^2} & \frac{X_t X_{t-1}}{(\beta+\alpha X_{t-1})^2} \\ \frac{X_t X_{t-1}}{(\beta+\alpha X_{t-1})^2} & \frac{X_t X_{t-1}^2}{(\beta+\alpha X_{t-1})^2} \end{pmatrix}.$$

From this we get

$$\sqrt{n-1}(P^{(inarch)}_{S,x_n}(\widehat{\theta}_{\mathrm{ML}}) - P^{(inarch)}_{S,x_n}(\theta_0)) \xrightarrow{d} \mathcal{N}(0, s^2), \tag{3.22}$$

where $s^2 = \nabla g(\theta_0)^\top I^{-1}(\theta_0) \nabla g(\theta_0)$ with

$$g(\theta_0) = \sum_{j \in S} \exp(-\beta - \alpha x_n) \frac{(\beta + \alpha x_n)^{x_n}}{j!}$$

20

and the gradient given by

$$\nabla g(\beta,\alpha)=\begin{pmatrix}\sum_{j\in S}\frac{1}{j!}\exp(-\beta-\alpha x_n)(\beta+\alpha x_n)^{j-1}(j-\beta-\alpha x_n)\\ \sum_{j\in S}\frac{1}{j!}\exp(-\beta-\alpha x_n)x_n(\beta+\alpha x_n)^{j-1}(j-\beta-\alpha x_n)\end{pmatrix}. \tag{3.23}$$

4. Practical Issues and Extensions

4.1. **Range Preserving Confidence Intervals.** It may happen that the confidence intervals obtained exceed the natural range of a probability measure, i.e., the lower bound of the confidence interval is less than zero or the upper bound exceeds one. This situation is more likely to occur when predictive sets S are considered for which $P(X_{n+1}\in S|X_n=x_n)$ is very close to the boundaries of $[0,1]$. To enforce the intervals to preserve the proper range, one way is to use the percentile method. This works by arranging the B bootstrap replicates $\widehat{P}^{*,(b)}_{S,x_n}$, $b=1,2,\ldots,B$, in increasing order, $\widehat{P}^{*,(1)}_{S,x_n},\ldots,\widehat{P}^{*,(B)}_{S,x_n,}$, and use $\widehat{P}^{*,(B\delta/2)}_{S,x_n}$ as the lower bound and $\widehat{P}^{*,(B(1-\delta/2))}_{S,x_n}$ as the upper bound of the confidence interval. In the case that $B\delta/2$ is not an integer, we define $m=\lfloor(B+1)\delta/2\rfloor$ and use $\widehat{P}^{*,(m)}_{S,x_n}$ as lower and $\widehat{P}^{*,(B+1-m)}_{S,x_n}$ as upper bound of the confidence interval.

4.2. **Larger Prediction Horizon and Higher Order Processes.** Consider the case where prediction is made for a horizon $h>1$. In this case, one can think of the prediction problem as that of obtaining a point or an interval estimator of the predictive probability

$$P\big(X_{n+j}\in S_j,\ j=1,2,\ldots,h|X_n=x_n\big), \tag{4.1}$$

where $S_j\subset\mathbb{R}$, $j=1,\ldots,h$, are user-selected sets of values that the process can take. (4.1) refers to the probability that for the future time points $t=n+1,\ldots,n+h$, the Markov process will visit consecutively the sets $S_1,\ldots,S_h$. Clearly, $S_j=S$ for all $j=1,\ldots,h$ as well as $P(X_{n+h}\in S|X_n=x_n)$, that is, $S_1=S_2=\cdots=S_{h-1}=range(X_0)$, are special cases of (4.1).

Due to the Markov property, the predictive probability (4.1) can be expressed as a sum of proper one step transition probabilities, that is,

$$\begin{aligned}&P\big(X_{n+j}\in S_j,\ j=1,2,\ldots,h|X_n=x_n\big)\\ &\quad=\sum_{y_h\in S_h}\sum_{y_{h-1}\in S_{h-1}}\cdots\sum_{y_1\in S_1}P_{X_{n+h}=y_h|X_{n+h-1}=y_{h-1}}P_{X_{n+h-1}=y_{h-1}|X_{n+h-2}=y_{h-2}}\cdots P_{X_{n+1}=y_1|X_n=x_n}.\end{aligned}$$

This implies that estimating parametrically or non-parametrically the one step transition probabilities $P_{X_{t+1}=y_k|X_t=y_{k-1}}$ for $k=1,\ldots,h$ and $y_0\equiv x_n$, which are involved in the above expression, a parametric or non-parametric estimator of the predictive probability $P\big(X_{n+j}\in S_j, j=1,2,\ldots,h|X_n=x_n\big)$ can be obtained. Furthermore, an asymptotic or bootstrap confidence interval for the predictive probability of interest can in principle be constructed following the approaches discussed in the previous sections.

When considering Markov processes of higher order $p > 1$, the goal is to construct a confidence interval for

$$P_{S,(x_n,\ldots,x_{n-p+1})} := P(X_{n+1} \in S | X_{n-p+1} = x_{n-p+1}, \ldots, X_n = x_n).$$

For this, the main construction principle remains the same for both the asymptotic and the bootstrap approaches discussed. Note that a non-parametric estimator of $P_{S,(x_n,\ldots,x_{n-p+1})}$ is given by

$$\widehat{P}^{(\text{npara})}_{S,(x_n,\ldots,x_{n-p+1})} = \frac{\sum_{t=1}^{n-p} \mathbf{1}_{\{X_{t+1}\in S, X_{t-p+1}=x_{n-p+1},\ldots,X_t=x_n\}}}{\sum_{t=1}^{n-p} \mathbf{1}_{\{X_{t-p+1}=x_{n-p+1},\ldots,X_t=x_n\}}}.$$

Analogously, a parametric estimator of the form $P^{(\text{para})}_{S,(x_n,\ldots,x_{n-p+1})}(\widehat{\theta})$ for $P^{(\text{para})}_{S,(x_n,\ldots,x_{n-p+1})}(\theta_0)$ can be for instance obtained by using a pth order INAR or INARCH model; see Section 3.2 and Section 3.3.

5. Simulations

We investigate the small sample performance of the asymptotic and the bootstrap methods to construct confidence intervals for the predictive probability. For the bootstrap, we distinguish between the construction principle of Algorithm 2.5 and its percentile versions as discussed in Section 4.1. We cover non-parametric and parametric setups focusing on predictions using INAR(1) and INARCH(1) data generating processes (DGPs), as elaborated in Section 3. To estimate the asymptotic variances involved in the non-parametric setup, we use the R package *stableGR* (Knudson and Vats, 2022). For the asymptotic approach in the parametric setup, we proceed as follows: In case an INAR(1) model is assumed, we use the asymptotic approach described in Section 3.2 by applying (3.15) and the approximated Hessian of the R function *constrOptim* (R Core Team, 2022). If an INARCH(1) model is assumed, we use the procedure described in Section 3.3.

We use $K = 500$ Monte Carlo samples and $B = 500$ bootstrap repetitions for four different sample sizes $n \in \{100, 500, 1000, 5000\}$. The significance level δ is set equal to 5%. We consider time series that stem from a Poi(1)-INAR(1) and from a NB(2,2/3)-INAR(1) model, both with parameter $\alpha = 0.5$ and from an INARCH(1) model with parameters $\alpha = 0.5$ and $\beta = 1$. If not stated otherwise, we consider $S = \{1, 2\}$. For simulation, estimation and bootstrapping of the INAR data, we use the R package *spINAR* (Faymonville et al., 2024). For the estimation of the INARCH data, we use the R package *tscount* (Liboschik et al., 2017).

5.1. **Asymptotic Confidence Intervals.** We apply the asymptotic procedures introduced in Section 2.2.1 and Section 2.2.2 to different data generating processes (DGPs) and construct

22

confidence intervals for P_{S,x_n}. We would like to emphasize that also in the parametric procedure introduced in Section 2.2.1, we construct a confidence interval for P_{S,x_n} only coinciding with $P_{S,x_n}(\theta_0)$ in case the assumed model for prediction is correct. To construct a confidence interval for $P_{S,x_n}(\theta_0)$, we would require an estimator for V_{θ_0}, see Proposition 2.2, which is not straightforward in general. To avoid the need of such an estimator, we explicitly refer to the bootstrap procedures in 2.3. The results obtained are presented in Table 4. First, consider the case where the DGP is a Poi(1)-INAR(1) process. Comparing coverages and mean lengths of the asymptotic confidence intervals, we see that the parametric approach using a Poi-INAR model for prediction performs best. The non-parametric approach keeps up well with respect to coverage, but it leads to much wider intervals. Furthermore and as expected, the parametric approach incorrectly using an INARCH model for prediction behaves worse and delivers confidence intervals the coverage of which decreases as the sample size n increases. Consider next the case where the data stems from an INAR(1) process with innovations having an overdispersed NB(2,2/3) distribution. In this case, both parametric models (Poi-INAR and INARCH) used for prediction are wrong, and this fact manifests itself in low coverages which decrease as n increases. The non-parametric approach performs best and shows a good coverage. The last considered case is that of time series stemming from an INARCH(1) process. Here again, using the wrong model for prediction leads to coverages which deteriorate as the sample size increases. At the same time and as expected, the parametric approach using the correct model performs best. The non-parametric approach lead to good results as this was the case in both previously considered DGPs.

To summarize, we see that concerning the calculation of a confidence interval for P_{S,x_n}, there is a trade-off between robustness and efficiency. The non-parametric approach is robust to model misspecification. Using a parametric approach is more efficient when the model assumptions made hold true, but it can lead to worse coverages if this is not the case because then $P_{S,x_n}(\widehat{\theta})$ is not a consistent estimator of P_{S,x_n} .

5.2. **Bootstrap Confidence Intervals.** We next investigate the finite sample performance of the bootstrap for constructing confidence intervals. We consider three of the four possible scenarios described in Table 2, respectively, Table 3. In the parametric case of 2.3.1, we construct confidence intervals for the parametric predictive probability $P^{(para)}_{S,x_n}(\theta_0)$. Recall that only if the parametric model used is correct, the predictive probabilities P_{S,x_n} and $P^{(para)}_{S,x_n}(\theta_0)$ coincide. In the non-parametric case of 2.3.2, we construct confidence intervals for the non-parametric predictive probability coinciding with P_{S,x_n}.

First of all, we investigate the parametric prediction setup and, in particular, the two cases listed in the second row of Table 2 and Table 3. For this purpose, a Poi-INAR(1) model is

23

Data	Prediction	n	100	500	1000	5000
Poi-INAR	Poi-INAR	cov	0.934	0.942	0.952	0.958
		ml	0.133	0.059	0.044	0.019
	INARCH	cov	0.440	0.352	0.354	0.122
		ml	0.101	0.043	0.034	0.014
	npara	cov	0.915	0.944	0.934	0.964
		ml	0.423	0.199	0.151	0.065
NB-INAR	Poi-INAR	cov	0.716	0.550	0.484	0.294
		ml	0.123	0.054	0.039	0.017
	INARCH	cov	0.548	0.190	0.140	0.086
		ml	0.105	0.045	0.033	0.014
	npara	cov	0.890	0.944	0.926	0.944
		ml	0.441	0.221	0.152	0.070
INARCH	Poi-INAR	cov	0.732	0.302	0.174	0.084
		ml	0.115	0.050	0.036	0.017
	INARCH	cov	0.932	0.918	0.918	0.944
		ml	0.101	0.042	0.031	0.014
	npara	cov	0.913	0.950	0.926	0.968
		ml	0.452	0.211	0.155	0.070

TABLE 4. Coverage (cov) and mean length (ml) of asymptotic confidence intervals for P_{S,x_n} for different sample sizes, different parametric and non-parametric approaches and significance level 0.05 when applied to different DGPs.

used for prediction and the following two scenarios for generating the bootstrap time series are considered: The bootstrap time series is generated non-parametrically, while the estimated Poi-INAR(1) model is used for prediction and the bootstrap time series is generated parametrically using the estimated Poi-INAR(1) model and this model also is used for prediction. We generate time series from the three different DGPs, that is, we use an INAR(1) process with $\alpha = 0.5$ and Poi(1) innovations, an INAR(1) process with $\alpha = 0.5$ and NB(2,2/3) innovations and an INARCH(1) process with $\alpha = 0.5$ and $\beta = 1$. Notice that all three processes possess an expected value of 2. To accurately approximate the unknown probability $P^{(para)}_{S,x_n}(\theta_0)$, we use a sample of $n = 10^6$ observations.

24

The results obtained are presented in Table 5. Observe that if the DGP does not coincide with the Poi-INAR(1) used for prediction, the bootstrap delivers good coverages which are approaching the nominal 95% for increasing n in the case where a non-parametric bootstrap procedure is used to generate the pseudo time series. However, when the bootstrap time series is generated parametrically using the (wrong) Poi-INAR model, the coverage is systematically lower. Also in terms of the mean length, there is a visible pattern if the model is misspecified. As the results show, the confidence intervals obtained in this case by applying a non-parametric bootstrap are always slightly larger than those obtained when a parametric bootstrap is used. This reflects the uncertainty associated with the fact that a wrong model is used to perform prediction. If the model is correctly specified, i.e., when we are in the case of a Poi-INAR DGP and the corresponding estimated model is used for prediction, both bootstrap procedures work well with respect to both coverage and length. It seems that in this case, essentially no price is paid for generating the bootstrap time series non-parametrically, that is, without using the (true) model. In summary, the proposed procedure in the below right corner of Table 3 is able to construct confidence intervals for $P_{S,x_n}^{(para)}(\theta_0)$ that retain the desired coverage.

We next consider the case of non-parametric prediction combined with non-parametric bootstrapping, i.e., we are referring to the first row and second column of Table 3. In addition, we also compare the results with the corresponding percentile version described in Section 4.1. Table 6 summarizes the results and shows that for increasing n and for all DGPs considered, the coverage is good and improves by getting closer to the desired 95% level. While the mean length of the intervals is large for small sample sizes, it decrease fast as n increases. Similar results are obtained using the percentile version.

5.3. **Range-preserving Intervals.** Table 6 displays the results when using the percentile version of the bootstrap confidence intervals. As pointed out in Section 4.1, the non-compliance with the natural range of probability can occur in situations where the considered set S is very (un)likely. To investigate such a scenario, we reconsider the Poi-INAR(1) DGP with $\alpha = 0.5$ and $\lambda = 1$, but now we define $S = \{10\}$. In the case of a Poi-INAR(1) DGP, the Poisson distribution of the innovations transfers to the observations leading to a Poi$(\lambda/(1-\alpha))$ marginal distribution, i.e., the observations are Poi(2) distributed. For $Y \sim$ Poi(2), we have $P(Y = 10) \approx 3.8 \cdot 10^{-5}$ illustrating the "unlikeliness" of S. Table 7 compares the results of the parametric asymptotic approach, the parametric bootstrap approach, where we use the parametric assumption for both estimating and bootstrapping, and its percentile version, where all three correctly assume a Poi-INAR(1) DGP. We see that the percentile intervals provide an increased and, consequently, more accurate coverage especially for small sample size n. This could be expected since we avoid the

25

Data	Prediction	Bootstrap	n	100	500	1000	5000
NB-INAR	Poi-INAR	para	cov	0.872	0.922	0.890	0.914
			ml	0.128	0.057	0.041	0.018
		npara	cov	0.892	0.942	0.930	0.932
			ml	0.139	0.062	0.044	0.019
INARCH	Poi-INAR	para	cov	0.892	0.902	0.922	0.924
			ml	0.122	0.053	0.039	0.018
		npara	cov	0.898	0.914	0.940	0.944
			ml	0.129	0.056	0.041	0.019
Poi-INAR	Poi-INAR	para	cov	0.900	0.930	0.946	0.956
			ml	0.130	0.058	0.043	0.019
		npara	cov	0.906	0.930	0.940	0.952
			ml	0.132	0.058	0.043	0.019

TABLE 5. Coverage (cov) and mean length (ml) of bootstrap confidence intervals for the predictive probability $P_{S,x_n}^{(para)}(\theta_0)$ when a Poi-INAR(1) model is used for prediction, i.e., for $P_{S,x_n}^{(inar)}(\lambda)$, where $\lambda = 1$ and λ the parameter of the Poisson distribution. A parametric (para) and a non-parametric (npara) bootstrap procedure is used and applied to three different DGPs and four different sample sizes.

violation of the natural range of probabilities. Asymptotically, the approaches do not differ much.

6. Real-Data Application

6.1. **Rotary Drilling Data.** In a first illustration of our proposed predictive inference procedure, we consider a time series of length $n = 417$ consisting of weekly counts of active rotary drilling rigs in Alaska (1990-1997)[4]. They are used to measure the demand of products from the drilling industry and have previously been analyzed in Weiß (2018) and Faymonville et al. (2025). In Figure 1, we see that the time series contains a lot of runs, i.e., same values for consecutive time points, which indicates a strong serial dependence together with a small innovations' mean. The latter is supported by the high and slowly decreasing autocorrelation level observed in the ACF plot. The PACF plot suggests a first-order autoregressive structure with

[4]phx.corporate-ir.net/phoenix.zhtml?c=79687&p=irol-rigcountsoverview

26

								percentile version			
Data	Prediction	Bootstrap	n	100	500	1000	5000	100	500	1000	5000
NB-INAR	npara	npara	cov	0.894	0.944	0.922	0.938	0.922	0.958	0.934	0.950
			ml	0.521	0.221	0.154	0.071	0.526	0.222	0.156	0.072
INARCH	npara	npara	cov	0.906	0.950	0.926	0.958	0.930	0.960	0.936	0.968
			ml	0.516	0.215	0.160	0.070	0.532	0.222	0.164	0.072
Poi-INAR	npara	npara	cov	0.922	0.940	0.922	0.954	0.938	0.956	0.936	0.958
			ml	0.489	0.210	0.155	0.065	0.493	0.212	0.157	0.065

TABLE 6. Coverage (cov) and mean length (ml) of bootstrap confidence intervals for P_{S,x_n} for different sample sizes and three different true DGPs. The confidence intervals are obtained by applying Algorithm 2.5 with the steps specified as in the upper right corner of Table 3.

						percentile version			
	n	100	500	1000	5000	100	500	1000	5000
Poi-INAR	cov	0.854	0.930	0.928	0.946	-	-	-	-
asy	ml ($\times 10^{-5}$)	8.5	3.7	3.8	1.1	-	-	-	-
Poi-INAR	cov	0.750	0.840	0.884	0.936	0.962	0.954	0.946	0.948
bs	ml ($\times 10^{-5}$)	9.6	3.8	3.8	1.1	9.9	4.0	3.9	1.1

TABLE 7. Coverage and mean length of the confidence intervals resulting from the three parametric approaches for different sample sizes and the different approaches applied on Poi(1)-INAR(1) DGP with $\alpha = 0.5$ in case of $S = \{10\}$.

$\widehat{\rho}(1) \approx 0.91$. An INAR(1) model appears to be reasonable. Indeed, testing the semi-parametric null hypothesis of an INAR(1) model, Faymonville et al. (2025) were not able to reject this hypothesis at 5% level. The data exhibits low counts and a dispersion index of approximately 1.11 suggesting approximate equidispersed counts.

Given that $x_n = 0$, a company in the drilling industry might be interested in the event that $x_{n+1} = 0$, i.e., that also in the next week there are no active rotary drilling rigs. As a consequence, this will result in a low demand of products of this company which may lead to a reduction of production activity. Apart from the set $S^{(1)} = \{0\}$, another set of interest could be the set $S^{(2)} = \{x : x \geq 2\}$. In this case, the company may want to hedge against suddenly more active rigs in the next week, which is associated with a higher need of replacement parts. In the following, we aim to estimate the predictive probability of these two sets and to construct

27

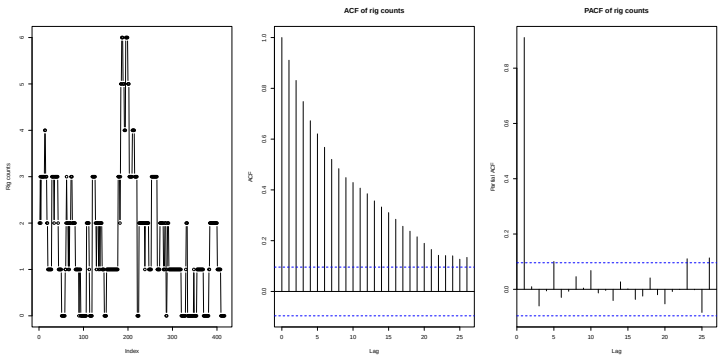

FIGURE 1. Plot of time series of weekly active rotary drillings and its corresponding (P)ACF.

a confidence interval for $P_{S^{(i)},x_n}$, $i = 1, 2$, at 95% level. Since $x_n = 0$ and taking into account the previous findings from the ACF and PACF plots, we may think of the set $S^{(2)}$ as an unlikely event. We apply the parametric and non-parametric bootstrap method to construct confidence intervals as described in Section 2.3.1 and Section 2.3.2. We also additionally calculate the percentile version of these confidence intervals; see Section 4.1. In the parametric case, we assume a Poisson distribution for the innovations which seems to be reasonable due to the equidispersed counts. The parametric assumption of an INAR(1) is both used for the point estimation and for the generation of the bootstrap time series. The same holds true also in the case where a non-parametric approach is used. Table 8 shows the results of the predictive probability of the two sets considered. As we see, the results for the parametric and the non-parametric procedure are close to each other and the confidence intervals obtained for the non-parametric prediction are, as expected, wider than those obtained in the parametric case. Furthermore, the percentile confidence intervals seem to be shorter and in the case of the set S_2, the benefits of these intervals become clear. While the lower bounds of the two bootstrap confidence intervals fall below 0, the percentile approach respects the natural range of the underlying parameter.

6.2. **Transaction Data.** In a second application, we consider a data set provided by the Deutsche Börse Group which has been discussed in Homburg et al. (2021) and Faymonville et al. (2025). For the period from February 2017 to August 2019, this data set contains $n = 404$ counts of transactions of structured products per trading day. Figure 2 shows a plot of the data along with the corresponding (P)ACF. Applying a goodness-of-fit test, Faymonville et al. (2025) were not able to reject the semi-parametric null hypothesis of an INAR(1) model at 5% level. Furthermore, they argued that due to overdispersion, a Poi-INAR(1) model would probably not be suitable for this data set. Instead, a Geo-INAR(1) model would be more

28

		Predictive Probabilities	Bootstrap CI	Percentile Version
$S^{(1)}$	para	0.8822	[0.8402, 0.9631]	[0.8485, 0.9143]
	npara	0.8764	[0.8309, 0.9681]	[0.7833, 0.9231]
$S^{(2)}$	para	0.0072	[-0.0067, 0.0115]	[0.0038, 0.0121]
	npara	0.0337	[-0.0182, 0.0674]	[0.0000, 0.0909]

TABLE 8. Estimated predictive probabilities and (percentile) bootstrap-based confidence intervals for the sets $S^{(1)} = \{0\}$ and $S^{(2)} = \{x : x \geq 2\}$ of the Rotary Drilling Data.

appropriate since the semi-parametrically estimated probability function of the innovations seems to be rather close to a geometric distribution. These findings motivate the application of a non-parametric and parametric prediction with bootstrap-based confidence intervals, where for obtaining these intervals, non-parametric and parametric generation of the bootstrap time series is applied. We aim to construct a 95% confidence interval for P_{S,x_n} and for $P_{S,x_n}(\theta_0)$ by considering two different sets of interests, $S^{(1)} = \{0\}$ and $S^{(2)} = \{2\}$. While the first one is associated with a lack of interest in the respective financial product, the second set speaks for a slight return to activity when $x_n = 0$. Both sets of interests concern questions related to risk management and investment strategies. In case of the parametric approaches, we separately consider the case of Poisson and geometrically distributed innovations. The results obtained are displayed in Table 9. Let us first discus these results from a goodness-of-fit point of view. By comparing the results of the parametric Po-INAR approach with those of the Geo-INAR and the non-parametric approaches, we see that the corresponding confidence intervals are almost disjoint. Additionally, when a geometric assumption for the innovations is used, the results seem more plausible and are very close to those obtained using a non-parametric method. In particular, the confidence intervals obtained using the Geo-INAR(1) model, are contained in the confidence intervals constructed using the non-parametric approach. This is in line with the findings in Faymonville et al. (2025), which argued in favor of a geometric distribution for the innovations. Let us spotlight the issue of model misspecification in the context of this data example. Toward this, consider the rows 1-4 and 5-9 of Table 9. We see that in the Geo-INAR case, the confidence intervals based on the parametric and the non-parametric bootstrap are very similar. This suggests that the uncertainty about the correctness of the parametric model associated with performing the prediction by using a Geo-INAR(1) model is almost negligible. In other words, the Geo-INAR(1) model describes this data set well, which seems not to be the case for the Poi-INAR model. The confidence intervals obtained using the latter model differ more to those obtained by the non-parametric method.

29

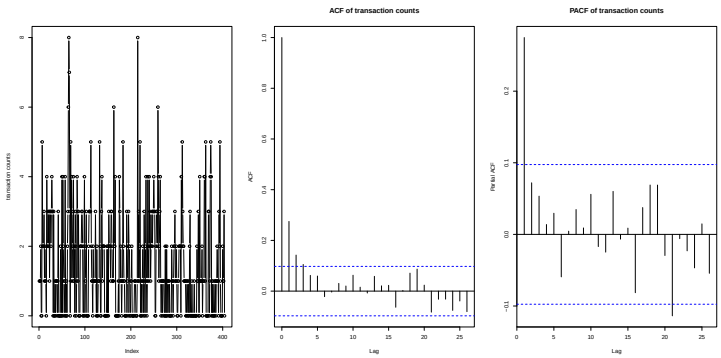

FIGURE 2. Plot of time series of transaction counts and its corresponding (P)ACF.

Set	Prediction	Bootstrap	Predictive Probability	Bootstrap CI
$S^{(1)}$	Poi-INAR	para	0.3203	[0.2720, 0.3676]
		npara		[0.2598, 0.3822]
	Geo-INAR	para	0.4929	[0.4469, 0.5290]
		npara		[0.4549, 0.5245]
	npara	npara	0.4884	[0.4023, 0.5752]
$S^{(2)}$	Poi-INAR	para	0.2076	[0.1841, 0.2343]
		npara		[0.1822, 0.2372]
	Geo-INAR	para	0.1267	[0.1189, 0.1365]
		npara		[0.1185, 0.1367]
	npara	npara	0.1318	[0.0676, 0.1824]

TABLE 9. Estimated predictive probabilities and bootstrap confidence intervals using different parametric and non-parametric approaches for the sets $S^{(1)} = \{0\}$ and $S^{(2)} = \{2\}$ of the Transaction Data.

7. CONCLUSIONS

The construction of prediction intervals is a well-developed topic in the literature. However, research is largely limited to the setup of continuous-valued time series. For discrete-valued time series, classical prediction intervals are not meaningful while prediction sets may not be able to keep a desired coverage level. To solve this problem, we proposed to reserve the prediction problem: Instead of constructing prediction intervals or sets given a desired coverage level, we argued to estimate predictive probabilities for user-selected subsets of values that the process can take in the future and to manage the uncertainty associated with the prediction by constructing confidence intervals for the corresponding predictive probabilities. Different

asymptotic and bootstrap-based procedures have been proposed for the construction of such confidence intervals, covering parametric and non-parametric setups and allowing to address the important case of possible model misspecification in implementing such a prediction. Whether a parametric or a non-parametric approach is used for prediction is a decision that the user has to make. The methodology proposed in this paper aims to provide a statistical way to quantify the uncertainty associated with the prediction made. The corresponding confidence intervals are constructed in such a way that even doubts about the appropriateness of the model used for prediction can properly be taken into account, which is an important issue in case the decision of the user is in favor of a parametric model.

Acknowledgments

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

The research of M. Faymonville and C. Jentsch was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 437270842 (Model Diagnostics for Count Time Series) and 520388526 (TRR 391: Spatio-temporal Statistics for the Transition of Energy and Transport, Project A03). The research of E. Paparoditis was also funded by the Cyprus Academy of Sciences, Letters, and Arts.

References

Al-Osh, M. A. and Alzaid, A. A. (1987). "First-order integer-valued autoregressive (INAR(1)) process". *Journal of Time Series Analysis* 8.3, pp. 261–275. DOI: 10.1111/j.1467-9892.1987.tb00438.x.

Alj, A., Azrak, R., and Mélard, G. (2014). "On conditions in central limit theorems for martingale difference arrays". *Economics Letters* 123.3, pp. 305–309. DOI: 10.1016/j.econlet.2014.03.008.

Andrews, D. W. (1992). "Generic uniform convergence". *Econometric Theory* 8.2, pp. 241–257. DOI: 10.1017/S0266466600012780.

Athreya, K. B. and Fuh, C.-D. (1992). "Bootstrapping markov chains: Countable case". *Journal of Statistical Planning and Inference* 33.3, pp. 311–331. DOI: 10.1016/0378-3758(92)90002-A.

Billingsley, P. (1961). *Statistical Inference for Markov Processes*. University of Chicago Press. DOI: 10.1002/bimj.19650070109.

Bradley, C. R. (2007). *Introduction to Strong Mixing Conditions, Volume 1.* Kendrick Press. ISBN: 978-0974042763.

Brown, B. M. (1971). "Martingale central limit theorems". *The Annals of Mathematical Statistics* 42.1, pp. 59–66. DOI: 10.1214/aoms/1177693494.

Davis, R. A., Holan, S. H., Lund, R., and Ravishanker, N., eds. (2016). *Handbook of Discrete-Valued Time Series*. Handbooks of Modern Statistical Methods. Boca Raton, FL: Chapman & Hall/CRC. DOI: 10.1201/b19485.

Derman, C. (1956). "Some asymptotic distribution theory for Markov chains with a denumerable number of states". *Biometrika* 43.3-4, pp. 285–294. DOI: 10.1093/biomet/43.3-4.285.

Drost, F., Van den Akker, R., and Werker, B. (2009). "Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models". *Journal of the Royal Statistical Society. Series B* 71, Part 2, pp. 467–485. DOI: doi={10.1111/j.1467-9868.2008.00687.x}.

Du, J.-G. and Li, Y. (1991). "The integer valued autoregressive (INAR(p)) model". *Journal of Time Series Analysis* 12.2, pp. 129–142. DOI: 10.1111/j.1467-9892.1991.tb00073.x.

Faymonville, M. and Jentsch, C. (2025). *Joint semi-parametric INAR bootstrap inference for model coefficients and innovation distribution.* arXiv: 2507.11124 [stat.ME]. URL: https://arxiv.org/abs/2507.11124.

Faymonville, M., Jentsch, C., and Weiß, C. H. (2025). "Semi-parametric goodness-of-fit testing for INAR models". *Bernoulli* 31.4, pp. 3213–3234. DOI: 10.3150/24-BEJ1844.

Faymonville, M., Riffo, J., Rieger, J., and Jentsch, C. (2024). "spINAR: An R package for semi-parametric and parametric estimation and bootstrapping of integer-valued autoregressive (INAR) models". *Journal of Open Source Software* 9.97, p. 5386. DOI: 10.21105/joss.05386. URL: https://doi.org/10.21105/joss.05386.

Ferland, R., Latour, A., and Oraichi, D. (2006). "Integer-valued GARCH processes". *Journal of Time Series Analysis* 27.6, pp. 923–942. DOI: 10.1111/j.1467-9892.2006.00496.x.

Fokianos, K. and Kedem, B. (2003). "Regression theory for categorical times series". *Statistical Science* 18, pp. 357–376. DOI: 10.1214/ss/1076102425.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). "Poisson autoregression". *Journal of the American Statistical Association* 104.488, pp. 1430–1439. DOI: 10.1198/jasa.2009.tm08270.

Freeland, R. K. and McCabe, B. P. M. (2004). "Forecasting discrete valued low count time series". *International Journal of Forecasting* 20.3, pp. 427–434. DOI: 10.1016/S0169-2070(03)00014-1.

Freeland, R. K. (1998). "Statistical analysis of discrete time series with application to the analysis of workers compensation claims data". Ph.D. Thesis. Canada: University of British Columbia.

Heinen, A. (2003). "Modelling time series count data: an autoregressive conditional Poisson model." *CORE Discussion Paper* 2003-63. DOI: 10.2139/ssrn.1117187.

Homburg, A., Weiß, C. H., Alwan, L. C., Frahm, G., and Göb, R. (2023). "Pmf forecasting for count processes: A comprehensive performance analysis". *Springer*. DOI: 10.1007/978-3-031-14197-3_6.

Homburg, A., Weiß, C., Frahm, G., Alwan, L. C., and Göb, R. (2021). "Analysis and forecasting of risk in count processes". *Journal of Risk and Financial Management* 14.4, p. 182. DOI: 10.3390/jrfm14040182.

Jacobs, P. A. and Lewis, P. A. W. (1983). "Stationary discrete autoregressive-moving average time series generated by mixtures". *Journal of Time Series Analysis* 4.1, pp. 19–36. DOI: 10.1111/j.1467-9892.1983.tb00354..

Jahn, M. and Weiß, C. H. (2024). "Nonlinear GARCH-type models for ordinal time series". *Stochastic Environmental Research and Risk Assessment* 38, pp. 637–649. DOI: 10.1007/s00477-023-02591-1.

Jentsch, C. and Reichmann, L. (2021). "Generalized binary vector autoregressive processes". *Journal of Time Series Analysis* 43.2, pp. 285–311. DOI: 10.1111/jtsa.12614.

Kedem, B. and Fokianos, K. (2002). "Regression Models for Binary Time Series". *Modeling Uncertainty*. Springer. ISBN: 978-0-471-36355-2.

Knudson, C. and Vats, D. (2022). *stableGR: A stable Gelman-Rubin diagnostic for Markov chain Monte Carlo*. R package version 1.2. URL: https://CRAN.R-project.org/package=stableGR.

Liboschik, T., Fokianos, K., and Fried, R. (2017). "tscount: An R package for analysis of count time series following generalized linear models". *Journal of Statistical Software* 82.5, pp. 1–51. DOI: 10.18637/jss.v082.i05.

Liu, M., Zhu, F., and Zhu, K. (2022). "Modeling normalcy-dominant ordinal time series: an application to air quality level". *Journal of Time Series Analysis* 43.3, pp. 460–478. DOI: 10.1111/jtsa.12625.

McKenzie, E. (1985). "Some simple models for discrete variate time series". *Water Resources Bulletin* 21.4, pp. 645–650. DOI: 10.1111/j.1752-1688.1985.tb05379.x.

Pan, L. and Politis, D. (2016). "Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregression". *Journal of Statistical Planning and Inference* 177, pp. 1–27. DOI: 10.1016/j.jspi.2014.10.003.

Peligrad, M. (2012). "Central limit theorem for triangular arrays of non-homogeneous Markov chains". *Probability Theory and Related Fields* 154, pp. 409–428. DOI: 10.1007/s00440-011-0371-6.

Pruscha, H. (1993). "Categorical time series with a recursive scheme and with covariates". *Statistics* 24, pp. 43–57. DOI: 10.1080/02331888308802388.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Roberts, G. and Rosenthal, J. (1997). "Geometric ergodicity and hybrid Markov chains". *Electronic Communications in Probability* 2, pp. 13–25. DOI: 10.1214/ECP.v2-981.

Steutel, F. W. and Van Harn, K. (1979). "Discrete analogues of self-decomposability and stability". *Annals of Probability* 7.5, pp. 893–899. DOI: 10.1214/aop/1176994950.

Weiß, C. H. (2020). "Distance-based analysis of ordinal data and ordinal time series". *Journal of the American Statistical Association* 115, pp. 1189–1200. DOI: 10.1080/01621459.2019.1604370.

Weiß, C. H. (2010). "The INARCH(1) model for overdispersed time series of counts". *Communications in Statistics - Simulation and Computation* 39.6, pp. 1269–1291. DOI: 10.1080/03610918.2010.490317.

Weiß, C. H. (2018). *An Introduction to Discrete-Valued Time Series*. 1st ed. Wiley. ISBN: 978-1119096962.

Zhu, F. and Wang, D. (2009). "Estimation and testing for Poisson autoregressive model". *Metrika* 73, pp. 211–230. DOI: 10.1007/s00184-009-0274-z.

Appendix A. Auxiliary Results and Proofs

Proof of Proposition 2.4: From Derman (1956), we get that

$$\sqrt{n}\big(\widehat{P}^{(npara)}_{X_{t+1}\in S|X_t=x_n} - P_{X_{t+1}\in S|X_t=x_n}\big) \xrightarrow{d} \mathcal{N}\Big(0,\sigma_S^2\Big), \tag{A.1}$$

where $\sigma_S^2 = a_S^T \Sigma_S a_S$. By making use of the Delta method and writing

$$\begin{aligned}\sqrt{n}\big(\widehat{P}^{(npara)}_{X_{t+1}\in S|X_t=x_n} - P_{X_{t+1}\in S|X_t=x_n}\big) =& \sqrt{n}\left(\frac{\widehat{Q}_{S,x_n}}{\widehat{Q}_{x_n}} - \frac{Q_{S,x_n}}{Q_{x_n}}\right)\\ =& \sqrt{n}\left(g(\widehat{Q}_{S,x_n},\widehat{Q}_{x_n}) - g(Q_{S,x_n},Q_{x_n})\right),\end{aligned}$$

where $g(a,b) = a/b$ with $\nabla g(a,b) = (1/b, -a/b^2)^\top$ for $b > 0$, the formula for the limiting variance $\sigma_S^2 = a^T\Sigma a$ is easily obtained, where

$$a_S = \nabla g(Q_{S,x_n}, Q_{x_n}) = \left(\frac{1}{Q_{x_n}}, -\frac{Q_{S,x_n}}{Q_{x_n}^2}\right)^\top$$

and

$$\Sigma_S = \begin{pmatrix}\Sigma_S^{(1,1)} & \Sigma_S^{(1,2)}\\ \Sigma_S^{(2,1)} & \Sigma_S^{(2,2)}\end{pmatrix} \tag{A.2}$$

with $(\Sigma_S^{(2,1)})^\top = \Sigma_S^{(1,2)}$ and

$$\begin{aligned}\Sigma_S^{(1,1)} &= \lim_{n\to\infty} nCov(\widehat{Q}_{S,x_n},\widehat{Q}_{S,x_n}) = \sum_{h=-\infty}^{\infty} Cov\left(\mathbf{1}_{\{X_{h+1}\in S, X_h=x_n\}}, \mathbf{1}_{\{X_1\in S, X_0=x_n\}}\right)\\ &= \sum_{h=-\infty}^{\infty}\big(P(X_{h+1}\in S, X_h = x_n, X_1\in S, X_0 = x_n) - P^2(X_1\in S, X_0 = x_n)\big),\\ \Sigma_S^{(1,2)} &= \lim_{n\to\infty} nCov(\widehat{Q}_{S,x_n},\widehat{Q}_{x_n}) = \sum_{h=-\infty}^{\infty} Cov\left(\mathbf{1}_{\{X_{h+1}\in S, X_h=x_n\}}, \mathbf{1}_{\{X_0=x_n\}}\right)\\ &= \sum_{h=-\infty}^{\infty}\big(P(X_{h+1}\in S, X_h = x_n, X_0 = x_n) - P(X_1\in S, X_0 = x_n)P(X_0=x_n)\big),\\ \Sigma_S^{(2,2)} &= \lim_{n\to\infty} nCov(\widehat{Q}_{x_n},\widehat{Q}_{x_n}) = \sum_{h=-\infty}^{\infty} Cov\left(\mathbf{1}_{\{X_h=x_n\}}, \mathbf{1}_{\{X_0=x_n\}}\right)\\ &= \sum_{h=-\infty}^{\infty}\big(P(X_h = x_n, X_0 = x_n) - P^2(X_0 = x_n)\big),\end{aligned}$$

where the infinite sums in all three terms $\Sigma_S^{(1,1)}, \Sigma_S^{(1,2)}$ and $\Sigma_S^{(2,2)}$ converge due to the geometric ergodicity of the process.

□

Proof of Theorem 2.8: (i) Relying on the estimated relative frequencies $\widehat{P}^{(npara)}_{X_{t+1}=j|X_t=i}$, the bootstrap procedure described in Step 2 of Algorithm 2.5 corresponds to the Markov bootstrap

procedure I of Athreya and Fuh (1992). For this bootstrap procedure, it is shown in Theorem 4 of the afore cited paper that

$$\sqrt{n}\big(\widehat{P}^{*,(npara)}_{X^*_{t+1}\in S|X^*_t=x_n} - \widehat{P}^{(npara)}_{X_{t+1}\in S|X_t=x_n}\big) \xrightarrow{d} \mathcal{N}\Big(0,\tau_S^2\Big), \tag{A.3}$$

in probability, where the variance τ_S^2 can be calculated as described in Section 3 of the afore cited paper. The result follows because

$$\sqrt{n}\big(\widehat{P}^{*,(npara)}_{S,x_n} - \widehat{P}^{(npara)}_{S,x_n}\big) = \sum_{y_k\in S}\sqrt{n}\big(\widehat{P}^{(npara)}_{X^*_{t+1}=y_k|X^*_t=x_n} - \widehat{P}^{(npara)}_{X_{t+1}=y_k|X_t=x_n}\big)$$

and the continuity of the limiting distribution.

(ii) Since $(X_t(\theta), t\in\mathbb{Z})$ is for every $\theta\in\Theta$ aperiodic, positive recurrent and irreducible, it has a unique stationary distribution denoted by $\pi_i(\theta)$ for $i\in range(X_0)$. The bootstrap time series $X_1^*,\ldots,X_n^*$ can be considered as stemming from the Markov chain $(X_t(\widehat{\theta}), t\in\mathbb{Z})$. Without loss of generality, assume that this bootstrap time series starts with the stationary distribution. Then, $P(X_t^*=i)=\pi_i(\widehat{\theta})$. Using Theorem 3, Theorem 4 and Remark 4 of Athreya and Fuh (1992) and in order to establish assertion (ii) of the theorem, it suffices to show that, as $n\to\infty$,

(a) $P_{X^*_{t+1}=j|X^*_t=i} \xrightarrow{P} P^{(para)}_{X_{t+1}=j|X_t=i}(\theta_0)$ for every $i,j\in range(X_0)$, and
(b) $P(X_t^*=i)\xrightarrow{P}\pi_i(\theta_0)$ for every $i\in range(X_0)$.

Observe that condition (iii) of Theorem 3 of Athreya and Fuh (1992) is satisfied since the bootstrap Markov chain $(X_t(\widehat{\theta}), t\in\mathbb{Z})$ is irreducible, positive recurrent and aperiodic. Now, (a) follows since for any $i,j\in range(X_0)$,

$$P_{X^*_{t+1}=j|X^*_t=i} = P^{(para)}_{X_{t+1}=j|X_t=i}(\widehat{\theta}) \xrightarrow{P} P^{(para)}_{X_{t+1}=j|X_t=i}(\theta_0)$$

by the fact that $\widehat{\theta}\xrightarrow{P}\theta_0$ and the continuity of the parametric one step transition probabilities with respect to θ. Similarly, $P(X_t^*=i)=\pi_i(\widehat{\theta})\xrightarrow{P}\pi_i(\theta_0)$. □

Proof of Theorem 3.1: By the continuity of $\mathrm{E}(\log(P^{(para)}_{X_t|X_{t-1}}(\theta))$ and in order to establish the assertion of the theorem, it suffices to show that, as $n\to\infty$,

$$\sup_{\theta\in\Theta}\left|\frac{1}{n-1}\sum_{t=2}^{n}\log\big(P^{(para)}_{X_t|X_{t-1}}(\theta)\big) - \mathrm{E}\big(\log P^{(para)}_{X_t|X_{t-1}}(\theta)\big)\right| \xrightarrow{P} 0, \tag{A.4}$$

since by Assumption 4(i), θ_0 is the unique maximum of $\mathrm{E}\big(\log\big(P^{(para)}_{X_t|X_{t-1}}(\theta)\big)\big)$. To see why (A.4) holds true, we verify the conditions stated in Theorem 1(a) of Andrews (1992). Let

$$G_n(\theta) := \frac{1}{n-1}\sum_{t=2}^{n}\log\big(P^{(para)}_{X_t|X_{t-1}}(\theta)\big) - \mathrm{E}\big(\log\big(P^{(para)}_{X_t|X_{t-1}}(\theta)\big)\big).$$

Θ is compact by Assumption 2(i). The pointwise convergence $G_n(\theta)\xrightarrow{P}0$ as $n\to\infty$, for any $\theta\in\Theta$, follows from the ergodicity of $\{X_t, t\in\mathbb{Z}\}$, see Assumption 1. It remains to show the

stochastic equicontinuity condition, that is, Assumption SE in the afore cited theorem. For this we have to show that for any $\epsilon > 0$ there exists $\eta > 0$ such that

$$\limsup_{n\to\infty} P\Big(\sup_{\theta\in\Theta} \sup_{\widetilde{\theta}\in B(\theta,\eta)} \big|G_n(\widetilde{\theta}) - G_n(\theta)\big| > \epsilon \Big) < \epsilon, \tag{A.5}$$

where $B(\theta,\eta)$ denotes the closed ball in Θ of radius $\eta \geq 0$ centered at θ. Notice that $\sup_{\theta\in\Theta} \big\|\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta)\| < \infty$ since $\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta)$ is continuous on a compact set. By Markov's inequality and the mean value theorem, we have for some θ' such that $\|\theta'-\theta\| \leq \|\widetilde{\theta}-\theta\|$, that

$$\begin{aligned}
\limsup_{n\to\infty} P\Big(\sup_{\theta\in\Theta} \sup_{\widetilde{\theta}\in B(\theta,\eta)} \big|G_n(\widetilde{\theta}) - G_n(\theta)\big| > \epsilon \Big) &\leq \frac{1}{\epsilon} \limsup_{n\to\infty} \mathrm{E} \sup_{\theta\in\Theta} \sup_{\widetilde{\theta}\in B(\theta,\eta)} \big|(\widetilde{\theta}-\theta)^\top \nabla_\theta G_n(\theta')\big| \\
&\leq \frac{\eta}{\epsilon} \limsup_{n\to\infty} \mathrm{E} \sup_{\theta\in\Theta} \big\|\nabla_\theta G_n(\theta)\big\| \\
&\leq \frac{\eta}{\epsilon}\, \mathrm{E} \sup_{\theta\in\Theta} \big\|\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta)\| \\
&\quad + \frac{\eta}{\epsilon} \sup_{\theta\in\Theta} \big\|\mathrm{E}(\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta))\| \\
&\leq \frac{2\eta}{\epsilon} \mathrm{E} \sup_{\theta\in\Theta} \big\|\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta)\| \\
&\leq \frac{2\eta}{\epsilon} C,
\end{aligned}$$

for a constant $C > 0$. Thus, for $0 < \eta < \epsilon^2/(2C)$, condition (A.5) is satisfied. □

Proof of Theorem 3.2:

(i). In this case, $X^*_1, \ldots, X^*_n$ is generated using a model from the class $\mathcal{M}_\theta$ with estimated parameter $\widehat{\theta}_{ML}$. The conditional maximum likelihood estimator $\widehat{\theta}^*_{ML}$ based on this bootstrap time series is defined as

$$\widehat{\theta}^*_{ML} = \mathrm{argmax}_{\theta\in\Theta} l_n(\theta|X^*_1) = \mathrm{argmax}_{\theta\in\Theta} \sum_{t=2}^{n} \log\big(P^{(para)}_{X^*_t|X^*_{t-1}}(\theta)\big).$$

Note that

$$\widehat{\theta}_{ML} = \mathrm{argmax}_{\theta\in\Theta} \mathrm{E}^*\big(\log\big(P^{(para)}_{X^*_t|X^*_{t-1}}(\theta)\big) = \mathrm{argmax}_{\theta\in\Theta} \mathrm{E}_{\widehat{\theta}_{ML}}\big(\log\big(P^{(para)}_{X^*_t|X^*_{t-1}}(\theta)\big), \tag{A.6}$$

where

$$\mathrm{E}^*\big(\log\big(P^{(para)}_{X^*_t|X^*_{t-1}}(\theta)\big) = \sum_r \sum_s \log(P^{(para)}_{X^*_t=r|X^*_{t-1}=s}(\theta)) P^{(para)}_{\widehat{\theta}_{ML}}(X^*_t = r, X^*_{t-1} = s)$$

and the double summation extends over the entire range of X_0. We use the expansion

$$\frac{1}{n-1} l''_n(\widetilde{\theta}^*_n|X^*_1)\sqrt{n-1}\big(\widehat{\theta}^*_{ML} - \widehat{\theta}_{ML}\big) = -\frac{1}{\sqrt{n-1}} l'_n(\widehat{\theta}_{ML}|X^*_1), \tag{A.7}$$

see also (3.1), for some $\widetilde{\theta}^*_n$ such that $\|\widetilde{\theta}^*_n - \widehat{\theta}_{ML}\| \leq \|\widehat{\theta}^*_{ML} - \widehat{\theta}_{ML}\|$. Let

$$\nabla_\theta \log(P^{(para)}_{X^*_t|X^*_{t-1}}(\theta))\big|_{\theta=\theta'} =: g(X^*_t, X^*_{t-1}; \theta').$$

Then, we have

$$\frac{1}{\sqrt{n-1}} l_n'(\widehat{\theta}_{ML}|X_1^*) = \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} g(X_t^*, X_{t-1}^*; \widehat{\theta}_{ML}), \tag{A.8}$$

and (A.8) is (conditional on $X_1, \ldots, X_n$) a sequence of martingale differences. This holds true since

$$\begin{aligned} &\mathrm{E}^*\big(\nabla_\theta \log(P^{(para)}_{X_t^*|X_{t-1}^*}(\theta))\big|_{\theta=\widehat{\theta}_{ML}} \big| X_{t-1}^*\big) \qquad (A.9)\\ &\quad = \sum_r \frac{1}{P^{(para)}_{X_t^*=r|X_{t-1}^*}(\widehat{\theta}_{ML})} \nabla_\theta P^{(para)}_{X_t^*=r|X_{t-1}^*}(\theta)\big|_{\widehat{\theta}_{ML}} P^{(para)}_{X_t^*=r|X_{t-1}^*}(\widehat{\theta}_{ML}) \\ &\quad = \sum_r \nabla_\theta P^{(para)}_{X_t^*=r|X_{t-1}^*}(\widehat{\theta}_{ML}) \\ &\quad = \nabla_\theta \Big(\sum_r P^{(para)}_{X_t^*=r|X_{t-1}^*}(\widehat{\theta}_{ML}) \Big) \\ &\quad = 0, \end{aligned}$$

where all sums go over the range of X_t^*. Notice that by (A.6) and the exchangeability of differentiation and expectation, see Assumption 4(iii), it is easily seen that $\mathrm{E}^*(g(X_t^*, X_{t-1}^*; \widehat{\theta}_{ML})) = 0$.

For the matrix of second-order partial derivatives and since the elements of the matrix $\nabla_\theta^2 \log(P^{(para)}_{X_t|X_{t-1}}(\theta))$ are Lipschitz continuous functions of θ, we get for the Frobenius norm $\|\cdot\|_F$,

$$\begin{aligned} \Big\| \Big(\frac{1}{n-1} l_n''(\widetilde{\theta}_n^*|X_1^*)\Big) - \Big(\frac{1}{n-1} l_n''(\theta_0|X_1^*)\Big) \Big\|_F &\le \Big\| \Big(\frac{1}{n-1} l_n''(\widetilde{\theta}_n^*|X_1^*)\Big) - \Big(\frac{1}{n-1} l_n''(\widehat{\theta}_{ML}|X_1^*)\Big) \Big\|_F \\ &\quad + \Big\| \Big(\frac{1}{n-1} l_n''(\widehat{\theta}_{ML}|X_1^*)\Big) - \Big(\frac{1}{n-1} l_n''(\theta_0|X_1^*)\Big) \Big\|_F \\ &\le O_P\big(\|\widehat{\theta}_{ML}^* - \widehat{\theta}_{ML}\|\big) + O_P\big(\|\widehat{\theta}_{ML} - \theta_0\|\big). \qquad (A.10) \end{aligned}$$

The second term of the last inequality goes to zero because $\widehat{\theta}_{ML} \xrightarrow{P} \theta_0$. The first term of (A.10) vanishes asymptotically by the triangular inequality and because $\|\widehat{\theta}_{ML} - \theta_0\| \xrightarrow{P} 0$ and $\|\widehat{\theta}_{ML}^* - \theta_0\| \xrightarrow{P} 0$, in probability. To establish the latter statement, it suffices to show that

$$\sup_{\theta\in\Theta} \Big| \frac{1}{n-1} \sum_{t=2}^{n} \log P^{(para)}_{X_t^*|X_{t-1}^*}(\theta) - \mathrm{E}_{\theta_0}\big(\log\big(P^{(para)}_{X_t|X_{t-1}}(\theta)\big)\big) \Big| \xrightarrow{P} 0, \tag{A.11}$$

where

$$\mathrm{E}_{\theta_0}\big(\log\big(P^{(para)}_{X_t|X_{t-1}}(\theta)\big) = \sum_r \sum_s \log P^{(para)}_{X_t=r|X_{t-1}=s}(\theta) P^{(para)}_{\theta_0}(X_t = r, X_{t-1} = s).$$

Notice that in the above set up, (X_t, X_{t-1}) follows the distribution generated by a parametric model with parameter θ_0 while (X_t^*, X_{t-1}^*) that of the same model but with parameter $\widehat{\theta}_{ML}$.

Using the notation

$$W_n^*(\theta) := \frac{1}{n-1} \sum_{t=2}^{n} \log P^{(para)}_{X_t^*|X_{t-1}^*}(\theta) - \mathrm{E}_{\theta_0}\big(\log\big(P^{(para)}_{X_t|X_{t-1}}(\theta)\big)\big)$$

and arguing as in the proof of (A.4), it suffices to show that $W_n^*(\theta) \to 0$ in probability for any $\theta \in \Theta$ and that $\eta > 0$ exists such that, in probability,

$$\limsup_{n\to\infty} P\Big(\sup_{\theta\in\Theta} \sup_{\widetilde{\theta}\in B(\theta,\eta)} \big|W_n^*(\widetilde{\theta}) - W_n^*(\theta)\big| > \epsilon \Big) < \epsilon, \tag{A.12}$$

for all $\epsilon > 0$. We have

$$\begin{aligned}
&\limsup_{n\to\infty} \mathrm{E}^* \sup_{\theta\in\Theta} \sup_{\widetilde{\theta}\in B(\theta,\eta)} \big|W_n^*(\widetilde{\theta}) - W_n^*(\theta)\big| \\
&\qquad \le \eta \limsup_{n\to\infty} \mathrm{E}^* \sup_{\theta\in\Theta} \|(n-1)^{-1} \sum_{t=2}^{n} \nabla_\theta \log P_{X_t^*|X_{t-1}^*}\| \\
&\qquad\quad + \eta \sup_{\theta\in\Theta} \|\mathrm{E}_{\theta_0}(\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta))\| \\
&\qquad \le \eta \limsup_{n\to\infty} \sum_r \sum_s \sup_{\theta\in\Theta} \|\nabla_\theta \log P^{(para)}_{X_2=r|X_1=s}(\theta)\| P^{(para)}_{\widehat{\theta}_{ML}}(X_2 = r, X_1 = s) \\
&\qquad\quad + \eta \sup_{\theta\in\Theta} \|\mathrm{E}_{\theta_0}(\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta))\| \\
&\qquad = \eta \sum_r \sum_s \sup_{\theta\in\Theta} \|\nabla_\theta \log P^{(para)}_{X_2=r|X_1=s}(\theta)\| P^{(para)}_{\theta_0}(X_2 = r, X_1 = s) \\
&\qquad\quad + \eta \sup_{\theta\in\Theta} \|\mathrm{E}_{\theta_0} \nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta)\| \\
&\qquad \le 2\eta\, \mathrm{E}_{\theta_0}\big(\sup_{\theta\in\Theta} \|\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta)\|\big).
\end{aligned}$$

Note that the equality before the last inequality follows because $\widehat{\theta}_{ML} \xrightarrow{P} \theta_0$ implies $P^{(para)}_{\widehat{\theta}_{ML}}(X_2 = r, X_1 = s) \xrightarrow{P} P^{(para)}_{\theta_0}(X_2 = r, X_1 = s)$ for every r, s. This together with the boundedness in probability of $\sup_{\theta\in\Theta} \|\nabla_\theta \log P^{(para)}_{X_2=r|X_1=s}(\theta)\|$ leads to

$$\begin{aligned}
&\sum_r \sum_s \sup_{\theta\in\Theta} \|\nabla_\theta \log P^{(para)}_{X_2=r|X_1=s}(\theta)\| \big| P^{(para)}_{\widehat{\theta}_{ML}}(X_2 = r, X_1 = s) - P^{(para)}_{\theta_0}(X_2 = r, X_1 = s)\big| \\
&\qquad\qquad \le O_P(1) \sum_r \sum_s \big| P^{(para)}_{\widehat{\theta}_{ML}}(X_2 = r, X_1 = s) - P^{(para)}_{\theta_0}(X_2 = r, X_1 = s)\big| \xrightarrow{P} 0,
\end{aligned}$$

where the last convergence follows by Scheffé's Theorem. Therefore, for $0 < \eta < \epsilon^2/2C$, where $C = \mathrm{E}_{\theta_0}\big(\sup_{\theta\in\Theta} \|\nabla_\theta \log P^{(para)}_{X_2|X_1}(\theta)\|\big) < \infty$, assertion (A.12) follows.

By the weak law of large numbers and (A.10), we have

$$\begin{aligned}
&\Big\| \mathrm{E}\big(\nabla_\theta^2 \log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big) - \Big(\frac{1}{n-1} l_n''(\widetilde{\theta}_n^*|X_1^*)\Big)\Big\|_F \\
&\qquad \le \Big\| E\big(\nabla_\theta^2 \log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big) - \Big(\frac{1}{n-1} l_n''(\theta_0|X_1)\Big)\Big\|_F \\
&\qquad\quad + \Big\| \Big(\frac{1}{n-1} l_n''(\theta_0|X_1)\Big) - \Big(\frac{1}{n-1} l_n''(\widehat{\theta}_{ML}^*|X_1^*)\Big)\Big\|_F \\
&\qquad\quad + \Big\| \Big(\frac{1}{n-1} l_n''(\widehat{\theta}_{ML}^*|X_1^*)\Big) - \Big(\frac{1}{n-1} l_n''(\widetilde{\theta}_n^*|X_1^*)\Big)\Big\|_F \\
&\qquad \xrightarrow{P} 0.
\end{aligned}$$

Invoking a central limit theorem for triangular arrays of martingale differences, see Brown (1971) and Alj et al. (2014), to the sequence

$$\frac{1}{\sqrt{n-1}}\sum_{t=2}^{n}\big(g(X_t^*,X_{t-1}^*;\widehat{\theta}_{ML})-E^*\big(g(X_t^*,X_{t-1}^*,;\hat{\theta}_{ML})\big),$$

also see (A.8), and using the fact that the matrix

$$\mathrm{E}_{\theta_0}\big(\nabla_\theta^2\log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big),$$

is positive definite, see Assumption 2, the assertion of the theorem follows.

Proof of assertion (ii). The proof of this assertion essentially follows the arguments used for establishing assertion (i) of the theorem with the main difference concerning the CLT used. Similar to (A.7), we have

$$\frac{1}{n-1}l_n''(\widetilde{\theta}_n^*|X_1^*)\sqrt{n-1}\big(\widehat{\theta}_{ML}^*-\widehat{\theta}_{ML}\big)=-\frac{1}{\sqrt{n-1}}l_n'(\widehat{\theta}_{ML}|X_1^*),\tag{A.13}$$

for some $\widetilde{\theta}_n^*$ such that $\|\widetilde{\theta}_n^*-\widehat{\theta}_{ML}\|\leq\|\widehat{\theta}_{ML}^*-\widehat{\theta}_{ML}\|$. Notice that the centering here by $\widehat{\theta}_{ML}$ is justified by the fact that,

$$\begin{aligned}\mathrm{argmax}_{\theta\in\Theta}\mathrm{E}^*(\log P_{X_t^*|X_{t-1}^*}(\theta))&=\mathrm{argmax}_{\theta\in\Theta}\sum_r\sum_s\log P_{X_t=r|X_{t-1}=s}(\theta)\widehat{P}^{(npara)}_{X_t=r|X_{t-1}=s}\\&=\mathrm{argmax}_{\theta\in\Theta}\sum_{t=2}^{n}\log P_{X_t|X_{t-1}}(\theta)=\widehat{\theta}_{ML}.\end{aligned}$$

According to (A.8), in (A.13) as well, we have

$$\frac{1}{\sqrt{n-1}}l_n'(\widehat{\theta}_{ML}|X_1^*)=\frac{1}{\sqrt{n-1}}\sum_{t=2}^{n}g(X_t^*,X_{t-1}^*;\widehat{\theta}_{ML})=:\frac{1}{\sqrt{n-1}}\sum_{t=2}^{n}Y_{t,n}^*,\tag{A.14}$$

where the pseudo time series $X_1^*,\ldots,X_n^*$ is generated using the non-parametrically estimated one step transition probabilities $\widehat{P}^{(npara)}_{X_{t+1}=x_j|X_t=x_i}$ for $x_i,x_j\in range(X_0)$ (instead of the parametrically estimated one step transition probabilities $P^{(para)}_{X_{t+1}=x_j|X_t=x_i}(\widehat{\theta}_{\mathrm{ML}})$ used in part (i)). In the following, we will use the central limit theorem of Peligrad (2012, Theorem 1) for triangular arrays of non-homogeneous Markov chains and adapt it to our (bootstrap) setup. Note that here we are dealing with the homogeneous case only, which simplifies arguments. By forming the two-dimensional Markov chain $((X_t,X_{t-1})',t\in\mathbb{Z})$, the latter shares important properties with $(X_t,t\in\mathbb{Z})$. In particular, the corresponding maximal correlation coefficient is still strictly smaller than 1.

Moreover, conditional on $X_1,\ldots,X_n$, the bootstrap process $(X_t^*,t\in\mathbb{Z})$ fulfills these properties as well with probability tending to one. We have

$$\rho_1^*=\sup_{f,g}\left\{\frac{\big|E^*(f(X_i^*)g(X_{i-1}^*))-E^*(f(X_i^*))E^*(g(X_{i-1}^*))\big|}{\sqrt{E^*(f^2(X_i^*))}\sqrt{E^*(g^2(X_{i-1}^*))}};\|f(X_i^*)\|_2^*<\infty,\ \|g(X_{i-1}^*)\|_2^*<\infty\right\},$$

where we used the notation $\|X^*\|_p^* = (E^*((X^*)^p))^{1/p}$ for $p > 1$. We show that, as $n \to \infty$, $\rho_1^* \to \rho_1$ in probability holds true, where ρ_1 is the corresponding coefficient of the underlying Markov process which is assumed to be strictly smaller than 1. After having achieved this, we can argue that $\rho_1^* < 1$ holds true in probability and we can make use of Peligrad (2012, Theorem 1) to establish a bootstrap CLT for (A.14). Toward this goal, we show that

$$|\rho_1^* - \rho_1| \to 0, \tag{A.15}$$

in probability, as $n \to \infty$. Consider the nominators of ρ_1^* and ρ_1. Then, we have

$$\begin{aligned}&|(E^*(f(X_i^*)g(X_{i-1}^*)) - E^*(f(X_i^*))E^*(g(X_{i-1}^*))) - (E(f(X_i)g(X_{i-1})) - E(f(X_i))E(g(X_{i-1})))|\\ \leq&|E^*(f(X_i^*)g(X_{i-1}^*)) - E(f(X_i)g(X_{i-1}))| + |E^*(f(X_i^*))E^*(g(X_{i-1}^*)) - E(f(X_i))E(g(X_{i-1}))|.\end{aligned}$$

We show that $|E^*(f(X_i^*)g(X_{i-1}^*)) - E(f(X_i)g(X_{i-1}))| \to 0$ and $|E^*(f(X_i^*))E^*(g(X_{i-1}^*)) - E(f(X_i))E(g(X_{i-1}))| \to 0$, where the latter is implied by $|E^*(f(X_i^*)) - E(f(X_i))| \to 0$. Similarly, we can deal with the denominator, where we have to show that $|E^*(f^2(X_i^*)) - E(f^2(X_i))| \to 0$. We only elaborate on the most cumbersome term $|E^*(f(X_i^*)g(X_{i-1}^*)) - E(f(X_i)g(X_{i-1}))|$ (all other terms can be dealt with analogously). We show that

$$E^*(f(X_i^*)g(X_{i-1}^*)) \to E(f(X_i)g(X_{i-1}))$$

in probability, as $n \to \infty$. For this purpose, let $\mathcal{N}$ denote the state space of $(X_t, t \in \mathbb{Z})$ and $\mathcal{N}_n^*$ the state space of $(X_t^*, t \in \mathbb{Z})$. Note that, for all $n \in \mathbb{N}$, we have $\mathcal{N}_n^* \subseteq \mathcal{N}$ with $\mathcal{N}_n^*$ being finite. Then,

$$\begin{aligned}E^*(f(X_i^*)g(X_{i-1}^*)) =& \sum_{r,s\in\mathcal{N}_n^*} f(r)g(s)P^*(X_i^* = r, X_{i-1}^* = s)\\ =& \sum_{r,s\in\mathcal{N}_n^*} f(r)g(s)\left(P^*(X_i^* = r, X_{i-1}^* = s) - P(X_i = r, X_{i-1} = s)\right)\\ &+ \sum_{r,s\in\mathcal{N}_n^*} f(r)g(s)P(X_i = r, X_{i-1} = s)\\ =& I_n + II_n.\end{aligned}$$

As $\mathcal{N}_n^* \to \mathcal{N}$ monotonically as $n \to \infty$, we get by the monotone convergence theorem that $II_n \to E(f(X_i)g(X_{i-1}))$ in probability. It remains to show that $I_n = o_p(1)$. We have

$$\begin{aligned}
I_n &= \sum_{r,s\in\mathcal{N}_n^*} f(r)g(s)\left(P^*(X_i^*=r, X_{i-1}^*=s) - P(X_i=r, X_{i-1}=s)\right) \\
&= \sum_{r,s\in\mathcal{N}} f(r)g(s)\left(P^*(X_i^*=r, X_{i-1}^*=s) - P(X_i=r, X_{i-1}=s)\right)\mathbb{1}(P^*(X^*=r)P^*(X^*=s)>0) \\
&= \sum_{r,s\in\mathcal{N}} \Big[f(r)g(s)\mathbb{1}(P^*(X^*=r)P^*(X^*=s)>0)\Big]\left(P^*(X_i^*=r, X_{i-1}^*=s) - P(X_i=r, X_{i-1}=s)\right) \\
&\leq \sum_{r,s\in\mathcal{N}} \Big|f(r)g(s)\mathbb{1}(P^*(X^*=r)P^*(X^*=s)>0)\Big| \cdot \left|P^*(X_i^*=r, X_{i-1}^*=s) - P(X_i=r, X_{i-1}=s)\right|.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\Big|f(r)g(s)\mathbb{1}(P^*(X^*=r)P^*(X^*=s)>0)\Big| \\
\leq &\sup_{r,s}\Big|f(r)g(s)\mathbb{1}(P^*(X^*=r)P^*(X^*=s)>0)\Big| \\
\leq &\sup_{r}\Big|f(r)\mathbb{1}(P^*(X^*=r)>0)\Big| \sup_{s}\Big|g(s)\mathbb{1}(P^*(X^*=s)>0)\Big| \\
\leq &\left(\max\{|f(X_1)|,\ldots,|f(X_n)|\}\right)^2.
\end{aligned}$$

Hence, it suffices to show that

$$\left(\max\{|f(X_1)|,\ldots,|f(X_n)|\}\right)^2 \sum_{r,s\in\mathcal{N}} \left|P^*(X_i^*=r, X_{i-1}^*=s) - P(X_i=r, X_{i-1}=s)\right| = o_P(1).$$

For this purpose, we aim to show

$$\frac{\left(\max\{|f(X_1)|,\ldots,|f(X_n)|\}\right)^2}{\sqrt{n}} = o_P(1) \tag{A.16}$$

$$\text{and } \sqrt{n}\sum_{r,s\in\mathcal{N}} \left|P^*(X_i^*=r, X_{i-1}^*=s) - P(X_i=r, X_{i-1}=s)\right| = O_P(1) \tag{A.17}$$

in the following. Starting with term (A.16), for any $x > 0$, we have

$$\begin{aligned}
P\left(\frac{(\max\{|f(X_1)|,\dots,|f(X_n)|\})^2}{\sqrt{n}} > x\right) =& P\left(\max\{|f(X_1)|,\dots,|f(X_n)|\} > x^{1/2}n^{1/4}\right)\\
=& P\left(\bigcup_{i=1}^{n}\left\{|f(X_i)| > x^{1/2}n^{1/4}\right\}\right)\\
\leq& \sum_{i=1}^{n} P\left(\left\{|f(X_i)| > x^{1/2}n^{1/4}\right\}\right)\\
=& nP\left(\left\{|f(X_1)|^k > (x^{1/2}n^{1/4})^k\right\}\right)\\
\leq& n\frac{E(|f(X_1)|^k)}{(x^{1/2}n^{1/4})^k}\\
=& \frac{n^{1-k/4}}{x^{k/2}}E(|f(X_1)|^k).
\end{aligned}$$

Hence, for $k > 4$, the last right-hand side converges to zero in probability, if $\mathrm{E}(|f(X_1)|^k) < \infty$ holds for all square-integrable functions f. That is, we need $8 + \delta$ moments, which holds as we have assumed $12 + \delta$ moments for the process $(X_t, t \in \mathbb{Z})$.

Continuing with (A.17), using $S = \{x_i, i \in \mathbb{N}\}$, we have

$$\begin{aligned}
&\sqrt{n}\sum_{r,s\in\mathcal{N}}|\widehat{P}(X_1 = r, X_0 = s) - P(X_1 = r, X_0 = s)|\\
\leq&\sqrt{n}\sum_{i,j=1}^{K}|\widehat{P}(X_1 = x_i, X_0 = x_j) - P(X_1 = x_i, X_0 = x_j)|\\
&+\sqrt{n}\sum_{i=K+1}^{\infty}\sum_{j=1}^{\infty}|\widehat{P}(X_1 = x_i, X_0 = x_j) - P(X_1 = x_i, X_0 = x_j)|\\
&+\sqrt{n}\sum_{i=1}^{\infty}\sum_{j=K+1}^{\infty}|\widehat{P}(X_1 = x_i, X_0 = x_j) - P(X_1 = x_i, X_0 = x_j)|\\
=&I_{K,n} + II_{K,n} + III_{K,n},
\end{aligned}$$

where K is some arbitrarily large positive integer. For all fixed K, we have that $I_{K,n} = O_P(1)$, because a multivariate CLT holds for all finitely many relative frequencies, when properly

centered and scaled with $\sqrt{n}$. Further, the second term $II_{K,n}$ can be bounded by

$$\begin{aligned}
&\sqrt{n}\sum_{i=K+1}^{\infty}\sum_{j=1}^{\infty}|\widehat{P}(X_1=x_i,X_0=x_j)|+\sqrt{n}\sum_{i=K+1}^{\infty}\sum_{j=1}^{\infty}|P(X_1=x_i,X_0=x_j)|\\
\leq&\sqrt{n}\sum_{i=K+1}^{\infty}\widehat{P}(X_1=x_i)+\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)\\
=&\sqrt{n}(1-\sum_{i=1}^{K}\widehat{P}(X_1=x_i))+\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)\\
=&\sqrt{n}(1-\sum_{i=1}^{K}\big(\widehat{P}(X_1=x_i)-P(X_1=x_i)\big))-\sqrt{n}\sum_{i=1}^{K}P(X_1=x_i)+\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)\\
\leq&\sqrt{n}(1-\sum_{i=1}^{K}\big(\widehat{P}(X_1=x_i)-P(X_1=x_i)\big))-\sqrt{n}\Big(1-\sum_{i=K+1}^{\infty}P(X_1=x_i)\Big)+\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)\\
=&-\sqrt{n}\sum_{i=1}^{K}\big(\widehat{P}(X_1=x_i)-P(X_1=x_i)\big)+\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)+\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)\\
\leq&O_P(1)+2\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)
\end{aligned}$$

for all K sufficiently large. Analogously, we get the same for term $III_{K,n}$. Hence, altogether, we have

$$\frac{(\max\{|f(X_1)|,\ldots,|f(X_n)|\})^2}{\sqrt{n}}O_P(1)=o_P(1)$$

and it remains to argue that, for all K, we also have

$$\frac{(\max\{|f(X_1)|,\ldots,|f(X_n)|\})^2}{\sqrt{n}}2\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)=o_P(1).$$

By the same arguments used above for (A.16), we get

$$\begin{aligned}
&\frac{(\max\{|f(X_1)|,\ldots,|f(X_n)|\})^2}{\sqrt{n}}2\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)\\
=&O_P\left(n^{1-k/4}\mathrm{E}(|f(X_1)|^k)\sqrt{n}\sum_{i=K+1}^{\infty}P(X_1=x_i)\right)\\
=&O_P\left(n^{(6-k)/4}\mathrm{E}(|f(X_1)|^k)\sum_{i=K+1}^{\infty}P(X_1=x_i)\right).
\end{aligned}$$

Hence, for $k>6$, the last right-hand side converges to zero in probability, if $\mathrm{E}(|f(X_1)|^k)<\infty$ holds for all square-integrable functions f. This is the case, as we have assumed $12+\delta$ moments for the process $(X_t, t\in\mathbb{Z})$.

Now since $\rho_1^*<1$ and having established that $\rho_1^*\to\rho_1$, in probability, there exists a $\widetilde{\rho}_1$ with $\rho_1<\widetilde{\rho}_1<1$ such that, with high probability, $\rho_1^*<\widetilde{\rho}_1<1$ holds true. Hence, we can use

Theorem 1 in Peligrad (2012). According to (A.14), we are concerned with

$$L_n^* := \sum_{t=2}^{n} Y_{t,n}^*, \tag{A.18}$$

where

$$Y_{t,n}^* = g_n(X_t^*, X_{t-1}^*) := g(X_t^*, X_{t-1}^*; \widehat{\theta}_{ML}) = \nabla_\theta \log(P^{(para)}_{X_t^*|X_{t-1}^*}(\theta))\big|_{\theta=\widehat{\theta}_{ML}}. \tag{A.19}$$

Moreover, as shown above, we have $\mathrm{E}^*(Y_{t,n}^*) = 0$ as well as $\mathrm{E}^*((Y_{t,n}^*)^2) < \infty$ and let $\widehat{\sigma}_n^2 := \mathrm{Var}^*(L_n^*)$. Then, conditional on $X_1, \ldots, X_n$, we have

$$\max_{1\le t\le n} |Y_{t,n}^*| = \max_{1\le t\le n} \nabla_\theta \log(P^{(para)}_{X_t^*|X_{t-1}^*}(\theta))\big|_{\theta=\widehat{\theta}_{ML}} =: C_n^*(\widehat{\theta}_{ML}), \tag{A.20}$$

where, for each fixed n and conditionally on $\widehat{\theta}_{ML}$, $C_n^* = C_n^*(\widehat{\theta}_{ML})$ is actually a finite (non-random) constant as $X_t^* \le \max\{X_1, \ldots, X_n\}$ by construction of the non-parametric bootstrap procedure. Also by adapting the notation in Peligrad (2012), let

$$\widehat{\lambda}_n := 1 - \widehat{\rho}_{n,1} := 1 - \rho_1^*. \tag{A.21}$$

Then, it remains to show that

$$\frac{C_n^*(1 + |\log(\widehat{\lambda}_n)|)}{\widehat{\lambda}_n \widehat{\sigma}_n} \xrightarrow{P} 0 \tag{A.22}$$

as $n \to \infty$, in order to prove that

$$\frac{L_n^*}{\widehat{\sigma}_n} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{in probability.}$$

Making use of the arguments from above, condition (A.22) holds. To see this, note that $\widehat{\lambda}_n > 0$ in probability as $\widehat{\lambda}_n < 1$ in probability, where the latter holds, because the bootstrap process $(X_t^*, t \in \mathbb{Z})$ shares the property of a maximal correlation coefficient smaller than 1 (in probability) with the process $(X_t, t \in \mathbb{Z})$, who fulfills this by assumption. Hence, we have

$$\frac{C_n^*(1 + |\log(\widehat{\lambda}_n)|)}{\widehat{\lambda}_n \widehat{\sigma}_n} = O_P\left(\frac{C_n^*}{\widehat{\sigma}_n}\right) = o_P(1) \tag{A.23}$$

by assumption (3.7) as $1/\widehat{\sigma}_n = O_p(1/\sqrt{n})$. Finally, we get

$$-\frac{1}{\sqrt{n-1}} l_n'(\widehat{\theta}_{ML}|X_1^*) = -\left(\frac{1}{n-1}\widehat{\sigma}_n^2\right)^{1/2} \frac{L_n^*}{\widehat{\sigma}_n} \xrightarrow{d} \mathcal{N}(0, \mathrm{E}\big(\nabla_\theta^2 \log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big)),$$

because of $\frac{1}{n-1}\widehat{\sigma}_n^2 \xrightarrow{P} \mathrm{E}\big(\nabla_\theta^2 \log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big)$ by the weak law of large numbers. Finally, similar to the arguments employed for part (i), we also have

$$\frac{1}{n-1} l_n''(\widetilde{\theta}_n^*|X_1^*) \xrightarrow{P} \mathrm{E}\big(\nabla_\theta^2 \log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big),$$

which completes the proof of part (ii) as $\mathrm{E}\big(\nabla_\theta^2 \log(P^{(para)}_{X_t|X_{t-1}}(\theta))\big|_{\theta=\theta_0}\big)$ is positive definite by Assumption 4.

□