No. 2008–53

EFFICIENT ESTIMATION OF AUTOREGRESSION PARAMETERS AND INNOVATION DISTRIBUTIONS FOR SEMIPARAMETRIC INTEGER-VALUED AR(*p*) MODELS

By Feike C. Drost, Ramon van den Akker, Bas J.M. Werker

May 2008

This is a revised version of CentER Discussion Paper No. 2007-23

March 2007

ISSN 0924-7815

Efficient estimation of autoregression parameters and innovation distributions for semiparametric integer-valued AR(p) models

Feike C. Drost

Econometrics and Finance group, CentER, Tilburg University, The Netherlands

Ramon van den Akker

Econometrics group, CentER, Tilburg University, The Netherlands

Bas J.M. Werker†

Econometrics and Finance group, CentER, Tilburg University, The Netherlands

Summary. Integer-valued autoregressive (INAR) processes have been introduced to model nonnegative integer-valued phenomena that evolve over time. The distribution of an INAR(p) process is essentially described by two parameters: a vector of autoregression coefficients and a probability distribution on the nonnegative integers, called an immigration or innovation distribution. Traditionally, parametric models are considered where the innovation distribution is assumed to belong to a parametric family. This paper instead considers a more realistic semiparametric INAR(p) model where there are essentially no restrictions on the innovation distribution. We provide an (semiparametrically) efficient estimator of both the autoregression parameters and the innovation distribution.

Keywords: count data, nonparametric maximum likelihood, infinite-dimensional Z-estimator, semiparametric efficiency

1. Introduction and notation

Al-Osh and Alzaid (1987) introduced the INAR(1) process to model nonnegative integer-valued phenomena that evolve in time. The INAR(1) process is defined by the recursion

$$X_t = \theta \circ X_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z}_+ = \mathbb{N} \cup \{0\}, \tag{1}$$

where

$$\theta \circ X_{t-1} = \sum_{j=1}^{X_{t-1}} Z_j^{(t)}.$$

Here an empty sum equals, by definition, 0. The variables $(Z_j^{(t)})_{j\in\mathbb{N},t\in\mathbb{Z}_+}$ are i.i.d. Bernoulli variables with success probability $\theta \in [0,1]$, independent of the i.i.d. innovation sequence $(\varepsilon_t)_{t\in\mathbb{Z}_+}$ with distribution G on $\mathbb{Z}_+$. The starting value X_{-1}, with distribution ν on $\mathbb{Z}_+$, is

†*Address for correspondence:* Bas J.M. Werker, Tilburg University, Econometrics and Finance group, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
E-mail: B.J.M.Werker@TilburgUniversity.nl

independent of $(\varepsilon_t)_{t\in\mathbb{Z}_+}$ and $(Z_j^{(t)})_{j\in\mathbb{N},t\in\mathbb{Z}_+}$. Display (1) can be interpreted as a branching process with immigration: X_t is composed of the surviving elements of X_{t-1} during the period $(t-1,t]$, $\theta\circ X_{t-1}$, and the number of immigrants during this period, ε_t. Each element of X_{t-1} survives with probability θ and its survival has no effect on the survival of the other elements, nor on the number of immigrants. In the literature on statistical inference for branching processes with immigration it is assumed that one observes both the X process and the ε process. We consider the empirically more common situation where the number of immigrants ε_t is not observed. Note, even if the true parameter θ would be known, the number of immigrants cannot be derived from the total X_t in the INAR(1) model.

The more general INAR(p) processes were first introduced by Al-Osh and Alzaid (1990) but Du and Li (1991) proposed a different setup. In the setup of Du and Li (1991) the autocorrelation structure of an INAR(p) process is the same as that of an AR(p) process, whereas it corresponds to the one of an ARMA($p,p-1$) process in the setup of Al-Osh and Alzaid (1990). The setup of Du and Li (1991) has been followed by most authors, and we use their setup as well. The INAR(p) process is an analogue of (1) with p lags. An INAR(p) process is recursively defined by,

$$X_t = \theta_1 \circ X_{t-1} + \theta_2 \circ X_{t-2} + \cdots + \theta_p \circ X_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z}_+, \tag{2}$$

where, for $i = 1,\ldots,p$,

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

Here $(Z_j^{(t,i)})_{j\in\mathbb{N},t\in\mathbb{Z}_+}$, $i \in \{1,\ldots,p\}$, are p mutually independent collections of i.i.d. Bernoulli variables with respective success probabilities $\theta_i \in [0,1]$, $i = 1,\ldots,p$, independent of the $\mathbb{Z}_+$-valued i.i.d. G-distributed innovations $(\varepsilon_t)_{t\in\mathbb{Z}_+}$. The starting value $(X_{-1},\ldots,X_{-p})^T$ is independent of $(\varepsilon_t)_{t\in\mathbb{Z}_+}$ and $(Z_j^{(t,i)})_{i\in\{1,\ldots,p\},j\in\mathbb{N},t\in\mathbb{Z}_+}$, and has distribution ν on $\mathbb{Z}_+^p$. The corresponding probability space is denoted by $(\Omega,\mathcal{F},\mathbb{P}_{\nu,\theta,G})$, where $\theta = (\theta_1,\ldots,\theta_p)^T$.

Applications of INAR processes in the medical sciences can be found in, for example, Franke and Seligmann (1993), Bélisle et al. (1998), and Cardinal et al. (1999); an application to psychometrics in Böckenholt (1999a), an application to environmentology in Thyregod et al. (1999); recent applications to economics in, for example, Böckenholt (1999b), Berglund and Brännäs (2001), Brännäs and Hellström (2001), Rudholm (2001), Böckenholt (2003), Brännäs and Shahiduzzaman (2004), Freeland and McCabe (2004), Gourieroux and Jasiak (2004), and McCabe and Martin (2005); and Pickands III and Stine (1997) and Ahn et al. (2000) considered queueing applications.

The statistical literature on INAR processes has concentrated on parametric models, i.e., G is assumed to belong to a parametric class of distributions, say $(G_\alpha|\alpha \in A \subset \mathbb{R}^q)$. For $p = 1$ and $G_\alpha = \text{Poisson}(\alpha)$ Franke and Seligmann (1993) analyzed maximum likelihood estimation. Du and Li (1991) and Freeland and McCabe (2005) derived the limit-distribution of the OLS-estimator of θ. Brännäs and Hellström (2001) considered GMM estimation, Silva and Oliveira (2005) proposed a frequency domain based estimator of θ, Silva and Silva (2006) considered a Yule-Walker estimator, Drost et al. (2006b) provided a computationally

attractive, asymptotically efficient estimator of (θ, α), and Jung et al. (2005) analyzed, by a Monte Carlo study, the finite sample behavior of several estimators for the INAR(1) case.

We consider a semiparametric model, where hardly any assumptions are made on G, and consider efficient estimation of (θ, G) from observations $X_{-p}, \ldots, X_n$. As far as we know, even inefficient estimation of G has not been addressed before. A possible explanation for this is that, even if $\theta_1, \ldots, \theta_p$ are known, observing $X_{t-p}, \ldots, X_t$ does not imply observing ε_t. Consequently, estimation of G cannot be based on residuals (as is the case for AR(p) processes). In fact, estimation of G can be viewed upon as a kind of deconvolution problem. However, estimation of the innovation distribution is, just as for standard AR models, an important topic. For INAR(p) processes this is even more important, since in some applications G has a clear physical interpretation. For example, Pickands III and Stine (1997) were interested in how often a physician prescribes a particular drug to new patients. The data are collected at the time of purchase, and so it is not possible to distinguish between new patient prescriptions and those of patients who have been using this medication. As a result, only the total prescriptions for a given drug for each doctor is observed. This can be modelled by an INAR(1) process, where the ε represent the number of new patients. In such examples, the parameter G is even the main parameter of interest.

Throughout the paper the number of lags, $p \in \mathbb{N}$, is fixed and known. To simplify the presentation, we gather all conditions needed for our results below in Assumption 1. Weaker conditions suffice for specific results, see for example Remark 2.2 concerning the consistency result of Theorem 2.1.

Assumption 1 Let $\tilde{\mathcal{G}}$ denote the set of all probability measures on $\mathbb{Z}_+$. We assume that $G = \mathcal{L}(\varepsilon_t) \in \mathcal{G} = \left\{ G \in \tilde{\mathcal{G}} : \ 0 < G(0) < 1; \ \mathbb{E}_G \varepsilon_t^{p+4} < \infty \right\}$. Furthermore we assume $\theta \in \Theta = \{\vartheta \in (0,1)^p : \ \sum_{i=1}^p \vartheta_i < 1\}$.

REMARK 1.1 The assumption $\theta \in (0,1)^p$ with $\sum_{i=1}^p \theta_i < 1$ is, see Lemma A.1, a sufficient condition for stationarity. For $p = 1$, inference in the nonstationary INAR(p) model, that is $\theta_1 \approx 1$, has been discussed in Ispány et al. (2003a, 2003b, 2005) and Drost et al. (2006a). The assumption $0 < G\{0\} < 1$ ensures the possibility of X becoming zero and not being always equal to 0, which is reasonable for virtually all applications. Finally, the $(p+4)$-th moment of G is needed in establishing weak convergence of certain empirical processes: the "size" of the class of functions involved increases with p, which explains the need for a more stringent condition for larger values of p.

The following notations are used. The Binomial distribution with parameters $\theta \in [0,1]$ and $n \in \mathbb{Z}_+$ is denoted by $\mathrm{Bin}_{n,\theta}$ ($\mathrm{Bin}_{0,\theta}$ is the Dirac-measure concentrated in 0) and $\mathrm{b}_{n,\theta}$ denotes the corresponding point mass function. For $G \in \mathcal{G}$, μ_G denotes the mean of G, σ_G^2 denotes its variance, and g its pdf. As usual, $\mathbb{E}_{\nu,\theta,G}(\cdot)$ is shorthand for $\int (\cdot)\, \mathrm{d}\mathbb{P}_{\nu,\theta,G}$. For (probability) measures F and G, $F * G$ denotes the convolution of F and G. Finally, $\mathbb{F} = (\mathcal{F}_t)_{t \geq -p}$ is the natural filtration generated by X, i.e. $\mathcal{F}_t = \sigma(X_{-p}, \ldots, X_t)$. Once more, note that in contrast to classical AR(p) processes, $\mathcal{F}_t \neq \sigma(X_{-p}, \ldots, X_{-1}, \varepsilon_0, \ldots, \varepsilon_t)$.

Before we discuss the contributions of this paper we recall some elementary properties of INAR processes, that will be used throughout. It immediately follows from (2) that, for

$t \in \mathbb{Z}_+$, the first two conditional moments are given by

$$\mathbb{E}_{\theta,G}\left[X_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{\theta,G}\left[X_t \mid X_{t-1}, \ldots, X_{t-p}\right] = \mu_G + \sum_{i=1}^{p} \theta_i X_{t-i} \in [0,\infty],$$

$$\mathrm{var}_{\theta,G}\left[X_t \mid \mathcal{F}_{t-1}\right] = \mathrm{var}_{\theta,G}\left[X_t \mid X_{t-1}, \ldots, X_{t-p}\right] = \sigma_G^2 + \sum_{i=1}^{p} \theta_i(1-\theta_i) X_{t-i} \in [0,\infty].$$

Hence an INAR(p) process has the same autoregression function as an AR(p) process. However, an INAR(p) process has conditional heteroskedasticity of autoregressive form, whereas the conditional variance is constant for AR(p) processes. Next we determine the conditional distribution of X_t given $\mathcal{F}_{t-1}$. From (2) it follows, for $t \in \mathbb{Z}_+$,

$$\mathbb{P}_{\theta,G}\left\{X_t = x_t \mid \mathcal{F}_{t-1}\right\} = \mathbb{P}_{\theta,G}\left\{X_t = x_t \mid X_{t-1}, \ldots, X_{t-p}\right\} = P^{\theta,G}_{(X_{t-1},\ldots,X_{t-p}),x_t},$$

where, for $x_{t-p}, \ldots, x_t \in \mathbb{Z}_+$, the transition-probability $P^{\theta,G}_{(x_{t-1},\ldots,x_{t-p}),x_t}$ is given by

$$\begin{aligned} P^{\theta,G}_{(x_{t-1},\ldots,x_{t-p}),x_t} &= \mathbb{P}_{\theta,G}\left\{\sum_{i=1}^{p} \theta_i \circ X_{t-i} + \varepsilon_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p}\right\} \\ &= \left(\mathrm{Bin}_{x_{t-1},\theta_1} * \cdots * \mathrm{Bin}_{x_{t-p},\theta_p} * G\right)\{x_t\}. \end{aligned}$$

Note that $X = (X_t)_{t \geq -p}$ is a p-th order Markov chain. Lemma A.1 gathers some auxiliary probabilistic results on the INAR(p) process. In particular, the lemma establishes, for $(\theta, G) \in \Theta \times \mathcal{G}$, the existence of a stationary solution $\nu_{\theta,G}$ and absolutely regular mixing with (at least) geometrically decreasing coefficients. Finally, Lemma A.1 proves a suitable Donsker type result for the empirical process of (X_t).

Next we discuss the contributions of our paper. Formally, we are interested in the experiments

$$\mathcal{E}^{(n)} = \left(\mathbb{Z}_+^{n+1+p}, 2^{\mathbb{Z}_+^{n+1+p}}, \left(\mathbb{P}^{(n)}_{\nu_{\theta,G},\theta,G} \mid \theta \in \Theta, G \in \mathcal{G}\right)\right), \quad n \in \mathbb{Z}_+,$$

where $\mathbb{P}^{(n)}_{\nu_{\theta,G},\theta,G}$ denotes the law of $(X_{-p}, \ldots, X_n)$, under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, on the measurable space $(\mathbb{Z}_+^{n+1+p}, 2^{\mathbb{Z}_+^{n+1+p}})$, with $\nu_{\theta,G}$ the stationary initial distribution (see Lemma A.1A).

Remark 1.2 Notice that the stationary distribution is taken as initial distribution. This enables us to use results from empirical processes theory for stationary time series. On the other hand, it complicates the semiparametric analysis: to obtain the LAN property we have to prove statistical negligibility of the initial value.

Compared to parametric models, the semiparametric model $\mathcal{E}^{(n)}$ is more general. However this comes at a cost: estimation in a semiparametric model is "at least as difficult" as in any parametric submodel. Although the OLS-estimator still yields an asymptotically normal estimator of θ in the semiparametric model (see Du and Li (1991)), it is not an efficient estimator of θ. This paper contributes a semiparametric efficient estimator of (θ, G). We stress that even inefficient estimation of G has not been considered before. Our estimator might be

viewed upon as a nonparametric maximum likelihood estimator (NPMLE). The monographs Bickel et al. (1998) and Van der Vaart (2000) (Chapter 25) are fairly complete accounts on the state of the art in semiparametric efficient estimation for i.i.d. models. Semiparametric efficiency considerations in time series originated by Kreiss (1987) for ARMA-type models, Drost et al. (1997) considered group models covering nonlinear location-scale time series, and Wefelmeyer (1996) considered models with general Markov type transitions. However, the semiparametric INAR(p) model cannot be analyzed by either of these approaches. This since it seems to be impossible to derive closed form formulas for the efficient influence operator. Nevertheless we are able to prove efficiency along the following lines. First we show that the NPMLE can be viewed upon as a solution to an infinite number of moment-conditions, i.e., as an infinite-dimensional Z-estimator. Following Van der Vaart (1995), who provides, for i.i.d. models, high-level conditions to prove efficiency of infinite-dimensional Z-estimators without having to calculate the efficient influence operator, we show that the NPMLE can be viewed upon as a Hadamard differentiable mapping of another estimator which is efficient for a certain artificial parameter. Since efficiency is retained under Hadamard differentiable maps (Van der Vaart (1991)) this can be exploited to obtain an efficiency proof. The main steps are proving Fréchet differentiability of the limiting estimating equation, and continuously invertibility of this derivative. These proofs are facilitated by "information-loss" representations of the transition-scores that were established by Drost et al. (2006b). Another important aspect is that the empirical estimating equation weakly converges, in an appropriate function space, to a Gaussian process. Since we are dealing with a Markovian structure, we rely on empirical processes for dependent data. Another crucial ingredient is that parametric submodels of the semiparametric model enjoy the local asymptotic normality (LAN) property.

The setup of the present paper is as follows. Section 2 introduces the NPMLE and discusses its consistency. In Section 3 we show that the NPMLE is a Z-estimator, i.e., it can be viewed upon as a solution to an infinite system of moment-conditions, and exploit this to derive the limiting distribution of the NPMLE. Section 4 proves that the NPMLE is efficient. Here we first show that parametric submodels have the LAN-property and that the NPMLE is regular. Next, the efficiency of the NPMLE follows from the regularity and the special representation of the limiting distribution. Finally, Section 5 discusses a small Monte Carlo simulation study and empirical application to analyze the finite sample behavior of the proposed estimator. Some auxiliary results are gathered in Appendix A. Some proofs have been organized, because of their length and technicality, separately in a Technical Appendix that is available online.

2. The estimator and consistency

In general, maximum likelihood is not directly applicable in semiparametric models. For the INAR model, due to the discreteness of G, nonparametric maximum likelihood estimation is feasible. We call an estimator $((\hat{\theta}_n, \hat{G}_n))_{n\in\mathbb{Z}_+}$ of (θ, G) a nonparametric maximum likelihood estimator (NPMLE) of (θ, G) if $(\hat{\theta}_n, \hat{G}_n)$ maximizes the *conditional* likelihood, i.e.,

$$\forall n \in \mathbb{Z}_+ : \quad (\hat{\theta}_n, \hat{G}_n) \in \operatorname*{argmax}_{(\theta,G)\in[0,1]^p\times\tilde{\mathcal{G}}} \prod_{t=0}^{n} P^{\theta,G}_{(X_{t-1},\ldots,X_{t-p}),X_t}. \tag{3}$$

REMARK 2.1 The conditional likelihood is used, since closed-form formulas for $\nu_{\theta,G}$ are only known for a few specific immigration distributions. Ignoring the information in the initial values has no consequences for the (asymptotic) efficiency of the NPMLE.

To guarantee the existence of a maximum likelihood estimator, we allow $(\hat{\theta}_n, \hat{G}_n)$ to take values outside $\Theta \times \mathcal{G}$. It is easy to see that $\hat{G}_n$ assigns all its mass to a subset of $\{u_-, \dots, u_+\}$, where

$$u_- = 0 \vee \min_{t=0,\dots,n} \left(X_t - \sum_{i=1}^{p} X_{t-i} \right), \text{ and } u_+ = \max_{t=0,\dots,n} X_t.$$

Now $(\hat{\theta}_n, \hat{G}_n)$ maximizes the likelihood if and only if the following holds: (i) $\hat{g}_n(k) = 0$ for $k < u_-$ and $k > u_+$, and (ii) $(\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,p}, \hat{g}_n(u_-), \dots, \hat{g}_n(u_+))$ is a solution to the (constrained) polynomial optimization problem:

$$\begin{aligned} \max_{\substack{x_1,\dots,x_p \\ z_{u_-},\dots,z_{u_+}}} \quad & \prod_{t=0}^{n} \sum_{e=0 \vee (X_t - \sum_{i=1}^{p} X_{t-i})}^{X_t} z_e \sum_{\substack{0 \le k_\ell \le X_{t-\ell},\, \ell=1,\dots,p \\ k_1+\cdots+k_p = X_t - e}} \prod_{\ell=1}^{p} \binom{X_{t-\ell}}{k_\ell} x_\ell^{k_\ell} (1-x_\ell)^{X_{t-\ell}-k_\ell} \\ \text{s.t.} \quad & 0 \le x_k \le 1 \text{ for } k = 1, \dots, p; \\ & z_j \ge 0 \text{ for } j = u_-, \dots, u_+; \\ & z_{u_-} + \cdots + z_{u_+} = 1. \end{aligned} \tag{4}$$

We stress that we nowhere (will) impose that such a maximum location is unique.

The next proposition, which follows by standard arguments, states that any maximum likelihood estimator is consistent. The proof is organized in the technical appendix.

Theorem 2.1 *For all $(\theta_0, G_0) \in \Theta \times \mathcal{G}$ and all initial probability measures ν_0 on $\mathbb{Z}_+^p$, any NPMLE $(\hat{\theta}_n, \hat{G}_n) = (\hat{\theta}_n, \hat{g}_n(0), \hat{g}_n(1), \dots)$, of (θ, G) is consistent in the following sense:*

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \text{ and } \sum_{k=0}^{\infty} |\hat{g}_n(k) - g_0(k)| \xrightarrow{p} 0, \text{ under } \mathbb{P}_{\nu_0,\theta_0,G_0}. \tag{5}$$

PROOF (OUTLINE): Let $(\hat{\theta}_n, \hat{G}_n)$ be a maximum likelihood estimator of (θ, G). To prove (5), it is well-known that it suffices to establish $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $\hat{g}_n(k) \xrightarrow{p} g_0(k)$ for all $k \in \mathbb{Z}_+$. We prove this by compactifying the parameter space and subsequently following the arguments of Wald's consistency theorem (see, for example, the proof of Theorem 5.14 in Van der Vaart (2000)). See the technical appendix for details. □

REMARK 2.2 Inspection of the proof of Theorem 2.1 shows that Assumption 1 is actually stronger than needed to establish (5). It suffices to assume that the innovation distribution G has finite mean μ_G and $G(0) = \mathbb{P}\{\varepsilon_t = 0\} < 1$.

3. Limit distribution

In this section we derive the limiting distribution of the NPMLE. First we show, in Section 3.1, that our NPMLE is actually a Z-estimator. To show this, we consider certain (artificial) submodels of the semiparametric model and exploit the fact that the NPMLE also maximizes the likelihood in these submodels. These submodels are such that the maximum is taken in a stationary point, which yields a score equation. Subsequently, this Z-estimator representation is used in Section 3.2 to derive the limiting distribution for $(\hat{\theta}_n, \hat{G}_n)$, which is represented as a transformation of a Gaussian process.

3.1. Likelihood equations

This section shows that $(\hat{\theta}_n, \hat{G}_n)$ can be viewed upon as an infinite-dimensional Z-estimator, i.e., $(\hat{\theta}_n, \hat{G}_n)$ solves an infinite number of moment conditions.

Fix the "truth" $(\theta_0, G_0) \in \Theta \times \mathcal{G}$. If $\hat{\theta}_n \in \Theta$ we obtain, since $(\hat{\theta}_n, \hat{G}_n)$ maximizes the likelihood and Θ is open,

$$\frac{1}{n}\sum_{t=0}^{n} \dot{\ell}_\theta(X_{t-p}, \ldots, X_t; \hat{\theta}_n, \hat{G}_n) = 0,$$

where, for $x_{t-p}, \ldots, x_t \in \mathbb{Z}_+$,

$$\dot{\ell}_\theta(x_{t-p}, \ldots, x_t; \theta, G) \quad = \quad \frac{\partial}{\partial \theta} \log\left(P^{\theta,G}_{(x_{t-1},\ldots,x_{t-p}),x_t}\right),$$

with the convention that $\dot{\ell}_\theta(x_{t-p}, \ldots, x_t; \theta, G) = 0$ if $P^{\theta,G}_{(x_{t-1},\ldots,x_{t-p})} = 0$. Drost et al. (2006b) derived, motivated by an "information-loss" interpretation of the model, that this θ-transition-score can be represented as,

$$\dot{\ell}_\theta(x_{t-p}, \ldots, x_t; \theta, G) = \begin{pmatrix} \mathbb{E}_{\theta,G}\left[\dot{s}_{X_{t-1},\theta_1}(\theta_1 \circ X_{t-1}) \mid X_t = x_t, \ldots, X_{t-p} = x_{t-p}\right] \\ \vdots \\ \mathbb{E}_{\theta,G}\left[\dot{s}_{X_{t-p},\theta_p}(\theta_p \circ X_{t-p}) \mid X_t = x_t, \ldots, X_{t-p} = x_{t-p}\right] \end{pmatrix},$$

where $\dot{s}_{n,\theta}(\cdot)$ is the score of a Binomial(n, θ) distribution, i.e.

$$\dot{s}_{n,\theta}(k) = \frac{k - n\theta}{\theta(1-\theta)}, \quad k \in \{0, \ldots, n\},\ n \in \mathbb{Z}_+.$$

This conditional expectation representation of the transition-score is heavily used later on.

Obtaining score-equations for the (infinite-dimensional) G-direction is more difficult. Construct (artificial) probability distributions on $\mathbb{Z}_+$, in direction $h : \mathbb{Z}_+ \to \mathbb{R}$ bounded, by

$$g_s(k) = g_s(k, h) = \left[1 + s\left(h(k) - \int h \,\mathrm{d}\hat{G}_n\right)\right] \hat{g}_n(k), \quad k \in \mathbb{Z}_+.$$

Note that $g_0 = \hat{g}_n$ and $G_s \in \tilde{\mathcal{G}}$ for all $|s| < (2\|h\|_\infty)^{-1}$. By construction $(\hat{\theta}_n, G_s)$ satisfies, for all s, the constraints of the optimization problem (4). Since $s = 0$ corresponds to the

NPMLE $(\hat{\theta}_n, \hat{G}_n)$, we obtain

$$0 = \frac{1}{n}\sum_{t=0}^{n} \frac{\partial}{\partial s} \log P^{\hat{\theta}_n, G_s}_{(X_{t-1},\ldots,X_{t-p}),X_t}\Big|_{s=0}.$$

To obtain a useful representation of this derivative, we note

$$\frac{\partial}{\partial s} \log P^{\theta, G_s}_{(x_{t-1},\ldots,x_{t-p}),x_t}\Big|_{s=0} = A_{\theta,\hat{G}_n} h(x_{t-p},\ldots,x_t) - \int h \,\mathrm{d}\hat{G}_n,$$

where

$$A_{\theta,G} h(x_{t-p},\ldots,x_t) = \mathbb{E}_{\theta,G}\left[h(\varepsilon_t) \mid X_t = x_t, \ldots, X_{t-p} = x_{t-p}\right], \quad x_{t-p},\ldots,x_t \in \mathbb{Z}_+.$$

Hence, we obtain a moment condition for every bounded function $h : \mathbb{Z}_+ \to \mathbb{R}$:

$$0 = \frac{1}{n}\sum_{t=0}^{n} \left(A_{\hat{\theta}_n,\hat{G}_n} h(X_{t-p},\ldots,X_t) - \int h \,\mathrm{d}\hat{G}_n \right).$$

Let $\mathcal{H}_1$ be the unit ball of $\ell^\infty(\mathbb{Z}_+)$, i.e., all functions $h : \mathbb{Z}_+ \to \mathbb{R}$ with $\sup_{e\in\mathbb{Z}_+} |h(e)| \le 1$. We will only use the moment conditions arising from $h \in \mathcal{H}_1$. We summarize these in an estimating equation $\Psi_n = (\Psi_{n1}, \Psi_{n2}) : (0,1)^p \times \tilde{\mathcal{G}} \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ defined by

$$\Psi_{n1}(\theta, G) = \frac{1}{n}\sum_{t=0}^{n} \dot{\ell}_\theta(X_{t-p},\ldots,X_t;\theta,G), \tag{6}$$

$$\Psi_{n2}(\theta, G)h = \frac{1}{n}\sum_{t=0}^{n} \left(A_{\theta,G} h(X_{t-p},\ldots,X_t) - \int h \,\mathrm{d}G \right), \quad h \in \mathcal{H}_1. \tag{7}$$

Note that $\Psi_{n2}(\theta, G)$ is indeed a random element of $\ell^\infty(\mathcal{H}_1)$, the set of bounded real-valued linear functionals on $\mathcal{H}_1$, since $\sup_{h\in\mathcal{H}_1} |\Psi_{n2}(\theta, G)h| \le 2$. From the discussion above we know that any NPMLE satisfies $\Psi_{n2}(\hat{\theta}_n, \hat{G}_n) = 0$, and $\mathbb{P}_{\nu_{\theta_0,G_0},\theta_0,G_0}\{\Psi_{n1}(\hat{\theta}_n, \hat{G}_n) = 0\} \to 1$ by the consistency result of Theorem 2.1.

For $(\theta_0, G_0) \in \Theta \times \mathcal{G}$ we introduce the "limit" of the estimating equation: $\Psi^{\theta_0,G_0} : (0,1)^p \times \mathcal{G} \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ by,

$$\Psi_1^{\theta_0,G_0}(\theta, G) = \mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0} \dot{\ell}_\theta(X_{-p},\ldots,X_0;\theta,G), \tag{8}$$

$$\Psi_2^{\theta_0,G_0}(\theta, G)h = \mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0} \left(A_{\theta,G} h(X_{-p},\ldots,X_0) - \int h \,\mathrm{d}G \right), \quad h \in \mathcal{H}_1. \tag{9}$$

It is easy to see that

$$\mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0} \Psi_1^{\theta_0,G_0}(\theta_0, G_0) = 0, \text{ and, for all } h \in \mathcal{H}_1, \ \mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0} \Psi_2^{\theta_0,G_0}(\theta_0, G_0)h = 0,$$

which is the usual result that, under the true probability measure, scores have expectation zero.

3.2. Asymptotic normality

In this section we exploit that the NPMLE can be seen as a solution to the estimating equation Ψ_n in (6)–(7) in order to derive its limiting distribution. Essentially we follow Huber's classical theorem on asymptotic normality of M-estimators. Compared to finite-dimensional parameters, we now have to deal with functional calculus instead of Euclidean calculus, and with empirical processes instead of weak convergence in Euclidean spaces.

First, we specify the chosen topology. Identify $G \in \mathcal{G}$ with its point mass function $\mathbb{Z}_+ \ni k \mapsto g(k) = G\{k\}$ and view the point mass functions as elements of the Banach space $\ell^1 = \ell^1(\mathbb{Z}_+)$, i.e. the space of real-valued sequences $(a_k)_{k\in\mathbb{Z}_+}$ for which $\|a\|_1 = \sum_{k\in\mathbb{Z}_+} |a_k| < \infty$. In the following, $\operatorname{lin}\mathcal{G}$ and its subsets are always regarded as subsets of $\ell^1(\mathbb{Z}_+)$. If no confusion can arise G will denote $G = (g(k))_{k\in\mathbb{Z}_+}$, and we write $\|G\|_1 = \|g\|_1$. Θ is equipped by the Euclidean topology, and we equip the product space $\mathbb{R}^p \times \ell^1(\mathbb{Z}_+)$ with the product topology, which can be metrized by the sum-norm $\|(\theta, G)\| = |\theta| + \|G\|_1$. Our parameter space, $\Theta \times \mathcal{G}$, is viewed upon as a subset of this Banach space $\mathbb{R}^p \times \ell^1(\mathbb{Z}_+)$. In this section we determine the limiting distribution of $\sqrt{n}((\hat{\theta}_n, \hat{G}_n) - (\theta, G))$, viewed upon as a random element in $\mathbb{R}^p \times \ell^1(\mathbb{Z}_+)$.

Lemma A.2 shows that the conditions to an infinite-dimensional version of Huber's theorem are satisfied. Part (L1) of Lemma A.2 shows that, for $(\theta_0, G_0) \in \Theta \times \mathcal{G}$, the limiting moment equations (8)–(9) are Fréchet-differentiable with derivative $\dot{\Psi}^0 = \dot{\Psi}^{\theta_0,G_0} : \operatorname{lin}([0,1]^p \times \mathcal{G}) \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ given by

$$\dot{\Psi}^0(\theta - \theta_0, G - G_0) = \left(\dot{\Psi}^0_{11}(\theta - \theta_0) + \dot{\Psi}^0_{12}(G - G_0), \dot{\Psi}^0_{21}(\theta - \theta_0) + \dot{\Psi}^0_{22}(G - G_0)\right), \quad (10)$$

where $\dot{\Psi}^0_{11} : \mathbb{R}^p \to \mathbb{R}^p$, $\dot{\Psi}^0_{12} : \operatorname{lin}\mathcal{G} \to \mathbb{R}^p$, $\dot{\Psi}^0_{21} : \mathbb{R}^p \to \ell^\infty(\mathcal{H}_1)$, and $\dot{\Psi}^0_{22} : \operatorname{lin}\mathcal{G} \to \ell^\infty(\mathcal{H}_1)$ are defined by

$$\dot{\Psi}^0_{11}(\theta - \theta_0) = -\left(\mathbb{E}_{\nu_0,\theta_0,G_0} \dot{\ell}_\theta \dot{\ell}_\theta^T(X_{-p}, \ldots, X_0; \theta_0, G_0)\right)(\theta - \theta_0), \quad (11)$$

$$\dot{\Psi}^0_{12}(G - G_0) = -\int \mathbb{E}_{\nu_0,\theta_0}\left[\dot{\ell}_\theta(X_{-p}, \ldots, X_0; \theta_0, G_0) \mid \varepsilon_0 = e\right] \mathrm{d}(G - G_0)(e), \quad (12)$$

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

$$\dot{\Psi}^0_{21}(\theta - \theta_0)h = -(\theta - \theta_0)^T \mathbb{E}_{\nu_0,\theta_0,G_0}\left[\dot{\ell}_\theta(X_{-p}, \ldots, X_0; \theta_0, G_0) A_{\theta_0,G_0} h(X_{-p}, \ldots, X_0)\right], \quad (13)$$

$$\dot{\Psi}^0_{22}(G - G_0)h = -\int \mathbb{E}_{\nu_0,\theta_0}\left[A_{\theta_0,G_0} h(X_{-p}, \ldots, X_0) \mid \varepsilon_0 = e\right] \mathrm{d}(G - G_0)(e), \quad (14)$$

with $\nu_0 = \nu_{\theta_0,G_0}$ and where we use the following version of conditional probabilities, for $G \in \tilde{\mathcal{G}}$ and $x_{-p}, \ldots, x_0, e \in \mathbb{Z}_+$,

$$\begin{aligned}\mathbb{P}_{\nu_0,\theta_0,G}\{X_{-p} = x_{-p}, \ldots, X_0 = x_0 \mid \varepsilon_0 = e\} &= \mathbb{P}_{\nu_0,\theta_0}\{X_{-p} = x_{-p}, \ldots, X_0 = x_0 \mid \varepsilon_0 = e\} \\ &= \nu_0\{(x_{-1}, \ldots, x_{-p})\}\left(\mathrm{Bin}_{x_{-p},\theta_p} * \cdots * \mathrm{Bin}_{x_{-1},\theta_1}\right)\{x_0 - e\}.\end{aligned}$$

Part (L2) of Lemma A.2 shows that the derivative $\dot{\Psi}^{\theta_0,G_0}$ is continuously invertible, which means that θ and g are locally identified. Subsequently, (L3) establishes weak convergence

of the moment conditions (6)–(7) to the limits (8)–(9), i.e.,

$$\mathbb{S}_n^{\theta_0,G_0} = \sqrt{n}(\Psi_n(\theta_0,G_0) - \Psi^{\theta_0,G_0}(\theta_0,G_0)) \rightsquigarrow \mathbb{S}^{\theta_0,G_0} \quad \text{in } \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1), \text{ under } \mathbb{P}_{\nu_0,\theta_0,G_0}, \tag{15}$$

where $\mathbb{S}^{\theta_0,G_0}$ is a tight, Borel measurable, Gaussian process. Finally (L4) of Lemma A.2 gives a convenient negligibility result. A combination of Theorem 2.1 and Lemma A.2 with an infinite-dimensional version of Huber's theorem yields the following theorem.

Theorem 3.1 *For $(\theta_0,G_0) \in \Theta \times \mathcal{G}$, any NPMLE $(\hat{\theta}_n,\hat{G}_n)$ satisfies*

$$\sqrt{n}\left((\hat{\theta}_n,\hat{G}_n) - (\theta_0,G_0)\right) = -\dot{\Psi}_{\theta_0,G_0}^{-1}\mathbb{S}_n^{\theta_0,G_0} + o(1;\mathbb{P}_{\nu_{\theta_0,G_0},\theta_0,G_0}) \rightsquigarrow -\dot{\Psi}_{\theta_0,G_0}^{-1}\mathbb{S}^{\theta_0,G_0}, \tag{16}$$

under $\mathbb{P}_{\nu_{\theta_0,G_0},\theta_0,G_0}$ in $\mathbb{R}^p \times \ell^1(\mathbb{Z}_+)$.

Proof: Theorem 2.1 and Lemma A.2 show that all conditions to Theorem 3.3.1 in Van der Vaart and Wellner (1993) are satisfied, which yields the result. □

4. Efficiency

In this section we prove efficiency of $(\hat{\theta}_n,\hat{G}_n)$. As mentioned in the introduction, our proof is nonstandard as it does not seem to be possible to obtain explicit expressions for the efficient influence operator. Fortunately, the special representation of the limiting distribution (Theorem 3.1) can be exploited to demonstrate efficiency. Basically, the argument is that the "score-process" $\mathbb{S}_n^{\theta,G}$ (see (15)) can be seen as an efficient estimator of a certain artificial parameter, and that efficiency is retained under Hadamard differentiable mappings.

4.1. Tangent space, regularity and the convolution theorem

It is well-known that the local structure of a model needs to be considered to obtain lower-bounds to the asymptotic precision of consistent estimators. Tangent spaces are the mathematical tool for this. The tangent set contains all scores that can be obtained from one-dimensional parametric submodels in the semiparametric model. Lemma A.3 shows that for the INAR(p) model, the tangent set, at $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, is given by

$$\mathcal{T}_{\theta,G}^0 = \left\{ a^T\dot{\ell}_\theta(X_{-p},\ldots,X_0;\theta,G) + A_{\theta,G}h(X_{-p},\ldots,X_0) - \int h\,\mathrm{d}G \;\middle|\; a \in \mathbb{R}^p,\, h \in \ell^\infty(\mathbb{Z}_+) \right\},$$

with $A_{\theta,G}$ as defined in Section 3.1. The tangent space is the $\mathrm{L}_2(\nu_{\theta,G} \otimes P^{\theta,G})$-closure of $\mathcal{T}_{\theta,G}^0$: $\mathcal{T}_{\theta,G} = \overline{\mathcal{T}_{\theta,G}^0}$.

Any result on (asymptotic) precision of estimators is restricted to a certain class of estimators. We follow the literature by considering "asymptotically unbiased" estimators. This concept, known under the name "regularity", is recalled first. An estimator T_n of (θ,G) is regular at $\mathbb{P}_{\nu_{\theta,G},\theta,G}$ if there exists a tight Borel measurable random element $\mathbb{L}$ in $\mathbb{R}^p \times \ell^1(\mathbb{Z}_+)$ such that for all $a \in \mathbb{R}^p$, $h \in \ell^\infty(\mathbb{Z}_+)$, we have,

$$\sqrt{n}\,(T_n - (\theta_n,G_n)) \rightsquigarrow \mathbb{L} \text{ under } \mathbb{P}_{\nu_n,\theta_n,G_n}, \tag{17}$$

where $\theta_n = \theta + a/\sqrt{n}$, $g_n = g(1 + (h - \int h \,\mathrm{d}G)/\sqrt{n})$, and $\nu_n = \nu_{\theta_n, G_n}$. An interpretation of (17) is that the limiting-distribution of T_n is not disturbed by vanishing perturbations in direction (a, h). An estimator T_n of (θ, G) is regular if it is regular at all $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, $(\theta, G) \in \Theta \times \mathcal{G}$.

Since Lemma A.3 establishes the LAN-property along parametric submodels of our semiparametric experiment $\mathcal{E}^{(n)}$, and it is straightforward to check pathwise differentiability, the following theorem is an immediate consequence of an infinite-dimensional analogue of the famous Hájek-Le Cam convolution theorem (see, for example, Bickel et al. (1998), Theorem 5.2.1, or Van der Vaart (1991), Theorem 2.1).

Theorem 4.1 *Let $(\theta, G) \in \Theta \times \mathcal{G}$ and let T_n be an estimator of (θ, G) which is regular at $\mathbb{P}_{\nu_{\theta,G},\theta,G}$. In particular,*

$$\mathcal{L}\left(\sqrt{n}\,(T_n - (\theta, G)) \mid \mathbb{P}_{\nu_{\theta,G},\theta,G}\right) \xrightarrow{w} \mathbb{Z} = \mathbb{Z}_{\theta,G,(T_n)_{n\in\mathbb{N}}}.$$

Then there exist independent random elements $\mathbb{L}_{\theta,G}$, which is a centered Gaussian process only depending on the model, and $\mathbb{N}_{\theta,G,(T_n)_{n\in\mathbb{N}}}$, which generally depends on both the model and the estimator, such that

$$\mathbb{Z}_{\theta,G,(T_n)_{n\in\mathbb{N}}} = \mathcal{L}(\mathbb{L}_{\theta,G} + \mathbb{N}_{\theta,G,(T_n)_{n\in\mathbb{N}}}).$$

So the scaled estimation error $\sqrt{n}(T_n - (\theta, G))$ can, in the limit, be represented by the convolution of the process $\mathbb{L}_{\theta,G}$ and $\mathbb{N}_{\theta,G,(T_n)_{n\in\mathbb{N}}}$. Since $\mathbb{L}_{\theta,G}$ only depends on the model and not on the estimator itself, it represents inevitable noise. Therefore an estimator is called efficient at $\mathbb{P}_{\nu_{\theta,G},\theta,G}$ if it is regular with limiting distribution $\mathbb{L}_{\theta,G}$. An estimator is efficient if it is efficient at all $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, $(\theta, G) \in \Theta \times \mathcal{G}$.

To claim efficiency of our NPMLE, we need this NPMLE to be a regular estimator itself. This is easily established.

Proposition 4.1 *Let $(\theta, G) \in \Theta \times \mathcal{G}$. Any NPMLE $((\hat{\theta}_n, \hat{G}_n))_{n\in\mathbb{Z}_+}$ is a regular estimator of (θ, G) at $\mathbb{P}_{\nu_{\theta,G},\theta,G}$.*

Proof: Using Le Cam's third lemma and Lemma A.3 it is easy to see (see also the proof of Theorem 2 in Van der Vaart (1995)) that $(\hat{\theta}_n, \hat{G}_n)$ is regular at $\mathbb{P}_{\nu_{\theta,G},\theta,G}$ if and only if the Fréchet derivative of the estimating equation, $\dot{\Psi}^{\theta,G}$ satisfies, for all $a \in \mathbb{R}^p$ and $h^* \in \ell^\infty(\mathbb{Z}_+)$ with $\mathbb{E}_G h^*(\varepsilon_1) = 0$,

$$\begin{aligned}\dot{\Psi}_1^{\theta,G}\,(a, (k \mapsto h^*(k)g(k))) \qquad\qquad (18)\\ = \quad -\mathbb{E}_{\nu_{\theta,G},\theta,G}\left(a^T \dot{\ell}_\theta(X_{-p}, \ldots, X_0; \theta, G)a + A_{\theta,G}h^*(X_{-p}, \ldots, X_0)\right)\dot{\ell}_\theta(X_{-p}, \ldots, X_0; \theta, G),\end{aligned}$$

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

$$\begin{aligned}\dot{\Psi}_2^{\theta,G}\,(a, (k \mapsto h^*(k)g(k)))\,h \qquad\qquad (19)\\ = \quad -\mathbb{E}_{\nu_{\theta,G},\theta,G}\left(a^T \dot{\ell}_\theta(X_{-p}, \ldots, X_0; \theta, G) + A_{\theta,G}h^*(X_{-p}, \ldots, X_0)\right)A_{\theta,G}h(X_{-p}, \ldots, X_0).\end{aligned}$$

Plugging in the definitions of $\dot{\Psi}_1^{\theta,G}$ and $\dot{\Psi}_2^{\theta,G}$, these displays are easily checked. □

REMARK 4.1 The displays (18)–(19) can be interpreted as the infinite-dimensional analogue of the information-matrix equality, i.e., the expectation of the outer-product of scores often equals minus the expectation of the Hessian of the log-likelihood.

4.2. Efficiency of the NPMLE

To prove efficiency we first recall the following characterization of efficiency. Fix $(\theta_0, G_0) \in \Theta \times \mathcal{G}$ and denote $\nu_0 = \nu_{\theta_0,G_0}$. Since $(\hat{\theta}_n, \hat{G}_n)$ is a regular estimator of (θ, G), we can conclude (see, for example, Bickel et al. (1998), Corollary 5.2.1) that $(\hat{\theta}_n, \hat{G}_n)$ is efficient at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, once we show that each component of $(\hat{\theta}_n, \hat{G}_n)$ is asymptotically linear at $\mathbb{P}_{\nu_0,\theta_0,G_0}$ with an influence function contained in the tangent space $\mathcal{T}_{\theta_0,G_0}$. More precise: there should exist $f_1, \ldots, f_p \in \mathcal{T}_{\theta_0,G_0}$ and h_k, $k \in \mathbb{Z}_+$ from $\mathcal{T}_{\theta_0,G_0}$ such that

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} \begin{pmatrix} f_1(X_{t-p}, \ldots, X_t) \\ \vdots \\ f_p(X_{t-p}, \ldots, X_t) \end{pmatrix} + o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}), \tag{20}$$

and for all $k \in \mathbb{Z}_+$,

$$\sqrt{n}(\hat{g}_n(k) - g(k)) = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} h_k(X_{t-p}, \ldots, X_t) + o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}). \tag{21}$$

Since we have no explicit formulas for $\dot{\Psi}^{-1}_{\theta_0,G_0}$ we cannot check directly whether this is the case. However, we will exploit the representation (see Theorem 3.1)

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_n - \theta_0 \\ (\hat{g}_n(k) - g_0(k))_{k \in \mathbb{Z}_+} \end{pmatrix} = -\dot{\Psi}^{-1}_{\theta_0,G_0} \mathbb{S}_n^{\theta_0,G_0} + o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}), \tag{22}$$

to demonstrate efficiency by an indirect argument. Recall that the finite-dimensional part of $\mathbb{S}_n^{\theta_0,G_0}$ is given by

$$\mathbb{S}_{n1}^{\theta_0,G_0} = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} \dot{\ell}_\theta(X_{t-p}, \ldots, X_t; \theta_0, G_0), \tag{23}$$

and the infinite-dimensional part by,

$$\mathbb{S}_{n2}^{\theta_0,G_0} h = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} \left(A_{\theta_0,G_0} h(X_{t-p}, \ldots, X_t) - \int h \, \mathrm{d}G_0 \right), \quad h \in \mathcal{H}_1. \tag{24}$$

So $\mathbb{S}_n^{\theta_0,G_0}$ is a process of certain elements of the tangent space. To prove efficiency we show that $\mathbb{S}_n^{\theta_0,G_0}$ can, at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, be seen as an efficient estimator of a certain artificial parameter and then exploit the representation (22) and that efficiency is retained under smooth mappings.

Theorem 4.2 *Any NPMLE $((\hat{\theta}_n, \hat{G}_n))_{n \in \mathbb{Z}_+}$ is an efficient estimator of (θ, G) within the experiments $\mathcal{E}^{(n)}$, $n \in \mathbb{Z}_+$. So we have (see Theorem 4.1), for all $(\theta_0, G_0) \in \Theta \times \mathcal{G}$,*

$$\mathcal{L}(\mathbb{L}_{\theta_0,G_0}) = \mathcal{L}(-\dot{\Psi}^{-1}_{\theta_0,G_0} \mathbb{S}^{\theta_0,G_0}).$$

PROOF: Introduce the *artificial* parameters (notice that we use ν_0 instead of $\nu_{\theta,G}$)

$$\Theta \times \mathcal{G} \ni (\theta, g) \mapsto \nu_1^{\theta_0,G_0}(\theta, g) = \mathbb{E}_{\nu_0,\theta,G} \dot{\ell}_\theta(X_{-p}, \ldots, X_0; \theta_0, G_0),$$

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

$$\Theta \times \mathcal{G} \ni (\theta, g) \mapsto \nu_h^{\theta_0,G_0}(\theta, g) = \mathbb{E}_{\nu_0,\theta,G} A_{\theta_0,G_0} h(X_{-p}, \ldots, X_0) - \int h \, \mathrm{d}G_0.$$

Hence $\nu_1^{\theta_0,G_0}(\theta_0, g_0) = \nu_h^{\theta_0,G_0}(\theta_0, g_0) = 0$. From (23) we conclude that, *at* $\mathbb{P}_{\nu_0,\theta_0,G_0}$, $\mathbb{S}_{n1}^{\theta_0,G_0}$ is an asymptotically linear estimator of $\nu_1^{\theta_0,G_0}(\theta, g)$ with influence function contained in $\mathcal{T}_{\theta_0,G_0}$. Similarly, from (24) we obtain that, *at* $\mathbb{P}_{\nu_0,\theta_0,G_0}$ and for $h \in \mathcal{H}_1$, $\mathbb{S}_{n2}^{\theta_0,G_0} h$ is an asymptotically linear estimator of $\nu_h^{\theta_0,G_0}(\theta, g)$, with influence function contained in $\mathcal{T}_{\theta_0,G_0}$. Consequently, these estimators are efficient *at* $\mathbb{P}_{\nu_0,\theta_0,G_0}$ once we show that they are regular at $\mathbb{P}_{\nu_0,\theta_0,G_0}$. Using Le Cam's third lemma and Lemma A.3 this regularity follows once we show that, for all $a \in \mathbb{R}^p$ and $f \in \ell^\infty(\mathbb{Z}_+)$ with $\mathbb{E}_{G_0} f(\varepsilon_1) = 0$,

$$\lim_{t \to 0} \frac{\nu_1^{\theta_0,G_0}(\theta + ta, g_0(1 + t(f - \int f \, \mathrm{d}G_0))) - \nu_1^{\theta_0,G_0}(\theta_0, g_0)}{t} =$$
$$\mathbb{E}_{\nu_0,\theta_0,G_0} \left(a^T \dot{\ell}_\theta(X_{-p}, \ldots, X_0; \theta_0, G_0) + A_{\theta_0,G_0} f(X_{-p}, \ldots, X_0) \right) \dot{\ell}_\theta(X_{-p}, \ldots, X_0),$$

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

$$\lim_{t \to 0} \frac{\nu_h^{\theta_0,G_0}(\theta + ta, g_0(1 + t(f - \int f \, \mathrm{d}G_0))) - \nu_h^{\theta_0,G_0}(\theta_0, g_0)}{t} =$$
$$\mathbb{E}_{\nu_0,\theta_0,G_0} \left(a^T \dot{\ell}_\theta(Z_0; \theta_0, G_0) + A_{\theta_0,G_0} f(Z_0) \right) \left(A_{\theta_0,G_0} h(Z_0) - \int h \, \mathrm{d}G_0 \right).$$

These relations are quite straightforward to check (see also the proof of Lemma A.3). Hence we conclude that, at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, $\mathbb{S}_{n1}^{\theta_0,G_0}$ is an efficient estimator of the parameter $(\theta, g) \mapsto \nu_1^{\theta_0,G_0}(\theta, g)$, and, for $h \in \mathcal{H}_1$, $\mathbb{S}_{n2}^{\theta_0,G_0} h$ is, at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, an efficient estimator of the parameter $(\theta, g) \mapsto \nu_h^{\theta_0,G_0}(\theta, g)$. Since we already established tightness of $\mathbb{S}_n^{\theta_0,G_0}$ (see Lemma A.2L3), and since marginal efficiency plus tightness is equivalent to efficiency, we conclude that $\mathbb{S}_n^{\theta_0,G_0}$ is, at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, an efficient estimator of the parameter $(\theta, g) \mapsto (\nu_1^{\theta_0,G_0}(\theta, g), (\nu_h^{\theta_0,G_0}(\theta, g))_{h \in \mathcal{H}_1})$. From (22) we obtain that, at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, $\sqrt{n}(\hat{\theta}_n - \theta_0, (\hat{g}_n(k) - g_0(k))_{k \in \mathbb{Z}_+})$ is a continuous, linear transformation of the efficient estimator $\mathbb{S}_n^{\theta_0,G_0}$. Since efficiency is retained under Hadamard differentiable mappings we conclude that, at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, $\sqrt{n}(\hat{\theta}_n - \theta_0, (\hat{g}_n(k) - g_0(k))_{k \in \mathbb{Z}_+})$ is an efficient estimator of a *certain* parameter (for details we refer to the proof of Theorem 3 in Van der Vaart (1995)). Hence, still at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, the influence functions of the components of $\sqrt{n}(\hat{\theta}_n - \theta_0, (\hat{g}_n(k) - g_0(k))_{k \in \mathbb{Z}_+})$ are contained in the tangent space $\mathcal{T}_{\theta_0,G_0}$, which yields (20) and (21). Since we already proved regularity this proves efficiency of the NPMLE at $\mathbb{P}_{\nu_0,\theta_0,G_0}$. □

5. Simulation experiment and empirical example

To enhance the interpretation and to investigate the validity of our theoretical results a small Monte Carlo study and empirical application is presented.

In the Monte Carlo study the finite sample behavior of the NPMLE is investigated. All simulations were carried out in `Matlab 6.5` and the NPMLE is computed using the optimization routine `fmincon`. As starting values for the optimization routine we use the OLS-estimator for θ and as starting value for G we use the uniform distribution on $\{0, \ldots, \max_{t=1,\ldots,n} X_t\}$. Due to the form of the likelihood the computational effort in the simulations is substantial. Therefore, the number of replications is limited to 2500, we only consider $p = 1$, and we only consider relatively small values of $\mu_G/(1-\theta_1)$. Four innovation distributions G are considered. Two of these choices are inspired by the estimates in the empirical application (see Table 3): Poisson(0.5) and Geometric($\exp(-0.5)$). We also consider the Poisson(1) and the Geometric($\exp(-1)$) distribution as innovation distributions. For each choice of the innovation distribution we consider three θ-values and two sample sizes: $\theta = 0.25$, 0.5, 0.75, and $n = 500$, 2000. Notice that the Poisson(μ) distribution assigns the same mass to 0 as the Geometric($\exp(-\mu)$) distribution, which explains the choice of parameters for the Geometric distributions. For the Poisson distribution it is well-known, and easy to check, that $\nu_{\theta,G} = \text{Poisson}(\mu_G/(1-\theta))$. Hence for Poisson innovations we use "exact" simulations for the initial value. For the Geometric innovation structures we let the chain start in the stationary mean (rounded to obtain an integer) and let it "run" for 250 periods. As first observation in our studies we use the value of the process at time 251.

Table 1 presents the results for $n = 500$, and Table 2 presents the results for $n = 2000$. To conserve space we only report the results for $\hat{g}_n(k)$ for $k = 0, \ldots, 5$. Comparing the entries in Table 1 with the corresponding entries in Table 2, we confirm the theoretical results developed before. First, even for the smaller sample, the NPMLE for θ is always more precise than the OLS estimator. The efficiency gain seems to be increasing in θ and runs up to 400%. This corroborates the result of Drost et al. (2006a) that shows that near unity the least-squares estimator does not even attain the optimal rate of convergence. Since estimation of G has not been considered before in the literature, the behavior of $\hat{g}_n$ is perhaps more interesting. We see that also for the smaller sample the probability estimates are unbiased. It appears that the standard errors of $\hat{g}_n$ tend to increase with θ. A possible explanation for this is the following. If the INAR(1) process drives to state 0, the next observation yields a direct observation on ε. The NPMLE exploits both these direct observations as well as the other observations for which we observe a (true) convolution of ε_t with $\theta_1 \circ X_{t-1}$. Asymptotically, we have $n\nu_{\theta,G}\{0\}$ direct observations on ε. Since $\nu_{\theta,G}\{0\}$ decreases as θ increases, we obtain less direct observations on ε as θ increases. So we have to deconvolute even more observations, which yields increasing standard errors. Comparing the Geometric distributions with their Poisson counterpart it seems that estimation of (θ, G) for Poisson innovations is more difficult than for Geometric innovations. Furthermore, the efficiency gain of $\hat{\theta}_n$ with respect to the OLS-estimator of θ is less large for Poisson innovations.

To demonstrate that the NPMLE is applicable in practice, we conclude this section with a simple empirical example based on ultra-high frequency data. We consider the IBM stock traded at the NYSE. We use quote data from the TAQ dataset for February 2005. In this month there were 19 trading days (on Monday February 21 the NYSE was closed because of Washington's Birthday). We remove all quotes that took place outside the opening hours; i.e. before 9.30 AM and after 4.00 PM. The variable of interest is the number of quotes per second, where we start the measurement at the first quote of the day and end at the last quote of the day. For the trading days in February 2005, the maximum number of quotes

Table 1. Simulation results for $n = 500$ (based on 2500 replications)

Parameter	Value	Estimator	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
			$\theta = 0.25$		$\theta = 0.5$		$\theta = 0.75$	
			$G = \text{Geometric}(\exp(-0.5))$					
		θ_n^{OLS}	0.2457	0.0482	0.4934	0.0441	0.7436	0.0317
		$\hat{\theta}_n$	0.2463	0.0391	0.4970	0.0315	0.7489	0.0178
$g(0)$	0.6065	$\hat{g}_n(0)$	0.6041	0.0290	0.6047	0.0311	0.6046	0.0339
$g(1)$	0.2387	$\hat{g}_n(1)$	0.2405	0.0259	0.2395	0.0291	0.2402	0.0336
$g(2)$	0.0939	$\hat{g}_n(2)$	0.0943	0.0165	0.0946	0.0187	0.0942	0.0209
$g(3)$	0.0369	$\hat{g}_n(3)$	0.0369	0.0105	0.0372	0.0117	0.0370	0.0132
$g(4)$	0.0145	$\hat{g}_n(4)$	0.0148	0.0068	0.0147	0.0078	0.0145	0.0084
$g(5)$	0.0057	$\hat{g}_n(5)$	0.0056	0.0043	0.0056	0.0049	0.0059	0.0051
			$G = \text{Poisson}(0.5)$					
		θ_n^{OLS}	0.2474	0.0494	0.4944	0.0447	0.7436	0.0335
		$\hat{\theta}_n$	0.2478	0.0470	0.4964	0.0364	0.7484	0.0210
$g(0)$	0.6065	$\hat{g}_n(0)$	0.6061	0.0297	0.6048	0.0318	0.6036	0.0347
$g(1)$	0.3033	$\hat{g}_n(1)$	0.3035	0.0276	0.3048	0.0304	0.3056	0.0342
$g(2)$	0.0758	$\hat{g}_n(2)$	0.0759	0.0149	0.0759	0.0161	0.0765	0.0167
$g(3)$	0.0126	$\hat{g}_n(3)$	0.0127	0.0062	0.0126	0.0064	0.0126	0.0069
$g(4)$	0.0016	$\hat{g}_n(4)$	0.0015	0.0022	0.0016	0.0024	0.0016	0.0024
$g(5)$	0.0002	$\hat{g}_n(5)$	0.0002	0.0007	0.0002	0.0008	0.0001	0.0006
			$G = \text{Geometric}(\exp(-1))$					
		θ_n^{OLS}	0.2475	0.0461	0.4960	0.0411	0.7419	0.0308
		$\hat{\theta}_n$	0.2466	0.0342	0.4971	0.0288	0.7478	0.0189
$g(0)$	0.3679	$\hat{g}_n(0)$	0.3660	0.0363	0.3643	0.0462	0.3598	0.0739
$g(1)$	0.2325	$\hat{g}_n(1)$	0.2327	0.0321	0.2347	0.0463	0.2377	0.0861
$g(2)$	0.1470	$\hat{g}_n(2)$	0.1478	0.0252	0.1474	0.0379	0.1478	0.0670
$g(3)$	0.0929	$\hat{g}_n(3)$	0.0927	0.0204	0.0929	0.0314	0.0934	0.0531
$g(4)$	0.0587	$\hat{g}_n(4)$	0.0588	0.0165	0.0590	0.0259	0.0591	0.0389
$g(5)$	0.0371	$\hat{g}_n(5)$	0.0378	0.0133	0.0371	0.0209	0.0376	0.0282
			$G = \text{Poisson}(1)$					
		θ_n^{OLS}	0.2460	0.0466	0.4947	0.0419	0.7427	0.0430
		$\hat{\theta}_n$	0.2443	0.0463	0.4956	0.0372	0.7450	0.0381
$g(0)$	0.3679	$\hat{g}_n(0)$	0.3657	0.0352	0.3626	0.0461	0.3586	0.0671
$g(1)$	0.3679	$\hat{g}_n(1)$	0.3676	0.0313	0.3709	0.0440	0.3740	0.0653
$g(2)$	0.1839	$\hat{g}_n(2)$	0.1851	0.0275	0.1849	0.0352	0.1841	0.0406
$g(3)$	0.0613	$\hat{g}_n(3)$	0.0624	0.0166	0.0625	0.0210	0.0619	0.0242
$g(4)$	0.0153	$\hat{g}_n(4)$	0.0153	0.0087	0.0154	0.0104	0.0161	0.0110
$g(5)$	0.0031	$\hat{g}_n(5)$	0.0030	0.0037	0.0030	0.0042	0.0031	0.0042

Table 2. Simulation results for $n = 2000$ (based on 2500 replications)

Parameter	Value	Estimator	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
			$\theta = 0.25$		$\theta = 0.5$		$\theta = 0.75$	
			$G = \text{Geometric}(\exp(-0.5))$					
		θ_n^{OLS}	0.2488	0.0247	0.4989	0.0228	0.7489	0.0164
		$\hat{\theta}_n$	0.2488	0.0194	0.4998	0.0157	0.7499	0.0088
$g(0)$	0.6065	$\hat{g}_n(0)$	0.6059	0.0145	0.6062	0.0160	0.6066	0.0165
$g(1)$	0.2387	$\hat{g}_n(1)$	0.2392	0.0129	0.2386	0.0148	0.2387	0.0165
$g(2)$	0.0939	$\hat{g}_n(2)$	0.0939	0.0084	0.0943	0.0097	0.0940	0.0102
$g(3)$	0.0369	$\hat{g}_n(3)$	0.0370	0.0052	0.0370	0.0059	0.0367	0.0067
$g(4)$	0.0145	$\hat{g}_n(4)$	0.0146	0.0033	0.0145	0.0038	0.0146	0.0042
$g(5)$	0.0057	$\hat{g}_n(5)$	0.0058	0.0021	0.0058	0.0024	0.0057	0.0026
			$G = \text{Poisson}(0.5)$					
		θ_n^{OLS}	0.2494	0.0245	0.4991	0.0248	0.7486	0.0222
		$\hat{\theta}_n$	0.2497	0.0231	0.4991	0.0206	0.7497	0.0180
$g(0)$	0.6065	$\hat{g}_n(0)$	0.6066	0.0148	0.6059	0.0199	0.6063	0.0205
$g(1)$	0.3033	$\hat{g}_n(1)$	0.3033	0.0135	0.3037	0.0162	0.3030	0.0177
$g(2)$	0.0758	$\hat{g}_n(2)$	0.0756	0.0074	0.0757	0.0082	0.0759	0.0085
$g(3)$	0.0126	$\hat{g}_n(3)$	0.0127	0.0031	0.0126	0.0033	0.0126	0.0033
$g(4)$	0.0016	$\hat{g}_n(4)$	0.0015	0.0011	0.0015	0.0012	0.0016	0.0012
$g(5)$	0.0002	$\hat{g}_n(5)$	0.0002	0.0003	0.0002	0.0004	0.0001	0.0003
			$G = \text{Geometric}(\exp(-1))$					
		θ_n^{OLS}	0.2493	0.0232	0.4990	0.0211	0.7484	0.0158
		$\hat{\theta}_n$	0.2490	0.0165	0.4995	0.0140	0.7494	0.0087
$g(0)$	0.3679	$\hat{g}_n(0)$	0.3678	0.0178	0.3672	0.0234	0.3655	0.0334
$g(1)$	0.2325	$\hat{g}_n(1)$	0.2327	0.0156	0.2333	0.0233	0.2341	0.0388
$g(2)$	0.1470	$\hat{g}_n(2)$	0.1470	0.0127	0.1467	0.0185	0.1474	0.0307
$g(3)$	0.0929	$\hat{g}_n(3)$	0.0925	0.0102	0.0930	0.0154	0.0933	0.0255
$g(4)$	0.0587	$\hat{g}_n(4)$	0.0594	0.0083	0.0588	0.0126	0.0587	0.0203
$g(5)$	0.0371	$\hat{g}_n(5)$	0.0369	0.0064	0.0370	0.0101	0.0371	0.0163
			$G = \text{Poisson}(1)$					
		θ_n^{OLS}	0.2492	0.0238	0.4972	0.0287	0.7486	0.0157
		$\hat{\theta}_n$	0.2490	0.0228	0.4977	0.0268	0.7491	0.0109
$g(0)$	0.3679	$\hat{g}_n(0)$	0.3676	0.0180	0.3663	0.0269	0.3661	0.0292
$g(1)$	0.3679	$\hat{g}_n(1)$	0.3675	0.0155	0.3678	0.0263	0.3688	0.0296
$g(2)$	0.1839	$\hat{g}_n(2)$	0.1844	0.0137	0.1838	0.0184	0.1844	0.0191
$g(3)$	0.0613	$\hat{g}_n(3)$	0.0616	0.0084	0.0613	0.0103	0.0615	0.0111
$g(4)$	0.0153	$\hat{g}_n(4)$	0.0153	0.0042	0.0156	0.0051	0.0155	0.0053
$g(5)$	0.0031	$\hat{g}_n(5)$	0.0030	0.0020	0.0030	0.0022	0.0031	0.0023

Table 3. Estimation results IBM

	Avg. Estimate	Std. Error
θ_n^{OLS}	0.2552	0.0159
$\hat{\theta}_n$	0.2307	0.0116
$\hat{g}_n(0)$	0.6385	0.0260
$\hat{g}_n(1)$	0.2440	0.0129
$\hat{g}_n(2)$	0.0844	0.0099
$\hat{g}_n(3)$	0.0239	0.0043
$\hat{g}_n(4)$	0.0066	0.0014
$\hat{g}_n(5)$	0.0018	0.0006

per second was on average 9.8, and the average number of quotes per second during the trading days was 0.68. For each trading day we estimate an INAR(1) model. In Table 3 we present the average of the parameter estimates and the standard errors of these estimates. To conserve space we only report the results for $\hat{g}_n(k)$ for $k = 0, \ldots, 5$. From the standard errors we see that the estimates for the different days are quite close. So, at least for February 2005, there seems to be some common structure in the arrival of quotes. The OLS estimates and the NPMLE estimates of θ are not too far away from each other, so this provides "no evidence" against the model. We have the following estimated autoregression $\mathbb{E}[X_t \mid X_{t-1}] = \theta X_{t-1} + \mu_G \approx 0.23X_{t-1} + 0.52$, and the following estimated conditional variance $\mathrm{var}[X_t \mid X_{t-1}] = \theta(1-\theta)+\sigma_G^2 \approx 0.18X_{t-1}+0.70$. Interpreting the INAR(1) model as a branching process with immigration, we can "decompose" the number of quotes per second into two parts. The first part, consists of quotes which are "offspring" of quotes in the previous second, and so models the predictable part. The estimated value for θ, which is about 0.23, means that a quote arriving at time t "generates" a new quote at period $t+1$ with probability 0.23. The estimates $\hat{g}_n(k)$ give the probability on k "new unpredictable" quotes. These estimates θ are confirmed by the autoregression results above as $\theta \approx 0.23$ and $\theta(1-\theta) \approx 0.18$. On the other hand, a Poisson innovation distribution seems to be rejected as $\hat{\mu}_G \neq \hat{\sigma}_G^2$. Also the estimated probabilities $\hat{g}(0), \hat{g}(1), \ldots$ do not follow a Poisson distribution. A geometric distribution possibly copes better with the IBM data. Of course, we do not advocate to impose a priori a Geometric distribution: G is a nuisance parameter and should be treated as such. Our proposed NPMLE is then optimal both for θ, in particular improving over OLS, and for G.

A. Some auxiliary results

To exploit the p-th order Markovian structure of the INAR(p) process (2), we introduce $Z = (Z_t)_{t\in\mathbb{Z}_+}$ defined by $Z_t = (X_t, X_{t-1}, \ldots, X_{t-p})^T$. Under $\mathbb{P}_{\nu,\theta,G}$ the process Z is a first-order Markov chain in $\mathbb{Z}_+^{p+1}$. It is easy to see that, in case $\theta \in \Theta$ and $G \in \mathcal{G}$, Z is irreducible and aperiodic. For notational convenience we also introduce $Y = (Y_t)_{t\in\mathbb{Z}_+}$ defined by $Y_t = (X_{t-1}, \ldots, X_{t-p})^T$.

The next lemma contains some auxiliary results. We briefly indicate their use in this paper. Part (A) establishes the existence of a stationary distribution for the INAR(p) process. Part (B) shows that the β-mixing numbers of Z are geometrically decreasing. This together with results from Doukhan et al. (1995) yields part (C). We will use Part (C) to demonstrate weak convergence of the infinite-dimensional part of the "score-process". Since it is

only in very special cases possible to derive closed form formulas for $\nu_{\theta,G}$, it is nontrivial to verify "negligibility of the initial value", which we need to prove that parametric submodels of the semiparametric model enjoy the LAN-property. Part (D) will allow us to prove this negligibility. The proof of the lemma is organized in the technical appendix.

Lemma A.1 *Let $\theta \in \Theta$ and $G \in \mathcal{G}$. The following results hold.*

(A) There exists a probability measure $\nu_{\theta,G}$ on $\mathbb{Z}_+^p$ such that X is a (strictly) stationary process under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$. Furthermore, the first moment of X_t, under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, is finite. If, for $k \in \mathbb{N}$, $\mathbb{E}_G \varepsilon_1^k < \infty$, then $\mathbb{E}_{\nu_{\theta,G},\theta,G} X_t^k < \infty$.

(B) Under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$ the β-mixing (also called: absolute regularity mixing) coefficients (for the definition see, for example, Doukhan (1994) page 3 and pages 87-88) of Z satisfy

$$\beta(n) \leq C\rho^n, \quad \text{for all } n \in \mathbb{N},$$

for some constant $C > 0$ and $0 < \rho < 1$.

(C) Let $\mathbb{Z}_n$ denote the empirical process of Z, i.e.

$$\mathbb{Z}_n f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} \left(f(Z_t) - \mathbb{E}_{\nu_{\theta,G},\theta,G} f(Z_0) \right), \quad \text{for } f : \mathbb{Z}_+^{p+1} \to \mathbb{R} : \ \mathbb{E}_{\nu_{\theta,G},\theta,G} f^2(Z_0) < \infty.$$

Let $\mathcal{F}$ be a collection of real-valued functions on $\mathbb{Z}_+^{p+1}$ with $\sup_{f \in \mathcal{F}} |f(x_{-p}, \ldots, x_0)| \leq C(1 + x_{-p} + \cdots + x_0)$ for some $C > 0$, and such that its bracketing numbers with respect to the $\mathrm{L}_2(\nu_{\theta,G} \otimes P^{\theta,G})$-norm, denoted by $N_{[]}(\delta, \mathcal{F})$, $\delta > 0$, satisfy

$$\int_0^1 x^{-a} \left(\log N_{[]}(x, \mathcal{F}) \right)^{1/2} \mathrm{d}x < \infty,$$

for some $a > 0$. Then the process $\{\mathbb{Z}_n f \mid f \in \mathcal{F}\}$ weakly converges, under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, in $\ell^\infty(\mathcal{F})$ to a tight Gaussian process.

(D) Define $V : \mathbb{Z}_+^p \to [1, \infty)$ by $V(x_{-1}, \ldots, x_{-p}) = 1 + \sum_{i=1}^p a_i x_{-i}$, where $a_i = \theta_i + \cdots + \theta_p$ for $i = 1, \ldots, p$. Let (θ_n, G_n) be a sequence in $\Theta \times \mathcal{G}$. Write δ_y for the Dirac measure on $y \in \mathbb{R}^p$. Then

$$\lim_{n \to \infty} \sup_{y \in \mathbb{Z}_+^p} \frac{\sup_{f : |f| \leq V} \left| \mathbb{E}_{\delta_y,\theta_n,G_n} f(Y_1) - \mathbb{E}_{\delta_y,\theta,G} f(Y_1) \right|}{V(y)} = 0$$

implies

$$\lim_{n \to \infty} \sup_{f : |f| \leq V} \left| \int f \, \mathrm{d}\nu_{\theta,G} - \int f \, \mathrm{d}\nu_{\theta_n,G_n} \right| = 0.$$

REMARK A.1 Stationarity in Part (A) can be established without the condition that $g(0) > 0$. In that case, the support of the stationary distribution $\nu_{\theta,G}$ is given by $\{\alpha, \alpha + 1, \ldots\}^p$, where $\alpha = \min\{k \in \mathbb{Z}_+ \mid g(k) > 0\}$. Furthermore, only the first moment of G needs to be finite.

REMARK A.2 Let us recall the definition of the bracketing numbers used in (C). A bracket is a pair of elements $[f, g]$ of $\mathcal{L}_2(\nu_{\theta,G} \otimes P^{\theta,G})$ such that $f \leq g$. For $\delta > 0$ the bracketing number $N_{[]}(\delta, \mathcal{F})$ is the smallest cardinality of collections $\mathcal{S}(\delta)$ of brackets such that for all $f \in \mathcal{F}$ there exists $[g, h] \in \mathcal{S}(\delta)$ such that $g \leq f \leq h$ and $\int (h-g)^2 \, \mathrm{d}(\nu_{\theta,G} \otimes P^{\theta,G}) \leq \delta^2$.

The following lemma is key to the derivation of the limiting distribution of the NPMLE in Section 3.2. As its proof is fairly technical and involved, we provide it in the Technical Appendix.

Lemma A.2 *Let $(\theta_0, G_0) \in \Theta \times \mathcal{G}$. Denote $\nu_0 = \nu_{\theta_0,G_0}$. Then the following properties hold.*

(L1) The map $\Psi^{\theta_0,G_0} : (0,1)^p \times \mathcal{G} \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$ is Fréchet-differentiable at (θ_0, G_0), i.e.

$$\|\Psi^{\theta_0,G_0}(\theta, G) - \Psi^{\theta_0,G_0}(\theta_0, G_0) - \dot{\Psi}^{\theta_0,G_0}(\theta - \theta_0, G - G_0)\| = o(\|(\theta, G) - (\theta_0, G_0)\|) \tag{25}$$

as $(\theta, G) \to (\theta_0, G_0)$ within $\Theta \times \mathcal{G}$ where $\dot{\Psi}^0 = \dot{\Psi}^{\theta_0,G_0} : \mathrm{lin}\,([0,1]^p \times \mathcal{G}) \to \mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$, defined by (10)-(14), is a continuous, linear mapping.

(L2) The inverse $\dot{\Psi}^{-1}_{\theta_0,G_0} : \mathrm{Range}(\dot{\Psi}^{\theta_0,G_0}) \to \mathrm{lin}\,(\Theta \times \mathcal{G})$ exists and is continuous. ($\dot{\Psi}^{-1}_{\theta_0,G_0}$ has a unique continuous extension to the closure of $\mathrm{Range}(\dot{\Psi}^{\theta_0,G_0})$, which we also denote by $\dot{\Psi}^{-1}_{\theta_0,G_0}$, and this operator is the inverse of the unique extension of $\dot{\Psi}_{\theta_0,G_0}$ to the closure of $\mathrm{lin}\,([0,1]^p \times \mathcal{G})$).

(L3) We have, for $\mathbb{S}_n^{\theta_0,G_0}$ defined by (15), $\mathbb{S}_n^{\theta_0,G_0} \rightsquigarrow \mathbb{S}^{\theta_0,G_0}$ in $\mathbb{R}^p \times \ell^\infty(\mathcal{H}_1)$, under $\mathbb{P}_{\nu_0,\theta_0,G_0}$, where $\mathbb{S}^{\theta_0,G_0}$ is a tight, Borel measurable, Gaussian process.

(L4) Let $(\hat{\theta}_n, \hat{G}_n)$, $n \in \mathbb{Z}_+$, be a NPMLE. We have

$$\sqrt{n}\left(\Psi_n - \Psi^{\theta_0,G_0}\right)(\hat{\theta}_n, \hat{G}_n) - \sqrt{n}\left(\Psi_n - \Psi^{\theta_0,G_0}\right)(\theta_0, G_0) = o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}).$$

PROOF (OUTLINE): Let us briefly comment on some elements of the proof of this lemma. (L1) The proof of (L1) is facilitated by the conditional expectation representations in the estimating equation Ψ^{θ_0,G_0}. In particular, we heavily exploit that, due to the chosen versions of conditional probabilities with respect to ε_t,

$$\mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G}\left[f(X_{t-p}, X_{t-p}, \ldots, X_t) \mid \varepsilon_t\right] = \mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0}\left[f(X_{t-p}, X_{t-p}, \ldots, X_t) \mid \varepsilon_t\right],$$

$\mathbb{P}_{\nu_0,\theta_0,G_0}$-a.s. for all $G \in \mathcal{G}$.
(L2) The proof of (L2) is decomposed in the following steps.

(1) In this step we show that we can rewrite some parts of the derivative $\dot{\Psi}^{\theta_0,G_0}$ as follows,

$$\dot{\Psi}_{12}^{\theta_0,G_0}(G - G_0) = -\int A^*_{\theta_0,G_0} \dot{\ell}_\theta(e) \, \mathrm{d}(G - G_0)(e),$$
$$\dot{\Psi}_{22}^{\theta_0,G_0}(G - G_0)h = -\int A^*_{\theta_0,G_0} A_{\theta_0,G_0} h(e) \, \mathrm{d}(G - G_0)(e), \quad h \in \mathcal{H}_1,$$

where $A^*_{\theta_0,G_0}$ is the L_2-adjoint of A_{θ_0,G_0}. This representation allows us to invoke results from Hilbert space theory.

(2) This step shows that to prove that $\dot{\Psi}^{\theta_0,G_0}$ has a continuous inverse, it suffices to prove that a certain operator from $\ell^\infty(\mathbb{Z}_+)$ into itself is onto and continuously invertible.
(3) This step shows that the operator from Step 2 is indeed onto and continuously invertible.

An important step in the proof of (3) is to exploit that

$$0 = \mathbb{E}_{\theta_0,G_0}[h(\varepsilon_0) \mid X_0 = e, X_{-1} = 0, \ldots, X_{-p} = 0] = h(e) \quad \forall e \in \mathbb{Z}_+,$$

which shows that 'local deviations in the immigration distributions' are identifiable from the scores. Unfortunately, it seems to be impossible to obtain an explicit formula for $\dot{\Psi}^{-1}_{\theta_0,G_0}$. This is related to the problem that it seems to be impossible to determine explicit expressions for the efficient influence operator.
(L3)-(L4) (L3) and (L4) are proved by an application of Lemma A.1C. In these proofs we need the existence of the $(p+4)$-th moment of G. □

The next lemma yields a tangent space: it shows that certain parametric submodels of the semiparametric INAR(p) model enjoy the LAN-property.

Lemma A.3 *Let $(\theta, G) \in \Theta \times \mathcal{G}$. Let $a \in \mathbb{R}^p$, and $h : \mathbb{Z}_+ \to \mathbb{R}$ bounded. Introduce probability measures G_τ by*

$$g_\tau(k) = g(k)\left[1 + \tau\left(h(k) - \int h \,\mathrm{d}G\right)\right], \quad k \in \mathbb{Z}_+, \quad |\tau| < \tilde{\epsilon} = (2\|h\|_\infty)^{-1}.$$

Note that, for $|\tau| < \tilde{\epsilon}$, $G_\tau \in \mathcal{G}$. Let $0 < \epsilon \leq \tilde{\epsilon}$ be such that $\theta + \tau a \in \Theta$ for $|\tau| \leq \epsilon$, and denote $\nu^\tau = \nu_{\theta+\tau a,G_\tau}$. Then the sequence of experiments

$$\mathcal{E}_n^{\theta,G}(a,h) = \left(\mathbb{Z}_+^{n+1+p}, 2^{\mathbb{Z}_+^{n+1+p}}, \left(\mathbb{P}^{(n)}_{\nu^\tau,\theta+\tau a,G_\tau} \mid \tau \in (-\epsilon,\epsilon)\right)\right), \quad n \in \mathbb{Z}_+,$$

has the LAN-property at $\tau = 0$ (recall that $Z_t = (X_t, \ldots, X_{t-p})^T$):

$$\log \frac{\mathrm{d}\mathbb{P}^{(n)}_{\nu_n,\theta_n,G_n}}{\mathrm{d}\mathbb{P}^{(n)}_{\nu_{\theta,G},\theta,G}} = \frac{1}{\sqrt{n}}\sum_{t=0}^{n} \begin{pmatrix} a^T & 1 \end{pmatrix} \begin{pmatrix} \dot{\ell}_\theta(Z_t;\theta,G) \\ A_{\theta,G}h(Z_t) - \int h \,\mathrm{d}G \end{pmatrix} - \frac{1}{2}\begin{pmatrix} a^T & 1 \end{pmatrix} J_{\theta,G,h} \begin{pmatrix} a \\ 1 \end{pmatrix} + o(1; \mathbb{P}_{\nu_{\theta,G},\theta,G}),$$

where $\theta_n = \theta + a/\sqrt{n}$, $G_n = G_{1/\sqrt{n}}$, and $\nu_n = \nu^{1/\sqrt{n}}$, and

$$J_{\theta,G,h} = \mathbb{E}_{\nu_{\theta,G},\theta,G}\begin{pmatrix} \dot{\ell}_\theta \dot{\ell}_\theta^T(Z_0;\theta,G) & \dot{\ell}_\theta(Z_0;\theta,G)\left(A_{\theta,G}h(Z_0) - \int h \,\mathrm{d}G\right) \\ \dot{\ell}_\theta^T(Z_0;\theta,G)\left(A_{\theta,G}h(Z_0) - \int h \,\mathrm{d}G\right) & \left(A_{\theta,G}h(Z_0) - \int h \,\mathrm{d}G\right)^2 \end{pmatrix}.$$

In this way we obtain a tangent set (which is already a linear space)

$$\mathcal{T}^0_{\theta,G} = \left\{ a^T \dot{\ell}_\theta(X_{-p},\ldots,X_0;\theta,G) + A_{\theta,G}h(X_{-p},\ldots,X_0) - \int h \,\mathrm{d}G \mid a \in \mathbb{R}^p,\, h \in \ell^\infty(\mathbb{Z}_+) \right\},$$

and the corresponding tangent space is the $\mathrm{L}_2(\nu_{\theta,G} \otimes P^{\theta,G})$-closure of $\mathcal{T}^0_{\theta,G}$: $\mathcal{T}_{\theta,G} = \overline{\mathcal{T}^0_{\theta,G}}$.

PROOF (OUTLINE): By an application of the main theorem of Drost et al. (2006b) the lemma is proved once we prove that $\nu_n\{X_{-p},\ldots,X_{-1}\} - \nu_{\theta,G}\{X_{-p},\ldots,X_{-1}\} \xrightarrow{p} 0$, under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$. By Lemma A.1D this follows if we show (recall that $Y_t = (X_{t-1},\ldots,X_{t-p})^T$)

$$\lim_{n\to\infty} \sup_{y\in\mathbb{Z}_+^p} \frac{\sup_{f:\,|f|\le V} \left|\mathbb{E}_{\delta_y,\theta_n,G_n} f(Y_1) - \mathbb{E}_{\delta_y,\theta,G} f(Y_1)\right|}{V(y)} = 0.$$

where $V(y) = 1 + \sum_{i=1}^p c_i y_i$, $c_i = \theta_i + \ldots, \theta_p$ for $i = 1,\ldots,p$. This is established by direct calculations; see the technical appendix for details. □

References

Ahn, S., Gyemin, L., and Jongwoo, J. (2000). Analysis of the M/D/1-type queue based on an integer-valued autoregressive process. *Operational Research Letters*, **27**, 235-241.

Al-Osh, M. and A. Alzaid (1987). First-order integer-valued autoregressive (INAR(1)) processes. *J. Time Ser. Anal.*, **8**, 261–275.

Al-Osh, M. and A. Alzaid (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *J. Appl. Prob.*, **27**, 314–324.

Alzaid, A. (1988). First-order integer-valued autoregressive (INAR(1)) process: distributional and regression properties. *Statist. Neerlandica*, **42**, 53–61.

Bélisle, P., L. Joseph, B. MacGibbon, D. Wolfson, and R. du Berger (1998). Change-point analysis of neuron spike train data. *Biometrics*, **54**, 113–123.

Berglund, E. and K. Brännäs (2001). Plants' entry and exit in Swedish municipalities. *Ann. Reg. Sci.*, **35**, 431–448.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1998). *Efficient and adaptive estimation for semiparametric models* (2nd ed.). Springer.

Böckenholt, U. (1999a). An INAR(1) negative multinomial regression model for longitudinal count data. *Psychometrika*, **64**, 53–67.

Böckenholt, U. (1999b). Mixed INAR(1) Poisson regression models: analyzing heterogeneity and serial dependencies in longitudinal count data. *J. Econometrics*, **89**, 317–338.

Böckenholt, U. (2003). Analysing state dependences in emotional experiences by dynamic count data models. *J. Roy. Statist. Soc. Ser. C*, **52**, 213–226.

Brännäs, K. and J. Hellström (2001). Generalized integer-valued autoregression. *Econometric Rev.*, **20**, 425–443.

Brännäs, K. and Q. Shahiduzzaman (2004). Integer-valued moving average modelling of the number of transactions in stocks. Working paper Umeå Economic Studies 637.

Cardinal, M., R. Roy, and J. Lambert (1999). On the application of integer-valued time series models for the analysis of disease incidence. *Stat. Med.*, **18**, 2025–2039.

Dion, J.-P., G. Gauthier, and A. Latour (1995). Branching processes with immigration and integer-valued time series. *Serdica Math. J.*, **21**, 123–136.

Doukhan, P. (1994). *Mixing: properties and examples* (1st ed.). Springer-Verlag: Lecture Notes in Statistics, **85**.

Doukhan, P., P. Massart, and E. Rio (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré, Probab. Stat.*, **32**, 393–427.

Drost, F., R. van den Akker, and B. Werker (2006a). An asymptotic analysis of nearly unstable INAR(1) models. *Tilburg University, CentER Discussion paper 2006-44*, http://arno.uvt.nl/show.cgi?fid=53920.

Drost, F., R. van den Akker, and B. Werker (2006b). Local asymptotic normality and efficient estimation for INAR(p) models. *Journal of Time Series Analysis*, forthcoming.

Drost, F., R. van den Akker, and B. Werker (2008). Note on Integer-Valued Bilinear Time Series Models. *Statist. Probab. Lett.*, **78**, 992–996.

Drost, F., C. Klaassen, and B. Werker (1997). Adaptive estimation in time-series models. *Ann. Statist.*, **25**, 786–818.

Du, J.-G. and Y. Li (1991). The integer valued autoregressive (INAR(p)) model. *J. Time Ser. Anal.*, **12**, 129–142.

Franke, J. and T. Seligmann (1993). Conditional maximum-likelihood estimates for INAR(1) processes and their applications to modelling epileptic seizure counts. In: T. Subba Rao (Ed.), *Developments in time series*, pp. 310–330. London: Chapman & Hall.

Freeland, R. and B. McCabe (2004). Analysis of low count time series data by Poisson autoregression. *J. Time Ser. Anal.*, **25**, 701–722.

Freeland, R. and B. McCabe (2005). Asymptotic properties of CLS estimators in the Poisson AR(1) model. *Statist. Probab. Lett.*, **73**, 147-153.

Gouriéroux, C. and J. Jasiak (2004). Heterogeneous INAR(1) model with application to car insurance. *Ins.: Mathematics Econ.*, **34**, 177–192.

Ispány, M., G. Pap, and M. van Zuijlen (2003a). Asymptotic inference for nearly unstable INAR(1) models. *J. Appl. Probab.*, **40**, 750–765.

Ispány, M., G. Pap, and M. van Zuijlen (2003b). Asymptotic behavior of estimators of the parameters of nearly unstable INAR(1) models. In: *Foundations of statistical inference*, eds. Y. Haitovsky, H. Lerche, and Y. Ritov, Physica, Heidelberg, pp. 193–204.

Ispány, M., G. Pap, and M. van Zuijlen (2005). Fluctuation limit of branching processes with immigration and estimation of the means. *Adv. Appl. Probab.*, **37**, 523–538.

Jung, R., G. Ronning, and A. Tremayne (2005). Estimation in conditional first order autoregression with discrete support. *Statist. Papers*, **46**, 195–224.

Jung, R., and A. Tremayne (2006). Binomial thinning models for integer time series. *Statistical Modelling*, **6**, 81–96.

Kartashov, N. (1985). Inequalities in theorems of ergodicity and stability for Markov chains with common phase space I. *Theory Probab. Appl.*, **30**, 247–259.

Kreiss, J.-P. (1987). On adaptive estimation in stationary ARMA processes. *Ann. Statist.*, **15**, 112–133.

Latour, A. (1998). Existence and stochastic structure of a nonnegative integer-valued autoregressive process. *J. Time Ser. Anal.*, **19**, 439–455.

McCabe, B., and Martin, G. (2005). Bayesian predictions of low count time series. *International Journal of Forecasting*, **21**, 315–330.

Meyn, S. and R. Tweedie (1994). *Markov Chains and Stochastic Stability* (2nd ed.). Springer-Verlag.

Neal, P. and Subba Rao, T. (2007). MCMC for integer-valued ARMA processes. *J. Time Ser. Anal.*, **28**, 92–110.

Pickands III, J. and R. Stine (1997). Estimation for an M/G/∞ queue with incomplete information. *Biometrika*, **84**, 295–308.

Rudholm, N. (2001). Entry and the number of firms in the Swedish pharmaceuticals market. *Rev. Ind. Organ.*, **19**, 351–364.

Silva, I. and M. Silva (2006). Asymptotic distribution of the Yule-Walker estimator for INAR(p) processes. *Statist. Probab. Lett.*, **76**, 1655-1663.

Silva, M. and V. Oliveira (2005). Difference equations for the higher order moments and cumulants of the INAR(p) model. *J. Time Ser. Anal.*, **26**, 17–36.

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

Van der Vaart, A. (1991). Efficiency and Hadamard differentiability. *Scand J. Statist.*, **18**, 63-75.

Van der Vaart, A. (1995). Efficiency of infinite dimensional M-estimators. *Statist. Neerlandica*, **49**, 9–30.

Van der Vaart, A. (2000). *Asymptotic Statistics* (1 ed.). Cambridge: Cambridge University Press.

Van der Vaart, A. and J. Wellner (1993). *Weak convergence and empirical processes* (2nd ed.). Springer-Verlag.

Wefelmeyer, W. (1996). Quasi-likelihood and optimal inference. *Ann. Statist.*, **24**, 405–422.

B. Technical appendix to "Efficient estimation of autoregression parameters and innovation distributions for semiparametric integer-valued AR(p) models"

This technical appendix contains the proofs of the results in "Efficient estimation of autoregression parameters and innovation distributions for semiparametric integer-valued AR(p) models". Proofs for the following results are gathered in this note: Lemma A.1, Theorem 2.1, Lemma A.2, and Lemma A.3. Let us briefly comment on these results and their proofs. Lemma A.1 contains auxiliary results, which are needed to prove the other results. Some parts are already known from the literature; the new parts are established by exploiting the V-uniform-ergodicity of an INAR process. Theorem 2.1 shows that the NPMLE is consistent. After a compactification of the parameter space, this consistency follows by Wald's method. Lemma A.2 is essential in establishing the limiting distribution of the NPMLE. The proof of this lemma is rather complicated and long. Finally, Lemma A.3 contains a LAN-result for parametric submodels of the semiparametric INAR(p) model. Apart from negligibility of the initial value, this result follows from the main theorem in Drost et al. (2006b). The negligibility is proved using Lemma A.1.D.

For notational convenience, δ_x denotes the Dirac measure concentrated in x. Also recall the notation $Z_t = (X_t, X_{t-1}, \ldots, X_{t-p})^T$ and $Y_t = (X_{t-1}, \ldots, X_{t-p})^T$.

B.1. Proof of Lemma A.1

Proof of (A): For the existence of the stationary distribution see Dion et al. (1995), Latour (1998), or Drost et al. (2007). The existence of moments follows from Drost et al. (2007).
Proof of (B): Notice first that $\nu_{\theta,G} \otimes P^{\theta,G}$ is the stationary distribution of Z, and that Z is an irreducible, aperiodic Markov chain on $\mathcal{Z} = \text{support}(\nu_{\theta,G} \otimes P^{\theta,G})$. Let Q^n denote the n-step transition-operator of Z (we drop the superscript θ, G). From well-known results on mixing-numbers for Markov chains (see, for example, Doukhan (1994) pages 87-89) it follows that it is sufficient to prove that there exists a function $A : \mathbb{Z}_+^{p+1} \to (0, \infty)$ such that $\int A \, \mathrm{d}(\nu_{\theta,G} \otimes P^{\theta,G}) < \infty$ and

$$\|Q^n(z, \cdot) - \nu_{\theta,G} \otimes P^{\theta,G}\|_{\mathrm{TV}} \leq A(z)\rho^n, \quad z \in \mathcal{Z}, \tag{26}$$

for some $0 < \rho < 1$, where $\| \cdot \|_{\mathrm{TV}}$ denotes the total variational norm of a signed measure. Recall (Meyn and Tweedie (1994) Chapter 16) that for Markov transition-probabilities P_1 and P_2 and a function $1 \leq V < \infty$ the V-norm distance between P_1 and P_2 is defined by $|||P_1 - P_2|||_V = \sup_{z \in \mathcal{Z}} \|P_1(z, \cdot) - P_2(z, \cdot)\|_V / V(z)$, where, for a signed measured μ, $\|\mu\|_V = \sup_{f: |f| \leq V} |\int f \, \mathrm{d}\mu|$. Introduce $V : \mathbb{Z}_+^{p+1} \to [1, \infty)$ by $V(z) = 1 + \sum_{i=1}^p a_i z_i$ with $a_i = \theta_i + \cdots + \theta_p$. Then it is straightforward to check (see also Drost et al. (2007)) that the following drift condition holds. There exists a constant $\beta > 0$ such that for all $z \in \mathcal{Z}$, except for some finite set, we have $\mathbb{E}_{\theta,G}[V(Z_t) \mid Z_{t-1} = z] - V(z) \leq -\beta V(z)$. We conclude from Meyn and Tweedie (1994) Theorem 16.01 that there exist constants $\rho < 1$ and $\tilde{C} < \infty$ such that for all $n \in \mathbb{Z}_+$ $|||Q^n - \nu_{\theta,G} \otimes P^{\theta,G}|||_V \leq \tilde{C}\rho^n$, i.e. Z is V-uniformly ergodic. Since $\mathbb{E}_{\nu_{\theta,G},\theta,G} V(Z_0) < \infty$ (by Lemma A.1A) and $V \geq 1$ (26) immediately follows, which concludes the proof of (A).
Proof of (C): this follows from Doukhan et al. (1995) Theorem 1 and Application 4. Take $r = 3/2$, notice that, using Markov's inequality and $\mathbb{E}_{\nu_{\theta,G},\theta,G} X_0^3 < \infty$ (by Lemma A.1A), the envelope belongs to $\Lambda_3(P) = \Lambda_{x\sqrt{x}}(P)$. Next, take $b > 3/2$ such that $b \geq 3/a$, and note

that there exists $C > 0$ such that $n^{-b} \geq C\rho^n$ for all $n \geq 1$.
Proof of (D): Notice first that $\nu_{\theta,G}$ is the stationary distribution of Y, and that Y is an irreducible, aperiodic Markov chain on $\mathbb{Z}_+^p$. Let $Q^{\theta,G}$ denote the transition-probabilities of Y and Q^n denotes the n-step transition-operator of Y (we drop the superscripts θ, G for the n step operator, since we only consider this operator at (θ, G)). Following the proof of (B) it follows that the Markov chain Y on $\mathbb{Z}_+^p$ is V-uniformly ergodic for $V(Y_t) = 1+\sum_{i=1}^p a_i X_{t-i}$, $a_i = \theta_i + \cdots + \theta_p$, i.e. there exist constants $C > 0$ and $0 < \rho < 1$ such that $|||Q^n - \nu_{\theta,G}|||_V \leq C\rho^n$ for all $n \in \mathbb{Z}_+$. Since Y is uniformly ergodic in the norm $|||\cdot|||_V$, an application of Kartashov (1985) Theorem B (it is easy to see that $|||\cdot|||_V$ satisfies the conditions) yields that Y is strongly stable in this norm: each transition-probability Q' in some neighborhood of Q has a unique stationary measure $\nu(Q')$ and $|||Q'_n - Q|||_V \to 0$ implies $\|\nu(Q'_n) - \nu_{\theta,G}\|_V \to 0$. This yields (D). □

B.2. Proof of consistency Theorem 2.1

Let $(\hat{\theta}_n, \hat{G}_n)$ be a nonparametric maximum likelihood estimator of (θ, G). It suffices to prove $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $\hat{g}_n(k) \xrightarrow{p} g_0(k)$ for all $k \in \mathbb{Z}_+$. We prove this pointwise convergence by an application of Wald's consistency proof. To that end, we first compactify the parameter space, starting with $\mathcal{G}$.

Introduce $\bar{\mathcal{G}}$: the class of all probability distributions on $\mathbb{Z}_+ \cup \{\infty\}$. Identify each $G \in \bar{\mathcal{G}}$ with the sequence $(g(k))_{k\in\mathbb{Z}_+}$. Notice that this correspondence is 1-to-1, since $g(\infty) = 1 - \sum_{k=0}^{\infty} g(k)$. As a result, $\bar{\mathcal{G}}$ is a subset of $[0,1]^{\mathbb{Z}_+}$ equipped with the norm $\|a\| = \sum_{k=0}^{\infty} 2^{-k}|a(k)|$, that is, we endow $[0,1]^{\mathbb{Z}_+}$ with the product topology. Notice that a sequence in $[0,1]^{\mathbb{Z}_+}$ converges if and only if all coordinates, which are sequences in $[0,1]$, converge. Using Helly's lemma (see, for example, Van der Vaart (2000) Lemma 1.5) it is an easy exercise to show that $\bar{\mathcal{G}}$ is a compact subset of $[0,1]^{\mathbb{Z}_+}$. For $G \in \bar{\mathcal{G}}$ define $P^{\theta,G}_{x,\infty} = 1 - \sum_{j\in\mathbb{Z}_+} P^{\theta,G}_{x,j} = g(\infty)$ for $x \in \mathbb{Z}_+^p$ and $P^{\theta,G}_{x,\infty} = 1$ if $\max_{i=1}^p x_i = \infty$.

Now, consider the parameter θ as well. Define $\bar{E} = [0,1]^p \times \bar{\mathcal{G}}$, and equip $\bar{E}$ with the "sum-distance" $d((\theta, G), (\theta', G')) = |\theta - \theta'| + \|(g(k))_{k\in\mathbb{Z}_+} - (g'(k))_{k\in\mathbb{Z}_+}\|$. $\bar{E}$ is the product of two compact spaces and, hence, itself compact.

Define

$$m^{\theta,G}(x_{-p}, \ldots, x_0) = \log P^{\theta,G}_{(x_{-1},\ldots,x_{-p}),x_0},$$

and the (random) function $M_n : \bar{E} \to [-\infty, \infty)$ by

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

From an appropriate law of large number for Markov chains, we find that M_n converges in probability to a (nonrandom) function $M : \bar{E} \to [-\infty, \infty)$ defined by

$$M(\theta, G) = \mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0} m^{\theta,G}(X_{-p}, \ldots, X_0).$$

Note that, by Lemma A.1A, the stationary distribution ν_{θ_0,G_0} indeed exists.

The following holds.

(A) For fixed $x_{-p},\dots,x_0 \in \mathbb{Z}_+$, the map $\bar{E} \ni (\theta, G) \mapsto m^{\theta,G}(x_{-p},\dots,x_0)$ is continuous. This is easy to see, since there appear only a finite number of $g(j)$'s in $P^{\theta,G}_{(x_{-1},\dots,x_{-p}),x_0}$.

(B) For all $x_{-p},\dots,x_0 \in \mathbb{Z}_+$ we have $m^{\theta,G}(x_{-p},\dots,x_0) \le \log(1) = 0$.

(C) The map $\bar{E} \ni (\theta, G) \mapsto M(\theta, G)$ has a unique maximum at (θ_0, G_0). Since we have the identification $P^{\theta,G}_{(X_{-1},\dots,X_{-p}),X_0} = P^{\theta_0,G_0}_{(X_{-1},\dots,X_{-p}),X_0}$ $\mathbb{P}_{\nu_{\theta_0,G_0},\theta_0,G_0}$-a.s. $\implies (\theta, G) = (\theta_0, G_0)$, this easily follows using the following well-known argument (use $\log x \le 2(\sqrt{x} - 1)$ for $x \ge 0$):

$$\begin{aligned} M(\theta,G) - M(\theta_0,G_0) &\le 2\mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0}\left(\sqrt{\frac{P^{\theta,G}_{(X_{-1},\dots,X_{-p}),X_0}}{P^{\theta_0,G_0}_{(X_{-1},\dots,X_{-p}),X_0}}} - 1\right) \\ &= 2\sum_{y\in\mathbb{Z}_+^p} \nu_{\theta_0,G_0}\{y\} \sum_{x_0=0}^{\infty} \sqrt{P^{\theta,G}_{y,x_0} P^{\theta_0,G_0}_{y,x_0}} - 2 \\ &\le -\sum_{y\in\mathbb{Z}_+^p} \nu_{\theta_0,G_0}\{y\} \sum_{x_0=0}^{\infty} \left(\sqrt{P^{\theta,G}_{y,x_0}} - \sqrt{P^{\theta_0,G_0}_{y,x_0}}\right)^2 \le 0. \end{aligned}$$

(D) $M_n(\hat{\theta}_n, \hat{G}_n) \ge M_n(\theta_0, G_0)$, since $(\hat{\theta}_n, \hat{G}_n)$ maximizes the likelihood.

Hence all conditions to Wald's consistency theorem hold (see, for example, the proof of Theorem 5.14 in Van der Vaart (2000), where the law of large numbers for the i.i.d. case has to be replaced by the result above. thus we obtain $d((\hat{\theta}_n, \hat{G}_n), (\theta_0, G_0)) \xrightarrow{p} 0$, which immediately yields $\hat{\theta}_n \xrightarrow{p} \theta_0$ and, for all $k \in \mathbb{Z}_+$, $\hat{g}_n(k) \xrightarrow{p} g_0(k)$. □

B.3. Proof of Lemma A.2

Throughout ν_0 is shorthand for ν_{θ_0,G_0}. If no confusion can arise, sub- and superscripts are sometimes dropped for notational convenience.

B.3.1. Proof of (L1)

To enhance readability the proof is decomposed in three steps. In the first step we show that $\dot{\Psi}$ is indeed linear and continuous. And in the second and third step we prove the Fréchet-differentiability of Ψ_1 and Ψ_2 respectively.

Step 1:

The linearity of $\dot{\Psi}$ is obvious. For the continuity, note that it suffices to prove that both $\dot{\Psi}_1$ and $\dot{\Psi}_2$ are continuous. We consider $\dot{\Psi}_1$ which is the sum of $\dot{\Psi}_{11}$ and $\dot{\Psi}_{12}$; the continuity of $\dot{\Psi}_2$ proceeds in the same way. Of course, $\dot{\Psi}_{11}$ is continuous. So the only thing left is to show that $\dot{\Psi}_{12}$ is continuous. It is easy to see that, here $\dot{\ell}_{\theta,i}$ refers to the ith coordinate of the p-vector $\dot{\ell}_\theta$,

$$\left|\dot{\ell}_{\theta,i}(x_{-p},\dots,x_0;\theta,G)\right| \le \frac{x_{-i}}{\theta_i(1-\theta_i)}, \tag{27}$$

which yields, using that ε_0 and X_{-i} are independent,

$$\left|\mathbb{E}_{\nu_0,\theta_0}\left[\dot{\ell}_{\theta,i}(X_{-p},\dots,X_0;\theta,G) \mid \varepsilon_0\right]\right| \le \frac{\mathbb{E}_{\nu_0} X_{-i}}{\theta_i(1-\theta_i)}.$$

Thus the map

$$\mathbb{Z}_+ \ni e \mapsto \left| \mathbb{E}_{\nu_0,\theta_0} \left[\dot{\ell}_\theta(x_{-p},\dots,x_0;\theta,G) \mid \varepsilon_0 = e \right] \right|$$

is bounded, say by C. This yields, for $H, G \in \operatorname{lin} \mathcal{G}$,

$$\begin{aligned} |\dot{\Psi}_{12}(G-H)| &= \left| \int \mathbb{E}_{\nu_0,\theta_0} \left[\dot{\ell}_\theta(X_{-p},\dots,X_0;\theta_0,G_0) \mid \varepsilon_0 = e \right] \mathrm{d}(H-G)(e) \right| \\ &\leq C \sum_{e=0}^{\infty} |h(e) - g(e)| = C\|H-G\|_1, \end{aligned}$$

which yields the continuity of $\dot{\Psi}_{12}$.

Step 2:
Rewrite,

$$\begin{aligned} \Psi_1(\theta,G) - \Psi_1(\theta_0,G_0) - \dot{\Psi}_{11}(\theta-\theta_0) - \dot{\Psi}_{12}(G-G_0) &= \Psi_1(\theta,G) - \Psi_1(\theta_0,G) - \dot{\Psi}_{11}(\theta-\theta_0) \\ &\quad + \Psi_1(\theta_0,G) - \Psi_1(\theta_0,G_0) - \dot{\Psi}_{12}(G-G_0). \end{aligned}$$

Let θ_n be a sequence in $[0,1]^p$ converging to θ_0 and G_n a sequence in $\mathcal{G}$ converging to G_0. In Step 2a we show that

$$\frac{\left| \Psi_1(\theta_n,G_n) - \Psi_1(\theta_0,G_n) - \dot{\Psi}_{11}(\theta_n-\theta_0) \right|}{|\theta_n-\theta_0| + \|G_n-G_0\|_1} \to 0, \tag{28}$$

and in Step 2b we show that

$$\frac{\left| \Psi_1(\theta_0,G_n) - \Psi_1(\theta_0,G_0) - \dot{\Psi}_{12}(G_n-G_0) \right|}{|\theta_n-\theta_0| + \|G_n-G_0\|_1} \to 0, \tag{29}$$

which will conclude the proof of Step 2.
Step 2a:
First we recall from Drost et al. (2006b) that the usual information-identity holds, i.e.

$$I_\theta(\theta_0,G_0) = \mathbb{E}_{\nu_0,\theta_0,G_0} \dot{\ell}_\theta \dot{\ell}_\theta^T(X_{-p},\dots,X_0;\theta_0,G_0) = -\mathbb{E}_{\nu_0,\theta_0,G_0} \frac{\partial}{\partial \theta^T} \dot{\ell}_\theta(X_{-p},\dots,X_0;\theta_0,G_0).$$

From the mean-value theorem we obtain, for $i = 1,\dots,p$,

$$\begin{aligned} &\dot{\ell}_{\theta,i}(X_{-p},\dots,X_0;\theta,G) - \dot{\ell}_{\theta,i}(X_{-p},\dots,X_0;\theta_0,G) \\ &\qquad = \frac{\partial}{\partial \theta^T} \dot{\ell}_{\theta,i}(X_{-p},\dots,X_0;\tilde{\theta}_i(\theta,G),G)(\theta-\theta_0), \end{aligned}$$

where $\tilde{\theta}_i(\theta,G) = \tilde{\theta}_i(X_{-p},\dots,X_0;\theta,G,\theta_0)$ is a point on the line segment between θ and θ_0. Let $J(X_{-p},\dots,X_0;\theta,G)$ be the $p \times p$ random matrix given by

$$J(X_{-p},\dots,X_0;\theta,G) = \begin{pmatrix} \frac{\partial}{\partial \theta^T} \dot{\ell}_{\theta,1}(X_{-p},\dots,X_0;\tilde{\theta}_1(\theta,G),G) \\ \vdots \\ \frac{\partial}{\partial \theta^T} \dot{\ell}_{\theta,p}(X_{-p},\dots,X_0;\tilde{\theta}_p(\theta,G),G) \end{pmatrix}.$$

It is easy to see, since we only have to deal with a finite number of $g(k)$'s, that we have for fixed $x_{-p},\ldots,x_0$, $J(x_{-p},\ldots,x_0;\theta_n,G_n)\to(\partial/\partial\theta^T)\dot\ell_\theta(X_{-p},\ldots,X_0;\theta_0,G_0)$. From Drost et al. (2006b) we have,

$$\left|\frac{\partial}{\partial\theta_j}\dot\ell_{\theta,i}(x_{-p},\ldots,x_0;\theta,G)\right|\le\frac{3}{2\theta_i(1-\theta_i)\theta_j(1-\theta_j)}(X_{-i}^2+X_{-j}^2),$$

which is $\mathbb{P}_{\nu_0,\theta_0,G_0}$-integrable. Thus, using dominated convergence, we obtain

$$\begin{aligned}&\frac{\left|\Psi_1(\theta_n,G_n)-\Psi_1(\theta_0,G_n)-\dot\Psi_{11}(\theta_n-\theta_0)\right|}{|\theta_n-\theta_0|}\\&\qquad\le\frac{\mathbb{E}_{\nu_0,\theta_0,G_0}\left|\left(I_\theta(\theta_0,G_0)+J(Z_0;\theta_n,G_n)\right)(\theta_n-\theta_0)\right|}{|\theta_n-\theta_0|}\to0,\end{aligned}$$

which yields (28).
Step 2b:
We have, using that $\mathbb{E}_{\nu_0,\theta_0,G}[\cdot\mid\varepsilon_0]$ does not depend on G,

$$\begin{aligned}&\Psi_1(\theta_0,G)-\Psi_1(\theta_0,G_0)-\dot\Psi_{12}(G-G_0)\\&\quad=\mathbb{E}_{\nu_0,\theta_0,G_0}\dot\ell_\theta(X_{-p},\ldots,X_0;\theta_0,G)+\mathbb{E}_G\mathbb{E}_{\nu_0,\theta_0}\left[\dot\ell_\theta(X_{-p},\ldots,X_0;\theta_0,G_0)\mid\varepsilon_0\right]\\&\quad=\mathbb{E}_{\nu_0,\theta_0,G_0}\dot\ell_\theta(X_{-p},\ldots,X_0;\theta_0,G)+\mathbb{E}_{\nu_0,\theta_0,G}\dot\ell_\theta(X_{-p},\ldots,X_0;\theta_0,G_0)\\&\quad=\mathbb{E}_{\nu_0}f(X_{-p},\ldots,X_{-1};G),\end{aligned}$$

where (using that $\mathbb{E}_{\nu_0,\theta_0,H}\left[\dot\ell_\theta(X_{-p},\ldots,X_0;\theta_0;H)\mid X_{-1},\ldots,X_{-p}\right]=0$ for $H\in\mathcal{G}$)

$$\begin{aligned}&f(X_{-p},\ldots,X_{-1};G)\\&=\sum_{x_0=0}^{\infty}\left(P_{Y_0,x_0}^{\theta_0,G}-P_{Y_0,x_0}^{\theta_0,G_0}\right)\left(\dot\ell_\theta(Y_0,x_0;\theta_0,G_0)-\dot\ell_\theta(Y_0,x_0;\theta_0,G)\right)\\&=\sum_{x_0=0}^{\infty}\sum_{k=0}^{\infty}(g(k)-g_0(k))\left(*_{i=1}^p\operatorname{Bin}_{X_{-i},\theta_{0,i}}\right)\{x_0-k\}\left(\dot\ell_\theta(Y_0,x_0;\theta_0,G_0)-\dot\ell_\theta(Y_0,x_0;\theta_0,G)\right)\\&=\sum_{k=0}^{\infty}(g(k)-g_0(k))\sum_{x_0=0}^{\infty}\left(*_{i=1}^p\operatorname{Bin}_{X_{-i},\theta_{0,i}}\right)\{x_0-k\}\left(\dot\ell_\theta(Y_0,x_0;\theta_0,G_0)-\dot\ell_\theta(Y_0,x_0;\theta_0,G)\right).\end{aligned}$$

From this we obtain the bound,

$$|f(X_{-1},\ldots,X_{-p};G)|\le\|G-G_0\|_1\sum_{x_0=0}^{X_{-1}+\cdots+X_{-p}}\left|\dot\ell_\theta(Y_0,x_0;\theta_0,G_0)-\dot\ell_\theta(Y_0,x_0;\theta_0,G)\right|.$$

Since G_n is a sequence in $\mathcal{G}$ converging to G_0, we obtain, for fixed $x_{-p},\ldots,x_{-1}$,

$$\sum_{x_0=0}^{x_{-1}+\cdots+x_{-p}}\left|\dot\ell_\theta(x_{-p},\ldots,x_{-1},x_0;\theta_0,G_0)-\dot\ell_\theta(x_{-p},\ldots,x_{-1},x_0;\theta_0,G_n)\right|\to0.$$

Furthermore, using (27),

$$\sum_{x_0=0}^{x_{-1}+\cdots+x_{-p}} \left|\dot{\ell}_\theta(X_{-p},\ldots,X_{-1},x_0;\theta_0,G_0) - \dot{\ell}_\theta(X_{-p},\ldots,X_{-1},x_0;\theta_0,G_n)\right| \leq 2\sum_{j=1}^{p} \frac{X_{-j}}{\theta_{0,j}(1-\theta_{0,j})}.$$

Thus $f(X_{-p},\ldots,X_{-1};G_n)/\|G_n - G_0\|_1$ converges $\mathbb{P}_{\nu_0}$-a.s. to 0, and is dominated by a $\mathbb{P}_{\nu_0}$-integrable function. An application of the dominated convergence theorem yields (29).

Step 3:
Rewrite,

$$\Psi_2(\theta,G) - \Psi_2(\theta_0,G_0) - \dot{\Psi}_{21}(\theta-\theta_0) - \dot{\Psi}_{22}(G-G_0) = \Psi_2(\theta,G) - \Psi_2(\theta_0,G) - \dot{\Psi}_{21}(\theta-\theta_0) + \Psi_2(\theta_0,G) - \Psi_2(\theta_0,G_0) - \dot{\Psi}_{22}(G-G_0).$$

Let θ_n be a sequence in $[0,1]^p$ converging to θ_0 and G_n a sequence in $\mathcal{G}$ converging to G_0. We will verify that

$$\frac{\sup_{h\in\mathcal{H}_1}\left|\Psi_2(\theta_n,G_n)h - \Psi_2(\theta_0,G_n)h - \dot{\Psi}_{21}(\theta_n-\theta_0)h\right|}{|\theta_n-\theta_0| + \|G_n-G_0\|_1} \to 0, \tag{30}$$

and,

$$\frac{\sup_{h\in\mathcal{H}_1}\left|\Psi_2(\theta_0,G_n)h - \Psi_2(\theta_0,G_0)h - \dot{\Psi}_{22}(G_n-G_0)h\right|}{|\theta_n-\theta_0| + \|G_n-G_0\|_1} \to 0, \tag{31}$$

which will conclude the proof.
Step 3a:
First note that

$$\Psi_2(\theta_n,G_n)h - \Psi_2(\theta_0,G_n)h - \dot{\Psi}_{21}(\theta_n-\theta_0)h = \mathbb{E}_{\nu_0,\theta_0,G_0}\left(A_{\theta_n,G_n}h(Z_0) - A_{\theta_0,G_n}h(Z_0) + A_{\theta_0,G_0}h(Z_0)\dot{\ell}_\theta^T(Z_0;\theta_0,G_0)(\theta_n-\theta_0)\right).$$

It is straightforward to check that, for $i=1,\ldots,p$,

$$\frac{\partial}{\partial\theta_i}A_{\theta,G}h(X_{-p},\ldots,X_0) = \mathbb{E}_{\theta,G}\left[h(\varepsilon_0)\dot{s}_{X_{-i},\theta_i}(\theta_i\circ X_{-i}) \mid X_0,\ldots,X_{-p}\right] - A_{\theta,G}h(X_{-p},\ldots,X_0)\dot{\ell}_{\theta,i}(X_{-p},\ldots,X_0;\theta,G),$$

and for $i,j=1,\ldots,p$,

$$\frac{\partial^2}{\partial\theta_j\partial\theta_i}A_{\theta,G}h(X_{-p},\ldots,X_0) = \mathbb{E}_{\theta,G}\left[h(\varepsilon_0)\dot{s}_{X_{-i},\theta_i}(\theta_i\circ X_{-i})\dot{s}_{X_{-j},\theta_j}(\theta_j\circ X_{-j})|X_0,\ldots,X_{-p}\right] - \mathbb{E}_{\theta,G}\left[h(\varepsilon_0)\dot{s}_{X_{-i},\theta_i}(\theta_i\circ X_{-i})|X_0,\ldots,X_{-p}\right]\dot{\ell}_{\theta,j}(X_{-p},\ldots,X_0;\theta,G)$$

$$
\begin{aligned}
&- A_{\theta,G}h(X_{-p},\ldots,X_0)\ddot{\ell}_{\theta,ij}(X_{-p},\ldots,X_0;\theta,G)\\
&- \dot{\ell}_{\theta,i}(X_{-p},\ldots,X_0;\theta,G)\frac{\partial}{\partial\theta_j}A_{\theta,G}h(X_{-p},\ldots,X_0)\\
&+ 1\{i=j\}\mathbb{E}_{\theta,G}\left[h(\varepsilon_0)\ddot{s}_{X_{-i},\theta_i}(\theta_i\circ X_{-i})\mid X_0,\ldots,X_{-p}\right],
\end{aligned}
$$

where $\ddot{s}_{n,\alpha}(k)=(\partial/\partial\alpha)\dot{s}_{n,\alpha}(k)$. Now it is easy, but a bit tedious, to see that there exists a constant $C_\theta>0$, which is bounded in θ in a neighborhood of θ_0 and not depending on h, such that, for $i,j=1,\ldots,p$,

$$
\left|\frac{\partial}{\partial\theta_i}A_{\theta,G}h(X_{-p},\ldots,X_0)\right|+\left|\frac{\partial^2}{\partial\theta_j\partial\theta_i}A_{\theta,G}h(X_{-p},\ldots,X_0)\right|\le C_\theta(X_{-i}^2+X_{-j}^2). \tag{32}
$$

A second order Taylor expansion in θ yields

$$
\begin{aligned}
&A_{\theta_n,G_n}h(Z_0)-A_{\theta_0,G_n}h(Z_0)+A_{\theta_0,G_n}h(Z_0)\dot{\ell}_\theta^T(Z_0;\theta_0,G_n)(\theta_n-\theta_0)\\
&\quad=\sum_{i=1}^p(\theta_{n,i}-\theta_{0,i})\mathbb{E}_{\theta_0,G_n}\left[h(\varepsilon_0)\dot{s}_{X_{-i},\theta_{0,i}}(\theta_i\circ X_{-i})\mid X_0,\ldots,X_{-p}\right]\\
&\qquad+\frac{1}{2}(\theta_n-\theta_0)^T\frac{\partial^2}{\partial\theta\partial\theta^T}A_{\tilde{\theta}_n,G_n}h(X_{-p},\ldots,X_0)(\theta_n-\theta_0),
\end{aligned}
$$

where $\tilde{\theta}_n$ is a random point on the line segment between θ_0 and θ_n (also depending on h, Z_0, and G_n). Using (32) it easily follows, using dominated convergence, that

$$
\sup_{h\in\mathcal{H}_1}\frac{\left|\mathbb{E}_{\nu_0,\theta_0,G_0}(\theta_n-\theta_0)^T\frac{\partial^2}{\partial\theta\partial\theta^T}A_{\tilde{\theta}_n,G_n}h(X_{-p},\ldots,X_0)(\theta_n-\theta_0)\right|}{|\theta_n-\theta_0|+\|G_n-G_0\|_1}\to 0.
$$

Hence we obtain (30) once we show that

$$
\begin{aligned}
&\sup_{h\in\mathcal{H}_1}\frac{\left|\sum_{i=1}^p(\theta_{n,i}-\theta_{0,i})\mathbb{E}_{\nu_0,\theta_0,G_0}\mathbb{E}_{\theta_0,G_n}\left[h(\varepsilon_0)\dot{s}_{X_{-i},\theta_{0,i}}(\theta_i\circ X_{-i})\mid X_0,\ldots,X_{-p}\right]\right|}{|\theta_n-\theta_0|+\|G_n-G_0\|_1}\\
&\qquad\to 0,
\end{aligned}\tag{33}
$$

and,

$$
\begin{aligned}
&\sup_{h\in\mathcal{H}_1}\frac{\left|\mathbb{E}_{\nu_0,\theta_0,G_0}\left(A_{\theta_0,G_n}h(Z_0)\dot{\ell}_\theta^T(Z_0;\theta_0,G_n)(\theta_n-\theta_0)-A_{\theta_0,G_0}h(Z_0)\dot{\ell}_\theta^T(Z_0;\theta_0,G_0)(\theta_n-\theta_0)\right)\right|}{|\theta_n-\theta_0|+\|G_n-G_0\|_1}\\
&\qquad\to 0
\end{aligned}\tag{34}
$$

both hold. It is easy to see that we have, for $i=1,\ldots,p$,

$$
\begin{aligned}
&\left|\mathbb{E}_{\theta_0,G_n}\left[h(\varepsilon_0)\dot{s}_{X_{-i},\theta_{0,i}}(\theta_i\circ X_{-i})\mid Z_0\right]-\mathbb{E}_{\theta_0,G_0}\left[h(\varepsilon_0)\dot{s}_{X_{-i},\theta_{0,i}}(\theta_i\circ X_{-i})\mid Z_0\right]\right|\\
&\quad\le\left|\frac{P^{\theta_0,G_0}_{(X_{-1},\ldots,X_{-p}),X_0}}{P^{\theta_0,G_n}_{(X_{-1},\ldots,X_{-p}),X_0}}-1\right|\frac{X_{-i}}{\theta_{0,i}(1-\theta_{0,i})}\\
&\qquad+\frac{\sum_{e=0}^{X_0}\sum_{k=0}^{X_{-i}}|g_n(e)-g_0(e)|\left(*_{j\neq i}\operatorname{Bin}_{X_{-j},\theta_{0,j}}\right)\{X_0-k-e\}\dot{s}_{X_{-i},\theta_{0,i}}(k)\,\mathrm{b}_{X_{-i},\theta_{0,i}}(k)}{P^{\theta_0,G_n}_{(X_{-1},\ldots,X_{-p}),X_0}},
\end{aligned}
$$

which for fixed $X_{-p},\dots,X_0$ converges to 0. Note that the left-hand-side of this display is bounded by the ν_0-integrable variable $2X_{-i}/(\theta_{0,i}(1-\theta_{0,i}))$. Since $\mathbb{E}_{\nu_0,\theta_0,G_0}h(\varepsilon_0)\dot{s}_{X_{-i},\theta_i}(\theta_i\circ X_{-i})=0$, by independence of ε_0 and $\theta_i\circ X_{-i}-\theta_{0,i}X_{-i}$, Display (33) easily follows using dominated convergence. In a similar fashion we obtain (34).
Step 3b:
Note first that we have

$$\begin{aligned}&\Psi_2(\theta_0,G_n)h-\Psi_2(\theta_0,G_0)h-\dot{\Psi}_{22}(G_n-G_0)h\\&\quad=\mathbb{E}_{\nu_0,\theta_0,G_0}A_{\theta_0,G_n}h(Z_0)-\int h\,\mathrm{d}G_n+\mathbb{E}_{\nu_0,\theta_0,G_n}A_{\theta_0,G_0}h(Z_0)-\int h\,\mathrm{d}G_0.\end{aligned}$$

It now follows that we have

$$\Psi_2(\theta_0,G_n)h-\Psi_2(\theta_0,G_0)h-\dot{\Psi}_{22}(G_n-G_0)h=\mathbb{E}_{\nu_0}f^h(X_{-p},\dots,X_{-1};G_n),$$

where

$$f^h(X_{-p},\dots,X_{-1};G_n)=\sum_{x_0=0}^{\infty}\left(P^{\theta_0,G_n}_{Y_0,x_0}-P^{\theta_0,G_0}_{Y_0,x_0}\right)\left(A_{\theta_0,G_0}h(Y_0,x_0)-A_{\theta_0,G_n}h(Y_0,x_0)\right).$$

Proceeding as in Step 2b we obtain the bound

$$|f^h(X_{-p},\dots,X_{-1};G_n)|\le\|G_n-G_0\|_1\sum_{x_0=0}^{X_{-p}+\cdots+X_{-1}}\left|A_{\theta_0,G_0}h(Y_0,x_0)-A_{\theta_0,G_n}h(Y_0,x_0)\right|.$$

Using that, for $x_0\in\{0,\dots,X_{-p}+\cdots+X_{-1}\}$,

$$\begin{aligned}&\sup_{h\in\mathcal{H}_1}\left|A_{\theta_0,G_n}h(X_{-p},\dots,X_{-1},x_0)-A_{\theta_0,G_0}h(X_{-p},\dots,X_{-1},x_0)\right|\\&\quad\le\left|\frac{P^{\theta_0,G_0}_{(X_{-p},\dots,X_{-1}),x_0}}{P^{\theta_0,G_n}_{(X_{-p},\dots,X_{-1}),x_0}}-1\right|\left|A_{\theta_0,G_0}h(X_{-p},\dots,X_{-1},x_0)\right|\\&\qquad+\frac{\sum_{e=0}^{x_0}|g_n(e)-g_0(e)|\left(*_{i=1}^{p}\operatorname{Bin}_{X_{-i},\theta_{0,i}}\right)\{x_0-e\}}{P^{\theta_0,G_n}_{(X_{-p},\dots,X_{-1}),x_0}},\end{aligned}$$

we see that for fixed $(X_{-p},\dots,X_{-1})$ $\sup_{h\in\mathcal{H}_1}|f^h(X_{-p},\dots,X_{-1};G_n)|/\|G_n-G_0\|_1\to 0$. Since $2(X_{-p}+\cdots+X_{-1})$ is an ν_0-integrable envelope for $\sup_{h\in\mathcal{H}_1}|f^h(X_{-p},\dots,X_{-1};G_n)|/\|G_n-G_0\|_1$, an application of dominated convergence yields (31).

B.3.2. Proof of (L2)

First we prove (L2) for the case $\operatorname{support}(G_0)=\mathbb{Z}_+$. To enhance readability we decompose the proof into the following steps.

(1) In this step we show that we can rewrite some parts of the derivative $\dot{\Psi}$ as follows,

$$\dot{\Psi}_{12}(G-G_0)=-\int A_0^*\dot{\ell}_\theta(e)\,\mathrm{d}(G-G_0)(e),\tag{35}$$

$$\dot{\Psi}_{22}(G-G_0)h = -\int A_0^* A_0 h(e)\,\mathrm{d}(G-G_0)(e), \quad h \in \mathcal{H}_1, \tag{36}$$

where A_0^* is the L_2-adjoint of $A_0 = A_{\theta_0,G_0}$. This representation allows us to invoke results from Hilbert space theory.

(2) This step shows that to prove that $\dot{\Psi}$ has a continuous inverse, it suffices to prove that a certain operator from $\ell^\infty(\mathbb{Z}_+)$ into itself is onto and continuously invertible.

(3) This step shows that the operator from Step 2 is indeed onto and continuously invertible.

Step 1:
Let $[\varepsilon]$ denote $\{f(\varepsilon_0) \mid f : \mathbb{Z}_+ \to \mathbb{R}, \mathbb{E}_{G_0} f^2(\varepsilon_0) < \infty\}$ equipped with the $\mathrm{L}_2(G_0)$ norm and let $[X]$ denote $\{f(X_{-p},\ldots,X_0) \mid f : \mathbb{Z}_+^{p+1} \to \mathbb{R}, \mathbb{E}_{\nu_0,\theta_0,G_0} f^2(X_{-p},\ldots,X_0) < \infty\}$ equipped with the $\mathrm{L}_2(\nu_0 \otimes P^{\theta_0,G_0})$ norm. It is not hard to see that both these spaces are, in fact, Hilbert spaces (that these spaces are already in their "a.s.-equivalence class form", follows from support$(G_0) = \mathbb{Z}_+$). We view upon A_0 as an operator from $[\varepsilon]$ into $[X]$. From the definition it is easy to see that A_0 is linear and continuous. Since A_0 is a continuous linear map between two Hilbert spaces, it has an adjoint map $A_0^* : [X] \to [\varepsilon]$ (which is a continuous linear map that satisfies and is uniquely determined by the equations $< A_0^* h_2, h_1 >_{[\varepsilon]} = < h_2, A_0 h_1 >_{[X]}$ for $h_1 \in [\varepsilon]$, $h_2 \in [X]$) given by

$$A_0^* f = A_0^* f(\varepsilon_0) = \mathbb{E}_{\nu_0,\theta_0}[f(X_{-p},\ldots,X_0) \mid \varepsilon_0].$$

Now, invoking the definitions of $\dot{\Psi}_{12}$ and $\dot{\Psi}_{22}$, (35) and (36) are immediate.

Step 2:
To prove that $\dot{\Psi}$ is continuously invertible, it suffices to prove that $\dot{\Psi}_{11} : \mathbb{R}^p \to \mathbb{R}^p$ and $\dot{V} = \dot{\Psi}_{22} - \dot{\Psi}_{21}\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12} : \operatorname{lin}\mathcal{G} \to \ell^\infty(\mathcal{H}_1)$ are both continuously invertible. The invertibility of $\dot{\Psi}_{11}$ is immediate, since the $p \times p$ Fisher information-matrix $I_{\theta_0} = \mathbb{E}_{\nu_0,\theta_0,G_0} \dot{\ell}_\theta \dot{\ell}_\theta^T(Z_0;\theta_0,G_0)$ is invertible (see Drost et al. (2006b)). To prove that $\dot{V}$ is continuously invertible is much harder. In this step, we will give an easier sufficient condition which is proved to hold true in Step 3. Introduce the operator $C : \mathcal{H}_1 \to [\varepsilon]$ by

$$Ch(e) = -\left[\mathbb{E}_{\nu_0,\theta_0,G_0} A_0 h(X_{-p},\ldots,X_0)\dot{\ell}_\theta^T(X_{-p},\ldots,X_0;\theta_0,G_0)\right] I_{\theta_0}^{-1}(A_0^*(\dot{\ell}_\theta(\cdot;\theta_0,G_0)))(e),$$

for $e \in \mathbb{Z}_+$, where $A_0^*(\dot{\ell}_\theta(\cdot;\theta_0,G_0)) = (A_0^*(\dot{\ell}_{\theta,1}(\cdot;\theta_0,G_0)),\ldots,A_0^*(\dot{\ell}_{\theta,p}(\cdot;\theta_0,G_0)))^T \in [\varepsilon]^p$. Then $\dot{V}$ can be rewritten as

$$\dot{V}(G-G_0)h = -\int \left(A_0^* A_0 h + Ch\right)(e)\,\mathrm{d}(G-G_0)(e), \quad h \in \mathcal{H}_1.$$

The mapping $\dot{V} : \operatorname{lin}\mathcal{G} \to \ell^\infty(\mathcal{H}_1)$ has a continuous inverse on its range if and only if there exists $\epsilon > 0$ such that

$$\|\dot{V}(G-G_0)\| = \sup_{h\in\mathcal{H}_1} |\dot{V}(G-G_0)h| \geq \epsilon \|G-G_0\|_1, \quad \text{for all } G \in \operatorname{lin}\mathcal{G}.$$

Notice that we have, since $(e \mapsto \operatorname{sgn}(g(e)-g_0(e))) \in \mathcal{H}_1$,

$$\|G-G_0\|_1 = \sum_{e=0}^{\infty} |g(e)-g_0(e)| \leq \sup_{h\in\mathcal{H}_1} \left|\int h\,\mathrm{d}(G-G_0)\right|.$$

Hence it suffices to prove that there exists $\epsilon > 0$ such that, for all $G \in \operatorname{lin} \mathcal{G}$,

$$\begin{aligned}\|\dot{V}(G - G_0)\| = \sup_{h \in \mathcal{H}_1} |\dot{V}(G - G_0)h| &= \sup_{h \in \mathcal{H}_1} \left| \int (A_0^* A + C)h \, \mathrm{d}(G - G_0) \right| \\ &\geq \epsilon \sup_{h \in \mathcal{H}_1} | \int h \, \mathrm{d}(G - G_0)|.\end{aligned}$$

Of course, a sufficient condition for this is given by $\epsilon \mathcal{H}_1 \subset \{(A_0^* A_0 + C)h \mid h \in \mathcal{H}_1\}$, which in turn holds if $B = A_0^* A_0 + C : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is onto and continuously invertible. To see this, first note that $\epsilon \mathcal{H}_1 \subset \{(A_0^* A_0 + C)h \mid h \in \mathcal{H}_1\}$ is equivalent to $\epsilon B^{-1} \mathcal{H}_1 \subset \mathcal{H}_1$. Since $\mathcal{H}_1$ is the unit-ball of $\ell^\infty(\mathbb{Z}_+)$ it thus suffices to show that there exists $\epsilon > 0$ such that $\|B^{-1}h\|_\infty \leq \epsilon^{-1}$ for all $h \in \mathcal{H}_1$. Since B^{-1} is continuous, there exists $\epsilon > 0$ such that $\|Bf\|_\infty \geq \epsilon \|f\|_\infty$ for all $f \in \ell^\infty(\mathbb{Z}_+)$. Taking $h \in \mathcal{H}_1$ and $f = B^{-1}h$ (which is possible, because B is onto), we indeed arrive at $\|B^{-1}h\|_\infty = \|f\|_\infty \leq \epsilon^{-1} \|Bf\|_\infty = \epsilon^{-1} \|h\|_\infty \leq \epsilon^{-1}$. Thus $\dot{\Psi}$ is continuously invertible if we prove that $A_0^* A_0 + C : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is onto and continuously invertible. This concludes Step 2.

Step 3:
In this step we prove that $B = A_0^* A_0 + C : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is onto and continuously invertible, which will conclude the proof of (L2). Notice that $C : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is a compact operator, since it has finite dimensional range. From functional analysis (see, for example, Van der Vaart (2000) Lemma 25.93), it is known that (all operators are defined on and take values in a common Banach space) the sum of a compact operator and a continuous operator, which is onto and has a continuous inverse, is continuously invertible and onto if the sum operator is 1-to-1. Thus it suffices to prove that $A_0^* A_0 : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is continuous, onto, and has a continuous inverse (Step 3a), and that B is one-to-one (Step 3b).
Step 3a:
The continuity of $A_0^* A_0$ is immediate,

$$\begin{aligned}\sup_{e \in \mathbb{Z}_+} |A_0^* A_0 h(e) - A_0^* A_0 h'(e)| &= \sup_{e \in \mathbb{Z}_+} |\mathbb{E}_{\nu_0, \theta_0} \left[\mathbb{E}_{\theta_0, G_0} \left[h(\varepsilon_0) - h'(\varepsilon_0) \mid X_0, \ldots, X_{-p} \right] \mid \varepsilon_0 = e \right]| \\ &\leq \sup_{e \in \mathbb{Z}_+} |h(e) - h'(e)|.\end{aligned}$$

Next we show that to prove that $A_0^* A_0 : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is onto and continuously invertible, it suffices to prove that $A_0^* A_0 : [\varepsilon] \to [\varepsilon]$ is onto and continuously invertible. If we already know that $A_0^* A_0 : [\varepsilon] \to [\varepsilon]$ is invertible, then $A_0^* A_0 : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is also invertible (since there are no "a.s.-problems" if support$(G_0) = \mathbb{Z}_+$). If $h \in \ell^\infty(\mathbb{Z}_+)$ it is clear that $A_0^* A_0 h \in \ell^\infty(\mathbb{Z}_+)$ Suppose next that $A_0^* A_0 h \in \ell^\infty(\mathbb{Z}_+)$. Since

$$\begin{aligned}A_0^* A_0 h(e) &= \sum_{y \in \mathbb{Z}_+^p} \sum_{x_0 = 0}^{\infty} \nu_0\{y\} \left(*_{i=1}^p \operatorname{Bin}_{y_i, \theta_{0,i}} \right) \{x_0 - e\} \mathbb{E}_{\theta_0, G_0}[h(\varepsilon_0) \mid Y_0 = y] \\ &\geq \nu_0\{0, \ldots, 0\} h(e),\end{aligned}$$

this implies $h \in \ell^\infty(\mathbb{Z}_+)$. Thus, since $A_0^* A_0 : [\varepsilon] \to [\varepsilon]$ is onto and $\ell^\infty(\mathbb{Z}_+) \subset [\varepsilon]$, $A_0^* A_0 : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is indeed onto. Thus $A_0^* A_0 : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is a linear continuous operator, whose range is a Banach space, we conclude, from Banach's theorem, that $A_0^* A_0 :$

$\ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is continuously invertible. Hence, the proof of Step 3a is complete once we show that $A_0^* A_0 : [\varepsilon] \to [\varepsilon]$ is onto and continuously invertible. First we show that $A_0 : [\varepsilon] \to R_2(A_0) \subset \mathrm{L}_2(\nu_0 \otimes P^{\theta_0,G_0})$ ($R_2(A_0)$ is the range of A_0, where we use the "subscript 2" to stress that we working in L_2) is one-to-one, i.e. that the null space of A_0 is trivial. Let $h : \mathbb{Z}_+ \to \mathbb{R}$ such that $\mathbb{E}_{G_0} h^2(\varepsilon_0) < \infty$ and

$$0 = \mathbb{E}_{\theta_0,G_0}[h(\varepsilon_0) \mid X_0, \ldots, X_{-p}] \quad \mathbb{P}_{\nu_0,\theta_0,G_0} - \text{a.s.}$$

Since $\text{support}(G_0) = \mathbb{Z}_+$, we can drop the "a.s." and we obtain

$$0 = \mathbb{E}_{\theta_0,G_0}[h(\varepsilon_0) \mid X_0 = e, X_{-1} = 0, \ldots, X_{-p} = 0] = h(e) \quad \forall e \in \mathbb{Z}_+$$

We see that $h(\varepsilon_0) = 0$ and hence A_0 is invertible, with inverse

$$(A_0^{-1} f)(\varepsilon_0) = f(0, \ldots, 0, \varepsilon_0).$$

Of course this is a linear operator. Moreover it is continuous since (remember that $P^{\theta_0,G_0}_{(0,\ldots,0),x_0} = g_0(x_0)$)

$$\begin{aligned}\mathbb{E}_{G_0}\left(A_0^{-1} f(\varepsilon_0) - A_0^{-1} f'(\varepsilon_0)\right)^2 &= \mathbb{E}_{G_0}\left(f(0,\ldots,0,\varepsilon_0) - f'(0,\ldots,0,\varepsilon_0)\right)^2 \\ &\leq \frac{1}{\nu_0\{0,\ldots,0\}} \mathbb{E}_{\nu_0,\theta_0,G_0}\left(f(X_{-p},\ldots,X_0) - f'(X_{-p},\ldots,X_0)\right)^2.\end{aligned}$$

Since $A_0 : [\varepsilon] \to R_2(A_0)$ is linear, continuous, one-to-one, and has a continuous inverse, we conclude from Banach's theorem that $R_2(A_0)$ is a closed subspace of $\mathrm{L}_2(\nu_0 \otimes P^{\theta_0,G_0})$. Since A_0 is one-to-one, and $R_2(A_0)$ is closed we conclude that the operator $A_0^* A_0 : [\varepsilon] \to [\varepsilon]$ is one-to-one, onto and has a continuous inverse (fact from Hilbert-space theory). This concludes Step 3a.

Step 3b:

In this step we show that $B : \ell^\infty(\mathbb{Z}_+) \to \ell^\infty(\mathbb{Z}_+)$ is one-to-one. This essentially follows from the proof of Lemma 25.92 in Van der Vaart (2000). For completeness we repeat the arguments, where we circumvent the need to consider the efficient information matrix for θ. Let $h \in \ell^\infty(\mathbb{Z}_+)$, with $Bh = 0$. We have to prove that $h = 0$. Introduce $\mathbb{R}^p \ni a = -I_{\theta_0}^{-1} \mathbb{E}_{\nu_0,\theta_0,G_0} A_0 h(Z_0) \dot{\ell}_\theta(Z_0; \theta_0, G_0)$, and notice that $Ch = a^T A_0^* \dot{\ell}_\theta(\cdot; \theta_0, G_0)$. Let $S = a^T \dot{\ell}_\theta(Z_0; \theta_0, G_0) + A_0 h(Z_0) - \int h \, \mathrm{d}G_0$. First we show that for $a \neq 0$ we have $\mathbb{E}_{\nu_0,\theta_0,G_0} S^2 > 0$. Suppose that $S = 0$ $\mathbb{P}_{\nu_0,\theta_0,G_0}$-a.s. Then conditioning on $X_{-p} = \cdots = X_{-1} = 0$ yields $h(e) - \int h \, \mathrm{d}G_0 = 0$ for all e. And we obtain, since I_{θ_0} is positive definite (Drost et al. (2006b) Theorem 3.1), $\mathbb{E}_{\nu_0,\theta_0,G_0} S^2 = a^T I_{\theta_0} a > 0$ for $a \neq 0$, which contradicts $\mathbb{E}_{\nu_0,\theta_0,G_0} S^2 = 0$. Conclude that we have, for $a \neq 0$,

$$0 < \mathbb{E}_{\nu_0,\theta_0,G_0} S^2 = \mathbb{E}_{\nu_0,\theta_0,G_0}\left(A_0 h(Z_0) - \int h \, \mathrm{d}G_0\right)^2 - a^T I_{\theta_0} a.$$

On the other hand $Bh = 0$, yields

$$\begin{aligned}0 &= \mathbb{E}_{\nu_0,\theta_0,G_0} h(\varepsilon_0) B h(\varepsilon_0) = \mathbb{E}_{\nu_0,\theta_0,G_0}(A_0 h(Z_0))^2 + a^T \mathbb{E}_{\nu_0,\theta_0,G_0} A_0 h(Z_0) \dot{\ell}_\theta(Z_0; \theta_0, G_0) \\ &\geq \mathbb{E}_{\nu_0,\theta_0,G_0}\left(A_0 h(Z_0) - \int h \, \mathrm{d}G_0\right)^2 - a^T I_{\theta_0} a.\end{aligned}$$

From the previous two displays we conclude $a = 0$, which by definition of a and C yields $Ch = 0$. Hence $A_0^* A_0 h = 0$, which, by Step 3a, yields $h = 0$. This concludes the proof.

So we have proved (L2) for the case support$(G_0) = \mathbb{Z}_+$. The proof for the general case uses exactly the same arguments, if we replace in the arguments where "a.s." plays a role $\mathbb{Z}_+$ by support(G_0). Recall that we always have, by assumption, $g_0(0) > 0$.

B.3.3. Proof of (L3)

The weak-convergence of $\sqrt{n}\left(\Psi_{n1} - \Psi_1^{\theta_0,G_0}\right)(\theta_0, G_0)$ follows from Lemma A.1C, since we are dealing with a finite function class and $|\dot{\ell}_{\theta,i}(Z_0; \theta_0, G_0)| \leq X_{-i}(\theta_{0,i}(1-\theta_{0,i}))^{-1}$, $i = 1, \ldots, p$. Hence, due to the form of $\sqrt{n}\left(\Psi_n - \Psi^{\theta_0,G_0}\right)(\theta_0, G_0)$, it suffices to prove that $\sqrt{n}\left(\Psi_{n2} - \Psi_2^{\theta_0,G_0}\right)(\theta_0, G_0)$ weakly converges, under $\mathbb{P}_{\nu_0,\theta_0,G_0}$, in $\ell^\infty(\mathcal{H}_1)$ to a tight Gaussian process. This can be reexpressed as the weak convergence of the empirical process $\{\mathbb{Z}_n f \mid f \in \mathcal{F}\}$, where $\mathcal{F} = \{\mathbb{Z}_+^{p+1} \ni (x_{-p}, \ldots, x_0) \mapsto A_0 h(x_{-p}, \ldots, x_0) \mid h \in \mathcal{H}_1\}$. We use Lemma A.1B to verify this. Let $\delta > 0$. Take $M_\delta = \lceil (8(p+1)\mathbb{E}_{\nu_0,\theta_0,G_0} X_0^{p+2})^{1/(p+2)} \delta^{-2/(p+2)} \rceil$. By Markov's inequality we have

$$\mathbb{P}_{\nu_0,\theta_0,G_0}\{\max_{i=0,\ldots,p} X_{-i} \geq M_\delta\} \leq \frac{\delta^2}{8}.$$

Next, form a grid of cubes with sides of length $\epsilon_\delta = \delta/2\sqrt{2}$ over $[-1,1]^{\{0,\ldots,M_\delta-1\}^{p+1}}$. This yields $N_\delta \leq \lceil 2/\epsilon_\delta \rceil^{M_\delta^{p+1}}$ points. Each point yields a mapping $f : \{0, \ldots, M_\delta - 1\}^{p+1} \to [-1, 1]$. We label these as $f_1, \ldots, f_{N_\delta}$. Since for $h \in \mathcal{H}_1$ we have $|A_0 h| \leq 1$, there exists $i \in \{1, \ldots, N_\delta\}$ such that $f_i(x_{-p}, \ldots, x_0) - \delta/2\sqrt{2} \leq A_0 h(x_{-p}, \ldots, x_0) \leq f_i(x_{-p}, \ldots, x_0) + \delta/2\sqrt{2}$ for $x_{-p}, \ldots, x_0 \leq M_\delta - 1$. Next we introduce mappings f_i^L, f_i^U, $i = 1, \ldots, N_\delta$, from $\mathbb{Z}_+^{p+1}$ into $[-1, 1]$ by $f_i^L = -1 \vee (f_i - \delta/2\sqrt{2})$ if $\max\{x_{-p}, \ldots, x_0\} \leq M_\delta - 1$, $f_i^L = -1$ for $\max\{x_{-p}, \ldots, x_0\} \geq M_\delta$, and $f_i^U = 1 \wedge (f_i + \delta/2\sqrt{2})$ if $\max\{x_{-p}, \ldots, x_0\} \leq M_\delta - 1$ and $f_i^U = 1$ if $\max\{x_{-p}, \ldots, x_0\} \geq M_\delta$. Conclude that for $h \in \mathcal{H}_1$ there exists $i \in \{1, \ldots, N_\delta\}$ such that $f_i^L \leq A_0 h \leq f_i^U$. So the brackets $\left[f_i^L, f_i^U\right]$, $i = 1, \ldots, N_\delta$, cover $\mathcal{F}$ and satisfy

$$\mathbb{E}_{\nu_0,\theta_0,G_0}(f_i^U - f_i^L)^2 \leq \left(\frac{\delta}{\sqrt{2}}\right)^2 + 4\mathbb{P}_{\nu_0,\theta_0,G_0}\{\max_{i=0,\ldots,p} X_{-i} \geq M_\delta\} \leq \delta^2.$$

Conclude that $N_{[]}(\delta, \mathcal{F}) \leq N_\delta$. Using $\log(x) \leq m(x^{1/m} - 1)$ for $x > 0$, $m \in \mathbb{N}$, it easily follows that we can find $a > 0$ such that $\int_0^1 x^{-a} \left(\log N_{[]}(x, \mathcal{F})\right)^{1/2} \mathrm{d}x < \infty$. Since the envelope of $\mathcal{F}$ is bounded by 2, an application of Lemma A.1C concludes the proof.

B.3.4. Proof of (L4)

In step A we prove

$$\sqrt{n}\left(\Psi_{n2} - \Psi_2^{\theta_0,G_0}\right)(\hat{\theta}_n, \hat{G}_n) - \sqrt{n}\left(\Psi_{n2} - \Psi_2^{\theta_0,G_0}\right)(\theta_0, G_0) = o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}), \tag{37}$$

and in step B we prove

$$\sqrt{n}\left(\Psi_{n1} - \Psi_1^{\theta_0,G_0}\right)(\hat{\theta}_n, \hat{G}_n) - \sqrt{n}\left(\Psi_{n1} - \Psi_1^{\theta_0,G_0}\right)(\theta_0, G_0) = o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}), \tag{38}$$

which will conclude the proof. Introduce for $\delta > 0$ $B_0(\delta) = \{(\theta, G) \in \Theta \times \mathcal{G} \mid |\theta - \theta_0| + \|G - G_0\|_1 \leq \delta\}$.

Step A: If we prove that there exists $\delta > 0$ such that

$$\lim_{n\to\infty} \sup_{h\in\mathcal{H}_1} \mathbb{E}_{\nu_0,\theta_0,G_0} \left(A_{\theta_n,G_n} h(X_{-p},\ldots,X_0) - A_{\theta_0,G_0} h(X_{-p},\ldots,X_0)\right)^2 = 0,$$

for all sequences (θ_n, G_n) in $\Theta \times \mathcal{G}$ converging to (θ_0, G_0), and that the empirical process $\{\mathbb{Z}_n f \mid f \in \mathcal{F}^\delta\}$ with $\mathcal{F}^\delta$ given by

$$\mathcal{F}^\delta = \{(x_{-p},\ldots,x_0) \mapsto A_{\theta,G} h(x_{-p},\ldots,x_0) - A_{\theta_0,G_0} h(x_{-p},\ldots,x_0) \mid h \in \mathcal{H}_1,\, (\theta, G) \in B_0(\delta)\},$$

weakly converges to a tight Gaussian process, then (37) follows from (the proof of) Lemma 3.3.5 in Van der Vaart and Wellner (1993). Since

$$\sup_{h\in\mathcal{H}_1} |A_{\theta_n,G_n} h(X_{-p},\ldots,X_0) - A_{\theta_0,G_0} h(X_{-p},\ldots,X_0)| \leq 2,$$

and since, for fixed $X_{-p},\ldots,X_0$,

$$\sup_{h\in\mathcal{H}_1} |A_{\theta_n,G_n} h(X_{-p},\ldots,X_0) - A_{\theta_0,G_0} h(X_{-p},\ldots,X_0)| \leq \left|\frac{P^{\theta_0,G_0}_{Y_0,X_0}}{P^{\theta_n,G_n}_{Y_0,X_0}} - 1\right| + \frac{\|G_n - G_0\|_1}{P^{\theta_n,G_n}_{Y_0,X_0}} \to 0,$$

the first condition easily follows by an application of the dominated convergence theorem. That the process $\{\mathbb{Z}_n f \mid f \in \mathcal{F}^\delta\}$ weakly converges to a tight Gaussian process follows by the same arguments as in the proof of (L3).

Step B: We consider the first coordinate. The others proceed in exactly the same way. If we prove that there exists $\delta > 0$ such that

$$\lim_{n\to\infty} \mathbb{E}_{\nu_0,\theta_0,G_0} \left(\dot{\ell}_{\theta,1}(X_{-p},\ldots,X_0;\theta_n,G_n) - \dot{\ell}_{\theta,1}(X_{-p},\ldots,X_0;\theta_0,G_0)\right)^2 = 0,$$

for all sequences (θ_n, G_n) in $\Theta \times \mathcal{G}$ converging to (θ_0, G_0), and that the empirical process $\{\mathbb{Z}_n f \mid f \in \mathcal{F}^\delta\}$ with $\mathcal{F}^\delta$ given by

$$\mathcal{F}^\delta = \left\{(x_{-p},\ldots,x_0) \mapsto \dot{\ell}_{\theta,1}(x_{-p},\ldots,x_0;\theta,G) - \dot{\ell}_{\theta,1}(x_{-p},\ldots,x_0;\theta_0,G_0) \mid (\theta, G) \in B_0(\delta)\right\},$$

converges weakly to a tight Gaussian process, then (38) follows from Lemma 3.3.5 in Van der Vaart and Wellner (1993). Choose $\delta > 0$ such that for all θ in the ball we have $(\theta_i(1-\theta_i))^{-1} \leq C$ for certain $C > 0$ and all $i = 1,\ldots,p$. The first condition easily follows using dominated convergence (use $4CX_{-1}^2$ as dominating function). We use Lemma A.1C to verify the second condition. Let $\eta > 0$. Take $M_\eta = \lceil \alpha^{1/(p+4)} \eta^{-2/(p+2)} \rceil$, where the constant α is given by $\alpha = (p+1)\left(8C^2 \mathbb{E}_{\nu_0,\theta_0,G_0} X_0^{p+4}\right)^{(p+4)/(p+2)}$. By Markov's inequality we have

$$\mathbb{P}_{\nu_0,\theta_0,G_0}\{\max_{i=0,\ldots,p} X_{-i} \geq M_\eta\} \leq \frac{\mathbb{E}_{\nu_0,\theta_0,G_0} X_0^{p+4}}{\left(8C^2 \mathbb{E}_{\nu_0,\theta_0,G_0} X_0^{p+4}\right)^{(p+4)/(p+2)}} \eta^{2\frac{p+4}{p+2}},$$

and using Hölder's inequality we now obtain

$$\mathbb{E}_{\nu_0,\theta_0,G_0} X_{-1}^2 1\{\max_{i=0,\ldots,p} X_{-i} \geq M_\eta\}$$
$$\leq \left(\mathbb{E}_{\nu_0,\theta_0,G_0} X_{-1}^{p+4}\right)^{2/(p+4)} \left(\mathbb{P}_{\nu_0,\theta_0,G_0}\{\max_{i=0,\ldots,p} X_{-i} \geq M_\eta\}\right)^{(p+2)/(p+4)} \leq \frac{\eta^2}{8C^2}. \quad (39)$$

Notice that for all $(\theta, G) \in B_0(\delta)$ we have

$$|\dot{\ell}_{\theta,1}(x_{-p},\ldots,x_0;\theta,G) - \dot{\ell}_{\theta,1}(x_{-p},\ldots,x_0;\theta_0,G_0)| \leq 2Cx_{-1}.$$

Next, form a grid of cubes with sides of length $\epsilon_\eta = \eta/2\sqrt{2}$ over $[-2CM_\eta, 2CM_\eta]^{\{0,\ldots,M_\eta-1\}^{p+1}}$. This yields $N_\eta \leq \lceil 4CM_\eta/\epsilon_\eta \rceil^{M_\eta^{p+1}}$ points. Each point yields a mapping $f : \{0,\ldots,M_\eta - 1\}^{p+1} \to [-2CM_\eta, 2CM_\eta]$. We label these as $f_1,\ldots,f_{N_\eta}$. So, for $(\theta, G) \in B_0(\delta)$, there exists $i \in \{1,\ldots,N_\eta\}$ such that, for $x_{-p},\ldots,x_0 \leq M_\eta - 1$,

$$f_i(x_{-p},\ldots,x_0) - \frac{\eta}{2\sqrt{2}} \leq \dot{\ell}_{\theta,1}(x_{-p},\ldots,x_0;\theta,G) - \dot{\ell}_{\theta,1}(x_{-p},\ldots,x_0;\theta_0,G_0)$$
$$\leq f_i(x_{-p},\ldots,x_0) + \frac{\eta}{2\sqrt{2}}.$$

Next we introduce mappings f_i^L, f_i^U, $i = 1,\ldots,N_\eta$, from $\mathbb{Z}_+^{p+1}$ into $\mathbb{R}$ by $f_i^L = -2CM_\eta \vee (f_i - \eta/2\sqrt{2})$ if $\max\{x_{-p},\ldots,x_0\} \leq M_\eta - 1$ and $f_i^L = -2Cx_{-1}$ if $\max\{x_{-p},\ldots,x_0\} \geq M_\eta$, and $f_i^U = 2CM_\eta \wedge (f_i + \eta/2\sqrt{2})$ if $\max\{x_{-p},\ldots,x_0\} \leq M_\eta - 1$ and $f_i^U = 2Cx_{-1}$ if $\max\{x_{-p},\ldots,x_0\} \geq M_\eta$. Conclude that for $(\theta, G) \in B_0(\delta)$ there exists $i \in \{1,\ldots,N_\eta\}$ such that $f_i^L \leq \dot{\ell}_{\theta,1}(\theta, G) - \dot{\ell}_{\theta,1}(\theta_0, G_0) \leq f_i^U$. So the brackets $\left[f_i^L, f_i^U\right]$, $i = 1,\ldots,N_\eta$, cover $\mathcal{F}^\delta$ and satisfy, by (39),

$$\mathbb{E}_{\nu_0,\theta_0,G_0}(f_i^U - f_i^L)^2 \leq \left(\frac{\eta}{\sqrt{2}}\right)^2 + 4C^2\mathbb{E}_{\nu_0,\theta_0,G_0} X_{-1}^2 1\{\max_{i=0,\ldots,p} X_{-i} \geq M_\eta\} \leq \eta^2.$$

Conclude that $N_{[]}(\eta, \mathcal{F}^\delta) \leq N_\eta$. Using $\log(x) \leq m(x^{1/m} - 1)$ for $x > 0$, $m \in \mathbb{N}$, it easily follows that we can find $a > 0$ such that $\int_0^1 x^{-a} \left(\log N_{[]}(x, \mathcal{F}^\delta)\right)^{1/2} \mathrm{d}x < \infty$. Since the envelope of $\mathcal{F}^\delta$ is bounded by the integrable variable $2CX_{-1}$, an application of Lemma A.1C concludes the proof. □

B.4. Proof of LAN Theorem A.3

By an application of the main theorem of Drost et al. (2006b) the lemma is proved once we prove that $\nu_n\{X_{-p},\ldots,X_{-1}\} - \nu_{\theta,G}\{X_{-p},\ldots,X_{-1}\} \xrightarrow{p} 0$, under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$. By Lemma A.1D this follows if we show (recall that $Y_t = (X_{t-1},\ldots,X_{t-p})^T$)

$$\lim_{n\to\infty} \sup_{y\in\mathbb{Z}_+^p} \frac{\sup_{f:\,|f|\leq V} \left|\mathbb{E}_{\delta_y,\theta_n,G_n} f(Y_1) - \mathbb{E}_{\delta_y,\theta,G} f(Y_1)\right|}{V(y)} = 0, \quad (40)$$

where $V(y) = 1 + \sum_{i=1}^p c_i y_i$, $c_i = \theta_i + \ldots, \theta_p$ for $i = 1,\ldots,p$. Straightforward computations yield

$$\mathbb{E}_{\delta_y,\theta_n,G_n} f(Y_1) - \mathbb{E}_{\delta_y,\theta,G} f(Y_1) = \frac{1}{\sqrt{n}} \mathbb{E}_{\delta_y,\theta_n,G}(h(\varepsilon_0) - \mathbb{E}_G h(\varepsilon_0)) f(Y_1)$$

$$+\int_0^{\frac{1}{\sqrt{n}}}\sum_{i=1}^{p} a_i \mathbb{E}_{\delta_y,\theta+\tau a,G} f(Y_1)\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})\,\mathrm{d}\tau.$$

We have, for a constant $C>0$, the bound

$$\begin{aligned}\sup_{f:\,|f|\le V}\left|\mathbb{E}_{\delta_y,\theta_n,G}(h(\varepsilon_0)-\mathbb{E}_G h(\varepsilon_0))f(Y_1)\right| &\le 2\|h\|_\infty\left(1+\sum_{i=2}^{p} c_i y_{i-1}+\mu_G+\left(\theta+\frac{a}{\sqrt{n}}\right)^T y\right)\\ &\le CV(y).\end{aligned}$$

Next let $i\in\{1,\dots,p\}$. Of course the supremum in

$$\sup_{f:\,|f|\le V}\left|\mathbb{E}_{\delta_y,\theta+\tau a,G} f(Y_1)\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})\right|$$

is taken for $f=V1_A-V1_{A^c}$, where $A=\{\dot{s}_{X_{-i},\theta_i}(\theta\circ X_{-i})>0\}$. Consequently, in the first equality we exploit $\mathbb{E}_{\delta_y,\theta+\tau a,G}\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})=0$,

$$\begin{aligned}&\sup_{f:\,|f|\le V}\left|\mathbb{E}_{\delta_y,\theta+\tau a,G} f(Y_1)\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})\right|\\ &\quad=\sup_{f:\,|f|\le V}\left|\mathbb{E}_{\delta_y,\theta+\tau a,G}(f(Y_1)-\mathbb{E}_{\delta_y,\theta+\tau a,G}f(Y_1))\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})\right|\\ &\quad=\mathbb{E}_{\delta_y,\theta+\tau a,G}1_A(V(Y_1)-\mathbb{E}_{\delta_y,\theta+\tau a,G}V(Y_1))\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})\\ &\qquad-\mathbb{E}_{\delta_y,\theta+\tau a,G}1_{A^c}(V(Y_1)-\mathbb{E}_{\delta_y,\theta+\tau a,G}V(Y_1))\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})\end{aligned}$$

(fill in V and use $\mathbb{E}_{\delta_y,\theta+\tau a,G}\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})=0$)

$$\begin{aligned}&=c_1\mathbb{E}_{\delta_y,\theta+\tau a,G}1_A(X_0-\mathbb{E}_{\delta_y,\theta+\tau a,G}X_0)\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})\\ &\quad-c_1\mathbb{E}_{\delta_y,\theta+\tau a,G}1_{A^c}(X_0-\mathbb{E}_{\delta_y,\theta+\tau a,G}X_0)\dot{s}_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})\\ &\le c_1\sqrt{\mathbb{E}_{\delta_y,\theta+\tau a,G}(X_0-\mathbb{E}_{\delta_y,\theta+\tau a,G}X_0)^2}\sqrt{\mathbb{E}_{\delta_y,\theta+\tau a,G}\dot{s}^2_{X_{-i},\theta_i+\tau a_i}(\theta_i\circ X_{-i})}\\ &=c_1\sqrt{\sigma_G^2+\sum_{j=1}^{p}(\theta_j+ta_j)y_j}\sqrt{\theta_i(1-\theta_i)y_i}\\ &\le CV(y),\end{aligned}$$

for a constant $C>0$. A combination of the previous four displays easily yields (40). □