Integer-valued time series

Integer-valued time series

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg, op gezag van de rector magnificus, prof.dr. F.A. van der Duyn Schouten, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op woensdag 7 november 2007 om 14.15 uur door

RAMON VAN DEN AKKER,

geboren op 21 februari 1980 te Amsterdam.

PROMOTOR: prof.dr. Bas J.M. Werker
COPROMOTOR: dr. Feike C. Drost

OVERIGE COMMISSIELEDEN: prof.dr. John H.J. Einmahl
prof.dr. Marc Hallin
prof.dr. Chris A.J. Klaassen
prof.dr. Brendan P.M. McCabe
dr. Johan J.J. Segers

THOMAS STIELTJES INSTITUTE
FOR MATHEMATICS

Voor Aitor

Acknowledgements

First of all I want to express my gratitude and appreciation to my academic advisors, Feico Drost and Bas Werker. I really enjoyed our cooperation and appreciate their kind support. I also enjoyed the collaboration with Johan Segers and Marc Hallin. A lot of thanks to John Einmahl, Marc Hallin, Chris Klaassen, Brendan McCabe, and Johan Segers for their willingness to serve on my thesis committee. I am grateful to my colleagues of the department of Econometrics & OR for the relaxed and stimulating working environment. In particular, I want to thank Alex, Hendri, and Ruud. Finally, but above all, I thank Vera, Aitor, Eddy, Grizzly, my parents, and Daan for their love and support.

Preface

This thesis consists of two parts. The first part contributes statistical methodology for nonnegative integer-valued time series. The second part of this thesis consists of two chapters. One chapter is concerned with the development of efficient estimators of the marginal distribution functions from multivariate data if one has knowledge on the dependence structure. The other chapter considers semiparametric estimation for general (continuous) time series models with innovations that are not necessarily independently and identically distributed. A short overview of both parts is presented below.

Part I In many sciences one encounters nonnegative discrete valued time series, often as counts of events or objects at consecutive points in time. Especially in economics and medicine many interesting variables are (nonnegative) integer-valued. For example: the number of transactions in SNS-Reaal during each day, the number of patients in a hospital at the end of the day, the number of claims an insurance company receives during each day, the number of epileptic seizures a patient suffers each day, etcetera. Hence the need for adequate probabilistic models and statistical techniques for nonnegative discrete valued time series is apparent. However, until the early eighties this area of research did not attract much attention. As possible explanation McKenzie (2003) mentions that modeling discrete valued time series is a challenging topic in time series analysis since most traditional representations of dependence become either impossible or impractical. The last two decades there have been attempts to develop suitable classes of models; the class of INteger-valued AutoRegressive (INAR) processes can, presently, be considered as the major model for discrete valued time series. Part I of this thesis contributes statistical methods for INAR processes. Chapter 1 contains some probabilistic results on INAR processes. The existence of a strictly stationary solution, the existence of moments under the stationary distribution, and the (uniform) ergodicity of INAR processes is investigated. In Chapter 2 parametric INAR processes are discussed: the innovation distribution G belongs to a parametric family. The main result of this chapter is that parametric INAR models enjoy the Local Asymptotic Normality (LAN) property. The proofs are made tractable by a certain representation of the transition-scores, which is motivated by an information-loss interpretation of the model. Furthermore, a new computationally attractive, asymptotically efficient estimator of the parameters is provided. Using a parametric model exposes the researcher to misspecification. Therefore Chapter 3 considers a semiparametric model, where hardly any assumptions are made on the innovation distribution. The focus is on efficient estimation of the Euclidean parameters as well as the distribution of the innovations. Even inefficient estimation of the innovation distribution has, to my best knowledge, not been addressed before. A possible explanation for this is that, even if the parameters are known, the innovations cannot be calculated from the observations. Consequently, estimation of the innovation distribution cannot be based on residuals (as is the case for AR processes). However, estimation of the innovation distribution is, just as for standard AR models, an important topic. For INAR processes this might be even more important, since in some applications the innovation distribution has a physical interpretation. We provide an estimator which might be viewed upon as a nonparametric maximum likelihood estimator. It turns out that we cannot prove efficiency by standard semiparametric methodology. Efficiency is proved by using the special representation of the limit distribution. In Chapters 2 and 3 the models only considered the 'stationary part' of the parameter space. To analyze the INAR model on the boundary of the parameter space, Chapter 4 considers a nearly nonstationary INAR(1) model and derives its limit experiment (in the Le Cam framework). The main result of this chapter is that this limit experiment is based on one observation from a Poisson distribution. This is rather surprising since limit experiments are usually Locally Asymptotically Quadratic (LAQ; see Jeganathan (1995) and Le Cam and Yang (1990)) and even non-regular models often enjoy a shift structure (see Hirano and Porter (2003a)), whereas the Poisson limit experiment does not enjoy these two properties. To illustrate the statistical consequences of the convergence to a Poisson limit experiment, we exploit this limit experiment to construct efficient estimators of the autoregression parameter in various models, and to construct an efficient test for the null hypothesis of a unit root. Related to this, we show that the Dickey-Fuller test for a unit root has no (local asymptotic) power.

Part II Chapter 5 discusses estimation of the marginals from a bivariate random sample. The only assumption on the marginals is that they are absolutely continuous. By Sklar's theorem, the joint distribution is uniquely determined by the copula (the dependence structure), and the marginal distributions. Of course, the marginal empirical distribution functions are $\sqrt{n}$-consistent esti-

mators of the marginal distribution functions. If the components are independent then these estimators are known to be efficient. We prove that, amongst smooth copulas, this is actually the only case that the marginal empirical distribution functions are efficient. So the natural question is how knowledge on the copula should be exploited to improve on the empirical distribution functions. Motivated by an empirical likelihood argument we provide a new estimator of the marginal distribution functions. Since the tangent space is the sum of two non-orthogonal spaces, traditional semiparametric arguments cannot be used to prove efficiency of our estimator. We derive, by ad hoc arguments, a special representation of the limiting distribution of our estimator. Using this representation we prove efficiency.
Chapter 6 derives semiparametric efficiency bounds for parametric components in general semiparametric time series models. The time series models considered are not, as is the case in the usual semiparametric time series approach, assumed to be driven by a sequence of independent innovations with an unknown distribution. Instead of this, the dependence between the innovations is seen as an additional nonparametric nuisance parameter. A Local Asymptotic Normality (LAN) result is, under quite natural and economical conditions, derived implying a lower bound on the asymptotic performance of (regular) estimators.

This thesis is partly based on the following research papers.

Chapter 1:
Drost, F. C., R. van den Akker, and B.J.M. Werker (2007). Note on integer-valued bilinear time series models, *CentER discussion Paper*, 2007-47, Tilburg University.

Chapter 2:
Drost, F. C., R. van den Akker, and B. J. M. Werker (2006). Local asymptotic normality and efficient estimation for INAR(p) models, *CentER discussion Paper*, 2006-45, Tilburg University.

Chapter 3:
Drost, F. C., R. van den Akker, and B.J.M. Werker (2007). Efficient estimation of autoregression parameters and innovation distribution for semiparametric nonnegative integer-valued AR(p) models, *CentER discussion Paper*, 2007-23, Tilburg University.

Chapter 4:
Drost, F. C., R. van den Akker, and B. J. M. Werker (2006). An asymptotic analysis of nearly unstable INAR(1) models, *CentER discussion Paper*, 2006-44, Tilburg University.

Chapter 5:
Segers, J. J. J., R. van den Akker, and B. J. M. Werker (2007). Efficient estimation of marginals by exploiting knowledge on the copula, *Working paper.*

Chapter 6:
Drost, F. C., R. van den Akker, and B. J. M. Werker (2007). Semiparametric efficiency bounds for time series models with non-i.i.d. innovations, *Working paper.*

Contents

Part I

Nonnegative integer-valued autoregressive processes

1 Preliminaries

In Section 1.1 we recall the definition of INAR processes. Section 1.2, the main part of this chapter, contributes some probabilistic results on INAR processes. In particular, conditions for the existence of a strictly stationary solution and the existence of moments under the stationary distribution are provided.

Definition 1.1

The nonnegative INteger-valued AutoRegressive process of the order 1 (INAR(1)) was introduced by Al-Osh and Alzaid (1987). The INAR(1) process is defined by the recursion,

$$X_t = \vartheta \circ X_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z}_+ = \mathbb{N} \cup \{0\}, \tag{1.1}$$

where[1],

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

The variables $(Z_j^{(t)})_{j \in \mathbb{N}, t \in \mathbb{Z}_+}$ are i.i.d Bernoulli distributed 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 $\mathbb{Z}_+$. Finally, the starting value X_{-1}, with distribution ν on $\mathbb{Z}_+$, is independent of $(\varepsilon_t)_{t \in \mathbb{Z}_+}$ and $(Z_j^{(t)})_{j \in \mathbb{N}, t \in \mathbb{Z}_+}$. The random variable $\vartheta \circ X_{t-1}$ is called the Binomial thinning of X_{t-1} (this operator was introduced by Steutel and Van Harn (1979) and, conditionally on X_{t-1}, it follows a Binomial distribution with success probability θ and number of trials equal to X_{t-1}). Display (1.1) can be interpreted as a branching process with immigration. The outcome

[1]An empty sum equals, by definition, 0. Although it would be more accurate to write $\vartheta^{(t)}\circ$ instead of $\vartheta\circ$, this superscript is, to keep in line with the literature, dropped.

X_t is composed of the surviving elements of X_{t-1} during the period $(t-1,t]$, $\vartheta \circ X_{t-1}$, 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 that, even if the true parameter θ would be known, the number of immigrants cannot be derived from the X process 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 (1.1) with p lags. An INAR(p) process is recursively defined by,

$$X_t = \vartheta_1 \circ X_{t-1} + \vartheta_2 \circ X_{t-2} + \cdots + \vartheta_p \circ X_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z}_+, \tag{1.2}$$

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

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

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 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 $(X_{-1},\ldots,X_{-p})'$ 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 $(\Omega,\mathscr{F},\mathbb{P}_{\nu,\theta,G})$, where $\theta = (\theta_1,\ldots,\theta_p)'$.

Without going into details, let us mention some empirical applications of INAR processes. Applications in the medical sciences can be found in, for example, Franke and Seligmann (1993) (epileptic seizure counts), Bélisle et al. (1998) (spike trains), and Cardinal et al. (1999) (infectious disease incidence). An application to psychometrics can be found in Böckenholt (1999a) (daily emotion experiences), an application to environmentology in Thyregod et al. (1999) (rainfall (rain data is most often collected by means of a tipping bucket rain gauge, which is a discrete sampler counting the number of times a bucket is filled in each sampling time interval)); recent applications to economics in, for example, Böckenholt (1999b), Berglund and Brännäs (2001) (number of plants in Swedish municipalities), Brännäs and Hellström (2001), Rudholm (2001), Böckenholt (2003), Freeland and McCabe (2004), Gouriéroux and Jasiak (2004) (number of claims an insurance company receives), and McCabe and Martin (2005); and Pickands III and Stine (1997) and Ahn et al. (2000) considered queueing applications.

Stationarity, moments & auxiliaries 1.2

This section introduces notation for Part I of this thesis, and provides some probabilistic results, which we need in later chapters.

Throughout Part I the number of lags, $p \in \mathbb{N}$, is fixed. The following notation is used: $\mathscr{G}$ denotes the set of all probability measures on $\mathbb{Z}_+ = \mathbb{N} \cup \{0\}$. 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), $\mathrm{b}_{n,\theta}$ denotes the corresponding point mass function, and δ_x denotes the Dirac measure concentrated in x. In general, we denote a probability measure on $\mathbb{Z}_+$ by a capital, and denote the associated probability mass function by the corresponding lower case. For $G \in \mathscr{G}$, μ_G denotes the mean of G, and σ_G^2 denotes its variance. 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, $\mathbb{F} = (\mathscr{F}_t)_{t \geq -p}$ is the filtration generated by X, i.e. $\mathscr{F}_t = \sigma\left(X_{-p}, \ldots, X_t\right)$. Note that, contrary to classical AR(p) processes, $\mathscr{F}_t \neq \sigma\left(X_{-p}, \ldots, X_{-1}, \varepsilon_0, \ldots, \varepsilon_t\right)$.

Next, we compute the first two conditional moments of an INAR(p) process to gain some insight in its dependence structure. It immediately follows from (1.2) that, for $t \in \mathbb{Z}_+$,

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

and,

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

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 (actually it is an ARCH(p) process), whereas the conditional variance is constant for AR(p) processes. For computations on higher-order moments we refer to Silva and Oliveira (2004, 2005).

Next we determine the conditional distribution of X_t given $\mathscr{F}_{t-1}$. From (1.2) it follows, for $t \in \mathbb{Z}_+$,

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

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,

$$\begin{aligned} P^{\theta,G}_{(x_{t-1},\ldots,x_{t-p}),x_t} &= \mathbb{P}_{\theta,G}\left\{\sum_{i=1}^{p}\vartheta_i \circ X_{t-i} + \varepsilon_t = x_t \mid X_{t-1} = x_{t-1},\ldots,X_{t-p} = x_{t-p}\right\} \\ &= \left(\mathrm{Bin}_{x_{t-1},\theta_1} * \cdots * \mathrm{Bin}_{x_{t-p},\theta_p} * G\right)\{x_t\}. \qquad (1.3) \end{aligned}$$

Note that $X = (X_t)_{t\geq -p}$ is a pth order Markov chain. To exploit this Markovian structure we introduce the $\mathbb{Z}_+^p$-valued process $Y = (Y_t)_{t\geq 0}$ defined by

$$Y_t = (X_{t-1}, X_{t-2},\ldots,X_{t-p})', \quad t \in \mathbb{Z}_+. \qquad (1.4)$$

Under $\mathbb{P}_{\nu,\theta,G}$ the process Y is a (first-order) Markov chain in $\mathbb{Z}_+^p$. It is easy to see that, in case $g(0) < 1$ and $\theta \in (0,1)^p$, the Markov chain Y is irreducible on $\{\alpha,\alpha+1,\ldots\}^p$, where $\alpha = \min\{k \in \mathbb{Z}_+ \mid g(k) > 0\}$. It is also easily seen that, under these conditions, the chain is also aperiodic.

Franke and Seligmann (1993) gave conditions for the existence of a (strictly) stationary INAR(1) process using generating functions. Du and Li (1991), Dion et al. (1995), and Latour (1998) proved the existence of a stationary INAR(p) process in case $\mathbb{E}_G\varepsilon_0^2 < \infty$ and $\sum_{i=1}^p \theta_i < 1$. Only using an elementary result on Markov chains, we give an alternative shorter proof.

Theorem 1.1. *For all $G \in \mathscr{G}$ with $g(0) \in [0,1)$, $\mu_G < \infty$, and $\theta \in (0,1)^p$ with $\sum_{i=1}^p \theta_i < 1$, there exists a probability measure $\nu_{\theta,G}$ on $\mathbb{Z}_+^p$ such that X is a strictly stationary process under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$. The support of $\nu_{\theta,G}$ is given by $\{\alpha,\alpha+1,\ldots\}^p$, where $\alpha = \min\{k \in \mathbb{Z}_+ \mid g(k) > 0\}$.*

Remark 1. Clearly, in case $g(0) = 1$, a strictly stationary solution is given by $X_t = 0$ for all t, i.e. $\nu_{\theta,G} = \delta_0$.

Proof.
First note that it suffices to prove that the Markov chain Y (see (1.4)) has a stationary distribution. We prove this for the case $g(0) > 0$. The case $g(0) = 0$ runs along the same lines.

By $Q^n_{i,j}$ we denote the n-step probability of moving from state i to j of the process Y, i.e., $Q^n_{i,j} = \mathbb{P}_{\delta_i,\theta,G}\{Y_n = j\}$, $i,j \in \mathbb{Z}_+^p$. Since, under the imposed conditions, Y is aperiodic and irreducible on $\mathbb{Z}_+^p$, it suffices, by, for example, Theorem 8.8 in Billingsley (1995), to prove that there exist states $i,j \in \mathbb{Z}_+^p$ for which $Q^n_{i,j}$

does not converge to 0 as $n \to \infty$.

It is easy to see that, for all $t \in \mathbb{Z}_+$, $\mathbb{E}_{\delta_0,\theta,G} X_t < \infty$ when $\mathbb{E}_G \varepsilon_0 < \infty$. We first show that we even have $\sup_{t\in\mathbb{Z}_+} \mathbb{E}_{\delta_0,\theta,G} X_t < \infty$. Note that this statement indeed holds true if we can show that $\mathbb{E}_{\delta_0,\theta,G} X_t \le \mu_G \sum_{j=0}^t \theta_*^j$, where $\theta_* = \sum_{i=1}^p \theta_i$ which is less than 1 by assumption. Obviously we have $\mathbb{E}_{\delta_0,\theta,G} X_{-1} = \cdots = \mathbb{E}_{\delta_0,\theta,G} X_{-p} = 0$ and $\mathbb{E}_{\delta_0,\theta,G} X_0 = \mu_G$. Hence the statement holds for $t \in \{-p,\dots,0\}$. Let $N \in \mathbb{Z}_+$. Assuming that $\mathbb{E}_{\delta_0,\theta,G} X_t \le \mu_G \sum_{j=0}^t \theta_*^j$ is valid for all $t \in \{-p,\dots,N\}$ we obtain

$$\mathbb{E}_{\delta_0,\theta,G} X_{N+1} = \mu_G + \sum_{i=1}^p \theta_i \mathbb{E}_{\delta_0,\theta,G} X_{N+1-i} \le \mu_G + \sum_{i=1}^p \theta_i \mu_G \sum_{j=0}^N \theta_*^j = \mu_G \sum_{j=0}^{N+1} \theta_*^j,$$

which concludes the induction argument.
Using $\sup_{t\in\mathbb{Z}_+} \mathbb{E}_{\delta_0,\theta,G} X_t < \infty$, Markov's inequality yields, for $M > 0$,

$$\sup_{t\in\mathbb{Z}_+} \mathbb{P}_{\delta_0,\theta,G}\left\{\max_{i=1,\dots,p} X_{t-i} > M\right\} \le \frac{p}{M} \sup_{t\in\mathbb{Z}_+} \mathbb{E}_{\delta_0,\theta,G} X_t.$$

Hence there exists $M \in \mathbb{N}$ such that, for all $t \in \mathbb{Z}_+$, $\mathbb{P}_{\delta_0,\theta,G}\left\{\max_{i=1,\dots,p} X_{t-i} \le M\right\} \ge 1/2$. Define $B_M = \{(x_1,\dots,x_p) \in \mathbb{Z}_+^p \mid \forall i \in \{1,\dots,p\} : x_i \le M\}$, then, for $n \ge 1$,

$$\begin{aligned} Q_{0,0}^{n+p} &= \mathbb{P}_{\delta_0,\theta,G}\{Y_{n+p} = 0\} \ge \mathbb{P}_{\delta_0,\theta,G}\{Y_n \in B_M, Y_{n+p} = 0\} \\ &= \sum_{\substack{(i_1,\dots,i_{n-1})\in\mathbb{Z}_+^{(n-1)p} \\ i_n\in B_M \\ (i_{n+1},\dots,i_{n+p-1})\in\mathbb{Z}_+^{(p-1)p}}} Q_{0,i_1}\cdots Q_{i_{n+p-1},0} \\ &= \sum_{\substack{(i_1,\dots,i_{n-1})\in\mathbb{Z}_+^{(n-1)p} \\ i_n\in B_M}} Q_{0,i_1}\cdots Q_{i_{n-1},i_n} Q_{i_n,0}^p. \end{aligned}$$

Using $i_n \in B_M$ we obtain the (very crude) bound $Q_{i_n,0}^p \ge \left[g(0)(1-\theta_*)^{pM}\right]^p$, where $\theta_* = \max_{i=1,\dots,p} \theta_i$. Since,

$$\sum_{(i_1,\dots,i_{n-1})\in\mathbb{Z}_+^{(n-1)p}, i_n\in B_M} Q_{0,i_1}\cdots Q_{i_{n-1},i_n} = \mathbb{P}_{\delta_0,\theta,G}\left\{X_{n-p} \le M,\dots,X_{n-1} \le M\right\} \ge \frac{1}{2},$$

we obtain, for all $n \in \mathbb{N}$,

$$\begin{aligned} Q_{0,0}^{n+p} &\ge \left[g(0)(1-\theta_*)^{pM}\right]^p \sum_{\substack{(i_1,\dots,i_{n-1})\in\mathbb{Z}_+^{(n-1)p} \\ i_n\in B_M}} Q_{0,i_1}\cdots Q_{i_{n-1},i_n} \\ &\ge \frac{1}{2}\left[g(0)(1-\theta_*)^{pM}\right]^p > 0, \end{aligned}$$

which concludes the proof. □

Only in special cases it is possible to derive explicit formulas for $\nu_{\theta,G}$. As an example: for $p = 1$ and $G = \text{Poisson}(\mu)$ we have $\nu_{\theta,G} = \text{Poisson}(\mu/(1-\theta_1))$ (this is well-known and is, using generating functions, easy to check). For this specific case it is immediate that, under the stationary distribution, all moments exist. In general this is not the case: if, for example, $\sigma_G^2 = \infty$ then, under $\nu_{\theta,G}$, X_0 cannot have a finite second moment. The next lemma gives sufficient conditions for the existence of moments under the stationary distribution. Oddly enough it appears that the existence of higher order moments has not been considered before.

Lemma 1.2.1. Let $G \in \mathscr{G}$ with $g(0) \in [0,1)$, and $\theta \in (0,1)^p$ with $\sum_{i=1}^p \theta_i < 1$. Then, for $k \in \mathbb{N}$, $\mathbb{E}_G \varepsilon_0^k < \infty$ if and only if $\mathbb{E}_{\nu_{\theta,G},\theta,G} X_0^k < \infty$.

Proof.
Of course, we only have to prove the 'only if'.
We give the proof for the case $g(0) > 0$, for $g(0) = 0$ the argument is almost similar. Under the assumption $\mathbb{E}_G \varepsilon_0^k < \infty$ the stationary distribution $\nu_{\theta,G}$ exists. Since Y is an irreducible, aperiodic Markov chain with stationary distribution $\nu_{\theta,G}$, we have $\mathscr{L}(Y_t \mid \mathbb{P}_{\delta_0,\theta,G}) \to \nu_{\theta,G}$. Hence (use the Portmanteau Lemma) we have,

$$\mathbb{E}_{\nu_{\theta,G},\theta,G} X_0^k \leq \liminf_{t\to\infty} \mathbb{E}_{\delta_0,\theta,G} X_t^k \leq \sup_{t\geq 0} \mathbb{E}_{\delta_0,\theta,G} X_t^k.$$

Thus it suffices to prove

$$\sup_{t\geq 0} \mathbb{E}_{\delta_0,\theta,G} X_t^k < \infty. \tag{1.5}$$

For $k = 1$ we have shown in the proof of Theorem 1.1 that $\mathbb{E}_{\delta_0,\theta,G} X_t$ is bounded in $t \in \mathbb{Z}_+$. Let $K \geq 2$. Suppose now that (1.5) holds for $k = 1,\ldots,K-1$. If we prove that then (1.5) also holds for $k = K$, then, by induction, the proof of the lemma is complete. So suppose (1.5) holds for $k = 1,\ldots,K-1$. If $Z_1,\ldots,Z_n$ are i.i.d. Bernoulli(θ) variables then we have, for $k \geq 2$, the bound (easily follows by elementary martingale theory; see, for example, Dharmadhikari et al. (1968)) $\mathbb{E}\left|\sum_{i=1}^n (Z_i - \theta)\right|^k \leq C_k n^{k/2}$, where the constant $C_k > 0$ only depends on k. Using that $\vartheta_i \circ X_{t-i}$, conditional on X_{t-i}, follows a Binomial(X_{t-i},θ) distribution, this yields the following inequality,

$$\begin{aligned}\mathbb{E}_{\delta_0,\theta,G} |\vartheta_i \circ X_{t-i} - \theta_i X_{t-i}|^K &= \mathbb{E}_{\delta_0,\theta,G}\left[\mathbb{E}_\theta\left[|\vartheta_i \circ X_{t-i} - \theta_i X_{t-i}|^K \mid X_{t-i}\right]\right] \\ &\leq C_K \mathbb{E}_{\delta_0,\theta,G} X_{t-i}^{K/2}.\end{aligned}$$

So, using the induction hypothesis ($K/2 \leq K-1$), we obtain

$$M = \|\varepsilon_0\|_K + p \sup_{t\in\mathbb{Z}_+,\, 1\leq i\leq p} \left(\mathbb{E}_{\delta_0,\theta,G} |\vartheta_i \circ X_{t-i} - \theta_i X_{t-i}|^K\right)^{1/K} < \infty,$$

where we denote $\|Z\|_K = \left(\mathbb{E}_{\delta_0,\theta,G}|Z|^K\right)^{1/K}$, i.e. the $\mathrm{L}_K(\mathbb{P}_{\delta_0,\theta,G})$ norm. We prove that, for all $s \geq 0$,

$$\|X_s\|_K \leq \frac{M}{1-\sum_{i=1}^p \theta_i}. \tag{1.6}$$

We have $\mathbb{E}_{\delta_0,\theta,G} X_{-i}^K = 0$ for $i = 1,\dots,p$, and $\mathbb{E}_{\delta_0,\theta,G} X_0^K = \|\varepsilon_0\|_K^K$. So (1.6) holds for $s = -p,\dots,0$. Let $t \in \mathbb{N}$. Suppose now that (1.6) holds for $-p \leq s \leq t-1$. We have,

$$\begin{aligned}
\|X_t\|_K &\leq \left\| X_t - \sum_{i=1}^p \theta_i X_{t-i} \right\|_K + \left\| \sum_{i=1}^p \theta_i X_{t-i} \right\|_K \\
&\leq \|\varepsilon_t\|_K + \sum_{i=1}^p \|\vartheta_i \circ X_{t-i} - \theta_i X_{t-i}\|_K + \sum_{i=1}^p \theta_i \|X_{t-i}\|_K \\
&\leq M + \sum_{i=1}^p \theta_i \frac{M}{1-\sum_{i=1}^p \theta_i} \leq \frac{M}{1-\sum_{i=1}^p \theta_i}.
\end{aligned}$$

Hence (1.6) holds for $s = t$. By induction we conclude that (1.6) holds for all $t \in \mathbb{Z}_+$. This completes the proof. □

In subsequent chapters we repeatedly have to deal with objects that are build of terms $f(X_{t-p},\dots,X_t)$, i.e. they depend on two consecutive observations on Y. Therefore we introduce the process $Z = (Z_t)_{t\geq 0}$, defined by

$$Z_t = (X_t,\dots,X_{t-p})', \quad t \in \mathbb{Z}_+. \tag{1.7}$$

It is easy to see that, in case $g(0) < 1$ and $\theta \in (0,1)^p$, Z is an irreducible, aperiodic Markov chain on the state space[2] $\mathscr{Z} = \mathrm{support}(\nu_{\theta,G} \otimes P^{\theta,G}) \subset \mathbb{Z}_+^{p+1}$. The next proposition contains some auxiliary results.

Proposition 1.2.1. Let $G \in \mathscr{G}$ with $g(0) \in (0,1)$, $\mu_G < \infty$, $\theta \in (0,1)^p$ with $\sum_{i=1}^p \theta_i < 1$. The following results hold.

1. Let ν a probability measure on $\mathbb{Z}_+^p$. If $h : \mathbb{Z}_+^{p+1} \to \mathbb{R}$ satisfies $\mathbb{E}_{\nu_{\theta,G},\theta,G} h^2(Z_0) < \infty$, then,

$$\frac{1}{\sqrt{n}} \sum_{t=0}^n \left(h(Z_t) - \mathbb{E}_{\theta,G}[h(Z_t) \mid \mathscr{F}_{t-1}]\right) \xrightarrow{d} \mathrm{N}(0,\sigma^2), \text{ under } \mathbb{P}_{\nu,\theta,G},$$

where σ^2 is given by,

$$\sigma^2 = \mathbb{E}_{\nu_{\theta,G},\theta,G} h^2(Z_0) - \mathbb{E}_{\nu_{\theta,G}} \left(\mathbb{E}_{\theta,G}[h(Z_0) \mid \mathscr{F}_{-1}]\right)^2.$$

[2]As usual, $\nu \otimes P^{\theta,G}$ denotes the joint distribution of $(X_{-p},\dots,X_0)$ under $\mathbb{P}_{\nu,\theta,G}$.

2. Let $C > 0$. The Markov chain Z is V_1-uniformly ergodic[3] for $V_1 : \mathbb{Z}_+^{p+1} \to [1,\infty)$ given by $V_1(Z_t) = 1 + C\sum_{i=0}^p X_{t-i}$. If also $\sigma_G^2 < \infty$ then Z is also V_2-uniformly ergodic for $V_2(Z_t) = 1 + C\left(\sum_{i=0}^p X_{t-i}\right)^2$.

3. Add the assumption $\sigma_G^2 < \infty$. Let ν a probability measure on $\mathbb{Z}_+^p$. Let K be a compact subset of $\mathbb{R}^k$. Let, for every $\kappa \in K$, $f(\cdot;\kappa) : \mathbb{Z}_+^{p+1} \to \mathbb{R}$ such that,

$$\sup_{\kappa\in K} \left| f(x_{-p},\dots,x_0;\kappa)\right| \leq C\left(\sum_{i=0}^p x_{-i}\right)^2,$$

for some constant $C > 0$, and for every $x_{-p},\dots,x_0 \in \mathbb{Z}_+$ the map $\kappa \mapsto f(x_{-p},\dots,x_0;\kappa)$ is continuous. Then we have, under $\mathbb{P}_{\nu,\theta,G}$

$$\sup_{\kappa\in K}\left|\frac{1}{n}\sum_{t=0}^n f(X_{t-p},\dots,X_t;\kappa) - \mathbb{E}_{\nu_{\theta,G},\theta,G} f(X_{-p},\dots,X_0;\kappa)\right| \xrightarrow{p} 0. \quad (1.8)$$

And, for $K \ni \kappa_n \to \kappa_0$,

$$\frac{1}{n}\sum_{t=0}^n f(X_{t-p},\dots,X_t;\kappa_n) \xrightarrow{p} \mathbb{E}_{\nu_{\theta,G},\theta,G} f(X_{-p},\dots,X_0;\kappa_0), \text{ under } \mathbb{P}_{\nu,\theta,G}. \quad (1.9)$$

4. Under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$ the β-mixing (also called: absolute regularity mixing) coefficients[4] of Z satisfy

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

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

5. Add the assumption $\mathbb{E}_G\varepsilon_1^3 < \infty$. Let $\mathbb{Z}_n$ denote the empirical process of Z, i.e. for $f : \mathbb{Z}_+^{p+1} \to \mathbb{R}$ satisfying $\mathbb{E}_{\nu_{\theta,G},\theta,G} f^2(Z_0) < \infty$,

$$\mathbb{Z}_n f = \frac{1}{\sqrt{n}}\sum_{t=0}^n \left(f(Z_t) - \mathbb{E}_{\nu_{\theta,G},\theta,G} f(Z_0)\right).$$

Let $\mathscr{F}$ be a collection of $\mathbb{R}$-valued functions on $\mathbb{Z}_+^{p+1}$ with, for some $C > 0$, $\sup_{f\in\mathscr{F}} |f(x_{-p},\dots,x_0)| \leq C(x_{-p} + \cdots + x_0)$, and such that its bracketing numbers[5] with respect to the $\mathrm{L}_2(\nu_{\theta,G} \otimes P^{\theta,G})$-norm, denoted by, $N_{[]}(\delta,\mathscr{F})$, $\delta > 0$, satisfy

$$\log N_{[]}(x,\mathscr{F}) = O(x^{-2\zeta}),$$

[3] Recall that the Markov chain Z is V-uniformly ergodic, with $V : \mathcal{Z} \to [1,\infty)$, if $\sup_{z\in\mathcal{Z}} \sup_{f:|f|\leq V} |\mathbb{E}_{\theta,G}[f(Z_t)|Z_0 = z] - \mathbb{E}_{\nu_{\theta,G},\theta,G} f(Z_0)|/V(z) \to 0$ as $t \to \infty$.

[4] For the definition see, for example, Davydov (1973) or Doukhan (1994, page 3 and pages 87-88).

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

with $\zeta \in (0,1)$. Then the process $\{\mathbb{Z}_n f \mid f \in \mathscr{F}\}$ weakly converges, under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, in $\ell^\infty(\mathscr{F})$ to a tight Gaussian process.

6. Define $V : \mathbb{Z}_+^p \to [1,\infty)$ by $V(x_{-1},\dots,x_{-p}) = 1+\sum_{i=1}^p a_i x_{-i}$, where $a_i = \theta_i + \cdots + \theta_p$ for $i = 1,\dots,p$. Let (θ_n, G_n) be a sequence with, for all $n \in \mathbb{N}$, $\theta_{n,i} > 0$ for all i, $\sum_{i=1}^p \theta_{n,i} < 1$, $g_n(0) \in (0,1)$ and $\mathbb{E}_{G_n}\varepsilon_0 < \infty$. Then

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

implies

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

Proof.
Proof of Part 1: Since Y is a positive recurrent Markov chain, this follows from Theorem 4.3.16 in Dacunha-Castelle and Duflo (1986) in case $\nu = \delta_y$. From this, the result extends to arbitrary ν by looking at pointwise convergence of characteristic functions, conditioning on the initial value and using dominated convergence.

Proof of Part 2: We consider the V_1-uniform ergodicity first. Introduce $V : \mathbb{Z}_+^{p+1} \to [1,\infty)$ given by $V(Z_t) = 1+\sum_{i=1}^{p+1} a_i X_{t+1-i}$, with $a_i = \theta_i + \cdots + \theta_p$ for $i = 1,\dots,p$ and $a_{p+1} = (1-a_1)/2$. If we verify that there exists a constant $\delta > 0$ such that we have, for all $z_{t-1} \in \mathcal{Z}$, except for some finite set, the inequality

$$\mathbb{E}_{\theta,G}\left[V(Z_t) \mid Z_{t-1} = z_{t-1} = (x_{t-1},\dots,x_{t-p-1})'\right] - V(z_{t-1}) \le -\delta V(z_{t-1}),$$

i.e. that a Foster-Lyanupov drift criterium holds, we obtain the first result from Meyn and Tweedie (1994, Theorem 16.01). Let $0 < \delta < (1-a_1)(\min_{j=1,\dots,p}\theta_j)/2 < 1$. We have,

$$\begin{aligned}(1-\delta)V(z_{t-1}) - \mathbb{E}_{\theta,G}[V(Z_t) \mid Z_{t-1} = z_{t-1}] &= a + (1-\delta)a_{p+1}x_{t-p-1} \\ &\quad + \sum_{i=1}^p [(1-\delta)a_i - c_i]\,x_{t-i},\end{aligned}$$

for some constant a and c_i given by $c_i = a_1\theta_i + a_{i+1}$, $i = 1,\dots,p$. We show that, for $i = 1,\dots,p$, $(1-\delta)a_i - c_i > 0$ which implies that

$$(1-\delta)V(z_{t-1}) - \mathbb{E}_{\theta,G}[V(Z_t) \mid Z_{t-1} = z_{t-1}] > 0,$$

outside a finite set. We have, for $i = 1,\dots,p-1$, (use $a_i = \theta_i + a_{i+1}$),

$$(1-\delta)a_i - a_1\theta_i - a_{i+1} = -\delta a_i + (1-a_1)\theta_i > -\delta + (1-a_1)\min_{j=1,\dots,p}\theta_j > 0,$$

and $(1-\delta)a_p - a_1\theta_p - a_{p+1} > -\delta + 0.5\theta_p(1-a_1)/2$, which concludes the proof of the V_1-uniform ergodicity. Next we prove the V_2-uniform ergodicity. Introduce $\tilde{V}:\mathbb{Z}_+^{p+1}\to[1,\infty)$ given by $\tilde{V}(Z_t) = 1+\left(\sum_{i=1}^{p+1} a_i X_{t+1-i}\right)^2$, with $a_i = \theta_i + \cdots + \theta_p$ for $i = 1,\ldots,p$, and $a_{p+1} = (1-a_1^2)\left(\min_j \theta_j\right)^2/4$. Let $0<\delta<a_{p+1}/4$. After some calculus we find,

$$\mathbb{E}_{\theta,G}\left[\tilde{V}(Z_t)\mid Z_{t-1}=z_{t-1}\right] = a + \sum_{i=1}^{p}\alpha_i X_{t-i} + \sum_{i=1}^{p}\sum_{j=1}^{p}\beta_{ij}X_{t-i}X_{t-j},$$

where $\beta_{ij} = a_1^2\theta_i\theta_j + \theta_i a_1 a_{j+1} + \theta_j a_1 a_{i+1} + a_{i+1}a_{j+1}$ for $i,j=1,\ldots,p$. We show that $(1-\delta)a_i a_j - \beta_{ij} > 0$ for $i,j = 1,\ldots,p$. Using that, for $i,j=1,\ldots,p-1$,

$$a_i a_j = \theta_i\theta_j + \theta_i(\theta_{j+1}+\cdots+\theta_p) + \theta_j(\theta_{i+1}+\cdots+\theta_p) + a_{i+1}a_{j+1},$$

and $a_i a_j < (\theta_i + \cdots + \theta_p)$, we obtain, for $i,j=1,\ldots,p-1$,

$$(1-\delta)a_i a_j - \beta_{ij} > (1-a_1^2)\theta_i\theta_j - \delta a_i a_j \geq (1-a_1^2)(\min_j \theta_j)^2 - \delta > 0.$$

For $i = 1,\ldots,p-1$ we have $(1-\delta)a_i a_p - \beta_{ip} > (1-a_1^2)\theta_i\theta_p - a_{p+1} - \delta > 0$, and $(1-\delta)a_p^2 - \beta_{pp} > -\delta + (1-a_1^2)\theta_p^2 - a_{p+1}(2a_1\theta_p + a_{p+1}) > 0$. We conclude that, outside a finite set, we have

$$(1-\delta)\tilde{V}(z_{t-1}) - \mathbb{E}_{\theta,G}\left[\tilde{V}(Z_t)\mid Z_{t-1}=z_{t-1}\right] > 0.$$

So another application of the drift criterion in Meyn and Tweedie (1994, Theorem 16.01) shows that Z is V_2-uniformly ergodic.

<u>Proof of Part 3:</u> Since Z is V_2-uniformly ergodic, a combination of Part 2 with Meyn and Tweedie (1994, Theorem 16.0.1) yields, for a constant $M>0$ and $\rho\in(0,1)$, such that for all $t\in\mathbb{Z}_+$,

$$\sup_{z\in\mathcal{Z}} \frac{\sup_{f:|f|\leq V_2}\left|\mathbb{E}_{\theta,G}[f(Z_t)|Z_0=z] - \mathbb{E}_{\nu_{\theta,G},\theta,G}f(Z_0)\right|}{V_2(z)} \leq M\rho^t.$$

Using that $V_2(Z_0)$ is $\mathbb{P}_{\nu_{\theta,G},\theta,G}$-integrable (by Lemma 1.1) we easily obtain

$$\lim_{t\to\infty}\sup_{f:|f|\leq V_2}\left|\mathbb{E}_{\nu,\theta,G}f(Z_t) - \mathbb{E}_{\nu_{\theta,G},\theta,G}f(Z_0)\right| = 0. \tag{1.10}$$

Display (1.10) yields, for $M>0$,

$$\lim_{t\to\infty}\mathbb{E}_{\nu,\theta,G}V_2(Z_t)1_{\{V_2(Z_t)\geq M\}} = \mathbb{E}_{\nu_{\theta,G},\theta,G}V_2(Z_0)1_{\{V_2(Z_0)\geq M\}}.$$

Hence we obtain, for $M>0$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n}\mathbb{E}_{\nu,\theta,G}V_2(Z_t)1_{\{V_2(Z_t)\geq M\}} = \mathbb{E}_{\nu_{\theta,G},\theta,G}V_2(Z_0)1_{\{V_2(Z_0)\geq M\}}.$$

Dominated convergence now yields,

$$\lim_{M\to\infty}\lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n}\mathbb{E}_{\nu,\theta,G}V_2(Z_t)1_{\{V_2(Z_t)\geq M\}}=0.$$

Hence Assumption DM in Andrews (1992) is satisfied. Assumption TSE-1B in that paper is also satisfied, by the compactness of K, by the continuity of $\kappa\mapsto f(z;\kappa)$, and because we have, from (1.10),

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n}\mathbb{P}_{\nu,\theta,G}\{Z_t\in A\}=\nu_{\theta,G}\otimes P^{\theta,G}(A),\quad \text{for } A\subset\mathbb{Z}_+^{p+1}.$$

A combination of the law of large numbers for Markov chains (see, for example, Dacunha-Castelle and Duflo (1986, Theorem 4.3.15)) with Theorem 4 in Andrews (1992) now yields,

$$\sup_{\kappa\in K}\left|\frac{1}{n}\sum_{t=0}^{n}f(X_{t-p},\dots,X_t;\kappa)-\mathbb{E}_{\nu,\theta,G}f(X_{-p},\dots,X_0;\kappa)\right|\xrightarrow{p}0,\text{ under }\mathbb{P}_{\nu,\theta,G}.$$

This yields (1.8), since (1.10) yields,

$$\sup_{\kappa\in K}\left|\mathbb{E}_{\nu,\theta,G}f(X_{t-p},\dots,X_t;\kappa)-\mathbb{E}_{\nu_{\theta,G},\theta,G}f(X_{-p},\dots,X_0;\kappa)\right|\to 0\text{ as }t\to\infty,$$

Display (1.9) follows by dominated convergence.

Proof of Part 4: Let Q^n denote the n-step transition-operator of Z (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

$$\|Q^n(z,\cdot)-\nu_{\theta,G}\otimes P^{\theta,G}\|_{\mathrm{TV}}\leq A(z)\rho^n,\quad z\in\mathcal{Z},\tag{1.11}$$

for some $0<\rho<1$, where $\|\cdot\|_{\mathrm{TV}}$ the total variational norm of a signed measure. By Part 2 Z is V_1-uniformly ergodic. Meyn and Tweedie (1994, Theorem 16.0.1) now yields, for a constant $M>0$ and $\rho\in(0,1)$, such that for all $t\in\mathbb{Z}_+$,

$$\sup_{z\in\mathcal{Z}}\frac{\sup_{f:|f|\leq V_2}\left|\mathbb{E}_{\theta,G}[f(Z_t)|Z_0=z]-\mathbb{E}_{\nu_{\theta,G},\theta,G}f(Z_0)\right|}{V_1(z)}\leq M\rho^t.$$

Since $\mathbb{E}_{\nu_{\theta,G},\theta,G}V_1(Z_0)<\infty$ (by Lemma 1.1) and $V\geq 1$ (1.11) immediately follows.

Proof of Part 5: this follows from Part (1) and Doukhan et al. (1995, Theorem 1, Application 4, and Display (2.16)). In their setup proceed as follows. Take $r=3/2$, notice that, using Markov's inequality and $\mathbb{E}_{\nu_{\theta,G}}X_0^3<\infty$ (by Lemma 1.1), the envelope belongs to $\Lambda_3(P) = \Lambda_{x\sqrt{x}}(P)$. Next, note that (2.16) in Doukhan et al. (1995) is satisfied, since we have $\sum_{n=1}^{\infty} n^{-1/2}\beta(n)^{(1-\zeta)(r-1)/(2r)} < \infty$, by Part 4.

Proof of Part 6: Analogous to the proof of Part 2 it follows that the Markov chain Y on $\mathbb{Z}_+^p$ is V-uniformly ergodic for $V(Y_t) = 1 + \sum_{i=1}^p a_i X_{t-i}$, $a_i = \theta_i + \cdots + \theta_p$. An application of Kartashov (1985, Theorem B) yields that Y is strongly stable in this norm, i.e. that Part 6 holds. □

Let us briefly comment on this proposition. Part 1 is stated for easy reference; its purpose is clear. In Chapter 2, where we discuss the LAN-property for parametric INAR models, we encounter remainder terms which we will handle with Part 3. We proved this uniform law of large numbers by exploiting a high level result of Andrews (1992). The V-uniform ergodicity, Part 2, which we prove by a drift criterium, appears to be new to the literature. Using this property, Part 4, Part 5 and Part 6 follow quite easily from the literature. Part 5 is used in Chapter 3 to demonstrate weak convergence of the infinite-dimensional part a 'score-process'. And Part 6 shows that, in appropriate topologies, the stationary distribution $\nu_{\theta,G}$ is a continuous mapping of (θ, G).

2 Parametric stationary INAR(p) models

This chapter considers parametric INAR(p) models: G belongs to a parametric class of distributions, say $(G_\alpha | \alpha \in A \subset \mathbb{R}^q)$. Estimators of the parameters are provided by several authors. For $p = 1$ and $G_\alpha = \text{Poisson}(\alpha)$, Franke and Seligmann (1993) analyzed maximum likelihood. 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 (2004) proposed a frequency domain based estimator of θ, and Silva and Silva (2006) considered a Yule-Walker estimator. Jung et al. (2005) analyzed, by a Monte Carlo study, the finite sample behavior of several estimators for the case $p = 1$. Zheng et al. (2006) analyzed random coefficient INAR(p) processes. And Enciso-Mora et al. (2006) and Neal and Subba Rao (2007) considered MCMC estimation. In this chapter we are interested in asymptotic efficient estimation of the parameters in an INAR(p) model. Maximum likelihood is, in general, considered to be computationally unattractive, since the transition-densities are convolutions of $p+1$ distributions. The main result of this chapter is that parametric INAR models enjoy the Local Asymptotic Normality property. A key step, which makes the analysis tractable, is a certain conditional expectation representation of the transition-scores. This representation is motivated by an information-loss interpretation of the model. As a consequence of the LAN-property, we obtain an efficient estimator of (θ, α) if there is available a $\sqrt{n}$-consistent estimator. We prove that such an initial estimator always exists. This yields a computationally attractive and efficient estimator.

2.1 Local Asymptotic Normality

We always restrict ourselves to the stationary parameter regime, i.e., $\theta \in (0,1)^p$ with $\sum_{i=1}^p \theta_i < 1$ (see Chapter 4 for the asymptotic structure of an INAR(1) model at the boundary of the parameter space). In a first model, the immigration-distribution G and the initial distribution ν are completely known. Observing $(X_{-p},\ldots,X_n)$ leads to the following sequence of statistical experiments

$$\mathscr{E}_1^{(n)}(\nu,G)=\left(\mathbb{Z}_+^{n+1+p},2^{\mathbb{Z}_+^{n+1+p}},\left(\mathbb{P}_{\nu,\theta,G}^{(n)}\mid\theta\in\Theta\right)\right),\quad n\in\mathbb{Z}_+,$$

where the initial distribution ν and the immigration distribution $G\in\mathscr{G}$ are fixed, $\Theta=\{\theta\in(0,1)^p\mid\sum_{i=1}^p\theta_i<1\}$, and $\mathbb{P}_{\nu,\theta,G}^{(n)}$ denotes the law of $(X_{-p},\ldots,X_n)$ on the measurable space $\left(\mathbb{Z}_+^{n+1+p},2^{\mathbb{Z}_+^{n+1+p}}\right)$ under $\mathbb{P}_{\nu,\theta,G}$. In this model G is completely known, but we also want to consider the case where G belongs to a parametric model, for example, $G=\text{Poisson}(\alpha)$. So let $A\subset\mathbb{R}^q$ and $\mathscr{G}_A=(G_\alpha)_{\alpha\in A}$ be a family of elements in $\mathscr{G}$, such that $\alpha\mapsto G_\alpha$ is sufficiently smooth (this will be made precise later). We then consider the sequence of experiments, induced by observing $(X_{-p},\ldots,X_n)$,

$$\mathscr{E}_2^{(n)}(\nu,\mathscr{G}_A)=\left(\mathbb{Z}_+^{n+1+p},2^{\mathbb{Z}_+^{n+1+p}},\left(\mathbb{P}_{\nu,\theta,\alpha}^{(n)}\mid\theta\in\Theta,\alpha\in A\right)\right),\quad n\in\mathbb{Z}_+,$$

where, for notational convenience, we abbreviate G_α in sub- and superscripts by α. In particular, ν_{θ,G_α} is denoted by $\nu_{\theta,\alpha}$. In this section we prove the LAN-property for the sequence of experiments $\mathscr{E}_2^{(n)}(\nu,\mathscr{G}_A)$, $n\in\mathbb{Z}_+$, immediately implying the LAN-property for the sequence of experiments $\mathscr{E}_1^{(n)}(\nu,G)$, $n\in\mathbb{Z}_+$.

Let $\mathscr{G}_A=(G_\alpha|\alpha\in A)$ be a parametric family of innovation distributions, where A is an open, convex subset of $\mathbb{R}^q$ such that,

(A1) the support of G_α does not depend on α and we have $0<g_\alpha(0)<1$;

(A2) for all $e\in\mathbb{Z}_+$ and $\alpha\in A$, the expressions,

$$\begin{aligned}h_\alpha(e)&=\frac{\partial}{\partial\alpha}\log\left(g_\alpha(e)\right)1_{(0,1]}\left(g_\alpha(e)\right)\in\mathbb{R}^q,\\ \dot h_\alpha(e)&=\frac{\partial^2}{\partial\alpha^T\partial\alpha}\log\left(g_\alpha(e)\right)1_{(0,1]}\left(g_\alpha(e)\right)\in\mathbb{R}^{q\times q},\end{aligned}$$

are defined and, for all $e\in\mathbb{Z}_+$, they are continuous in α;

(A3) for every $(\theta,\alpha)\in\Theta\times A$, there exists $\delta>0$ and a constant $C>0$ such that

$$\sup_{(\tilde\theta,\tilde\alpha):|(\tilde\theta,\tilde\alpha)-(\theta,\alpha)|<\delta}\mathbb{E}_{\tilde\theta,\tilde\alpha}\left[|h_{\tilde\alpha}(\varepsilon_0)|^2\mid X_0,\ldots,X_{-p}\right]\le C\left(\sum_{i=0}^p X_{-i}\right)^2,\tag{2.1}$$

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

$$\sup_{(\tilde\theta,\tilde\alpha):|(\tilde\theta,\tilde\alpha)-(\theta,\alpha)|<\delta}\mathbb{E}_{\tilde\theta,\tilde\alpha}\left[\left|\dot h_{\tilde\alpha,ij}(\varepsilon_0)\right|\mid X_0,\ldots,X_{-p}\right]\le C\left(\sum_{i=0}^p X_{-i}\right)^2,\tag{2.2}$$

where $\dot h_{\alpha,ij}(e)$ is the (i,j)-entry of the matrix $\dot h_\alpha(e)$;

(A4) the information-equality $\mathbb{E}_\alpha h_\alpha h_\alpha^T(\varepsilon_0) = -\mathbb{E}_\alpha \dot{h}_\alpha(\varepsilon_0)$ is satisfied, and the $q \times q$ matrix $\mathbb{E}_\alpha h_\alpha h_\alpha^T(\varepsilon_0)$ is non-singular and continuous in α;

(A5) $\mathbb{E}_\alpha \varepsilon_0^2 < \infty$ for $\alpha \in A$;

(A6) $G_\alpha = G_{\alpha'}$ implies $\alpha = \alpha'$.

Remark 2. Assumption (A1) is necessary to make sure that the INAR process can reach state 0. This is a reasonable assumption for virtually all applications. From a technical point of view, this assumption will help us to prove invertibility of the Fisher information.

Remark 3. It is well-known that Assumptions (A2) and (A4) ensure that $\mathcal{G}_A$ is differentiable in quadratic mean with score $h_\alpha(\varepsilon_0)$ (see, for example, Lemma 7.6 in Van der Vaart (2000)) and consequently $\mathbb{E}_\alpha h(\varepsilon_0) = 0$, see the proof of Theorem 7.2 in Van der Vaart (2000).

Remark 4. Assumptions (A1)-(A6) are of the Cramér-type. Conditions (2.1) and (2.2) in Assumption (A3) are rather awkward. A simple sufficient condition is given by $|h_{\alpha,i}(e)| \le a_\alpha + c_\alpha e$ and $|\dot{h}_{\alpha,ij}| \le b_\alpha + d_\alpha e^2$ for $a_\alpha, b_\alpha, c_\alpha$ and d_α that are (locally) bounded in α. Now it is easy to see that the (in the literature often-used) example $A = (0,\infty)$ and $G_\alpha = \text{Poisson}(\alpha)$ satisfies the conditions above. We note that in (2.1) and (2.2) the upper-bound $C\left(\sum_{i=0}^p X_{-i}\right)^2$ can be replaced by $\mathbb{P}_{\nu_{\theta,\alpha},\theta,\alpha}$-integrable variables $M_1^{\theta,\alpha}$ and $M_2^{\theta,\alpha}$ in case the initial distribution ν has finite support (in that case ergodicity instead of V_2-uniform ergodicity (see Proposition 1.2.1.2) suffices).

To see that the sequence of experiments $(\mathcal{E}_2^{(n)}(\nu,\mathcal{G}_A))_{n\in\mathbb{Z}_+}$ has the LAN-property, we need to determine the asymptotic behavior of a localized log-likelihood ratio. To that end we first write down the likelihood. By the p-th order Markov-structure, the likelihood is given by,

$$L_n(\theta,\alpha \mid X_{-p},\ldots,X_n) = \nu\{X_{-1},\ldots,X_{-p}\} \prod_{t=0}^{n} P^{\theta,\alpha}_{(X_{t-1},\ldots,X_{t-p}),X_t}.$$

Since the likelihood is extremely smooth in (θ,α), it seems to be appropriate to establish the LAN-property directly, using a Taylor-expansion. This is the path we take. To obtain useful expressions for the transition-scores for θ and α, we briefly discuss how we can view upon the model as an information-loss model. Suppose that, instead of just observing $X_{-p},\ldots,X_n$, we would also be able to observe $\vartheta_i \circ X_{t-i}$, $i = 1,\ldots,p$, $t = 0,\ldots,n$. Then $\varepsilon_t = X_t - \sum_{i=1}^p \vartheta_i \circ X_{t-i}$ also belongs to the information set at time t, just as in the classical AR(p) model. In our model, with only observations on $X_{-p},\ldots,X_n$, this does not hold true; there is loss of information. The 'information-loss principle', see for example Le Cam and Yang (1988) or Bickel et al. (1998, Proposition A.5.5), suggests that the transition-score for θ_i in the model where we only observe $X_{t-p},\ldots,X_t$, equals the conditional expectation, given $X_{t-p},\dots,X_t$, of the transition-score for θ_i in the model with also observations on $\vartheta_i \circ X_{t-i}$, $i=1,\dots,p$. It is not difficult to see that the transition-score for θ_i in the model with the additional observations $\vartheta_i \circ X_{t-i}$ is nothing but the score of a $\mathrm{Bin}_{X_{t-i},\theta_i}$ distribution. Recall that the score of a $\mathrm{Bin}_{x,\theta}$ distribution is given by, for $\theta \in (0,1)$,

$$\dot{s}_{x,\theta}(k) = \left(\frac{\partial}{\partial\theta}\log \mathrm{b}_{x,\theta}(k)\right) 1_{(0,1]}(\mathrm{b}_{x,\theta}(k)) = \frac{k-\theta x}{\theta(1-\theta)} 1_{\{0,\dots,x\}}(k), \quad x \in \mathbb{Z}_+. \tag{2.3}$$

Hence, the information-loss structure suggests that the transition-score for θ_i in our model equals,

$$\mathbb{E}_{\theta,\alpha}\left[\dot{s}_{X_{t-i},\theta_i}(\vartheta_i \circ X_{t-i}) \mid X_t,\dots,X_{t-p}\right].$$

Similarly, the transition-score for α is conjectured to be equal to,

$$\mathbb{E}_{\theta,\alpha}\left[h_\alpha(\varepsilon_t) \mid X_t,\dots,X_{t-p}\right].$$

One way to make this reasoning precise, is to show that the model is differentiable in quadratic mean with respect to (θ,α). Instead, since the model is extremely smooth, we may derive the transition-scores directly by calculating the partial derivatives of $\log P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}$ with respect to both θ and α. It is easy to see that, for $x_{t-p},\dots,x_t \in \mathbb{Z}_+$, $i=1,\dots,p$, $\theta\in(0,1)^p$, we have,

$$\begin{aligned}
\dot{\ell}_{\theta,i}(x_{t-p},\dots,x_{t-1},x_t;\theta,\alpha) &= \frac{\partial}{\partial\theta_i}\log\left(P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}\right) 1_{(0,1]}\left(P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}\right)\\
&= \frac{\sum_k \dot{s}_{x_{t-i},\theta_i}(k)\,\mathrm{b}_{x_{t-i},\theta_i}(k)\left(G_\alpha *_{j\neq i} \mathrm{Bin}_{x_{t-j},\theta_j}\right)\{x_t-k\}}{P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}} 1_{(0,1]}\left(P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}\right)\\
&= \mathbb{E}_{\theta,\alpha}\left[\dot{s}_{X_{t-i},\theta_i}(\vartheta_i\circ X_{t-i}) \mid X_t = x_t,\dots,X_{t-p}=x_{t-p}\right],
\end{aligned} \tag{2.4}$$

where we put $\mathbb{E}_{\theta,\alpha}\left[\cdot \mid X_t = x_t,\dots,X_{t-p}=x_{t-p}\right]=0$ if $\mathbb{P}_{\nu,\theta,\alpha}\{X_{t-p}=x_{t-p},\dots,X_t = x_t\}=0$. Similarly we find, for $x_{t-p},\dots,x_t\in\mathbb{Z}_+$, and $i=1,\dots,q$,

$$\begin{aligned}
\dot{\ell}_{\alpha,i}(x_{t-p},\dots,x_{t-1},x_t;\theta,\alpha) &= \frac{\partial}{\partial\alpha_i}\log\left(P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}\right) 1_{(0,1]}\left(P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}\right)\\
&= \frac{\sum_e h_{\alpha,i}(e)\,g_\alpha(e)\left(*_{j=1,\dots,p}\mathrm{Bin}_{x_{t-j},\theta_j}\right)\{x_t-e\}}{P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}} 1_{(0,1]}\left(P^{\theta,\alpha}_{(x_{t-1},\dots,x_{t-p}),x_t}\right)\\
&= \mathbb{E}_{\theta,\alpha}\left[h_{\alpha,i}(\varepsilon_t) \mid X_t = x_t,\dots,X_{t-p}=x_{t-p}\right].
\end{aligned} \tag{2.5}$$

For the case $p=1$ and $G_\alpha = \mathrm{Poisson}(\alpha)$, representation (2.4) was also found by Freeland and McCabe (2004). Although we established (2.4) and (2.5) also by direct calculations, we stress that the structure is due to the information-loss interpretation of the model. From the representation it immediately follows that the score is a martingale. If we would not have the representations available, this would be a tedious matter. A Taylor-expansion of the localized log-likelihood ratio, a martingale central limit theorem, and a law of large numbers now suggest that the sequence of experiments $(\mathscr{E}_2^{(n)}(\nu,\mathscr{G}_A))_{n\in\mathbb{Z}_+}$ has the LAN-property. The following theorem gives the precise result.

Theorem 2.1. *Let $\mathscr{G}_A \subset \mathscr{G}$ satisfy Assumptions (A1)-(A5), ν a probability measure on $\mathbb{Z}_+^p$, and $(\theta,\alpha) \in \Theta \times A$. Then the sequence of experiments $(\mathscr{E}_2^{(n)}(\nu,\mathscr{G}_A))_{n\in\mathbb{Z}_+}$ has the LAN-property in (θ,α), i.e. for every $u=(u_1,u_2) \in \mathbb{R}^p \times \mathbb{R}^q$ the following expansion holds,*

$$\log \frac{\mathrm{d}\mathbb{P}^{(n)}_{\nu,\theta+u_1/\sqrt{n},\alpha+u_2/\sqrt{n}}}{\mathrm{d}\mathbb{P}^{(n)}_{\nu,\theta,\alpha}}(X_{-p},\dots,X_n) = \log \frac{L_n\left(\theta+\frac{u_1}{\sqrt{n}},\alpha+\frac{u_2}{\sqrt{n}} \mid X_{-p},\dots,X_n\right)}{L_n\left(\theta,\alpha \mid X_{-p},\dots,X_n\right)}$$
$$= u^T S_n - \frac{1}{2}u^T J u + R_n,$$

where the score (also called central sequence),

$$S_n = S_n(\theta,\alpha) = \frac{1}{\sqrt{n}}\sum_{t=0}^{n}\begin{pmatrix}\dot{\ell}_\theta(X_{t-p},\dots,X_t;\theta,\alpha)\\ \dot{\ell}_\alpha(X_{t-p},\dots,X_t;\theta,\alpha)\end{pmatrix}, \tag{2.6}$$

satisfies

$$S_n \xrightarrow{d} \mathrm{N}(0,J), \text{ under } \mathbb{P}_{\nu,\theta,\alpha}. \tag{2.7}$$

The Fisher-information defined by,

$$J = J(\theta,\alpha) = \begin{pmatrix} J_\theta & J_{\theta,\alpha}\\ J_{\alpha,\theta} & J_\alpha\end{pmatrix}$$
$$= \begin{pmatrix}\mathbb{E}_{\nu_{\theta,\alpha},\theta,\alpha}\dot{\ell}_\theta\dot{\ell}_\theta^T(X_{-p},\dots,X_0;\theta,\alpha) & \mathbb{E}_{\nu_{\theta,\alpha},\theta,\alpha}\dot{\ell}_\theta\dot{\ell}_\alpha^T(X_{-p},\dots,X_0;\theta,\alpha)\\ \mathbb{E}_{\nu_{\theta,\alpha},\theta,\alpha}\dot{\ell}_\alpha\dot{\ell}_\theta^T(X_{-p},\dots,X_0;\theta,\alpha) & \mathbb{E}_{\nu_{\theta,\alpha},\theta,\alpha}\dot{\ell}_\alpha\dot{\ell}_\alpha^T(X_{-p},\dots,X_0;\theta,\alpha)\end{pmatrix},$$

is non-singular, and $R_n = R_n(u,\theta,\alpha) \xrightarrow{p} 0$ under $\mathbb{P}_{\nu,\theta,G}$.

Remark 5. If one wants to draw the initial value, $(X_{-1},\dots,X_{-p})'$, according to the stationary distribution, one considers the sequence of experiments $\tilde{\mathscr{E}}_2^{(n)}(\mathscr{G}_A) = (\mathbb{Z}_+^{n+1+p}, 2^{\mathbb{Z}_+^{n+1+p}}, (\mathbb{P}^{(n)}_{\nu_{\theta,\alpha},\theta,\alpha} \mid \theta \in \Theta, \alpha \in A))$, $n \in \mathbb{Z}_+$. If the conditions in Theorem 2.1 are satisfied and if the initial value is negligible: for all $u=(u_1,u_2) \in \mathbb{R}^p \times \mathbb{R}^q$ we have $\nu_{\theta+u_1/\sqrt{n},\alpha+u_2/\sqrt{n}}\{X_{-1},\dots,X_{-p}\} - \nu_{\theta,\alpha}\{X_{-1},\dots,X_{-p}\} = o(\mathbb{P}_{\nu_{\theta,\alpha},\theta,\alpha};1)$, then we also have the LAN-property for $(\tilde{\mathscr{E}}_2^{(n)}(\mathscr{G}_A))_{n\in\mathbb{Z}_+}$. In case $p=1$, $A=(0,\infty)$, and $G_\alpha = \text{Poisson}(\alpha)$, it is easy to see, using generating functions, that $\nu_{\theta,\alpha} = \text{Poisson}(\alpha/(1-\theta))$. For this case the negligibility of the initial value readily follows. See the proof of Lemma 3.3.1 how to verify, in general, the negligibility.

Remark 6. For the case $p = 1$ and $G_\alpha = \text{Poisson}(\alpha)$, the non-singularity of J was obtained, via direct calculation, by Franke and Seligmann (1993).

Proof.
Using Assumption (A5) on $\mathcal{G}_A$ and Lemma 1.2.1 we obtain $\mathbb{E}_{\nu_0,\theta,\alpha} X_0^2 < \infty$, where, for notational convenience, we denote $\nu_0 = \nu_{\theta,\alpha}$.

Expansion of log-likelihood ratio: Let $u = (u_1, u_2) \in \mathbb{R}^p \times \mathbb{R}^q$, $u \neq 0$ (the case $u = 0$ is, of course, trivial). Since $\Theta \times A$ is open and convex we obtain, by Taylor's theorem,

$$\log \frac{L_n\left(\theta + \frac{u_1}{\sqrt{n}}, \alpha + \frac{u_2}{\sqrt{n}} \mid X_{-p}, \dots, X_n\right)}{L_n\left(\theta, \alpha \mid X_{-p}, \dots, X_n\right)} = u^T S_n(\theta, \alpha) - \frac{1}{2} u^T J_n(\tilde{\theta}_n, \tilde{\alpha}_n) u, \quad (2.8)$$

where $(\tilde{\theta}_n, \tilde{\alpha}_n)$ is a random point on the line-segment between (θ, α) and $(\theta + u_1/\sqrt{n}, \alpha + u_2/\sqrt{n})$ and

$$J_n(\theta, \alpha) = -\frac{1}{\sqrt{n}} \frac{\partial}{\partial(\theta, \alpha)^T} S_n(\theta, \alpha). \quad (2.9)$$

First, we give some auxiliary calculations in Part 0. Part 1 shows that $S_n(\theta, \alpha) \xrightarrow{d} \mathrm{N}(0, J)$ under $\mathbb{P}_{\nu,\theta,\alpha}$, in Part 2 we prove that $J_n(\tilde{\theta}_n, \tilde{\alpha}_n) \xrightarrow{p} J$ under $\mathbb{P}_{\nu,\theta,\alpha}$, and, finally, in Part 3 we prove the non-singularity of J.

Part 0: auxiliary calculations In this part we show that certain expressions are integrable, which is needed in Step 1 and Step 2.
It is easy to see that, for $\theta \in (0,1)$, $\ell \in \mathbb{N}$, we have

$$\frac{\partial^\ell}{\partial \theta^\ell} \log \mathrm{b}_{x,\theta}(k) = (-1)^{\ell+1} (\ell - 1)! \frac{k}{\theta^\ell} - (\ell - 1)! \frac{x - k}{(1-\theta)^\ell},$$

and hence

$$\left| \frac{\partial^\ell}{\partial \theta^\ell} \log \mathrm{b}_{x,\theta}(k) \right| \leq (\ell - 1)! x \left(\frac{1}{(1-\theta)^\ell} \vee \frac{1}{\theta^\ell} \right) \leq (\ell - 1)! \frac{x}{(1-\theta)^\ell \theta^\ell}. \quad (2.10)$$

For notational convenience we denote

$$\dot{s}_{x,\theta}(k) = \frac{\partial}{\partial \theta} \log \mathrm{b}_{x,\theta}(k), \text{ and } \ddot{s}_{x,\theta}(k) = \frac{\partial^2}{\partial \theta^2} \log \mathrm{b}_{x,\theta}(k).$$

From (2.4) and (2.10) we obtain the bound

$$\left| \dot{\ell}_{\theta,i}(X_{-p}, \dots, X_0; \theta, \alpha) \right| \leq \frac{1}{\theta_i (1 - \theta_i)} X_{-i}. \quad (2.11)$$

From Assumption (A3) on $\mathcal{G}_A$ we obtain $\delta > 0$. If necessary, decrease δ such that the ball round θ with radius δ is a subset of Θ. Of course, this has no influence on the validity of (2.1) and (2.2). Using (2.11) and Cauchy-Schwarz, we obtain, for $i, j = 1, \dots, p$,

$$\mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha}):|(\tilde{\theta},\tilde{\alpha})-(\theta,\alpha)|<\delta} \left|\dot{\ell}_{\theta,i}\dot{\ell}_{\theta,j}(X_{-p},\dots,X_0;\tilde{\theta},\tilde{\alpha})\right|$$
$$\leq M_\theta \sqrt{\mathbb{E}_{\nu_0,\theta,\alpha} X_{-i}^2 \mathbb{E}_{\nu_0,\theta,\alpha} X_{-j}^2} = M_\theta \mathbb{E}_{\nu_0,\theta,\alpha} X_0^2 < \infty, \tag{2.12}$$

for some constant $M_\theta > 0$. Using (2.1) from Assumption (A3) on $\mathcal{G}_A$ we obtain, for $i, j = 1, \dots, q$,

$$\mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha}):|(\tilde{\theta},\tilde{\alpha})-(\theta,\alpha)|<\delta} |\dot{\ell}_{\alpha,i}\dot{\ell}_{\alpha,j}(X_{-p},\dots,X_0;\tilde{\theta},\tilde{\alpha})|$$
$$\leq \mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha})\in B_\delta} \sqrt{\mathbb{E}_{\tilde{\theta},\tilde{\alpha}}\left[|h_{\tilde{\alpha},i}(\varepsilon_0)|^2 \mid X_0,\dots,X_{-p}\right] \mathbb{E}_{\tilde{\theta},\tilde{\alpha}}\left[|h_{\tilde{\alpha},j}(\varepsilon_0)|^2 \mid X_0,\dots,X_{-p}\right]}$$
$$\leq C\mathbb{E}_{\nu_0,\theta,\alpha}\left(\sum_{i=0}^{p} X_{-i}\right)^2 < \infty, \tag{2.13}$$

where $B_\delta = \{(\tilde{\theta},\tilde{\alpha}) \mid |(\tilde{\theta},\tilde{\alpha}) - (\theta,\alpha)| < \delta\}$. Using Cauchy-Schwarz, (2.10) and (2.1) from Assumption (A3) on $\mathcal{G}_A$ we also have, for $i = 1, \dots, p$, $j = 1, \dots, q$,

$$\mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha}):|(\tilde{\theta},\tilde{\alpha})-(\theta,\alpha)|<\delta} \left|\dot{\ell}_{\theta,i}\dot{\ell}_{\alpha,j}(X_{-p},\dots,X_0;\tilde{\theta},\tilde{\alpha})\right| < \infty. \tag{2.14}$$

In the same way as we derived (2.4) we obtain the representations,

$$\frac{\frac{\partial^2}{\partial\theta_i^2} P^{\theta,\alpha}_{(X_{-1},\dots,X_{-p}),X_0}}{P^{\theta,\alpha}_{(X_{-1},\dots,X_{-p}),X_0}} = \mathbb{E}_{\theta,\alpha}\left[\ddot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) + \dot{s}^2_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) \mid X_0,\dots,X_{-p}\right],$$

and for $i \neq j$,

$$\frac{\frac{\partial^2}{\partial\theta_j\partial\theta_i} P^{\theta,\alpha}_{(X_{-1},\dots,X_{-p}),X_0}}{P^{\theta,\alpha}_{(X_{-1},\dots,X_{-p}),X_0}} = \mathbb{E}_{\theta,\alpha}\left[\dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i})\dot{s}_{X_{-j},\theta_j}(\vartheta_j \circ X_{-j}) \mid X_0,\dots,X_{-p}\right].$$

Using (2.10) we obtain the bound, for $i, j = 1, \dots, p$,

$$\left|\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} P^{\theta,\alpha}_{(X_{-1},\dots,X_{-p}),X_0}}{P^{\theta,\alpha}_{(X_{-1},\dots,X_{-p}),X_0}}\right| \leq \frac{1}{\theta_i(1-\theta_i)}\frac{1}{\theta_j(1-\theta_j)}\left(X_{-i}^2 + X_{-j}^2\right),$$

which, since $\mathbb{E}_{\nu_0,\theta,\alpha} X_t^2 < \infty$, implies, for $i, j = 1, \dots, p$,

$$\mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha}):|(\tilde{\theta},\tilde{\alpha})-(\theta,\alpha)|<\delta} \left|\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} P^{\tilde{\theta},\tilde{\alpha}}_{(X_{-1},\dots,X_{-p}),X_0}}{P^{\tilde{\theta},\tilde{\alpha}}_{(X_{-1},\dots,X_{-p}),X_0}}\right| < \infty. \tag{2.15}$$

In the same way as we derived (2.5) we obtain the representation, for $i, j = 1, \ldots, q$,

$$\frac{\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} P^{\theta,\alpha}_{(X_{-1},\ldots,X_{-p}),X_0}}{P^{\theta,\alpha}_{(X_{-1},\ldots,X_{-p}),X_0}} = \mathbb{E}_{\theta,\alpha}\left[\dot{h}_{\alpha,ji}(\varepsilon_0) + h_{\alpha,j}(\varepsilon_0) h_{\alpha,i}(\varepsilon_0) \mid X_0, \ldots, X_{-p}\right].$$

Using (2.1) from Assumption (A3) on $\mathscr{G}_A$ we obtain,

$$\begin{aligned}
&\mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha}):|(\tilde{\theta},\tilde{\alpha})-(\theta,\alpha)|<\delta} \left|\mathbb{E}_{\tilde{\theta},\tilde{\alpha}}\left[h_{\tilde{\alpha},j}(\varepsilon_0) h_{\tilde{\alpha},i}(\varepsilon_0) \mid X_0, \ldots, X_{-p}\right]\right| \\
&\leq \mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha})\in B_\delta} \sqrt{\mathbb{E}_{\tilde{\theta},\tilde{\alpha}}\left[|h_{\tilde{\alpha},j}(\varepsilon_0)|^2 \mid X_0, \ldots, X_{-p}\right] \mathbb{E}_{\tilde{\theta},\tilde{\alpha}}\left[|h_{\tilde{\alpha},i}(\varepsilon_0)|^2 \mid X_0, \ldots, X_{-p}\right]} \\
&\leq C\mathbb{E}_{\nu_0,\theta,\alpha}\left(\sum_{i=0}^{p} X_{-i}\right)^2 < \infty,
\end{aligned}$$

where $B_\delta = \{(\tilde{\theta}, \tilde{\alpha}) \mid |(\tilde{\theta}, \tilde{\alpha}) - (\theta, \alpha)| < \delta\}$. Hence, an combination with (2.2) from Assumption (A3), yields, for $i, j = 1, \ldots, q$,

$$\mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha}):|(\tilde{\theta},\tilde{\alpha})-(\theta,\alpha)|<\delta} \left|\frac{\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} P^{\tilde{\theta},\tilde{\alpha}}_{(X_{-1},\ldots,X_{-p}),X_0}}{P^{\tilde{\theta},\tilde{\alpha}}_{(X_{-1},\ldots,X_{-p}),X_0}}\right| < \infty. \tag{2.16}$$

Next we compute for $i = 1, \ldots, p$, $j = 1, \ldots, q$, the representation,

$$\frac{\frac{\partial^2}{\partial \alpha_j \partial \theta_i} P^{\theta,\alpha}_{(X_{-1},\ldots,X_{-p}),X_0}}{P^{\theta,\alpha}_{(X_{-1},\ldots,X_{-p}),X_0}} = \mathbb{E}_{\theta,\alpha}\left[h_{\alpha,j}(\varepsilon_0) \dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) \mid X_0, \ldots, X_{-p}\right],$$

which, using (2.10) and (2.1), yields,

$$\begin{aligned}
&\mathbb{E}_{\nu_0,\theta,\alpha} \sup_{(\tilde{\theta},\tilde{\alpha}):|(\tilde{\theta},\tilde{\alpha})-(\theta,\alpha)|<\delta} \left|\mathbb{E}_{\tilde{\theta},\tilde{\alpha}}\left[h_{\tilde{\alpha},j}(\varepsilon_0) \dot{s}_{X_{-i},\tilde{\theta}_i}(\vartheta_i \circ X_{-i}) \mid X_0, \ldots, X_{-p}\right]\right| \\
&\qquad \leq M_\theta \sqrt{\mathbb{E}_{\nu_0,\theta,\alpha} X_0^2} \sqrt{C\mathbb{E}_{\nu_0,\theta,\alpha}\left(\sum_{i=0}^{p} X_{-i}\right)^2} < \infty.
\end{aligned} \tag{2.17}$$

<u>Part 1: the score</u> From (2.4) it follows that,

$$\begin{aligned}
&\mathbb{E}_{\theta,\alpha}\left[\dot{\ell}_{\theta,i}(X_{t-p}, \ldots, X_t; \theta, \alpha) \mid X_{t-1}, \ldots, X_{t-p}\right] \\
&\qquad = \mathbb{E}_{\theta,\alpha}\left[\dot{s}_{X_{t-i},\theta_i}(\vartheta_i \circ X_{t-i}) \mid X_{t-1}, \ldots, X_{t-p}\right] = 0,
\end{aligned} \tag{2.18}$$

since $\vartheta_i \circ X_{t-i}$, conditional on $X_{t-p}, \ldots, X_{t-1}$, has expectation $\theta_i X_{t-i}$. From (2.5) it follows that,

$$\mathbb{E}_{\theta,\alpha}\left[\dot{\ell}_{\alpha,j}(X_{t-p}, \ldots, X_t; \theta, \alpha) \mid X_{t-1}, \ldots, X_{t-p}\right]$$

$$= \mathbb{E}_{\theta,\alpha}\left[h_{\alpha,j}(\varepsilon_t) \mid X_{t-1},\dots,X_{t-p}\right] = 0, \tag{2.19}$$

since ε_t is independent of $X_{t-p},\dots,X_{t-1}$ and $\mathbb{E}_\alpha h_{\alpha,j}(\varepsilon_0) = 0$. Let $w = (w_1, w_2) \in \mathbb{R}^p \times \mathbb{R}^q$. From (2.18) and (2.19) it follows that,

$$\mathbb{E}_{\theta,\alpha}\left[w_1^T \dot{\ell}_\theta(X_{t-p},\dots,X_t;\theta,\alpha) + w_2^T \dot{\ell}_\alpha(X_{t-p},\dots,X_t;\theta,\alpha) \mid X_{t-1},\dots,X_{t-p}\right] = 0,$$

and, by (2.12) and (2.13),

$$\mathbb{E}_{\nu_0,\theta,\alpha}\left[w_1^T \dot{\ell}_\theta(X_{-p},\dots,X_0;\theta,\alpha) + w_2^T \dot{\ell}_\alpha(X_{-p},\dots,X_0;\theta,\alpha)\right]^2 = w^T J w < \infty.$$

Hence we have, by Proposition 1.2.1A,

$$\frac{1}{\sqrt{n}}\sum_{t=0}^{n}\left[w_1^T \dot{\ell}_\theta(X_{t-p},\dots,X_t;\theta,\alpha) + w_2^T \dot{\ell}_\alpha(X_{t-p},\dots,X_t;\theta,\alpha)\right] \xrightarrow{d} w^T \mathrm{N}(0,J),$$

under $\mathbb{P}_{\nu,\theta,\alpha}$. Display (2.7) now follows by applying the Cramér-Wold device, which concludes Part 1.

Part 2: the Fisher information In this part we prove that $J_n(\tilde{\theta}_n,\tilde{\alpha}_n) \xrightarrow{p} J$ under $\mathbb{P}_{\nu,\theta,\alpha}$, where

$$J_n(\theta,\alpha) = \begin{pmatrix} J_n^\theta & J_n^{\theta,\alpha} \\ J_n^{\alpha,\theta} & J_n^\alpha \end{pmatrix}$$
$$= -\frac{1}{n}\sum_{t=0}^{n}\begin{pmatrix} \frac{\partial}{\partial\theta^T}\dot{\ell}_\theta(X_{t-p},\dots,X_t;\theta,\alpha) & \frac{\partial}{\partial\alpha^T}\dot{\ell}_\theta(X_{t-p},\dots,X_t;\theta,\alpha) \\ \frac{\partial}{\partial\theta^T}\dot{\ell}_\alpha(X_{t-p},\dots,X_t;\theta,\alpha) & \frac{\partial}{\partial\alpha^T}\dot{\ell}_\alpha(X_{t-p},\dots,X_t;\theta,\alpha) \end{pmatrix}. \tag{2.20}$$

Using Assumption (A2) on $\mathscr{G}_A$ it is easy to see that for $x_{-p},\dots,x_0 \in \mathbb{Z}_+$ the following mappings are all continuous: $(\theta,\alpha) \mapsto (\partial/\partial\theta)\log\dot{\ell}_\theta(x_{-p},\dots,x_0;\theta,\alpha)$, $(\theta,\alpha) \mapsto (\partial/\partial\theta)\log\dot{\ell}_\alpha(x_{-p},\dots,x_0;\theta,\alpha)$, $(\theta,\alpha) \mapsto (\partial/\partial\alpha)\log\dot{\ell}_\theta(x_{-p},\dots,x_0;\theta,\alpha)$ and $(\theta,\alpha) \mapsto (\partial/\partial\alpha)\log\dot{\ell}_\alpha(x_{-p},\dots,x_0;\theta,\alpha)$. Since we already proved (2.12), (2.13), (2.14), (2.15), (2.16), and (2.17), it is sufficient, by Proposition 1.2.1.3, to prove that we have $J_n(\theta,\alpha) \xrightarrow{p} J$.
First we consider the diagonal of J_n^θ. For $i \in \{1,\dots,p\}$, the calculations in Part 0 and a Markov law of large numbers (see Dacunha-Castelle and Duflo (1986, Theorem 4.3.15)), yield,

$$J_{n,ii}^\theta \xrightarrow{p} -\mathbb{E}_{\nu_0,\theta,\alpha}\mathbb{E}_{\theta,\alpha}\left[\ddot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) + \dot{s}^2_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) \mid X_0,\dots,X_{-p}\right]$$
$$+\mathbb{E}_{\nu_0,\theta,\alpha}\dot{\ell}^2_{\theta,i}(X_{-p},\dots,X_0;\theta,\alpha) = J_{ii}^\theta,$$

where the last equality follows from,

$$\mathbb{E}_{\nu_0,\theta,\alpha}\left[\ddot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) + \dot{s}^2_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i})\right]$$

$$= \mathbb{E}_{\nu_0,\theta,\alpha}\mathbb{E}_{\theta,\alpha}\left[\ddot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) + \dot{s}^2_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) \mid X_{-1},\dots,X_{-p}\right] = 0,$$

which is standard once one realizes that $\vartheta_i \circ X_{-i}$ given $X_{-p},\dots,X_{-1}$ is $\mathrm{Bin}_{X_{-i},\theta_i}$ distributed. Next we consider the off-diagonal elements of J^θ. Let $i \neq j$. Applying the representations in Part 0 and a Markov law of large numbers gives,

$$J^\theta_{n,ij} \xrightarrow{p} -\mathbb{E}_{\nu_0,\theta,\alpha}\mathbb{E}_{\theta,\alpha}\left[\dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i})\dot{s}_{X_{-j},\theta_j}(\vartheta_j \circ X_{-j}) \mid X_0,\dots,X_{-p}\right]$$
$$+ \mathbb{E}_{\nu_0,\theta,\alpha}\dot{\ell}_{\theta,i}\dot{\ell}_{\theta,j}(X_{-p},\dots,X_0;\theta,\alpha) = J^\theta_{ij},$$

since,

$$\mathbb{E}_{\nu_0,\theta,\alpha}\dot{s}_{X_{-i},\theta_i}(\theta_i \circ X_{-i})\dot{s}_{X_{-j},\theta_j}(\vartheta_j \circ X_{-j})$$
$$= \mathbb{E}_{\nu_0,\theta,\alpha}\mathbb{E}_{\theta,\alpha}\left[\dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i})\dot{s}_{X_{-j},\theta_j}(\vartheta_j \circ X_{-j}) \mid X_{-1},\dots,X_{-p}\right] = 0,$$

because $\vartheta_i \circ X_{-i}$ and $\vartheta_j \circ X_{-j}$ given $X_{-p},\dots,X_{-1}$ are mean-zero and independent. Next we consider the block $J^{\theta,\alpha}_n$ (by symmetry this also yields the result for the block $J^{\alpha,\theta}_n$). Using the representations derived in Part 0 and a law of large numbers for Markov chains we obtain,

$$J^{\theta,\alpha}_{n,ij} \xrightarrow{p} -\mathbb{E}_{\nu_0,\theta,\alpha}\mathbb{E}_{\theta,\alpha}\left[h_{\alpha,j}(\varepsilon_0)\dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) \mid X_0,\dots,X_{-p}\right]$$
$$+ \mathbb{E}_{\nu_0,\theta,\alpha}\dot{\ell}_{\alpha,j}\dot{\ell}_{\theta,i}(X_{-p},\dots,X_0;\theta,\alpha) = I^{\theta,\alpha}_{ij},$$

since,

$$\mathbb{E}_{\nu_0,\theta,\alpha}\left[\mathbb{E}_{\theta,\alpha}\left[h_{\alpha,j}(\varepsilon_0)\dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) \mid X_0,\dots,X_{-p}\right]\right]$$
$$= \mathbb{E}_{\nu_0,\theta,\alpha}h_{\alpha,j}(\varepsilon_0)\dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) = 0,$$

because $h_{\alpha,j}(\varepsilon_0)$ and $\dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i})$ are independent and have mean zero. Finally we treat J^α_n. Using the representations in Part 0 and the law of large numbers again, we obtain

$$J^\alpha_n \xrightarrow{p} -\mathbb{E}_{\nu_0,\theta,\alpha}\mathbb{E}_{\theta,G}\left[\dot{h}_\alpha(\varepsilon_0) + h_\alpha h^T_\alpha(\varepsilon_0) \mid X_0,\dots,X_{-p}\right]$$
$$+ \mathbb{E}_{\nu_0,\theta,\alpha}\dot{\ell}_\alpha\dot{\ell}^T_\alpha(X_{-p},\dots,X_0;\theta,\alpha) = J^\alpha,$$

since, by Assumption (A4) on $\mathscr{G}_A$,

$$\mathbb{E}_{\nu_0,\theta,\alpha}\mathbb{E}_{\theta,\alpha}\left[\dot{h}_\alpha(\varepsilon_0) + h_\alpha h^T_\alpha(\varepsilon_0) \mid X_0,\dots,X_{-p}\right] = \mathbb{E}_\alpha\dot{h}_\alpha(\varepsilon_0) + \mathbb{E}_\alpha h_\alpha h^T_\alpha(\varepsilon_0) = 0.$$

Part 3: non-singularity of J Finally we prove that J is non-singular. First we prove that J^α is non-singular. If J^α would be singular we would have,

$$a_2^T\dot{\ell}_\alpha\left(X_{-p},\dots,X_0;\theta,\alpha\right) = 0 \quad \mathbb{P}_{\nu_0,\theta,\alpha}\text{-a.s. for certain } a_2 \in \mathbb{R}^q\backslash\{0\}.$$

Note that we have, for all $k \in \text{support}(G_\alpha)$, $\mathbb{P}_{\nu_0,\theta,\alpha}\left\{X_{-p} = \cdots = X_{-1} = 0, X_0 = k\right\} > 0$, and on the event $E_k = \left\{X_{-p} = \cdots = X_{-1} = 0, X_0 = k\right\}$ we have $\varepsilon_0 = k$. Hence, for $k \in \text{support}(G_\alpha)$, we obtain on the event E_k

$$0 = a_2^T \dot{\ell}_\alpha\left(X_{-p},\ldots,X_0;\theta,\alpha\right) = a_2^T \mathbb{E}_{\theta,\alpha}\left[h(\varepsilon_0) \mid X_0,\ldots,X_{-p}\right] = a_2^T h_\alpha(k),$$

which contradicts Assumption (A4) on $\mathcal{G}_A$ that $\mathbb{E}_\alpha h_\alpha(\varepsilon_0) h_\alpha^T(\varepsilon_0)$ is non-singular. Hence J^α is indeed non-singular.
Suppose that (a_1, a_2) is such that

$$a_1^T \dot{\ell}_\theta\left(X_{-p},\ldots,X_0;\theta,\alpha\right) + a_2^T \dot{\ell}_\alpha\left(X_{-p},\ldots,X_0;\theta,\alpha\right) = 0 \quad \mathbb{P}_{\nu_0,\theta,\alpha}\text{-a.s.} \tag{2.21}$$

Let $i \in \{1,\ldots,p\}$ and note that for $k \in \mathbb{Z}_+$ the event

$$\{X_j = 0 \text{ for } j \in \{-p,\ldots,0\}\backslash\{-i\},\ X_{-i} = k\}$$

has positive probability under $\mathbb{P}_{\nu_0,\theta,\alpha}$ and that on this event we have,

$$\dot{\ell}_{\theta,i}\left(X_{-p},\ldots,X_0;\theta,\alpha\right) = \mathbb{E}_{\theta,\alpha}\left[\dot{s}_{X_{-i},\theta_i}(\vartheta_i \circ X_{-i}) \mid X_0,\ldots,X_{-p}\right] = -\frac{\theta_i k}{\theta_i(1-\theta_i)},$$
$$\dot{\ell}_{\theta,j}\left(X_{-p},\ldots,X_0;\theta,\alpha\right) = \mathbb{E}_{\theta,\alpha}\left[\dot{s}_{X_{-j},\theta_j}(\vartheta_j \circ X_{-j}) \mid X_0,\ldots,X_{-p}\right] = 0, \quad \text{for } j \neq i,$$

and,

$$\dot{\ell}_{\alpha,m}(X_{-p},\ldots,X_0;\theta,\alpha) = \mathbb{E}_{\theta,\alpha}\left[h_{\alpha,m}(\varepsilon_0) \mid X_0,\ldots,X_{-p}\right] = h_{\alpha,m}(0).$$

Hence we obtain from (2.21), for $k \in \mathbb{Z}_+$, $i = 1,\ldots,p$, the equality,

$$\frac{-a_{1,i}\theta_i k}{\theta_i(1-\theta_i)} + a_2^T h_\alpha(0) = 0,$$

which is only possible if $a_1 = 0$. Hence $a_1 = 0$, so from (2.21) we get

$$a_2^T \dot{\ell}_\alpha\left(X_{-p},\ldots,X_0;\theta,\alpha\right) = 0 \quad \mathbb{P}_{\nu_0,\theta,\alpha}\text{-a.s.}$$

This is only possible if $a_2 = 0$, since we already proved that J^α is non-singular. Thus $(a_1, a_2) = 0$, and we conclude that J is non-singular. □

If we want to consider the sequence of experiments $\mathcal{E}_1^{(n)}(\nu, G)$, $n \in \mathbb{Z}_+$, we can always embed G in a parametric model $\mathcal{G}_A$ which satisfies Assumptions (A1)-(A5). Then an application of the preceding theorem with $u_2 = 0$ immediately yields the following corollary.

Corollary 2.2. *Let $\theta \in \Theta$, let $G \in \mathcal{G}$ with $\mathbb{E}_G\varepsilon_0^2 < \infty$, and $g(0) \in (0,1)$, and let ν be a probability measure on $\mathbb{Z}_+^p$. Then the sequence of experiments $(\mathcal{E}_1^{(n)}(\nu, G))_{n\in\mathbb{Z}_+}$ has the LAN-property in θ, i.e. for every $u \in \mathbb{R}^p$ the following expansion holds,*

$$\log \frac{\mathrm{d}\mathbb{P}^{(n)}_{\nu,\theta+\frac{u}{\sqrt{n}},G}}{\mathrm{d}\mathbb{P}^{(n)}_{\nu,\theta,G}}(X_{-p},\ldots,X_n) = u^T S_n^\theta - \frac{1}{2} u^T J_\theta u + R_n,$$

where $S_n^\theta = n^{-1/2}\sum_{t=0}^n \dot{\ell}_\theta(X_{t-p},\ldots,X_t;\theta,G) \xrightarrow{d} \mathrm{N}(0, J_\theta)$ under $\mathbb{P}_{\nu,\theta,G}$, $J_\theta = J_\theta(\theta, G)$ is invertible, and $R_n = R_n(u,\theta,G) \xrightarrow{p} 0$ under $\mathbb{P}_{\nu,\theta,G}$.

2.2 Efficient estimation

This section provides efficient estimators of the parameters in an INAR(p) model based on the ubiquitous one-step update method.

2.2.1 Innovation distribution is known

In case $\mu_G < \infty$, an initial estimator of θ is the OLS-estimator,

$$\hat{\theta}_n^G = \begin{pmatrix} \sum_{t=0}^n X_{t-1}^2 & \dots & \sum_{t=0}^n X_{t-1}X_{t-p} \\ \vdots & \ddots & \vdots \\ \sum_{t=0}^n X_{t-p}X_{t-1} & \dots & \sum_{t=0}^n X_{t-p}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=0}^n X_{t-1}(X_t - \mu_G) \\ \vdots \\ \sum_{t=0}^n X_{t-p}(X_t - \mu_G) \end{pmatrix}.$$

If we assume the existence of a third moment of X_0 under the stationary distribution (which is, by Lemma 1.2.1, equivalent to imposing $\mathbb{E}_G \varepsilon_0^3 < \infty$), $\hat{\theta}_n^G$ yields a $\sqrt{n}$-consistent estimator of θ. The following proposition is well-known (see Du and Li (1991)).

Proposition 2.2.1. Let $\theta \in \Theta$, ν a probability measure on $\mathbb{Z}_+^p$, $G \in \mathscr{G}$ with $g(0) \in (0,1)$ and $\mathbb{E}_G \varepsilon_0^3 < \infty$. Then $\sqrt{n}\left(\hat{\theta}_n^G - \theta\right)$ converges in distribution under $\mathbb{P}_{\nu,\theta,G}$.

Next, we apply the one-step-Newton-Raphson-method to update this initial $\sqrt{n}$-consistent estimator into an efficient estimator. To state this theorem, we need the concept of a discretized estimator. For $n \in \mathbb{N}$ make a grid of cubes, with sides of length $1/\sqrt{n}$, over $\mathbb{R}^p$ and, given $\hat{\theta}_n^G$, define $\hat{\theta}_n^{G,*}$ to be the midpoint of the cube into which $\hat{\theta}_n$ has fallen (for ties take one of the possibilities). Then $\hat{\theta}_n^{G,*}$ is also $\sqrt{n}$-consistent and is called a discretized version of $\hat{\theta}_n$.

Theorem 2.3. *Let ν a probability measure on $\mathbb{Z}_+^p$, $G \in \mathscr{G}$ with $g(0) \in (0,1)$ and $\mathbb{E}_G \varepsilon_0^3 < \infty$. Let $\hat{\theta}_n^*$ be a discretized version of $\hat{\theta}_n$. Then*

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

where,

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

is an efficient estimator of θ in the sequence of experiments $(\mathscr{E}_1^{(n)}(\nu,G))_{n\in\mathbb{Z}_+}$. Moreover, $\hat{J}_{n,\theta}^{-1}$ is a consistent estimator of the asymptotic covariance matrix of $\hat{\theta}_n^G$, i.e.

$$\hat{J}_{n,\theta}^{-1} \xrightarrow{p} J_\theta^{-1}, \text{ under } \mathbb{P}_{\nu,\theta,G}.$$

Remark 7. Instead of $\hat{\theta}_n^G$, any other $\sqrt{n}$-consistent estimator of θ can be used.

Remark 8. If one can find a $\sqrt{n}$-consistent initial estimator of θ under the weaker assumption $\mathbb{E}_G \varepsilon_0^2 < \infty$, the condition $\mathbb{E}_G \varepsilon_0^3 < \infty$ may be replaced by $\mathbb{E}_G \varepsilon_0^2 < \infty$.

The proof of this theorem runs along the same lines as the proof of Theorem 2.4.

Innovation distribution belongs to a parametric model 2.2.2

To use the OLS-estimator as an initial estimator of θ we need the existence of a third moment of X_t under the stationary distribution. Therefore we replace, in this section, Assumption (A5) on $\mathscr{G}_A$ by,

(A5′) for all $\alpha \in A$: $\mathbb{E}_\alpha \varepsilon_0^3 < \infty$.

This yields, by Lemma 1.2.1, the existence of a third moment of X_0 under the stationary distribution. Just as for the case G known, OLS yields a $\sqrt{n}$-consistent estimator of (θ, μ_G) (see, for example, Du and Li (1991)).

Proposition 2.2.2. Let $\theta \in \Theta$, ν a probability measure on $\mathbb{Z}_+^p$, $G \in \mathscr{G}$ with $\mathbb{E}_G \varepsilon_0^3 < \infty$ and $g(0) \in (0,1)$. Then $\left(\sqrt{n}\left(\hat{\theta}_n - \theta\right), \sqrt{n}\left(\hat{\mu}_{G,n} - \mu_G\right)\right)$ converges in distribution under $\mathbb{P}_{\nu,\theta,G}$, where,

$$\begin{pmatrix} \hat{\mu}_{G,n} \\ \hat{\theta}_n \end{pmatrix} = \begin{pmatrix} n & \sum_{t=0}^n X_{t-1} & \cdots & \sum_{t=0}^n X_{t-p} \\ \sum_{t=0}^n X_{t-1} & \sum_{t=0}^n X_{t-1}^2 & \cdots & \sum_{t=0}^n X_{t-1} X_{t-p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=0}^n X_{t-p} & \sum_{t=0}^n X_{t-p} X_{t-1} & \cdots & \sum_{t=0}^n X_{t-p}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=0}^n X_t \\ \sum_{t=0}^n X_{t-1} X_t \\ \vdots \\ \sum_{t=0}^n X_{t-p} X_t \end{pmatrix}.$$

Note that $\hat{\mu}_{G,n}$ yields a $\sqrt{n}$-consistent estimator of α for the popular choice $\mathscr{G}_A = (\text{Poisson}(\alpha) \mid \alpha > 0)$, since then $\mu_{G_\alpha} = \alpha$.

For other specific choices of G_α, it might by easy to find a (moment-based) estimator of α. This is the approach we recommend. However, it would be reassuring to know that a $\sqrt{n}$-consistent estimator of α always exists. The following observation is the key to the general existence of a $\sqrt{n}$-consistent estimator of α. Although we do not observe the innovation process $(\varepsilon_t)_{t \in \mathbb{Z}_+}$, we have observations on some innovations (if $g(0) > 0$), since

$$X_t 1\{X_{t-1} = 0, \ldots, X_{t-p} = 0\} = \varepsilon_t. \tag{2.22}$$

By Assumptions (A1)-(A6) $\mathscr{G}_A$ is an identified regular parametric model (see Definition 2.1.1 and Proposition 2.1.1 in Bickel et al. (1998)). By a theorem by Le Cam (see, e.g., Theorem 2.5.1 in Bickel et al. (1998)) there exists an 'estimator' $T_n = t_n(\varepsilon_1, \ldots, \varepsilon_n)$ of α such that $\sqrt{n}(T_n - \alpha)$ is tight under $\mathbb{P}_{\nu,\theta,\alpha}$ for all $\alpha \in A$. Using display (2.22) we could use such an 'estimator' to construct a $\sqrt{n}$-consistent estimator of α.

Proposition 2.2.3. Let $\mathscr{G}_A \subset \mathscr{G}$ satisfy Assumptions (A1)-(A6), ν a probability measure on $\mathbb{Z}_+^p$, and $(\theta, \alpha) \in \Theta \times A$. Let

$$\tau_0 = 0, \ \tau_k = \inf\{t > \tau_{k-1} \mid X_{t-p} = \cdots = X_{t-1} = 0\}, \ k \in \mathbb{N},$$

and

$$N_n = \max\{j \in \mathbb{Z}_+ \mid \tau_j \le n\}.$$

Then $\hat{\alpha}_n = t_{N_n}\left(X_{\tau_1},\ldots,X_{\tau_{N_n}}\right)$, defines a $\sqrt{n}$-consistent estimator of α. In particular, if for some $\sigma^2 > 0$,

$$\sqrt{n}\,(t_n(\varepsilon_1,\ldots,\varepsilon_n)-\alpha) \xrightarrow{d} \mathrm{N}(0,\sigma^2), \text{ under } \mathbb{P}_{\nu,\theta,\alpha},$$

we have,

$$\sqrt{n}\,(\hat{\alpha}_n-\alpha) \xrightarrow{d} \mathrm{N}\left(0,\frac{\sigma^2}{\nu_{\theta,\alpha}\{0,\ldots,0\}}\right), \text{ under } \mathbb{P}_{\nu,\theta,\alpha}.$$

Proof.
By Prohorov's theorem it suffices to prove that there exists a subsequence n_k such that $\sqrt{n_k}(\hat{\alpha}_{n_k}-\alpha)$ converges in distribution.
Note first that, by a law of large numbers for Markov chains (see, for example Dacunha-Castelle and Duflo (1986, Theorem 4.3.15)),

$$\frac{N_n}{n} \to \nu_{\theta,\alpha}\{0,\ldots,0\} > 0, \quad \mathbb{P}_{\nu,\theta,\alpha}-\text{a.s.}$$

Let, for $u \in \mathbb{R}^q$,

$$\phi_n(u) = \mathbb{E}_\alpha \exp\left(iu^T(\sqrt{n}(t_n(\varepsilon_1,\ldots,\varepsilon_n)-\alpha))\right).$$

Since $t_n(\varepsilon_1,\ldots,\varepsilon_n)$ is a $\sqrt{n}$-consistent estimator of α, there exists, by Prohorov's theorem, a subsequence n_k such that $\sqrt{n_k}(t_{n_k}(\varepsilon_1,\ldots,\varepsilon_{n_k})-\alpha)$ converges in distribution under $\mathbb{P}_{\nu,\theta,\alpha}$. Hence for all $u \in \mathbb{R}^q$,

$$\lim_{k\to\infty} \phi_{n_k}(u) = \phi(u),$$

where ϕ is a characteristic function of an $\mathbb{R}^q$-valued random variable, which we denote by Z. Using the strong Markov property, it is not very hard to see that $(X_{\tau_k})_{k\in\mathbb{N}}$ are i.i.d. G-distributed independent of N_n. Hence (use dominated convergence),

$$\lim_{k\to\infty} \mathbb{E}_{\nu,\theta,\alpha} \exp\left(iu^T(\sqrt{N_{n_k}}(\hat{\alpha}_{n_k}-\alpha))\right) = \lim_{k\to\infty} \mathbb{E}_{\nu,\theta,\alpha}\phi_{N_{n_k}}(u) = \mathbb{E}_{\nu,\theta,\alpha}\phi(u) = \phi(u),$$

which yields,

$$\sqrt{N_{n_k}}(\hat{\alpha}_{n_k}-\alpha) \xrightarrow{d} Z, \text{ under } \mathbb{P}_{\nu,\theta,\alpha} \text{ as } k\to\infty.$$

Now,

$$\sqrt{n_k}(\hat{\alpha}_{n_k}-\alpha) = \sqrt{\frac{n_k}{N_{n_k}}}\sqrt{N_{n_k}}(\hat{\alpha}_{n_k}-\alpha) \xrightarrow{d} \frac{1}{\sqrt{\nu_{\theta,\alpha}\{0,\ldots,0\}}}Z,$$

under $\mathbb{P}_{\nu,\theta,\alpha}$ as $k\to\infty$, which concludes the proof. □

Remark 9. For the construction in the proposition it is essential that the process can drive to state 0 for which it is necessary that the immigration distribution assigns positive mass to state 0. If the immigration distribution does not assign mass to state 0 the situation is more complicated. However, notice that, conditional on the past, the law of X_t is the convolution of the immigration distribution with binomial distributions. The parameters in these binomial distributions can be estimated by OLS and we can also estimate the transition-probabilities from our observations. So the idea is that, in general, α can be estimated by a deconvolution argument.

Since we have a $\sqrt{n}$-consistent estimator of (θ,α), we can update this estimator into an efficient estimator.

Theorem 2.4. *Let ν a probability measure on $\mathbb{Z}_+^p$, and $\mathcal{G}_A \subset \mathcal{G}$ satisfying Assumptions (A1)-(A6) with (A5′) instead of (A5). Let $(\hat\theta_n, \hat\alpha_n)$ be a $\sqrt{n}$-consistent estimator of (θ,α) and $(\hat\theta_n^*, \hat\alpha_n^*)$ a discretized version of it. Then,*

$$\begin{pmatrix}\hat\theta_n^{**}\\ \hat\alpha_n^{**}\end{pmatrix} = \begin{pmatrix}\hat\theta_n^{*}\\ \hat\alpha_n^{*}\end{pmatrix} + \frac{1}{n}\sum_{t=0}^{n} \hat J_n^{-1} \begin{pmatrix}\dot\ell_\theta(X_{t-p},\dots,X_t;\hat\theta_n^*,\hat\alpha_n^*)\\ \dot\ell_\alpha(X_{t-p},\dots,X_t;\hat\theta_n^*,\hat\alpha_n^*)\end{pmatrix},$$

with,

$$\hat J_n = \begin{pmatrix}\frac{1}{n}\sum_{t=0}^{n}\dot\ell_\theta\dot\ell_\theta^T(X_{t-p},\dots,X_t;\hat\theta_n^*,\hat\alpha_n^*) & \frac{1}{n}\sum_{t=0}^{n}\dot\ell_\theta\dot\ell_\alpha^T(X_{t-p}\dots,,X_t;\hat\theta_n^*,\hat\alpha_n^*)\\ \frac{1}{n}\sum_{t=0}^{n}\dot\ell_\alpha\dot\ell_\theta^T(X_{t-p},\dots,X_t;\hat\theta_n^*,\hat\alpha_n^*) & \frac{1}{n}\sum_{t=0}^{n}\dot\ell_\alpha\dot\ell_\alpha^T(X_{t-p},\dots,X_t;\hat\theta_n^*,\hat\alpha_n^*)\end{pmatrix},$$

is an efficient estimator of (θ,α) in the sequence of experiments $(\mathcal{E}_2^{(n)}(\nu,\mathcal{G}_A))_{n\in\mathbb{Z}_+}$. Moreover, $\hat J_n^{-1}$ yields a consistent estimator of the asymptotic covariance matrix of $(\hat\theta_n^{**},\hat\alpha_n^{**})$, i.e.,

$$\hat J_n^{-1} \xrightarrow{p} J^{-1}, \text{ under } \mathbb{P}_{\nu,\theta,\alpha}.$$

Remark 10. The same comments as after Theorem 2.3 apply.

Proof.
Let $(\theta,\alpha)\in\Theta\times A$. To prove that $(\hat\theta_n^{**},\hat\alpha_n^{**})$ is efficient at (θ,α) it suffices (see, for example, Theorem 2.3.1 in Bickel et al. (1998)) to prove that it is asymptotically linear in the efficient influence function at (θ,α), i.e.

$$\sqrt{n}\begin{pmatrix}\hat\theta_n^{**}-\theta\\ \hat\alpha_n^{**}-\alpha\end{pmatrix} = J^{-1}(\theta,\alpha)S_n(\theta,\alpha) + o(\mathbb{P}_{\nu,\theta,\alpha};1).$$

If we can show that the following conditions hold,

(C1) $S_n(\theta,\alpha)$ converges in distribution under $\mathbb{P}_{\nu,\theta,\alpha}$;

(C2) for every deterministic sequence $(\theta_n,\alpha_n) = (\theta_0,\alpha_0) + O(1/\sqrt{n})$ we have,

$$S_n(\theta_n,\alpha_n) - S_n(\theta,\alpha) + J(\theta,\alpha)\sqrt{n}\begin{pmatrix}\theta_n-\theta\\ \alpha_n-\alpha\end{pmatrix} \xrightarrow{p} 0, \text{ under } \mathbb{P}_{\nu,\theta,\alpha};$$

C3 $\hat{J}_n \xrightarrow{p} J(\theta,\alpha)$ under $\mathbb{P}_{\nu,\theta,\alpha}$,

then we obtain, from Theorem 5.48 in Van der Vaart (2000) ($(\hat{\theta}_n^*, \hat{\alpha}_n^*)$ is consistent and discretized) the desired result.
Condition 1 has already been proved in Part 1 of the proof of Theorem 2.1; Condition 3 is proved in Part 0 and Part 2 of the proof of Theorem 2.1. Let $(\theta_n, \alpha_n) = (\theta, \alpha) + O(n^{-1/2})$ be a deterministic sequence. From the proof of Theorem 2.1 we have,

$$S_n(\theta_n, \alpha_n) = S_n(\theta, \alpha) - J_n(\tilde{\theta}_n, \tilde{\alpha}_n)\sqrt{n}\begin{pmatrix}\theta_n - \theta \\ \alpha_n - \alpha\end{pmatrix},$$

where $(\tilde{\theta}_n, \tilde{\alpha}_n)$ is a point between (θ, α) and (θ_n, α_n). Using Part 0 in the proof of Theorem 2.1 and Proposition 1.2.1.3 we obtain $J_n(\tilde{\theta}_n, \tilde{\alpha}_n) \xrightarrow{p} J(\theta, \alpha)$ under $\mathbb{P}_{\nu,\theta,\alpha}$. This yields Condition 2, which concludes the proof. □

3 Semiparametric stationary INAR(p) models

In the previous chapter we discussed parametric INAR(p) models, i.e. the innovation distribution G is assumed to belong to a (smooth) parametric family. However, this exposes the researcher to possible misspecification. Therefore, one wants to consider a more realistic model. This chapter consider a semiparametric model, where hardly any assumptions are made on G. We focus on 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). Estimation of the innovation distribution is however, just as for standard AR models, an important topic. For INAR(p) processes this might be even more important, since in some applications G has a 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 modeled by an INAR(1) process, where the ε represent the number of new patients. In such examples the parameter G is the main parameter of interest.

Just as in the previous chapter, we restrict ourselves to the 'stationary parameter regime', i.e. $\Theta = \{\theta \in (0,1)^p \mid \sum_{i=1}^p \theta_i < 1\}$. Formally, we are interested in the experiments

$$\mathscr{E}^{(n)} = \left(\mathbb{Z}_+^{n+1+p}, 2^{\mathbb{Z}_+^{n+1+p}}, \left(\mathbb{P}^{(n)}_{\nu_{\theta,G},\theta,G} \mid \theta \in \Theta, G \in \mathscr{G}_{p+4}\right)\right), \quad n \in \mathbb{Z}_+,$$

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 $(\mathbb{Z}_+^{n+1+p}, 2^{\mathbb{Z}_+^{n+1+p}})$, and where $\mathscr{G}_{p+4}$ denotes the set of all probability distributions G on $\mathbb{Z}_+$ with finite $(p+4)$th moment and $0 < G\{0\} < 1$, and $\nu_{\theta,G}$ is the stationary initial distribution (see Theorem 1.1). Let us comment on the model assumptions on G. The assumption $0 < G\{0\} < 1$ ensures that it is possible that X becomes zero (and is not always equal to 0), which is reasonable for virtually all applications. Perhaps the assumption that the $(p+4)$th moment of G is finite appears to be odd at first sight. We need this assumption in establishing weak convergence of certain empirical processes. The size of the class of functions involved increases with p, which explains that we need a more stringent condition for larger p.

Compared to parametric models, the semiparametric model $\mathscr{E}^{(n)}$ 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 (see Du and Li (1991)) in the semiparametric model, it is not an efficient estimator of θ. This paper contributes a semiparametric efficient estimator of (θ, G). We stress once more 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 (1987b) 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 model cannot be analyzed by these approaches. The main problem is that one needs to have explicit expressions for the efficient influence operator. For the present model it however seems to be impossible to obtain a closed form formula for this efficient influence operator. Nevertheless we are able to prove efficiency. This proceeds 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. For i.i.d. models Van der Vaart (1995) gives high-level conditions to prove efficiency of infinite-dimensional Z-estimators without having to calculate the efficient influence operator. The basic idea is that often a 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 (1991b)) this can be exploited to obtain an efficiency proof. As we show, the i.i.d. framework of Van der Vaart (1995) extends to our Markovian setting. 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, which we established in Chapter 2. 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, essentially established in Chapter 2, is that parametric submodels of the semiparametric model enjoy the local asymptotic normality (LAN) property.

The setup of the rest of this chapter is as follows. Section 3.1 introduces the NPMLE and discusses its consistency. In Section 3.2 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. Here the main steps are the Fréchet differentiability of the limiting estimating equation, and the continuously invertibility of this operator. Section 3.3 proves that the NPMLE is efficient. Here we first show that parametric submodels have the LAN-property and that the NPMLE is regular. Next, following Van der Vaart (1995), the efficiency of the NPMLE follows from the regularity and the special representation of the limiting distribution. Finally, Section 3.4 discusses a small Monte Carlo simulation study and empirical application to analyze the finite sample behavior of the proposed estimator.

3.1 The estimator

In general, maximum likelihood estimation is not (directly) applicable in semiparametric models. For the INAR(p) 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.

$$\forall n \in \mathbb{Z}_+ : \quad (\hat{\theta}_n, \hat{G}_n) \in \operatorname*{argmax}_{(\theta,G)\in[0,1]^p\times\mathcal{G}} \prod_{t=0}^{n} P^{\theta,G}_{(X_{t-1},\dots,X_{t-p}),X_t}. \tag{3.1}$$

Note that, to guarantee the existence of a maximum likelihood estimator, we allow $(\hat{\theta}_n, \hat{G}_n)$ to take values outside $\Theta\times\mathcal{G}_{p+4}$. It is easy to see that, when it exists, $\hat{G}_n$ 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 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 optimization problem

$$\begin{aligned}
\max_{\substack{x_1,\dots,x_p\\ z_{u_-},\dots,z_{u_+}}} \quad & \prod_{t=0}^{n} \sum_{e=0\vee(X_t-\sum_{i=1}^p X_{t-i})}^{X_t} z_e \sum_{\substack{0\le k_\ell\le X_{t-\ell}\\ \ell=1,\dots,p\\ k_1+\dots+k_p=X_t-e}} \prod_{\ell=1}^{p} \binom{X_{t-\ell}}{k_\ell} x_\ell^{k_\ell}(1-x_\ell)^{X_{t-\ell}-k_\ell}\\
\text{s.t.} \quad & x_k \ge 0 \text{ for } k=1,\dots,p;\\
& x_k \le 1 \text{ for } k=1,\dots,p;\\
& z_j \ge 0 \text{ for } j=u_-,\dots,u_+;\\
& z_{u_-}+\dots+z_{u_+}=1.
\end{aligned} \tag{3.2}$$

Thus maximizing the likelihood is equivalent to optimizing a certain polynomial on a compact set. Hence a (global) maximum location indeed exists. 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.

Proposition 3.1.1. Let ν be a probability measure on $\mathbb{Z}_+^p$, $\theta_0 \in \Theta$, and $G_0 \in \mathscr{G}$ with $\mu_{G_0} < \infty$ and $g_0(0) < 1$. Then 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,

$$\hat\theta_n \xrightarrow{p} \theta_0 \text{ and } \sum_{k=0}^{\infty} \left|\hat g_n(k) - g_0(k)\right| \xrightarrow{p} 0, \text{ under } \mathbb{P}_{\nu,\theta_0,G_0}. \tag{3.3}$$

Proof.
Let $(\hat\theta_n, \hat G_n)$ be a maximum likelihood estimator of (θ, G). It is easy to see, and well-known, that to prove (3.3) 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 that this pointwise convergence holds by an application of Wald's consistency proof. This method works best for compact parameter spaces. Therefore we introduce $\tilde{\mathscr{G}}$: the class of all probability distributions on $\mathbb{Z}_+ \cup \{\infty\}$. Associate to each $G \in \tilde{\mathscr{G}}$ the sequence $(g(k))_{k\in\mathbb{Z}_+}$. Notice that this correspondence is 1-to-1, since $g(\infty) = 1 - \sum_{k=0}^{\infty} g(k)$. So we can regard $\tilde{\mathscr{G}}$ as a subset of $[0,1]^{\mathbb{Z}_+}$ equipped with the norm $\|a\| = \sum_{k=0}^{\infty} 2^{-k}|a(k)|$, i.e. 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 $\tilde{\mathscr{G}}$ is a compact subset of $[0,1]^{\mathbb{Z}_+}$. Define $\bar E = [0,1]^p \times \tilde{\mathscr{G}}$, and equip $\bar E$ with the 'sum-distance' $d((\theta,G),(\theta',G')) = |\theta-\theta'| + \|(g(k))_{k\in\mathbb{Z}_+} - (g'(k))_{k\in\mathbb{Z}_+}\|$, and note that $\bar E$ is compact. For $G \in \tilde{\mathscr{G}}$ define $P_{x,\infty}^{\theta,G} = 1 - \sum_{j\in\mathbb{Z}_+} P_{x,j}^{\theta,G} = g(\infty)$ for $x \in \mathbb{Z}_+^p$ and $P_{x,\infty}^{\theta,G} = 1$ if $\max_{i=1}^p x_i = \infty$. Define $m^{\theta,G}(x_{-p},\dots,x_0) = \log P^{\theta,G}_{(x_{-1},\dots,x_{-p}),x_0}$. And define the (random) function $M_n : \bar E \to [-\infty,\infty)$ by

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

and the function $M : \bar{E} \to [-\infty, \infty)$ by (by Theorem 1.1 ν_{θ_0,G_0} exists)

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

The following holds.

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

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

(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 $P^{\theta,G}_{(X_{-1},\ldots,X_{-p}),X_0} = P^{\theta_0,G_0}_{(X_{-1},\ldots,X_{-p}),X_0}$ $\mathbb{P}_{\nu_{\theta_0,G_0},\theta_0,G_0}$-a.s. $\implies$ $(\theta, G) = (\theta_0, G_0)$, this easily follows using the following well-known argument (recall that $Y_t = (X_{t-1},\ldots,X_{t-p})'$, and use $\log x \le 2(\sqrt{x}-1)$ for $x \ge 0$):

$$\begin{aligned} M(\theta,G) - M(\theta_0,G_0) &\le 2\mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0}\left(\sqrt{\frac{P^{\theta,G}_{Y_0,X_0}}{P^{\theta_0,G_0}_{Y_0,X_0}}} - 1\right) \\ &= 2\sum_{y \in \mathbb{Z}_+^p} \nu_{\theta_0,G_0}\{y\} \sum_{x_0=0}^{\infty} \sqrt{P^{\theta,G}_{y,x_0} P^{\theta_0,G_0}_{y,x_0}} - 2 \\ &\le -\sum_{y \in \mathbb{Z}_+^p} \nu_{\theta_0,G_0}\{y\} \sum_{x_0=0}^{\infty} \left(\sqrt{P^{\theta,G}_{y,x_0}} - \sqrt{P^{\theta_0,G_0}_{y,x_0}}\right)^2 \le 0. \end{aligned}$$

(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) (in this proof the law of large numbers for the i.i.d. case has to be replaced by an appropriate strong law of large numbers for Markov chains). Hence we obtain $d((\hat{\theta}_n, \hat{G}_n), (\theta_0, G_0)) \xrightarrow{p} 0$, which easily yields $\hat{\theta}_n \xrightarrow{p} \theta_0$ and, for all $k \in \mathbb{Z}_+$, $\hat{g}_n(k) \xrightarrow{p} g_0(k)$. □

Limit distribution 3.2

Next we investigate whether the NPMLE has a limiting distribution. To this end, we first have to specify which topology we use. We identify $G \in \mathscr{G}$ 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, $\operatorname{lin}\mathscr{G}$ 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, $\Theta \times \mathscr{G}_{p+4}$, 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 $\mathbb{R}^p \times \ell^1(\mathbb{Z}_+)$.

3.2.1 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.

To show that the NPMLE is a Z-estimator, we consider certain (artificial) submodels of the semiparametric model and subsequently exploit the fact that the maximum likelihood estimator also maximizes, by construction, the likelihood in these submodels. These submodels are such that the maximum is taken in a stationary point, which yields a score equation.

Fix the 'truth' $(\theta_0, G_0) \in \Theta \times \mathscr{G}_{p+4}$. And fix a realization ω, which yields the data $X_1(\omega), \ldots, X_n(\omega)$ and $(\hat\theta_n(\omega), \hat G_n(\omega))$, the realization of the maximum likelihood estimator. If $\hat\theta_n(\omega) \in \Theta$ we obtain, since $(\hat\theta_n(\omega), \hat G_n(\omega))$ maximizes the likelihood and Θ is open,

$$\frac{1}{n}\sum_{t=0}^{n} \dot\ell_\theta(X_{t-p}(\omega), \ldots, X_t(\omega); \hat\theta_n(\omega), \hat G_n(\omega)) = 0,$$

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

$$\dot\ell_\theta(x_{t-p}, \ldots, x_t; \theta, G) = \frac{\partial}{\partial\theta}\log\left(P^{\theta,G}_{(x_{t-1},\ldots,x_{t-p}),x_t}\right) 1\{P^{\theta,G}_{(x_{t-1},\ldots,x_{t-p}),x_t} > 0\}.$$

By Proposition 3.1.1 we have $\mathbb{P}_{\nu_{\theta_0,G_0},\theta_0,G_0}\{\hat\theta_n \in \Theta\} \to 1$. In Chapter 2 we showed, motivated by an 'information-loss' interpretation of the model, that this θ-part of the transition-score can be represented as,

$$\dot\ell_\theta(x_{t-p}, \ldots, x_t; \theta, G) = \begin{pmatrix} \mathbb{E}_{\theta,G}\left[\dot s_{X_{t-1},\theta_1}(\vartheta_1 \circ X_{t-1}) \mid X_t = x_t, \ldots, X_{t-p} = x_{t-p}\right] \\ \vdots \\ \mathbb{E}_{\theta,G}\left[\dot s_{X_{t-p},\theta_p}(\vartheta_p \circ X_{t-p}) \mid X_t = x_t, \ldots, X_{t-p} = x_{t-p}\right] \end{pmatrix},$$

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

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

The conditional expectation representation of the transition-score is heavily used later on. Obtaining score-equations for the G-direction is more difficult.

Construct (artificial) probability distributions on $\mathbb{Z}_+$, in direction $h : \mathbb{Z}_+ \to \mathbb{R}$, $h \neq 0$ and bounded, by

$$g_s(k,\omega) = g_s(k,h,\omega) = \left[1 + s\left(h(k) - \int h \, \mathrm{d}\hat{G}_n(\omega)\right)\right] \hat{g}_n(k,\omega),$$

for $k \in \mathbb{Z}_+$, $|s| < (2\|h\|_\infty)^{-1}$. Note that $g_0 = \hat{g}_n$ and $G_s(\omega) \in \mathscr{G}_{p+4}$ for all s. By construction $(\hat{\theta}_n(\omega), G_s(\omega))$ satisfies, for all s, the constraints of the optimization problem (3.2). Since $s = 0$ corresponds to $(\hat{\theta}_n(\omega), \hat{G}_n(\omega))$, which is a global maximum location *if* the outcome is ω, we obtain

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

To obtain a useful representation of this derivative, we recall from Chapter 2 that we have the representation,

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

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

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

Hence we obtain, if the realization is ω,

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{t=0}^{n} \frac{\partial}{\partial s} \log P^{\hat{\theta}_n(\omega), G_s(\omega)}_{(X_{t-1},\ldots,X_{t-p}),X_t} \bigg|_{s=0} \\ &= \frac{1}{n} \sum_{t=0}^{n} \left(A_{\hat{\theta}_n(\omega),\hat{G}_n(\omega)} h(X_{t-p},\ldots,X_t) - \int h \, \mathrm{d}\hat{G}_n(\omega) \right). \end{aligned}$$

Since this holds for all realizations ω (for different realizations different paths $s \mapsto G_s(\omega)$ are used) we obtain, for every bounded function $h : \mathbb{Z}_+ \to \mathbb{R}$, a moment condition:

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

Let $\mathscr{H}_1$ be the unit ball of $\ell^\infty(\mathbb{Z}_+)$, i.e. all functions $h : \mathbb{Z}_+ \to \mathbb{R}$ that satisfy $\sup_{e \in \mathbb{Z}_+} |h(e)| \leq 1$. We will only use the moment conditions arising from $h \in \mathscr{H}_1$. We summarize these in an estimating equation $\Psi_n = (\Psi_{n1}, \Psi_{n2}) : [0,1]^p \times \mathscr{G}_{p+4} \to \mathbb{R}^p \times \ell^\infty(\mathscr{H}_1)$ defined by

$$\begin{aligned} \Psi_{n1}(\theta, G) &= \frac{1}{n} \sum_{t=0}^{n} \dot{\ell}_\theta(X_{t-p},\ldots,X_t; \theta, G), \\ \Psi_{n2}(\theta, G) h &= \frac{1}{n} \sum_{t=0}^{n} \left(A_{\theta,G} h(X_{t-p},\ldots,X_t) - \int h \, \mathrm{d}G \right), \quad h \in \mathscr{H}_1. \end{aligned}$$

Indeed, $\Psi_{n2}(\theta, G)$ is an element of $\ell^\infty(\mathscr{H}_1)$ since $\sup_{h\in\mathscr{H}_1}|\Psi_{n2}(\theta,G)h| \leq 2$. From the discussion above we know that any maximum likelihood estimator satisfies $\Psi_{n2}(\hat\theta_n, \hat G_n) = 0$, and from Proposition 3.1.1 we have $\mathbb{P}_{\nu_{\theta_0,G_0},\theta_0,G_0}\{\Psi_{n1}(\hat\theta_n,\hat G_n) = 0\} \to 1$. For $(\theta_0, G_0) \in \Theta \times \mathscr{G}_{p+4}$ we introduce the 'limit' of the estimating equation: $\Psi^{\theta_0,G_0} : [0,1]^p \times \mathscr{G}_{p+4} \to \mathbb{R}^p \times \ell^\infty(\mathscr{H}_1)$ by,

$$\Psi_1^{\theta_0,G_0}(\theta, G) = \mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0}\dot\ell_\theta(X_{-p},\ldots,X_0;\theta,G),$$
$$\Psi_2^{\theta_0,G_0}(\theta, G)h = \mathbb{E}_{\nu_{\theta_0,G_0},\theta_0,G_0}\left(A_{\theta,G}h(X_{-p},\ldots,X_0) - \int h\,\mathrm{d}G\right), \quad h \in \mathscr{H}_1.$$

It is easy to see that

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

which is the usual result that, under the truth, scores have expectation zero.

3.2.2 Asymptotic normality

In this section we exploit that the NPMLE can be seen as a solution to the estimating equation Ψ_n. The following lemma is the key result of this chapter. It establishes conditions to an asymptotic normality theorem for infinite-dimensional M-estimators. Compared to a semiparametric analysis where one only wants to estimate the Euclidean part of the parameter, we now have to deal with functional calculus instead of Euclidean calculus, and with empirical processes instead of weak convergence in Euclidean spaces.

Lemma 3.2.1. Let $(\theta_0, G_0) \in \Theta \times \mathscr{G}_{p+4}$. Denote $\nu_0 = \nu_{\theta_0,G_0}$. Then the following properties hold.

(L1) The map $\Psi^{\theta_0,G_0} : [0,1]^p \times \mathscr{G}_{p+4} \to \mathbb{R}^p \times \ell^\infty(\mathscr{H}_1)$ is Fréchet-differentiable at (θ_0, G_0), i.e.

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

as $(\theta, G) \to (\theta_0, G_0)$ within $\Theta\times\mathscr{G}_{p+4}$ where $\dot\Psi^0 = \dot\Psi^{\theta_0,G_0} : \mathrm{lin}\left([0,1]^p \times \mathscr{G}_{p+4}\right) \to \mathbb{R}^p \times \ell^\infty(\mathscr{H}_1)$ is a continuous, linear mapping given by

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

where $\dot\Psi^0_{11} : \mathbb{R}^p \to \mathbb{R}^p$, $\dot\Psi^0_{12} : \mathrm{lin}\,\mathscr{G}_{p+4} \to \mathbb{R}^p$, $\dot\Psi^0_{21} : \mathbb{R}^p \to \ell^\infty(\mathscr{H}_1)$, and $\dot\Psi^0_{22} : \mathrm{lin}\,\mathscr{G}_{p+4} \to \ell^\infty(\mathscr{H}_1)$ are defined by

$$\dot\Psi^0_{11}(\theta-\theta_0) = -\left(\mathbb{E}_{\nu_0,\theta_0,G_0}\dot\ell_\theta\dot\ell_\theta^T(X_{-p},\ldots,X_0;\theta_0,G_0)\right)(\theta-\theta_0),$$
$$\dot\Psi^0_{12}(G-G_0) = -\int \mathbb{E}_{\nu_0,\theta_0}\left[\dot\ell_\theta(X_{-p},\ldots,X_0;\theta_0,G_0) \mid \varepsilon_0 = e\right]\mathrm{d}(G-G_0)(e),$$

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

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

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

where we use the following version of conditional probabilities, for $G \in \mathscr{G}$ and $x_{-p},\dots,x_0,e \in \mathbb{Z}_+$,

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

(L2) The inverse $\dot\Psi^{-1}_{\theta_0,G_0}$: Range$(\dot\Psi^{\theta_0,G_0}) \to \mathrm{lin}\left(\Theta \times \mathscr{G}_{p+4}\right)$ exists and is continuous[1].

(L3) We have, under $\mathbb{P}_{\nu_0,\theta_0,G_0}$,

$$\mathbb{S}^{\theta_0,G_0}_n = \sqrt{n}(\Psi_n(\theta_0,G_0) - \Psi^{\theta_0,G_0}(\theta_0,G_0)) \rightsquigarrow \mathbb{S}^{\theta_0,G_0} \quad \text{in } \mathbb{R}^p \times \ell^\infty(\mathscr{H}_1),$$

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

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

The next subsection is devoted to the proof of the lemma. Let us briefly comment on some elements of this proof. 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,

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

for all $G \in \mathscr{G}$. These representations are also crucial in the proof of (L2). 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. The process $\mathbb{S}^{\theta_0,G_0}_n$ can be interpreted as a 'score process', since its marginals are elements of the tangent space (see Section 3.3). Since all conditions to an infinite-dimensional version of Huber's classical theorem on asymptotic normality of M-estimators hold, we obtain the next theorem.

[1] $\dot\Psi^{-1}_{\theta_0,G_0}$ has a unique continuous extension to the closure of 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 $\mathrm{lin}\left([0,1]^p \times \mathscr{G}_{p+4}\right)$.

Theorem 3.1. *For* $(\theta_0, G_0) \in \Theta \times \mathcal{G}_{p+4}$ *we have*

$$\sqrt{n}\big((\hat\theta_n, \hat G_n) - (\theta_0, G_0)\big) = -\dot\Psi^{-1}_{\theta_0,G_0}\mathbb{S}_n^{\theta_0,G_0} + o(1;\mathbb{P}_{\nu_{\theta_0,G_0},\theta_0,G_0})$$
$$\rightsquigarrow -\dot\Psi^{-1}_{\theta_0,G_0}\mathbb{S}^{\theta_0,G_0}, \tag{3.4}$$

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

Proof.
Proposition 3.1.1 and Lemma 3.2.1 show that all conditions to Theorem 3.3.1 in Van der Vaart and Wellner (1993) are satisfied, which yields the result. □

3.2.3 Proof of Lemma 3.2.1

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. In order to conserve space we sometimes use the processes $Y_t = (X_{t-1},\dots,X_{t-p})'$ and $Z_t = (X_t,\dots,X_{t-p})'$, $t \geq 0$.

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. From Chapter 2 we have, here $\dot\ell_{\theta,i}$ refers to the ith coordinate of the p-vector $\dot\ell_\theta$,

$$\left|\dot\ell_{\theta,i}(x_{-p},\dots,x_0;\theta,G)\right| \leq \frac{x_{-i}}{\theta_i(1-\theta_i)}, \tag{3.5}$$

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

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

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 $H, G \in \operatorname{lin}\mathcal{G}_{p+4}$,

$$|\dot\Psi_{12}(G-H)| = \left|\int \mathbb{E}_{\nu_0,\theta_0}\left[\dot\ell_\theta(X_{-p},\dots,X_0;\theta_0,G_0) \mid \varepsilon_0 = e\right] \mathrm{d}(H-G)(e)\right|$$

$$\leq C \sum_{e=0}^{\infty} |h(e) - g(e)| = C\|H - G\|_1,$$

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

Step 2:
Rewrite,

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

Let θ_n be a sequence in $[0,1]^p$ converging to θ_0 and G_n a sequence in $\mathcal{G}_{p+4}$ converging to G_0. In Step 2a we show that

$$\frac{\left|\Psi_1(\theta_n, G_n) - \Psi_1(\theta_0, G_n) - \dot{\Psi}_{11}(\theta_n - \theta_0)\right|}{|\theta_n - \theta_0| + \|G_n - G_0\|_1} \to 0, \tag{3.6}$$

and in Step 2b we show that

$$\frac{\left|\Psi_1(\theta_0, G_n) - \Psi_1(\theta_0, G_0) - \dot{\Psi}_{12}(G_n - G_0)\right|}{|\theta_n - \theta_0| + \|G_n - G_0\|_1} \to 0, \tag{3.7}$$

which will conclude the proof of Step 2.
Step 2a:
From Chapter 2 we recall that the usual information-identity holds, i.e.

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

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

$$\dot{\ell}_{\theta,i}(Z_0; \theta, G) - \dot{\ell}_{\theta,i}(Z_0; \theta_0, G) = \frac{\partial}{\partial \theta^T} \dot{\ell}_{\theta,i}(X_{-p}, \ldots, X_0; \tilde{\theta}_i(\theta, G), G)(\theta - \theta_0),$$

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

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

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 Chapter 2 we have the bound,

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

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

$$\frac{\left|\Psi_1(\theta_n, G_n) - \Psi_1(\theta_0, G_n) - \dot{\Psi}_{11}(\theta_n - \theta_0)\right|}{|\theta_n - \theta_0|} \leq \frac{\mathbb{E}_{\nu_0,\theta_0,G_0} \left|(I_\theta(\theta_0, G_0) + J(Z_0; \theta_n, G_n))(\theta_n - \theta_0)\right|}{|\theta_n - \theta_0|} \to 0,$$

which yields (3.6).

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 $H \in \mathscr{G}$)

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

From this we obtain the bound,

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

Since G_n is a sequence in $\mathscr{G}_{p+4}$ converging to G_0, we obtain, for fixed $x_{-p}, \ldots, x_{-1}$,

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

Furthermore, using (3.5),

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

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 (3.7).

Step 3:
Rewrite,

$$\begin{aligned}\Psi_2(\theta,G)-\Psi_2(\theta_0,G_0)-\dot\Psi_{21}(\theta-\theta_0)-\dot\Psi_{22}(G-G_0)\\ =\Psi_2(\theta,G)-\Psi_2(\theta_0,G)-\dot\Psi_{21}(\theta-\theta_0)\\ +\Psi_2(\theta_0,G)-\Psi_2(\theta_0,G_0)-\dot\Psi_{22}(G-G_0).\end{aligned}$$

Let θ_n be a sequence in $[0,1]^p$ converging to θ_0 and G_n a sequence in $\mathscr{G}_{p+4}$ converging to G_0. We will verify that

$$\frac{\sup_{h\in\mathscr{H}_1}\left|\Psi_2(\theta_n,G_n)h-\Psi_2(\theta_0,G_n)h-\dot\Psi_{21}(\theta_n-\theta_0)h\right|}{|\theta_n-\theta_0|+\|G_n-G_0\|_1}\to 0, \tag{3.8}$$

and,

$$\frac{\sup_{h\in\mathscr{H}_1}\left|\Psi_2(\theta_0,G_n)h-\Psi_2(\theta_0,G_0)h-\dot\Psi_{22}(G_n-G_0)h\right|}{|\theta_n-\theta_0|+\|G_n-G_0\|_1}\to 0, \tag{3.9}$$

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

$$\begin{aligned}&\Psi_2(\theta_n,G_n)h-\Psi_2(\theta_0,G_n)h-\dot\Psi_{21}(\theta_n-\theta_0)h\\ &\quad=\mathbb{E}_{\nu_0,\theta_0,G_0}\left(A_{\theta_n,G_n}h(Z_0)-A_{\theta_0,G_n}h(Z_0)+A_{\theta_0,G_0}h(Z_0)\dot\ell_\theta^T(Z_0;\theta_0,G_0)(\theta_n-\theta_0)\right).\end{aligned}$$

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

$$\begin{aligned}\frac{\partial}{\partial\theta_i}A_{\theta,G}h(X_{-p},\dots,X_0)&=\mathbb{E}_{\theta,G}\left[h(\varepsilon_0)\dot s_{X_{-i},\theta_i}(\vartheta_i\circ X_{-i})\mid X_0,\dots,X_{-p}\right]\\ &\quad-A_{\theta,G}h(X_{-p},\dots,X_0)\dot\ell_{\theta,i}(X_{-p},\dots,X_0;\theta,G),\end{aligned}$$

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

$$\begin{aligned}&\frac{\partial^2}{\partial\theta_j\partial\theta_i}A_{\theta,G}h(X_{-p},\dots,X_0)\\ &\quad=\mathbb{E}_{\theta,G}\left[h(\varepsilon_0)\dot s_{X_{-i},\theta_i}(\vartheta_i\circ X_{-i})\dot s_{X_{-j},\theta_j}(\vartheta_j\circ X_{-j})|X_0,\dots,X_{-p}\right]\\ &\qquad-\mathbb{E}_{\theta,G}\left[h(\varepsilon_0)\dot s_{X_{-i},\theta_i}(\vartheta_i\circ X_{-i})|X_0,\dots,X_{-p}\right]\dot\ell_{\theta,j}(X_{-p},\dots,X_0;\theta,G)\\ &\qquad-A_{\theta,G}h(X_{-p},\dots,X_0)\ddot\ell_{\theta,ij}(X_{-p},\dots,X_0;\theta,G)\\ &\qquad-\dot\ell_{\theta,i}(X_{-p},\dots,X_0;\theta,G)\frac{\partial}{\partial\theta_j}A_{\theta,G}h(X_{-p},\dots,X_0)\\ &\qquad+1\{i=j\}\mathbb{E}_{\theta,G}\left[h(\varepsilon_0)\ddot s_{X_{-i},\theta_i}(\vartheta_i\circ X_{-i})\mid X_0,\dots,X_{-p}\right],\end{aligned}$$

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

$$\left|\frac{\partial}{\partial\theta_i} A_{\theta,G} h(X_{-p},\ldots,X_0)\right| + \left|\frac{\partial^2}{\partial\theta_j\partial\theta_i} A_{\theta,G} h(X_{-p},\ldots,X_0)\right| \le C_\theta(X_{-i}^2 + X_{-j}^2). \qquad (3.10)$$

A second order Taylor expansion in θ yields

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

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 (3.10) it easily follows, using dominated convergence, that

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

Hence we obtain (3.8) once we show that

$$\frac{\sup_{h\in\mathscr{H}_1}\left|\sum_i(\theta_{ni}-\theta_{0i})\mathbb{E}_{\nu_0,\theta_0,G_0}\mathbb{E}_{\theta_0,G_n}\left[h(\varepsilon_0)\dot{s}_{X_{-i},\theta_{0i}}(\vartheta_i\circ X_{-i}) \mid Z_0\right]\right|}{|\theta_n-\theta_0| + \|G_n-G_0\|_1} \to 0, \qquad (3.11)$$

and,

$$\begin{aligned} \sup_{h\in\mathscr{H}_1}\left|\mathbb{E}_{\nu_0,\theta_0,G_0}\left(A_{\theta_0,G_n}h(Z_0)\dot{\ell}_\theta^T(Z_0;\theta_0,G_n) - A_{\theta_0,G_0}h(Z_0)\dot{\ell}_\theta^T(Z_0;\theta_0,G_0)\right)(\theta_n-\theta_0)\right| \\ = o(|\theta_n-\theta_0| + \|G_n-G_0\|_1), \qquad (3.12) \end{aligned}$$

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

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

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 $2X_{-i}/(\theta_{0,i}(1-\theta_{0,i}))$. By independence of ε_0 and $\vartheta_i\circ X_{-i}-\theta_{0,i}X_{-i}$ we obtain $\mathbb{E}_{\nu_0,\theta_0,G_0}h(\varepsilon_0)\dot{s}_{X_{-i},\theta_i}(\vartheta_i\circ X_{-i})=0$. Display (3.11) now easily follows using dominated convergence. In a similar fashion we obtain (3.12).

Step 3b:

Note first that we have

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

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

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

we see that for fixed $(X_{-p},\dots,X_{-1})$ $\sup_{h\in\mathscr{H}_1}|f^h(X_{-p},\dots,X_{-1};G_n)|/\|G_n-G_0\|_1\to 0$. Since $\sup_{h\in\mathscr{H}_1}|f^h(X_{-p},\dots,X_{-1};G_n)|/\|G_n-G_0\|_1$ is bounded by $2(X_{-p}+\cdots+X_{-1})$ which is ν_0-integrable, dominated convergence yields (3.9).

Proof of (L2)

First we prove (L2) for the case 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,

$$\dot{\Psi}_{12}(G-G_0) = -\int A_0^*\dot{\ell}_\theta(e)\,\mathrm{d}(G-G_0)(e), \tag{3.13}$$

$$\dot{\Psi}_{22}(G-G_0)h = -\int A_0^* A_0 h(e)\,\mathrm{d}(G-G_0)(e), \quad h \in \mathscr{H}_1, \tag{3.14}$$

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 $\dot{\Psi}_{22}$, (3.13) and (3.14) 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 $\dot{V} = \dot{\Psi}_{22} - \dot{\Psi}_{21}\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12} : \mathrm{lin}\,\mathscr{G}_{p+4} \to \ell^\infty(\mathscr{H}_1)$ 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 Theorem 2.1). 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 $C : \mathscr{H}_1 \to [\varepsilon]$ by

$$Ch(e) = -\left[\mathbb{E}_{\nu_0,\theta_0,G_0} A_0 h(Z_0)\dot{\ell}_\theta^T(Z_0;\theta_0,G_0)\right] I_{\theta_0}^{-1}(A_0^*(\dot{\ell}_\theta(\cdot;\theta_0,G_0)))(e),$$

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

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

The mapping $\dot{V}: \mathrm{lin}\mathscr{G}_{p+4} \to \ell^\infty(\mathscr{H}_1)$ has a continuous inverse on its range if and only if there exists $\epsilon > 0$ such that

$$\|\dot{V}(G-G_0)\| = \sup_{h\in\mathscr{H}_1} |\dot{V}(G-G_0)h| \geq \epsilon \|G-G_0\|_1, \quad \text{for all } G \in \mathrm{lin}\mathscr{G}_{p+4}.$$

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

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

Hence it suffices to prove that there exists $\epsilon > 0$ such that, for all $G \in \mathrm{lin}\mathscr{G}_{p+4}$,

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

Of course, a sufficient condition for this is $\epsilon\mathscr{H}_1 \subset \left\{(A_0^* A_0 + C)h \mid h \in \mathscr{H}_1\right\}$, 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 $\epsilon\mathscr{H}_1 \subset \left\{(A_0^* A_0 + C)h \mid h \in \mathscr{H}_1\right\}$ is equivalent to $\epsilon B^{-1}\mathscr{H}_1 \subset \mathscr{H}_1$. Since $\mathscr{H}_1$ 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 $h \in \mathscr{H}_1$. 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 $h \in \mathscr{H}_1$ 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,

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

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

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

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

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

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

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

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

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

Of course this is a linear operator. Moreover it is continuous since (remember that $P^{\theta_0,G_0}_{(0,\dots,0),x_0} = \mathrm{g}_0(x_0)$)

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

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 $L_2(\nu_0 \otimes P^{\theta_0,G_0})$. 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 (Theorem 2.1), $\mathbb{E}_{\nu_0,\theta_0,G_0} S^2 = a^T I_{\theta_0} a > 0$ 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

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

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, $g_0(0) > 0$.

Proof of (L3)

Proposition 1.2.1.1 yields the weak-convergence of $\sqrt{n}\left(\Psi_{n1} - \Psi_1^{\theta_0,G_0}\right)(\theta_0,G_0)$, since we are dealing with a finite function class and since we have the bound $|\dot\ell_{\theta,i}(Z_0;\theta_0,G_0)| \leq X_{-i}(\theta_{0,i}(1-\theta_{0,i}))^{-1}$, $i = 1,\ldots,p$. 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 weak convergence, under $\mathbb{P}_{\nu_0,\theta_0,G_0}$, of the process $\sqrt{n}\left(\Psi_{n2} - \Psi_2^{\theta_0,G_0}\right)(\theta_0,G_0)$ in $\ell^\infty(\mathcal{H}_1)$ to a tight Gaussian process. This can be reexpressed as the weak convergence of the empirical process

$\{\mathbb{Z}_n f \mid f \in \mathscr{F}\}$, where $\mathscr{F} = \{\mathbb{Z}_+^{p+1} \ni (x_{-p},\ldots,x_0) \mapsto A_0 h(x_{-p},\ldots,x_0) \mid h \in \mathscr{H}_1\}$. We use Proposition 1.2.1.5 to verify this. Let $\delta > 0$. By Markov's inequality we have

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

for $M_\delta = \lceil (8(p+1)\mathbb{E}_{\nu_0,\theta_0,G_0} X_0^{p+2})^{1/(p+2)}\delta^{-2/(p+2)} \rceil$. 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 $h \in \mathscr{H}_1$ 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 $h \in \mathscr{H}_1$ there exists $i \in \{1,\ldots,N_\delta\}$ such that $f_i^L \leq A_0 h \leq f_i^U$. So the brackets $[f_i^L, f_i^U]$, $i = 1,\ldots,N_\delta$, cover $\mathscr{F}$ 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 $N_{[\,]}(\delta,\mathscr{F}) \leq N_\delta$. Using $\log(x) \leq m(x^{1/m} - 1)$ for $x > 0$, $m \in \mathbb{N}$, it follows that we can find $\zeta < 1$ such that $\log N_{[\,]}(x,\mathscr{F}) = O(x^{-2\zeta})$. Since the envelope of $\mathscr{F}$ is bounded by 2, an application of Proposition 1.2.1.5 concludes the proof.

Proof of (L4)

In step A we prove

$$\sqrt{n}\left(\Psi_{n2} - \Psi_2^{\theta_0,G_0}\right)(\hat{\theta}_n,\hat{G}_n) - \sqrt{n}\left(\Psi_{n2} - \Psi_2^{\theta_0,G_0}\right)(\theta_0,G_0) = o(1;\mathbb{P}_{\nu_0,\theta_0,G_0}), \quad (3.15)$$

and in step B we prove

$$\sqrt{n}\left(\Psi_{n1} - \Psi_1^{\theta_0,G_0}\right)(\hat{\theta}_n,\hat{G}_n) - \sqrt{n}\left(\Psi_{n1} - \Psi_1^{\theta_0,G_0}\right)(\theta_0,G_0) = o(1;\mathbb{P}_{\nu_0,\theta_0,G_0}), \quad (3.16)$$

which will conclude the proof. Introduce for $\delta > 0$ $B_0(\delta) = \{(\theta,G) \in \Theta \times \mathscr{G}_{p+4} \mid |\theta - \theta_0| + \|G - G_0\|_1 \leq \delta\}$.

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

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

for all sequences (θ_n, G_n) in $\Theta \times \mathscr{G}_{p+4}$ converging to (θ_0, G_0), and that the empirical process $\{\mathbb{Z}_n f \mid f \in \mathscr{F}^\delta\}$ with $\mathscr{F}^\delta$ given by

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

weakly converges to a tight Gaussian process, then (3.15) follows from (the proof of) Lemma 3.3.5 in Van der Vaart and Wellner (1993). Since

$$\sup_{h\in\mathscr{H}_1} |A_{\theta_n,G_n}h(X_{-p},\dots,X_0) - A_{\theta_0,G_0}h(X_{-p},\dots,X_0)| \le 2,$$

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

$$\sup_{h\in\mathscr{H}_1} |A_{\theta_n,G_n}h(Z_0) - A_{\theta_0,G_0}h(Z_0)| \le \left|\frac{P^{\theta_0,G_0}_{Y_0,X_0}}{P^{\theta_n,G_n}_{Y_0,X_0}} - 1\right| + \frac{\|G_n - G_0\|_1}{P^{\theta_n,G_n}_{Y_0,X_0}} \to 0,$$

the first condition easily follows by an application of the dominated convergence theorem. That the process $\{\mathbb{Z}_n f \mid f \in \mathscr{F}^\delta\}$ 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

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

for all sequences (θ_n, G_n) in $\Theta \times \mathscr{G}_{p+4}$ converging to (θ_0, G_0), and that the empirical process $\{\mathbb{Z}_n f \mid f \in \mathscr{F}^\delta\}$ with $\mathscr{F}^\delta$ given by

$$\mathscr{F}^\delta = \left\{z_0 \mapsto \dot\ell_{\theta,1}(x_{-p},\dots,x_0;\theta,G) - \dot\ell_{\theta,1}(x_{-p},\dots,x_0;\theta_0,G_0) \mid (\theta,G) \in B_0(\delta)\right\},$$

converges weakly to a tight Gaussian process, then (3.16) follows from (the proof of) Lemma 3.3.5 in Van der Vaart and Wellner (1993). Choose $\delta > 0$ such that for all θ in the ball we have $(\theta_i(1-\theta_i))^{-1} \le C$ for certain $C > 0$ and all $i = 1,\dots,p$. The first condition easily follows using dominated convergence (use $4CX_{-1}^2$ as dominating function). We use Proposition 1.2.1.5 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

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

and using Hölder's inequality we now obtain

$$\mathbb{E}_{\nu_0,\theta_0,G_0}X_{-1}^2 1\{\max_{i=0,\dots,p} X_{-i} \ge M_\eta\}$$

$$\leq \left(\mathbb{E}_{\nu_0,\theta_0,G_0} X_{-1}^{p+4}\right)^{2/(p+4)} \left(\mathbb{P}_{\nu_0,\theta_0,G_0}\{\max_{i=0,\dots,p} X_{-i} \geq M_\eta\}\right)^{(p+2)/(p+4)} \leq \frac{\eta^2}{8C^2}. \quad (3.17)$$

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

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

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

$$f_i(z_0) - \frac{\eta}{2\sqrt{2}} \leq \dot{\ell}_{\theta,1}(z_0;\theta,G) - \dot{\ell}_{\theta,1}(z_0;\theta_0,G_0) \leq f_i(z_0) + \frac{\eta}{2\sqrt{2}}.$$

Next we introduce mappings f_i^L, f_i^U, $i = 1,\dots,N_\eta$, from $\mathbb{Z}_+^{p+1}$ into $\mathbb{R}$ by $f_i^L = -2CM_\eta \vee (f_i - \eta/2\sqrt{2})$ if $\max\{x_{-p},\dots,x_0\} \leq M_\eta - 1$ and $f_i^L = -2Cx_{-1}$ in case $\max\{x_{-p},\dots,x_0\} \geq M_\eta$, and $f_i^U = 2CM_\eta \wedge (f_i + \eta/2\sqrt{2})$ if $\max\{x_{-p},\dots,x_0\} \leq M_\eta - 1$ and $f_i^U = 2Cx_{-1}$ if $\max\{x_{-p},\dots,x_0\} \geq M_\eta$. Conclude that for $(\theta,G) \in B_0(\delta)$ there exists $i \in \{1,\dots,N_\eta\}$ 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 $[f_i^L, f_i^U]$, $i = 1,\dots,N_\eta$, cover $\mathscr{F}^\delta$ and satisfy, by (3.17),

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

Conclude that $N_{[]}(\eta, \mathscr{F}^\delta) \leq N_\eta$. Using $\log(x) \leq m(x^{1/m} - 1)$ for $x > 0$, $m \in \mathbb{N}$, it easily follows that we can find $\zeta < 1$ such that $\log N_{[]}(x, \mathscr{F}^\delta) = O(x^{-2\zeta})$. Since the envelope of $\mathscr{F}^\delta$ is bounded by the integrable variable $2CX_{-1}$, an application of Proposition 1.2.1.5 concludes the proof.

3.3 Efficiency

In this section we prove efficiency of $(\hat{\theta}_n, \hat{G}_n)$. As mentioned in the introduction it is a nonstandard problem to demonstrate efficiency. This since 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' $\mathbb{S}_n^{\theta,G}$ can be seen as an efficient estimator of a certain artificial parameter, and that efficiency is retained under Hadamard differentiable mappings.

It is well-known that the local structure of a model needs to be considered to obtain lower-bounds to the precision of estimators. Tangent spaces are the mathematical tool for this. The next lemma yields a tangent space: it shows that certain parametric submodels enjoy the LAN-property.

Lemma 3.3.1. Let $(\theta, G) \in \Theta \times \mathscr{G}_{p+4}$. 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 $|\tau| < \tilde{\epsilon}$, $G_\tau \in \mathscr{G}_{p+4}$. Let $0 < \epsilon \le \tilde{\epsilon}$ be such that $\theta + \tau a \in \Theta$ for $|\tau| \le \epsilon$, and denote $\nu^\tau = \nu_{\theta+\tau a, G_\tau}$. Then the sequence of experiments

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

has the LAN-property at $\tau = 0$ (recall that $Z_t = (X_t, \ldots, X_{t-p})'$):

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

where $\theta_n = \theta + a/\sqrt{n}$, $G_n = G_{1/\sqrt{n}}$, and $\nu_n = \nu^{1/\sqrt{n}}$, and where $J_{\theta,G,h}$ is given by,

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

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

$$\mathscr{T}^0_{\theta,G} = \left\{a^T\dot{\ell}_\theta(Z_0;\theta,G) + A_{\theta,G}h(Z_0) - \int h\,\mathrm{d}G \mid a \in \mathbb{R}^p, h \in \ell^\infty(\mathbb{Z}_+)\right\},$$

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

Proof.
By an application of Theorem 2.1 the lemma is proved once we prove that the initial value satisfies $\nu_n\{X_{-p},\ldots,X_{-1}\} - \nu_{\theta,G}\{X_{-p},\ldots,X_{-1}\} \xrightarrow{p} 0$, under $\mathbb{P}_{\nu_{\theta,G},\theta,G}$. By Proposition 1.2.1.6 this follows if we show (recall that $Y_t = (X_{t-1},\ldots,X_{t-p})'$)

$$\lim_{n\to\infty}\sup_{y\in\mathbb{Z}_+^p}\frac{\sup_{f:|f|\le V}\left|\mathbb{E}_{\delta_y,\theta_n,G_n}f(Y_1) - \mathbb{E}_{\delta_y,\theta,G}f(Y_1)\right|}{V(y)} = 0, \tag{3.18}$$

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

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

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

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

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

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

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

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

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

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

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

for a constant $C>0$. A combination of the previous four displays easily yields (3.18). □

Now we are able to recall the concept of a regular estimator for (θ,G): 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,

$$\sqrt{n}(T_n-(\theta_n,G_n))\rightsquigarrow\mathbb{Z}\text{ under }\mathbb{P}_{\nu_n,\theta_n,G_n},\tag{3.19}$$

where $\theta_n = \theta + a/\sqrt{n}$, $g_n = g(1 + (h - \int h \, dG)/\sqrt{n})$, and $\nu_n = \nu_{\theta_n, G_n}$. An interpretation of (3.19) 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 $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, $(\theta, G) \in \Theta \times \mathcal{G}_{p+4}$. Since Lemma 3.3.1 established 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 (1991b) Theorem 2.1).

Theorem 3.2. *Let $(\theta, G) \in \Theta \times \mathcal{G}_{p+4}$, and T_n 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 $\mathbb{P}_{\nu_{\theta,G},\theta,G}$, $(\theta, G) \in \Theta \times \mathcal{G}_{p+4}$.

Using Le Cam's third lemma and Lemma 3.3.1 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 $\mathbb{E}_G h^*(\varepsilon_1) = 0$,

$$\dot{\Psi}_1^{\theta,G}\left(a, \left(k \mapsto h^*(k) g(k)\right)\right) = -\mathbb{E}_{\nu_{\theta,G},\theta,G}\left(a^T \dot{\ell}_\theta(Z_0; \theta, G) a + A_{\theta,G} h^*(Z_0)\right) \dot{\ell}_\theta(Z_0; \theta, G),$$

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

$$\dot{\Psi}_2^{\theta,G}\left(a, \left(k \mapsto h^*(k) g(k)\right)\right) h = -\mathbb{E}_{\nu_{\theta,G},\theta,G}\left(a^T \dot{\ell}_\theta(Z_0; \theta, G) + A_{\theta,G} h^*(Z_0)\right) A_{\theta,G} h(Z_0).$$

These displays 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. Plugging in the definitions of $\dot{\Psi}_1^{\theta,G}$ and $\dot{\Psi}_2^{\theta,G}$, these displays are easily checked. We organize the result in the following proposition.

Proposition 3.3.1. Let $(\theta, G) \in \Theta \times \mathscr{G}_{p+4}$. Any NPMLE $((\hat{\theta}_n, \hat{G}_n))_{n\in\mathbb{Z}_+}$ is a regular estimator of (θ, G) at $\mathbb{P}_{\nu_{\theta,G},\theta,G}$.

To prove efficiency we first recall the following characterization of efficiency. Fix $(\theta_0, G_0) \in \Theta \times \mathscr{G}_{p+4}$ 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 $\mathscr{T}_{\theta_0,G_0}$. More precise: there should exist $f_1, \dots, f_p \in \mathscr{T}_{\theta_0,G_0}$ and h_k, $k \in \mathbb{Z}_+$ from $\mathscr{T}_{\theta_0,G_0}$ such that

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} \begin{pmatrix} f_1(X_{t-p}, \dots, X_t) \\ \cdots \\ f_p(X_{t-p}, \dots, X_t) \end{pmatrix} + o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}), \tag{3.20}$$

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

$$\sqrt{n}(\hat{g}_n(k) - g(k)) = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} h_k(X_{t-p}, \dots, X_t) + o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}). \tag{3.21}$$

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)

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_n - \theta_0 \\ (\hat{g}_n(k) - g_0(k))_{k\in\mathbb{Z}_+} \end{pmatrix} = -\dot{\Psi}^{-1}_{\theta_0,G_0} \mathbb{S}_n^{\theta_0,G_0} + o(1; \mathbb{P}_{\nu_0,\theta_0,G_0}), \tag{3.22}$$

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

$$\mathbb{S}_{n1}^{\theta_0,G_0} = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} \dot{\ell}_\theta(X_{t-p}, \dots, X_t; \theta_0, G_0), \tag{3.23}$$

and the infinite-dimensional part by,

$$\mathbb{S}_{n2}^{\theta_0,G_0} h = \frac{1}{\sqrt{n}} \sum_{t=0}^{n} \left(A_{\theta_0,G_0} h(X_{t-p}, \dots, X_t) - \int h \, \mathrm{d}G_0 \right), \quad h \in \mathscr{H}_1. \tag{3.24}$$

So $\mathbb{S}_n^{\theta_0,G_0}$ is a process of certain elements of the tangent space. Introduce the *artificial* parameters (notice that we use ν_0 instead of $\nu_{\theta,G}$)

$$\Theta \times \mathscr{G}_{p+4} \ni (\theta, g) \mapsto \nu_1^{\theta_0,G_0}(\theta, g) = \mathbb{E}_{\nu_0,\theta,G} \dot{\ell}_\theta(X_{t-p}, \dots, X_0; \theta_0, G_0),$$

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

$$\Theta \times \mathscr{G}_{p+4} \ni (\theta, g) \mapsto \nu_h^{\theta_0,G_0}(\theta, g) = \mathbb{E}_{\nu_0,\theta,G} A_{\theta_0,G_0} h(X_{t-p}, \dots, X_0) - \int h \, \mathrm{d}G_0.$$

And note that $\nu_1^{\theta_0,G_0}(\theta_0, g_0) = \nu_h^{\theta_0,G_0}(\theta_0, g_0) = 0$. From (3.23) we now see 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 $\mathcal{T}_{\theta_0,G_0}$. And, from (3.24) we see 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* $\mathbb{P}_{\nu_0,\theta_0,G_0}$ one we show that they are regular at $\mathbb{P}_{\nu_0,\theta_0,G_0}$. Using Le Cam's third lemma and Lemma 3.3.1 this regularity follows once we show that for all $a \in \mathbb{R}^p$ and $f \in \ell^\infty(\mathbb{Z}_+)$ with $\mathbb{E}_{G_0} f(\varepsilon_1) = 0$, we have,

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

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

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

which are quite straightforward to check (see also the proof of Lemma 3.3.1). 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 3.2.1L3), and 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 (3.22) we see 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 $\sqrt{n}(\hat{\theta}_n - \theta_0, (\hat{g}_n(k) - g_0(k))_{k\in\mathbb{Z}_+})$, at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, an efficient estimator of *a certain* parameter (for details we refer to the proof of Theorem 3 in Van der Vaart (1995)). Hence 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, at $\mathbb{P}_{\nu_0,\theta_0,G_0}$, contained in the tangent space $\mathcal{T}_{\theta_0,G_0}$, which yields (3.20) and (5.19). Since we already proved regularity this proves efficiency of the NPMLE at $\mathbb{P}_{\nu_0,\theta_0,G_0}$. Since $(\theta_0, G_0) \in \Theta \times \mathcal{G}_{p+4}$ was arbitrary, we obtain the following theorem.

Theorem 3.3. *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 Lemma 3.2.1 and Theorem 3.2), for all* $(\theta, G) \in \Theta \times \mathcal{G}_{p+4}$,

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

3.4 Monte Carlo study & Empirical Application

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 3.3): 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 the state 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 3.1 presents the results for $n = 500$, and Table 3.2 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 3.1 with the corresponding entries in Table 3.2, 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 200%. This corroborates the result of Chapter 4 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 $\vartheta_1 \circ X_{t-1}$. Asymp-

Table 3.1: 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 = 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 = 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 = 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 = 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

totically, 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 Mon-

Table 3.2: 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 3.3: 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

day 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 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 3.3 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 $\hat{\mathbb{E}}[X_t \mid X_{t-1}] \approx 0.24 + 0.52$, and the following estimated conditional variance $\hat{\mathrm{var}}[X_t \mid X_{t-1}] \approx 0.18X_{t-1} + 0.70$. 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 0.24, means that a quote arriving at time t 'generates' a new quote at period $t+1$ with probability 0.25. The estimates $\hat{g}_n(k)$ give the probability on k 'new unpredictable' quotes.

4 The limit experiment of nearly unstable INAR(1) models

Recall that the INAR(1) process starting at 0 is defined by $X_0 = 0$ and the recursion,

$$X_t = \vartheta \circ X_{t-1} + \varepsilon_t, \quad t \in \mathbb{N}, \tag{4.1}$$

where,

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

Here $(Z_j^{(t)})_{j\in\mathbb{N},t\in\mathbb{N}}$ is a collection of i.i.d. Bernoulli variables with success probability $\theta \in [0,1]$, independent of the i.i.d. innovation sequence $(\varepsilon_t)_{t\in\mathbb{N}}$ with distribution G on $\mathbb{Z}_+ = \mathbb{N} \cup \{0\}$. All these variables are defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P}_{\theta,G})$. If we work with fixed G, we drop the subscript G. From Theorem 1.1 we know that, if $\theta \in [0,1)$ and $\mathbb{E}_G\varepsilon_1 < \infty$, which is called the 'stable' case, there exists an initial distribution, $\nu_{\theta,G}$, such that X is stationary if $\mathscr{L}(X_0) = \nu_{\theta,G}$. Of course, the INAR(1) process is non-stationary if $\theta = 1$: under $\mathbb{P}_1$ the process X is nothing but a standard random walk with drift on $\mathbb{Z}_+$ (but note that X is nondecreasing under $\mathbb{P}_1$). We call this situation 'unstable' or say that the process has a 'unit root'. Although the unit root is on the boundary of the parameter space, it is an important parameter value since Hellström (2001) documented that in many applications the estimates of θ are close to 1.

As before, we denote the law of $(X_0,\ldots,X_n)$ under $\mathbb{P}_{\theta,G}$ (on the measurable space $(\mathscr{X}_n, \mathscr{A}_n) = (\mathbb{Z}_+^{n+1}, 2^{\mathbb{Z}_+^{n+1}})$) by $\mathbb{P}_{\theta,G}^{(n)}$. In the applications in this chapter we mainly consider two sets of assumptions on G: (i) G is known or (ii) G is completely unknown (apart from some regularity conditions). For expository reasons, let us, for the moment, focus on the case that G is completely known and that the goal is to estimate θ. We use 'local-to-unity' asymptotics to take the 'increasing statistical difficulty' in the neighborhood of the unit root into account, i.e. we consider local alternatives to the unit root in such a way that the increasing degree of difficulty to discriminate between these alternatives and the unit root compensates the increase of information contained in the sample as the number of observations grows. This approach is well-known; it traces back to the work of Chan and Wei (1987) and Philips (1987), who studied the behavior of a given estimator (OLS) in a nearly unstable AR(1) setting, and Jeganathan (1995), whose results yield the asymptotic structure of nearly unstable AR models. Following this approach, we introduce the sequence of nearly unstable INAR experiments

$$\mathscr{E}_n(G) = \left(\mathscr{X}_n, \mathscr{A}_n, \left(\mathbb{P}^{(n)}_{1-h/n^2} \mid h \geq 0\right)\right), \quad n \in \mathbb{N}.$$

The 'localizing rate' n^2 will become apparent later on. It is surprising that the localizing rate is n^2, since for the classical nearly unstable AR(1) model one has rate $n\sqrt{n}$ (non-zero intercept) or n (no intercept). Suppose that we have found an estimator $\widehat{h}_n$ with 'nice properties', then this corresponds to the estimate $\widehat{\theta}_n = 1 - \widehat{h}_n/n^2$ of θ in the experiment of interest. To our knowledge, Ispány et al. (2003b) were the first to study estimation in a nearly unstable INAR(1) model. These authors study the behavior of the OLS estimator and they use a localizing rate n instead of n^2. However, as we will see shortly, n^2 is indeed the proper localizing rate and in Proposition 4.3.4 we show that the OLS estimator is an exploding estimator in $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$, i.e. it has not even the 'right' rate of convergence. The question then arises how we should estimate h. Instead of analyzing the asymptotic behavior of a given estimator, we derive the asymptotic structure of the experiments themselves by determining the limit experiment (in the Le Cam sense) of $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$. This limit experiment gives bounds to the accuracy of inference procedures and suggests how to construct efficient ones.

The main goal of this chapter is to determine the limit experiment of $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$. Remember that (see, for example, Le Cam (1986), Le Cam and Yang (1990), Van der Vaart (1991a), Shiryaev and Spokoiny (1999) or Van der Vaart (2000, Chapter 9)), the sequence of experiments $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$ is said to converge to a limit experiment (in Le Cam's weak topology) $\mathscr{E} = (\mathscr{X}, \mathscr{A}, (\mathbb{Q}_h \mid h \geq 0))$ if, for every finite subset $I \subset \mathbb{R}_+$ and every $h_0 \in \mathbb{R}_+$, we have

$$\left(\frac{\mathrm{d}\mathbb{P}^{(n)}_{1-\frac{h}{n^2}}}{\mathrm{d}\mathbb{P}^{(n)}_{1-\frac{h_0}{n^2}}}\right)_{h\in I} \xrightarrow{d} \left(\frac{\mathrm{d}\mathbb{Q}_h}{\mathrm{d}\mathbb{Q}_{h_0}}\right)_{h\in I}, \text{ under } \mathbb{P}^{(n)}_{1-\frac{h_0}{n^2}}.$$

To see that it is indeed reasonable to expect n^2 as the proper localizing rate, we briefly discuss the case of geometrically distributed innovations (in the remainder we treat general G). In case $G = \text{Geometric}(1/2)$, i.e., G puts mass $(1/2)^{k+1}$

at $k \in \mathbb{Z}_+$, it is an easy exercise[1] to verify for $h > 0$,

$$\frac{\mathrm{d}\mathbb{P}^{(n)}_{1-\frac{h}{r_n}}}{\mathrm{d}\mathbb{P}^{(n)}_1} \xrightarrow{p} \begin{cases} 0 & \text{if } \frac{r_n}{n^2} \to 0, \\ \exp\left(-\frac{hG(0)\mathbb{E}_G\varepsilon_1}{2}\right) = \exp\left(-\frac{h}{4}\right) & \text{if } \frac{r_n}{n^2} \to 1, \\ 1 & \text{if } \frac{r_n}{n^2} \to \infty, \end{cases} \quad \text{under } \mathbb{P}_1.$$

This simple calculation has two important implications. First, it indicates that n^2 is indeed the proper localizing rate. Intuitively, if we go faster than n^2 we cannot distinguish $\mathbb{P}^{(n)}_{1-h/r_n}$ from $\mathbb{P}^{(n)}_1$, and if we go slower we can distinguish $\mathbb{P}^{(n)}_{1-h/r_n}$ perfectly from $\mathbb{P}^{(n)}_1$. Secondly, since $\exp(-h/4) < 1$ we cannot, by Le Cam's first lemma, hope, in general, for contiguity of $\mathbb{P}^{(n)}_{1-h/n^2}$ with respect to $\mathbb{P}^{(n)}_1$ (Remark 12 after Theorem 4.1 gives an example of sets that yield this non-contiguity). This lack of contiguity is unfortunate for several reasons. Most importantly, if we would have contiguity the limiting behavior of the random vectors $(\mathrm{d}\mathbb{P}^{(n)}_{1-h/n^2}/\,\mathrm{d}\mathbb{P}^{(n)}_1)_{h\in I}$ determines the limit experiment, whereas we need to consider the behavior of $(\mathrm{d}\mathbb{P}^{(n)}_{1-h/n^2}/\,\mathrm{d}\mathbb{P}^{(n)}_{1-h_0/n^2})_{h\in I}$ for all $h_0 \geq 0$. So to be clear: the preceding display does not yet yield the limit experiment for this Geometric case. And it implies that the global sequence of experiments has not the common Local Asymptotic Quadratic structure (see Jeganathan (1995)) at $\theta = 1$. This differs from the traditional AR(1) process $Y_0 = 0$, $Y_t = \mu + \theta Y_{t-1} + u_t$, u_t i.i.d. $\mathrm{N}(0, \sigma^2)$, with $\mu \neq 0$ and σ^2 known, that enjoys this LAQ property at $\theta = 1$: the limit experiment at $\theta = 1$ is the usual normal location experiment (i.e., the model is Locally Asymptotically Normal) and the localizing rate is $n^{3/2}$. The limit experiment at $\theta = 1$ for $Y_0 = 0$, $Y_t = \theta Y_{t-1} + u_t$, u_t i.i.d. $\mathrm{N}(0, \sigma^2)$, with σ^2 known, does not have the LAN-structure; the limit experiment is of the Locally Asymptotically Brownian Functional type (see Jeganathan (1995)) and the localizing rate is n. Thus although the INAR(1) process and the traditional AR(1) process both are a random walk with drift at $\theta = 1$, their statistical properties 'near $\theta = 1$' are very different. In Section 4.2 we prove that the limit-experiment of $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$ corresponds to one draw from a Poisson distribution with mean $hG(0)\mathbb{E}_G\varepsilon_1/2$, i.e.

$$\mathscr{E}(G) = \left(\mathbb{Z}_+, 2^{\mathbb{Z}_+}, \left(\text{Poisson}\left(\frac{hG(0)\mathbb{E}_G\varepsilon_1}{2}\right) \mid h \geq 0\right)\right).$$

We indeed recognize $\exp(-hG(0)\mathbb{E}_G\varepsilon_1/2)$ as the likelihood ratio at h relative to $h_0 = 0$ in the experiment $\mathscr{E}(G)$. Due to the lack of enough smoothness of the likelihood ratios around the unit root, this convergence of experiments is not obtained by the usual (general applicable) techniques, but by a direct approach. Since the transition probability is the convolution of a Binomial distribution with G and the fact that certain Binomial experiments converge to

[1]The Geometric distribution allows us, using Newton's Binomial formula, to obtain explicit expressions for the transition-probabilities from X_{t-1} to X_t if $X_t \geq X_{t-1}$: $P^{\theta,G}_{X_{t-1},X_t} = 2^{-(X_t+1)}(1+\theta)^{X_{t-1}}$.

a Poisson limit experiment (see Remark 16 after Theorem 4.2 for the precise statement), one might be tempted to think that the convergence $\mathscr{E}_n(G) \to \mathscr{E}(G)$ follows, in some way, from this convergence. Remark 16 after Theorem 4.2 shows that this reasoning is not valid.

The remainder of this chapter is organized as follows. In Section 4.1 we discuss some preliminary properties which provide insight in the behavior of a nearly unstable INAR(1) process. The main result of this chapter is stated and proved in Section 4.2. Section 4.3 uses this result to analyze some estimation and testing problems. In Section 4.3.1 we consider efficient estimation of h, the deviation from a unit root, in the nearly unstable case for two settings. The first setting, discussed in Section 4.3.1, treats the case that the immigration distribution G is completely known. The second setting, analyzed in Section 4.3.1, considers a semiparametric model, where hardly any conditions on G are imposed. Since the INAR(1) process is a particular branching process with immigration, this also partially solves the question (see Wei and Winnicki (1990)) how to estimate the offspring mean efficiently.. Furthermore, we show in Section 4.3.1 that the OLS-estimator, considered by Ispány et al. (2003b, 2003a, 2005), is explosive. In Section 4.3.2 we provide an efficient estimator of θ in the 'global' model. Finally, we discuss testing for a unit root in Section 4.3.3. We show that the traditional Dickey-Fuller test has no (local) power, but that an intuitively obvious test is efficient.

4.1 Preliminaries

This section discusses some basic properties of nearly unstable INAR(1) processes. Besides giving insight in the behavior of a nearly unstable INAR(1) process, these properties are a key input in the next sections.

In this chapter we focus on the statistical properties of the INAR(1) process for parameter values θ close to one. To this end it is convenient to use another representation of the transition-probabilities (1.3). Since, conditional on $X_{t-1} = x_{t-1}$, the random variables ε_t and $\vartheta \circ X_{t-1}$ are independent, and $X_{t-1} - \vartheta \circ X_{t-1}$, 'the number of deaths during $(t-1,t]$', follows a Binomial$(X_{t-1}, 1-\theta)$ distribution, we obtain, for $x_{t-1}, x_t \in \mathbb{Z}_+$,

$$\begin{aligned} P^{\theta}_{x_{t-1},x_t} &= \mathbb{P}_\theta\{X_t = x_t \mid X_{t-1} = x_{t-1}\} \\ &= \sum_{k=0}^{x_{t-1}} \mathbb{P}_\theta\{\varepsilon_t = x_t - x_{t-1} + k, X_{t-1} - \vartheta \circ X_{t-1} = k \mid X_{t-1} = x_{t-1}\} \\ &= \sum_{k=0}^{x_{t-1}} \mathrm{b}_{x_{t-1},1-\theta}(k) g(\Delta x_t + k), \end{aligned}$$

where $\Delta x_t = x_t - x_{t-1}$, and $g(i) = 0$ for $i < 0$. Under $\mathbb{P}_1$ we have $X_t = \mu_G t + \sum_{i=1}^{t}(\varepsilon_i - \mu_G)$, and $P^1_{x_{t-1},x_t} = g(\Delta x_t)$, $x_{t-1}, x_t \in \mathbb{Z}_+$. Hence, under $\mathbb{P}_1$, an INAR(1) process is nothing but a random walk with drift.

The next proposition is basic, but often applied in the sequel.

Proposition 4.1.1. If $\sigma_G^2 < \infty$, we have, for $h \geq 0$,

$$\lim_{n\to\infty} \mathbb{E}_{1-\frac{h}{n^2}} \left[\frac{1}{n^2}\sum_{t=1}^{n} X_t - \frac{\mu_G}{2}\right]^2 = 0, \tag{4.2}$$

and we have, for $\alpha > 0$ and every sequence $(\theta_n)_{n\in\mathbb{N}}$ in $[0,1]$,

$$\lim_{n\to\infty} \frac{1}{n^{3+\alpha}} \sum_{t=1}^{n} \mathbb{E}_{\theta_n} X_t^2 = 0. \tag{4.3}$$

Proof.
We obviously have, $\operatorname{var}_1\left(\sum_{t=1}^n X_t\right) = O(n^3)$ and $\lim_{n\to\infty} n^{-2}\sum_{t=1}^n \mathbb{E}_1 X_t = \mu_G/2$, which yields (4.2) for $h = 0$. Next, we prove (4.2) for $h > 0$. Straightforward calculations show, for $\theta < 1$,

$$\mathbb{E}_\theta \sum_{t=1}^{n} X_t = \mu_G \sum_{t=1}^{n} \frac{1-\theta^t}{1-\theta} = \mu_G\left[\frac{n}{1-\theta} - \frac{\theta-\theta^{n+1}}{(1-\theta)^2}\right],$$

which yields

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n^2}\mathbb{E}_{1-\frac{h}{n^2}} \sum_{t=1}^{n} X_t \\
&= \lim_{n\to\infty} \frac{\mu_G}{n^2}\left[\frac{n}{h/n^2} - \frac{1-\frac{h}{n^2} - \left[1-(n+1)\frac{h}{n^2} + \frac{(n+1)n}{2}\frac{h^2}{n^4} + o\left(\frac{1}{n^2}\right)\right]}{h^2/n^4}\right] \\
&= \frac{\mu_G}{2}.
\end{aligned} \tag{4.4}$$

To treat the variance of $n^{-2}\sum_{t=1}^n X_t$, we use the following simple relations, see also Ispány et al. (2003b), for $0 < \theta < 1$, $s, t \geq 1$,

$$\begin{aligned}
\operatorname{cov}_\theta(X_t, X_s) &= \theta^{|t-s|}\operatorname{var}_\theta X_{t\wedge s}, \\
\operatorname{var}_\theta X_t &= \frac{1-\theta^{2t}}{1-\theta^2}\sigma_G^2 + \frac{(\theta-\theta^t)(1-\theta^t)}{1-\theta^2}\mu_G \leq (\sigma_G^2+\mu_G)\frac{1-\theta^{2t}}{1-\theta^2}.
\end{aligned} \tag{4.5}$$

From this we obtain

$$\operatorname{var}_{1-\frac{h}{n^2}}\left(\frac{1}{n^2}\sum_{t=1}^{n} X_t\right) = \frac{1}{n^4}\sum_{t=1}^{n}\left(1 + 2\sum_{s=t+1}^{n}\left(1-\frac{h}{n^2}\right)^{s-t}\right)\operatorname{var}_{1-\frac{h}{n^2}} X_t$$

$$\leq \frac{1}{n} 2n(\sigma_G^2 + \mu_G) \frac{1}{n^2} \frac{1}{1 - \left(1 - \frac{h}{n^2}\right)^2} \frac{1}{n} \sum_{t=1}^{n} \left(1 - \left(1 - \frac{h}{n^2}\right)^{2t}\right) \to 0,$$

as $n \to \infty$. Together with (4.4) this completes the proof of (4.2) for $h > 0$. To prove (4.3), note that $X_t \leq \sum_{i=1}^{t} \varepsilon_i$. Hence $\mathbb{E}_{\theta_n} X_t^2 \leq \mathbb{E}_1 X_t^2 = \sigma_G^2 t + \mu_G^2 t^2$, which yields the desired conclusion. □

Remark 11. Convergence in probability for the case $h > 0$ in (4.2) cannot be concluded from the convergence in probability in (4.2) for $h = 0$ by contiguity arguments. The reason is (see Remark 12 after the proof of Theorem 4.1) that $\mathbb{P}_{1-h/n^2}^{(n)}$ is not contiguous with respect to $\mathbb{P}_1^{(n)}$.

Next, we consider the thinning process $(\vartheta \circ X_{t-1})_{t\geq 1}$. Under $\mathbb{P}_{1-h/n^2}$, $X_{t-1} - \vartheta \circ X_{t-1}$, conditional on X_{t-1}, follows a Binomial$(X_{t-1}, h/n^2)$ distribution. So we expect that there do not occur many 'deaths' in any time-interval $(t-1, t]$. The following proposition gives a precise statement, where we use the notation, for $h \geq 0$ and $n \in \mathbb{N}$,

$$A_n^h = \left\{ z \in \mathbb{Z}_+ \;\middle|\; \frac{h(z+1)}{n^2} < \frac{1}{2} \right\}, \quad \mathscr{A}_n^h = \{(X_0, \ldots, X_{n-1}) \in A_n^h \times \cdots \times A_n^h\}. \quad (4.6)$$

The reasons for the introduction of these sets are the following. By Proposition 4.4.1 we have, for $x \in A_n^h$, $\sum_{k=r}^{x} \mathrm{b}_{x,h/n^2}(k) \leq 2\,\mathrm{b}_{x,h/n^2}(r)$ for $r = 2, 3$ and terms of the form $(1 - \frac{h}{n^2})^{-2}$ can be bounded neatly, without having to make statements of the form 'for n large enough', or having to refer to 'up to a constant depending on h'. Furthermore, recall the notation $\Delta X_t = X_t - X_{t-1}$.

Proposition 4.1.2. Assume G satisfies $\sigma_G^2 < \infty$. Then we have for all sequences $(\theta_n)_{n\in\mathbb{N}}$ in $[0,1]$, $h \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}_{\theta_n}(\mathscr{A}_n^h) = 1. \quad (4.7)$$

And for $h \geq 0$ we have,

$$\lim_{n\to\infty} \mathbb{P}_{1-\frac{h}{n^2}} \{\exists t \in \{1, \ldots, n\} : X_{t-1} - \vartheta \circ X_{t-1} \geq 2\} = 0. \quad (4.8)$$

Proof.
For a sequence $(\theta_n)_{n\in\mathbb{N}}$ in $[0,1]$, (4.3) implies

$$\mathbb{P}_{\theta_n}\left\{\exists 0 \leq t \leq n : X_t > n^{7/4}\right\} \leq \frac{1}{n^{7/2}} \sum_{t=1}^{n} \mathbb{E}_{\theta_n} X_t^2 \to 0 \text{ as } n \to \infty. \quad (4.9)$$

From this we easily obtain (4.7).
To obtain (4.8) note that, for $X_{t-1} \in A_n^h$ we have, using the bound (4.43),

$$\mathbb{P}_{1-\frac{h}{n^2}} \{X_{t-1} - \vartheta \circ X_{t-1} \geq 2 \mid X_{t-1}\} = \sum_{k=2}^{X_{t-1}} \mathrm{b}_{X_{t-1}, \frac{h}{n^2}}(k) \leq 2\,\mathrm{b}_{X_{t-1}, \frac{h}{n^2}}(2) \leq \frac{h^2 X_{t-1}^2}{n^4}.$$

By (4.3) this yields,

$$\lim_{n\to\infty} \mathbb{P}_{1-\frac{h}{n^2}} \left(\{\exists t \in \{1,\ldots,n\} : X_{t-1} - \vartheta \circ X_{t-1} \geq 2\} \cap \mathscr{A}_n^h\right)$$
$$\leq \lim_{n\to\infty} \frac{h^2}{n^4} \sum_{t=1}^{n} \mathbb{E}_{1-\frac{h}{n^2}} X_{t-1}^2 = 0.$$

Since we already showed $\lim_{n\to\infty} \mathbb{P}_{1-h/n^2}(\mathscr{A}_n^h) = 1$, this yields (4.8). □

As main result of this section, we derive the limit distribution of the number of downward movements of X during $[0,n]$. The probability that the Bernoulli variable $1\{\Delta X_t < 0\}$ equals one is small. Intuitively the dependence over time of this indicator-process is not too strong, so it is not unreasonable to expect that a ‘Poisson law of small numbers' holds. As the following theorem shows, this is indeed the case.

Theorem 4.1. *Assume that G satisfies $\sigma_G^2 < \infty$. Then we have, for $h \geq 0$,*

$$\sum_{t=1}^{n} 1\{\Delta X_t < 0\} \xrightarrow{d} \text{Poisson}\left(\frac{hg(0)\mu_G}{2}\right), \text{ under } \mathbb{P}_{1-\frac{h}{n^2}}. \tag{4.10}$$

Proof.
If $g(0) = 0$ then $\Delta X_t < 0$ implies $X_{t-1} - \vartheta \circ X_{t-1} \geq 2$. Hence, from (4.8) it follows that $\sum_{t=1}^{n} 1\{\Delta X_t < 0\} \xrightarrow{p} 0$ under $\mathbb{P}_{1-h/n^2}$. Since the Poisson distribution with mean 0 concentrates all its mass at 0, this yields the result. The cases $h = 0$ or $g(0) = 1$ (recall $X_0 = 0$) are also trivial.
So we consider the case $h > 0$ and $0 < g(0) < 1$. For notational convenience, abbreviate $\mathbb{P}_{1-h/n^2}$ by $\mathbb{P}_n$ and $\mathbb{E}_{1-h/n^2}$ by $\mathbb{E}_n$. Put $Z_t = 1\{\Delta X_t = -1, \varepsilon_t = 0\}$, and notice that

$$0 \leq 1\{\Delta X_t < 0\} - Z_t = 1\{\Delta X_t \leq -2\} + 1\{\Delta X_t = -1, \varepsilon_t \geq 1\}.$$

From (4.8) it now follows that

$$0 \leq \sum_{t=1}^{n} 1\{\Delta X_t < 0\} - \sum_{t=1}^{n} Z_t \leq 2 \sum_{t=1}^{n} 1\{X_{t-1} - \vartheta \circ X_{t-1} \geq 2\} \xrightarrow{p} 0, \text{ under } \mathbb{P}_n.$$

Thus it suffices to prove that $\sum_{t=1}^{n} Z_t \xrightarrow{d} \text{Poisson}(hg(0)\mu_G/2)$ under $\mathbb{P}_n$. We do this by applying Lemma 4.4.1. Introduce random variables Y_n, where Y_n follows a Poisson distribution with mean $\lambda_n = \sum_{t=1}^{n} \mathbb{E}_n Z_t$. And let Z follow a Poisson distribution with mean $hg(0)\mu_G/2$. From Lemma 4.4.1 we obtain the bound

$$\sup_{A \subset \mathbb{Z}_+} \left| \mathbb{P}_n \left\{ \sum_{t=1}^{n} Z_t \in A \right\} - \Pr\{Y_n \in A\} \right|$$

$$\leq \sum_{t=1}^{n} (\mathbb{E}_n Z_t)^2 + \sum_{t=1}^{n} \mathbb{E}_n |\mathbb{E}_n [Z_t - \mathbb{E}_n Z_t \mid Z_1, \ldots, Z_{t-1}]|.$$

If we prove that

$$\begin{aligned}
(i) &\quad \sum_{t=1}^{n} (\mathbb{E}_n Z_t)^2 \to 0,\\
(ii) &\quad \sum_{t=1}^{n} \mathbb{E}_n Z_t \to \frac{hg(0)\mu_G}{2},\\
(iii) &\quad \sum_{t=1}^{n} \mathbb{E}_n |\mathbb{E}_n [Z_t - \mathbb{E}_n Z_t \mid Z_1, \ldots, Z_{t-1}]| \to 0,
\end{aligned}$$

all hold as $n \to \infty$, then the result follows since we then have, for all $z \in \mathbb{R}$,

$$\begin{aligned}
\left| \mathbb{P}_n \left\{ \sum_{t=1}^{n} Z_t \leq z \right\} - \Pr(Z \leq z) \right| &\leq \left| \mathbb{P}_n \left\{ \sum_{t=1}^{n} Z_t \leq z \right\} - \Pr\{Y_n \leq z\} \right| \\
&\quad + |\Pr\{Y_n \leq z\} - \Pr(Z \leq z)| \to 0.
\end{aligned}$$

First we tackle (i). Notice that, use that, conditional on X_{t-1}, ε_t and $X_{t-1} - \vartheta \circ X_{t-1} \sim \mathrm{Bin}_{X_{t-1}, h/n^2}$ are independent,

$$\begin{aligned}
\mathbb{E}_n Z_t = \mathbb{P}_n\{\varepsilon_t = 0, X_{t-1} - \vartheta \circ X_{t-1} = 1\} &= \frac{hg(0)}{n^2} \mathbb{E}_n X_{t-1} \left(1 - \frac{h}{n^2}\right)^{X_{t-1}-1} \\
&\leq \frac{hg(0)}{n^2} \mathbb{E}_n X_{t-1}.
\end{aligned}$$

Then, using (4.3), (i) easily follows,

$$\lim_{n\to\infty} \sum_{t=1}^{n} (\mathbb{E}_n Z_t)^2 \leq \lim_{n\to\infty} \frac{h^2 g^2(0)}{n^4} \sum_{t=1}^{n} \mathbb{E}_n X_{t-1}^2 = 0.$$

Next we consider (ii). If we prove the relation,

$$\lim_{n\to\infty} \left| \frac{1}{n^2} \sum_{t=1}^{n} \mathbb{E}_n X_{t-1} - \frac{1}{n^2} \sum_{t=1}^{n} \mathbb{E}_n X_{t-1} \left(1 - \frac{h}{n^2}\right)^{X_{t-1}-1} \right| = 0,$$

it is immediate that (ii) follows from (4.2). To prove the previous display, we introduce $B_n = \{\forall t \in \{1, \ldots, n\} : X_t \leq n^{7/4}\}$ with $\lim_{n\to\infty} \mathbb{P}_n(B_n) = 1$ (see (4.9)). On the event B_n we have $n^{-2} X_t \leq n^{-1/4}$ for $t = 1, \ldots, n$. This yields

$$\begin{aligned}
0 &\leq \mathbb{E}_n X_{t-1} \left(1 - \left(1 - \frac{h}{n^2}\right)^{X_{t-1}-1}\right) \\
&\leq \mathbb{E}_n X_{t-1} \left(1 - \left(1 - \frac{h}{n^2}\right)^{X_{t-1}}\right) 1_{B_n} + \mathbb{E}_n X_{t-1} 1_{B_n^c}
\end{aligned}$$

$$\leq \mathbb{E}_n\left[1_{B_n}X_{t-1}\sum_{j=1}^{X_{t-1}}\binom{X_{t-1}}{j}\left(\frac{h}{n^2}\right)^j\right]+\mathbb{E}_nX_{t-1}1_{B_n^c}$$
$$\leq \frac{1}{n^{1/4}}\exp(h)\mathbb{E}_nX_{t-1}+\mathbb{E}_nX_{t-1}1_{B_n^c}.$$

Using $\mathbb{P}_n(B_n)\to 1$ and (4.2) we now obtain,

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{t=1}^n\mathbb{E}_nX_{t-1}1_{B_n^c}\leq \lim_{n\to\infty}\sqrt{\mathbb{E}_n\left(\frac{1}{n^2}\sum_{t=1}^nX_{t-1}\right)^2\mathbb{P}_n(B_n^c)}=\sqrt{\left(\frac{\mu_G}{2}\right)^2\cdot 0}=0.$$

By (4.3) we have $\lim_{n\to\infty}n^{-9/4}\sum_{t=1}^n\mathbb{E}_nX_{t-1}=0$. Combination with the previous two displays yields the result.

Finally, we prove (iii). Let $\mathscr{F}^\varepsilon=(\mathscr{F}_t^\varepsilon)_{t\geq 1}$ and $\mathscr{F}^X=(\mathscr{F}_t^X)_{t\geq 0}$ be the filtrations generated by $(\varepsilon_t)_{t\geq 1}$ and $(X_t)_{t\geq 0}$ respectively, i.e. $\mathscr{F}_t^\varepsilon=\sigma(\varepsilon_1,\dots,\varepsilon_t)$ and $\mathscr{F}_t^X=\sigma(X_0,\dots,X_t)$. Note that we have, for $t\geq 2$,

$$\begin{aligned}&\mathbb{E}_n|\mathbb{E}_n[Z_t-\mathbb{E}_nZ_t\mid Z_1,\dots,Z_{t-1}]|\\ &\quad\leq \mathbb{E}_n\left|\mathbb{E}_n\left[Z_t-\mathbb{E}_nZ_t\mid \mathscr{F}_{t-1}^\varepsilon,\mathscr{F}_{t-1}^X\right]\right|\\ &\quad=\frac{hg(0)}{n^2}\mathbb{E}_n\left|X_{t-1}\left(1-\frac{h}{n^2}\right)^{X_{t-1}-1}-\mathbb{E}_nX_{t-1}\left(1-\frac{h}{n^2}\right)^{X_{t-1}-1}\right|.\end{aligned}\tag{4.11}$$

Using the reverse triangle-inequality we obtain

$$\begin{aligned}&\left|\mathbb{E}_n\left|X_{t-1}\left(1-\frac{h}{n^2}\right)^{X_{t-1}-1}-\mathbb{E}_nX_{t-1}\left(1-\frac{h}{n^2}\right)^{X_{t-1}-1}\right|-\mathbb{E}_n|X_{t-1}-\mathbb{E}_nX_{t-1}|\right|\\ &\quad\leq \mathbb{E}_n\left|X_{t-1}\left(1-\left(1-\frac{h}{n^2}\right)^{X_{t-1}-1}\right)-\mathbb{E}_nX_{t-1}\left(1-\left(1-\frac{h}{n^2}\right)^{X_{t-1}-1}\right)\right|\\ &\quad\leq 2\mathbb{E}_nX_{t-1}\left(1-\left(1-\frac{h}{n^2}\right)^{X_{t-1}-1}\right).\end{aligned}$$

We have already seen in the proof of (ii) that

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{t=1}^n\mathbb{E}_nX_{t-1}\left(1-\left(1-\frac{h}{n^2}\right)^{X_{t-1}-1}\right)=0$$

holds. A combination of the previous two displays with (4.11) now easily yields the bound

$$\sum_{t=1}^n\mathbb{E}_n|\mathbb{E}_n[Z_t-\mathbb{E}_nZ_t\mid Z_1,\dots,Z_{t-1}]|\leq o(1)+\frac{hg(0)}{n^2}\sum_{t=1}^n\sqrt{\mathrm{var}_n X_{t-1}}.\tag{4.12}$$

From (4.5) we have, for $\theta<1$, $\mathrm{var}_\theta X_t\leq\left(\sigma_G^2+\mu_G\right)(1-\theta^{2t})(1-\theta^2)^{-1}$. And for $1\leq t\leq n$ we have $0\leq 1-\left(1-h/n^2\right)^{2t}\leq n^{-1}\exp(2h)$. Now we easily obtain, as

$n \to \infty$,

$$\frac{1}{n^2}\sum_{t=1}^{n}\sqrt{\mathrm{var}_n X_{t-1}} \leq \sqrt{\sigma_G^2+\mu_G}\sqrt{\frac{1}{n^2}\frac{1}{1-\left(1-\frac{h}{n^2}\right)^2}\frac{1}{n}n}\sqrt{\frac{\exp(2h)}{n}} \to 0.$$

A combination with (4.12) yields (iii). This concludes the proof. □

Remark 12. Since $\sum_{t=1}^{n} 1\{\Delta X_t < 0\}$ equals zero under $\mathbb{P}_1^{(n)}$ and converges in distribution to a non-degenerated limit under $\mathbb{P}_{1-h/n^2}^{(n)}$ ($h>0$, $0<g(0)<1$), we see that $\mathbb{P}_{1-h/n^2}^{(n)}$ is not contiguous with respect to $\mathbb{P}_1^{(n)}$ for $h>0$.

4.2 The limit experiment: one observation from a Poisson distribution

For easy reference, we introduce the following assumption.

Assumption 1. A probability distribution G on $\mathbb{Z}_+$ is said to satisfy Assumption 1 if one of the following two condition holds.

(1) support$(G) = \{0,\dots,M\}$ for some $M \in \mathbb{N}$;

(2) support$(G) = \mathbb{Z}_+$, $\sigma_G^2 < \infty$ and g is eventually decreasing, i.e. there exists $M \in \mathbb{N}$ such that $g(k+1) \leq g(k)$ for $k \geq M$.

The rest of this section is devoted to the following theorem.

Theorem 4.2. *Suppose G satisfies Assumption 1. Then the limit experiment of* $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$ *is given by*

$$\mathscr{E}(G) = \left(\mathbb{Z}_+, 2^{\mathbb{Z}_+}, (\mathbb{Q}_h \mid h \geq 0)\right),$$

where $\mathbb{Q}_h = \text{Poisson}\left(hg(0)\mu_G/2\right)$.

Remark 13. Notice that the likelihood-ratios for this Poisson limit experiment are given by,

$$\frac{\mathrm{d}\mathbb{Q}_h}{\mathrm{d}\mathbb{Q}_{h_0}}(Z) = \exp\left(-\frac{(h-h_0)g(0)\mu_G}{2}\right)\left(\frac{h}{h_0}\right)^Z, \tag{4.13}$$

for $h \geq 0$, $h_0 > 0$ and,

$$\frac{\mathrm{d}\mathbb{Q}_h}{\mathrm{d}\mathbb{Q}_0}(Z) = \exp\left(-\frac{hg(0)\mu_G}{2}\right)1_{\{0\}}(Z), \tag{4.14}$$

for $h \geq 0$.

Remark 14. Usually limit experiments are Locally Asymptotically Quadratic (see Jeganathan (1995) and Le Cam and Yang (1990)) and even non-regular models often enjoy a shift structure (see Hirano and Porter (2003a)), whereas the Poisson limit experiment does not enjoy these two properties. As discussed in the introduction, the nearly unstable AR(1) model yields LAQ limit experiments. The theorem is indeed rather surprising since Ispány et al. (2003b) established a functional limit theorem with a Ornstein-Uhlenbeck limit process from which one would conjecture a standard LAQ-type limit experiment.

Proof.
To determine the limit-experiment we need to determine the limit-distribution of the log-likelihood ratios, $h, h_0 \geq 0$,

$$\mathscr{L}_n(h,h_0) = \log \frac{\mathrm{d}\mathbb{P}^{(n)}_{1-\frac{h}{n^2}}}{\mathrm{d}\mathbb{P}^{(n)}_{1-\frac{h_0}{n^2}}} = \sum_{t=1}^{n} \log \frac{P^{1-\frac{h}{n^2}}_{X_{t-1},X_t}}{P^{1-\frac{h_0}{n^2}}_{X_{t-1},X_t}},$$

under $\mathbb{P}_{1-h_0/n^2}$. Note that for $h_0 > 0$ $\mathscr{L}_n(0,h_0) = -\infty$, and thus $\mathrm{d}\mathbb{P}^{(n)}_0 / \mathrm{d}\mathbb{P}^{(n)}_{1-h_0/n^2} = 0$, if $\sum_{t=1}^{n} 1\{\Delta X_t < 0\} > 0$. Because $\mathscr{L}_n(h,h_0)$ is complicated to analyze, we first make a suitable approximation of this object. Split the transition-probability $P^{1-h/n^2}_{x_{t-1},x_t}$ into a leading term,

$$L_n(x_{t-1},x_t,h) = \begin{cases} \sum_{k=-\Delta x_t}^{-\Delta x_t+1} \mathrm{b}_{x_{t-1},\frac{h}{n^2}}(k) g(\Delta x_t + k) & \text{if } \Delta x_t < 0, \\ \sum_{k=0}^{1} \mathrm{b}_{x_{t-1},\frac{h}{n^2}}(k) g(\Delta x_t + k) & \text{if } \Delta x_t \geq 0, \end{cases}$$

and a remainder term,

$$R_n(x_{t-1},x_t,h) = \begin{cases} \sum_{k=-\Delta x_t+2}^{x_{t-1}} \mathrm{b}_{x_{t-1},\frac{h}{n^2}}(k) g(\Delta x_t + k) & \text{if } \Delta x_t < 0, \\ \sum_{k=2}^{x_{t-1}} \mathrm{b}_{x_{t-1},\frac{h}{n^2}}(k) g(\Delta x_t + k) & \text{if } \Delta x_t \geq 0. \end{cases}$$

We introduce a simpler version of $\mathscr{L}_n(h,h_0)$ in which the remainder terms are removed,

$$\tilde{\mathscr{L}}_n(h,h_0) = \sum_{t=1}^{n} \log \frac{L_n(X_{t-1},X_t,h)}{L_n(X_{t-1},X_t,h_0)}.$$

The difference between $\tilde{\mathscr{L}}_n(h,h_0)$ and $\mathscr{L}_n(h,h_0)$ is negligible. To enhance readability we organize this result and its proof in a lemma.

Lemma 4.2.1. Suppose G satisfies Assumption 1. We have, for $h, h_0 \geq 0$,

$$\tilde{\mathscr{L}}_n(h,h_0) = \mathscr{L}_n(h,h_0) + o\left(\mathbb{P}_{1-\frac{h_0}{n^2}}; 1\right). \tag{4.15}$$

Proof.
We obtain, for $h > 0, h_0 \geq 0$, from the inequality $\left|\log((a+b)/(c+d)) - \log(a/c)\right| \leq b/a + d/c$ for $a, c > 0$, $b, d \geq 0$, the bound

$$\left|\mathscr{L}_n(h,h_0) - \tilde{\mathscr{L}}_n(h,h_0)\right| \leq \sum_{t=1}^{n} \left\{ \frac{R_n(X_{t-1},X_t,h)}{L_n(X_{t-1},X_t,h)} + \frac{R_n(X_{t-1},X_t,h_0)}{L_n(X_{t-1},X_t,h_0)} \right\}, \quad (4.16)$$

$\mathbb{P}_{1-h_0/n^2}$-a.s. It is easy to see, since $\mathrm{b}_{n,0}(k) = 0$ if $k > 0$, that, for $h_0 > 0$, $\mathscr{L}_n(0,h_0)$ and $\tilde{\mathscr{L}}_n(0,h_0)$ both equal minus infinity if $\sum_{t=1}^n 1\{\Delta X_t < 0\} \geq 1$. If $\sum_{t=1}^n 1\{\Delta X_t < 0\} = 0$ we have

$$\left|\mathscr{L}_n(0,h_0) - \tilde{\mathscr{L}}_n(0,h_0)\right| \leq \sum_{t=1}^{n} \frac{R_n(X_{t-1},X_t,h_0)}{L_n(X_{t-1},X_t,h_0)} \quad \mathbb{P}_{1-\frac{h_0}{n^2}} - \text{a.s.}$$

Thus if we show that

$$\sum_{t=1}^{n} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}},$$

holds for $h' = h$ and $h' = h_0$ the lemma is proved (exclude the case $h' = 0$ and $h_0 > 0$, which need not be considered). We split the expression in the previous display into four nonnegative parts

$$\sum_{t=1}^{n} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} = \sum_{t:\Delta X_t \leq -2} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} + \sum_{t:\Delta X_t = -1} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')}$$
$$+ \sum_{t:0 \leq \Delta X_t \leq M} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} + \sum_{t:\Delta X_t > M} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')}.$$

Since $\Delta X_t \leq -2$ implies $X_{t-1} - \vartheta \circ X_{t-1} \geq 2$ an application of (4.8) yields

$$\sum_{t:\Delta X_t \leq -2} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}}.$$

Next we treat the terms for which $\Delta X_t = -1$. If $h_0 = 0$ we do not have such terms (under $\mathbb{P}_{1-h_0/n^2}$), and remember that the case $h' = 0$ and $h_0 > 0$ need not be considered. So we only need to consider this term for $h', h_0 > 0$. On the event $\mathscr{A}_n^{h'}$ (see (4.6) for the definition of this event), an application of (4.43) yields,

$$\sum_{t:\Delta X_t = -1} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} \leq \sum_{t:\Delta X_t = -1} \frac{\sum_{k=3}^{X_{t-1}} \mathrm{b}_{X_{t-1},\frac{h'}{n^2}}(k)}{g(0)\,\mathrm{b}_{X_{t-1},\frac{h'}{n^2}}(1)}$$
$$\leq 2\sum_{t=1}^{n} \frac{\frac{X_{t-1}^3}{3!}\frac{h'^3}{n^6}\left(1-\frac{h'}{n^2}\right)^{X_{t-1}-3}}{g(0)X_{t-1}\frac{h'}{n^2}\left(1-\frac{h'}{n^2}\right)^{X_{t-1}-1}} 1\{X_{t-1} \geq 1\}$$

$$\leq \frac{4h'^2}{3g(0)n^4}\sum_{t=1}^{n} X_{t-1}^2,$$

since $(1-h'/n^2)^{-2} \leq 4$ by definition of $\mathscr{A}_n^{h'}$ (see (4.6) for the definition of this set). From (4.3) and (4.7) it now easily follows that we have

$$\sum_{t:\Delta X_t=-1} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}}.$$

Next, we analyze the terms for which $0 \leq \Delta X_t \leq M$. We have, by (4.43), on the event $\mathscr{A}_n^{h'}$,

$$\begin{aligned}\sum_{t:0\leq\Delta X_t\leq M} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} &\leq \sum_{t:0\leq\Delta X_t\leq M} \frac{\sum_{k=2}^{X_{t-1}} \mathrm{b}_{X_{t-1},\frac{h'}{n^2}}(k)g(\Delta X_t+k)}{g(\Delta X_t)\,\mathrm{b}_{X_{t-1},\frac{h'}{n^2}}(0)}\\ &\leq \frac{2}{m^*}\sum_{t:0\leq\Delta X_t\leq M} \frac{\mathrm{b}_{X_{t-1},\frac{h'}{n^2}}(2)}{\mathrm{b}_{X_{t-1},\frac{h'}{n^2}}(0)}\\ &\leq \frac{4h'^2}{m^* n^4}\sum_{t=1}^{n} X_{t-1}^2,\end{aligned}$$

where $m^* = \min\{g(k)|0 \leq k \leq M\} > 0$. Now (4.3), and (4.7) yield the desired convergence,

$$\sum_{t:0\leq\Delta X_t\leq M} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}}.$$

Finally, we discuss the terms for which $\Delta X_t > M$. If the support of G was given by $\{0,\dots,M\}$ there are no such terms. So we only need to consider the case, where the support of G is $\mathbb{Z}_+$. Since g is non-increasing on $\{M, M+1,\dots\}$, we have, by (4.43),

$$R_n(X_{t-1},X_t,h') \leq 2g(\Delta X_t)\,\mathrm{b}_{X_{t-1},\frac{h'}{n^2}}(2), \quad X_{t-1} \in A_n^{h'},$$

which yields, for $X_{t-1} \in A_n^{h'}$,

$$0 \leq \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} \leq \frac{2g(\Delta X_t)\frac{X_{t-1}^2}{2}\frac{h'^2}{n^4}\left(1-\frac{h'}{n^2}\right)^{X_{t-1}-2}}{g(\Delta X_t)\left(1-\frac{h'}{n^2}\right)^{X_{t-1}}} \leq \frac{4h'^2}{n^4}X_{t-1}^2.$$

From (4.3), and (4.7) it now easily follows that we have

$$\sum_{t:\Delta X_t\geq M} \frac{R_n(X_{t-1},X_t,h')}{L_n(X_{t-1},X_t,h')} \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}}.$$

This concludes the proof of the lemma. □

Hence, the limit-distribution of the random vector $(\mathscr{L}_n(h,h_0))_{h\in I}$, for a finite subset $I\subset\mathbb{R}_+$, is the same as the limit-distribution of $(\tilde{\mathscr{L}}_n(h,h_0))_{h\in I}$. It easily follows, using (4.8), that $\tilde{\mathscr{L}}_n(h,h_0)$ can be decomposed as

$$\begin{aligned}\tilde{\mathscr{L}}_n(h,h_0) &= \sum_{t=1}^{n}\frac{X_{t-1}-2}{n^2}\log\left(\frac{1-\frac{h}{n^2}}{1-\frac{h_0}{n^2}}\right)^{n^2}\\ &\quad + S_n^+(h,h_0)+S_n^-(h,h_0)+o\left(\mathbb{P}_{1-\frac{h_0}{n^2}};1\right),\end{aligned} \tag{4.17}$$

where $S_n^+(h,h_0)=\sum_{t:\Delta X_t\geq 0}W_{tn}^+$ and $S_n^-(h,h_0)=\sum_{t:\Delta X_t=-1}W_{tn}^-$, are defined by,

$$W_{tn}^+=\log\left[\frac{g(\Delta X_t)\left(1-\frac{h}{n^2}\right)^2+X_{t-1}\frac{h}{n^2}\left(1-\frac{h}{n^2}\right)g(\Delta X_t+1)}{g(\Delta X_t)\left(1-\frac{h_0}{n^2}\right)^2+X_{t-1}\frac{h_0}{n^2}\left(1-\frac{h_0}{n^2}\right)g(\Delta X_t+1)}\right],$$

and

$$W_{tn}^-=\log\left[\frac{X_{t-1}\frac{h}{n^2}\left(1-\frac{h}{n^2}\right)g(0)+\frac{X_{t-1}(X_{t-1}-1)}{2}\frac{h^2}{n^4}g(1)}{X_{t-1}\frac{h_0}{n^2}\left(1-\frac{h_0}{n^2}\right)g(0)+\frac{X_{t-1}(X_{t-1}-1)}{2}\frac{h_0^2}{n^4}g(1)}\right].$$

First, we treat the first term in (4.17). By (4.2) we have,

$$\log\left[\left(\frac{1-\frac{h}{n^2}}{1-\frac{h_0}{n^2}}\right)^{n^2}\right]\frac{1}{n^2}\sum_{t=1}^{n}(X_{t-1}-2)\xrightarrow{p}-\frac{(h-h_0)\mu_G}{2},\text{ under }\mathbb{P}_{1-\frac{h_0}{n^2}}. \tag{4.18}$$

Next, we discuss the behavior of $S_n^+(h,h_0)$, the second term of (4.17). This is the content of the next lemma.

Lemma 4.2.2. Suppose G satisfies Assumption 1. We have, for $h,h_0\geq 0$,

$$S_n^+(h,h_0)\xrightarrow{p}\frac{(h-h_0)(1-g(0))\mu_G}{2},\text{ under }\mathbb{P}_{1-\frac{h_0}{n^2}}. \tag{4.19}$$

Proof.
We write,

$$S_n^+(h,h_0)=\sum_{t:\Delta X_t\geq 0}\log\left[1+U_{tn}^+\right],$$

where

$$U_{tn}^+=\frac{g(\Delta X_t)\left[\frac{h^2-h_0^2}{n^4}-2\frac{h-h_0}{n^2}\right]+X_{t-1}g(\Delta X_t+1)\left[\frac{h-h_0}{n^2}-\frac{h^2-h_0^2}{n^4}\right]}{g(\Delta X_t)\left(1-\frac{h_0}{n^2}\right)^2+X_{t-1}g(\Delta X_t+1)\frac{h_0}{n^2}\left(1-\frac{h_0}{n^2}\right)}.$$

Notice that, for n large enough,

$$U_{tn}^{+2} \leq \frac{2\left(g^2(\Delta X_t)\left[\frac{h^2-h_0^2}{n^4} - 2\frac{h-h_0}{n^2}\right]^2 + X_{t-1}^2 g^2(\Delta X_t+1)\left[\frac{h-h_0}{n^2} - \frac{h^2-h_0^2}{n^4}\right]^2\right)}{g^2(\Delta X_t)\left(1-\frac{h_0}{n^2}\right)^4}$$
$$\leq \frac{C}{n^4}\left(X_{t-1}^2+1\right),$$

for some constant C, where we used that $e \mapsto g(e+1)/g(e)$ is bounded. From (4.3) we obtain,

$$\lim_{n\to\infty} \mathbb{E}_{1-\frac{h_0}{n^2}} \sum_{t:\Delta X_t \geq 0} U_{tn}^{+2} \leq 0 + \lim_{n\to\infty} \mathbb{E}_{1-\frac{h_0}{n^2}} \frac{C}{n^4} \sum_{t=1}^{n} X_{t-1}^2 = 0.$$

Hence

$$\sum_{t:\Delta X_t\geq 0} U_{tn}^{+2} \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-h_0/n^2}, \tag{4.20}$$

and

$$\lim_{n\to\infty} \mathbb{P}_{1-\frac{h_0}{n^2}} \left\{\max_{t:\Delta X_t\geq 0} |U_{tn}^+| \leq 1/2\right\} = 1, \tag{4.21}$$

since

$$\mathbb{P}_n\{\exists t: \Delta X_t \geq 0, |U_{tn}^+| > 1/2\} \leq \mathbb{P}_n\left\{\sum_{t:\Delta X_t\geq 0} U_{tn}^{+2} > \frac{1}{4}\right\} \to 0.$$

Using the expansion $\log(1+x) = x + r(x)$, where the remainder term r satisfies $|r(x)| \leq 2x^2$ for $|x| \leq 1/2$, we obtain from (4.20) and (4.21),

$$S_n^+(h,h_0) = \sum_{t:\Delta X_t\geq 0} \log\left[1+U_{tn}^+\right] = \sum_{t:\Delta X_t\geq 0} U_{tn}^+ + o\left(\mathbb{P}_{1-\frac{h_0}{n^2}};1\right).$$

Thus the problem reduces to determining the asymptotic behavior of $\sum_{t:\Delta X_t\geq 0} U_{tn}^+$. Note that,

$$\sum_{t:\Delta X_t\geq 0} U_{tn}^+ = \sum_{t:\Delta X_t\geq 0} \frac{X_{t-1}g(\Delta X_t+1)\left[\frac{h-h_0}{n^2} - \frac{h^2-h_0^2}{n^4}\right]}{g(\Delta X_t)\left(1-\frac{h_0}{n^2}\right)^2 + X_{t-1}g(\Delta X_t+1)\frac{h_0}{n^2}\left(1-\frac{h_0}{n^2}\right)}$$
$$+\, o\left(\mathbb{P}_{1-\frac{h_0}{n^2}};1\right).$$

Using that $e \mapsto g(e+1)/g(e)$ is bounded and (4.3), we obtain

$$\sum_{t:\Delta X_t\geq 0} \left| \frac{X_{t-1}g(\Delta X_t+1)\left[\frac{h-h_0}{n^2} - \frac{h^2-h_0^2}{n^4}\right]}{g(\Delta X_t)\left(1-\frac{h_0}{n^2}\right)^2 + X_{t-1}g(\Delta X_t+1)\frac{h_0}{n^2}\left(1-\frac{h_0}{n^2}\right)}\right.$$

$$-\frac{(h-h_0)}{n^2}\frac{X_{t-1}g(\Delta X_t+1)}{g(\Delta X_t)}\Bigg|$$
$$\le \frac{C}{n^4}\sum_{t=1}^n X_{t-1}^2 \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}}.$$

Thus the previous three displays and (4.8) yield

$$S_n^+(h,h_0) = \frac{h-h_0}{n^2}\sum_{t=1}^n X_{t-1}\frac{g(\Delta X_t+1)}{g(\Delta X_t)}1\{\Delta X_t \ge 0,\, X_{t-1}-\vartheta\circ X_{t-1}\le 1\}$$
$$+\, o\left(\mathbb{P}_{1-\frac{h_0}{n^2}};1\right).$$

Finally, we will show that, under $\mathbb{P}_{1-h_0/n^2}$,

$$\frac{1}{n^2}\sum_{t=1}^n X_{t-1}\frac{g(\Delta X_t+1)}{g(\Delta X_t)}1\{\Delta X_t\ge 0,\, X_{t-1}-\vartheta\circ X_{t-1}\le 1\}\xrightarrow{p}\frac{(1-g(0))\mu_G}{2}, \quad (4.22)$$

which will conclude the proof. For notational convenience we introduce

$$Z_t = \frac{g(\Delta X_t+1)}{g(\Delta X_t)}1\{\Delta X_t\ge 0,\, X_{t-1}-\vartheta\circ X_{t-1}\le 1\}$$
$$= \frac{g(\varepsilon_t+1)}{g(\varepsilon_t)}1\{X_{t-1}-\vartheta\circ X_{t-1}=0\}+\frac{g(\varepsilon_t)}{g(\varepsilon_t-1)}1\{\varepsilon_t\ge 1,\, X_{t-1}-\vartheta\circ X_{t-1}=1\}.$$

Using that ε_t is independent of $X_{t-1}-\vartheta\circ X_{t-1}$ we obtain

$$\mathbb{E}_{1-\frac{h_0}{n^2}}[Z_t \mid X_{t-1}-\vartheta\circ X_{t-1}] = (1-g(0))1\{X_{t-1}-\vartheta\circ X_{t-1}=0\}$$
$$+\,1\{X_{t-1}-\vartheta\circ X_{t-1}=1\}\mathbb{E}\frac{g(\varepsilon_t)}{g(\varepsilon_t-1)}1\{\varepsilon_t\ge 1\},$$

where we used that $\mathbb{E}g(\varepsilon_1+1)/g(\varepsilon_1)=1-g(0)$ and $\mathbb{E}1\{\varepsilon_1\ge 1\}g(\varepsilon_1)/g(\varepsilon_1-1)<\infty$, since we assumed that g is eventually decreasing. So we have

$$Z_t - \mathbb{E}_{1-\frac{h_0}{n^2}}[Z_t\mid X_{t-1}-\vartheta\circ X_{t-1}]$$
$$=\left(\frac{g(\varepsilon_t+1)}{g(\varepsilon_t)}-\mathbb{E}\frac{g(\varepsilon_t+1)}{g(\varepsilon_t)}\right)1\{X_{t-1}-\vartheta\circ X_{t-1}=0\}$$
$$+\left(\frac{g(\varepsilon_t)}{g(\varepsilon_t-1)}1\{\varepsilon_t\ge 1\}-\mathbb{E}\frac{g(\varepsilon_t)}{g(\varepsilon_t-1)}1\{\varepsilon_t\ge 1\}\right)1\{X_{t-1}-\vartheta\circ X_{t-1}=1\}.$$

From this it is not hard to see that we have, for $t\in\mathbb{N}$,

$$\mathbb{E}_{1-\frac{h_0}{n^2}}X_{t-1}\left(Z_t-\mathbb{E}_{1-\frac{h_0}{n^2}}[Z_t\mid X_{t-1}-\vartheta\circ X_{t-1}]\right)=0,$$

for $s < t$,

$$\mathbb{E}_{1-\frac{h_0}{n^2}} X_{t-1}\left(Z_t - \mathbb{E}_{1-\frac{h_0}{n^2}}\left[Z_t \mid X_{t-1} - \vartheta \circ X_{t-1}\right]\right) X_{s-1}\left(Z_s - \mathbb{E}_{1-\frac{h_0}{n^2}}\left[Z_s \mid X_{s-1} - \vartheta \circ X_{s-1}\right]\right) = 0.$$

and,

$$\mathbb{E}_{1-\frac{h_0}{n^2}}\left(Z_t - \mathbb{E}_{1-\frac{h_0}{n^2}}\left[Z_t \mid X_{t-1} - \vartheta \circ X_{t-1}\right]\right)^2 \leq C, \tag{4.23}$$

for $C = 2(\mathrm{var}\left(g(\varepsilon_1+1)/g(\varepsilon_1)\right) + \mathrm{var}\left(1_{\{\varepsilon_t \geq 1\}} g(\varepsilon_1)/g(\varepsilon_1 - 1)\right)$, which is finite by Assumption 1. Thus, by (4.3), it follows that

$$\begin{aligned}
&\mathbb{E}_{1-\frac{h_0}{n^2}}\left(\frac{1}{n^2}\sum_{t=1}^{n} X_{t-1}\left(Z_t - \mathbb{E}_{1-\frac{h_0}{n^2}}\left[Z_t \mid X_{t-1} - \vartheta \circ X_{t-1}\right]\right)\right)^2 \\
&\quad = \frac{1}{n^4}\sum_{t=1}^{n} \mathbb{E}_{1-\frac{h_0}{n^2}} X_{t-1}^2\left(Z_t - \mathbb{E}_{1-\frac{h_0}{n^2}}\left[Z_t \mid X_{t-1} - \vartheta \circ X_{t-1}\right]\right)^2 \\
&\quad \leq \frac{C}{n^4}\sum_{t=1}^{n} \mathbb{E}_{1-\frac{h_0}{n^2}} X_{t-1}^2 \to 0.
\end{aligned}$$

Hence (4.22) is equivalent to,

$$\frac{1}{n^2}\sum_{t=1}^{n} X_{t-1}\mathbb{E}_{1-\frac{h_0}{n^2}}\left[Z_t \mid X_{t-1} - \vartheta \circ X_{t-1}\right] \xrightarrow{p} \frac{(1-g(0))\mu_G}{2}, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}}. \tag{4.24}$$

Since, by (4.3),

$$\begin{aligned}
\frac{1}{n^2}\sum_{t=1}^{n} \mathbb{E}_{1-\frac{h_0}{n^2}} X_{t-1} 1\{X_{t-1} - \vartheta \circ X_{t-1} = 1\} &= \frac{h_0}{n^4}\sum_{t=1}^{n} \mathbb{E}_{1-\frac{h_0}{n^2}} X_{t-1}^2\left(1 - \frac{h_0}{n^2}\right)^{X_{t-1}-1} \\
&\leq \frac{h_0}{n^4}\sum_{t=1}^{n} \mathbb{E}_{1-\frac{h_0}{n^2}} X_{t-1}^2 \to 0,
\end{aligned}$$

we have, using (4.8),

$$\begin{aligned}
&\left|\frac{1}{n^2}\sum_{t=1}^{n} X_{t-1}\mathbb{E}_{1-\frac{h_0}{n^2}}\left[Z_t \mid X_{t-1} - \vartheta \circ X_{t-1}\right] - \frac{1-g(0)}{n^2}\sum_{t=1}^{n} X_{t-1}\right| \\
&\quad \leq \left|\mathbb{E}\frac{g(\varepsilon_t)}{g(\varepsilon_t - 1)} 1\{\varepsilon_t \geq 1\} - (1-g(0))\right| \frac{1}{n^2}\sum_{t=1}^{n} X_{t-1} 1\{X_{t-1} - \vartheta \circ X_{t-1} = 1\} \\
&\qquad + \frac{1-g(0)}{n^2}\sum_{t=1}^{n} X_{t-1} 1\{X_{t-1} - \vartheta \circ X_{t-1} \geq 2\} \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}},
\end{aligned}$$

we conclude (4.24), which concludes the proof of the lemma. □

Finally, we discuss the term $S_n^-(h,h_0)$ in (4.17). Under $\mathbb{P}_1$ this term is not present, so we only need to consider $h_0>0$. We organize the result and its proof in the following lemma.

Lemma 4.2.3. Suppose G satisfies Assumption 1. We have, for $h_0>0$, $h\geq 0$,

$$S_n^-(h,h_0) = \log\left[\frac{h}{h_0}\right]\sum_{t=1}^n 1\{\Delta X_t<0\} + o\left(\mathbb{P}_{1-\frac{h_0}{n^2}};1\right), \tag{4.25}$$

where we set $\log(0)=-\infty$ and $-\infty\cdot 0=0$.

Proof.
First we consider $h=0$. From the definition of $S_n^-(0,h_0)$ we see that $S_n^-(0,h_0)=0$ if $\sum_{t=1}^n 1\{\Delta X_t<0\}=0$ (since an empty sum equals zero by definition). And if $\sum_{t=1}^n 1\{\Delta X_t<0\}\geq 1$ we have $S_n^-(0,h_0)=-\infty$ (since $W_{tn}^-=-\infty$ for $h=0$). This concludes the proof for $h=0$.
So we now consider $h>0$. We rewrite

$$W_{tn}^- = \log\left[\frac{\frac{h}{h_0}\left(\frac{1-\frac{h}{n^2}}{1-\frac{h_0}{n^2}}\right)+\frac{X_{t-1}-1}{2n^2}\frac{h^2g(1)}{g(0)h_0\left(1-\frac{h_0}{n^2}\right)}}{1+\frac{X_{t-1}-1}{2n^2}\frac{h_0g(1)}{g(0)\left(1-\frac{h_0}{n^2}\right)}}\right].$$

By (4.8), the proof is finished, if we show that

$$\sum_{t:\Delta X_t=-1}\left|W_{tn}^- - \log\left[\frac{h}{h_0}\right]\right| \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}}.$$

Using the inequality $|\log((a+b)/(c+d))-\log(a/c)|\leq b/a+d/c$ for $a,c>0$, $b,d\geq 0$, we obtain

$$\begin{aligned}\left|W_{tn}^- - \log\left[\frac{h}{h_0}\right]\right| &\leq \left|W_{tn}^- - \log\left[\frac{h}{h_0}\left(\frac{1-\frac{h}{n^2}}{1-\frac{h_0}{n^2}}\right)\right]\right| + O(n^{-2})\\ &\leq \frac{X_{t-1}-1}{2n^2}\left[\frac{h^2g(1)}{g(0)h_0\left(1-\frac{h_0}{n^2}\right)}\left(\frac{h}{h_0}\left(\frac{1-\frac{h}{n^2}}{1-\frac{h_0}{n^2}}\right)\right)^{-1}+\frac{h_0g(1)}{g(0)\left(1-\frac{h_0}{n^2}\right)}\right]+O(n^{-2}).\end{aligned}$$

Hence, it suffices to show that

$$\sum_{t:\Delta X_t=-1}\frac{X_{t-1}}{n^2}\xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h_0}{n^2}}.$$

Note first that we have, by (4.8),

$$0\leq \frac{1}{n^2}\sum_{t=1}^n X_{t-1}1\{\Delta X_t=-1\} = \frac{1}{n^2}\sum_{t=1}^n X_{t-1}1\{\Delta X_t=-1,\varepsilon_t=0\}+o\left(\mathbb{P}_{1-\frac{h_0}{n^2}};1\right).$$

We show that the expectation of the first term on the right-hand side in the previous display converges to zero, which will conclude the proof. We have, by (4.3),

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n^2}\sum_{t=1}^{n}\mathbb{E}_{1-\frac{h_0}{n^2}}X_{t-1}1\{\Delta X_t=-1,\varepsilon_t=0\}\\ &=\lim_{n\to\infty}\frac{h_0}{n^4}\sum_{t=1}^{n}\mathbb{E}_{1-\frac{h_0}{n^2}}g(0)X_{t-1}^2\left(1-\frac{h_0}{n^2}\right)^{X_{t-1}-1}\\ &\leq\lim_{n\to\infty}\frac{h_0 g(0)}{n^4}\sum_{t=1}^{n}\mathbb{E}_{1-\frac{h_0}{n^2}}X_{t-1}^2=0,\end{aligned}$$

which concludes the proof of the lemma. □

To complete the proof of the theorem, note that we obtain from Lemma 4.2.1, (4.17), (4.18), Lemma 4.2.2 and Lemma 4.2.3,

$$\begin{aligned}\mathscr{L}_n(h,h_0)&=\tilde{\mathscr{L}}_n(h,h_0)+o\left(\mathbb{P}_{1-\frac{h_0}{n^2}};1\right)\\ &=-\frac{(h-h_0)\,g(0)\mu_G}{2}+\log\left[\frac{h}{h_0}\right]\sum_{t=1}^{n}1\{\Delta X_t<0\}+o\left(\mathbb{P}_{1-\frac{h_0}{n^2}};1\right),\end{aligned}$$

where we interpret $\log(0)=-\infty$, $\log(0)\cdot 0=0$ and $\log(h/0)\sum_{t=1}^{n}1\{\Delta X_t<0\}=0$ when $h_0=0$, $h>0$. Hence, Theorem 4.1 implies that, for a finite subset $I\subset\mathbb{R}_+$,

$$(\mathscr{L}_n(h,h_0))_{h\in I}\xrightarrow{d}\left(\log\frac{\mathrm{d}\mathbb{Q}_h}{\mathrm{d}\mathbb{Q}_{h_0}}(Z)\right)_{h\in I},\text{ under }\mathbb{P}_{1-\frac{h_0}{n^2}},$$

which concludes the proof. □

Remark 15. In the proof we have seen that,

$$\log\frac{\mathrm{d}\mathbb{P}^{(n)}_{1-\frac{h}{n^2}}}{\mathrm{d}\mathbb{P}^{(n)}_{1-\frac{h_0}{n^2}}}=-\frac{(h-h_0)\,g(0)\mu_G}{2}+\log\left[\frac{h}{h_0}\right]\sum_{t=1}^{n}1\{\Delta X_t<0\}+o(\mathbb{P}_{1-h_0/n^2};1).$$

So, heuristically, we can interpret $\sum_{t=1}^{n}1\{\Delta X_t<0\}$ as an 'approximately sufficient statistic'.

Remark 16. It is straightforward to see that the experiments

$$\mathscr{B}_n^0=\left(\mathbb{Z}_+,2^{\mathbb{Z}_+},\left(\text{Binomial}\left(n,\frac{h}{n}\right)\mid h\geq 0\right)\right),$$

and

$$\mathscr{B}_n^1=\left(\mathbb{Z}_+,2^{\mathbb{Z}_+},\left(\text{Binomial}\left(n,1-\frac{h}{n}\right)\mid h\geq 0\right)\right),$$

$n \in \mathbb{N}$, both converge to the Poisson experiment

$$\mathscr{P} = \left(\mathbb{Z}_+, 2^{\mathbb{Z}_+}, (\text{Poisson}(h) \mid h \geq 0)\right).$$

Since the law of X_t, given X_{t-1}, is the convolution of G with a Binomial(X_{t-1}, θ) distribution, one might be tempted to think that the convergence of experiments $\mathscr{E}_n(G) \to \mathscr{E}(G)$ somehow follows from the convergence $\mathscr{B}_n^1 \to \mathscr{P}$. However, a similar reasoning would yield that the sequence of experiments

$$\mathscr{E}_n^0(G) = \left(\mathbb{Z}_+^{n+1}, 2^{\mathbb{Z}_+^{n+1}}, \left(\mathbb{P}^{(n)}_{\frac{h}{\sqrt{n}}} \mid h \geq 0\right)\right), \quad n \in \mathbb{N},$$

converges to some Poisson experiment. This is not the case. As Proposition 4.3.6 shows, the sequence $\left(\mathscr{E}_n^0(G)\right)_{n\in\mathbb{N}}$ converges to the normal location experiment $(\mathbb{R}, \mathscr{B}(\mathbb{R}), (\mathrm{N}(h,\tau) \mid h \geq 0))$, for some $\tau > 0$.

Remark 17. An obvious question is whether we can expect a similar limit experiment for higher order INAR processes (similar to the classical AR(p) processes, we say that an INAR(p) process has a unit root in case $\sum_{i=1}^p \theta_i = 1$). However, deriving the limit experiment for nearly unstable higher order INAR processes seems to be extremely challenging due to the complicated form of the transition-probabilities. But, intuitively, there is no reason to expect a Poisson limit experiment. To great extent the Poisson limit experiment for the INAR(1) model is coming from the property that the process is non-decreasing under the unit root. In a unit root INAR(2) setting we need not have such a property, since for, e.g., $\theta_1 = \theta_2 = 1/2$, the process can move down as well as up.

4.3 Applications

This section considers the following applications as an illustration of the statistical consequences of the convergence of experiments. In Section 4.3.1 we discuss efficient estimation of h, the deviation from a unit root, in the nearly unstable case for two settings. The first setting, discussed in Section 4.3.1, treats the case that the immigration distribution G is completely known. And the second setting considers a semiparametric model, where hardly any conditions on G are imposed. In Section 4.3.2 we provide an efficient estimator of θ in the 'global' INAR model. Finally, we discuss testing for a unit root in Section 4.3.3.

4.3.1 Efficient estimation of h in nearly unstable INAR models

G known

In this section G is assumed to be known. So we consider the sequence of experiments $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$. As before, we denote the observation from the limit experiment $\mathscr{E}(G)$ by Z, and $\mathbb{Q}_h = \text{Poisson}(hg(0)\mu_G/2)$.

Since we have established convergence of $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$ to $\mathscr{E}(G)$, an application of the Le Cam-Van der Vaart Asymptotic Representation Theorem yields the following proposition.

Proposition 4.3.1. Suppose G satisfies Assumption 1. If $(T_n)_{n\in\mathbb{N}}$ is a sequence of estimators of h in the sequence of experiments $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$ such that $\mathscr{L}(T_n \mid \mathbb{P}_{1-h/n^2}) \to Z_h$ for all $h \geq 0$, then there exists a map $t : \mathbb{Z}_+ \times [0,1] \to \mathbb{R}$ such that $Z_h = \mathscr{L}(t(Z,U) \mid \mathbb{Q}_h \times \text{Uniform}[0,1])$ (i.e. U is distributed uniformly on $[0,1]$ and independent of the observation Z from the limit experiment $\mathscr{E}(G)$).

Proof.
Under the stated conditions the sequence of experiments $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$ converges to the Poisson limit experiment $\mathscr{E}(G)$ (by Theorem 4.2). Since this experiment is dominated by counting measure on $\mathbb{Z}_+$, the result follows by applying the Le Cam-Van der Vaart Asymptotic Representation Theorem (see, for instance, Theorem 3.1 in Van der Vaart (1991a) or Theorem 9.3 in Van der Vaart (2000)). □

Thus, to any set of limit-laws of an estimator there is a randomized estimator in the limit experiment which has the same set of laws. If the asymptotic performance of an estimator is considered to be determined by its sets of limit laws, the limit experiment thus gives a lower bound to what is possible: along the sequence of experiments you cannot do better than the best procedure in the limit experiment.

To discuss efficient estimation we need to prescribe what we judge to be optimal in the Poisson limit experiment. Often a normal location experiment is the limit experiment. For such a normal location experiment, i.e. estimate h on basis of one observation Y from $\mathrm{N}(h,\tau)$ (τ known), it is natural to restrict to location-equivariant estimators. For this class one has a convolution-property (see, for example, Van der Vaart (2000, Proposition 8.4) or Janssen and Ostrovski (2005)): the law of every location-equivariant estimator T of h can be decomposed as $T \overset{d}{=} Y + V$, where V is independent of Y. This yields, by Anderson's lemma (see, for example, Lemma 8.5 in Van der Vaart (2000)), efficiency of Y (within the class of location-equivariant estimators) for all bowl-shaped loss functions. More general, there are convolution-results for shift-experiments (see, for example, Hirano and Porter (2003b)). However, the Poisson limit experiment $\mathscr{E}(G)$ has not a natural shift structure. In such a Poisson setting it seems reasonable to minimize variance amongst the unbiased estimators.

Proposition 4.3.2. Suppose G is such that $0 < g(0) < 1$ and $\mu_G < \infty$. In the experiment,

$$\mathscr{E}(G) = \left(\mathbb{Z}_+, 2^{\mathbb{Z}_+}, \left(\mathbb{Q}_h = \text{Poisson}(hg(0)\mu_G/2) \mid h \geq 0\right)\right),$$

the unbiased estimator $2Z/g(0)\mu_G$ minimizes the variance amongst all randomized estimators $t(Z,U)$ for which $\mathbb{E}_h t(Z,U) = h$ for all $h \geq 0$, i.e.

$$\mathrm{var}_h\, t(Z,U) \geq \mathrm{var}_h \left(\frac{2Z}{g(0)\mu_G} \right) = \frac{2h}{g(0)\mu_G} \text{ for all } h \geq 0.$$

Proof.
This is an immediate consequence of the Lehmann-Scheffé theorem. □

A combination of this proposition with Proposition 4.3.1 yields a variance lower-bound to asymptotically unbiased estimators in the sequence of experiments $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$.

Proposition 4.3.3. Suppose G satisfies Assumption 1. If $(T_n)_{n\in\mathbb{N}}$ is an estimator of h in the sequence of experiments $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$ such that $\mathscr{L}(T_n|\mathbb{P}_{1-h/n^2}) \to Z_h$ with $\int z\, \mathrm{d}Z_h(z) = h$ for all $h \geq 0$, then we have

$$\int (z-h)^2\, \mathrm{d}Z_h(z) \geq \frac{2h}{g(0)\mu_G}, \text{ for all } h \geq 0. \tag{4.26}$$

Proof.
By Proposition 4.3.1 there exists a randomized estimator $t(Z,U)$ in the limit experiment such that $Z_h = \mathscr{L}(t(Z,U) \mid \mathbb{Q}_h \times \mathrm{Uniform}[0,1])$. Hence $\mathbb{E}_h t(Z,U) = h$ and $\mathrm{var}_h\, t(Z,U) = \int (z-h)^2\, \mathrm{d}Z_h(z)$. Now the result follows from Proposition 4.3.2. □

It is not unnatural to restrict to estimators that satisfy $\mathscr{L}(T_n \mid \mathbb{P}_{1-h/n^2}) \to Z_h$. We make the additional restriction that $\int z\, \mathrm{d}Z_h(Z) = h$, i.e. the *limit*-distribution is unbiased. Now, based on the previous proposition, it is natural to call an estimator in this class efficient if it attains the variance-bound (4.26). To demonstrate efficiency of a given estimator, one only needs to show that it belongs to the class of asymptotically unbiased estimators, and that it attains the bound.

First we discuss the OLS estimator. Let $\theta_n = 1 - h/n^2$. Rewriting $X_t = \vartheta \circ X_{t-1} + \varepsilon_t = \mu_G + \theta_n X_{t-1} + u_t$ for $u_t = \varepsilon_t - \mu_G + \vartheta \circ X_{t-1} - \theta_n X_{t-1}$, we obtain the autoregression $X_t - \mu_G = \theta_n X_{t-1} + u_t$, which can also be written as $n^2(X_t - X_{t-1} - \mu_G) = h(-X_{t-1}) + n^2 u_t$ (note that indeed $\mathbb{E}_{\theta_n} u_t = \mathbb{E}_{\theta_n} X_{t-1} u_t = 0$). So the OLS estimator of θ_n is given by,

$$\widehat{\theta}_n^{\mathrm{OLS}} = \frac{\sum_{t=1}^n X_{t-1}(X_t - \mu_G)}{\sum_{t=1}^n X_{t-1}^2}, \tag{4.27}$$

and the OLS estimator of h is given by,

$$\widehat{h}_n^{\mathrm{OLS}} = -\frac{n^2 \sum_{t=1}^n X_{t-1}(X_t - X_{t-1} - \mu_G)}{\sum_{t=1}^n X_{t-1}^2} = n^2\left(1 - \widehat{\theta}_n^{\mathrm{OLS}}\right).$$

Ispány et al. (2003b) analyzed the asymptotic behavior of the OLS estimator under localizing rate n. However, since the convergence of experiments takes place at rate n^2, we analyze the behavior of the OLS estimator also under localizing rate n^2. The next proposition gives this behavior.

Proposition 4.3.4. If $\mathbb{E}_G\varepsilon_1^3 < \infty$, then we have, for all $h \geq 0$,

$$\left|\widehat{h}_n^{\mathrm{OLS}}\right| \xrightarrow{p} \infty, \text{ under } \mathbb{P}_{1-\frac{h}{n^2}}.$$

Proof.
Let $h \geq 0$ and set $\theta_n = 1 - h/n^2$, $\mathbb{P}_n = \mathbb{P}_{\theta_n}$, and $\mathbb{E}_n(\cdot) = \mathbb{E}_{\theta_n}(\cdot)$. We have

$$n^{3/2}\left(\widehat{\theta}_n^{\mathrm{OLS}} - \theta_n\right) = \frac{n^{-3/2}\sum_{t=1}^n X_{t-1}\left(\varepsilon_t - \mu_G + \vartheta \circ X_{t-1} - \theta_n X_{t-1}\right)}{n^{-3}\sum_{t=1}^n X_{t-1}^2}.$$

We prove that,

$$n^{-3/2}\sum_{t=1}^n X_{t-1}\left(\vartheta \circ X_{t-1} - \theta_n X_{t-1}\right) \xrightarrow{p} 0, \text{ under } \mathbb{P}_n, \tag{4.28}$$

$$n^{-3}\sum_{t=1}^n X_{t-1}^2 \xrightarrow{p} \frac{\mu_G^2}{3}, \text{ under } \mathbb{P}_n, \tag{4.29}$$

$$n^{-3/2}\sum_{t=1}^n X_{t-1}\left(\varepsilon_t - \mu_G\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{\sigma_G^2\mu_G^2}{3}\right), \text{ under } \mathbb{P}_n, \tag{4.30}$$

all hold, which yields,

$$n^{3/2}\left(\widehat{\theta}_n^{\mathrm{OLS}} - \theta_n\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{3\sigma_G^2}{\mu_G^2}\right), \text{ under } \mathbb{P}_n, \tag{4.31}$$

which in turn will yield the result, since,

$$\left|\widehat{h}_n^{\mathrm{OLS}}\right| = \sqrt{n}\left|-n^{3/2}\left(\widehat{\theta}_n^{\mathrm{OLS}} - \theta_n\right) + \frac{h}{\sqrt{n}}\right|.$$

First, we treat (4.28). Since

$$\begin{aligned}\mathbb{E}_n X_{t-1}|\vartheta \circ X_{t-1} - \theta_n X_{t-1}| &\leq \mathbb{E}_n X_{t-1}^2 |Z_1^{(t)} - \theta_n| = \mathbb{E}_n X_{t-1}^2 \mathbb{E}_n |Z_1^{(t)} - \theta_n| \\ &= 2\theta_n(1-\theta_n)\mathbb{E}_n X_{t-1}^2 \leq \frac{2h}{n^2}\mathbb{E}_n X_{t-1}^2,\end{aligned}$$

we obtain, using (4.3),

$$\frac{1}{n^{3/2}}\sum_{t=1}^n \mathbb{E}_n X_{t-1}|\vartheta \circ X_{t-1} - \theta_n X_{t-1}| \leq \frac{2h}{n^{7/2}}\sum_{t=1}^n \mathbb{E}_n X_{t-1}^2 \to 0,$$

which implies (4.28). Next, we discuss (4.29). Introduce $S_t = \sum_{i=1}^t \varepsilon_i$ and $Y_t = S_t - X_t$. Notice that Y_t is nonnegative, $Y_s = Y_{s-1} + (X_{s-1} - \vartheta \circ X_{s-1})$ for $s \geq 1$, $Y_0 = 0$, and thus $Y_t = \sum_{i=1}^t (X_{i-1} - \vartheta \circ X_{i-1})$. Decompose $X_t^2 = Y_t^2 + S_t^2 - 2S_t Y_t$. It is straightforward to check that $n^{-3}\sum_{t=1}^n S_t^2 \xrightarrow{p} \mu_G^2/3$, under $\mathbb{P}_n$. To obtain (4.29), it thus suffices to prove that $n^{-3}\sum_{t=1}^n \mathbb{E}_n Y_t^2 \to 0$ and $n^{-3}\sum_{t=1}^n S_t Y_t \xrightarrow{p} 0$ under $\mathbb{P}_n$. First notice that, for a constant $C > 0$,

$$\mathbb{E}_n(X_{j-1} - \vartheta \circ X_{j-1})X_{i-1} \leq \sqrt{\mathbb{E}_n X_{i-1}^2 \mathbb{E}_n (X_{j-1} - \vartheta \circ X_{j-1})^2} \leq Ci\left(\frac{\sqrt{h}}{n}\sqrt{j} + \frac{h}{n^2} j\right).$$

Now we obtain, use that conditional on X_t, $X_t - \vartheta \circ X_t$ has a Binomial$(X_t, h/n^2)$ distribution,

$$\begin{aligned}
\mathbb{E}_n Y_t^2 &= \sum_{i=1}^t \sum_{j=1}^t \mathbb{E}_n (X_{i-1} - \vartheta \circ X_{i-1})(X_{j-1} - \vartheta \circ X_{j-1}) \\
&= \sum_{i=1}^t \mathbb{E}_n (X_{i-1} - \vartheta \circ X_{i-1})^2 + 2\sum_{i=1}^t \sum_{j=1}^{i-1} \frac{h}{n^2} \mathbb{E}_n (X_{j-1} - \vartheta \circ X_{j-1}) X_{i-1} \\
&\leq \sum_{i=1}^t \left(\frac{h}{n^2}\mathbb{E}_n X_{i-1} + \frac{h^2}{n^4}\mathbb{E}_n X_{i-1}^2\right) + \frac{2Ch}{n^2}\sum_{i=1}^t \sum_{j=1}^{i-1} i\left(\frac{\sqrt{h}}{n}\sqrt{j} + \frac{h}{n^2} j\right).
\end{aligned}$$

Since $n^{-4}\sum_{t=1}^n \sum_{s=1}^n \mathbb{E}_n X_s X_t$ converges by (4.2), we obtain $n^{-3}\sum_{t=1}^n \mathbb{E}_n Y_t^2 \xrightarrow{p} 0$, under $\mathbb{P}_n$. Furthermore, we have,

$$\begin{aligned}
\frac{1}{n^3}\sum_{t=1}^n \mathbb{E}_n S_t Y_t &\leq \frac{1}{n^3}\sum_{t=1}^n \sqrt{\mathbb{E}_n S_t^2 \mathbb{E}_n Y_t^2} \leq \frac{\sqrt{\mu^2 + \sigma^2}}{n^3}\sum_{t=1}^n t\sqrt{\mathbb{E}_n Y_t^2} \\
&\leq \frac{\sqrt{\mu^2 + \sigma^2}}{n^3}\sqrt{\frac{n(2n+1)(n+1)}{6}}\sqrt{\sum_{t=1}^n \mathbb{E}_n Y_t^2} \to 0,
\end{aligned}$$

which concludes the proof of (4.29).

Finally, we treat (4.30). By a martingale central limit theorem for arrays (see Theorem 3.2, Corollary 3.1 and the remark after that corollary in Hall and Heyde (1980)) we have (4.30), if the following two conditions are satisfied,

$$\frac{1}{n^3}\sum_{t=1}^n X_{t-1}^2 \mathbb{E}_n\left[(\varepsilon_t - \mu_G)^2 \mid X_{t-1}, \ldots, X_0\right] \xrightarrow{p} \frac{\sigma_G^2 \mu_G^2}{3}, \text{ under } \mathbb{P}_n, \tag{4.32}$$

and for all $\epsilon > 0$,

$$\frac{1}{n^3}\sum_{t=1}^n X_{t-1}^2 \mathbb{E}_n\left[(\varepsilon_t - \mu_G)^2 1\{X_{t-1}|\varepsilon_t - \mu_G| > \epsilon n^{3/2}\} \mid X_{t-1}, \ldots, X_0\right] \xrightarrow{p} 0,, \tag{4.33}$$

under $\mathbb{P}_n$. Since ε_t is independent of X_{t-1} (4.32) immediately follows from (4.29). To see that the Lindeberg condition (4.33) is satisfied, notice that, using the independence of ε_t and X_{t-1}, Hölder's inequality, and Markov's inequality, we have

$$\begin{aligned}
&\mathbb{E}_n\left[\left(\varepsilon_t-\mu_G\right)^2 1\{X_{t-1}|\varepsilon_t-\mu_G|>\epsilon n^{3/2}\}\mid X_{t-1}\right] \\
&\qquad\le \left(\mathbb{E}_G(\varepsilon_1-\mu_G)^3\right)^{2/3}\left(\mathbb{P}_n\left[\left|\varepsilon_t-\mu_G\right|>\frac{\epsilon n^{3/2}}{X_{t-1}}\mid X_{t-1}\right]\right)^{1/3} \\
&\qquad\le \left(\mathbb{E}_G(\varepsilon_1-\mu_G)^3\right)^{2/3}\left(\frac{X_{t-1}^3\mathbb{E}_G(\varepsilon_1-\mu_G)^3}{\epsilon^3 n^{9/2}}\right)^{1/3} = \frac{X_{t-1}\mathbb{E}_G(\varepsilon_1-\mu_G)^3}{\epsilon n^{3/2}},
\end{aligned}$$

which yields,

$$\begin{aligned}
&\frac{1}{n^3}\sum_{t=1}^n X_{t-1}^2\mathbb{E}_n\left[\left(\varepsilon_t-\mu_G\right)^2 1\{X_{t-1}|\varepsilon_t-\mu_G|>\epsilon n^{3/2}\}\mid X_{t-1}\right] \\
&\qquad\qquad\le \frac{\mathbb{E}_G(\varepsilon_1-\mu_G)^3}{\epsilon n^{9/2}}\sum_{t=1}^n X_{t-1}^3 \xrightarrow{p} 0 \text{ under } \mathbb{P}_n,
\end{aligned}$$

since it is easily checked that $n^{-(4+\alpha)}\sum_{t=1}^n X_{t-1}^3 \xrightarrow{p} 0$, under $\mathbb{P}_n$, for $\alpha>0$. This concludes the proof. □

Remark 18. A similar result holds for the OLS-estimator in the model where G is unknown.

Thus the OLS estimator explodes. How should we estimate h then? Recall, that we interpreted $\sum_{t=1}^n 1\{\Delta X_t<0\}$ as an approximately sufficient statistic for h. Hence, it is natural to try to construct an efficient estimator based on this statistic. Using Theorem 4.1 we see that this is indeed possible.

Corollary 4.3. *Let G satisfy Assumption 1. The estimator,*

$$\widehat{h}_n = \frac{2\sum_{t=1}^n 1\{\Delta X_t<0\}}{g(0)\mu_G}, \tag{4.34}$$

is an efficient estimator of h in the sequence $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$.

A semiparametric model

So far we assumed that G is known. In this section, where we instead consider a semiparametric model, we hardly impose conditions on G (see, for example, Wefelmeyer (1996) for semiparametric stationary Markov models). The dependence of $\mathbb{P}_\theta$ upon G is made explicit by adding a subscript: $\mathbb{P}_{\theta,G}$. Formally, we consider the sequence of experiments,

$$\mathscr{E}_n = \left(\mathbb{Z}_+^{n+1}, 2^{\mathbb{Z}_+^{n+1}}, \left(\mathbb{P}^{(n)}_{1-\frac{h}{n^2},G}\mid (h,G)\in[0,\infty)\times\mathscr{G}\right)\right), \quad n\in\mathbb{N},$$

where $\mathscr{G}$ is the set of all distributions on $\mathbb{Z}_+$ that satisfy Assumption 1.

The goal is to estimate h efficiently. Here efficient, just as in the previous section, means asymptotically unbiased with minimal variance. Since the semiparametric model is more realistic, the estimation of h becomes more difficult. As we will see, the situation for our semiparametric model is quite fortunate: we can estimate h with the same asymptotic precision as in the case that G is known. In semiparametric statistics this is called adaptive estimation.

The efficient estimator for the case that G is known cannot be used anymore, since it depends on $g(0)$ and μ_G. The obvious idea is to replace these objects by estimates. The next proposition provides consistent estimators.

Proposition 4.3.5. Let $h \geq 0$ and G satisfy $\sigma_G^2 < \infty$. Then we have,

$$\widehat{g}_n(0) = \frac{1}{n}\sum_{t=1}^{n} 1\{X_t = X_{t-1}\} \xrightarrow{p} g(0) \text{ and } \widehat{\mu}_{G,n} = \frac{X_n}{n} \xrightarrow{p} \mu_G, \text{ under } \mathbb{P}_{1-\frac{h}{n^2},G}.$$

Proof.
Notice first that we have,

$$\frac{1}{n}\sum_{t=1}^{n}(X_{t-1} - \vartheta \circ X_{t-1}) \xrightarrow{p} 0, \text{ under } \mathbb{P}_{1-\frac{h}{n^2},G}, \tag{4.35}$$

since, condition on X_{t-1} and use (4.2),

$$0 \leq \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{1-\frac{h}{n^2},G}(X_{t-1} - \vartheta \circ X_{t-1}) = \frac{h}{n^3}\sum_{t=1}^{n}\mathbb{E}_{1-\frac{h}{n^2},G}X_{t-1} \to 0.$$

Using that $|1\{X_t = X_{t-1}\} - 1\{\varepsilon_t = 0\}| = 1$ only if $X_{t-1} - \vartheta \circ X_{t-1} \geq 1$, we easily obtain, by using (4.35),

$$\begin{aligned}\left|\widehat{g}_n(0) - \frac{1}{n}\sum_{t=1}^{n} 1\{\varepsilon_t = 0\}\right| &\leq \frac{1}{n}\sum_{t=1}^{n} 1\{X_{t-1} - \vartheta \circ X_{t-1} \geq 1\} \\ &\leq \frac{1}{n}\sum_{t=1}^{n}(X_{t-1} - \vartheta \circ X_{t-1}) \xrightarrow{p} 0.\end{aligned}$$

Now the result for $\hat{g}_n(0)$ follows by applying the weak law of large numbers to $n^{-1}\sum_{t=1}^{n} 1\{\varepsilon_t = 0\}$. Next, consider $\widehat{\mu}_{G,n}$. We have, use (4.35) and the weak law of large numbers for $n^{-1}\sum_{t=1}^{n}\varepsilon_t$,

$$\widehat{\mu}_{G,n} = \frac{X_n}{n} = \frac{1}{n}\sum_{t=1}^{n}(X_t - X_{t-1}) = \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t - \frac{1}{n}\sum_{t=1}^{n}(X_{t-1} - \vartheta \circ X_{t-1}) \xrightarrow{p} \mu_G,$$

under $\mathbb{P}_{1-h/n^2,G}$, which concludes the proof. □

From the previous proposition we have $\widehat{h}_n - \tilde{h}_n \xrightarrow{p} 0$, under $\mathbb{P}_{1-h/n^2,G}$, where

$$\tilde{h}_n = \frac{2\sum_{t=1}^n 1\{\Delta X_t < 0\}}{\widehat{g}_n(0)\widehat{\mu}_{G,n}}.$$

This implies that estimation of h in the semiparametric experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$ is not harder than estimation of h in $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$. In semiparametric parlor: the semiparametric problem is adaptive to $\mathscr{G}$. The precise statement is given in the following corollary; the proof is trivial.

Corollary 4.4. *If $(T_n)_{n\in\mathbb{N}}$ is a sequence of estimators in the semiparametric sequence of experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$ such that $\mathscr{L}\left(T_n \mid \mathbb{P}_{1-h/n^2,G}\right) \to Z_{h,G}$, $\int z\,\mathrm{d}Z_{h,G}(z) = h$ for all $(h,G) \in [0,\infty) \times \mathscr{G}$, then we have*

$$\int (z-h)^2\,\mathrm{d}Z_{h,G}(z) \geq \frac{2h}{g(0)\mu_G} \text{ for all } (h,G) \in [0,\infty) \times \mathscr{G}.$$

The estimator $\tilde{h}_n$ satisfies the conditions and achieves the variance bound.

Efficient estimation in the global model in case G is known 4.3.2

For convenience we introduce $\mathscr{X}_n = \mathbb{Z}_+^{n+1}$ and $\mathscr{A}_n = 2^{\mathbb{Z}_+^{n+1}}$, and the following assumption.

Assumption 2. A probability distribution G on $\mathbb{Z}_+$ is said to satisfy Assumption 2 if $g(k) > 0$ for all $k \in \mathbb{Z}_+$, $\mathbb{E}_G\varepsilon_1^3 < \infty$, and $\sum_{k=1}^\infty g^2(k-1)/g(k) < \infty$.

So far we considered nearly unstable INAR experiments. This section considers global experiments for the case G known, i.e.

$$\mathscr{D}_n(G) = \left(\mathscr{X}_n, \mathscr{A}_n, \left(\mathbb{P}_\theta^{(n)} \mid \theta \in [0,1]\right)\right), \quad n \in \mathbb{N}.$$

The goal is to estimate the autoregression parameter θ efficiently.

We already analyzed the 'stable' sequence of experiments

$$\mathscr{D}_n^{(0,1)}(G) = \left(\mathscr{X}_n, \mathscr{A}_n, \left(\mathbb{P}_\theta^{(n)} \mid \theta \in (0,1)\right)\right), \quad n \in \mathbb{N},$$

in Chapter 2. Under Assumption 2 it follows from Theorem 2.1 that these experiments are of the Local Asymptotic Normal form (at $\sqrt{n}$-rate). Recall that an estimator T_n of θ is regular if for all $\theta \in (0,1)$ there exists a law L_θ such that for all $h \in \mathbb{R}$,

$$\mathscr{L}\left(\sqrt{n}\left(T_n - \left(\theta + \frac{h}{\sqrt{n}}\right)\right) \mid \mathbb{P}_{\theta+h/\sqrt{n}}\right) \to L_\theta,$$

i.e. vanishing perturbations do not influence the limiting distribution (or more accurately: the associated estimators in the local limit experiment are location-equivariant). For LAN experiments, the Hájek-Le Cam convolution theorem tells us that for every regular estimator T_n of θ we have: $L_\theta = \mathrm{N}(0, I_\theta^{-1}) \oplus \Delta_{\theta,(T_n)}$, where $I_\theta > 0$ (which does not depend on the estimator, and thus is unavoidable noise) is the Fisher-information (see Theorem 2.1 for the formula). Since $\Delta_{\theta,(T_n)}$ is additional noise, one calls a regular estimator efficient if $\Delta_{\theta,(T_n)}$ is degenerated at $\{0\}$. Section 2.2 provides an (computationally attractive) efficient estimator of θ by updating the OLS estimator into an efficient estimator. Let us recall this estimator. Let $\hat{\theta}_n^*$ be a discretized version of $\hat{\theta}_n^{\mathrm{OLS}}$ (for $n \in \mathbb{N}$ make a grid of intervals with lengths $1/\sqrt{n}$, over $\mathbb{R}$ and, given $\hat{\theta}_n^{\mathrm{OLS}}$, define $\hat{\theta}_n^*$ to be the midpoint of the interval into which $\hat{\theta}_n^{\mathrm{OLS}}$ falls). Then,

$$\theta_n^{(0,1)} = \hat{\theta}_n^* + \frac{1}{n} \sum_{t=1}^{n} \hat{I}_{\theta,n}^{-1} \dot{\ell}_\theta(X_{t-1}, X_t; \hat{\theta}_n^*, G), \tag{4.36}$$

where,

$$\hat{I}_{n,\theta} = \frac{1}{n} \sum_{t=1}^{n} \dot{\ell}_\theta^2(X_{t-1}, X_t; \hat{\theta}_n^*, G),$$

is an efficient estimator of θ in the sequence of experiments $\mathscr{D}_n^{(0,1)}(G)$, $n \in \mathbb{N}$.

The difference between $\mathscr{D}_n^{(0,1)}(G)$ and $\mathscr{D}_n(G)$ is that in $\mathscr{D}_n(G)$ the full parameter space is used. To consider estimation in the full model, we also need to consider the local asymptotic structure of $\mathscr{D}_n(G)$ at $\theta = 0$ and $\theta = 1$. For $\theta = 1$ we have already done this by determining the limit experiment of $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$. The next proposition shows that for $\theta = 0$ the situation is standard: we have the LAN-property.

Proposition 4.3.6. Suppose G satisfies Assumption 2. Then $(\mathscr{D}_n(G))_{n\in\mathbb{N}}$ has the LAN-property at $\theta = 0$, i.e. for $h \geq 0$ we have,

$$\sum_{t=1}^{n} \log \frac{P_{X_{t-1},X_t}^{h/\sqrt{n}}}{P_{X_{t-1},X_t}^{0}} = \sum_{t=1}^{n} \log \frac{P_{X_{t-1},X_t}^{h/\sqrt{n}}}{g(X_t)} = hS_n^0 - \frac{h^2}{2} I_0 + o(\mathbb{P}_0; 1), \tag{4.37}$$

where,

$$I_0 = (\sigma_G^2 + \mu_G^2)\mathbb{E}_G \left(\frac{g(\varepsilon_1) - g(\varepsilon_1 - 1)}{g(\varepsilon_1)} \right)^2,$$

$$S_n^0 = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} -X_{t-1} \left(\frac{g(X_t) - g(X_t - 1)}{g(X_t)} \right) \xrightarrow{d} \mathrm{N}(0, I_0), \text{ under } \mathbb{P}_0.$$

Proof.
Note first that under $\mathbb{P}_0$ we have $X_t = \varepsilon_t$. Since we are localizing at $\theta = 0$, the

following representation of the transition probabilities is convenient, $P^{\theta}_{x_{t-1},x_t} = \sum_{k=0}^{x_{t-1}} \mathrm{b}_{x_{t-1},\theta}(k) g(x_t - k)$. Using the inequality $\log((a+b)/c) - \log(a/c) \leq b/a$ for $a, c > 0$, $b \geq 0$ we obtain, for $h > 0$,

$$\log \frac{P^{h/\sqrt{n}}_{X_{t-1},X_t}}{g(X_t)} - \log \frac{\sum_{k=0}^{2} \mathrm{b}_{X_{t-1},\frac{h}{\sqrt{n}}}(k) g(X_t - k)}{g(X_t)} \leq R_t = \frac{\sum_{k=3}^{X_{t-1}} \mathrm{b}_{X_{t-1},\frac{h}{\sqrt{n}}}(k) g(X_t - k)}{\sum_{k=0}^{2} \mathrm{b}_{X_{t-1},\frac{h}{\sqrt{n}}}(k) g(X_t - k)}.$$

On the event $A_n = \left\{\forall\, t \in \{1, \ldots, n\} : h\varepsilon_t < \sqrt{n}\right\}$ we have for some constant $K \geq 0$, using (4.43) and the assumption that G is eventually decreasing,

$$R_t \leq \frac{2K \mathrm{b}_{X_{t-1},\frac{h}{\sqrt{n}}}(3)}{\left(1 - \frac{h}{\sqrt{n}}\right)^{X_{t-1}}} \leq \frac{K h^3 X_{t-1}^3}{3n\sqrt{n}\left(1 - \frac{h}{\sqrt{n}}\right)^3}.$$

Using $\mathbb{E}_G \varepsilon_1^3 < \infty$ and Markov's inequality, it is easy to see that $\lim_{n\to\infty} \mathbb{P}_0(A_n^c) = 0$. From this it easily follows that $\sum_{t=1}^n R_t \xrightarrow{p} 0$ under $\mathbb{P}_0$. We decompose,

$$\begin{aligned} L_{tn} = \log \frac{\sum_{k=0}^{2} \mathrm{b}_{X_{t-1},\frac{h}{\sqrt{n}}}(k) g(X_t - k)}{g(X_t)} &= (X_{t-1} - 2) \log\left(1 - \frac{h}{\sqrt{n}}\right) \\ &\quad + \log\left(1 + A_n + B_{tn} + C_{tn}\right), \end{aligned}$$

where,

$$A_n = -\frac{2h}{\sqrt{n}} + \frac{h^2}{n}, \quad B_{tn} = \frac{h}{\sqrt{n}} X_{t-1} \left(1 - \frac{h}{\sqrt{n}}\right) \frac{g(X_t - 1)}{g(X_t)},$$

and,

$$C_{tn} = \frac{X_{t-1}(X_{t-1} - 1)}{2} \frac{h^2}{n} \frac{g(X_t - 2)}{g(X_t)}.$$

From here on the proof continues in the classical way. Using the Taylor expansion $\log(1 + x) = x - x^2/2 + x^2 r(x)$, where r satisfies $r(x) \to 0$ as $x \to 0$, we make the decomposition,

$$\begin{aligned} \log(1 + A_n + B_{tn} + C_{tn}) = {} & A_n + B_{tn} + C_{tn} + R_{tn} \\ & - \frac{1}{2}\left(A_n^2 + B_{tn}^2 + C_{tn}^2 + 2A_n B_{tn} + 2A_n C_{tn} + 2B_{tn} C_{tn}\right), \end{aligned}$$

where $R_{tn} = (A_n + B_{tn} + C_{tn})^2\, r(A_n + B_{tn} + C_{tn})$. It is easy to see that the terms $\sum_{t=1}^n C_{tn}^2$, $\sum_{t=1}^n B_{tn} C_{tn}$ and $\sum_{t=1}^n A_n C_{tn}$ are all $o(\mathbb{P}_0; 1)$. Furthermore, we have,

$$\sum_{t=1}^n \left\{(X_{t-1} - 2) \log\left(1 - \frac{h}{\sqrt{n}}\right) + A_n - \frac{1}{2} A_n^2\right\} = -\frac{h}{\sqrt{n}} \sum_{t=1}^n X_{t-1} - \frac{h^2}{2n} \sum_{t=1}^n X_{t-1} + o(\mathbb{P}_0; 1),$$

and,

$$-\frac{h}{\sqrt{n}}\sum_{t=1}^{n} X_{t-1} + \sum_{t=1}^{n} B_{tn} = hS_n^0 - \frac{h^2}{n}\sum_{t=1}^{n} X_{t-1}\frac{g(X_t-1)}{g(X_t)}.$$

Combining the previous displays we obtain,

$$L_{tn} = hS_n^0 + \sum_{t=1}^{n}\left\{C_{tn} - \frac{1}{2}B_{tn}^2 - A_n B_{tn} - \frac{h^2}{2n}X_{t-1} - \frac{h^2}{n}X_{t-1}\frac{g(X_t-1)}{g(X_t)}\right\} + R_{tn} + o(\mathbb{P}_0;1).$$

By the law of large numbers we have (note that $\mathbb{E}_0 g(X_t-i)/g(X_t) = 1$, $i = 1,2$), $\sum_{t=1}^n C_{tn} \xrightarrow{p} h^2\left(\sigma_G^2 + \mu_G^2 - \mu_G\right)/2$, $\sum_{t=1}^n A_n B_{tn} \xrightarrow{p} -2h^2\mu_G$, $(1/n)\sum_{t=1}^n X_{t-1} \xrightarrow{p} \mu_G$, $(1/n)\sum_{t=1}^n X_{t-1}g(X_t-1)/g(X_t) \xrightarrow{p} \mu_G$, and $\sum_{t=1}^n B_{tn}^2 \xrightarrow{p} h^2(\sigma_G^2+\mu_G^2)\mathbb{E}(g^2(\varepsilon_1-1)/g^2(\varepsilon_1))$ under $\mathbb{P}_0$. Thus, once we show that $\sum_{t=1}^n R_{tn} = o(\mathbb{P}_0;1)$ the proposition is proved. Using the inequality $(x+y+z)^2 \le 9(x^2+y^2+z^2)$ we easily obtain $\sum_{t=1}^n (A_n + B_{tn} + C_{tn})^2 = O(\mathbb{P}_0;1)$. And using Markov's inequality it is easy to see that, for $\epsilon > 0$, $\mathbb{P}_0\{\max_{1\le t\le n}|A_n+B_{tn}+C_{tn}| > \epsilon\} \le \sum_{t=1}^n \mathbb{P}_0\{|A_n+B_{tn}+C_{tn}| > \epsilon\} \to 0$. Thus $\sum_{t=1}^n (A_n+B_{tn}+C_{tn})^2 r(A_n+B_{tn}+C_{tn}) \xrightarrow{p} 0$ under $\mathbb{P}_0$, which concludes the proof. □

Remark 19. The meaning of this LAN-result is that the sequence,

$$\left(\mathscr{X}_n, \mathscr{A}_n, \left(\mathbb{P}^{(n)}_{h/\sqrt{n}} \mid h \ge 0\right)\right), \quad n \in \mathbb{N},$$

of local experiments, converges to the experiment $\left((\mathbb{R},\mathscr{B}(\mathbb{R}), \left(\mathrm{N}\left(h, I_0^{-1}\right) \mid h \ge 0\right)\right)$. Note that we are dealing here with a 'one-sided' LAN-result, i.e. we only consider h positive. As a consequence, it is not possible to apply the standard results for experiments with the LAN-structure directly (this, since these are formulated for interior points of the parameter space). Since we do not want to discuss this issue further, we consider asymptotically centered estimators with minimal asymptotic variance as a best estimator at $\theta = 0$ (see below). We note that the 'information-loss principle', which we used in Chapter 2 to establish the LAN-property for $\theta \in (0,1)$, cannot be used here since the score of a Binomial distribution does not exist (in the usual sense) at $\theta = 0$. Finally we point out that the proposition actually holds under the weaker assumption $\sigma_G^2 < \infty$ instead of $\mathbb{E}_G\varepsilon_1^3 < \infty$. This requires a finer, lengthier analysis of the Taylor expansions using the Borel-Cantelli lemma. Since this proposition is just an input in the proof of Proposition 4.3.7 where we need $\mathbb{E}_G\varepsilon_1^3 < \infty$ anyway, we give here the simpler proof under $\mathbb{E}_G\varepsilon_1^3 < \infty$.

Now we completed the picture of the local asymptotic structures of $(\mathscr{D}_n(G))_{n\in\mathbb{N}}$ we can discuss efficient estimation. First, we describe the class of estimators in which we are interested. We consider estimators T_n that satisfy,

(i) $(\theta = 0)$ for all $h \geq 0$,

$$\mathscr{L}\left(\sqrt{n}\left(T_n - \frac{h}{\sqrt{n}}\right) \mid \mathbb{P}_{h/\sqrt{n}}\right) \to L_h, \text{ with } \int z \,\mathrm{d}L_h(z) = 0, \tag{4.38}$$

(ii) $(0 < \theta < 1)$ T_n is regular, i.e. for all $h \in \mathbb{R}$,

$$\mathscr{L}\left(\sqrt{n}\left(T_n - \left(\theta + \frac{h}{\sqrt{n}}\right)\right) \mid \mathbb{P}_{\theta + h/\sqrt{n}}\right) \to L_\theta, \tag{4.39}$$

(ii) $(\theta = 1)$ for all $h \geq 0$,

$$\mathscr{L}\left(n^2\left(T_n - \left(1 - \frac{h}{n^2}\right)\right) \mid \mathbb{P}_{1-h/n^2}\right) \to R_h \text{ with } \int z \,\mathrm{d}R_h(z) = 0. \tag{4.40}$$

So for $\theta \in (0,1)$ we ask for regularity which we discussed earlier. For $\theta = 0$ and $\theta = 1$ we only ask for a limiting distribution with mean zero. For any such estimator we have (the first inequality follows by arguments completely analogue to the derivation of the third inequality, we already discussed the second statement, and the third follows from Proposition 4.3.3 by taking $\widehat{h}_n = n^2(1 - T_n)$ as estimator of h),

$$\int z^2 \,\mathrm{d}L_h(z) \geq I_0^{-1}, \quad L_\theta = \mathrm{N}(0, I_\theta^{-1}) \oplus \Delta_{\theta,(T_n)}, \quad \int z^2 \,\mathrm{d}R_h(z) \geq \frac{2h}{g(0)\mu_G}, \tag{4.41}$$

for all $h \geq 0, \theta \in (0,1)$. Hence it is natural to call an estimator in the global model efficient if it satisfies (4.38)-(4.40) with $L_\theta = \mathrm{N}(0, I_\theta^{-1})$, $\int z^2 \,\mathrm{d}L_h(z) = I_0^{-1}$, and $\int z^2 \,\mathrm{d}R_h(z) = 2h/g(0)\mu_G$ for all $h \geq 0, \theta \in (0,1)$.

Proposition 4.3.7. Suppose G satisfies Assumptions 1 and 2. Let $\alpha, \beta \in (0, 1/2)$, and $c_\alpha, c_\beta > 0$. The estimator,

$$\begin{aligned}\widehat{\theta}_n = \theta_n^0 1\left\{|\widehat{\theta}_n^{\mathrm{OLS}}| \leq c_\alpha n^{-\alpha}\right\} + \theta_n^{(0,1)} 1\left\{|\widehat{\theta}_n^{\mathrm{OLS}}| > c_\alpha n^{-\alpha}, |\widehat{\theta}_n^{\mathrm{OLS}} - 1| > c_\beta n^{-\beta}\right\} \\ + \theta_n^1 1\left\{|\widehat{\theta}_n^{\mathrm{OLS}} - 1| \leq c_\beta n^{-\beta}\right\},\end{aligned}$$

where $\theta_n^{(0,1)}$ is defined in (4.36) and,

$$\theta_n^0 = \frac{1}{\sqrt{n}} I_0^{-1} S_n^0, \quad \theta_n^1 = 1 - \frac{2\sum_{t=1}^n 1\{\Delta X_t < 0\}}{n^2 g(0)\mu_G},$$

is an efficient estimator of θ in the sequence of experiments $(\mathscr{D}_n(G))_{n\in\mathbb{N}}$.

Proof.
From Le Cam's third lemma and Proposition 4.3.6 it easily follows that θ_n^0 satisfies (4.38) and attains its variance lower-bound in (4.41). Since $\theta_n^{(0,1)}$ is an

efficient estimator in the 'stable experiments' $(\mathscr{D}_n^{(0,1)}(G))_{n\in\mathbb{N}}$ it follows, by definition, that $\theta_n^{(0,1)}$ satisfies (4.39) and attains the convolution lower-bound in (4.41). And it is also clear (from Corollary 4.3) that θ_n^1 satisfies (4.40) and attains its variance lower-bound in (4.41). Thus it suffices to show that $\sqrt{n}\left(\widehat{\theta}_n - \theta_n^0\right) \xrightarrow{p} 0$ under $\mathbb{P}_{h/\sqrt{n}}$ for all $h \geq 0$, $\sqrt{n}\left(\widehat{\theta}_n - \theta_n^{(0,1)}\right) \xrightarrow{p} 0$ under $\mathbb{P}_{\theta+h/\sqrt{n}}$ for all $\theta \in (0,1)$, $h \in \mathbb{R}$, and $n^2\left(\widehat{\theta}_n - \theta_n^1\right) \xrightarrow{p} 0$ under $\mathbb{P}_{1-h/n^2}$ for all $h \geq 0$. It is an easy exercise, using a martingale central limit theorem, to show that $\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{OLS}} - (\theta + h/\sqrt{n})\right)$ converges to a normal distribution under $\mathbb{P}_{\theta+h/\sqrt{n}}$ for all $\theta \in [0,1)$ and $h \in \mathbb{R}$ (for $\theta = 0$ we only consider $h \geq 0$). And from (4.31) we have that $n^{3/2}(\widehat{\theta}_n^{\mathrm{OLS}} - (1 - h/n^2))$ converges to a normal distribution under $\mathbb{P}_{1-h/n^2}$ for $h \geq 0$. This implies that $n^\alpha \widehat{\theta}_n^{\mathrm{OLS}} \xrightarrow{p} 0$ under $\mathbb{P}_{h/\sqrt{n}}$ and $\mathbb{P}_{h/\sqrt{n}}\left\{\left|\widehat{\theta}_n^{\mathrm{OLS}} - 1\right| \leq c_\beta n^{-\beta}\right\} \to 0$ for $h \geq 0$, $n^\alpha \widehat{\theta}_n^{\mathrm{OLS}} \xrightarrow{p} \infty$ and $n^\beta \left|\widehat{\theta}_n^{\mathrm{OLS}} - 1\right| \xrightarrow{p} \infty$ under $\mathbb{P}_{\theta+h/\sqrt{n}}$, for $\theta \in (0,1)$, $h \in \mathbb{R}$, and we have $n^\beta\left(\widehat{\theta}_n^{\mathrm{OLS}} - 1\right) \xrightarrow{p} 0$ under $\mathbb{P}_{1-h/n^2}$ and $\mathbb{P}_{1-h/n^2}\left\{\left|\widehat{\theta}_n^{\mathrm{OLS}}\right| \leq c_\alpha n^{-\alpha}\right\} \to 0$ for $h \geq 0$. This concludes the proof. □

4.3.3 Testing for a unit root

This section discusses testing for a unit root in an INAR(1) model. We consider the case that G is known and satisfies Assumption 1.

In the global experiments $\mathscr{D}_n(G) = (\mathscr{X}_n, \mathscr{A}_n, (\mathbb{P}_\theta^{(n)} \mid \theta \in [0,1]))$, $n \in \mathbb{N}$, we want to test the hypothesis $\mathrm{H}_0 : \theta = 1$ versus $\mathrm{H}_1 : \theta < 1$. In other words, we want to test the null hypothesis of a unit root. Hellström (2001) considered this problem, from the perspective that one wants to use standard (that is, OLS) software routines in the testing. He derives, by Monte Carlo simulations, the finite sample null-distributions for a Dickey-Fuller test of a random walk with Poisson distributed errors. This (standard) Dickey-Fuller test statistic is given by the *usual* (i.e. non-corrected) t-test that the slope parameter equals 1, i.e.

$$\tau_n = \frac{\widehat{\theta}_n^{\mathrm{OLS}} - 1}{\sqrt{\sigma_G^2\left(\sum_{t=1}^n X_{t-1}^2\right)^{-1}}},$$

where $\widehat{\theta}_n^{\mathrm{OLS}}$ is given by (4.27). Under H_0, i.e. under $\mathbb{P}_1$, we have (we are now dealing with a random walk with drift), $\tau_n \xrightarrow{d} \mathrm{N}(0,1)$. Hence, the size $\alpha \in (0,1)$ Dickey-Fuller test rejects H_0 if and only if $\tau_n < \Phi^{-1}(\alpha)$. To analyze the performance of a test, one needs to consider the local asymptotic behavior of the test. Since $\mathscr{E}_n(G) \to \mathscr{E}(G)$ we should consider the performance of τ_n along the sequence $\mathscr{E}_n(G)$. The following proposition shows, however, that the asymptotic probability that the null hypothesis is rejected equals α for all alternatives. Hence, the standard Dickey-Fuller test has no power.

Proposition 4.3.8. If $\mathbb{E}_G\varepsilon_1^3 < \infty$ we have for all $h \geq 0$,

$$\tau_n \xrightarrow{d} \mathrm{N}(0,1), \text{ under } \mathbb{P}_{1-\frac{h}{n^2}},$$

which yields

$$\lim_{n\to\infty} \mathbb{P}_{1-\frac{h}{n^2}}\left(\text{reject } \mathrm{H}_0\right) = \alpha.$$

Proof.
From (4.29) and (4.31) the result easily follows. □

So the standard Dickey-Fuller test for a unit root does not behave well in the nearly unstable INAR(1) setting. In our sequence of experiments $\mathscr{E}_n(G)$, $n \in \mathbb{N}$, we propose the intuitively obvious tests

$$\psi_n(X_0,\ldots,X_n) = \begin{cases} \alpha, & \text{if } \sum_{t=1}^n 1\{\Delta X_t < 0\} = 0, \\ 1, & \text{if } \sum_{t=1}^n 1\{\Delta X_t < 0\} \geq 1, \end{cases}$$

i.e. reject H_0 if the process ever moves down and reject H_0 with probability α if there are no downward movements. We will see that this obvious test is in fact efficient.

To discuss efficiency of tests, we recall the implication of the Le Cam-Van der Vaart asymptotic representation theorem to testing (see, for example, Theorem 7.2 in Van der Vaart (1991a)). Let $\alpha \in (0,1)$ and ϕ_n be a sequence of tests in $(\mathscr{E}_n(G))_{n\in\mathbb{N}}$ such that $\limsup_{n\to\infty} \mathbb{E}_1\phi_n(X_0,\ldots,X_n) \leq \alpha$. Then we have

$$\limsup_{n\to\infty} \mathbb{E}_{1-\frac{h}{n^2}}\phi_n(X_0,\ldots,X_n) \leq \sup_{\phi\in\Phi_\alpha} \mathbb{E}_h\phi(Z) \text{ for all } h > 0,$$

where Φ_α is the collection of all level α tests for testing $\mathrm{H}_0 : h = 0$ versus $\mathrm{H}_1 : h > 0$ in the Poisson limit experiment $\mathscr{E}(G)$. If we have equality in the previous display, it is natural to call a test ϕ_n efficient. It is obvious that the uniform most powerful test in the Poisson limit experiment is given by

$$\phi(Z) = \begin{cases} \alpha, & \text{if } Z = 0, \\ 1, & \text{if } Z \geq 1. \end{cases}$$

Its power function is given by $\mathbb{E}_0\phi(Z) = \alpha$ and $\mathbb{E}_h\phi(Z) = 1-(1-\alpha)\exp(-hg(0)\mu_G/2)$. Using Theorem 4.1 we find

$$\lim_{n\to\infty} \mathbb{E}_1\psi_n(X_0,\ldots,X_n) = \alpha,$$

and,

$$\lim_{n\to\infty} \mathbb{E}_{1-\frac{h}{n^2}}\psi_n(X_0,\ldots,X_n) = 1-(1-\alpha)\exp\left(-\frac{hg(0)\mu_G}{2}\right) \text{ for } h > 0.$$

We conclude that the test ψ_n is indeed efficient.

4.4 Auxiliaries

This section contains auxiliary results from the literature that are specific to this chapter.

The following ‘tail-result’ for the Binomial distribution is basic (see, for instance, Feller (1968) pages 150-151), but since it is heavily applied, we recall it here for convenience.

Proposition 4.4.1. Let $m \in \mathbb{N}$, $p \in (0,1)$. If $r > mp$, we have

$$\sum_{k=r}^{m} \mathrm{b}_{m,p}(k) \le \mathrm{b}_{m,p}(r)\frac{r(1-p)}{r-mp}. \tag{4.42}$$

So, if $1 > mp$, we have for $r = 2,3$,

$$\sum_{k=r}^{m} \mathrm{b}_{m,p}(k) \le 2\,\mathrm{b}_{m,p}(r). \tag{4.43}$$

There is a large literature on Poisson approximation of the distribution of sums of dependent indicator variables with small success probabilities. Results of this kind are usually called ‘Poisson laws of small numbers’. For our application the following theorem by Serfling (1975) is the most convenient.

Lemma 4.4.1. Let $Z_1,\ldots,Z_n$ (possibly dependent) 0-1 valued random variables and set $S_n = \sum_{t=1}^{n} Z_t$. Let Y be Poisson distributed with mean $\sum_{t=1}^{n} \mathbb{E}Z_t$. Then we have

$$\sup_{A\subset\mathbb{Z}_+} |\mathbb{P}\{S_n \in A\} - \mathbb{P}\{Y \in A\}| \le \sum_{t=1}^{n} (\mathbb{E}Z_t)^2 + \sum_{t=1}^{n} \mathbb{E}|\mathbb{E}[Z_t \mid Z_1,\ldots,Z_{t-1}] - \mathbb{E}Z_t|.$$

Part II

Other essays

5 Efficient estimation of marginals by exploiting knowledge on the copula

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent random pairs with unknown (bivariate) distribution function H. The only knowledge we have on H is that its copula is C, i.e. $H(x, y) = C(F(x), G(y))$, $(x, y) \in \mathbb{R}^2$, where F and G are the (unknown) marginal distribution functions of H. These marginals can, of course, be estimated by their empirical distribution functions. However, we show that, amongst smooth copulas, these estimators are only efficient for the independence copula. This chapter shows how to exploit the information on the dependence structure in an optimal way.

5.1 Introduction

A (bivariate) copula C is the restriction to $[0, 1]^2$ of a bivariate distribution function with uniform marginal distributions; see Joe (1997) or Nelsen (1999) for an extensive introduction to copulas. Copulas are extremely attractive in studies where one needs to construct a multivariate model, since copulas allow to separate the modeling of the dependence between the coordinates from the modeling of the marginal distributions. More precise, if C is a copula and F and G are univariate distribution functions, then it is easy to show that $H(x, y) = C(F(x), G(y))$, $x, y \in \mathbb{R}$, defines a (bivariate) distribution function with marginal distributions F and G. The reverse is known as A. Sklar's (1959) theorem (see, for example, Nelsen (1999)): for a bivariate distribution function H with marginal distribution functions F, G, there is a copula C such that $H(x, y) = C(F(x), G(y))$, $x, y \in \mathbb{R}$. Moreover, if F and G are continuous the copula C is unique. Hence, when one has to model the distribution of a bivariate random variable, one can separate without loss of generality the modeling of the marginal distributions from the modeling of the dependence structure, by choosing a pair of marginal distribution functions and a copula. So a copula can be viewed upon as a margin-free description of dependence. This could be an explanation of the popularity of copulas in (financial) econometrics. Without going into details we mention some recent financial applications: Li (2000), Embrechts et al. (2003a), Embrechts et al. (2003b), Genest et al. (2005), Junker and May (2005), and Hu (2006).

Throughout this chapter, $\mathscr{F}$ denotes the collection of all distribution functions on $\mathbb{R}$, and $\mathscr{F}_{\mathrm{ac}}$ denotes the subset of all absolutely continuous distribution functions. Given a copula C and distribution functions F and G, define the probability measure $\mathbb{P}^C_{F,G}$ on $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$ by

$$\mathbb{P}^C_{F,G}\big((-\infty, x] \times (-\infty, y]\big) = C(F(x), G(y)), \quad x, y \in \mathbb{R}.$$

Typical applications of copula models select specific parametric forms for the dependence structure, i.e. C_θ, $\theta \in \Theta \subset \mathbb{R}^m$, and the marginals (for example, normal distributions or t-distributions). Next the question arises how we should estimate the parameters from a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. An obvious approach is to maximize the joint likelihood, either directly or in two steps as proposed by Joe (2005). However, inappropriate choices for the marginals could invalidate the estimation of the dependence parameter θ, i.e. expose the researcher to possible misspecification of the marginals. To avoid this problem, margin-free ad hoc estimates of θ were, in special contexts, developed by, amongst others, Clayton and Cuzick (1985), Oakes (1982, 1986), Genest (1987), and Hougaard (1989). Oakes (1994) introduced a semiparametric estimator of θ. This omnibus procedure consists of replacing, in the likelihood, F and G by the marginal empirical distribution functions. Genest et al. (1995), Shih and Louis (1995), and Tsukahara (2005) established consistency and asymptotic normality of this estimator. As an aside, we note that Chen and Fan (2006) showed that this estimation method also works in a univariate (first-order) Markov context. An obvious and interesting question is for which copulas the omnibus estimator constitutes a semiparametric efficient estimator of θ. Klaassen and Wellner (1997) proved that the omnibus procedure is efficient for the normal copula family, and is asymptotically equivalent to the Van der Waerden normal scores rank correlation estimator. Genest and Werker (2002) characterized the efficiency of the omnibus procedure. Amongst popular copula families, only two instances of semiparametric efficiency are identified: the case of independence and the normal copula model.

Instead of focusing on (efficient) estimation of the dependence parameter, we focus on using the knowledge on the dependence structure to construct improvements of the marginal empirical distribution functions. As far as we know Klaassen and Wellner (1997) were the first to consider efficient estimation of the marginals in a copula model: 'It would be very interesting to know information bounds and efficient estimators for estimation of the marginal distribution functions F and G in the bivariate normal copula model treated here, or in other copula models.' In this chapter we consider the following two models for one observation (X, Y):

$$\mathscr{P}(C, G_0) = \left(\mathbb{P}^C_{F,G_0} \mid F \in \mathscr{F}_{\text{ac}}\right), \quad G_0 \in \mathscr{F}_{\text{ac}}, \quad \text{and,} \quad \mathscr{P}(C) = \left(\mathbb{P}^C_{F,G} \mid F, G \in \mathscr{F}_{\text{ac}}\right).$$

So in the model $\mathscr{P}(C)$ the copula is known, and in the model $\mathscr{P}(C, G_0)$ the second marginal is also known. For the model $\mathscr{P}(C, G_0)$ we consider efficient estimation of the parameter F, seen as an element of the space $\ell^\infty(\mathbb{R})$, from a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. And for the model $\mathscr{P}(C)$ the goal is to develop an efficient estimator of the parameter (F, G), seen as an element of $\ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R})$. Although interesting from a theoretical point of view, the assumption that the copula is known is not realistic. Our results can however be used to construct efficient estimators of (F, G) in the model

$$\mathscr{P} = \left(\mathbb{P}^{C_\theta}_{F,G} \mid F, G \in \mathscr{F}_{\text{ac}}, \theta \in \Theta \subset \mathbb{R}^m\right),$$

where $(C_\theta \mid \theta \in \Theta)$ is a 'smooth' family of copulas: if one has an efficient estimator of θ available, then the 'plug-in principle' constitutes, see Klaassen and Putter (2005), an efficient estimator of (F, G). We also believe that the idea of our estimation technique and efficiency proof can be extended to yield a direct efficient estimator of the parameter (θ, F, G) in the model $\mathscr{P}$. However, since the semiparametric analysis of the models $\mathscr{P}(C, G_0)$ and $\mathscr{P}(C)$ is already nonstandard, we think it is reasonable to concentrate first on these models. Furthermore the analysis of the models $\mathscr{P}(C, G_0)$ and $\mathscr{P}(C)$ can be considered as the complement of Bickel et al. (1991), and Peng and Schick (2002, 2004, 2005). who considered efficient estimation of some aspect of the bivariate distribution function if one has complete information on one or both marginals and no further information.

Recently, and independently from our work, Chen et al. (2006) proposed an elegant sieve maximum likelihood estimation procedure for semiparametric copula models. Their procedure approximates the infinite-dimensional unknown marginal densities by linear combinations of finite-dimensional known basis functions, and then maximizes the joint likelihood with respect to the copula parameter and the sieve parameters of the approximating marginal densities. To prove that this approximation is valid they require a more restricted class of marginals than $\mathscr{F}_{\text{ac}}$: the support should be a compact interval or the whole real line, and the square roots of the densities should satisfy certain differentiability conditions. Relying on general theory, developed by Shen (1997), they provide conditions under which their sieve estimation method provides efficient estimates of real-valued smooth functionals of (θ, F, G). Unfortunately a sieve estimation method always brings ambiguity in the estimation method since one has to select a finite subset of the class of basis functions, and different shapes of the density require different sieves. Our estimators, which are motivated by an empirical likelihood argument, do not suffer this ambiguity. Furthermore, it is not completely clear what restrictions the conditions of Chen et al. (2006) put on the copula C.

The setup of the remainder of this chapter is as follows. Section 5.2 discusses some preliminaries. Section 5.3 analyzes the model $\mathscr{P}(C, G_0)$. In Section 5.3.1 we discuss the lower-bound to the asymptotic variance of regular estimators of F. We also prove that, amongst smooth copulas, the independence copula, $C(u, v) = uv$, is the only copula for which the empirical distribution function of $X_1, \ldots, X_n$ is an efficient estimator of F. Furthermore, it is discussed that, in general, we cannot obtain explicit expressions for the efficient influence operator which makes it hard to prove efficiency of an estimator. Section 5.3.2 introduces our estimator. We show that it can be computed by solving a linear system of $n+1$ equations in $n+1$ variables. In Section 5.3.3 we determine the limiting distribution of our estimator. The limiting distribution is a Gaussian process, but it seems to be impossible to obtain explicit formulas for its covariance process. Section 5.3.4 proves, using the special representation of the limiting distribution, efficiency of our estimator. In Section 5.4 we analyze the model $\mathscr{P}(C)$. From the efficiency point of view, the major difference to Section 5.3 is that the tangent space of the model is now the sum of two non-orthogonal spaces. At first sight this complicates the semiparametric analysis even further. However, the trick we used to prove efficiency in the model $\mathscr{P}(C, G_0)$ extends to the model $\mathscr{P}(C)$.

5.2 Assumptions and Notation

This section gives a precise description of our primitive assumption on the copula. Moreover, we introduce some notation we will use later on.

First we discuss our assumptions on the copula.

Assumptions on C

(C1) C is absolutely continuous w.r.t. Lebesgue measure. There is a version of its density, c, which is strictly positive on $(0,1)^2$.

(C2) The density c is two times continuously differentiable on $(0,1)^2$. Hence,

$$\dot{\ell}_i(u_1, u_2) = \frac{\partial}{\partial u_i} \log c(u_1, u_2), \quad u_1, u_2 \in (0,1),\ i = 1,2, \tag{5.1}$$

$$\ddot{\ell}_{ij}(u_1, u_2) = \frac{\partial^2}{\partial u_i \partial u_j} \log c(u_1, u_2), \quad u_1, u_2 \in (0,1),\ i, j = 1,2, \tag{5.2}$$

are well-defined. We impose that these objects are C-integrable and we impose,

$$\int \dot{\ell}_i(u_1,u_2)c(u_1,u_2)\,\mathrm{d}u_{3-i} = 0, \quad u_i \in (0,1),\ i=1,2.$$

Define

$$I_{11}(x) = \int_0^1 \dot{\ell}_1^2(x,y)c(x,y)\,\mathrm{d}y, \quad I_{22}(y) = \int_0^1 \dot{\ell}_2^2(x,y)c(x,y)\,\mathrm{d}x.$$

We also impose,

$$I_{ii}(u_i) = -\int \ddot{\ell}_{ii}(u_1,u_2)c(u_1,u_2)\,\mathrm{d}u_{3-i}, \quad u_i \in (0,1),\ i=1,2,$$

and, for some constant $M>0$,

$$I_{ii}(u) \le \frac{M}{(u(1-u))^2}, \quad u\in(0,1),\ i=1,2.$$

(C3) Define $r:(0,1)\to\mathbb{R}$ by $r(u)=u(1-u)$.

(i) There exists $M>0$ and $\alpha\in[0,1)$, such that, for $i=1,2$,

$$\left|\dot{\ell}_i(u_1,u_2)\right| \le \frac{M}{r(u_i)r^{\alpha}(u_{3-i})}, \quad (u_1,u_2)\in(0,1)^2.$$

(ii) For $i=1,2$, there exists $M\ge 0$, $\epsilon\in(0,1/2]$ such that for all $(u_1,u_2),(u_1',u_2')\in(0,1)^2$,

$$\left|\ddot{\ell}_{ii}(u_1,u_2)-\ddot{\ell}_{ii}(u_1',u_2')\right| \le M\left(\frac{|u_i-u_i'|}{r^3(u_i)r^{1/2-\epsilon}(u_{3-i})} + \frac{|u_{3-i}-u_{3-i}'|}{r^2(u_i)r(u_{3-i})}\right),$$

and,

$$\left|\ddot{\ell}_{12}(u_1,u_2)-\ddot{\ell}_{12}(u_1',u_2')\right| \le M\frac{|u_1-u_1'|+|u_2-u_2'|}{r(u_1)r(u_2)}.$$

Let us briefly discuss the assumptions. Assumptions (C1) and (C2) are standard in the semiparametric literature on copulas (the integrability condition on $\ddot{\ell}_{ij}$ can be relaxed; however this is beyond the scope of the thesis). Since for some copulas the $\dot{\ell}_i$'s and $\ddot{\ell}_{ij}$'s are not (extendable to be) bounded on $[0,1]^2$, we allow for explosive behavior on the boundary of $[0,1]^2$. The last part of Assumption (C2) and Assumption (C3) put restrictions on this boundary behavior. Since the copula is fixed, we denote, for notational convenience, $\mathbb{P}^C_{F,G}$ from now on by $\mathbb{P}_{F,G}$, and expectations with respect to $\mathbb{P}_{F,G}$ by $\mathbb{E}_{F,G}$. For $F,G\in\mathscr{F}_{\mathrm{ac}}$ the measure $\mathbb{P}_{F,G}$ has density (with respect to Lebesgue measure):

$$p_{F,G}(x,y) = c(F(x),G(y))f(x)g(y), \quad x,y\in\mathbb{R}.$$

The following invariance argument is heavily exploited (without notice) in the sequel. If $F, G \in \mathscr{F}_{\text{ac}}$ then, under $\mathbb{P}_{F,G}$, $(F(X), G(Y)) \sim C = \mathbb{P}_{\text{Un}[0,1],\text{Un}[0,1]}$. Consequently,

$$\mathbb{E}_{F,G} f(F(X), G(Y)) = \mathbb{E}_{\text{Un}[0,1],\text{Un}[0,1]} f(X, Y) = \int f(x, y)\, \mathrm{d}C(x, y).$$

We denote the marginal empirical distributions by F_n and G_n, and the bivariate empirical distribution function by H_n, for $x, y \in \mathbb{R}$,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le x\},\ G_n(y) = \frac{1}{n}\sum_{i=1}^{n} 1\{Y_i \le y\},$$

and,

$$H_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le x, Y_i \le y\}.$$

and we introduce scaled versions of F_n and G_n by

$$\tilde{F}_n(x) = \frac{n}{n+1} F_n(x), \quad x \in \mathbb{R}, \quad \tilde{G}_n(y) = \frac{n}{n+1} G_n(y), \quad y \in \mathbb{R}.$$

The (right-continuous with left-hand limits) empirical copula is defined by

$$C_n(u, v) = \frac{1}{n}\sum_{i=1}^{n} 1\{F_n(X_i) \le u, G_n(Y_i) \le v\}, \quad (u, v) \in [0,1]^2.$$

For $f : [0,1]^2 \to \mathbb{R}$, we have the identity

$$\int f(F_n(x), G_n(y))\, \mathrm{d}H_n(x, y) = \int_{[0,1]^2} f(u, v)\, \mathrm{d}C_n(x, y).$$

Since we want to allow for integrands that are explosive on the boundary of $[0,1]^2$, it is more convenient to use a 'shifted' version of C_n:

$$\tilde{C}_n(u, v) = \frac{1}{n}\sum_{i=1}^{n} 1\left\{\tilde{F}_n(X_i) \le u, \tilde{G}_n(Y_i) \le v\right\}, \quad u, v \in [0,1].$$

Now we have, for $f : (0,1)^2 \to \mathbb{R}$, the identity

$$\int f(\tilde{F}_n(x), \tilde{G}_n(y))\, \mathrm{d}H_n(x, y) = \int_{[0,1]^2} f(u, v)\, \mathrm{d}\tilde{C}_n(x, y).$$

Gaenssler and Stute (1987), Van der Vaart and Wellner (1993, Section 3.9.4.4), and Fermanian et al. (2004) considered weak convergence[1] of the process $(u, v) \mapsto$

[1] Since we may encounter objects that are not sufficiently measurable we rely throughout on the weak convergence theory à la Hoffman-Jørgensen (see, for example, Van der Vaart and Wellner (1993)).

$\sqrt{n}(C_n - C)(u, v)$. Recently, Biau and Wegkamp (2005) provided a maximal inequality for the empirical copula process indexed by certain sets. Our analysis in the following sections would be facilitated if there were also results available for the weak convergence of $\{\sqrt{n}\int f\, \mathrm{d}(\tilde{C}_n - C) \mid f \in \mathscr{F}\}$, the shifted empirical copula process indexed by a collection functions. Since results of this kind are not available we rely on Taylor expansions to invoke weak convergence of the standard empirical process.

Copula and second marginal known 5.3

In this section the model for one observation (X, Y) is

$$\mathscr{P}(C, G_0) = \left(\mathbb{P}_{F,G_0} \mid F \in \mathscr{F}_{\mathrm{ac}}\right),$$

where besides the copula C, satisfying the assumptions described below, the second marginal $G_0 \in \mathscr{F}_{\mathrm{ac}}$ is known. We study efficient estimation of the parameter F, seen as an element of the space $\ell^\infty(\mathbb{R})$, from an i.i.d. sample (X_i, Y_i) $i = 1, \ldots, n$.

Throughout Section 5.3 we impose on C, besides Assumption $(C1)$ the following additional asssumption:

> $c > 0$ on $[0,1]^2$ and for all $v \in [0,1]$ the mapping $u \mapsto \log c(u, v)$ is two times differentiable. Furthermore the derivatives, considered as functions on $[0,1]^2$, are continuous.

Hence $\dot{\ell}_1$ and $\ddot{\ell}_{11}$ (see (5.1)-(5.2)) are finite on $[0,1]^2$. As mentioned before, this assumption is too strong for many interesting copulas. However, we impose this assumption in this section to make the analysis more transparent. In the next section, for which this section provides intuition, we impose the assumptions (C1)-(C3) which allow for exploding $\dot{\ell}_i$'s and $\ddot{\ell}_{ij}$'s.

Information lower-bound & inefficiency of F_n 5.3.1

To be able to state the convolution theorem, which gives a lower-bound to the precision of regular estimators, we first recall the necessary notions from semiparametric theory. For details we refer to Bickel et al. (1998) and Van der Vaart (2000, Chapter 25). To obtain asymptotic bounds to the precision of estimators, it is well-known that one has to consider the local structure of the model, which is described by the tangent space. Fix $F_0 \in \mathscr{F}_{\mathrm{ac}}$. We describe how to construct a tangent space for the model $\mathscr{P}(C, G_0)$ at $\mathbb{P}_{F_0,G_0}$. Let $k : \mathbb{R} \to \mathbb{R}$ be defined by $k(z) = 2/(1 + \exp(-2z))$. Let $v \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$; here $\mathrm{L}_2^0(\mathrm{Un}[0,1])$ is the subset of $\mathrm{L}_2(\mathrm{Un}[0,1])$ for which $\int_0^1 a(u)\,\mathrm{d}u = 0$. Next define for $t \in (-1,1)$ the densities

$$f_t^v(x) = c_f^v(t)\, k\left(t v(F_0(x))\right) f_0(x), \quad x \in \mathbb{R}, \tag{5.3}$$

where c_f^v is such defined that f_t^v are indeed densities[2]. The densities f_t^v induce distribution functions $F_t^v \in \mathscr{F}_{\mathrm{ac}}$, and $t \mapsto F_t^v$ passes F_0 at $t = 0$. We introduce the 'score-operator' $\dot{\ell}_F : \mathrm{L}_2^0(\mathrm{Un}[0,1]) \to \mathrm{L}_2(\mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]})$ by (see Proposition 4.7.5 in Bickel et al. (1998)),

$$\dot{\ell}_F v(X,Y) = v(X) + \dot{\ell}_1(X,Y)\int_0^X v(z)\,\mathrm{d}z. \tag{5.4}$$

The following proposition yields a tangent space at $\mathbb{P}_{F_0,G_0}$.

Lemma 5.3.1. Let $v \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$. Then the path $t \mapsto F_t^v$, as defined by (5.3) has score $\dot{\ell}_F v(F_0(X), G_0(Y))$ at $t = 0$:

$$\lim_{t\to 0}\int_0^1\int_0^1\left(\frac{\sqrt{p_t(x,y)}-\sqrt{p_0(x,y)}}{t} - \frac{1}{2}\dot{\ell}_F v(F_0(x),G_0(y))\sqrt{p_0(x,y)}\right)^2 \mathrm{d}x\,\mathrm{d}y = 0,$$

where $p_t(x,y) = p_{F_t^v,G_0}(x,y)$. This yields the following tangent set for the model $\mathscr{P}(C,G_0)$ at $\mathbb{P}_{F_0,G_0}$,

$$\mathscr{T}(\mathbb{P}_{F_0,G_0} \mid \mathscr{P}(C,G_0)) = \left\{\dot{\ell}_F v(F_0(X),G_0(Y)) \mid v \in \mathrm{L}_2^0(\mathrm{Un}[0,1])\right\},$$

which is a closed linear subspace of $\mathrm{L}_2(\mathbb{P}_{F_0,G_0})$.

Proof.
The part on the score is essentially Proposition 4.7.4 in Bickel et al. (1998). The tangent space is closed, since the operator $\dot{\ell}_F : \mathrm{L}_2^0(\mathrm{Un}[0,1]) \to \mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}$ has closed range (see Bickel et al. (1998, Proposition 4.7.5). □

Next we recall the concept of a regular estimator. An estimator (not necessarily measurable) F_n^* of F is regular at $\mathbb{P}_{F_0,G_0}$ along the submodel $t \mapsto \mathbb{P}_{F_t^v,G_0}$ through $\mathbb{P}_{F_0,G_0}$ if there exists a tight Borel measurable element L_0 in $\ell^\infty(\mathbb{R})$ such that for all $u_n \to u \in \mathbb{R}$, we have

$$\sqrt{n}\left(F_n^* - F^v_{u_n/\sqrt{n}}\right) \xrightarrow{d} L_0, \text{ under } \mathbb{P}_{F^v_{u_n/\sqrt{n}},G_0} \text{ in } \ell^\infty(\mathbb{R}).$$

An estimator F_n^* of F is (semiparametrically) regular in the model $\mathscr{P}(C,G_0)$ at $\mathbb{P}_{F_0,G_0}$ if it is regular along all submodels $t \mapsto \mathbb{P}_{F_t^v,G_0}$, $v \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$, through $\mathbb{P}_{F_0,G_0}$. Finally, an estimator is regular for the model $\mathscr{P}(C,G_0)$, if it is regular at every $\mathbb{P}_{F,G_0}$, $F \in \mathscr{F}_{\mathrm{ac}}$. Our parameter of interest is described by the mapping (by Sklar's theorem this is indeed a mapping) $\nu : \mathscr{P}(C,G_0) \to \ell^\infty(\mathbb{R})$ defined by $\nu(\mathbb{P}_{F,G_0}) = ((F(x))_{x\in\mathbb{R}})$. Fix $F_0 \in \mathscr{F}_{\mathrm{ac}}$. We need the pathwise derivative of ν along the paths that generate the tangent space $\mathscr{T}(\mathbb{P}_{F_0,G_0} \mid \mathscr{P}(C,G_0))$. For a path $t \mapsto F_t^v$, defined by (5.3), it is an easy exercise to show $t^{-1}\left(\nu(\mathbb{P}_{F_t^v,G_0}) - \nu(\mathbb{P}_{F_0,G_0})\right) \to$

[2]It is trivial to check that we have $0 < k \le 2$, $0 < k' \le 4$, $0 < k'/k \le 2$ and $k(0) = k'(0) = 1$, $c_f^v(0) = 1$, $t \mapsto c_f^v(t)$ is continuously differentiable with $c_f^{v'}(0) = 0$.

$\nu'_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}_F v(F_0(X), G_0(Y)))$ in $\ell^\infty(\mathbb{R})$ as $t \to 0$. Here the pathwise derivative is given by $\nu'_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}_F v(F_0(X), G_0(Y)))(z) = \int_0^{F_0(z)} v(u)\,\mathrm{d}u$, $z \in \mathbb{R}$. Since this operator, seen as map from $\mathcal{T}(\mathbb{P}_{F_0,G_0} \mid \mathcal{P}(C, G_0))$ into $\ell^\infty(\mathbb{R})$, is continuous there exist, using the Riesz representation theorem for each coordinate, unique elements $\nu^*_{z,\mathbb{P}_{F_0,G_0}} \in \mathcal{T}(\mathbb{P}_{F_0,G_0} | \mathcal{P}(C, G_0))$ such that for all $v \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$:

$$\nu'_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}_F v(F_0(X), G_0(Y)))(z) = \mathbb{E}_{F_0,G_0} \nu^*_{z,\mathbb{P}_{F_0,G_0}}(X, Y)\dot{\ell}_F v(F_0(X), G_0(Y))$$

Let us now recall the convolution theorem (see, for example, Van der Vaart (1991b, Theorem 2.1) or Bickel et al. (1998, Theorem 5.2.1)).

Theorem 5.1. *Let $F_0 \in \mathcal{F}_{ac}$. If F_n^* is a regular estimator of F at $\mathbb{P}_{F_0,G_0}$ in the model $\mathcal{P}(C, G_0)$ with limit distribution W (under $\mathbb{P}_{F_0,G_0}$), then there exist tight Borel measurable elements L and N in $\ell^\infty(\mathbb{R})$ such that*

$$\mathcal{L}(W) = \mathcal{L}(L + N),$$

where L and N are independent and L is a mean 0 Gaussian process whose covariances are determined by the efficient influence operator $z \mapsto \nu^*_{z,\mathbb{P}_{F_0,G_0}}$.

Proof. Since the tangent space (Lemma 5.3.1) is linear and ν is pathwise differentiable all conditions of Bickel et al. (1998) Theorem 5.2.1 are met. □

Since $\mathcal{L}(L)$ is determined by the model only, via its efficient influence operator, it represents inevitable noise. Therefore, it is natural to call an estimator efficient at $\mathbb{P}_{F_0,G_0}$ if it is regular at $\mathbb{P}_{F_0,G_0}$ and if its limiting distribution (under $\mathbb{P}_{F_0,G_0}$) is given by L. An estimator of F is efficient (in the model $\mathcal{P}(C, G_0)$) if it is efficient at all $\mathbb{P}_{F_0,G_0}$.

Remark 1. It is easy to see that ν^*_{z,F_0,G_0} is the projection of $1\{X \le z\} - F_0(z)$ (the influence function of the empirical distribution function evaluated at z, i.e. $F_n(z)$) on $\mathcal{T}(\mathbb{P}_{F_0,G_0} \mid \mathcal{P}(C, G_0))$. It however seems to be impossible to obtain, in general, explicit expressions for $\nu^*_{z,\mathbb{P}_{F_0,G_0}}$, so it seems to be difficult to verify efficiency of a proposed estimator.

The next theorem shows that, amongst smooth absolutely continuous copulas, the independence copula is the only copula for which the empirical distribution function of $X_1, \ldots, X_n$ constitutes an efficient estimator of F.

Theorem 5.2. *Let C a copula satisfying the assumptions stated at the beginning of Section 5.3, and $F_0 \in \mathcal{F}_{ac}$. Let $z \in \mathbb{R}$ such that $1 > F_0(z) > 0$ Then $F_n(z)$ is an efficient estimator of $F(z)$ in the model $\mathcal{P}(C, G_0)$ at $\mathbb{P}_{F_0,G_0}$ if and only if $C(u, v) = uv$.*

Proof. Let $F_0 \in \mathscr{F}_{\mathrm{ac}}$.
'$\Longleftarrow$' It is easy to check that for all $z \in \mathbb{R}$ we have $\nu^*_{z,\mathbb{P}_{F_0,G_0}} = 1\{X \le z\} - F_0(z)$. Now efficiency directly follows from Bickel et al. (1998) Corollary 5.2.1. Of course, efficiency of F_n in the model where one has i.i.d. observations from $F \in \mathscr{F}_{\mathrm{ac}}$ unknown is well known; see, for example, Bickel et al. (1998) Example 5.3.1.
'$\Longrightarrow$' Since $F_n(z)$ is an efficient estimator of $F(z)$, the influence function of $F_n(z)$, $x \mapsto 1\{x \le z\} - F_0(z)$, belongs to the tangent space $\mathscr{T}(\mathbb{P}_{F_0,G_0}|\mathscr{P}(C,G_0))$, i.e. there exists $a^z \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$ such that

$$1\{X \le z\} - F_0(z) = \dot{\ell}_F a^z(F_0(X), G_0(Y)) \quad \text{a.s.} \tag{5.5}$$

Since $\mathbb{E}_{F_0,G_0}[\dot{\ell}_F a(F_0(X), G_0(Y)) \mid X] = a(F_0(X))$ for all $a \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$ we obtain

$$a^z(F_0(X)) = 1\{X \le z\} - F_0(z), \quad \text{a.s.}$$

Hence we must have, $\mathbb{P}_{F_0,G_0}$-a.s.,

$$0 = \dot{\ell}_1(F_0(X), G_0(Y)) \int_{w=0}^{F_0(X)} a^z(w)\,\mathrm{d}w = \dot{\ell}_1(F_0(X), G_0(Y))(F_0(X \wedge z) - F_0(z)F_0(X)),$$

which, by continuity of F_0 and G_0, implies $\dot{\ell}_1 = 0$, which yields $c_1 = 0$ and consequently the mappings $x \mapsto c(x,y)$ are constant for fixed y. Since $\int c(x,y)\,\mathrm{d}x = 1$ we conclude $c = 1$, which completes the proof. □

So F_n is an efficient estimator in case $C(u,v) = uv$, but in general inefficient. Thus the obvious question is whether it is possible to construct an efficient estimator for general copulas.

5.3.2 The estimator

This section introduces our estimator. We start by considering a kind of empirical likelihood method (see, for example, Owen (2001)), and after several simplifying approximations we will arrive at our estimator. This proposed estimator is computationally attractive: the only computational difficulty is to determine a solution to a linear system of $n+1$ equations in $n+1$ variables. We prove uniform consistency of this estimator.

Since the (unknown) joint distribution H has copula C and G_0 as second marginal, it seems natural to impose that the estimated joint distribution has copula C and G_0 as second marginal as well. Inspired by the literature on empirical likelihood this leads to the idea to consider estimators F_n^* that maximize the 'empirical-copula-likelihood'

$$\mathscr{E}_n(F, G_0) = \frac{1}{n}\sum_{i=1}^n \log \mathbb{P}_{F,G_0}\{(X_i, Y_i)\},$$

over $F \in \mathscr{F}$. For the moment, ignore the existence of a maximum. The next proposition shows that we only have to consider estimators that concentrate on the data.

Proposition 5.3.1. If F_n^* maximizes $\mathscr{E}_n(\cdot, G_0)$ over $\mathscr{F}$, then F_n^* is concentrated on the data, i.e. F_n^* assigns mass 1 to the set $\{X_1, \ldots, X_n\}$.

Proof.
Let F_n^1 be a maximizer of $\mathscr{E}_n(\cdot, G_0)$. Define the distribution function F_n^*, concentrated on $\{X_1, \ldots, X_n\}$, by, for $i = 1, \ldots, n$,

$$F_n^*(X_{i:n}) = \begin{cases} F_n^1(X_{i:n}) & \text{if } i < n, \\ 1 & \text{if } i = n. \end{cases}$$

Now the proposition is certainly proved, once we show that, for all $i, j = 1, \ldots, n$, we have $\mathbb{P}_{F_n^*, G_0}\{(X_{i:n}, Y_{j:n})\} \geq \mathbb{P}_{F_n^1, G_0}\{(X_{i:n}, Y_{j:n})\}$, with equality for all i, j only if F_n^1 is already concentrated on the data. For convenience, denote $X_{0:n} = Y_{0:n} = -\infty$. Then, for $1 \leq i, j < n$, we have

$$\begin{aligned}
\mathbb{P}_{F_n^*, G_0}\{(X_{i:n}, Y_{j:n})\} &= C\left(F_n^*(X_{i:n}), G_0(Y_{j:n})\right) - C\left(F_n^*(X_{i-1:n}), G_0(Y_{j:n})\right) \\
&\quad - C\left(F_n^*(X_{i:n}), G_0(Y_{j-1:n})\right) + C\left(F_n^*(X_{i-1:n}), G_0(Y_{j-1:n})\right) \\
&= C\left(F_n^1(X_{i:n}), G_0(Y_{j:n})\right) - C\left(F_n^1(X_{i-1:n}), G_0(Y_{j:n})\right) \\
&\quad - C\left(F_n^1(X_{i:n}), G_0(Y_{j-1:n})\right) + C\left(F_n^1(X_{i-1:n}), G_0(Y_{j-1:n})\right) \\
&= \mathbb{P}_{F_n^1, G_0}\left((X_{i-1:n}, X_{i:n}] \times (Y_{j-1:n}, Y_{j:n}]\right) \\
&\geq \mathbb{P}_{F_n^1, G_0}\{(X_{i:n}, Y_{j:n})\},
\end{aligned}$$

where the last inequality is an equality if and only if the probability measure $\mathbb{P}_{F_n^1, G_0}$ does not assign mass to the set $\left((X_{i-1:n}, X_{i:n}] \times (Y_{j-1:n}, Y_{j:n}]\right) \setminus \{(X_{i:n}, Y_{j:n})\}$. Similarly, we find for $i = n$ and $j < n$,

$$\begin{aligned}
\mathbb{P}_{F_n^*, G_0}\{(X_{n:n}, Y_{j:n})\} &= \mathbb{P}_{F_n^1, G_0}\left((X_{n-1:n}, \infty) \times (Y_{j-1:n}, Y_{j:n}]\right) \\
&\geq \mathbb{P}_{F_n^1, G_0}\{(X_{n,n}, Y_{j:n})\},
\end{aligned}$$

where the last inequality is an equality if and only if the probability measure $\mathbb{P}_{F_n^1, G_0}$ does not assign mass to the set $\left((X_{n-1:n}, \infty) \times (Y_{j-1:n}, Y_{j:n}]\right) \setminus \{(X_{n:n}, Y_{j:n})\}$. Analogous inequalities hold for the cases $i < n$, $j = n$, and $i = j = n$. □

Therefore we restrict ourselves to estimators of the form

$$F_n^*(x) = \sum_{i=1}^{n} p_i^{(n)} 1\{X_i \leq x\}, \quad x \in \mathbb{R},$$

i.e. $p_i^{(n)}$ is the mass that F_n^* assigns to the point $\{X_i\}$ (if there are no ties in the data (which happens with probability 1)). Now we would like to maximize

$\mathscr{E}_n(\cdot, G_0)$ as a function of $p_1^{(n)}, \dots, p_n^{(n)}$ under the nonnegativity constraints $p_i^{(n)} \geq 0$ $(i = 1, \dots, n)$, and the equality constraint $p_1^{(n)} + \cdots + p_n^{(n)} = 1$. This is a highly nonlinear constrained optimization problem in n variables; to reduce the computational complexity we make several approximations. Recall that $p_{F,G}(x, y) = c(F(x), G(y)) f(x) g(y)$ for $F, G \in \mathscr{F}_{\text{ac}}$. As a first simplification, we replace $f(X_i)$ by $p_i^{(n)}$, which motivates the following approximation to $\mathscr{E}_n(F_n^*, G_0)$ (up to a constant):

$$\tilde{\mathscr{E}}_n^{G_0}(F_n^*) = \frac{1}{n} \sum_{i=1}^n \log p_i^{(n)} + \frac{1}{n} \sum_{i=1}^n \log c\left(F_n^*(X_i), G_0(Y_i)\right).$$

Although this approximation already simplifies life, we are still dealing with a highly nonlinear constrained optimization problem in n variables. We approximate $\log p_i^{(n)} = \log(1 + (np_i^{(n)} - 1)) - \log(n)$ by $(np_i^{(n)} - 1) - \frac{1}{2}(np_i^{(n)} - 1)^2 - \log n$. The motivation for this approximation is that we think of our estimators as being 'close to' the empirical distribution function. Inspired by a Taylor expansion, the equality $\sum_{i=1}^n (np_i^{(n)} - 1) = 0$, and motivated by the equality $\mathbb{E}_{F_0,G_0}[\dot{\ell}_1^2(F_0(X), G_0(Y)) \mid X] = -\mathbb{E}_{F_0,G_0}[\ddot{\ell}_{11}(F_0(X), G_0(Y)) \mid X]$, we take as our objective function

$$\begin{aligned}
\mathscr{L}_n^{G_0}(F_n^*) = \frac{1}{n} \sum_{i=1}^n \dot{\ell}_1(i) \left(F_n^*(X_i) - F_n(X_i)\right) - \frac{1}{2n} \sum_{i=1}^n \dot{\ell}_1^2(i) \left(F_n^*(X_i) - F_n(X_i)\right)^2 \\
- \frac{1}{2n} \sum_{i=1}^n \left(nF_n^*\{X_i\} - 1\right)^2,
\end{aligned}$$

where we use the abbreviations

$$\dot{\ell}_1(i) = \dot{\ell}_1\left(F_n(X_i), G_0(Y_i)\right), \quad \ddot{\ell}_{11}(i) = \ddot{\ell}_{11}\left(F_n(X_i), G_0(Y_i)\right), \quad i = 1, \dots, n.$$

Next, we consider the constrained quadratic optimization problem

$$\begin{aligned}
&\max_{F_n^*} \quad \mathscr{L}_n^{G_0}(F_n^*), \\
&\text{s.t.} \quad F_n^* \text{ probability distribution concentrated on } \{X_1, \dots, X_n\}. \qquad (5.6)
\end{aligned}$$

Since we have to maximize a continuous function (of n variables) on a compact set a maximum indeed exists. We propose to estimate F by *a* global maximum $\hat{F}_n$ of (5.6). Notice that for the independence copula, i.e. $C(u, v) = uv$, we find $\hat{F}_n = F_n$. In general it seems impossible to obtain 'explicit' expressions for $\hat{F}_n$. Although it is possible to determine a solution to (5.6) by numerical routines for constrained quadratic optimizations, we will show that a solution can be found by determining a solution (which is, with probability tending to 1, unique) to a linear system of $n + 1$ equations in $n + 1$ variables. Besides yielding a computationally attractive estimator, we will use these Lagrange equations in the next section to obtain the limit distribution of $\hat{F}_n$. The next propositions, which give a consistency result for $\hat{F}_n$, will allow us to get rid of the inequality constraints in (5.6). This will allow us to show that $\hat{F}_n$ is indeed a solution to a linear system of equations.

Proposition 5.3.2. Let $F_0 \in \mathscr{F}_{\text{ac}}$. Let, for $n \in \mathbb{N}$, $\hat{F}_n$ a global maximum location of (5.6). Then we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \left(n\hat{F}_n\{X_i\} - 1\right)^2 = 0, \quad \mathbb{P}_{F_0,G_0}\text{-a.s.}$$

Proof.
Notice that $\mathscr{L}_n^{G_0}(F_n) = 0$. Since $\hat{F}_n$ maximizes, by definition, $\mathscr{L}_n^{G_0}$ we thus have $\mathscr{L}_n^{G_0}(\hat{F}_n) \geq 0$ which yields

$$\begin{aligned}\frac{1}{2n}\sum_{i=1}^{n} \left(n\hat{F}_n\{X_i\} - 1\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_1(i)\left(\hat{F}_n(X_i) - F_n(X_i)\right) \\ - \frac{1}{2n}\sum_{i=1}^{n} \dot{\ell}_1^2(i)\left(\hat{F}_n(X_i) - F_n(X_i)\right)^2.\end{aligned}$$

Since $\ddot{\ell}_{11}$ is bounded, say by $C > 0$, we obtain

$$\begin{aligned}\sup_{F,G\in\mathscr{F}} &\left|\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_1(i)\,(F(X_i) - G(X_i)) - \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_1(F_0(X_i), G_0(Y_i))\,(F(X_i) - G(X_i))\right| \\ &\leq C\|F_n - F_0\|_\infty \to 0, \quad \mathbb{P}_{F_0,G_0}\text{-a.s.}\end{aligned}$$

Since $\mathbb{E}_{F_0,G_0}[\dot{\ell}_1(F_0(X), G_0(Y))|X] = 0$ it follows that $\mathbb{E}_{F_0,G_0}\dot{\ell}_1(F_0(X), G_0(Y))(F(X) - G(X)) = 0$ for all $F, G \in \mathscr{F}$. It is well-known that the class of monotone functions from $\mathbb{R}$ into $[0,1]$ has for all $\epsilon > 0$ a finite $L_1(\mathbb{Q})$-ϵ-bracketing number for all probability measures $\mathbb{Q}$ on the real line. From this it easily follows that the class of functions $\{(x, y) \mapsto \dot{\ell}_1(F_0(x), G_0(y))(F(x) - G(y)) \mid F, G \in \mathscr{F}\}$ has for all $\epsilon > 0$ a finite $L_1(\mathbb{P}_{F_0,G_0})$-$\epsilon$-bracketing number. A combination of the previous display with the Glivenko-Cantelli theorem thus yields

$$\sup_{F,G\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_1(i)\,(F(X_i) - G(X_i))\right| \to 0, \quad \mathbb{P}_{F_0,G_0}\text{-a.s.}$$

Hence we obtain

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{2n}\sum_{i=1}^{n} \left(n\hat{F}_n\{X_i\} - 1\right)^2 \\ &\leq 0 + \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_1(i)\left(\hat{F}_n(X_i) - F_n(X_i)\right) = 0, \quad \mathbb{P}_{F_0,G_0}\text{-a.s.}\end{aligned}$$

□

Proposition 5.3.2 immediately yields uniform consistency of $\hat{F}_n$. The following corollary gives the precise statement.

Corollary 5.3. *Let the setting be the same as in the previous proposition. Then we have*

$$\lim_{n\to\infty} \|\hat{F}_n - F_0\|_\infty = 0, \quad \mathbb{P}_{F_0,G_0}\text{-a.s.}$$

Proof.
This immediately follows from the previous proposition and the classic Glivenko-Cantelli theorem:

$$\begin{aligned}\|\hat{F}_n - F_0\|_\infty &\le \|\hat{F}_n - F_n\|_\infty + \|F_n - F_0\|_\infty \le \frac{1}{n}\sum_{i=1}^n |n\hat{F}_n\{X_i\} - 1| + \|F_n - F_0\|_\infty \\ &\le \sqrt{\frac{1}{n}\sum_{i=1}^n (n\hat{F}_n\{X_i\} - 1)^2} + \|F_n - F_0\|_\infty \to 0, \quad \mathbb{P}_{F_0,G_0}\text{-a.s.}\end{aligned}$$

□

The following proposition shows that the inequality constraints in the first order conditions are indeed not binding.

Proposition 5.3.3. Let the setting be the same as in the previous proposition. Let $F_0 \in \mathscr{F}_{\text{ac}}$. Then the probability $\mathbb{P}_{F_0,G_0}(\mathscr{A}_n)$, with $\mathscr{A}_n = \{\min_{i=1,\ldots,n} \hat{F}_n\{X_i\} > 0\}$, converges to 1.

Proof. In the following we always work on the event that there are no ties, which has probability 1. Let $p_{(i)}^{(n)}$ denotes the mass on $\{X_{i:n}\}$, $i = 1,\ldots,n$. Fix $i \in \{1,\ldots,n-1\}$, and define pointmasses on $\{X_{1:n},\ldots,X_{n:n}\}$ by $p_{(j)}^t = p_{(j)}^{(n)}$ for $j \notin \{i, i+1\}$, $p_{(i)}^t = p_{(i)}^{(n)} + t$, and $p_{(i+1)}^t = p_{(i+1)}^{(n)} - t$. Consider the following three cases. Case (i): if $p_{(i)}^{(n)} \wedge p_{(i+1)}^{(n)} > 0$ this defines a probability measure for $|t| < \eta$ for some $\eta > 0$; Case (ii): if $p_{(i)}^{(n)} = 0$, $p_{(i+1)}^{(n)} > 0$ it defines a probability measure for $0 \le t \le \eta$ for some $\eta > 0$; Case (iii): if $p_{(i)}^{(n)} > 0$, $p_{(i+1)}^{(n)} = 0$ it defines a probability measure for $-\eta \le t \le 0$ for some $\eta > 0$. Note that the resulting distribution function satisfies $F_t(X_{j:n}) = \hat{F}_n(X_{j:n})$ for $j \neq i$, and $F_t(X_{i:n}) = \hat{F}_n(X_{i:n}) + t$. In Case (i) we have $(\partial/\partial t)\mathscr{L}_n^{G_0}(F_t)\,|_{t=0} = 0$, in Case (ii) $(\partial/\partial t)\mathscr{L}_n^{G_0}(F_t)\,|_{t=0} \le 0$, and in Case (iii) $(\partial/\partial t)\mathscr{L}_n^{G_0}(F_t)\,|_{t=0} \ge 0$. So in Case (i) we obtain

$$0 = \frac{1}{n}\left\{\dot{\ell}_1^*(i) - \dot{\ell}_1^{*2}(i)(\hat{F}_n(X_{i:n}) - F_n(X_{i:n}))\right\} + np_{(i+1)}^{(n)} - np_{(i)}^{(n)},$$

where $R_{1:n}^{-1}$ denotes the inverse permutation of the ranks of $\{X_1,\ldots,X_n\}$, i.e. $R_{1:n}^{-1}(i) = k$ if and only if $R_k^X = i$, and

$$\dot{\ell}_1^*(i) = \dot{\ell}_1\left(F_n(X_{i:n}), G_0(Y_{R_{1:n}^{-1}(i)})\right), \quad i = 1,\ldots,n,$$

are the 'X-ranked versions' of $\dot{\ell}_1$. Hence we obtain the bound

$$|np^{(n)}_{(i+1)} - np^{(n)}_{(i)}| \leq \frac{|\dot{\ell}^*_1(i)| + |\dot{\ell}^{*2}_1(i)|}{n}.$$

It is easy to check that this inequality also holds for Case (ii) and Case (iii), and of course also if $p^{(n)}_{(i)} = p^{(n)}_{(i+1)} = 0$. Next note that

$$\frac{1}{n}\sum_{r=0}^{s-1}\left\{|\dot{\ell}^*_1(i+r)| + |\dot{\ell}^{*2}_1(i+r)|\right\}$$
$$= \frac{1}{n}\sum_{j=1}^{n}(|\dot{\ell}_1(j)|(1+|\dot{\ell}_1(j)|)1\left\{\frac{i}{n} \leq F_n(X_j) \leq \frac{i+s-1}{n}\right\}.$$

Using the Glivenko-Cantelli theorem and some basic arguments, it follows that for every $\epsilon > 0$ there exists $1 > \delta > 0$ such that

$$\limsup_{n\to\infty} \max_{\substack{i=1,\dots,n \\ s=1,\dots,(n-i)\wedge\lceil n\delta\rceil}} \left|np^{(n)}_{(i+s)} - np^{(n)}_{(i)}\right| \leq \epsilon \quad \mathbb{P}_{F_0,G_0}\text{-a.s.} \tag{5.7}$$

Let

$$\mathscr{B}_n = \left\{ \max_{\substack{i=1,\dots,n \\ s=1,\dots,(n-i)\wedge\lceil n\delta\rceil}} \left|np^{(n)}_{(i+s)} - np^{(n)}_{(i)}\right| \leq \epsilon \right\}.$$

On the event $\mathscr{A}^c_n$ there exists $1 \leq i \leq n$ with $p^{(n)}_i = 0$. Then there are, on the event $\mathscr{B}_n$, $\lceil n\delta\rceil$ scaled probabilities $np^{(n)}_{i+s}$ less than or equal to ϵ. But then also

$$\frac{1}{n}\sum_{i=1}^{n}(n\hat{F}_n\{X_i\} - 1)^2 \geq \delta(1-\epsilon)^2.$$

Using Proposition 5.3.2 it now easily follows that indeed $\mathbb{P}_{F_0,G_0}(\mathscr{A}_n) \to 1$. □

For our asymptotic analysis it is more convenient to work with the relative deviations of the pointmasses $\hat{F}_n\{X_i\}$ from $1/n$. Introduce

$$a^{(n)}_i = n\hat{F}_n\{X_i\} - 1, \quad i = 1,\dots,n.$$

Now the optimization problem (5.6) can be reformulated in terms of $a^{(n)} = (a^{(n)}_1,\dots,a^{(n)}_n)'$. To this end we introduce the $n \times n$ matrices

$$A = \left(1\{X_j \geq X_i\}\right)_{i,j=1,\dots,n}, \text{ and } M_{11} = \begin{pmatrix} -\dot{\ell}^2_1(1) & & 0 \\ & \ddots & \\ 0 & & -\dot{\ell}^2_1(n) \end{pmatrix},$$

and the n-vector $M_1 = (\dot\ell_1(1),\dots,\dot\ell_1(n))'$. Then (5.6) is equivalent to the optimization problem

$$\begin{aligned}\max_{\tilde a^{(n)}\in\mathbb{R}^n}\quad & \mathscr{L}_n^{G_0}(\tilde a^{(n)}) = \frac{1}{2}\tilde a^{(n)'}\left[-I_n + \frac{1}{n^2}AM_{11}A'\right]\tilde a^{(n)} + \frac{1}{n}(AM_1)'\tilde a^{(n)}\\ \text{s.t.}\quad & \tilde a_1^{(n)} + \cdots + \tilde a_n^{(n)} = 0,\\ & \tilde a_i^{(n)} \ge -1 \quad (i = 1,\dots,n),\end{aligned} \tag{5.8}$$

where I_n denotes the $n \times n$ identity matrix. It trivially follows that $\mathscr{L}_n^{G_0}$ is concave, so $a^{(n)}$ is the unique solution to (5.8) (and hence (5.6) has unique optimizer $\hat F_n$). Furthermore, the corresponding Kuhn-Tucker conditions are necessary and sufficient. By Proposition 5.3.3, the probability of the event $\mathscr{A}_n = \{\min_{i=1,\dots,n} a_i^{(n)} > -1\}$ tends to 1 as $n \to \infty$. On the event $\mathscr{A}_n$ the inequality constraints in (5.8) are not binding. Because the target function is concave the following set of Lagrange equations has, on the event $\mathscr{A}_n$, $(a^{(n)}, \kappa_n^*)$ as unique solution.

$$\begin{cases}\dfrac{\partial \mathscr{L}_n^{G_0}}{\partial \tilde a_k^{(n)}} - \kappa = 0, \quad (k = 1,\dots,n),\\ \tilde a_1^{(n)} + \cdots + \tilde a_n^{(n)} = 0.\end{cases} \tag{5.9}$$

Here κ is the Lagrange multiplier corresponding to the constraint $\tilde a_1^{(n)} + \cdots + \tilde a_n^{(n)} = 0$. The relevant partial derivatives are given by (if there are no ties), for $k = 1,\dots,n$,

$$\begin{aligned} n\frac{\partial}{\partial \tilde a_k^{(n)}}\mathscr{L}_n^{G_0} = {} & \frac{1}{n}\sum_{i=1}^n \dot\ell_1(i)1\{X_i \ge X_k\}\\ & - \frac{1}{n^2}\sum_{\ell=1}^n \tilde a_\ell^{(n)} \sum_{i=1}^n \dot\ell_1^2(i)1\{X_i \ge X_k\}1\{X_i \ge X_\ell\} - \tilde a_k^{(n)},\end{aligned} \tag{5.10}$$

The system of Lagrange equations can be rewritten as the following linear system of $n+1$ equations in $n+1$ variables

$$\begin{pmatrix}\frac{1}{n^2}AM_{11}A' - I_n & -e_n\\ e_n' & 0\end{pmatrix}\begin{pmatrix}\tilde a^{(n)}\\ \kappa\end{pmatrix} = \begin{pmatrix}-\frac{1}{n}AM_1\\ 0\end{pmatrix}, \tag{5.11}$$

where e_n is a n-vector of ones.

So with probability converging to 1 the linear system (5.11) has $(a^{(n)}, \kappa_n^*)$ as unique solution.

5.3.3 The limit distribution

In this section we derive the limiting distribution of $(\hat F_n)_{n\in\mathbb{N}}$. First we reformulate the Lagrange equations of the previous section in operator notation. Next we show that a limiting version of this operator is continuously invertible. Using this result the limiting distribution of $\hat{F}_n$ is obtained.

Recall from the previous section that $a_i^{(n)} = n\hat{F}_n\{X_i\} - 1$ for $i = 1,\dots,n$. Introduce the function $a^{(n)} : [0,1] \to [-1, n-1]$ as follows: $a^{(n)}(0) = 0$, and

$$a^{(n)}(u) = a_i^{(n)} \text{ for } u \in \left(F_n(X_i) - \frac{1}{n}, F_n(X_i)\right], \quad i = 1,\dots,n.$$

So we have the relation $n\hat{F}_n\{X_i\} = 1 + a^{(n)}(F_n(X_i))$ a.s. Introduce $A_n : [0,1] \to [-1,1]$ by

$$A_n(u) = \int_0^u a^{(n)}(z)\,\mathrm{d}z, \quad u \in [0,1],$$

and note that we have

$$A_n(F_n(x)) = \hat{F}_n(x) - F_n(x), \quad x \in \mathbb{R}, \quad \text{a.s..} \tag{5.12}$$

Using that $\sum_{k=1}^n a_k^{(n)} = 0$ we solve for the Lagrange multiplier κ in (5.9), and substitute this in the Lagrange equations for $a_k^{(n)}$, $k = 1,\dots,n$. In this way we see that, with probability tending to 1, the following equations hold, for $k = 1,\dots,n$,

$$\begin{aligned} 0 = \frac{1}{n}\sum_{i=1}^n \dot{\ell}_1(i)\,(1\{X_i \geq X_k\} - F_n(X_i)) \\ - \frac{1}{n}\sum_{i=1}^n \dot{\ell}_1^2(i)\left(\hat{F}_n(X_i) - F_n(X_i)\right)(1\{X_i \geq X_k\} - F_n(X_i)) - a_k^{(n)}. \end{aligned}$$

This implies that almost surely the following equation holds, for all $v \in (0,1]$,

$$\begin{aligned} -a^{(n)}(v) - \frac{1}{n}\sum_{i=1}^n \dot{\ell}_1^2(F_n(X_i), G_0(Y_i))A_n(F_n(X_i))\,(1\{F_n(X_i) \geq v\} - F_n(X_i)) \\ = -\frac{1}{n}\sum_{i=1}^n \dot{\ell}_1(F_n(X_i), G_0(Y_i))\,(1\{F_n(X_i) \geq v\} - F_n(X_i)). \end{aligned}$$

Integrating v over $[0,u]$ yields almost surely, for all $u \in [0,1]$,

$$\begin{aligned} -A_n(u) - \frac{1}{n}\sum_{i=1}^n (F_n(X_i) \wedge u - F_n(X_i)u)A_n(F_n(X_i))\dot{\ell}_1^2(F_n(X_i), G_0(Y_i)) \\ = -\frac{1}{n}\sum_{i=1}^n (F_n(X_i) \wedge u - F_n(X_i)u)\dot{\ell}_1(F_n(X_i), G_0(Y_i)). \end{aligned} \tag{5.13}$$

To facilitate our asymptotic analysis we introduce the operator[3] $\dot{\Psi}_n : \mathrm{D}[0,1] \to \mathrm{D}[0,1]$ by

$$\dot{\Psi}_n(h)(u) = -h(u) - \frac{1}{n}\sum_{i=1}^n (F_n(X_i) \wedge u - F_n(X_i)u)h(F_n(X_i))\dot{\ell}_1^2(F_n(X_i), G_0(Y_i)),$$

[3] As usual, $\mathrm{D}[0,1]$ denotes the space of real-valued cadlag functions on $[0,1]$, equipped with the supremum norm.

and we introduce the operator $\dot{\Psi}_C : \mathrm{D}[0,1] \to \mathrm{D}[0,1]$, which can be interpreted as the 'limit-version' of $\dot{\Psi}_n$, by

$$\dot{\Psi}_C(h)(u) = -h(u) - \int (x \wedge u - xu)\, h(x) \dot{\ell}_1^2(x,y)\, \mathrm{d}C(x,y). \tag{5.14}$$

Note that the operator $\dot{\Psi}_C$ only depends on the copula, and not on the marginals.

Lemma 5.3.2. The operator $\dot{\Psi}_C : \mathrm{D}[0,1] \to \mathrm{D}[0,1]$ is onto and one-to-one, and the inverse $\dot{\Psi}_C^{-1}$ is continuous.

Proof.
In this proof expectations are always taken with respect to $\mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}$. For notational convenience we will drop subscripts if no confusion is possible. For clarity we break the proof into two propositions. First, we prove that $\dot{\Psi}$ is invertible on its range. Next, we prove that $\dot{\Psi}$ is onto and that the inverse is continuous.

Proposition 5.3.4. $\dot{\Psi}_C$ is one-to-one.

Proof.
Since $\dot{\Psi}$ is linear, we have to show that the null space of $\dot{\Psi}$ is trivial. So let $h \in \mathrm{D}[0,1]$ be such that $\dot{\Psi} h = 0$, or written more explicitly,

$$0 = -h(u) - \mathbb{E}\,(u \wedge X - uX)\, \dot{\ell}_1^2(X,Y) h(X), \qquad \forall u \in [0,1]. \tag{5.15}$$

Plugging in $u = 0$, and $u = 1$ we see that necessarily

$$h(0) = h(1) = 0. \tag{5.16}$$

Notice that,

$$\begin{aligned} P(u) = -\mathbb{E}(u \wedge X - uX)\dot{\ell}_1^2(X,Y)h(X) &= u\mathbb{E}X\dot{\ell}_1^2(X,Y)h(X) \\ &- \int_{x=0}^{u}\int_{y=0}^{1} x\dot{\ell}_1^2(x,y)h(x)c(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &- u\int_{x=u}^{1}\int_{y=0}^{1} \dot{\ell}_1^2(x,y)h(x)c(x,y)\,\mathrm{d}y\,\mathrm{d}x, \end{aligned}$$

is two times continuously differentiable with second derivative given by $P''(u) = I_{11}(u)h(u)$. So from (5.15) it follows that h is two times continuously differentiable. Using this and the boundary conditions (5.16) we see that any solution h to (5.15) is a solution to the following Sturm-Liouville (see, for example, Tricomi (1985)) differential equation

$$\begin{cases} -h''(u) + h(u)I_{11}(u) = 0, \\ h(0) = h(1) = 0. \end{cases}$$

It is easy to see that this differential equation has unique solution $h = 0$. Multiply the differential equation by h, use partial integration and the boundary conditions to arrive at $\int_0^1 (h'(u))^2 \,\mathrm{d}u + \int_0^1 h^2(u) I_{11}(u) \,\mathrm{d}u = 0$. Since $I_{11} \geq 0$ this yields $h' = 0$, hence the boundary conditions yield $h = 0$. Thus we conclude that $\dot{\Psi}$ is one-to-one. □

The next proposition will conclude the proof of the lemma.

Proposition 5.3.5. $\dot{\Psi}_C$ is onto and the inverse $\dot{\Psi}_C^{-1} : \mathrm{D}[0,1] \to \mathrm{D}[0,1]$ is continuous.

Proof.
The proof uses the following result from the Fredholm theory of linear operators (see, for example, Van der Vaart (2000, Lemma 25.93) or Rudin (1973, pp. 99-103)).

Lemma 5.3.3. Let $\mathbb{B}$ a Banach space. Let $J : \mathbb{B} \to \mathbb{B}$ continuous, onto and continuously invertible, and $K : \mathbb{B} \to \mathbb{B}$ a compact operator. Then $J + K$ is onto and continuously invertible if $J + K$ is one-to-one.

We can write $\dot{\Psi} = -I + K$, where I is the identity, i.e. $Ih = h$, and where $K : \mathrm{D}[0,1] \to \mathrm{D}[0,1]$ is defined by

$$K(h)(u) = -\mathbb{E}(u \wedge X - uX)\dot{\ell}_1^2(X,Y)h(X), \quad u \in [0,1].$$

So, using the previous proposition, the proof of the present proposition is complete, once we show that K is a compact operator. Note that the range of K is a subset of[4] $\mathrm{C}[0,1]$. Remember that (Compactness criterion (see, for example, Kreyszig (1978))) K is compact if and only if the following holds: for any sequence $(h_n)_{n\in\mathbb{N}}$, in $\mathrm{D}[0,1]$ for which $\|h_n\|_\infty$ is bounded, the sequence Kh_n, in $\mathrm{C}[0,1]$, has a convergent subsequence. So let h_n, $n \in \mathbb{N}$, a sequence in $\mathrm{D}[0,1]$ with $\sup_n \|h_n\|_\infty \leq C$. We have to show that the sequence Kh_n in $\mathrm{C}[0,1]$ has a convergent subsequence. Let us remember (a basic version of) the Arzelà-Ascoli theorem (see, for example, Kreyszig (1978))).

Lemma 5.3.4. A bounded equicontinuous sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathrm{C}[0,1]$ has a subsequence which converges[5].

Thus the proof is complete if we show that the sequence Kh_n is bounded and equicontinuous. Boundedness is immediate since $\sup_n \|h_n\|_\infty \leq C$. And equicontinuity immediately follows from the estimate $\left|K(h_n)(u) - K(h_n)(u')\right| \leq C|u - u'|\mathbb{E}(1+X)\dot{\ell}_1^2(X,Y)$. □

□

[4]As usual, $\mathrm{C}[0,1]$ denotes the space of continuous functions on $[0,1]$ equipped with the supremum norm.

[5]Recall: a sequence x_n in $\mathrm{C}[0,1]$ is equicontinuous if for all $\epsilon > 0$ there is $\delta > 0$ such that for all n, for all $u, u' \in [0,1]$ with $|u - u'| < \delta$ we have $|x_n(u) - x_n(u')| < \epsilon$.

In general it seems to be impossible to obtain 'explicit' expressions for $\dot\Psi_C^{-1}$. Next we derive the limiting distribution of $\hat F_n$.

Theorem 5.4. *Let $F_0 \in \mathscr{F}_{ac}$, and let, for $n \in \mathbb{N}$, $\hat F_n$ a maximum of (5.6). Then, under $\mathbb{P}_{F_0,G_0}$, the following holds.*
The process $\mathbb{S}_n^{F_0,G_0} = \left(\mathbb{S}_n^{F_0,G_0}(u)\right)_{u\in[0,1]}$, with

$$\mathbb{S}_n^{F_0,G_0}(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{1\{F_0(X_i) \le u\} - u + \dot\ell_1(F_0(X_i), G_0(Y_i))(F_0(X_i)\wedge u - F_0(X_i)u)\},$$

weakly converges in $\ell^\infty([0,1])$ to a tight zero-mean Gaussian process, denoted by $\mathbb{S}^C = (\mathbb{S}^C(u))_{u\in[0,1]}$, whose covariance function only depends on C.
And we have, in $\ell^\infty(\mathbb{R})$,

$$\begin{aligned}\left(\sqrt{n}\left(\hat F_n(x) - F_0(x)\right)\right)_{x\in\mathbb{R}} &= -\left((\dot\Psi_C^{-1}\mathbb{S}_n^{F_0,G_0})\circ F_0(x)\right)_{x\in\mathbb{R}} + o(1;\mathbb{P}_{F_0,G_0})\\ &\rightsquigarrow -\left((\dot\Psi_C^{-1}\mathbb{S}^C)\circ F_0(x)\right)_{x\in\mathbb{R}}.\end{aligned}$$

Remark 2. Notice that the coordinates of $\mathbb{S}_n^{F_0,G_0}$ are build of elements of the tangent space: taking $v(\cdot) = 1\{\cdot \le u\} - u$ yields $n^{-1/2}\sum_{i=1}^n \dot\ell_F v(F_0(X_i), G_0(Y_i)) = \mathbb{S}_n^{F_0,G_0}(u)$. We will exploit this in the next section to prove efficiency of $\hat F_n$.

Proof.
We split the proof into two parts. In Part A we prove weak convergence of $\mathbb{S}_n^{F_0,G_0}$, and in Part B we prove weak convergence of $\sqrt{n}(\hat F_n - F_0)$.

Part A
This is an easy exercise. The quantile transformation yields $\mathscr{L}(\mathbb{S}_n^{F_0,G_0} \mid \mathbb{P}_{F_0,G_0}) = \mathscr{L}(\mathbb{S}_n^{\mathrm{Un}[0,1],\mathrm{Un}[0,1]} \mid \mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]})$. So it suffices to consider uniform marginals. We introduce the classes $\mathscr{A}_1 = \{(x,y) \mapsto 1\{x \le u\} - u \mid u \in [0,1]\}$, and $\mathscr{A}_2 = \{(x,y) \mapsto \dot\ell_1(x,y)(u\wedge x - ux) \mid u \in [0,1]\}$. Notice that $\mathbb{E}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}\left[\dot\ell_1(X,Y) \mid X\right] = 0$. So it suffices to show that the pairwise-sum $\mathscr{A}_1 + \mathscr{A}_2$ has the Donsker property. Since $|\sup_{a\in\mathscr{A}_1\cup\mathscr{A}_2}\int a\,\mathrm{d}C| < \infty$, this pairwise-sum is indeed Donsker if $\mathscr{A}_1$ and $\mathscr{A}_2$ are both Donsker. Of course, $\mathscr{A}_1$ is Donsker. Since $\dot\ell_1$ is bounded and we have the bound $|\dot\ell_1(x,y)(u\wedge x - ux) - \dot\ell_1(x,y)(u'\wedge x - u'x)| \le 2|\dot\ell_1(x,y)||u - u'|$, it follows that $\mathscr{A}_2$ is indeed Donsker (see, for example, Van der Vaart (2000) page 271).

Part B
First we show that it suffices to consider $F_0 = G_0 = \mathrm{Un}[0,1]$. Recall from (5.12) that $\hat F_n(x) = F_n(x) + A_n(F_n(x))$, $x \in \mathbb{R}$ a.s. Introduce $U_i = F_0(X_i)$, $i = 1,\dots,n$. Then U_i, $i = 1,\dots,n$, are i.i.d. Un[0,1] distributed. Let F_n^U denote the empirical distribution function of $U_1,\dots,U_n$; we have $F_n(x) = F_n^U(F_0(x))$, $x \in \mathbb{R}$. Next note that

A_n only depends on $X_1, \ldots, X_n$ by their ranks. Hence if we compute A_n^U, resulting from $\hat{F}_n^U$, our estimator calculated from the data $(U_1, Y_1), \ldots, (U_n, Y_n)$, we have $A_n = A_n^U$ a.s. Hence we have $\hat{F}_n(x) = (\hat{F}_n^U \circ F_0)(x)$. Thus, by the continuous mapping theorem, it suffices to prove the theorem for $F_0 = \mathrm{Un}[0,1]$. Furthermore, it is easy to see that $\mathscr{L}(\hat{F}_n \mid \mathbb{P}_{F_0,G_0}) = \mathscr{L}(\hat{F}_n \mid \mathbb{P}_{F_0,\mathrm{Un}[0,1]})$. So it is indeed sufficient to prove the theorem for $F_0 = G_0 = \mathrm{Un}[0,1]$. Denote $\mathbb{P} = \mathbb{P}_{F_0,G_0}$.

We will need the following lemma several times.

Lemma 5.3.5. Suppose $\mathscr{F}$ is a $\mathbb{P}$-Donsker class of functions, and that $(f_n^u)_{u\in[0,1]}$ and $(g_n^u)_{u\in[0,1]}$ are sequence of random functions with values in $\mathscr{F}$ that satisfy

$$\sup_{u\in[0,1]} \int (f_n^u(x,y) - g_n^u(x,y))^2 \,\mathrm{d}C(x,y) = o(1;\mathbb{P}).$$

Then we have

$$\sup_{u\in[0,1]} \sqrt{n} \left| \int (f_n^u(x,y) - g_n^u(x,y)) \,\mathrm{d}(H_n - C)(x,y) \right| = o(1;\mathbb{P}).$$

Proof. Since $\mathscr{F}$ is $\mathbb{P}$-Donsker (which entails asymptotic equicontinuity of the empirical process) we have, for every $\epsilon > 0$,

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left\{ \sup_{f,g\in\mathscr{F}:\mathrm{var}_{\mathbb{P}}(f-g)(X,Y)\le\delta^2} \sqrt{n} \left| \int (f-g)\,\mathrm{d}(H_n - C) \right| > \epsilon \right\} = 0.$$

Let $\eta, \epsilon > 0$. There exists $\delta = \delta_{\eta,\epsilon}$ and $N = N_\delta$ such that for all $n \ge N$,

$$\mathbb{P}\left\{ \sup_{f,g\in\mathscr{F}:\mathrm{var}_{\mathbb{P}}(f-g)(X,Y)\le\delta^2} \sqrt{n} \left| \int (f-g)\,\mathrm{d}(H_n - C) \right| > \epsilon \right\} \le \frac{\eta}{2},$$

and there exists $M = N_\delta$ such that, for all $n \ge M$,

$$\mathbb{P}\left\{ \sup_{u\in[0,1]} \int (h_n^u(x,y))^2 \,\mathrm{d}C(x,y) > \delta^2 \right\} \le \frac{\eta}{2},$$

with $h_n^u(x,y) = f_n^u(x,y) - g_n^u(x,y)$. Hence we have for $n \ge M \vee N$,

$$\begin{aligned}
&\mathbb{P}\left\{ \sup_{u\in[0,1]} \sqrt{n} \left| \int h_n^u(x,y)\,\mathrm{d}(H_n - C)(x,y) \right| > \epsilon \right\} \\
&\qquad \le \mathbb{P}\left\{ \sup_{f\in\mathscr{F}:\mathrm{var}_{\mathbb{P}} f(X,Y)\le\delta^2} \sqrt{n} \left| \int f\,\mathrm{d}(H_n - C) \right| > \epsilon \right\} \\
&\qquad\quad + \mathbb{P}\left\{ \sup_{u\in[0,1]} \int (h_n^u(x,y))^2 \,\mathrm{d}C(x,y) > \delta^2 \right\} \le \eta,
\end{aligned}$$

which concludes the proof. □

With probability converging to 1 the process A_n satisfies (5.13), and by definition of $\dot{\Psi}_n$ this can be written as, for $u \in [0,1]$,

$$\dot{\Psi}_n(A_n)(u) = -\frac{1}{n}\sum_{i=1}^{n}(F_n(X_i)\wedge u - F_n(X_i)u)\dot{\ell}_1(F_n(X_i), G_0(Y_i)). \qquad (5.17)$$

For convenience we ignore the asymptotically null-set on which (5.13) does not hold. In the following all expectations and probabilities are calculated under $\mathbb{P} = \mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}$. To enhance readability some parts of the proof are organized in propositions.

We will prove that

$$\left(\sqrt{n}\left(\hat{F}_n(u) - u\right)\right)_{u\in[0,1]} = -\dot{\Psi}_C^{-1}\mathbb{S}_n^{\mathrm{Un}[0,1],\mathrm{Un}[0,1]} + o(1;\mathbb{P}),$$

from which the result, by a combination of Lemma 5.3.2 with the continuous mapping theorem, follows. Recall that we have $\hat{F}_n(u) - u = F_n(u) - u + A_n(F_n(u))$. The next proposition shows that we may replace $A_n(F_n(u)))$ by $A_n(u)$.

Proposition 5.3.6. Under the conditions of the theorem we have

$$\sup_{u\in[0,1]} \sqrt{n}\,|A_n(F_n(u)) - A_n(u)| \xrightarrow{p} 0, \text{ under } \mathbb{P}.$$

Proof.
For notational convenience introduce

$$\alpha_n(x,u) = F_n(x)\wedge F_n(u) - F_n(x)\wedge u + F_n(x)\,(u - F_n(u))\,.$$

And note that

$$\sup_{u\in[0,1]} |\alpha_n(X_i,u)| \le 2\|F_n - F_0\|_\infty.$$

From (5.13) we have

$$A_n(F_n(u)) - A_n(u) = \frac{1}{n}\sum_{i=1}^{n}\dot{\ell}_1(i)\alpha_n(X_i,u) - \frac{1}{n}\sum_{i=1}^{n}A_n(F_n(X_i))\dot{\ell}_1^2(i)\alpha_n(X_i,u).$$

Using that $\dot{\ell}_1$ is bounded, say by $C > 0$, we obtain

$$\sup_{u\in[0,1]} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|A_n(F_n(X_i))||\dot{\ell}_1^2(i)||\alpha_n(X_i,u)| \le 2C^2\|A_n\|_\infty\sqrt{n}\|F_n - F_0\|_\infty \xrightarrow{p} 0,$$

since, by Corollary 5.3, $\|A_n\|_\infty \xrightarrow{p} 0$ a.s. and $\sqrt{n}\|F_n - F_0\|_\infty = O(1;\mathbb{P})$. Since the class of non-decreasing functions from $\mathbb{R} \to [0,1]$ is a Donsker class it easily follows (using permanence of the Donsker property) that the class of functions

$$\mathscr{B} = \left\{(0,1)^2 \ni (x,y) \mapsto \dot{\ell}_1(x,y)(F(x)\wedge u - F(x)u) \mid F \in \mathscr{F}, u \in [0,1]\right\},$$

is $\mathbb{P}$-Donsker. Since

$$\int (\dot{\ell}_1(x,y)\alpha_n(x,u))^2 \, \mathrm{d}C(x,y) \le 4\|F_n - F_0\|_\infty^2 \int \dot{\ell}_1^2(x,y)\, \mathrm{d}C(x,y) \xrightarrow{p} 0,$$

Lemma 5.3.5 yields

$$\sup_{u\in[0,1]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_1(i)\alpha_n(X_i,u) \right| \xrightarrow{p} 0,$$

which completes the proof. □

Hence we have

$$\hat{F}_n(u) - u = F_n(u) - u + A_n(u) + R_n(u), \quad u \in [0,1],$$

where $\sup_{u\in[0,1]} \sqrt{n}|R_n(u)| \xrightarrow{p} 0$. Applying $\dot{\Psi}_C$ to both sides yields (let I denote the identity on $\mathrm{D}[0,1]$), since $\dot{\Psi}_C$ is continuous,

$$\dot{\Psi}_C\left(\hat{F}_n - I\right) = \dot{\Psi}_C\left(F_n - I\right) + \dot{\Psi}_C(A_n) + o(1/\sqrt{n};\mathbb{P}). \tag{5.18}$$

The next proposition establishes $\dot{\Psi}_C(A_n) = \dot{\Psi}_n(A_n) + o(1/\sqrt{n};\mathbb{P})$, which we will exploit to invoke (5.13).

Proposition 5.3.7. Under the conditions of the theorem we have

$$\sqrt{n}\left(\dot{\Psi}_n A_n - \dot{\Psi}_C A_n\right) \xrightarrow{p} 0, \text{ in } \ell^\infty([0,1]), \text{ under } \mathbb{P}.$$

Proof.
We have, from the definitions of $\dot{\Psi}_n$ and $\dot{\Psi}_C$,

$$\sqrt{n}(\dot{\Psi}_n A_n - \dot{\Psi}_C A_n)(u) = -\frac{1}{\sqrt{n}} \sum_{i=1}^n \Big\{(F_n(X_i) \wedge u - F_n(X_i)u)\, A_n(F_n(X_i))\dot{\ell}_1^2(i) - \int (x\wedge u - xu)\, A_n(x)\dot{\ell}_1^2(x,y)\, \mathrm{d}C(x,y)\Big\}.$$

Using $\|A_n\|_\infty \xrightarrow{p} 0$, $\sqrt{n}\|F_n - F_0\|_\infty = O(1;\mathbb{P})$, a Taylor expansion and the previous proposition we obtain

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (F_n(X_i)\wedge u - F_n(X_i)u)\, A_n(F_n(X_i))\dot{\ell}_1^2(i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i \wedge u - X_i u)\, A_n(X_i)\dot{\ell}_1^2(X_i,Y_i) + S_n(u),$$

where $\sup_{u\in[0,1]} |S_n(u)| \xrightarrow{p} 0$. Using $\|A_n\|_\infty \xrightarrow{p} 0$ an application of Lemma 5.3.5 yields the result. □

Hence a combination of the proposition with (5.18) yields

$$\dot{\Psi}_C\left(\hat{F}_n - I\right) = \dot{\Psi}_C(F_n - I) + \dot{\Psi}_n(A_n) + o(1/\sqrt{n};\mathbb{P}).$$

We have, from the definition of $\dot{\Psi}_C$,

$$\begin{aligned}(\dot{\Psi}_C(F_n - I))(u) &= -(F_n(u) - u) - \int (x \wedge u - xu)(F_n(x) - x)\dot{\ell}_1^2(x, y)\,\mathrm{d}C(x, y)\\ &= -(F_n(u) - u) + \int (x \wedge u - xu)(F_n(x) - x)\ddot{\ell}_{11}(x, y)\,\mathrm{d}C(x, y),\end{aligned}$$

where we used $\mathbb{E}[\ddot{\ell}_{11}(X, Y) \mid X] = -\mathbb{E}[\dot{\ell}_1^2(X, Y) \mid X]$.

Proposition 5.3.8. Under the conditions to the theorem we have, for $u \in [0, 1]$,

$$(\dot{\Psi}_C\left(\sqrt{n}(F_n - I)\right))(u) + (\dot{\Psi}_n\sqrt{n}A_n)(u) = -\mathbb{S}_n^{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}(u) + S_n(u),$$

where $\|S_n\|_\infty \xrightarrow{p} 0$.

Proof.
From (5.17) we have, for $u \in [0, 1]$,

$$\sqrt{n}\dot{\Psi}_n(A_n)(u) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i \wedge u - X_i u)\dot{\ell}_1(F_n(X_i), Y_i) - r_{n1}(u),$$

where

$$r_{n1}(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (F_n(X_i) \wedge u - X_i \wedge u + X_i u - F_n(X_i)u)\dot{\ell}_1(F_n(X_i), Y_i),$$

satisfies, by an application of Lemma 5.3.5, $\|r_{n1}\|_\infty \xrightarrow{p} 0$. Next using a Taylor expansion we obtain

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i \wedge u - X_i u)\dot{\ell}_1(F_n(X_i), G_0(Y_i)) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i \wedge u - X_i u)\dot{\ell}_1(X_i, Y_i)\\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i \wedge u - X_i u)(F_n(X_i) - X_i)\ddot{\ell}_{11}(F_{i,n}, Y_i),\end{aligned}$$

where $F_{i,n}$ is on the line segment between X_i and $F_n(X_i)$. Since $\ddot{\ell}_{11}$ is uniformly continuous and $\sqrt{n}\|F_n - F_0\|_\infty = O(1;\mathbb{P})$ we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i \wedge u - X_i u)(F_n(X_i) - X_i)\ddot{\ell}_{11}(F_{i,n}, Y_i)\\ &\quad= \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i \wedge u - X_i u)(F_n(X_i) - X_i)\ddot{\ell}_{11}(X_i, Y_i) + r_{n2}(u),\end{aligned}$$

where $r_{n2} \xrightarrow{p} 0$ in $\ell^\infty([0,1])$. Hence we obtain

$$
\begin{aligned}
(\dot\Psi_C\left(\sqrt{n}(F_n - I)\right))(u) &+ (\dot\Psi_n \sqrt{n} A_n)(u) \\
&= -\mathbb{S}_n^{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}(u) - r_{n1}(u) - r_{n2}u \\
&\quad - \Bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i \wedge u - X_i u)(F_n(X_i) - X_i)\ddot\ell_{11}(X_i, Y_i) \\
&\quad - \int (x \wedge u - xu)(F_n(x) - x)\ddot\ell_{11}(x,y)\,\mathrm{d}C(x,y)\Bigg).
\end{aligned}
$$

An application of Lemma 5.3.5 easily yields the result. □

A combination of (5.18) with the previous two propositions finally yields

$$\dot\Psi_C\left(\sqrt{n}(\hat F_n - I)\right) = -\mathbb{S}_n^{\mathrm{Un}[0,1],\mathrm{Un}[0,1]} + o(1;\mathbb{P}).$$

□

Efficiency proof 5.3.4

In this section we establish efficiency of $\hat F_n$. As mentioned in Section 5.3.1 it is a nonstandard problem to demonstrate efficiency. Fortunately, the special representation of the limiting distribution (Theorem 5.4) can be exploited to prove efficiency. Basically, following Van der Vaart (1995), the argument is that the 'score-process' $\mathbb{S}_n^{F_0,G_0}$ can be seen as an efficient estimator of a certain artificial parameter, and that efficiency is retained under Hadamard differentiable mappings.

Since we were not able to derive explicit formulas for the lower-bound to the asymptotic variances of regular estimators of F, we cannot prove efficiency of $\hat F_n$ by comparing the limiting distribution with the lower bound. We will exploit the special representation of the limiting distribution.

Theorem 5.5. *The estimator $(\hat F_n)_{n\in\mathbb{N}}$ is an efficient estimator of F in the model $\mathscr{P}(C, G_0)$.*

Proof.
Fix $F_0 \in \mathscr{F}_{\mathrm{ac}}$. We will show that $\hat F_n$ is an efficient estimator of F at $\mathbb{P}_{F_0,G_0}$.

First we recall the following characterization of efficiency (see, for example, Bickel et al. (1998) Corollary 5.2.1): $\hat F_n$ is an efficient estimator of F at $\mathbb{P}_{F_0,G_0}$ if and only if the following holds:

(E1) $\hat F_n$ is a regular estimator of F at $\mathbb{P}_{F_0,G_0}$;

(E2) for all $x \in \mathbb{R}$, $\hat{F}_n(x)$ is asymptotically linear at $\mathbb{P}_{F_0,G_0}$ with an influence function contained in the tangent space $\mathcal{T}(\mathbb{P}_{F_0,G_0} \mid \mathcal{P}(C,G_0))$. More precise: there should exist $h_x \in \mathcal{T}(\mathbb{P}_{F_0,G_0} \mid \mathcal{P}(C,G_0))$ such that, for all $x \in \mathbb{R}$,

$$\sqrt{n}(\hat{F}_n(x) - F_0(x)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_x(X_i, Y_i) + o(1; \mathbb{P}_{F_0,G_0}). \tag{5.19}$$

Using Le Cam's third lemma and Lemma 5.3.1 the regularity easily follows.

Proposition 5.3.9. Let $F_0 \in \mathcal{F}_{\mathrm{ac}}$. Then $\hat{F}_n$ is a regular estimator of F, in the model $\mathcal{P}(C,G_0)$, at $\mathbb{P}_{F_0,G_0}$.

Proof.
Let $v \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$, and consider the path $t \mapsto F_t^v$ through F_0 (see (5.3)). Let $\alpha_n \to \alpha \in \mathbb{R}$, and denote $F^{(n)} = F^v_{\alpha_n/\sqrt{n}}$, $\mathbb{P}^{(n)} = \mathbb{P}_{F^{(n)},G_0}$. And let, for $u \in [0,1]$, $v_u(x) = 1\{x \le u\} - u$. Using Le Cam's third lemma and Lemma 5.3.1 it is easy to see (see also the proof of Theorem 2 in Van der Vaart (1995)) that, under $\mathbb{P}^{(n)}$,

$$\mathbb{S}_n^{F_0,G_0} \xrightarrow{d} \left(\mathbb{S}^C(u) + \alpha \mathbb{E}_{F_0,G_0} \dot{\ell}_F v(F_0(X), G_0(Y)) \dot{\ell}_F v_u(F_0(X), G_0(Y))\right)_{u \in [0,1]}.$$

Using $\mathbb{E}_{F_0,G_0}[\dot{\ell}_1(F_0(X), G_0(Y)) \mid X] = 0$ yields

$$\begin{aligned}
&\mathbb{E}_{F_0,G_0} \dot{\ell}_F v(F_0(X), G_0(Y)) \dot{\ell}_F v_u(F_0(X), G_0(Y)) \\
&\quad = \mathbb{E}_{F_0} v(F_0(X)) \left(1\{F_0(X) \le u\} - u\right) \\
&\qquad + \mathbb{E}_{F_0,G_0} \dot{\ell}_1^2(F_0(X), G_0(Y)) (F_0(X) \wedge u - F_0(X) u) \int_0^{F_0(X)} v(z)\, \mathrm{d}z \\
&\quad = \int_{z=0}^{u} v(z)\, \mathrm{d}z + \int \left(\dot{\ell}_1^2(x,y)(x \wedge u - xu) \int_0^x v(z)\, \mathrm{d}z \right) \mathrm{d}C(x,y).
\end{aligned} \tag{5.20}$$

After some calculus we find

$$\sup_{u \in [0,1]} \left| \sqrt{n}(F^{(n)}(F_0^{-1}(u)) - u) - \alpha \int_0^u v(z)\, \mathrm{d}z \right| \to 0,$$

and, for $u \in [0,1]$,

$$\begin{aligned}
\dot{\Psi}_C \left(x \mapsto \int_0^x v(z)\, \mathrm{d}z \right)(u) = &- \int_0^u v(z)\, \mathrm{d}z \\
&- \int \dot{\ell}_1^2(x,y)(x \wedge u - xu) \int_0^x v(z)\, \mathrm{d}z\, \mathrm{d}C(x,y).
\end{aligned}$$

By a combination of Lemma 5.3.1 with Le Cam's first lemma we have $o(1; \mathbb{P}_{F_0,G_0}) = o(1; \mathbb{P}^{(n)})$, so we obtain from Theorem 5.4, under $\mathbb{P}^{(n)}$,

$$\dot{\Psi}_C \left(\sqrt{n} \left(\hat{F}_n(F_0^{-1}(u)) - F^{(n)}(F_0^{-1}(u)) \right)_{u \in [0,1]} \right)$$

$$
\begin{aligned}
&= \dot\Psi_C\Big(\sqrt{n}\big(\hat F_n(F_0^{-1}(u)) - u\big)_{u\in[0,1]}\Big) - \dot\Psi_C\Big(\sqrt{n}\big(F^{(n)}(F_0^{-1}(u)) - u\big)_{u\in[0,1]}\Big)\\
&= \Big(-\mathbb{S}_n^{F_0,G_0}(w)\Big)_{w\in[0,1]}\\
&\quad + \left(\alpha\left(\int_0^w v(z)\,\mathrm{d}z + \int \dot\ell_1^2(x,y)(x\wedge w - xw)\int_0^x v(z)\,\mathrm{d}z\,\mathrm{d}C(x,y)\right)\right)_{w\in[0,1]}\\
&\quad + o(1;\mathbb{P}^{(n)})\\
&\xrightarrow{d} \Big(-\mathbb{S}^C(w) - \alpha\mathbb{E}_{F_0,G_0}\dot\ell_F v(F_0(X),G_0(Y))\dot\ell_F v_w(F_0(X),G_0(Y))\\
&\qquad + \alpha\int_0^w v(z)\,\mathrm{d}z + \int \dot\ell_1^2(x,y)(x\wedge w - xw)\int_0^x v(z)\,\mathrm{d}z\,\mathrm{d}C(x,y)\Big)_{w\in[0,1]}.
\end{aligned}
$$

Invoking (5.20) we conclude $\dot\Psi_C\big(\sqrt{n}\big(\hat F_n(F_0^{-1}(u)) - F^{(n)}(F_0^{-1}(u))\big)\big)_{u\in[0,1]} \xrightarrow{d} -\mathbb{S}^C$ under $\mathbb{P}^{(n)}$. An application of the continuous mapping theorem yields (notice that $\hat F_n(F_0^{-1}(F_0(x))) = \hat F_n(x)$ and $F^{(n)}(F_0^{-1}(F_0(x))) = F^{(n)}(x)$), under $\mathbb{P}^{(n)}$, $\sqrt{n}(\hat F_n - F_0) \xrightarrow{d} -(\dot\Psi_C^{-1}\mathbb{S}^C)\circ F_0$. This concludes the proof. □

So we can conclude efficiency of $\hat F_n$ once we show that (E2) holds. Since we have no explicit formulas for $\dot\Psi_C^{-1}$ we cannot check directly whether this is the case. However, we will exploit the representation of the limiting distribution to demonstrate efficiency by an indirect argument. We already noticed that the components of the process $\mathbb{S}_n^{F_0,G_0}$ are composed of elements of the tangent space. Let, for $u \in [0,1]$, $v^u = 1\{\cdot \le u\} - u$. Introduce, for $u \in [0,1]$), the *artificial* parameters

$$
\mathscr{F}_{\mathrm{ac}} \ni H \mapsto v_u^{F_0}(H) = \mathbb{E}_{H,G_0}\dot\ell_F v^u(F_0(X),G_0(Y)).
$$

And note that $v_u^{F_0}(F_0) = 0$. Conclude that, *at* $\mathbb{P}_{F_0,G_0}$, $\mathbb{S}_n^{F_0,G_0}(u)$ is an asymptotically linear estimator of $v_u^{F_0}(F)$, with influence function contained in $\mathscr{T}(\mathbb{P}_{F_0,G_0} \mid \mathscr{P}(C,G_0))$. Consequently, these estimators are efficient *at* $\mathbb{P}_{F_0,G_0}$ one we show that they are regular at $\mathbb{P}_{F_0,G_0}$. Let $v \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$, and consider the path $t \mapsto F_t^v$ through F_0. Let $\alpha_n \to \alpha \in \mathbb{R}$, and denote $F^{(n)} = F^v_{\alpha_n/\sqrt{n}}$, $\mathbb{P}^{(n)} = \mathbb{P}_{F^{(n)},G_0}$. Using Le Cam's third lemma and Lemma 5.3.1 this regularity follows once we show

$$
\lim_{n\to\infty}\sqrt{n}\Big(v_u^{F_0}(F^v_{\alpha_n/\sqrt{n}}) - v_u^{F_0}(F_0)\Big) = \alpha\mathbb{E}_{F_0,G_0}\dot\ell_F v(F_0(X),G_0(Y))\dot\ell_F v^u(F_0(X),G_0(Y)),
$$

which easily follows. Hence we conclude that, at $\mathbb{P}_{F_0,G_0}$, $\mathbb{S}_n^{F_0,G_0}(u)$ is, at $\mathbb{P}_{F_0,G_0}$, an efficient estimator of the parameter $F \mapsto v_u^{F_0}(F)$. Since we already established tightness of $\mathbb{S}_n^{F_0,G_0}$, and since marginal efficiency plus tightness is equivalent to efficiency, we conclude that $\mathbb{S}_n^{F_0,G_0}$ is, at $\mathbb{P}_{F_0,G_0}$, an efficient estimator of the parameter $F \mapsto (v_u^{F_0}(F))_{u\in[0,1]}$. Thus we see that, at $\mathbb{P}_{F_0,G_0}$, $\sqrt{n}(\hat F_n - F_0)$ is a continuous, linear transformation of the efficient estimator $\mathbb{S}_n^{F_0,G_0}$. Since efficiency is retained under Hadamard differentiable mappings we conclude that

$\sqrt{n}(\hat{F}_n - F_0)$, at $\mathbb{P}_{F_0,G_0}$, an efficient estimator of *a certain* parameter that vanishes at $\mathbb{P}_{F_0,G_0}$; for details we refer to the proof of Theorem 3 in Van der Vaart (1995). Hence the influence functions of the components of $\sqrt{n}(\hat{F}_n - F_0)$ are, at $\mathbb{P}_{F_0,G_0}$, contained in the tangent space $\mathscr{T}(\mathbb{P}_{F_0,G_0}|\mathscr{P}(C,G_0))$, which yields (E2). Since we already proved regularity this proves efficiency of the NPMLE at $\mathbb{P}_{F_0,G_0}$. □

5.4 Copula known

In this section the model for one observation (X,Y) is

$$\mathscr{P}(C) = \left(\mathbb{P}_{F,G} \mid F, G \in \mathscr{F}_{\text{ac}}\right),$$

where the copula C satisfies Assumptions (C1)-(C3). We study efficient estimation of the parameter (F,G), seen as an element of the space $\ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R})$, from an i.i.d. sample (X_i, Y_i) $i = 1, \ldots, n$.

5.4.1 Tangent space & inefficiency of (F_n, G_n)

In this section we derive the tangent space and show that, amongst the copulas satisfying Assumptions (C1)-(C3), the independence copula is the only copula for which the marginal empirical distribution functions are an efficient estimator of (F,G).

Fix $F_0, G_0 \in \mathscr{F}_{\text{ac}}$. Let $v, w \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$. Just as in (5.3) we define densities, for $t \in (-1,1)$,

$$f_t^v(x) = c_f^v(t) k(tv(F_0(x))) f_0(x), \quad g_t^w(y) = c_g^w(t) k\left(tw(G_0(y))\right) g_0(y), \tag{5.21}$$

which induce distribution functions $F_t^v, G_t^w \in \mathscr{F}_{\text{ac}}$, and the paths $t \mapsto F_t^v$, $t \mapsto G_t^w$ pass F_0 and G_0 at $t=0$. Analogous to the score operator (5.4) we define the score operator for the second coordinate, $\dot{\ell}_G : \mathrm{L}_2^0(\mathrm{Un}[0,1]) \to \mathrm{L}_2(\mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]})$ by (see Proposition 4.7.5 in Bickel et al. (1998)),

$$\dot{\ell}_G w(X,Y) = w(Y) + \dot{\ell}_2(X,Y) \int_0^Y w(z)\,\mathrm{d}z.$$

The following proposition yields a tangent space at $\mathbb{P}_{F_0,G_0}$.

Lemma 5.4.1. Let $v, w \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$. Then the path $t \mapsto (F_t^v, G_t^w)$, as defined by (5.21) has the following score at $t=0$,

$$\dot{\ell}_{F_0,G_0}^{v,w}(X,Y) = \dot{\ell}_F v(F_0(X), G_0(Y)) + \dot{\ell}_G w(F_0(X), G_0(Y)),$$

i.e.,

$$\lim_{t\to 0} \int_{y=0}^1 \int_{x=0}^1 \left(\frac{\sqrt{p_t(x,y)} - \sqrt{p_0(x,y)}}{t} - \frac{1}{2} \dot{\ell}_{F_0,G_0}^{v,w}(x,y) \sqrt{p_0(x,y)} \right)^2 \mathrm{d}x\,\mathrm{d}y = 0,$$

where, $p_t(x,y) = p_{F_t^v, G_t^w}(x,y)$. This yields the following tangent set at (F_0, G_0),

$$\mathscr{T}(\mathbb{P}_{F_0,G_0} \mid \mathscr{P}(C)) = \left\{ \dot{\ell}^{v,w}_{F_0,G_0}(X,Y) \mid v,w \in \mathrm{L}_2^0(\mathrm{Un}[0,1]) \right\},$$

which is a closed linear subspace of $\mathrm{L}_2(\mathbb{P}_{F_0,G_0})$.

Proof.
The part on the score is Proposition 4.7.4 in Bickel et al. (1998). The closedness of the tangent space follows from Proposition 4.7.6 and Theorem A.4.2.B in Bickel et al. (1998). □

An estimator (not necessarily measurable) (F_n^*, G_n^*) of (F,G) is regular at $\mathbb{P}_{F_0,G_0}$ along the submodel $t \mapsto \mathbb{P}_{F_t^v, G_t^w}$ through $\mathbb{P}_{F_0,G_0}$ if there exists a tight Borel measurable element L_0 in $\ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R})$ such that for all $u_n^1 \to u_1 \in \mathbb{R}$, $u_n^2 \to u_2 \in \mathbb{R}$ we have

$$\sqrt{n}\left((F_n^*, G_n^*) - (F^v_{u_n^1/\sqrt{n}}, G^w_{u_n^2/\sqrt{n}}) \right) \xrightarrow{d} L_0, \text{ under } \mathbb{P}_{F^v_{u_n^1/\sqrt{n}}, G^w_{u_n^2/\sqrt{n}}} \text{ in } \ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R}).$$

An estimator (F_n^*, G_n^*) of (F,G) is (semiparametrically) regular at $\mathbb{P}_{F_0,G_0}$ if it is regular along all submodels $t \mapsto \mathbb{P}_{F_t^v, G_t^w}$, $v,w \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$, through $\mathbb{P}_{F_0,G_0}$. Finally, an estimator is regular for the model $\mathscr{P}(C)$, if it is regular at every $(F,G) \in \mathscr{F}_{\mathrm{ac}} \times \mathscr{F}_{\mathrm{ac}}$. Our parameter of interest is described by the mapping $\nu : \mathscr{P}(C) \to \ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R})$ defined by $\nu(\mathbb{P}_{F,G}) = \big((F(x))_{x\in\mathbb{R}}, (G(y))_{y\in\mathbb{R}}\big)$. Fix $F_0, G_0 \in \mathscr{F}_{\mathrm{ac}}$. We need the pathwise derivative of ν along the paths that generate the tangent space $\mathscr{T}(\mathbb{P}_{F_0,G_0} \mid \mathscr{P}(C))$. For a path $t \mapsto (F_t^v, G_t^w)$, defined by (5.21), it is an easy exercise to show $t^{-1}\left(\nu(\mathbb{P}_{F_t^v,G_t^w}) - \nu(\mathbb{P}_{F_0,G_0}) \right) \to \nu'_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0})$ in $\ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R})$ as $t \to 0$. Here $\nu'_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0}) = (\nu^{1'}_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0}), \nu^{2'}_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0}))$ is defined by

$$\nu^{1'}_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0})(x) = \int_0^{F_0(x)} v(z)\,\mathrm{d}z, \quad \nu^{2'}_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0})(y) = \int_0^{G_0(y)} w(z)\,\mathrm{d}z, \quad x,y \in \mathbb{R}.$$

By the Riesz representation theorem there exist unique elements $\nu^{1*}_{x,\mathbb{P}_{F_0,G_0}}$ and $\nu^{2*}_{y,\mathbb{P}_{F_0,G_0}}$ in $\mathscr{T}(\mathbb{P}_{F_0,G_0} \mid \mathscr{P}(C))$ such that, for all $v,w \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$.

$$\nu'_{\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0})(x,y) = \begin{pmatrix} \nu'_{1,\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0})(x) \\ \nu'_{2,\mathbb{P}_{F_0,G_0}}(\dot{\ell}^{v,w}_{F_0,G_0})(y) \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{F_0,G_0} \nu^{1*}_{x,\mathbb{P}_{F_0,G_0}} \dot{\ell}^{v,w}_{F_0,G_0}(X,Y) \\ \mathbb{E}_{F_0,G_0} \nu^{2*}_{y,\mathbb{P}_{F_0,G_0}} \dot{\ell}^{v,w}_{F_0,G_0}(X,Y) \end{pmatrix}.$$

Now the convolution theorem yields a lower bound to the precision of regular estimators: if (F_n^*, G_n^*) is a regular estimator of (F,G) at $\mathbb{P}_{F_0,G_0}$ with limit distribution W, then there exist tight Borel measurable elements L and N in $\ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R})$ such that $\mathscr{L}(W) = \mathscr{L}(L+N)$, where L and N are independent and L is a mean 0 Gaussian process whose covariances are determined by the efficient influence operator $(x,y) \mapsto (\nu^{1*}_{x,\mathbb{P}_{F_0,G_0}}, \nu^{2*}_{y,\mathbb{P}_{F_0,G_0}})$. Since $\mathscr{L}(L)$ is determined by the model only, via its efficient influence operator, it represents inevitable noise. Therefore, it is natural to call an estimator efficient at $\mathbb{P}_{F_0,G_0}$ if it is regular at $\mathbb{P}_{F_0,G_0}$ and if its limiting distribution (under $\mathbb{P}_{F_0,G_0}$) is given by L. An estimator of (F,G) is efficient if it is efficient at all $\mathbb{P}_{F_0,G_0}$. The next proposition shows that, amongst the copulas satisfying Assumptions (C1)-(C3), the independence copula is the only copula for which (F_n,G_n) constitutes an efficient estimator of (F,G). Compared to the proof of Theorem 5.2 the present proof uses a more advanced argument, since the tangent space is now the sum of two non-orthogonal spaces.

Theorem 5.6. *Let C a copula satisfying Assumptions (C1)-(C3). Then (F_n,G_n) is an efficient estimator of (F,G) if and only if $C(u,v)=uv$.*

Proof. Let $F_0,G_0\in\mathscr{F}_{\mathrm{ac}}$.
Using Corollary 5.2.1 in Bickel et al. (1998) and the 'transformation of axes' structure of the tangent space, it is easy to see that (F_n,G_n) is efficient at $\mathbb{P}_{F_0,G_0}$ if and only if (F_n,G_n) is efficient at $\mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}$. Therefore we only consider uniform margins in the sequel of the proof. Since no confusion can arise we drop the subscripts related to the margins.
'$\Longleftarrow$' It is easy to check that for all $\alpha,\beta\in[0,1]$ we have $\nu_\alpha^{1*}=1\{X\le\alpha\}-\alpha$ and $\nu_\beta^{2*}=1\{Y\le\beta\}-\beta$. Now efficiency directly follows from Bickel et al. (1998) Corollary 5.2.1.
'$\Longrightarrow$' Since F_n is an efficient estimator of F, for all $\alpha\in[0,1]$, the influence function of $F_n(\alpha)$, $x\mapsto 1\{x\le\alpha\}-\alpha$, belongs to the tangent space $\mathscr{T}(\mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}\mid\mathscr{P}(C))$, i.e. there exists $a^\alpha,b^\alpha\in\mathrm{L}_2^0(\mathrm{Un}[0,1])$ such that

$$1\{X\le\alpha\}-\alpha=\dot{\ell}_F a^\alpha(X,Y)+\dot{\ell}_G b^\alpha(X,Y)\quad\text{a.s.}\tag{5.22}$$

Using $\mathbb{E}[\dot{\ell}_F a(X,Y)+\dot{\ell}_G b(X,Y)\mid X]=a(X)$ and $\mathbb{E}[\dot{\ell}_F a(X,Y)+\dot{\ell}_G b(X,Y)\mid Y]=b(Y)$ for $a,b\in\mathrm{L}_2^0(\mathrm{Un}[0,1])$ (easy to check using partial integration) we obtain

$$a^\alpha(x)=1\{x\le\alpha\}-\alpha,\quad b^\alpha(y)=\int_{z=0}^{\alpha}c(z,y)\,\mathrm{d}z-\alpha.$$

A combination with (5.22) yields, for all $x,y\in(0,1)$, $\alpha\in(0,1)$ (since all functions involved are continuous the 'a.s.' disappears),

$$-\dot{\ell}_1(x,y)(x\wedge\alpha-x\alpha)=\int_{z=0}^{\alpha}c(z,y)\,\mathrm{d}z-\alpha+\dot{\ell}_2(x,y)(C(\alpha,y)-\alpha y).\tag{5.23}$$

In case $x<\alpha$ differentiating both sides of (5.23) with respect to x yields

$$-(1-\alpha)\left(x\ddot{\ell}_{11}(x,y)+\dot{\ell}_1(x,y)\right)=\ddot{\ell}_{12}(x,y)(C(\alpha,y)-\alpha y),\tag{5.24}$$

and in case $x>\alpha$ we have

$$-\alpha\left(\ddot{\ell}_{11}(x,y)(1-x)-\dot{\ell}_1(x,y)\right)=\ddot{\ell}_{12}(x,y)(C(\alpha,y)-\alpha y).\tag{5.25}$$

Fix $x, y \in (0,1)$. Since all objects involved are continuous, we obtain, by letting $\alpha \downarrow x$ in (5.24) and $\alpha \uparrow x$ in (5.25),

$$(1-x)\left(x\ddot{\ell}_{11}(x,y)+\dot{\ell}_1(x,y)\right) = x\left(\ddot{\ell}_{11}(x,y)(1-x)-\dot{\ell}_1(x,y)\right).$$

Trivially, this yields $\dot{\ell}_1(x,y) = 0$. Hence $c_1(x,y) = 0$. So $x \mapsto c(x,y)$ is constant. This yields $c(x,y) = 1$. □

Remark 3. From the proof we see that actually a stronger result holds: F_n (G_n) is an efficient estimator of F (G) only for the independence copula, i.e. we only need efficiency of one marginal to conclude that the copula must be the independence copula. Also notice that compared to Theorem 5.2 we now need efficiency of all $F_n(z)$, $z \in \mathbb{R}$, to conclude that the copula is the independence copula.

The estimator 5.4.2

Following the motivation in Section 5.3.2 it is natural to take as estimator of (F,G) a maximum, denoted by (F_n^*, G_n^*), of the constrained quadratic optimization problem

$$\begin{aligned} \max_{F,G} \quad & \mathscr{L}_n(F,G), \\ \text{s.t.} \quad & F \text{ probability distribution concentrated on } \{X_1,\dots,X_n\}, \\ & G \text{ probability distribution concentrated on } \{Y_1,\dots,Y_n\}. \end{aligned} \tag{5.26}$$

where the objective function is given by, for $F,G \in \mathscr{F}$,

$$\begin{aligned} \mathscr{L}_n(F,G) = {} & \frac{1}{n}\sum_{i=1}^n \left\{\dot{\ell}_1(i)\,(F(X_i)-F_n(X_i)) + \dot{\ell}_2(i)\,(G(Y_i)-G_n(Y_i))\right\} \\ & - \frac{1}{2n}\sum_{i=1}^n \left\{\dot{\ell}_1^2(i)\,(F(X_i)-F_n(X_i))^2 + \dot{\ell}_2^2(i)\,(G(Y_i)-G_n(Y_i))^2\right\} \\ & + \frac{1}{n}\sum_{i=1}^n \ddot{\ell}_{12}(i)\,(F(X_i)-F_n(X_i))\,(G(Y_i)-G_n(Y_i)) \\ & - \frac{1}{2n}\sum_{i=1}^n (nF\{X_i\}-1)^2 - \frac{1}{2n}\sum_{i=1}^n (nG\{Y_i\}-1)^2, \end{aligned}$$

where we use the abbreviations[6], for $k,\ell = 1,2$, $i = 1,\dots,n$,

$$\dot{\ell}_k(i) = \dot{\ell}_k\left(\tilde{F}_n(X_i), \tilde{G}_n(Y_i)\right), \quad \ddot{\ell}_{k\ell}(i) = \ddot{\ell}_{k\ell}\left(\tilde{F}_n(X_i), \tilde{G}_n(Y_i)\right).$$

Notice that we now use the rescaled empirical distribution functions, since the $\dot{\ell}_i$'s and $\ddot{\ell}_{ij}$'s are possible not defined on the boundary of $[0,1]^2$. Unfortunately,

[6] Please note that $\dot{\ell}_1(i)$ and $\ddot{\ell}_{11}(i)$ differ from those we used in Section 5.3.

we are not able to prove, in general, consistency of $(\hat{F}_n, \hat{G}_n)$. It seems impossible to use the traditional arguments for consistency of M-estimators. This since our criterion function is not smooth enough. Therefore we will develop, inspired by the previous section, an alternative estimator.

Remark 4. If one is interested in the model where the copula is known and the (absolutely continuous) marginals are unknown, but equal, it is natural to take as estimator a maximum of the optimization problem (5.26) with the additional constraint $F = G$. Then: if the copula satisfies $\ddot{\ell}_{12} \leq 0$ (for example a Gaussian copula with negative correlation coefficient), consistency can be proved analogous to the proof of Proposition 5.3.2.

To state our estimator we first introduce an analogue of (5.14). Define the operator $\dot{\Psi}_C = (\dot{\Psi}_C^1, \dot{\Psi}_C^2) : \mathrm{D}[0,1] \times \mathrm{D}[0,1] \to \mathrm{D}[0,1] \times \mathrm{D}[0,1]$ by, for $u, v \in [0,1]$,

$$\dot{\Psi}_C^1(h_1, h_2)(u) = -h_1(u) - \int (x \wedge u - xu)\left(h_1(x)\dot{\ell}_1^2(x,y) - h_2(y)\ddot{\ell}_{12}(x,y)\right)\mathrm{d}C(x,y),$$

$$\dot{\Psi}_C^2(h_1, h_2)(v) = -h_2(y) - \int \left(y \wedge v - yv\right)\left(h_2(y)\dot{\ell}_2^2(x,y) - h_1(x)\ddot{\ell}_{12}(x,y)\right)\mathrm{d}C(x,y).$$

The following lemma is the analogue of Lemma 5.3.2.

Lemma 5.4.2. The operator $\dot{\Psi}_C : \mathrm{D}[0,1] \times \mathrm{D}[0,1] \to \mathrm{D}[0,1] \times \mathrm{D}[0,1]$ is onto and one-to-one, and the inverse $\dot{\Psi}_C^{-1}$ is continuous.

Proof.
In Step A we show that $\dot{\Psi}_C$ is one-to-one, and in Step B we show that $\dot{\Psi}_C$ is onto and that the inverse $\dot{\Psi}_C^{-1}$ is continuous. In the following we drop the subscript C, and all expectations are taken under $\mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}$.

Step A Since $\dot{\Psi}$ is linear, we have to show that the null space of $\dot{\Psi}$ is trivial. So let $(h_1, h_2) \in \mathrm{D}[0,1] \times \mathrm{D}[0,1]$ be such that $\dot{\Psi}(h_1, h_2) = 0$:

$$0 = -h_1(u) - \mathbb{E}(u \wedge X - uX)\left(\dot{\ell}_1^2(X,Y)h_1(X) - \ddot{\ell}_{12}(X,Y)h_2(Y)\right), \quad u \in [0,1], \quad (5.27)$$

$$0 = -h_2(v) - \mathbb{E}(v \wedge Y - vY)\left(\dot{\ell}_2^2(X,Y)h_2(Y) - \ddot{\ell}_{12}(X,Y)h_1(X)\right), \quad v \in [0,1]. \quad (5.28)$$

Plugging in $u = 0$, $v = 0$, $u = 1$ and $v = 1$ we see that necessarily

$$h_1(0) = h_2(0) = h_1(1) = h_2(1) = 0. \quad (5.29)$$

Notice that,

$$\begin{aligned} P_1(u) &= -\mathbb{E}(u \wedge X - uX)\left(\dot{\ell}_1^2(X,Y)h_1(X) - \ddot{\ell}_{12}(X,Y)h_2(Y)\right) \\ &= u\mathbb{E}X\left(\dot{\ell}_1^2(X,Y)h_1(X) - \ddot{\ell}_{12}(X,Y)h_2(Y)\right) \\ &\quad - \int_{x=0}^{u}\int_{y=0}^{1} x\left(\dot{\ell}_1^2(x,y)h_1(x) - \ddot{\ell}_{12}(x,y)h_2(y)\right)c(x,y)\,\mathrm{d}y\,\mathrm{d}x \end{aligned}$$

$$- u \int_{x=u}^{1} \int_{y=0}^{1} \left(\dot{\ell}_1^2(x,y) h_1(x) - \ddot{\ell}_{12}(x,y) h_2(y) \right) c(x,y) \, \mathrm{d}y \, \mathrm{d}x,$$

is two times differentiable with

$$P_1''(u) = I_{11}(u) h_1(u) - \int_{y=0}^{1} \ddot{\ell}_{12}(u,y) h_2(y) c(u,y) \, \mathrm{d}y.$$

So from (5.27) it follows that h_1 is two times continuously differentiable. We get the same result for h_2. Using this and the boundary conditions (5.29) we see that any solution (h_1, h_2) to the system (5.27)-(5.28) is a solution to the following system of differential equations

$$\begin{cases} h_1''(u) - h_1(u) I_{11}(u) + \int_0^1 h_2(y) \ddot{\ell}_{12}(u,y) c(u,y) \, \mathrm{d}y = 0, \\ h_2''(v) - h_2(v) I_{22}(v) + \int_0^1 h_1(x) \ddot{\ell}_{12}(x,v) c(x,v) \, \mathrm{d}x = 0, \\ h_1(0) = h_2(0) = h_1(1) = h_2(1) = 0. \end{cases} \tag{5.30}$$

We will show that this system of differential equations has unique solution $h_1 = h_2 = 0$. That implies that our system of interest has unique solution $h_1 = h_2 = 0$, which concludes the proof of Step A. In the proof of Proposition 5.3.4 we encountered a certain one-dimensional version of the system (5.30). For that case we were able to prove directly that the system only has the trivial solution. For the present system it seems not possible to extend that argument. We give an indirect proof. As we will show, the system (5.30) is exactly the homogenous system corresponding to the system (4.57)-(4.58) in Klaassen and Wellner (1997) with the same boundary conditions. Since a solution of their system yields a certain efficient score, which is unique, it follows that indeed $h_1 = h_2 = 0$. Let us make some brief remarks to gain a better understanding of the arguments on pages 65-67 in Klaassen and Wellner (1997) (we use their notation). In Klaassen and Wellner (1997) the copula depends on a Euclidean parameter θ and they want to calculate the efficient score for θ, i.e. project the score $\dot{\ell}_\theta = (\partial/\partial\theta) \log c_\theta$ on the sum-space $\mathscr{R}(\dot{l}_g) + \mathscr{R}(\dot{l}_h)$. Since this sum space $\mathscr{R}(\dot{l}_g) + \mathscr{R}(\dot{l}_h)$ is indeed closed under our assumptions, the projection is unique and is, by the ACE method (Proposition A.4.1 in Bickel et al. (1998)), completely characterized by (4.42) and (4.43) in Klaassen and Wellner (1997). Before formula (4.49) they operate (4.42) by $\dot{l}_g^T$, the adjoint of $\dot{l}_g$. Note that (4.42) and (4.49)-(4.50) are indeed equivalent (note that sofar they are working in $L_2[0,1]$) since $\dot{\ell}_g^T$ is invertible (see Bickel et al. (1998) Proposition 4.7.2B). Hence we see that the efficient score for θ is completely characterized by the system of differential equations (4.57)-(4.58). Since the efficient score is unique, there is only one solution to this system. Letting A correspond to h_1 and B to h_2, we obtain from our boundary conditions for A' and B': $\int_0^1 A'(x) \, \mathrm{d}x = \int_0^1 B'(y) \, \mathrm{d}y = 0$, So our system of differential equations can indeed be interpreted as the homogenous system corresponding to (4.57)-(4.58) in Klaassen and Wellner (1997). Since their system has a unique solution, our system has only the trivial solution.

Step B This proceeds completely analogous to the proof of Proposition 5.3.5. □

Next we introduce the processes $\tilde{\mathbb{S}}_{n1} = (\tilde{\mathbb{S}}_{n1}(u))_{u\in[0,1]}$ and $\tilde{\mathbb{S}}_{n2} = (\tilde{\mathbb{S}}_{n2}(v))_{v\in[0,1]}$ by

$$\tilde{\mathbb{S}}_{n1}(u) = -\frac{1}{n}\sum_{i=1}^{n}(F_n(X_i)\wedge u - F_n(X_i)u)\dot{\ell}_1\left(\tilde{F}_n(X_i),\tilde{G}_n(Y_i)\right),$$
$$\tilde{\mathbb{S}}_{n2}(v) = -\frac{1}{n}\sum_{i=1}^{n}(G_n(Y_i)\wedge v - G_n(Y_i)v)\dot{\ell}_2\left(\tilde{F}_n(X_i),\tilde{G}_n(Y_i)\right).$$

Denote $\tilde{\mathbb{S}}_n = (\tilde{\mathbb{S}}_{n1},\tilde{\mathbb{S}}_{n2}) \in \mathrm{D}[0,1]\times\mathrm{D}[0,1]$. By Lemma 5.4.2 it is possible introduce processes $A_n, B_n \in \mathrm{D}[0,1]$ by

$$(A_n,B_n) = \dot{\Psi}_C^{-1}\tilde{\mathbb{S}}_n.$$

Notice that the processes A_n and B_n only depend on the copula, and not on the marginals. Inspired by (5.12) we now introduce our estimator $(\hat{F}_n,\hat{G}_n)$:

$$\hat{F}_n(x) = F_n(x) + A_n(F_n(x)),\;\; x\in\mathbb{R},\quad \hat{G}_n(y) = G_n(y) + B_n(G_n(y)),\;\; y\in\mathbb{R}.$$

Note that $\hat{F}_n,\hat{G}_n \in \ell^\infty(\mathbb{R})$, but they need not be distribution functions. The next proposition establishes consistency.

Proposition 5.4.1. Let $F_0, G_0 \in \mathscr{F}_{\mathrm{ac}}$. Then we have

$$\|\hat{F}_n - F_0\|_\infty + \|\hat{G}_n - G_0\|_\infty \to 0,\quad \mathbb{P}_{F_0,G_0}\text{-a.s.}$$

Proof.
Of course, it suffices to prove that $\|A_n\|_\infty + \|B_n\|_\infty \to 0$ a.s. Since (A_n,B_n) depends on (X_i,Y_i), $i=1,\dots,n$, only by (R_i^X,R_i^Y), $i=1,\dots,n$, it suffices to prove the proposition under $\mathbb{P}=\mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}$. Since $\dot{\Psi}_C^{-1}$ is continuous it suffices to prove that $\|\tilde{\mathbb{S}}_n\|_\infty \to 0$ $\mathbb{P}$-a.s. We prove $\|\tilde{\mathbb{S}}_{n1}\|_\infty \to 0$ $\mathbb{P}$-a.s.; the second component of $\tilde{\mathbb{S}}_n$ proceeds in exactly the same way. By the mean value theorem we have,

$$\begin{aligned}\tilde{\mathbb{S}}_{n1}(u) = &-\frac{1}{n}\sum_{i=1}^{n}(F_n(X_i)\wedge u - F_n(X_i)u)\dot{\ell}_1(X_i,Y_i)\\ &-\frac{1}{n}\sum_{i=1}^{n}(F_n(X_i)\wedge u - F_n(X_i)u)\left\{\ddot{\ell}_{11}(F_{ni},G_{ni})(\tilde{F}_n(X_i)-X_i)\right.\\ &\qquad\left.+\ddot{\ell}_{12}(F_{ni},G_{ni})(\tilde{G}_n(Y_i)-Y_i)\right\}. \end{aligned}\tag{5.31}$$

where (F_{ni},G_{ni}) is a point on the line segment between $(\tilde{F}_n(X_i),\tilde{G}_n(Y_i))$ and (X_i,Y_i), $i=1,\dots,n$. It is easy to see, using that $\dot{\ell}_1$ is square-integrable, Cauchy-Schwarz, and that the class of monotone functions from $\mathbb{R}$ into $[0,1]$ has for all

$\epsilon > 0$ a finite $L_2(\mathbb{Q})$-ϵ-bracketing number for all probability measures $\mathbb{Q}$ on the real line, that the class of functions $\mathscr{A} = \{(0,1)^2 \ni (x,y) \mapsto (F(x) \wedge u - F(x)u)\dot{\ell}_1(x,y) \mid F \in \mathscr{F},\ u \in [0,1]\}$ satisfies for all $\epsilon > 0$ $N_{[]}(\epsilon, \mathscr{A}, L_1(\mathbb{P})) < \infty$. Since $\mathbb{E}[h(X)\dot{\ell}_1(X,Y) \mid X] = 0$, we thus have, by the Glivenko-Cantelli theorem,

$$\sup_{u\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} (F_n(X_i) \wedge u - F_n(X_i)u)\dot{\ell}_1(X_i, Y_i) \right| \le \sup_{a\in\mathscr{A}} \left| \frac{1}{n}\sum_{i=1}^{n} a(X_i, Y_i) \right| \to 0,$$

$\mathbb{P}_{F_0,G_0}$-a.s. Next we show that

$$\sup_{u\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} (F_n(X_i) \wedge u - F_n(X_i)u)\ddot{\ell}_{11}(F_{ni}, G_{ni})(\tilde{F}_n(X_i) - X_i) \right| \xrightarrow{p} 0. \quad (5.32)$$

This will conclude the proof, since the last term in (5.31) can be handled similarly. Let $\alpha \in (5/12, 1/2)$. Recall that the weighted empirical processes $\sqrt{n}(F_n - I)/r^\alpha$ and $\sqrt{n}(G_n - I)/r^\alpha$ converge in distribution under $\mathbb{P}$ (here I denotes the identity on $[0,1]$). Using Lemma 2.10.14 in Van der Vaart and Wellner (1993) we conclude that, $\mathbb{P}$-a.s.,

$$\sup_{x\in[0,1]} \left| \frac{F_n(x) - x}{r^{2\alpha}(x)} \right| + \sup_{y\in[0,1]} \left| \frac{G_n(y) - y}{r^{2\alpha}(y)} \right| \to 0.$$

The probability of the event

$$\mathscr{E}_n = \left\{ \sup_{x\in[0,1]} \left| \frac{F_n(x) - x}{r^{2\alpha}(x)} \right| + \sup_{y\in[0,1]} \left| \frac{G_n(y) - y}{r^{2\alpha}(y)} \right| < 1 \right\}$$

thus converges to 1 as $n \to \infty$. Using Assumption (C3) we have

$$\left|\ddot{\ell}_{11}(F_{ni}, G_{ni}) - \ddot{\ell}_{11}(X_i, Y_i)\right| \le M \left(\frac{\left\|(F_n - I)/r^{2\alpha}\right\|_\infty r^{2\alpha}(X_i)}{r^3(X_i) r^{1/2-\epsilon}(Y_i)} + \frac{\left\|(G_n - I)/r^{2\alpha}\right\|_\infty r^{2\alpha}(Y_i)}{r^2(X_i) r^1(Y_i)} \right).$$

On the event $\mathscr{E}_n$, we have the bound $0 \le u \wedge F_n(X_i) - uF_n(X_i) \le 3r^{2\alpha}(X_i)$. Decomposing $\ddot{\ell}_{11}(F_{ni}, G_{ni}) = \ddot{\ell}_{11}(X_i, Y_i) + \ddot{\ell}_{11}(F_{ni}, G_{ni}) - \ddot{\ell}_{11}(X_i, Y_i)$ we obtain, on the event $\mathscr{E}_n$, the bound

$$\begin{aligned} &\sup_{u\in[0,1]} \frac{1}{n}\sum_{i=1}^{n} \left|(F_n(X_i) \wedge u - F_n(X_i)u)\ddot{\ell}_{11}(F_{ni}, G_{ni})(\tilde{F}_n(X_i) - X_i)\right| \\ &\quad \le \frac{\|\tilde{F}_n - I\|_\infty}{n} \sum_{i=1}^{n} |\ddot{\ell}_{11}(X_i, Y_i)| + \left\| \frac{F_n - I}{r^{2\alpha}} \right\|_\infty^2 \frac{3}{n}\sum_{i=1}^{n} \frac{r^{6\alpha}(X_i)}{r^3(X_i) r^{1/2-\epsilon}(Y_i)} \\ &\quad + \left\| \frac{F_n - I}{r^{2\alpha}} \right\|_\infty \left\| \frac{G_n - I}{r^{2\alpha}} \right\|_\infty \frac{1}{n}\sum_{i=1}^{n} \frac{r^{4\alpha}(X_i) r^{2\alpha}(Y_i)}{r^2(X_i) r(Y_i)} = o(1; \mathbb{P}), \end{aligned}$$

since $\mathbb{E} r^{-(3-6\alpha)}(X_i) r^{-(1/2-\epsilon)}(Y_i)$, and $\mathbb{E} r^{-(2-4\alpha)}(X_i) r^{-(1-2\alpha)}(Y_i)$ are finite (Cauchy-Schwarz). Thus we conclude that (5.32) indeed holds. □

The next theorem presents the limiting distribution of our estimator.

Theorem 5.7. *Let $F_0, G_0 \in \mathscr{F}_{ac}$ and C satisfy Assumptions (C1)-(C3). Then, under $\mathbb{P}_{F_0,G_0}$, the following holds.*
The process $\mathbb{S}_n^{F_0,G_0} = \left(\left(\mathbb{S}_{n1}^{F_0,G_0}(u)\right)_{u\in[0,1]}, \left(\mathbb{S}_{n2}^{F_0,G_0}(v)\right)_{v\in[0,1]}\right)$, with

$$\mathbb{S}_{n1}^{F_0,G_0}(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{1\{F_0(X_i) \le u\} - u + \dot\ell_1(F_0(X_i), G_0(Y_i))(F_0(X_i) \wedge u - F_0(X_i)u)\right\},$$
$$\mathbb{S}_{n2}^{F_0,G_0}(v) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{1\{G_0(Y_i) \le v\} - v + \dot\ell_2(F_0(X_i), G_0(Y_i))(G_0(Y_i) \wedge v - G_0(Y_i)v)\right\},$$

weakly converges in $\ell^\infty([0,1]) \times \ell^\infty([0,1])$ to a tight zero-mean Gaussian process, denoted by $\mathbb{S}^C = \left((\mathbb{S}_1^C(u))_{u\in[0,1]}, (\mathbb{S}_2^C(v))_{v\in[0,1]}\right)$ whose covariance function only depends on C.
And we have, in $\ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R})$,

$$\sqrt{n}\begin{pmatrix}\left(\hat F_n(x) - F_0(x)\right)_{x\in\mathbb{R}} \\ \left(\hat G_n(y) - G_0(y)\right)_{y\in\mathbb{R}}\end{pmatrix} = -\left((\dot\Psi_C^{-1}\mathbb{S}_n^{F_0,G_0}) \circ (F_0, G_0)(x,y)\right)_{x,y\in\mathbb{R}} + o(1;\mathbb{P}_{F_0,G_0})$$
$$\rightsquigarrow -\left((\dot\Psi_C^{-1}\mathbb{S}^C) \circ (F_0, G_0)(x,y)\right)_{x,y\in\mathbb{R}}.$$

Proof.
The weak convergence of $\mathbb{S}_n^{F_0,G_0}$ follows by the same arguments as in part A of the proof to Theorem 5.4 (see also the proof of Proposition 5.4.2). So we only have to prove the weak convergence of $\sqrt{n}(\hat F_n - F_0, \hat G_n - G_0)$. Since $\tilde{\mathbb{S}}_n$ depends on $(X_1, Y_1), \ldots, (X_n, Y_n)$ only by the ranks $(R_1^X, R_1^Y), \ldots (R_n^X, R_n^Y)$, it is easy to see that it suffices to consider $F_0 = G_0 = \mathrm{Un}[0,1]$. In the following all probabilities and expectations are calculated under $\mathbb{P} = \mathbb{P}_{\mathrm{Un}[0,1],\mathrm{Un}[0,1]}$.

We will prove that

$$\sqrt{n}\dot\Psi_C\begin{pmatrix}\hat F_n - F_0 \\ \hat G_n - G_0\end{pmatrix} = -\mathbb{S}_n^{\mathrm{Un}[0,1],\mathrm{Un}[0,1]} + o(1;\mathbb{P}),$$

which, by continuity of $\dot\Psi_C^{-1}$, yields the result. We will show that,

$$\sqrt{n}\dot\Psi_C^1\begin{pmatrix}\hat F_n - F_0 \\ \hat G_n - G_0\end{pmatrix} = -\mathbb{S}_{n1}^{\mathrm{Un}[0,1],\mathrm{Un}[0,1]} + o(1;\mathbb{P}).$$

The second coordinate proceeds in exactly the same way.

We start by analyzing the structure of $\tilde{\mathbb{S}}_{n1}$.
Proposition 5.4.2. We have,

$$\sqrt{n}\tilde{\mathbb{S}}_{n1}(u) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i \wedge u - X_i u)\dot\ell_1(X_i, Y_i)$$

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i \wedge u - X_i u)\left\{\ddot{\ell}_{11}(X_i,Y_i)(F_n(X_i)-X_i)\right.$$
$$\left.+\ddot{\ell}_{12}(X_i,Y_i)(G_n(Y_i)-Y_i)\right\} + o(1;\mathbb{P}).$$

Proof.
For notational convenience we introduce

$$\alpha_n(x,u) = (u \wedge F_n(x) - x \wedge u - u(F_n(x)-x)),$$

and note that $|\alpha_n(x,u)| \le 2\|F_n - F_0\|_\infty$. Using the mean value theorem we obtain the expansion

$$\begin{aligned}\sqrt{n}\tilde{\mathbb{S}}_{n1}(u) = &-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i \wedge u - X_i u)\dot{\ell}_1(X_i,Y_i) - r_{n1}(u)\\ &-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i \wedge u - X_i u)\left\{\ddot{\ell}_{11}(X_i,Y_i)(F_n(X_i)-X_i)\right.\\ &\qquad\left.+\ddot{\ell}_{12}(X_i,Y_i)(G_n(Y_i)-Y_i)\right\}\\ &- r_{n2}(u) - r_{n3}(u) + r_{n4}(u),\end{aligned}$$

with,

$$\begin{aligned}r_{n1}(u) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\alpha_n(X_i,u)\dot{\ell}_1(X_i,Y_i),\\ r_{n2}(u) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(F_n(X_i)\wedge u - F_n(X_i)u)\left\{\left(\ddot{\ell}_{11}(F_{ni},G_{ni}) - \ddot{\ell}_{11}(X_i,Y_i)\right)\left(\tilde{F}_n(X_i)-X_i\right)\right.\\ &\qquad\left.+\left(\ddot{\ell}_{12}(F_{ni},G_{ni}) - \ddot{\ell}_{12}(X_i,Y_i)\right)\left(\tilde{G}_n(Y_i)-Y_i\right)\right\},\\ r_{n3}(u) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\alpha_n(X_i,u)\left(\ddot{\ell}_{11}(X_i,Y_i)(\tilde{F}_n(X_i)-X_i) + \ddot{\ell}_{12}(X_i,Y_i)(\tilde{G}_n(Y_i)-Y_i)\right),\\ r_{n4}(u) &= \frac{1}{n^{3/2}}\sum_{i=1}^{n}(X_i\wedge u - X_i u)\left\{\ddot{\ell}_{11}(X_i,Y_i)F_n(X_i) + \ddot{\ell}_{12}(X_i,Y_i)G_n(Y_i)\right\},\end{aligned}$$

where (F_{ni}, G_{ni}) is a (random) point on the line segment between (X_i, Y_i) and $(\tilde{F}_n(X_i), \tilde{G}_n(Y_i))$, $i = 1,\dots,n$. The proposition is proved once we show $\|r_{n1}\|_\infty + \|r_{n2}\|_\infty + \|r_{n3}\|_\infty + \|r_{n4}\|_\infty \xrightarrow{p} 0$. $\|r_{n4}\|_\infty \xrightarrow{p} 0$ is trivial, since the expectation of $|\ddot{\ell}_{1i}|(X,Y)$ is finite for $i = 1,2$. $\|r_{n3}\|_\infty \xrightarrow{p} 0$ easily follows from

$$\|r_{n3}\|_\infty \le 2\sqrt{n}\|\tilde{F}_n - F_0\|_\infty\left(\frac{\|F_n - F_0\|_\infty}{n}\sum_{i=1}^{n}|\ddot{\ell}_{11}(X_i,Y_i)| + \frac{\|G_n - G_0\|_\infty}{n}\sum_{i=1}^{n}|\ddot{\ell}_{12}(X_i,Y_i)|\right).$$

Next we discuss $\|r_{n1}\|_\infty$. Since the class of non-decreasing functions from $\mathbb{R} \to [0,1]$ is a Donsker class it easily follows (using permanence of the Donsker property) that the class of functions

$$\mathcal{B} = \left\{(0,1)^2 \ni (x,y) \mapsto \dot{\ell}_1(x,y)(F(x)\wedge u - F(x)u) \mid F \in \mathcal{F}, u \in [0,1]\right\},$$

is $\mathbb{P}$-Donsker. Since

$$\int (\dot\ell_1(x,y)\alpha_n(x,u))^2 \,\mathrm{d}C(x,y) \le 4\|F_n - F_0\|_\infty^2 \int \dot\ell_1^2(x,y)\,\mathrm{d}C(x,y) \xrightarrow{p} 0,$$

Lemma 5.3.5 yields $\|r_{n1}\|_\infty \xrightarrow{p} 0$ under $\mathbb{P}$. Finally, we discuss r_{n2}. We only discuss the first part (the $\ddot\ell_{11}$ part); the second part follows by a similar argument. Notice first, with ϵ from Assumption (C3ii), that we can find $p_1, p_2, q_1, q_2 > 1$ and $\alpha \in [0, 1/2)$ such that $p_1^{-1} + q_1^{-1} = p_2^{-1} + q_2^{-1} = 1$, $p_1(3-5\alpha) < 1$, $q_1(1/2-\epsilon) < 1$, $p_2(2-3\alpha) < 1$, $q_2(1-2\alpha) < 1$ (which we need to apply Hölder's inequality. Remember that the weighted empirical processes $r^{-\alpha}\sqrt{n}(\tilde F_n - F_0)$ and $r^{-\alpha}\sqrt{n}(\tilde G_n - F_0)$ weakly converge, and that $\|(F_n - F_0)/r^{2\alpha}\|_\infty + \|(G_n - G_0)/r^{2\alpha}\|_\infty \xrightarrow{p} 0$. Thus the probability of the event

$$\mathscr{E}_n = \left\{ \sup_{x\in[0,1]} \left| \frac{F_n(x) - x}{r^{2\alpha}(x)} \right| + \sup_{y\in[0,1]} \left| \frac{G_n(y) - y}{r^{2\alpha}(y)} \right| < 1 \right\}.$$

converges to 1 as $n \to \infty$. Recall, on $\mathscr{E}_n$, the bound, $0 \le u \wedge F_n(X_i) - uF_n(X_i) \le 3r^{2\alpha}(X_i)$. We now obtain, on the event $\mathscr{E}_n$, the bound,

$$\begin{aligned} &\sup_{u\in[0,1]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (F_n(X_i)\wedge u - F_n(X_i)u)\left(\ddot\ell_{11}(F_{ni}, G_{ni}) - \ddot\ell_{11}(X_i, Y_i)\right)\left(\tilde F_n(X_i) - X_i\right) \right| \\ &\qquad \le \|\sqrt{n}(\tilde F_n - F_0)/r^\alpha\|_\infty \frac{3}{n} \sum_{i=1}^n r^{3\alpha}(X_i) \left|\ddot\ell_{11}(F_{ni}, G_{ni}) - \ddot\ell_{11}(X_i, Y_i)\right|. \end{aligned}$$

So the proof of the proposition is complete once we show that

$$\frac{1}{n}\sum_{i=1}^n r^{3\alpha}(X_i)\left|\ddot\ell_{11}(F_{ni}, G_{ni}) - \ddot\ell_{11}(X_i, Y_i)\right| 1_{\mathscr{E}_n} \xrightarrow{p} 0.$$

Using Assumption (C3ii) we obtain (use Hölder's inequality in the last step)

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n r^{3\alpha}(X_i)\left|\ddot\ell_{11}(F_{ni}, G_{ni}) - \ddot\ell_{11}(X_i, Y_i)\right| 1_{\mathscr{E}_n} \\ &\qquad \le M\left(\left\| \frac{F_n - F_0}{r^{2\alpha}} \right\|_\infty \frac{1}{n}\sum_{i=1}^n r^{-(3-5\alpha)}(X_i) r^{-(1/2-\epsilon)}(Y_i) \right. \\ &\qquad\qquad \left. + \left\| \frac{G_n - G_0}{r^{2\alpha}} \right\|_\infty \frac{1}{n}\sum_{i=1}^n r^{-(2-3\alpha)}(X_i) r^{-(1-2\alpha)}(Y_i) \right) \\ &\qquad = M(o(1;\mathbb{P})O(1;\mathbb{P}) + o(1;\mathbb{P})O(1;\mathbb{P})) = o(1;\mathbb{P}), \end{aligned}$$

which concludes the proof. □

The next proposition is the analogue of Proposition 5.3.6.

Proposition 5.4.3. We have

$$\sup_{u\in[0,1]} \sqrt{n}|A_n(F_n(u)) - A_n(u)| + \sup_{u\in[0,1]} \sqrt{n}|B_n(G_n(u)) - B_n(u)| \xrightarrow{p} 0, \text{ under } \mathbb{P}.$$

Proof.
By definition of (A_n, B_n) we have $(A_n, B_n) = \dot{\Psi}_C^{-1}\tilde{\mathbb{S}}_n$. Operating by $\dot{\Psi}_C$ on both sides thus yields $\dot{\Psi}_C(A_n, B_n) = \tilde{\mathbb{S}}_n$. Next invoking the definition of $\dot{\Psi}_C$ we obtain, for all $z \in [0,1]$,

$$-A_n(z) + \int (x \wedge z - xz)\left(\ddot{\ell}_{11}(x,y)A_n(x) + \ddot{\ell}_{12}(x,y)B_n(y)\right) \mathrm{d}C(x,y) = \tilde{\mathbb{S}}_{n1}(z),$$

and,

$$-B_n(z) + \int \left(y \wedge z - yz\right)\left(\ddot{\ell}_{22}(x,y)B_n(y) + \ddot{\ell}_{12}(x,y)A_n(x)\right) \mathrm{d}C(x,y) = \tilde{\mathbb{S}}_{n2}(z).$$

We will prove $\sup_{u\in[0,1]} \sqrt{n}|A_n(F_n(u)) - A_n(u)| \xrightarrow{p} 0$ under $\mathbb{P}$ from the first display. The proof for B_n proceeds in the same way by using the second display. From the first display we obtain, for $u \in [0,1]$,

$$\begin{aligned} A_n(u) - A_n(F_n(u)) &= \tilde{\mathbb{S}}_{n1}(F_n(u)) - \tilde{\mathbb{S}}_{n1}(u) \\ &\quad - \int \alpha_n(X_i, u)\left(\ddot{\ell}_{11}(x,y)A_n(x) + \ddot{\ell}_{12}(x,y)B_n(y)\right) \mathrm{d}C(x,y), \end{aligned}$$

where we denote, for $x, u \in [0,1]$,

$$\alpha_n(x,u) = (x \wedge F_n(u) - x \wedge u - x(F_n(u) - u)).$$

Since $|\alpha_n(X_i,u)| \le 2\|F_n - F_0\|_\infty$ and $\|A_n\|_\infty + \|B_n\|_\infty \xrightarrow{p} 0$ (Proposition 5.4.1) we obtain, under $\mathbb{P}$,

$$\begin{aligned} \sqrt{n}&\left|\int \alpha_n(X_i,u)\left(\ddot{\ell}_{11}(x,y)A_n(x) + \ddot{\ell}_{12}(x,y)B_n(y)\right) \mathrm{d}C(x,y)\right| \\ &\le 2\sqrt{n}\|F_n - F_0\|_\infty \left(\|A_n\|_\infty \int |\ddot{\ell}_{11}|(x,y)\,\mathrm{d}C(x,y) \right. \\ &\qquad \left. + \|B_n\|_\infty \int |\ddot{\ell}_{12}|(x,y)|\,\mathrm{d}C(x,y)\right) \xrightarrow{p} 0. \end{aligned}$$

So the result follows once we prove $\sup_{u\in[0,1]} \sqrt{n}\left|\tilde{\mathbb{S}}_{n1}(F_n(u)) - \tilde{\mathbb{S}}_{n1}(u)\right| \xrightarrow{p} 0$ under $\mathbb{P}$. Using Proposition 5.4.2 we obtain,

$$\begin{aligned} \sqrt{n}\left(\tilde{\mathbb{S}}_{n1}(F_n(u)) - \tilde{\mathbb{S}}_{n1}(u)\right) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha_n(X_i,u)\dot{\ell}_1(X_i,Y_i) \\ &\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha_n(X_i,u)\left\{\ddot{\ell}_{11}(X_i,Y_i)(F_n(X_i) - X_i) + \ddot{\ell}_{12}(X_i,Y_i)(G_n(Y_i) - Y_i)\right\} \end{aligned}$$

$$+ o(1;\mathbb{P}).$$

By the proof of Proposition 5.4.2 we have

$$\sup_{u\in[0,1]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha_n(X_i,u)\dot{\ell}_1(X_i,Y_i)\right| = o(1;\mathbb{P}).$$

So the result now follows from the bound

$$\begin{aligned}&\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n \alpha_n(X_i,u)\left\{\ddot{\ell}_{11}(X_i,Y_i)(F_n(X_i)-X_i)+\ddot{\ell}_{12}(X_i,Y_i)(G_n(Y_i)-Y_i)\right\}\right|\\ &\le 2\sqrt{n}\|F_n-F_0\|_\infty\left(\|F_n-F_0\|_\infty\frac{1}{n}\sum_{i=1}^n|\ddot{\ell}_{11}(X_i,Y_i)|+\|G_n-G_0\|_\infty\frac{1}{n}\sum_{i=1}^n|\ddot{\ell}_{12}(X_i,Y_i)|\right)\\ &\xrightarrow{p} 0.\end{aligned}$$

□

Recall that $(\hat{F}_n,\hat{G}_n)=(F_n,G_n)+(A_n\circ F_n, B_n\circ G_n)$. By the proposition we thus have $(\hat{F}_n,\hat{G}_n)=(F_n,G_n)+(A_n,B_n)+o(1/\sqrt{n};\mathbb{P})$. Operating both sides by $\dot{\Psi}_C$, using that $\dot{\Psi}_C$ is continuous, and using the definition of (A_n,B_n) we obtain

$$\begin{aligned}\dot{\Psi}_C\begin{pmatrix}\hat{F}_n-F_0\\ \hat{G}_n-G_0\end{pmatrix} &= \dot{\Psi}_C\begin{pmatrix}F_n-F_0\\ G_n-G_0\end{pmatrix}+\dot{\Psi}_C\begin{pmatrix}A_n\\ B_n\end{pmatrix}+o(1/\sqrt{n};\mathbb{P})\\ &= \dot{\Psi}_C\begin{pmatrix}F_n-F_0\\ G_n-G_0\end{pmatrix}+\tilde{\mathbb{S}}_n+o(1/\sqrt{n};\mathbb{P}).\end{aligned}$$

From the definition of $\dot{\Psi}_C$ we obtain, for $u\in[0,1]$,

$$\begin{aligned}\dot{\Psi}_C^1(F_n-F_0,G_n-G_0)(u) &= -\left(\frac{1}{n}\sum_{i=1}^n 1\{X_i\le u\}-u\right)\\ &\quad+\int (x\wedge u-xu)\left\{(F_n(x)-x)\ddot{\ell}_{11}(x,y)+(G_n(y)-y)\ddot{\ell}_{12}(x,y)\right\}\mathrm{d}C(x,y).\end{aligned}$$

So, by Proposition 5.4.2 the proof is complete once we show that

$$\sup_{u\in[0,1]}\sqrt{n}\left|\int (x\wedge u-xu)(F_n(x)-x)\ddot{\ell}_{11}(x,y)\,\mathrm{d}(H_n-C)(x,y)\right|\xrightarrow{p}0, \quad (5.33)$$

and,

$$\sup_{u\in[0,1]}\sqrt{n}\left|\int (x\wedge u-xu)(G_n(y)-y)\ddot{\ell}_{12}(x,y)\,\mathrm{d}(H_n-C)(x,y)\right|\xrightarrow{p}0. \quad (5.34)$$

We prove (5.33); (5.34) follows by similar arguments. Let $\mathscr{H}$ the class of distributions on $[0,1]^2$ for which $\int |\ddot{\ell}_{11}(x,y)|\,\mathrm{d}H(x,y) < \int |\ddot{\ell}_{11}(x,y)|\,\mathrm{d}C(x,y)+1$. Notice first that we have,

$$\begin{aligned}
\sqrt{n}\int & (x\wedge u - xu)(F_n(x)-x)\ddot{\ell}_{11}(x,y)\,\mathrm{d}(H_n - C)(x,y) \\
&= \sqrt{n}\int\int (x\wedge u - xu)\ddot{\ell}_{11}(x,y)1_{[0,x]}(z)\,\mathrm{d}(H_n-C)(x,y)\,\mathrm{d}(F_n-F_0)(z) \\
&= \sqrt{n}\int f_u^{H_n}(z)\,\mathrm{d}(F_n-F_0)(z),
\end{aligned}$$

with, for $u\in[0,1]$ and $H\in\mathscr{H}$, $f_u^H:[0,1]\to\mathbb{R}$ defined by

$$f_u^H(z) = \int (x\wedge u - xu)\ddot{\ell}_{11}(x,y)1_{[0,x]}(z)\,\mathrm{d}(H-C)(x,y).$$

And let $\mathscr{A} = \{f_u^H \mid u\in[0,1], H\in\mathscr{H}\}$. From the bound

$$|f_u^H(z) - f_{u'}^H(z)| \le 2(2\int |\ddot{\ell}_{11}(x,y)|\,\mathrm{d}C(x,y)+1)|u-u'|$$

it follows (see, e.g., Example 19.7 in Van der Vaart (2000)) that $\mathscr{A}$ is a F_0-Donsker class. Using the law of large numbers it follows that $\mathbb{P}\{H_n\in\mathscr{H}\}\to 1$, so $\mathbb{P}\{\forall\, u\in[0,1]: f_u^{H_n}\in\mathscr{A}\}\to 1$. It is easy to show that the class of functions $\mathscr{B} = \{(x,y)\mapsto \ddot{\ell}_{11}(x,y)(u\wedge x - ux)1_{[0,x]}(z) \mid u,z\in[0,1]\}$ is a $\mathbb{P}$-Glivenko-Cantelli class, which implies

$$\int_0^1 \sup_{u\in[0,1]} (f_u^{H_n}(z))^2\,\mathrm{d}z \le (\sup_{b\in\mathscr{B}} |\int b\,\mathrm{d}(F_n-F_0)|)^2 \xrightarrow{p} 0.$$

An application of Lemma 5.3.5 now yields (5.33). □

Efficiency proof 5.4.3

This section is the analogue of Section 5.3.4. We prove that our estimator $(\hat{F}_n,\hat{G}_n)$ is efficient. The arguments are completely similar to the proof of Theorem 5.5. The only complication is that we have to show that the artificial parameters we use are indeed well-defined. In Section 5.3.4 the artificial parameters were automatically well-defined, because there we dealt with copulas for which there are no problems on the boundary of $[0,1]^2$.

Theorem 5.8. *Let C a copula satisfying Assumptions (C1)-(C3). Then the estimator $(\hat{F}_n, G_n)_{n\in\mathbb{N}}$ is an efficient estimator of the parameter (F,G) in the model $\mathscr{P}(C)$.*

Proof.
Fix $(F_0, G_0) \in \mathscr{F}_{\mathrm{ac}}$. The proof is completely analogous to the proof of Theorem 5.5. As mentioned above we should only verify that the artificial parameters (the expectation of certain scores calculated *at* $\mathbb{P}_{F_0,G_0}$) are well-defined. Actually, it suffices to prove that the artificial parameter is well-defined for the paths generating the tangent space. We consider the artificial parameter for $\hat{F}_n$. Fix $v, w \in \mathrm{L}_2^0(\mathrm{Un}[0,1])$ and use the paths $t \mapsto F_t^v$ and $t \mapsto G_t^w$ through F_0 and G_0 (see Section 5.4.1). This yields the path $t \mapsto \mathbb{P}_t = \mathbb{P}_{F_t^v, G_t^w}$ through $\mathbb{P}_{F_0,G_0}$. Fix $u \in (0,1)$. We have to verify whether, at least for small t, the mapping

$$t \mapsto \mathbb{E}_{\mathbb{P}_t}\left[1\{F_0(X) \leq u\} - u + \dot{\ell}_1(F_0(X), G_0(Y))(F_0(X \wedge u) - F_0(X)F_0(u))\right],$$

is well-defined. Of course, the term $1\{F_0(X) \leq u\} - u$ does not give any problems. We have to deal with the second part. Since $(F_0(X \wedge u) - F_0(X)u) \leq F_0(X)(1 - F_0(X))$ we obtain from Assumption (C3)

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_t}\left|\dot{\ell}_1(F_0(X), G_0(Y))(F_0(X \wedge u) - F_0(X)F_0(u))\right| &\leq M_u \mathbb{E}_{G_t} \frac{1}{(G_0(Y)(1 - G_0(Y)))^\alpha} \\
&= M_u \int_{\mathbb{R}} c_{g_0}^w(t) k(tw(G_0(y))) g_0(y) \frac{1}{(G_0(y)(1 - G_0(y)))^\alpha}\, \mathrm{d}y \\
&= c_{g_0}^w(t) \int_0^1 k(tw(q)) \frac{1}{(q(1-q))^\alpha}\, \mathrm{d}q < 2c_{g_0}^w(t) \int_0^1 \frac{1}{(q(1-q))^\alpha}\, \mathrm{d}q < \infty.
\end{aligned}$$

□

6 Semiparametric efficiency bounds for time series models with non-i.i.d. innovations

This chapter derives semiparametric efficiency bounds for parametric components in general semiparametric time series models. The time series models are not assumed to be driven by a sequence of independent innovations with unknown distribution as is the case in the usual semiparametric time series approach. Instead of this, the dependence between the innovations is seen as a nonparametric nuisance parameter in addition to the marginal distribution of the innovations. We obtain a Local Asymptotic Normality (LAN) result under quite natural and economical conditions implying a lower bound on the asymptotic performance of (regular) estimators.

6.1 Introduction

The availability of large data sets is rapidly growing, especially in finance, as well as computing power to analyze them. If these data are confronted with classical parametric financial and econometric models, it is clear that these are misspecified. Semiparametric and nonparametric models are popular alternatives. Usually, from a practitioners point of view, some finite dimensional parameter is of interest. For example the mean or median as a measure of location, the Value at Risk as a measure of risk, etcetera. The question arises how to efficiently estimate such quantities in general semi- and nonparametric models. Quite often, the efficiency issue is considered as being less important given the enormous amount of data. While this may be the case in simple parametric models, the standard deviations of simple, for example moment type, estimators will substantially increase in general time series models due to the presence of infinite-dimensional nuisance parameters. These standard deviations can sometimes be substantially reduced using a semiparametrically efficient estimator[1].

To study what is best asymptotically in a semiparametric model, one needs a bound on the asymptotic performance of estimators in the presence of the infinite-dimensional nuisance parameter. Many models enjoy the Local Asymptotic Normality (LAN) property. Then the Hájek-Le Cam convolution theorem yields a bound to the precision of regular estimators. For the i.i.d. case, accounts on the present theory along these lines are Bickel et al. (1998) and Van der Vaart (2000, Chapter 25). Survey papers in an econometric setting are, for example, Robinson (1988), Newey (1990) and Stoker (1991). In financial data, of course, the time dimension also plays an important role. Drost et al. (1997) and Koul and Schick (1997) have developed a unified theory for time series models with independently and identically distributed innovations. This covers, for example, semiparametric ARMA models (Kreiss (1987a, 1987b)) and semiparametric GARCH models (Engle and Gonzalez-Rivera (1991), Linton (1993), and Drost and Klaassen (1997)). Steigerwald (1992) and Jeganathan (1995) have also obtained results for more general time series. Efficient rank-based inference for semiparametric time series models with i.i.d. innovations was considered by Hallin and Werker (2001) and Hallin et al. (2004). And Wefelmeyer (1996) obtained efficiency results in a Markovian context when only some moments are given and the innovations are assumed to be martingale differences.

Recent work in applied financial econometrics shows that the assumption of i.i.d. innovations does not hold when using standard semiparametric time series models, see Engle (2000), Drost and Werker (2004), and Gouriéroux et al. (2004). Volatility is for example time varying. Usually GARCH type models or stochastic volatility models are quite suitable to pick up the time-varying nature of the first two conditional moments with only a few additional parameters. However, the implications of this parametric model of volatility for higher order conditional moments are not reflected in the data. More precisely formulated, the conditional distribution of the errors cannot be described by just a functional form of the conditional volatility and a fixed nonparametric distribution. The description of the conditional distribution of the innovations is more delicate. To cover this problem we will take a more general approach by taking the whole distribution of the errors as a nuisance parameter. It is the purpose of this chapter to infer how the nonparametric nature of the conditional error distribution influences the estimation problem for the parametric component. In the most general case innovations are just martingales, nothing is assumed on the conditional distribution of the innovations. On the other extreme side, we have the 'classical' semiparametric time series model with i.i.d.

[1] In special occasions it is even possible to estimate the Euclidean parameter of interest adaptively.

innovations (the conditional distribution of the innovations is constant). One can imagine several cases in between where the conditional distribution of the innovations is not completely free.

In Section 6.2, we introduce the time series model in its general form by describing the possible dependence structures of the innovations. The observations will be obtained from the observations via an adapted time-varying group operator. Examples are the location-group, the scale-group, and the location-scale group. The regularity conditions needed for the efficiency result, are outlined in Section 6.3. These assumptions are related to the assumptions in time series models with i.i.d. innovations. Our assumptions for the general class of time series models (build from possibly *dependent* innovations) are minimal in the sense that they reduce to the assumptions for the i.i.d. case if attention is restricted to those models. The main results are presented in Section 6.4. We present a Local Asymptotic Normality (LAN) Theorem in case the infinite-dimensional nuisance parameter concerning the error distribution is known. From this we derive the efficient score for the semiparametric problem. According to the Hájek-Le Cam Convolution Theorem regular estimators which attain this bound are efficient. Loosely speaking this means that every other asymptotically normal estimator will have a larger variance than this lower bound. The proofs are based on a general LAN result which we recall in Appendix 6.5. Of course, the bound is only of value if one can construct an estimator attaining this bound. Since a general construction is extremely difficult in a time series setting, this chapter fully focuses on the derivation of a lower-bound.

Setup 6.2

This section (extensively) describes the model for a sequence of observations $(Y_t)_{t\in\mathbb{N}}$. To simplify exposition we will start by explaining the innovation structure underlying the model in Section 6.2.1. In Section 6.2.2 we build the observations $(Y_t)_{t\in\mathbb{N}}$ from the innovations by an adapted time-varying group operator. Special cases are the location-, the scale-, and the location-scale-group. This (structural) construction induces the probability distribution of the observations[2].

Innovation structure 6.2.1

Let X be the set of exogenous random variables, whose law is allowed to depend on the parameters, and let $(\varepsilon_t)_{t\in\mathbb{N}}$ be some sequence of innovations. Let $\mathscr{F}_0 = \sigma(X)$ denote the information set generated by the exogenous variables

[2] Of course, one could alternatively immediately define the probability measures for the observation process and, as a consequence, derive the implications for the innovations.

and let $\mathscr{F}_t = \sigma(X, \varepsilon_1, \ldots, \varepsilon_t)$ be the σ-field corresponding to the information at time t, $t \geq 1$.

As in any model, we have to describe the distributional structure of the innovation sequence $(\varepsilon_t)_{t\in\mathbb{N}}$. This will be done via the conditional distribution of ε_{t+1} given the information $\mathscr{F}_t$ until time $t \in \mathbb{Z}_+$. To make the role of the dependence structure of the innovation process in the efficiency analysis visible, we can assume, without loss of generality, that some sub σ-field $\mathscr{H}_t$ of $\mathscr{F}_t$ is given such that the conditional distribution of ε_{t+1} given the (smaller) σ-field $\mathscr{H}_t$ coincides with the conditional distribution given all information $\mathscr{F}_t$ available at time t. So, the additional information contained in $\mathscr{F}_t$ is not helpful to determine a better forecast of the error distribution of ε_{t+1}. Note that we do not require that the sequence $(\mathscr{H}_t)_{t\in\mathbb{Z}_+}$ is increasing like $(\mathscr{F}_t)_{t\in\mathbb{Z}_+}$, we merely impose that $\mathscr{H}_t$ is a sub σ-field of $\mathscr{F}_t$, $t \in \mathbb{Z}_+$. Before providing some examples, we introduce some additional notation to describe this requirement on the conditional error distributions:

$$G_t = \mathscr{L}(\varepsilon_{t+1}|\mathscr{F}_t) = \mathscr{L}(\varepsilon_{t+1}|\mathscr{H}_t) \in \mathscr{G} \text{ (a.s.)}, \; t \in \mathbb{Z}_+. \tag{6.1}$$

Here $\mathscr{G}$ is a given class of distributions and $(\mathscr{H}_t)_{t\in\mathbb{Z}_+}$, with $\mathscr{H}_t \subset \mathscr{F}_t$, is a given sequence of σ-fields. The conditional distributions given in (6.1) are not restricted, yet. However, to be able to derive lower bounds, we will need some assumptions that avoid a large difference between conditional distributions if the conditioning variables (including the parameters) are close. The assumptions will be presented in Section 6.3 once we have completed the statistical model of our observations. To illustrate the generality of this innovation structure we give some examples.

1. Traditional models with independent error sequences $(\varepsilon_t)_{t\in\mathbb{N}}$ are obtained by taking $\mathscr{H}_t = \{\emptyset, \Omega\}$. In this way the conditional distributions specified in (6.1) do not depend on past observations. Requiring stationarity yields i.i.d.-ness.

2. Quite another model will be obtained by letting $\mathscr{H}_t$ be as large as possible, that is $\mathscr{H}_t = \mathscr{F}_t$. If $\mathscr{G}$ consists of zero mean distributions this results in a model with martingale difference innovations.

3. As a model somewhere in between these two extremes consider a Markovian setting, where the conditional error distribution is only allowed to depend on the last observation. In this model we may take $\mathscr{H}_t = \sigma(\varepsilon_t)$.

Just as in classical semiparametric problems, the class of distributions $\mathscr{G}$ will be large typically, although we do not exclude a parametric class of distributions. For example, one might consider the class of all zero mean distributions (satisfying some weak differentiability and integrability conditions). In semiparametric time series models with i.i.d. innovations *one* unknown element $G \in \mathscr{G}$, the marginal distribution of the innovations, serves as a nonparametric nuisance parameter. The present set-up is much more complicated since, at each point in time, we may pick, depending upon past observations, another distribution from $\mathscr{G}$ (however, if one requires stationarity of $(\varepsilon_t)_{t\in\mathbb{Z}_+}$, then G_t cannot be chosen arbitrarily). In a (time-homogenous) Markovian context we actually only have to deal with two (infinite-dimensional) parameters: the law of ε_1 (conditional on X) and the transition-probability operator. In general, the sequence of conditional error distributions $(G_t)_{t\in\mathbb{Z}_+}$ will serve as the nonparametric nuisance parameter in our semiparametric model of the observations. In a semiparametric setting parts of this parameter may have to be estimated from the data to obtain an estimate of the efficient score function. Of course, imposing stationarity of the innovation process can help to accomplish this.

Summarizing: the general description of the error structure allows us to study a big variety of time series models, including models with i.i.d. errors, models with martingale difference innovations, Markovian innovations, and all kinds of situations in between.

Group structure 6.2.2

Having described the structure of the innovations in our statistical model, we describe how the observations are constructed from these innovations. To build our observations $(Y_t)_{t\in\mathbb{N}}$ from the real-valued innovation sequence $(\varepsilon_t)_{t\in\mathbb{N}}$ let a group of measurable transformations be given, $\{a_u \mid u \in \mathscr{U}\}$, where $a_u : \mathbb{R} \to \mathbb{R}$, and $\mathscr{U} \subset \mathbb{R}^m$ is open. Let u_e be a unitary element, i.e. $a_{u_e}(e) = e$. It is assumed throughout that this group satisfies the following smoothness conditions.

Assumption 1. The group $\{a_u \mid u \in \mathscr{U}\}$ with $\mathscr{U} \subset \mathbb{R}^m$ open satisfies the following conditions.

(a) The mapping $u \mapsto a_u$ is one-to-one. Consequently, there exists for all $u_0, u_1 \in \mathscr{U}$ a unique element $\omega(u_0, u_1) \in \mathscr{U}$ such that $a_{\omega(u_0,u_1)} = a_{u_0}^{-1} \circ a_{u_1}$. Notice that u_e is unique and that $\omega(u, u) = u_e$.

(b) For each $u_0 \in \mathscr{U}$ the mapping $\mathscr{U} \ni u \mapsto \omega(u_0, u)$ is continuously differentiable at u_0 with derivative $\dot\omega(u_0) \in \mathbb{R}^{m\times m}$, i.e. $|\omega(u_0, u+h) - u_e - \dot\omega(u_0)h| = o(|h|)$ as $|h| \to 0$.

(c) $\dot\omega(u_e) = I$, i.e. the $m \times m$ identity matrix.

(d) With λ denoting Lebesgue measure, the measure $\lambda \circ a_u^{-1}$ is equivalent to λ. Denote

$$j(e; u) = \frac{\mathrm{d}\left(\lambda \circ a_u^{-1}\right)}{\mathrm{d}\lambda}(e).$$

Remark 1. Actually, Assumptions (a)-(b) can be replaced by the weaker, but bit more complex, Conditions (i)a-c on pages 90-91 in Bickel et al. (1998) which allow for locally invertible parametrizations $u \mapsto a_u$. Assumption (c) is not restrictive, since we can achieve it after a reparametrization.

Examples 1-3 below satisfy Assumption 1. Let $(U_t(\theta))_{t\in\mathbb{Z}_+}$, be a sequence of $\mathscr{U}$-valued, $\mathscr{F}_t$-adapted random vectors. Here $\theta \in \Theta$, with $\Theta \subset \mathbb{R}^k$ open, is our parameter of interest, while the sequence of conditional distributions $(G_t)_{t\in\mathbb{Z}_+}$ of the innovations will be our nuisance parameter. The observations are defined via

$$Y_{t+1} = a_{U_t(\theta)}(\varepsilon_{t+1}), \quad t \in \mathbb{Z}_+. \tag{6.2}$$

In (financial) econometrics we very often encounter the following examples of the group-model.

1. *Location group*: $a_u : \mathbb{R} \to \mathbb{R}$ defined by $a_u(e) = u + e$, $\mathscr{U} = \mathbb{R}$. This yields $Y_{t+1} = U_t(\theta) + \varepsilon_{t+1}$.
2. *Scale group*: $a_u : \mathbb{R} \to \mathbb{R}$ defined by $a_u(e) = ue$, $\mathscr{U} = \mathbb{R}_{++}$. This yields $Y_{t+1} = U_t(\theta)\varepsilon_{t+1}$.
3. *Location-scale group*: $a_{u_1,u_2} : \mathbb{R} \to \mathbb{R}$ defined by $a_{u_1,u_2}(e) = u_1 + u_2 e$, $\mathscr{U} = \mathbb{R} \times \mathbb{R}_{++}$. This yields $Y_{t+1} = U_{t1}(\theta) + U_{t2}(\theta)\varepsilon_{t+1}$.

From (6.1) and (6.2) it is clear that the conditional distribution of Y_{t+1} given $\mathscr{F}_t$ is given by

$$P_{U_t(\theta),G_t} = \mathscr{L}(Y_{t+1}|\mathscr{F}_t) = G_t \circ a^{-1}_{U_t(\theta)} \text{ (a.s.)}, \ t \in \mathbb{Z}_+. \tag{6.3}$$

If G_t has density g_t (with respect to Lebesgue measure) then (under Part (d) of Assumption 1) the density of $P_{U_t(\theta),G_t}$ is given by

$$p_t(y; U_t(\theta), G_t) = g_t(a^{-1}_{U_t(\theta)}(y))\, j(y; U_t(\theta)). \tag{6.4}$$

For the location-, the scale- and the location-scale-group $p_t(\cdot; U_t(\theta), G_t)$ is respectively given by

$$g_t(\cdot - U_t(\theta)), \quad \frac{1}{U_t(\theta)} g_t\left(\frac{\cdot}{U_t\theta}\right), \text{ and } \frac{1}{U_{t2}(\theta)} g_t\left(\frac{\cdot - U_{t1}(\theta)}{U_{t2}(\theta)}\right).$$

For a fixed absolutely continuous distribution G with differentiable density the standard parametric group model

$$\left(G \circ a_u^{-1} \mid u \in \mathscr{U}\right), \tag{6.5}$$

has score given by (see pages 90-91 in Bickel et al. (1998)),

$$\phi(y; u, G) = \dot{\omega}^T(u)\,\dot{l}(a_u^{-1} y; G), \text{ with } \dot{l}(y; G) = \frac{\partial}{\partial u} \log\left(g(a_u^{-1} y)\, j(y; u)\right)\Big|_{u=u_e}.$$

For the location-, scale- and location-scale group the objects $\dot{l}$ and $\dot{\omega}$ are respectively given by,

$$
\begin{aligned}
\text{Location:}\quad & \dot{l}(y;G) = -\frac{g'}{g}(y), \text{ and } \dot{\omega}(u) = 1,\\
\text{Scale:}\quad & \dot{l}(y;G) = -\left(1 + y\frac{g'}{g}(y)\right), \text{ and } \dot{\omega}(u) = \frac{1}{u},\\
\text{Location-scale:}\quad & \dot{l}(y;G) = \begin{pmatrix} -\frac{g'}{g}(y) \\ -\left(1 + y\frac{g'}{g}(y)\right) \end{pmatrix} \quad \text{and } \dot{\omega}(u_1,u_2) = \frac{1}{u_2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\end{aligned}
$$

In our statistical model we thus have the flexibility of choosing both a suitable group structure and a suitable time varying parametrization of the group operator. If time would be fixed, this would induce an i.i.d. group model with $u \in \mathscr{U}$ as the parameter of interest and $G \in \mathscr{G}$ the unknown nuisance structure. Semiparametric i.i.d. group models have been studied in Bickel et al. (1998, pp. 88-103). Adaptive time series group models with i.i.d. innovations are discussed in Drost et al. (1997). Here we allow for time series group models with general innovation structures. This concludes the set-up of our statistical model. To enhance readability we conclude this section with examples of models that fit into our setup.

Examples 6.2.3

Example 6.1. *(Classical AR(1) model)*

$$Y_{t+1} = \theta Y_t + \varepsilon_{t+1}.$$

This is clearly a model with a location group structure and the time varying location parameter is given by $U_t(\theta) = \theta Y_t$. In a traditional parametric model one often assumes that the errors are i.i.d. normal, that is the set $\mathscr{G}$ consists of normal distributions centered around zero and the σ-field $\mathscr{H}_t$ is the trivial σ-field. A semiparametric model is obtained by enlarging the set $\mathscr{G}$ to the class of all (symmetric) distributions centered around zero. If the i.i.d. assumption is considered to be too restrictive, one may enlarge the σ-field $\mathscr{H}_t$. Take, for example, the set of all events that are Y_t-measurable, i.e. $\mathscr{H}_t = \sigma(Y_t)$.

Example 6.2. *(Random coefficient AR(1) model)*

$$Y_{t+1} = f(Y_t;\theta) Y_t + \varepsilon_{t+1}.$$

Just as in the previous model this is a location group model with the time varying location parameter given by $U_t(\theta) = f(Y_t;\theta) Y_t$. The choice between a parametric model or a semiparametric model or between i.i.d. innovations or, for example, martingale innovations can be treated similarly as in the previous example.

Example 6.3. *(AR(1)-ARCH(1) model)*

$$Y_{t+1} = \alpha Y_t + S_t \varepsilon_{t+1}, \quad \text{with} \quad S_t^2 = \psi + \beta Y_t^2.$$

In contrast to the previous two examples, we cannot formulate the present model as a location model, since in that case the imposed structure of the conditional variance cannot be recovered. Therefore, we take the location-scale operator as the group structure with $a_{u_1,u_2}(e) = u_1 + u_2 e$. The time varying location parameter is given by $U_{t1}(\alpha,\beta,\psi) = \alpha Y_t$ and the scale parameter is determined by $U_{t2}(\alpha,\beta,\psi) = S_t$. The remaining details can be treated in the same way as in the previous examples.

Remark 2. Our setup also contains models with independent but not identically distributed observations. However, a Bayesian setup where the infinite-dimensional nuisance parameter is random (see Bickel and Klaassen (1986)) is not contained in our setup.

6.3 Assumptions

To be able to derive an asymptotic bound on the performance of regular estimators of θ we will need several assumptions. This section discusses these assumptions in some detail. These assumptions will be used in Section 6.4.1 to prove that the log-likelihood ratios corresponding to observations of the process $(Y_t)_{t\in\mathbb{N}}$ are Locally Asymptotically Normal (LAN). This LAN-property will yield, via the Hájek-Le Cam convolution theorem, a notion of efficiency in parametric models where the nuisance parameter $(G_t)_{t\in\mathbb{Z}_+}$ is completely known. In turn this result is a key input in Section 6.4.2 where we obtain efficiency bounds for the semiparametric model.

The law of the process Y is determined by $\mathscr{L}(X)$, θ, and $(G_t)_{t\in\mathbb{Z}_+}$. It is allowed that $\mathscr{L}(X)$ depends on θ and $(G_t)_{t\in\mathbb{Z}_+}$. Since we work in this section with a fixed nuisance structure $(G_t)_{t\in\mathbb{Z}_+}$, we denote, for notational convenience, the underlying probability measure by $\mathbb{P}_\theta$.

In our first condition, we impose the technical assumption ensuring that exogenous and/or starting variables in $\mathscr{F}_0$ are indeed exogenous in the sense that they contain almost no information about the parameter θ.

Assumption 2. Let $\theta_0 \in \Theta$ and $\theta_n = \theta_0 + h_n/\sqrt{n}$ with $h_n \to h_0$. Let $\Lambda_n^X = \Lambda_n^X(\theta_n,\theta_0)$ denote the likelihood ratio of θ_n with respect to θ_0 of the law of the exogenous variables X. Then, under $\mathbb{P}_{\theta_0}$,

$$\Lambda_n^X \xrightarrow{p} 1, \quad \text{as } n \to \infty.$$

Remark 3. If the law of the exogenous variables X does not depend on θ, the assumption is trivially satisfied. For Markovian structures which are 'sufficiently ergodic' this assumption often can be verified using stability results for Markov chains (see Kartashov (1985)).

Recall from Subsection 6.2.2 that, if time would be fixed, model (6.3) yields the i.i.d. group model (6.5) with parameter $u \in \mathscr{U}$, with, under sufficient regularity, score given by $\varphi(y; u, G) = \dot{\omega}^T(u)\,\dot{l}\left(a_u^{-1} y; G\right)$. Assumption 3 requires that this core model is indeed regular (at each point in time) with respect to the parameter $u \in \mathscr{U}$.

Assumption 3. For each $G \in \mathscr{G}$, the model $Q_G = \left(G \circ a_u^{-1} \mid u \in \mathscr{U}\right)$ is regular with respect to Lebesgue measure (see Bickel et al. (1998, Chapter 2)) with score given by $\varphi(y; u, G)$ and positive definite Fisher-information given by $J(u; G) = \dot{\omega}^T(u) J(G) \dot{\omega}(u)$, where $J(G) = \int \dot{l}\dot{l}^T(e; G)\,\mathrm{d}G(e) \in \mathbb{R}^{m\times m}$.

Remark 4. If each $G \in \mathscr{G}$ has a differentiable density with $J(u; G)$ invertible then, by Lemma 4.2.1 in Bickel et al. (1998), Assumption 3 is satisfied.

Assumption 3 is the standard one in an i.i.d. semiparametric model without time-varying parameters/distributions and in our general set-up this assumption seems to be a natural starting point for our regularity conditions. Assumption 3 implies that the densities $p(\cdot; u, G)$ satisfy the following properties:

1. The square-root density $s(\cdot; u, G) = \sqrt{p(\cdot; u, G)}$ is Fréchet differentiable in $L_2(\lambda)$ with derivative $\dot{s}(\cdot; u, G) = \frac{1}{2}\varphi(\cdot; u, G)\sqrt{p(\cdot; u, G)}$, i.e.

$$\lim_{h\to 0} \frac{\int \left(s\left(y; u+h, G\right) - s\left(y; u, G\right) - h^T \dot{s}\left(y; u, G\right)\right)^2 \mathrm{d}y}{|h|^2} = 0. \tag{6.6}$$

2. The mapping $u \to \dot{s}(\cdot; u, G)$ from $\mathscr{U}$ to $L_2(\lambda)$ is continuous.

The function φ would be the score function for the parameter $u = U_t(\theta)$ in our statistical model if a random sample would be taken at a fixed time point t. However, we only have available one observation at each point in time and the parameter $u = U_t(\theta)$ is time-varying. Therefore it is not enough to consider the score function with respect to u. We have to infer the score function with respect to the parameter θ. The chain rule suggests that the score of observation $t+1$, should be given by, for $t \in \mathbb{Z}_+$,

$$\begin{aligned} S_{t+1}(\theta) &= \dot{U}_t(\theta)\,\varphi\left(Y_{t+1}; U_t(\theta), G_t\right) = \dot{V}_t(\theta)\,\dot{l}\left(a_{U_t(\theta)}^{-1} Y_{t+1}; G_t\right) \\ &= \dot{V}_t(\theta)\,\dot{l}\left(\varepsilon_{t+1}; G_t\right), \end{aligned} \tag{6.7}$$

where the $k \times m$ matrix $\dot{U}_t(\theta)$ is a (kind of) derivative of $U_t^T(\theta)$, see Assumption 5 below, and where $\dot{V}_t(\theta) = \dot{U}_t(\theta)\,\dot{\omega}^T\left(U_t(\theta)\right) \in \mathbb{R}^{k\times m}$. By regularity of the core model it is immediate that the total score is a martingale.

Proposition 6.3.1. Let $\theta_0 \in \Theta$. If the process $(\dot{V}_t(\theta_0))_{t\geq 0}$ is $\mathbb{P}_{\theta_0}$-square-integrable, then $(\tilde{S}_t(\theta_0))_{t\geq 0}$, defined by $\tilde{S}_0(\theta_0) = 0$ and $\tilde{S}_n(\theta_0) = \sum_{t=1}^n S_t(\theta_0)$, $n \in \mathbb{N}$, is a $\mathbb{P}_{\theta_0}$-martingale (w.r.t. $(\mathscr{F}_t)_{t\geq 0}$).

Proof. This easily follows by conditioning and using that the score in a regular parametric model (Assumption 3) has (conditional) mean zero. □

The next assumptions will assure that the score has a normal limiting distribution[3].

Assumption 4. Let $\theta_0 \in \Theta$. The following conditions hold under $\mathbb{P}_{\theta_0}$.

1. The process $(\dot{V}_t(\theta_0))_{t\geq 0}$ is $\mathbb{P}_{\theta_0}$-square-integrable.

2. The following law of large numbers holds, for a non-singular $k \times k$ matrix $I(\theta_0)$,

$$\frac{1}{n}\sum_{t=0}^{n} \dot{V}_t(\theta_0)\, J(G_t)\, \dot{V}_t^T(\theta_0) \xrightarrow{p} I(\theta_0). \tag{6.8}$$

3. For all $\epsilon > 0$ the following Lindeberg conditions hold,

$$\frac{1}{n}\sum_{t=0}^{n} \|\dot{V}_t(\theta_0) J^{1/2}(G_t)\|^2 1_{\{\|\dot{V}_t(\theta_0) J^{1/2}(G_t)\| > \epsilon\sqrt{n}\}} \xrightarrow{p} 0, \tag{6.9}$$

and,

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\theta_0}\left[|S_t(\theta_0)|^2 1_{\{|S_t(\theta_0)| > \epsilon\sqrt{n}\}} \mid \mathscr{F}_{t-1}\right] \xrightarrow{p} 0. \tag{6.10}$$

Note that Assumptions 1-4 are exactly the standard conditions in i.i.d. semiparametric group models. There these assumptions imply a LAN Theorem. It will be clear that we need some additional conditions to ensure that the time-varying nature of our problem will not disturb the LAN property. The additional assumptions we present below seem to be quite minimal since they are trivially met in the situation where u and G are not depending on time. We will try as much as possible to distinguish between differentiability conditions on the 'parameter' $U_t(\theta)$ (Assumption 5) and smoothness conditions on the conditional distributions G_t (Assumption 6). Nevertheless, the conditional Fisher information $J(G_t)$ will enter in all our conditions because the model does not exclude that either the Fisher information or its inverse has unbounded terms. In the proof of the main theorem, these two additional assumptions are needed to obtain (6.15). Alternatively one might replace both conditions by this equation

[3]For a $m \times n$ matrix A we denote $\|A\| = \sup\{|Ax| \mid |x| \leq 1\}$.

which merges the conditions on differentiability of the parameter and smoothness of the distribution. The next assumption imposes 'differentiability' conditions with respect to the time varying 'parameters' $\omega\left(U_t\left(\theta_0\right), U_t\left(\theta_n\right)\right)$ and $U_t\left(\theta_n\right)$. The pre-multiplication with $J^{1/2}\left(G_t\right)$ is done, as mentioned before, to correct for possible unboundedly growing Fisher information matrices.

Assumption 5. Let $\theta_0 \in \Theta$, and $\theta_n = \theta_0 + u_n/\sqrt{n}$ with $u_n \to u_0$. As $n \to \infty$ we have, with $\omega_{nt} = \omega(U_t(\theta_0), U_t(\theta_n))$,

$$\sum_{t=0}^{n} \left| J^{1/2}\left(G_t\right)\left(\omega_{nt} - u_e - \dot{\omega}\left(U_t\left(\theta_0\right)\right)\left(U_t\left(\theta_n\right) - U_t\left(\theta_0\right)\right)\right)\right|^2 = o(1; \mathbb{P}_{\theta_0}), \tag{6.11}$$

$$\sum_{t=0}^{n} \left| J^{1/2}\left(G_t\right)\dot{\omega}\left(U_t\left(\theta_0\right)\right)\left(U_t\left(\theta_n\right) - U_t\left(\theta_0\right) - \dot{U}_t^T\left(\theta_0\right)\left(\theta_n - \theta_0\right)\right)\right|^2 = o(1; \mathbb{P}_{\theta_0}), \tag{6.12}$$

where $\dot{U}_t^T\left(\theta_0\right)$ is implicitly defined as an appropriate differential of $U_t\left(\theta_0\right)$.

Remark 5. For the location-group (6.11) is trivially satisfied, since for this group $\omega(u_0, u_1) = u_1 - u_0$, $u_e = 0$, and $\dot{\omega}(u) = 1$. If $U_t(\theta)$ is linear in the parameters (6.12) is trivial by taking $\dot{U}_t(\theta)$ equal to the gradient of $U_t(\theta)$. For i.i.d. innovations (6.11) is satisfied, since, by the imposed group structure, we have, for each fixed $G \in \mathscr{G}$,

$$\left| J^{1/2}(G)\left(\omega\left(u, u+h\right) - u_e - \dot{\omega}\left(u\right)h\right)\right| = o\left(|h|\right).$$

Remark 6. If we combine (6.11) and (6.12) we obtain

$$\sum_{t=0}^{n} \left| J^{1/2}\left(G_t\right)\left(\omega_{nt} - u_e - \dot{\omega}\left(U_t\left(\theta_0\right)\right)\dot{U}_t^T(\theta_0)(\theta_n - \theta_0)\right)\right|^2 = o(1; \mathbb{P}_{\theta_0}). \tag{6.13}$$

The behavior of the time-varying conditional distribution functions is restricted by a uniformity condition on the function $\dot{s}$ in item 2 above. By the group character we need this condition only at $u = u_e$. Here deviations from the unitary element are not measured by the usual distance in the Euclidean space, but again the Fisher information matrix $J\left(G_t\right)$ is used to standardize.

Assumption 6. Let $\theta_0 \in \Theta$. Introduce, for $t \in \mathbb{Z}_+$ and $\delta > 0$,

$$M_t(\delta) = \sup_{|h| \le \delta} \int \left| J^{-1/2}\left(G_t\right)\left(\dot{s}\left(y; u_e + J^{-1/2}\left(G_t\right)h, G_t\right) - \dot{s}\left(y; u_e, G_t\right)\right)\right|^2 \mathrm{d}y.$$

There exists $\eta > 0$ and a $\mathbb{P}_{\theta_0}$-integrable variable B such that $\sup_{t\ge 0, 0\le\delta\le\eta} M_t(\delta) \le B$, and for $\delta_n \to 0$ we have $\sup_{0\le t\le n} M_t(\delta_n) \xrightarrow{p} 0$.

Remark 7. In case the innovations are i.i.d. the assumption is trivially satisfied since Assumption 3 implies that $u \mapsto \dot{s}(\cdot; u, G)$ is continuous.

6.4 Main Results

6.4.1 Parametric LAN theorem

Suppose that we observe $X, Y_1, \ldots, Y_n$. We are considering estimation of θ in the presence of the infinite dimensional nuisance structure $(G_t)_{t\in\mathbb{Z}_+}$. However, in this subsection we will fix this nuisance parameter and in the resulting parametric model we will derive a bound on the asymptotic performance of regular estimators of θ. In the next subsection we will discuss the consequences on the bound when the nuisance structure is unknown.

To be able to derive such a bound in the parametric model we have to show that the log-likelihood ratios of the observed random variables are locally asymptotically normal (LAN). Let $\mathbb{P}_\theta^{(n)}$ denote the law of $X, Y_1, \ldots, Y_n$ when the Euclidean parameter equals θ. The likelihood ratio statistic of the observations $X, Y_1, \ldots, Y_n$ for θ_n with respect to θ_0 is given by

$$\frac{\mathrm{d}\mathbb{P}_{\theta_n}^{(n)}}{\mathrm{d}\mathbb{P}_{\theta_0}^{(n)}} = \Lambda_n^X(\theta_n, \theta_0) \prod_{t=1}^n \frac{p_{t-1}(Y_t; U_{t-1}(\theta_n), G_{t-1})}{p_{t-1}(Y_t; U_{t-1}(\theta_0), G_{t-1})}.$$

The next theorem shows that these likelihood ratios are of the LAN form.

Theorem 6.1. *Under Assumptions* 1-6 *the statistical model defined by* (6.3) *satisfies the LAN condition with scores* (6.7)*, i.e. for* $\theta_0 \in \Theta$ *and* $\theta_n = \theta_0 + u_n/\sqrt{n}$ *with* $u_n \to u_0 \in \mathbb{R}^k$ *we have,*

$$\log \frac{\mathrm{d}\mathbb{P}_{\theta_n}^{(n)}}{\mathrm{d}\mathbb{P}_{\theta_0}^{(n)}} = \frac{u_0^T}{\sqrt{n}} \sum_{t=1}^n S_t(\theta_0) - \frac{1}{2} u_0^T I(\theta_0)\, u_0 + o\left(1; \mathbb{P}_{\theta_0}^{(n)}\right). \tag{6.14}$$

Proof.
We use Theorem 6.4 to prove the theorem. Set $\tilde{P}_n = \mathbb{P}_{\theta_n}^{(n)}$, $P_n = \mathbb{P}_{\theta_0}^{(n)}$, $\mathscr{F}_{nt} = \mathscr{F}_t$, $S_{nt} = S_t(\theta_0)$ for $t \geq 1$, and $h_n = \sqrt{n}(\theta_n - \theta_0) = u_n$. Notice that, in the notation of Theorem 6.4,

$$LR_{n0} = \Lambda_n^X(\theta_n, \theta_0), \text{ and for } 1 \leq t \leq n,\ LR_{nt} = \frac{p_{t-1}(Y_t; U_{t-1}(\theta_n), G_{t-1})}{p_{t-1}(Y_t; U_{t-1}(\theta_0), G_{t-1})}.$$

Thus the fourth condition of Theorem 6.4 is satisfied by Assumption 2. In the following we denote $p_t(\cdot; u) = p_t(\cdot; u, G_t)$, and use the same notational convention for other objects depending on G_t. For $1 \leq t \leq n$ we have,

$$R_{nt} = \sqrt{LR_{nt}} - 1 - \frac{1}{2} h_n^T \frac{S_{nt}}{\sqrt{n}} = \sqrt{\frac{p_{t-1}(Y_t; U_{t-1}(\theta_n))}{p_{t-1}(Y_t; U_{t-1}(\theta_0))}} - 1 - \frac{1}{2}(\theta_n - \theta_0)^T S_t(\theta_0).$$

First we check the first set of Assumptions (6.25)-(6.27). Condition (6.25) is immediate by Proposition 6.3.1. From Assumption 4 Condition (6.27) is immediate,

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{P_n}\left[S_{n,t}S_{n,t}^T \mid \mathscr{F}_{n,t-1}\right]=\frac{1}{n}\sum_{t=0}^{n-1}\dot V_t(\theta_0)J(G_t)\dot V_t^T(\theta_0)\xrightarrow{p} I(\theta_0)\,,\text{ as } n\to\infty,$$

and Condition (6.26) holds by Assumption 4. Next notice that,

$$\sum_{t=1}^{n}\mathbb{E}_{P_n}\left[R_{nt}^2 \mid \mathscr{F}_{n,t-1}\right]$$
$$=\sum_{t=0}^{n-1}\int_{A_t}\left|p_t^{1/2}\left(y;U_t(\theta_n)\right)-p_t^{1/2}\left(y;U_t(\theta_0)\right)-\frac{u_n^T}{\sqrt{n}}\dot U_t(\theta_0)\dot s_t\left(y;U_t(\theta_0)\right)\right|^2\mathrm{d}y.$$

with $A_t=\left\{p_t\left(\cdot;U_t(\theta_0)\right)>0\right\}$. And, since $\dot s_t\left(\cdot;U_t(\theta_0)\right)=0$ as $p_t\left(\cdot;U_t(\theta_0)\right)=0$, we have, with $B_t=\left\{p_t\left(\cdot;U_t(\theta_0)\right)=0\right\}$,

$$\sum_{t=1}^{n}\mathbb{E}_{P_n}\left(1-LR_{nt}\right)=\sum_{t=0}^{n-1}\int_{B_t}p_t\left(y;U_t(\theta_n)\right)\mathrm{d}y$$
$$=\sum_{t=0}^{n-1}\int_{B_t}\left|p_t^{1/2}\left(y;U_t(\theta_n)\right)-p_t^{1/2}\left(y;U_t(\theta_0)\right)-\frac{u_n^T}{\sqrt{n}}\dot U_t(\theta_0)\dot s_t\left(y;U_t(\theta_0)\right)\right|^2\mathrm{d}y.$$

Consequently, (6.28) and (6.29) will both follow once we prove that

$$\sum_{t=0}^{n}\int\left|s_t\left(U_t(\theta_n)\right)-s_t\left(U_t(\theta_0)\right)-\frac{u_n^T}{\sqrt{n}}\dot U_t(\theta_0)\dot s_t\left(U_t(\theta_0)\right)\right|^2(y)\,\mathrm{d}y=o\left(1;\mathbb{P}_{\theta_0}\right).\quad(6.15)$$

Using that (see the proof of Lemma 4.2.1 in Bickel et al. (1998)),

$$p_t(y;u)=p_t(a_{U_t(\theta_0)}^{-1}y;\omega(U_t(\theta_0),u))\,j(y,U_t(\theta_0)),$$

we obtain by the substitution $y=a_{U_t(\theta_0)}^{-1}z$, and using that $\omega(U_t(\theta_0),U_t(\theta_0))=u_e$, and $\dot\omega(u_e)=I$,

$$\sum_{t=0}^{n}\int\left|s_t\left(z;U_t(\theta_n)\right)-s_t\left(z;U_t(\theta_0)\right)-(\theta_n-\theta_0)^T\dot U_t(\theta_0)\dot s_t\left(z;U_t(\theta_0)\right)\right|^2\mathrm{d}z$$
$$=\sum_{t=0}^{n}\int\left|s_t\left(y;\omega_{nt}\right)-s_t\left(y;u_e\right)-(\theta_n-\theta_0)^T\dot U_t(\theta_0)\dot\omega^T(U_t(\theta_0))\dot s_t\left(y;u_e\right)\right|^2\mathrm{d}y,$$

with $\omega_{nt}=\omega(U_t(\theta_0),U_t(\theta_n))$. Plugging in $\dot s_t(y;u_e)=2^{-1}\dot l(y;G_t)s_t(y;u_e)$ we now obtain,

$$\leq 2\sum_{t=0}^{n}\int\left|s_t\left(y;\omega_{nt}\right)-s_t\left(y;u_e\right)-(\omega_{nt}-u_e)^T\dot s_t\left(y;u_e\right)\right|^2\mathrm{d}y$$

$$+\frac{1}{2}\sum_{t=0}^{n}\int\left|\left(\omega_{nt}-u_e-\dot\omega(U_t(\theta_0))\dot U_t^T(\theta_0)\,(\theta_n-\theta_0)\right)^T\dot l\left(y;G_t\right)s_t\left(y;u_e\right)\right|^2\mathrm{d}y,$$

using Assumption 3 and Proposition A.5.3 in Bickel et al. (1998) we now obtain,

$$=2\sum_{t=0}^{n}\int\left|(\omega_{nt}-u_e)^T\int_0^1\left\{\dot s_t\left(y;u_e+\lambda\left(\omega_{nt}-u_e\right)\right)-\dot s_t\left(y;u_e\right)\right\}\mathrm{d}\lambda\right|^2\mathrm{d}y$$
$$+\frac{1}{2}\sum_{t=0}^{n}\left|J^{1/2}(G_t)\left(\omega_{nt}-u_e-\dot\omega(U_t(\theta_0))\dot U_t^T(\theta_0)\,(\theta_n-\theta_0)\right)\right|^2$$

The second term on the right-hand side of (6.15) converges to zero in probability by Assumption 5 and (6.13). The proof of the theorem is complete once we show that the first term of this equation also converges to zero. Decompose,

$$|J^{1/2}(G_t)(\omega_{tn}-u_e)|\le|J^{1/2}(G_t)(\omega_{tn}-u_e-\dot\omega(U_t(\theta_0))\dot U_t^T(\theta_0)(\theta_n-\theta_0))|$$
$$+|J^{1/2}(G_t)\dot V_t^T(\theta_0)(\theta_n-\theta_0)|.$$

Now a combination of (6.8) and (6.13) immediately yields,

$$\sum_{t=0}^{n}|J^{1/2}(G_t)(\omega_{tn}-u_e)|^2=O(1;\mathbb{P}_{\theta_0}). \tag{6.16}$$

And a combination of (6.9) and (6.13) easily yields,

$$\max_{0\le t\le n}|J^{1/2}(G_t)(\omega_{tn}-u_e)|\xrightarrow{p}0. \tag{6.17}$$

For notational convenience we denote in the following $J_t=J(G_t)$. We have,

$$\sum_{t=0}^{n}\int\left|(\omega_{nt}-u_e)^T\int_0^1\left\{\dot s_t\left(y;u_e+\lambda\left(\omega_{nt}-u_e\right)\right)-\dot s_t\left(y;u_e\right)\right\}\mathrm{d}\lambda\right|^2\mathrm{d}y$$
$$\le\sum_{t=0}^{n}\left|J_t^{1/2}\left(\omega_{nt}-u_e\right)\right|^2\int\left|J_t^{-1/2}\int_0^1\left\{\dot s_t\left(y;u_e+\lambda\left(\omega_{nt}-u_e\right)\right)-\dot s_t\left(y;u_e\right)\right\}\mathrm{d}\lambda\right|^2\mathrm{d}y$$
$$\le\sum_{t=0}^{n}\left|J_t^{1/2}\left(\omega_{nt}-u_e\right)\right|^2\int_0^1\int\left|J_t^{-1/2}\left\{\dot s_t\left(y;u_e+\lambda\left(\omega_{nt}-u_e\right)\right)-\dot s_t\left(y;u_e\right)\right\}\right|^2\mathrm{d}y\,\mathrm{d}\lambda$$
$$\le\sup_{0\le t\le n}M_t\left(\max_{0\le t\le n}|J_t^{1/2}(\omega_{tn}-u_e)|\right)\sum_{t=0}^{n}\left|J^{1/2}(G_t)\left(\omega_{nt}-u_e\right)\right|^2,$$

with $M_t(\max_{0\le t\le n}|J_t^{1/2}(\omega_{tn}-u_e)|)$ from Assumption 6. Hence a combination of (6.16) and (6.17) with Assumption 6, and the dominated convergence theorem (for convergence in probability) yields,

$$\sum_{t=0}^{n}\int\left|(\omega_{nt}-u_e)^T\int_0^1\left\{\dot s_t\left(y;u_e+\lambda\left(\omega_{nt}-u_e\right)\right)-\dot s_t\left(y;u_e\right)\right\}\mathrm{d}\lambda\right|^2\mathrm{d}y\xrightarrow{p}0,$$

which concludes the proof. □

Recall that an estimator $t_n = t_n(X, Y_1, \dots, Y_n)$ of θ is regular if for all $\theta_0 \in \Theta$ and all $h_n \to h$ we have,

$$\sqrt{n}(t_n - (\theta_0 + h_n/\sqrt{n})) \xrightarrow{d} L_{\theta_0}, \quad \text{under } \mathbb{P}_{\theta_0 + h_n/\sqrt{n}}.$$

Regularity of an estimator can be interpreted as a kind of uniform convergence, in shrinking neighbourhoods, to the limiting distribution. Since we have obtained the LAN-property the Hájek-Le Cam convolution theorem holds.

Corollary 6.2. *Make the same assumptions as in the previous theorem. If $(t_n)_{n\in\mathbb{N}}$ is a regular estimator of θ then, under $\mathbb{P}_{\theta_0}$,*

$$\sqrt{n}(t_n - (\theta_0 + h_n/\sqrt{n})) \xrightarrow{d} L_{\theta_0} = \mathrm{N}(0, I^{-1}(\theta_0)) \oplus Z_{(t_n),\theta_0}.$$

Hence the limiting variance of a regular estimator is at least $I^{-1}(\theta)$, hence this gives a lower bound to the asymptotic precision of regular estimators.

Semiparametric lower bound 6.4.2

In the previous subsection we derived a lower bound for estimating θ in case the nuisance structure as defined in (6.1) is known. In this paragraph we investigate the influence of not knowing the sequence of conditional distribution functions $(G_t)_{t\in\mathbb{Z}_+}$ belonging to the set $\mathcal{G}$. Before presenting the mathematical details leading to a lower bound in the presence of this infinite dimensional nuisance structure, we give some intuition leading to the lower bound on the (asymptotic) variance of regular estimators for the parameter θ. The previous subsection showed that the lower bound, for estimation of θ, in the parametric model with the nuisance structure $(G_t)_{t\in\mathbb{Z}_+}$ known is determined by the central sequence $(S_t(\theta))_{t\in\mathbb{N}}$. In our heuristic calculation of the lower bound in the semiparametric problem, we have to project the elements of this central sequence onto the tangent space with respect to this unknown nuisance structure.

An important building block of the central sequence $(S_t(\theta))_{t\in\mathbb{N}}$ is the score $\phi(y; u, G_t) = \dot{\omega}^T(u)\dot{l}\left(a_u^{-1}y; G_t\right)$ of the parameter u in the group model with time fixed and $G_t \in \mathcal{G}$ known. In the semiparametric i.i.d. group model, given by $\left(G \circ a_u^{-1} \mid u \in \mathcal{U}, G \in \mathcal{G}\right)$, with $\varepsilon \sim G$ and $Y = a_u(\varepsilon)$, where $G \in \mathcal{G}$ is considered as an unknown nuisance parameter, the efficient score function at $u = u_e$, $\dot{l}^*(\varepsilon; G)$, takes over the role that the score function at $u = u_e$, $\dot{l}(\varepsilon; G)$, plays in the parametric i.i.d. group model. See Section 4.2 of Bickel et al. (1998) for a detailed discussion. Denote the tangent space[4] of this semiparametric i.i.d. group model by

[4] Let us, intuitively, recall the meaning of an element of this tangent space. Let $h: \mathbb{R} \to \mathbb{R}$ be bounded and such that $g_\eta(e) = g(e)(1 + \eta h(e))$ defines a density in $\mathcal{G}$. So in a neighborhood of 0 $\eta \mapsto g_\eta$ is a path in $\mathcal{G}$ that passes g at $\eta = 0$. This yields a path $\eta \mapsto p(y; u, G_\eta)$ through the semiparametric group model. The score of this path is given by, using (6.4), $(\partial/\partial\eta) \log p(Y; u, G_\eta)\big|_{t=0} = h(\varepsilon)$. This shows that the element $h(\varepsilon)$ belongs to the tangent space.

$\mathcal{T}_G$. This tangent space $\mathcal{T}_G$ is a $L_2(G)$-closed linear subspace of the space $\{h(\varepsilon) \mid \int h\,\mathrm{d}G = 0, \int h^2\,\mathrm{d}G < \infty\}$. By definition, the efficient score $\dot{\omega}^T(u)\dot{l}^*(\varepsilon; G)$ is the projection (in $L_2(G)$) of the score on the orthoplement of $\mathcal{T}_G$. Hence the projection of the score $\dot{\omega}^T(u)\dot{l}(\varepsilon; G)$ on $\mathcal{T}_G$ follows from $\Pi\left(\dot{l}(\varepsilon; G) \mid \mathcal{T}_G\right) = \dot{l}(\varepsilon; G) - \dot{l}^*(\varepsilon; G)$. Corresponding to the information at $u = u_e$, $J(G) = \int \dot{l}\dot{l}^T(e; G)\,\mathrm{d}G(e)$, we define the efficient information matrix at $u = u_e$, $J^*(G) = \int \dot{l}^*\dot{l}^{*T}(e; G)\,\mathrm{d}G(e)$. Since the location-scale group is of special interest in (financial) econometrics, we will provide these efficient score functions in the following example.

Example 6.4. We consider four cases, the pure location model, the pure scale model, the pure scale model with zero mean, and the location scale model,

$$\begin{aligned}
&Y_{t+1} = \mu_t + \varepsilon_{t+1}, \quad \int e\,\mathrm{d}G_t(e) = 0,\\
&Y_{t+1} = \sigma_t\varepsilon_{t+1}, \quad \int e^2\,\mathrm{d}G_t(e) = 1,\\
&Y_{t+1} = \sigma_t\varepsilon_{t+1}, \quad \int e\,\mathrm{d}G_t(e) = 0, \quad \int e^2\,\mathrm{d}G_t(e) = 1,\\
&Y_{t+1} = \mu_t + \sigma_t\varepsilon_{t+1}, \quad \int e\,\mathrm{d}G_t(e) = 0, \quad \int e^2\,\mathrm{d}G_t(e) = 1.
\end{aligned}$$

Recall that the scores for the location and/or the scale parameter at $u = u_e$ are given by $\dot{l}(e; G) = -(g'/g)(e)$ and $\dot{l}(e; G) = -\{1 + e(g'/g)(e)\}$, respectively. In these location-scale problems, where apart from the model assumption (6.1) and some regularity conditions, nothing is known about the conditional error distributions, the efficient score functions are uniquely defined by the aforementioned restrictions. Put $\sigma^2 = \mathbb{E}_G\varepsilon^2$, $\gamma = \mathbb{E}_G\varepsilon^3$, $\kappa = \mathbb{E}_G\varepsilon^4$. It is easily verified that the efficient score functions at $u = u_e$, $\dot{l}^*(\varepsilon; G)$, in the respective models are given by,

$$\begin{aligned}
\dot{l}^*(\varepsilon; G) &= \sigma^{-2}\varepsilon,\\
\dot{l}^*(\varepsilon; G) &= 2[\kappa - 1]^{-1}\left\{\varepsilon^2 - 1\right\},\\
\dot{l}^*(\varepsilon; G) &= 2\left[\kappa - \gamma^2 - 1\right]^{-1}\left\{\varepsilon^2 - \gamma\varepsilon - 1\right\},\\
\dot{l}^*(\varepsilon; G) &= \frac{1}{\kappa - 1 - \gamma^2}\begin{bmatrix} (\kappa - 1)\varepsilon - \gamma\left(\varepsilon^2 - 1\right) \\ 2(\varepsilon^2 - \gamma\varepsilon - 1) \end{bmatrix}.
\end{aligned}$$

The efficient information matrices $J^*(G_t)$ can be simply calculated by evaluating the expected outerproduct of the score under G_t. It follows that these matrces are given by the leading coefficients of the scores above. Sometimes more information is available about the conditional error distributions, for example it is known that they are symmetric about zero. This symmetry condition on the conditional error distributions also implies that the corresponding tangent space is restricted to symmetric functions. This also affects the projection of $\dot{l}(\varepsilon_{t+1}; G_t)$ onto this smaller tangent space. In the *symmetric* location problem there is even adaptivity i.e. $\dot{l}^*(\varepsilon_{t+1}; G_t) = \dot{l}(\varepsilon_{t+1}; G_t)$. The efficient score function in the *symmetric* scale problems do not alter since $\gamma_t = 0$ in the symmetric problem. In the *symmetric* location-scale problem the efficient score functions of the corresponding *symmetric* location problem and the *symmetric* scale problem should be stacked. In a similar manner one can treat higher order moment conditions, fixed moments, etc., see Bickel et al. (1998) for more examples.

In our semiparametric time series setting we have to project the component $S_{t+1}(\theta)$ of the central sequence on the tangent space at time $t+1$. Let us first discuss this tangent space. Fix $h_t(\varepsilon_{t+1}) \in \mathscr{T}_{G_t}$, where h_t is allowed to depend on the information in $\mathscr{H}_t$. Since $h_t(\varepsilon_{t+1})$ is a score, it is often possible to find a path $\eta \mapsto g_{\eta,t}$ in $\mathscr{G}$ which passes g_t at $\eta = 0$ and with $(\partial/\partial\eta)\log p_t(\varepsilon_{t+1}; U_t(\theta), G_\eta)\,|_{\eta=0} = h_t(\varepsilon_{t+1})$. Thus the time $t+1$ tangent space is given by (a subset[5] of) $\mathscr{T}_{t+1} = \{h_t(\varepsilon_{t+1}) \mid h_t \in \mathscr{T}_{G_t}\}$. Thus this tangent space is a subset of all zero mean, $\mathscr{H}_t$-measurable, square integrable functions of ε_{t+1}. We want to calculate the projection of $S_{t+1}(\theta)$, on this tangent space (in L_2). To this end, we first introduce

$$H_{t+1}(\varepsilon_{t+1},\theta) = \tilde{V}_t(\theta)\,\Pi\left(\dot{l}(\varepsilon_{t+1}; G_t) \mid \mathscr{T}_{G_t}\right) = \tilde{V}_t(\theta)\left(\dot{l}(\varepsilon_{t+1}; G_t) - \dot{l}^*(\varepsilon_{t+1}; G_t)\right),$$

with (the expectation taken under the parameters θ and $(G_t)_{t\in\mathbb{Z}_+}$),

$$\tilde{V}_t(\theta) = \mathbb{E}\left[\dot{V}_t(\theta)\,|\mathscr{H}_t\right].$$

Since $\tilde{V}_t(\theta)$ is $\mathscr{H}_t$-measurable and $\Pi\left(\dot{l}(\varepsilon_{t+1}; G_t) \mid \mathscr{T}_{G_t}\right)$ belongs (componentwise) to $\mathscr{T}_{t+1}$, is is clear that $H_{t+1}(\varepsilon_{t+1},\theta)$ is (componentwise) in $\mathscr{T}_{t+1}$. We now show that $H_{t+1}(\varepsilon_{t+1},\theta)$ is the projection of $S_{t+1}(\theta)$ onto $\mathscr{T}^*_{t+1}$. Introduce $S^*_t(\theta) = S_t(\theta) - H_t(\varepsilon_t,\theta)$. Decompose,

$$S^*_{t+1}(\theta) = (\dot{V}_t(\theta) - \tilde{V}_t(\theta))\dot{l}(\varepsilon_{t+1}; G_t) + \tilde{V}_t(\theta)\dot{l}^*(\varepsilon_{t+1}; G_t).$$

We will show that $S^*_{t+1}(\theta)$ is orthogonal (in L_2) to $\mathscr{T}_{t+1}$. Let $h_t(\varepsilon_{t+1}) \in \mathscr{T}_{t+1}$. In the following expectations are taken under the parameters θ and $(G_t)_{t\in\mathbb{Z}_+}$. Using that the conditional distribution of ε_{t+1} given $\mathscr{F}_t$ is given by G_t and since $h_t(\varepsilon_{t+1})$ is orthogonal to $\dot{l}^*(\varepsilon_{t+1}; G_t)$ in $L_2(G_t)$, we obtain

$$\mathbb{E}\left[h_t(\varepsilon_{t+1})\tilde{V}_t(\theta)\dot{l}^*(\varepsilon_{t+1}; G_t) \mid \mathscr{F}_t\right] = \tilde{V}_t(\theta)\int h_t(e)\dot{l}^*(e; G_t)\,\mathrm{d}G_t(e) = 0.$$

And we have,

$$\mathbb{E}h_t(\varepsilon_{t+1})(\dot{V}_t(\theta) - \tilde{V}_t(\theta))\dot{l}(\varepsilon_{t+1}; G_t) = \mathbb{E}(\dot{V}_t(\theta) - \tilde{V}_t(\theta))\mathbb{E}\left[h_t(\varepsilon_{t+1})\dot{l}(\varepsilon_{t+1}; G_t) \mid \mathscr{F}_t\right],$$

[5]If one puts stationarity of the innovation process into the model, then this puts extra restrictions on the conditional densities which may yield a smaller tangent space. This explains the inclusion.

using that h_t only depends on $\mathscr{F}_t$ via $\mathscr{H}_t$ and that the law of ε_{t+1} depends only on $\mathscr{F}_t$ via $\mathscr{H}_t$ we now obtain,

$$= \mathbb{E}(\dot{V}_t(\theta) - \tilde{V}_t(\theta))\mathbb{E}\left[h_t(\varepsilon_{t+1})\dot{l}(\varepsilon_{t+1};G_t) \mid \mathscr{H}_t\right]$$
$$= \mathbb{E}\mathbb{E}\left[\dot{V}_t(\theta) - \tilde{V}_t(\theta) \mid \mathscr{H}_t\right]\mathbb{E}\left[h_t(\varepsilon_{t+1})\dot{l}(\varepsilon_{t+1};G_t) \mid \mathscr{H}_t\right] = 0,$$

by definition of $\tilde{V}_t(\theta)$. Hence $S^*_{t+1}(\theta)$ is indeed orthogonal to $\mathscr{T}_{t+1}$. This determines $\left(S^*_t(\theta)\right)_{t\in\mathbb{N}}$ as central sequence for the semiparametric model and the corresponding lower bound is given by the inverse of the probability limit $I^*(\theta) = \text{plim}_{n\to\infty}\frac{1}{n}\sum_{t=1}^n S^*_t(\theta)\,S^*_t(\theta)^T$.

Remark 8. Note that $I(\theta) = I^*(\theta)$, which is a necessary condition for adaptive estimation, in case $\text{plim}_{n\to\infty}\frac{1}{n}\sum_{t=1}^n H_t(\varepsilon_t,\theta)H_t^T(\varepsilon_t,\theta) = 0$. This is certainly the case if $\tilde{V}_t(\theta) = 0$ for all t. As an example, this holds for an AR(1) model with i.i.d. mean zero innovations. Another sufficient condition for $I(\theta) = I^*(\theta)$ is $l^*(\varepsilon_{t+1};G_t) = l(\varepsilon_{t+1};G_t)$, which is, for example the case for the location group and $\mathscr{G}$ a subset of symmetric distrubutions centered around 0.

To formalize these heuristic arguments it suffices to construct a parametric least-favorable submodel. Let us recall what this means. First introduce, for $\theta \in \Theta$ and a nuisance structure $(G_t)_{t\in\mathbb{Z}_+}$, the probability measure $\mathbb{P}_{\theta,(G_t)}$ that generates the observations. If we can find, at a fixed $\mathbb{P}_{\theta,(G_t)}$, a parametric submodel of the semiparametric model that contains $\mathbb{P}_{\theta,(G_t)}$, and has the LAN-property at $\mathbb{P}_{\theta,(G_t)}$ with information bound for θ equal to $I^*(\theta)^{-1}$, then this submodel is least favorable, i.e. it determines the most difficult local direction for estimation of θ. Since an estimator in the semiparametric model is by definition regular if it is regular along all parametric submodels, it then immediately follows that $I^*(\theta)^{-1}$ is indeed a lower bound to the asymptotic variance of regular estimators of θ.

We will present high-level assumptions that yield such a least-favorable submodel. The first assumption essentially requires that the standard i.i.d. semiparametric group model has a least favorable submodel, and that the time-varying nature does not disturb this property.

Assumption 7. Let $\theta_0 \in \Theta$, and $(G_t)_{t\in\mathbb{Z}_+}$ such that $\mathbb{P}_{\theta_0,(G_t)}$ belongs to the semiparametric model. There exists $\epsilon > 0$ such that, for all $t \in \mathbb{Z}_+$, we can find a path $(-\epsilon,\epsilon) \ni \eta \mapsto g_{t,\eta}$ in $\mathscr{G}$ such that the following conditions hold.

1. At $\eta = 0$ the path passes g_t, i.e. $g_{t,0} = g_t$.

2. The path is an allowed nuisance structure in the semiparametric model: for all θ in a neighborhood of θ_0 and all $\eta \in (-\epsilon,\epsilon)$ the probability measure $\mathbb{P}_{\theta,(G_{t,\eta})}$ belongs to the semiparametric model.

3. For all $\eta_n = v_n/\sqrt{n}$ with $v_n \to v_0 \in \mathbb{R}^k$ the following expansion holds,

$$\sum_{t=0}^{n} \log \frac{g_{t,\eta_n}}{g_t}(\varepsilon_{t+1}) = \frac{v_0^T}{\sqrt{n}} \sum_{t=0}^{n} H_{t+1}(\varepsilon_{t+1};\theta_0) - \frac{1}{2n}\sum_{t=0}^{n}(v_0^T H_{t+1}(\varepsilon_{t+1};\theta_0))^2 + o(1;\mathbb{P}_{(G_t)}).$$

Remark 9. Items 1 and 3 require that the standard i.i.d. semiparametric group model has a least favorable submodel satisfying the LAN-property, and that the time varying nature does not disturb this LAN-expansion. If G_t may be chosen arbitrarily from $\mathscr{G}$ then item 2 is automatically satisfied. This is, for example, the case for i.i.d. innovations or for Markovian innovation structures for which one only makes assumptions on the transition-density. However, if one wants to put stationarity or mixing conditions on the innovations, item 2 needs to be checked.

Remark 10. In semiparametrics, one typically wants to have the set $\mathscr{G}$ as large as possible to avoid possible misspecifications. If $\mathscr{G}$ would consist of all densities we could take the densities

$$g_{t,\eta}(e) = c_t(\eta) g_t(e) \psi\left(\eta^T H_{t+1}(e,\theta_0)\right), \quad e \in \mathbb{R}, \eta \in \mathbb{R}, \tag{6.18}$$

where $c_t(\eta) = 1/\mathbb{E}_{G_t}\psi\left(\eta^T H_{t+1}(\varepsilon_{t+1},\theta_0)\right)$ is the constant such that the left-hand side is a (conditional) density and[6] $\psi(e) = 1 + \frac{\pi}{2}\arctan(e)$. Now item 3 usually follows by a second order Taylor expansion (the negligibility of the remainder term follows if $n^{-1}\sum_{t=1}^{n}|v_0^T H_{t+1}(\varepsilon_{t+1})|^2 = O(1;\mathbb{P}_{(G_t)})$). Quite often, some general restrictions on the set $\mathscr{G}$ are still useful or even necessary to be able to identify the finite dimensional parameter of interest. In the examples of Section 6.1 symmetry and/or moment conditions are mentioned. It is clear that a symmetry condition on the conditional densities is automatically transformed to the same symmetry condition on the score $H_t(\varepsilon_{t+1},\theta_0)$ and (6.18) thus remains a valid submodel. Moment conditions are more delicate since moment conditions are not necessarily preserved. This problem can be handled along the lines of Example 3 on pp.53–55 of Bickel et al. (1998).

Fix θ_0 and a nuisance structure $(G_t)_{t\in\mathbb{Z}_+}$. Using Assumption 7 we obtain a path $(\theta,\eta) \mapsto \mathbb{P}_{\theta,(G_{t,\eta})}$ in our semiparametric model which passes $\mathbb{P}_{\theta_0,(G_t)}$ at $(\theta,\eta) = (\theta_0, 0)$ (please note that this does not change the interpretation of θ, since we are still dealing with a group model). We will show that $\theta \mapsto \mathbb{P}_{\theta,(G_{t,\theta_0-\theta})}$ yields a least favorable submodel (at $\mathbb{P}_{\theta_0,(G_t)}$). We need some further regularity conditions to be able to derive the desired LAN property for this proposed submodel. Assumption 8 requires a law of large numbers and a central limit theorem for the efficient score. Often this can be verified by an application of the martingale

[6]Note that ψ is a bounded, smooth function $\psi(0) = \psi'(0) = 1, \psi''(0) = 0$, and ψ'/ψ bounded.

central limit theorem (see, for example, Hall and Heyde (1980)). Assumption 9 is a smoothness condition on the empirical efficient Fisher information. Verification can usually be done via a Taylor expansion of $H_{t+1}(a^{-1}_{U_t(\theta)}Y_{t+1},\theta_0)$ around θ_0. Condition (6.19) is a continuity condition, and Condition (6.20) is a first order expansion of $H_{t+1}(a^{-1}_{U_t(\theta)}Y_{t+1},\theta_0)$ where the form of the 'derivative' is based on the classical assumption that the Fisher information can be obtained by either taking the expectation of the outerproduct of the scores or by minus the expectation of the derivative of the score.

Assumption 8. Let θ_0 and $(G_t)_{t\in\mathbb{Z}_+}$ be such that $\mathbb{P}_{\theta_0,(G_t)}$ is an element of the semiparametric model. Under $\mathbb{P}_{\theta_0,(G_t)}$ the efficient score satisfies the weak law of large numbers,

$$\frac{1}{n}\sum_{t=1}^{n} S_t^*(\theta_0)S_t^*(\theta_0)^T \xrightarrow{p} I^*(\theta_0),$$

and the central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} S_t^*(\theta_0) \xrightarrow{d} \mathrm{N}(0, I^*(\theta_0)).$$

Assumption 9. Let θ_0 and $(G_t)_{t\in\mathbb{Z}_+}$ be such that $\mathbb{P}_{\theta_0,(G_t)}$ is an element of the semiparametric model. For all sequences $\theta_n = \theta_0 + u_n/\sqrt{n}$, with $u_n \to u_0$, we have

$$\frac{1}{n}\sum_{t=0}^{n}\left|H_{t+1}\left(a^{-1}_{U_t(\theta_n)}Y_{t+1},\theta_0\right) - H_{t+1}\left(a^{-1}_{U_t(\theta_0)}Y_{t+1},\theta_0\right)\right|^2 = o(1;\mathbb{P}_{\theta_0,(G_t)}), \tag{6.19}$$

and,

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{t=0}^{n}H_{t+1}\left(a^{-1}_{U_t(\theta_n)}Y_{t+1},\theta_0\right) &= \frac{1}{\sqrt{n}}\sum_{t=0}^{n}H_{t+1}\left(a^{-1}_{U_t(\theta_0)}Y_{t+1},\theta_0\right)\\ &\quad - \frac{1}{n}\sum_{t=0}^{n}H_{t+1}\left(a^{-1}_{U_t(\theta_0)}Y_{t+1},\theta_0\right)S_t^T(\theta_0)\,u_0 + o(1;\mathbb{P}_{\theta_0,(G_t)}).\end{aligned} \tag{6.20}$$

Finally, we prove that the submodel $\theta \mapsto \mathbb{P}_{\theta,(G_{t,\theta_0-\theta})}$ satisfies the LAN property at $\theta = \theta_0$ with information equal to $I^*(\theta_0)$, which implies, as discussed above, that $I^*(\theta)$ is indeed a lower bound to the asymptotic variance of regular estimators of θ. Let $\mathbb{P}^{(n)}_{\theta,\eta}$ denote the law of $X, Y_1,\dots,Y_n$ under $\mathbb{P}_{\theta,(G_{t,\eta})}$, and let $\mathbb{P}^{(n)}_{\theta} = \mathbb{P}^{(n)}_{\theta,\theta_0-\theta}$.

Theorem 6.3. *Let $\theta_0 \in \Theta$ and $(G_t)_{t\in\mathbb{Z}_+}$ such that $\mathbb{P}_{\theta_0,(G_t)_{t\in\mathbb{Z}_+}}$ belongs to the semiparametric model. Under Assumptions*[7] *1-9 the statistical submodel $\theta \mapsto \mathbb{P}_{\theta,(G_{t,\theta_0-\theta})}$,*

[7] Assume that Assumption 2 also holds, at $\mathbb{P}_{\theta_0,(G_t)}$, for the submodel $\theta \mapsto \mathbb{P}_{\theta,(G_{t,\theta_0-\theta})}$.

has the LAN property at $\theta = \theta_0$, i.e. for all sequences $\theta_n = \theta_0 + u_n/\sqrt{n}$, with $u_n \to u_0$, we have

$$\log \frac{d\mathbb{P}^{(n)}_{\theta_n}}{d\mathbb{P}^{(n)}_{\theta_0}} = \frac{u_0^T}{\sqrt{n}} \sum_{t=1}^{n} S_t^*(\theta_0) - \frac{1}{2n} \sum_{t=1}^{n} \left(u_0^T S_t^*(\theta_0)\right)^2 + o\left(1; \mathbb{P}^{(n)}_{\theta_0}\right)$$
$$\xrightarrow{d} \mathrm{N}\left(-\frac{1}{2} u_0^T I^*(\theta_0) u_0, u_0^T I^*(\theta_0) u_0\right).$$

Proof.
Notice first that, with $\eta_n = \theta_0 - \theta_n = -u_n/\sqrt{n}$,

$$\log \frac{d\mathbb{P}^{(n)}_{\theta_n}}{d\mathbb{P}^{(n)}_{\theta_0}} = \log \frac{d\mathbb{P}^{(n)}_{\theta_n,\eta_n}}{d\mathbb{P}^{(n)}_{\theta_0,0}} = \log \frac{d\mathbb{P}^{(n)}_{\theta_n,0}}{d\mathbb{P}^{(n)}_{\theta_0,0}} + \log \frac{d\mathbb{P}^{(n)}_{\theta_n,\eta_n}}{d\mathbb{P}^{(n)}_{\theta_n,0}}. \tag{6.21}$$

From Theorem 6.1 we have,

$$\log \frac{d\mathbb{P}^{(n)}_{\theta_n,0}}{d\mathbb{P}^{(n)}_{\theta_0,0}} = \frac{u_0^T}{\sqrt{n}} \sum_{t=1}^{n} S_t(\theta_0) - \frac{1}{2n} \sum_{t=1}^{n} (u_0^T S_t(\theta_0))^2 + o\left(1; \mathbb{P}^{(n)}_{\theta_0}\right). \tag{6.22}$$

Under $\mathbb{P}_{\theta_n,0}$ we have, using Assumption 2, the identity $\varepsilon_{t+1} = a^{-1}_{U_t(\theta_n)} Y_{t+1}$ and Assumption 7,

$$\begin{aligned}
\log \frac{d\mathbb{P}^{(n)}_{\theta_n,\eta_n}}{d\mathbb{P}^{(n)}_{\theta_n,0}} &= o(1; \mathbb{P}^{(n)}_{\theta_n}) + \sum_{t=0}^{n-1} \log \frac{g_{t,\eta_n}(a^{-1}_{U_t(\theta_n)} Y_{t+1})}{g_t(a^{-1}_{U_t(\theta_n)} Y_{t+1})} \\
&= o(1; \mathbb{P}^{(n)}_{\theta_n}) - \frac{u_0^T}{\sqrt{n}} \sum_{t=0}^{n-1} H_{t+1}(a^{-1}_{U_t(\theta_n)} Y_{t+1}; \theta_0) \\
&\quad - \frac{1}{2n} \sum_{t=0}^{n-1} (u_0^T H_{t+1}(a^{-1}_{U_t(\theta_n)} Y_{t+1}; \theta_0))^2.
\end{aligned}$$

By a combination of Le Cam's first lemma with Theorem 6.1, we may replace, in the previous display, the term $o(1; \mathbb{P}^{(n)}_{\theta_n})$ by $o(1; \mathbb{P}^{(n)}_{\theta_0})$. Using Assumption 9 we now obtain

$$\begin{aligned}
\log \frac{d\mathbb{P}^{(n)}_{\theta_n,\eta_n}}{d\mathbb{P}^{(n)}_{\theta_n,0}} &= -\frac{u_0^T}{\sqrt{n}} \sum_{t=0}^{n-1} H_{t+1}\left(a^{-1}_{U_t(\theta_0)} Y_{t+1}, \theta_0\right) - \frac{1}{2n} \sum_{t=0}^{n-1} \left(u_0^T H_{t+1}\left(a^{-1}_{U_t(\theta_0)} Y_{t+1}, \theta_0\right)\right)^2 \\
&\quad + \frac{1}{n} \sum_{t=0}^{n-1} u_0^T H_t\left(a^{-1}_{U_t(\theta_0)} Y_{t+1}, \theta_0\right) S_t(\theta_0)^T u_0 + o\left(1; \mathbb{P}^{(n)}_{\theta_0}\right)
\end{aligned} \tag{6.23}$$

Inserting (6.22) and (6.23) into (6.21) proves the desired asymptotic linearity of the loglikelihood ratio:

$$\log \frac{d\mathbb{P}^{(n)}_{\theta_n}}{d\mathbb{P}^{(n)}_{\theta_0}} = \frac{u_0^T}{\sqrt{n}} \sum_{t=1}^{n} S_t^*(\theta_0) - \frac{1}{2n} \sum_{t=1}^{n} \left(u_0^T S_t^*(\theta_0)\right)^2 + o\left(1; \mathbb{P}^{(n)}_{\theta_0}\right).$$

Now Assumption 8 completes the proof. □

6.5 Appendix: a general LAN theorem

In this appendix we provide a general setup to derive a (uniform) LAN theorem for models with dependent observations. Theorem 6.4 generalizes the results of Roussas (1972). A similar result was obtained by McNeney and Wellner (2000, Theorem 3.1).

Let, for each $n \in \mathbb{N}$, $(\Omega_n, \mathscr{F}_n)$ be a measurable space on which two probability measures $\tilde{\mathbb{P}}_n$ and $\mathbb{P}_n$ are defined. Let, for each $n \in \mathbb{N}$, $\mathscr{F}_{n0} \subset \cdots \subset \mathscr{F}_{nn} \subset \mathscr{F}_n$, be a sequence of increasing σ-fields. On these σ-fields we define, for $n \in \mathbb{N}$, the probability measures $\tilde{P}_n = \tilde{\mathbb{P}}_n|_{\mathscr{F}_{nn}}$, $P_n = \mathbb{P}_n|_{\mathscr{F}_{nn}}$, and for $t = 0, \ldots, n$, $\tilde{P}_{nt} = \tilde{\mathbb{P}}_n|_{\mathscr{F}_{nt}}$ and $P_{nt} = \mathbb{P}_n|_{\mathscr{F}_{nt}}$. Denote the Lebesgue decomposition of $\tilde{P}_{nt}$ on P_{nt} (with respect to $\mathscr{F}_{nt}$) by (L_{nt}, N_{nt}), i.e. $\tilde{P}_{nt}(A) = \int_A L_{nt}\,\mathrm{d}P_{nt} + \tilde{P}_{nt}(A \cap N_{nt})$, and $P_{nt}(N_{nt}) = 0$ for all $A \in \mathscr{F}_{nt}$. Under P_n, the likelihood ratio statistic LR_n for $\tilde{P}_n$ with respect to P_n is, by definition, given by L_{nn}. Put $LR_{n0} = L_{n0}$ and define the conditional likelihood ratio contribution of the t-th observation by

$$LR_{nt} = \frac{L_{nt}}{L_{n,t-1}}, \quad t = 1, \ldots, n,$$

with the convention $0/0 = 1$. Then, the likelihood ratio statistic can be decomposed as

$$LR_n = \prod_{t=0}^{n} LR_{nt}, \quad P_n\text{-a.s.}$$

This equality follows from the fact that, under P_n, $\{L_{nt} : 0 \le t \le n\}$ is a supermartingale with respect to the filtration $\{\mathscr{F}_{nt} : 0 \le t \le n\}$ (which is easy to check) and by repeated application of the following trivial proposition with $X = L_{nt}$, $Y = L_{n,t-1}$, and $\mathscr{F} = \mathscr{F}_{n,t-1}$, $t = 1, \ldots, n$.

Proposition 6.5.1. Suppose X is a nonnegative, integrable random variable and Y a $\mathscr{F}$-measurable random variable satisfying $Y \ge \mathbb{E}[X|\mathscr{F}]$. Then, $X1_{\{Y=0\}} \overset{\text{a.s.}}{=} 0$.

Proof. This follows from $0 \le \mathbb{E}X1_{\{Y=0\}} = \mathbb{E}\mathbb{E}[X|\mathscr{F}]\,1_{\{Y=0\}} \le \mathbb{E}Y1_{\{Y=0\}} = 0$. □

For general models, this concludes the general description of the likelihood ratio statistic as the product of conditional contributions. In the following theorem, we develop general criteria which allow for a LAN result.

Theorem 6.4. *Suppose that there exists $k \in \mathbb{N}$, such that for each $n \in \mathbb{N}$, there exist $\mathscr{F}_{nt}$-measurable mappings $S_{nt} : \Omega_n \to \mathbb{R}^k$, $R_{nt} : \Omega_n \to \mathbb{R}$, $t = 1, \ldots, n$, such that the conditional likelihood ratio contribution LR_{nt} can be written as*

$$LR_{nt} = \left(1 + \frac{1}{2}\frac{h_n^T}{\sqrt{n}} S_{nt} + R_{nt}\right)^2, \tag{6.24}$$

where, with $\mathscr{F}_{n,-1} = \{\emptyset, \Omega_n\}$,

1. *$h_n \to h_0 \in \mathbb{R}^k$, as $n \to \infty$*

2. *for each $n \in \mathbb{N}$, $\{S_{nt} : 1 \le t \le n\}$ is a P_n-square integrable martingale difference array with respect to the filtration $\{\mathscr{F}_{nt} : 0 \le t \le n\}$ satisfying the conditional Lindeberg condition and the WLLN for the squared conditional moments, i.e. for some $k \times k$ non-singular matrix I and for all $\epsilon > 0$, we have,*

$$\mathbb{E}_{P_n}\left[S_{nt} \mid \mathscr{F}_{n,t-1}\right] = 0, \quad t = 1, \dots, n, \tag{6.25}$$

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{P_n}\left[|S_{nt}|^2 1_{\{|S_{nt}|>\epsilon\sqrt{n}\}} \mid \mathscr{F}_{n,t-1}\right] \xrightarrow{p} 0, \quad \text{under } P_n \text{ as } n \to \infty, \tag{6.26}$$

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{P_n}\left[S_{nt}S_{nt}^T \mid \mathscr{F}_{n,t-1}\right] \xrightarrow{p} I, \quad \text{under } P_n \text{ as } n \to \infty. \tag{6.27}$$

3. *the remainder terms R_{nt} and the null-sets N_{nt} from the Lebesgue decomposition of P_n on $\tilde{P}_n$ are sufficiently small,*

$$\sum_{t=1}^{n} \mathbb{E}_{P_n}\left[R_{nt}^2 \mid \mathscr{F}_{n,t-1}\right] \xrightarrow{p} 0, \quad \text{under } P_n \text{ as } n \to \infty, \tag{6.28}$$

$$\sum_{t=1}^{n} \left(1 - \mathbb{E}_{P_n}\left[LR_{nt} \mid \mathscr{F}_{n,t-1}\right]\right) \xrightarrow{p} 0, \quad \text{under } P_n \text{ as } n \to \infty. \tag{6.29}$$

4. *the first term is asymptotically negligible, i.e. $LR_{n0} \xrightarrow{p} 1$ under P_n.*

Then the model satisfies the Uniform Local Asymptotic Normality (ULAN) condition, i.e.

$$\log LR_n = \frac{h_n^T}{\sqrt{n}}\sum_{t=1}^{n} S_{nt} - \frac{1}{2n}\sum_{t=1}^{n} h_n^T S_{nt} S_{nt}^T h_n + o(1; P_n) \tag{6.30}$$

$$\xrightarrow{d} \mathrm{N}\left(-\frac{1}{2}h_0^T I h_0, h_0^T I h_0\right), \quad \text{under } P_n \text{ as } n \to \infty. \tag{6.31}$$

Proof.
Rewrite the likelihood ratio statistic as the two leading terms in (6.30) and some remainder terms,

$$\begin{aligned}\log LR_n = \sum_{t=0}^{n} \log LR_{nt} &= o(1; P_n) + \sum_{t=1}^{n} \frac{h_n^T}{\sqrt{n}} S_{nt} - \frac{1}{2}\sum_{t=1}^{n}\left(\frac{h_n^T}{\sqrt{n}} S_{nt}\right)^2 \\ &+ 2\sum_{t=1}^{n}\left(R_{nt} - \mathbb{E}_{P_n}\left[R_{nt} \mid \mathscr{F}_{n,t-1}\right]\right) - \sum_{t=1}^{n} R_{nt}^2 - \sum_{t=1}^{n} \frac{h_n^T}{\sqrt{n}} S_{nt} R_{nt} \\ &+ 2\sum_{t=1}^{n}\left\{\mathbb{E}_{P_n}\left[R_{nt} \mid \mathscr{F}_{n,t-1}\right] + \frac{1}{8}\left(\frac{h_n^T}{\sqrt{n}} S_{nt}\right)^2\right\}\end{aligned}$$

$$+\sum_{t=1}^{n} r\left(\frac{1}{2}\frac{h_n^T}{\sqrt{n}}S_{nt}+R_{nt}\right), \tag{6.32}$$

where $r(x) = 2\log|1+x| - 2x + x^2$. To prove the expansion (6.30), we show that the five remainder terms at the right-hand side all converge to zero in probability. First we recall the following implication of Theorem 2.23 and Corollary 3.1 in Hall and Heyde (1980) (see Drost et al. (1997, Lemma 2.2), for additional details).

Lemma 6.5.1. If the square-integrable process $\{X_{nt} : 1 \le t \le n\}$ adapted to the filtration $(\mathscr{F}_{nt})_{0\le t\le n}$ satisfies, under P_n, $\sum_{t=1}^{n} \mathbb{E}_{P_n}\left[X_{nt}^2 \mid \mathscr{F}_{n,t-1}\right] \xrightarrow{p} 0$, then, under P_n,

$$\sum_{t=1}^{n} X_{nt}^2 \xrightarrow{p} 0, \quad \text{and} \quad \sum_{t=1}^{n}\left(X_{nt} - \mathbb{E}_{P_n}\left[X_{nt} \mid \mathscr{F}_{n,t-1}\right]\right) \xrightarrow{p} 0.$$

□

Since $(L_{nt})_{0\le t\le n}$ is a P_n-supermartingale we have $\mathbb{E}_{P_n} LR_{nt} \le 1$. Since S_{nt} is P_n-square integrable we see, from (6.24), that R_{nt} is P_n-square integrable. From Lemma 6.5.1 and Condition (6.28), we now immediately obtain, under P_n,

$$\sum_{t=1}^{n}\left(R_{nt} - \mathbb{E}_{P_n}\left[R_{nt} \mid \mathscr{F}_{n,t-1}\right]\right) \xrightarrow{p} 0, \text{ and } \sum_{t=1}^{n} R_{nt}^2 \xrightarrow{p} 0.$$

Next we show that the third remainder $n^{-1/2}\sum_{t=1}^{n} h_n^T S_{nt} R_{nt} \xrightarrow{p} 0$, under P_n. First note that Conditions (6.25)-(6.27) and Theorem 2.23 of Hall and Heyde (1980) imply the unconditional version of (6.27): $n^{-1}\sum_{t=1}^{n} h_n^T S_{nt} S_{nt}^T h_n \xrightarrow{p} h_0^T I h_0$. Thus an application of the Cauchy-Schwarz inequality, combined with the previously obtained $\sum_{t=1}^{n} R_{nt}^2 \xrightarrow{p} 0$, yields the desired convergence of this remainder term. To prove the negligibility of the fourth remainder term in (6.32), observe that, (6.24), (6.25), (6.27), (6.28), and the Cauchy-Schwarz inequality, yield

$$\begin{aligned}
\sum_{t=1}^{n}\left(\mathbb{E}_{P_n}\left[LR_{nt} \mid \mathscr{F}_{n,t-1}\right] - 1\right) &= \sum_{t=1}^{n} \mathbb{E}_{P_n}\left[\frac{h_n^T}{\sqrt{n}} S_{nt} \mid \mathscr{F}_{n,t-1}\right] + 2\sum_{t=1}^{n} \mathbb{E}_{P_n}\left[R_{nt} \mid \mathscr{F}_{n,t-1}\right] \\
&\quad + \sum_{t=1}^{n} \mathbb{E}_{P_n}\left[\left(\frac{1}{2}\frac{h_n^T}{\sqrt{n}} S_{nt}\right)^2 \mid \mathscr{F}_{n,t-1}\right] + \sum_{t=1}^{n} \mathbb{E}_{P_n}\left[R_{nt}^2 \mid \mathscr{F}_{n,t-1}\right] \\
&\quad + 2\sum_{t=1}^{n} \mathbb{E}_{P_n}\left[\left(\frac{1}{2}\frac{h_n^T}{\sqrt{n}} S_{nt}\right) R_{nt} \mid \mathscr{F}_{n,t-1}\right] \\
&= 2\sum_{t=1}^{n} \mathbb{E}_{P_n}\left[R_{nt} \mid \mathscr{F}_{n,t-1}\right] + \frac{1}{4} h_0^T I h_0 + o(1; P_n).
\end{aligned}$$

Thus, by (6.29),

$$\sum_{t=1}^{n} \mathbb{E}_{P_n}\left[R_{nt} \mid \mathscr{F}_{n,t-1}\right] \xrightarrow{p} -\frac{1}{8} h_0^T I h_0.$$

Substituting this result into the fourth remainder term and using the unconditional version of (6.27) yields convergence to zero. To show that the final remainder term in (6.32) is negligible, we first show that,

$$\max_{t=1,\dots,n}\left|\frac{h_n^T}{\sqrt{n}}S_{nt}+R_{nt}\right|\xrightarrow{p}0, \tag{6.33}$$

$$\sum_{t=1}^{n}\left|\frac{h_n^T}{\sqrt{n}}S_{nt}+R_{nt}\right|^3\xrightarrow{p}0. \tag{6.34}$$

Observe that, for all $a, b \in \mathbb{R}$,

$$|a+b|^2\, I_{\{|a+b|>\epsilon\}} \le 4\,|a|^2\, I_{\{|a|>\epsilon/2\}} + 4\,|b|^2\, I_{\{|b|>\epsilon/2\}}.$$

Let $\epsilon > 0$ and $\eta > 0$, then by a result due to Dvoretzky (see Hall and Heyde (1980), Lemma 2.5) and by (6.26) [or by (6.25) if $h_0 = 0$] and (6.28),

$$\begin{aligned}
&P_n\left\{\max_{t=1,\dots,n}\left|\frac{1}{2}\frac{h_n^T}{\sqrt{n}}S_{nt}+R_{nt}\right|>\epsilon\right\}\\
&\le \eta + P_n\left\{\sum_{t=1}^{n}P_n\left(\left\{\left|\frac{1}{2}\frac{h_n^T}{\sqrt{n}}S_{nt}+R_{nt}\right|>\epsilon\right\}\middle|\,\mathscr{F}_{n,t-1}\right)>\eta\right\}\\
&\le P_n\left\{\sum_{t=1}^{n}\mathbb{E}_{P_n}\left[\left|\frac{1}{2}\frac{h_n^T}{\sqrt{n}}S_{nt}\right|^2 1_{\{|h_n^T S_{nt}|>\sqrt{n}\epsilon\}}\mid \mathscr{F}_{n,t-1}\right]+\sum_{t=1}^{n}\mathbb{E}_{P_n}\left[R_{nt}^2\mid\mathscr{F}_{n,t-1}\right]>\frac{\epsilon^2\eta}{4}\right\}\\
&\quad+\eta\to\eta \text{ as } n\to\infty.
\end{aligned}$$

This implies (6.33) since η is arbitrary. Equation (6.34) is obtained from this result by taking out the maximum (which tends to zero) and by observing that the remaining quadratic term is bounded in probability (use the arguments leading to the convergence of the third remainder term). By (6.33) it suffices to derive the behavior of the final remainder term on the event $\left\{\left|\frac{1}{2}\frac{h_n^T}{\sqrt{n}}S_{nt}+R_{nt}\right|\le\frac{1}{2}\right\}$. On this set this remainder term is bounded by, using $\left|\log(1+x)-x+\frac{1}{2}x^2\right|\le\frac{2}{3}x^3$ for $|x|\le\frac{1}{2}$,

$$\left|\sum_{t=1}^{n} r\left(\frac{1}{2}\frac{h_n^T}{\sqrt{n}}S_{nt}+R_{nt}\right)\right|\le\frac{4}{3}\sum_{t=1}^{n}\left(\frac{1}{2}\frac{h_n^T}{\sqrt{n}}S_{nt}+R_{nt}\right)^3.$$

Convergence to zero is obtained from (6.34). This completes the proof of the expansion (6.30). The convergence to the normal distribution in (6.31) follows immediately from Corollary 3.1 of Hall and Heyde (1980). This completes the proof of Theorem 6.4. □

Remark 11. Let us discuss the assumptions in Theorem 6.4 shortly. First of all note that Assumption 1 ensures the validity of a martingale central limit theorem. Assumption 2 allows for the expansion of the logarithmic likelihood. For an appreciation of using an expansion of the *square root* of the likelihood ratio, see Pollard (1997).

Bibliography

Ahn, S., L. Gyemin, and J. Jongwoo (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. *Journal of Time Series Analysis*, 8:261–275.

——— (1990): An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, 27:314–324.

Andrews, D. (1992): Generic uniform convergence. *Econometric Theory*, 8:241–257.

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. *The Annals of Regional Science*, 35:431–448.

Biau, G. and M. Wegkamp (2005): A note on minimum distance estimation of copula densities. *Statistics & Probability Letters*, 73:105–114.

Bickel, P. and C. Klaassen (1986): Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location. *Advances in Applied Mathematics*, 7:55–69.

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

Bickel, P., Y. Ritov, and J. Wellner (1991): Efficient estimation of linear functionals of a probability measure P with known marginal distributions. *Annals of Statistics*, 19:1316–1346.

Billingsley, P. (1995): *Probability and Measure*. Wiley series in probability and mathematical statistics, 2nd edition.

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

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

——— (2003): Analysing state dependences in emotional experiences by dynamic count data models. *Journal of the Royal Statistical Society: Series C*, 52:213–226.

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

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

Chan, N. and C. Wei (1987): Asymptotic inference for nearly nonstationary AR(1) processes. *Annals of Statistics*, 15:1050–1063.

Chen, X. and Y. Fan (2006): Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130:307–335.

Chen, X., Y. Fan, and V. Tsyrennikov (2006): Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101:1228–1240.

Clayton, D. and J. Cuzick (1985): Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society: Series C*, 148:82–117.

Dacunha-Castelle, D. and M. Duflo (1986): *Probability and Statistics*. Springer-Verlag, 1st edition.

Davydov, Y. (1973): Mixing conditions for Markov chains. *Theory of Probability and its Applications*, 18:312–328.

Dharmadhikari, S., V. Fabian, and K. Jogdeo (1968): Bounds on the moments of martingales. *The Annals of Mathematical Statistics*, 39:1719–1723.

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

Doukhan, P. (1994): *Mixing: properties and examples.* Springer-Verlag: Lecture notes in Statistics.

Doukhan, P., P. Massart, and E. Rio (1995): Invariance principles for absolutely regular empirical processes. *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, 32:393–427.

Drost, F. and C. Klaassen (1997): Efficient estimation in semiparametric GARCH models. *Journal of Econometrics*, 81:193–221.

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

Drost, F. and B. Werker (2004): Semiparametric duration models. *Journal of Business & Economic Statistics*, 22:40–50.

Du, J.-G. and Y. Li (1991): The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis*, 12:129–142.

Embrechts, P., A. Hoeing, and A. Juri (2003a): Using copulae to bound the value-at-risk for functions of dependent risks. *Finance & Stochastics*, 7:145–167.

Embrechts, P., F. Lindskog, and A. McNeil (2003b): Modeling dependence with copulas and applications to risk management. In S. Rachev, editor, *Handbook of Heavy Tailed Distributions in Finance*, pages 329–384.

Enciso-Mora, V., P. Neal, and T. Subba Rao (2006): Efficient order selection algorithms for integer valued ARMA processes. *Research Report No. 18, Probability and Statistics Group School of Mathematics, The University of Manchester.*

Engle, R. (2000): The econometrics of ultra-high-frequency data. *Econometrica*, 68:1–22.

Engle, R. and G. Gonzalez-Rivera (1991): Semiparametric ARCH Models. *Journal of Business & Economic Statistics*, 9:345–360.

Feller, W. (1968): *An introduction to probability theory and its applications: volume I.* New York, John Wiley & Sons, 3rd edition.

Fermanian, J.-D., D. Radulović, and M. Wegkamp (2004): Weak Convergence of Empirical Copula Processes. *Bernoulli*, 10:847–860.

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, editor, *Developments in time series*, pages 310–330. London: Chapman & Hall.

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

——— (2005): Asymptotic properties of CLS estimators in the Poisson AR(1) model. *Statistics & Probability Letters*, 73:147–153.

Gaenssler, P. and W. Stute (1987): *Seminar on empirical processes*. Birkhäuser.

Genest, C. (1987): Frank's family of bivariate distributions. *Biometrika*, 74:549–555.

Genest, C., K. Ghoudi, and L.-P. Rivest (1995): A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552.

Genest, C., R. van den Goorbergh, and B. Werker (2005): Multivariate option pricing using dynamic copula models. *Insurance: Mathematics and Economics*, 37:101–114.

Genest, C. and B. Werker (2002): Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In C. Cuadras and A. Rodríguez-Lallena, editors, *Distributions with given marginals and statistical modelling*, pages 103–112. Kluwer Academic Publishers, Dordrecht.

Gouriéroux, C., E. Ghysels, and J. Jasiak (2004): Stochastic volatility duration models. *Journal of Econometrics*, 119:413–433.

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

Hall, P. and C. Heyde (1980): *Martingale limit theory and its applications*. Academic Press, New York.

Hallin, M., C. Vermandele, and B. Werker (2004): Semiparametrically efficient inference based on signs and ranks for median-restricted models. *Tilburg University, CentER discussion paper*, 2004-11.

Hallin, M. and B. Werker (2001): Semiparametric efficiency, distribution-freeness, and invariance. *Bernoulli*, 9:137–165.

Hellström, J. (2001): Unit root testing in integer-valued AR(1) models. *Economic Letters*, 70:9–14.

Hirano, K. and J. Porter (2003a): Asymptotic efficiency in parametric structural models with parameter dependent support. *Econometrica*, 71:1307–1338.

——— (2003b): Efficiency in asymptotic shift experiments. *Working paper.*

Hougaard, P. (1989): Fitting a multivariate failure time distribution. *IEEE Transactions on Reliability*, 38:444–448.

Hu, L. (2006): Dependence patterns across financial markets: a mixed copula approach. *Applied Financial Economics*, 16:717–729.

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

——— (2003b): Asymptotic inference for nearly unstable INAR(1) models. *Journal of Applied Probability*, 40:750–765.

——— (2005): Fluctuation limit of branching processes with immigration and estimation of the means. *Advances in Applied Probability*, 37:523–538.

Janssen, A. and V. Ostrovski (2005): The convolution theorem of Hájek and Le Cam - revisited. *Working paper.*

Jeganathan, P. (1995): Some aspects of asymptotic theory with applications to time series models. *Econometric Theory*, 11:818–887.

Joe, H. (1997): *Multivariate models and dependence concepts.* Chapmann & Hall, London, 1st edition.

——— (2005): Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

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

Junker, M. and A. May (2005): Meaurement of aggregate risk with copulas. *Econometrics Journal*, 8:428–454.

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

Klaassen, C. and H. Putter (2005): Efficient estimation of Banach parameters in semiparametric models. *Annals of Statistics*, 33:307–346.

Klaassen, C. and J. Wellner (1997): Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3:55–77.

Koul, H. and A. Schick (1997): Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli*, 3:247–277.

Kreiss, J.-P. (1987a): On adaptive estimation in autoregressive models when there are nuisance functions. *Statistics & Decisions*, 5:59–76.

——— (1987b): On adaptive estimation in stationary ARMA processes. *Annals of Statistics*, 15:112–133.

Kreyszig, E. (1978): *Introductory functional analysis with applications.* Wiley, 1st edition.

Latour, A. (1998): Existence and stochastic structure of a nonnegative integer-valued autoregressive process. *Journal of Time Series Analysis*, 19:439–455.

Le Cam, L. (1986): *Asymptotic methods in statistical decision theory.* Springer series in statistics, 1st edition.

Le Cam, L. and G. Yang (1988): On the preservation of local asymptotic normality under information loss. *Annals of Statistics*, 16:483–520.

——— (1990): *Asymptotics in Statistics: some basic concepts.* Springer series in statistics, 1st edition.

Li, D. (2000): On default correlation: a copula function approach. *Journal of Fixed Income*, 9:43–54.

Linton, O. (1993): Adaptive estimation in ARCH models. *Econometric Theory*, 9:99–135.

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

McKenzie, E. (2003): Discrete variate time series. In C. Rao and D. Shanbhag, editors, *Handbook of Statistics*, pages 573–606. Amsterdam: Elsevier Science.

McNeney, B. and J. Wellner (2000): Application of convolution theorems in semiparametric models with non-i.i.d. data. *Journal of Statistical Planning and Inference*, 91:441–480.

Meyn, S. and R. Tweedie (1994): *Markov chains and stochastic stability.* Springer-Verlag, 2nd edition.

Neal, P. and T. Subba Rao (2007): MCMC for integer-valued ARMA processes. *Journal of Time Series Analysis*, 28:92–110.

Nelsen, R. (1999): *An introduction to copulas.* Springer-Verlag, New York, 1st edition.

Newey, W. (1990): Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5:99–135.

Oakes, D. (1982): A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B*, 44:414–422.

——— (1986): Semiparametric inference in a model for association in bivariate survival data. *Biometrika*, 73:353–361.

——— (1994): Multivariate survival distributions. *Journal of Nonparametric Statistics*, 3:343–354.

Owen, A. (2001): *Empirical likelihood.* Chapman & Hall, 1st edition.

Peng, H. and A. Schick (2002): On efficient estimation of linear functionals of a bivariate distribution with known marginals. *Statistics & Probability Letters*, 59:83–91.

——— (2004): Efficient estimation of linear functionals of a bivariate distribution with equal, but unknown, marginals: the minimum chi-square approach. *Statistics & Decisions*, 22:301–318.

——— (2005): Efficient estimation of linear functionals of a bivariate distribution with equal, but unknown marginals: the least-squares approach. *Journal of Multivariate Analysis*, 95:385–409.

Philips, P. (1987): Towards a unified asymptotic theory for autoregression. *Biometrika*, 74:535–547.

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

Pollard, D. (1997): Another look at differentiability in quadratic mean. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 305–314. Springer-Verlag, New York.

Robinson, P. (1988): Semiparametric econometrics: a survey. *Journal of Applied Econometrics*, 3:35–51.

Roussas, G. (1972): *Contiguity of probability measures: some applications in statistics.* Cambridge tracts in mathematics and mathematical physics 63, Cambridge.

Rudholm, N. (2001): Entry and the number of firms in the Swedish pharmaceuticals market. *Review of Industrial Organization*, 19:351–364.

Rudin, W. (1973): *Functional analysis.* McGraw-Hill, New York, 1st edition.

Serfling, R. (1975): A general Poisson approximation theorem. *Annals of Probability,* 3:726–731.

Shen, X. (1997): On methods of sieves and penalization. *Annals of Statistics,* 25:2555–2591.

Shih, J. and T. Louis (1995): Inferences on the association parameter in copula models for bivariate survival data. *Biometrics,* 51:1384–1399.

Shiryaev, A. and V. Spokoiny (1999): *Statistical experiments and decisions: asymptotic theory.* World Scientific Publishing Company, 1st edition.

Silva, I. and M. Silva (2006): Asymptotic distribution of the Yule-Walker estimator for INAR(p) processes. *Statistics & Probability Letters,* 76:1655–1663.

Silva, M. and V. Oliveira (2004): Difference equations for the higher order moments and cumulants of the INAR(1) model. *Journal of Time Series Analysis,* 25:317–333.

——— (2005): Difference equations for the higher-order moments and cumulants of the INAR(p) model. *Journal of Time Series Analysis,* 26:17–36.

Steigerwald, D. (1992): Adaptive estimation in time series regression models. *Journal of Econometrics,* 54:251–275.

Steutel, F. and K. Van Harn (1979): Discrete analogues of self-decomposability and stability. *Annals of Probability,* 7:893–899.

Stoker, T. (1991): *Lectures on semiparametric econometrics.* CORE Foundation, Louvain-la-Neuve, Belgium.

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

Tricomi, F. (1985): *Integral equations.* Dover Publications, 2nd edition.

Tsukahara, H. (2005): Semiparametric estimation in copula models. *The Canadian Journal of Statistics,* 33:357–375.

Van der Vaart, A. (1991a): An asymptotic representation theorem. *International Statistical Review,* 59:97–121.

——— (1991b): Efficiency and Hadamard differentiability. *Scandinavian Journal of Statistics,* 18:63–75.

——— (1995): Efficiency of infinite dimensional M-estimators. *Statistica Neerlandica*, 49:9–30.

——— (2000): *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1st edition.

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

Wefelmeyer, W. (1996): Quasi-likelihood and optimal inference. *Annals of Statistics*, 24:405–422.

Wei, C. and J. Winnicki (1990): Estimation of the means in the branching process with immigration. *Annals of Statistics*, 18:1757–1773.

Zheng, H., I. Basawa, and S. Datta (2006): Inference for pth-order random coefficient integer-valued autoregressive processes. *Journal of Time Series Analysis*, 27:411–440.

Nederlandse samenvatting

Dit proefschrift bestaat uit twee delen. Deel I bevat contributies aan de literatuur over geheeltallige tijdreeksmodellen. Het tweede deel bestaat uit twee hoofdstukken: een hoofdstuk over copula-modellen en een hoofdstuk over semiparametrische tijdreeksmodellen.

Deel I Veel interessante variabelen in de economische wetenschappen, maar bijvoorbeeld ook in de medische wetenschappen en biologie, kunnen opgevat worden als een niet-negatieve, geheeltallige tijdreeks. Denk bijvoorbeeld aan het aantal transacties in het aandeel SNS-Reaal per dag, het aantal patiënten in een ziekenhuis gemeten aan het einde van iedere dag, etcetera. Het belang van adequate modellen en statistische technieken voor dergelijke processen behoeft dus geen betoog. Tot eind jaren zeventig werd echter, relatief gezien, weinig onderzoek verricht op dit gebied. Een verklaring hiervoor is dat het construeren van adequate probabilistische modellen veel lastiger is dan voor continue data. De laatste twintig jaar zijn er verschillende probabilistische modellen voorgesteld. Deel I van dit proefschrift ontwikkelt statistische methoden voor een van de meest, in empirische applicaties, gebruikte modellen: de klasse van INAR processen. Deze kunnen gezien worden als een niet-negatief geheeltallig analogon van de bekende (continue) autoregressieve (AR) processen.

Hoofdstuk 1 presenteert enkele probabilistische resultaten voor INAR processen, welke gebruikt worden in latere hoofdstukken van Deel 1. In het bijzonder worden condities gegeven voor de existentie van een stationair INAR proces, en de existentie van momenten en (uniforme) ergodiciteit.

In **Hoofdstuk 2** wordt de structuur van parametrische, stationaire INAR modellen bekeken. De conclusie van dit hoofdstuk is dat, onder zekere gladheidsvoorwaarden, deze modellen de Lokale Asymptotisch Normale (LAN) structuur hebben. Een zeer belangrijk ingrediënt in het bewijs van dit resultaat is dat we de overgangsscores kunnen representeren als conditionele verwachtin-

gen. De LAN-structuur is van belang voor Hoofdstuk 3. Bovendien volgt uit deze structuur een nieuwe schatter die, behalve asymptotisch efficiënt, ook attractief is uit computationeel oogpunt (ten opzichte van de meest aannemelijke schatter).

In **Hoofdstuk 3** worden semiparametrische INAR modellen bestudeerd. In deze semiparametrische modellen wordt de verdelingsveronderstelling op de innovatie-structuur nagenoeg losgelaten. Dit geeft een groter en dus realistischer model. De prijs hiervoor is dat het schatten van een parameter moeilijker is dan het schatten van dezelfde parameter in een parametrisch deelmodel. In dit hoofdstuk wordt een schatter, van zowel de Euclidische parameter als de puntmassa functie van de innovaties, voorgesteld, die geïnterpreteerd kan worden als een niet-parametrische meest aannemelijke schatter. Asymptotische efficiëntie van deze schatter wordt bewezen.

In Hoofdstukken 1-3 wordt gekeken naar stationaire modellen. Om het effect van niet-stationairiteit op de statistische eigenschappen te onderzoeken, wordt in **Hoofdstuk 4** het limiet-experiment van een onstabiel INAR proces afgeleid. Het resultaat is zeer opmerkelijk, aangezien het limiet-experiment niet de gebruikelijke equivariantie- en kwadratische structuur heeft. De statistische implicaties van dit resultaat worden ook besproken. In het bijzonder wordt aangetoond dat de Dickey-Fuller toets (asymptotisch en lokaal) geen onderscheidend vermogen heeft, terwijl, dankzij het limiet-experiment, aangetoond wordt dat een intuïtieve toets asymptotisch optimaal is.

Deel II

In **Hoofdstuk 5** wordt het efficiënt schatten van de marginale verdelingsfuncties op basis van een aselecte steekproef uit een bivariate verdeling, waarvan de copula bekend is en de marginalen onbekend zijn, bestudeerd. Er wordt aangetoond dat, in het algemeen, de marginale empirische verdelingsfuncties niet efficiënt zijn. Op basis van de empirische aannemelijkheidsfunctie wordt een schatter voorgesteld die de kennis over de copula uitbuit. Asymptotische optimaliteit van deze schatter wordt aangetoond.

Chapter 6 leidt semiparametrische ondergrenzen af voor Euclidische componenten in algemene (continue) tijdreeksmodellen met een groep-structuur. In deze modellen wordt niet, zoals gebruikelijk, aangenomen dat de innovaties onderling onafhankelijk en identiek verdeeld zijn. In plaats hiervan wordt de afhankelijkheidsstructuur als een (extra) oneindig-dimensionale hinderparameter gezien.