Efficient Probabilistic Forecasts for Counts*

Brendan P.M. McCabe‡, Gael M. Martin‡ and David Harris§

September 21, 2010

Abstract

Efficient probabilistic forecasts of integer-valued random variables are derived. The optimality is achieved by estimating the forecast distribution nonparametrically over a given broad model class and proving asymptotic (nonparametric) efficiency in that setting. The method is developed within the context of the integer autoregressive class (INAR) of models, which is a suitable class for any count data that can be interpreted as a queue, stock, birth and death process or branching process. The theoretical proofs of asymptotic efficiency are supplemented by simulation results that demonstrate the overall superiority of the nonparametric estimator relative to a misspecified parametric alternative, in large but finite samples. The method is applied to counts of stock market iceberg orders. A subsampling method is used to assess sampling variation in the full estimated forecast distribution and a proof of its validity is given.

KEYWORDS: Nonparametric Maximum Likelihood Estimation; Nonparametric Efficiency; Probabilistic Forecasting; INAR Model Class; Subsampling; Iceberg Stock Market Orders.

JEL CODES: C14, C22, C53.

1 Introduction

In this paper we propose an approach to forecasting count time series data modelled by the integer autoregressive (INAR) class. Forecasts that are coherent with the discrete sample space and that quantify all uncertainty associated with future counts are produced by estimating forecast distributions over all horizons. An ex-ante optimal estimator for the INAR class is derived by treating the arrivals process nonparametrically and proving the

*This research has been supported by an Australian Research Council (ARC) Discovery Grant and an ARC Future Fellowship. The authors would like to thank the Editor, a co-editor and two anonymous referees for very detailed and constructive comments on an earlier draft of the paper. The empirical data has been kindly supplied by Joachim Grammig, Robert Jung and Andy Tremayne, with the permission of the Deutsche Börse.

†Management School, University of Liverpool, UK.

‡Corresponding author: Department of Econometrics and Business Statistics, Monash University, Australia. Email: gael.martin@monash.edu.

§Department of Economics, University of Melbourne, Australia.
asymptotic (nonparametric) efficiency of the estimated forecast distributions. Subsampling
(Politis, Romano and Wolf, 1999) is used to provide a novel technique for assessing and
visualizing sampling variation over the entire forecast distribution. The technique parallels
the conventional prediction interval for a scalar point forecast, but automatically ensures
that the non-negativity and summation to unity properties of probabilities hold. A proof of
the theoretical validity of the subsampling method is given.

The INAR class is a behavioural/structural model of a very large collection of count
data time series. In brief, any data series that may be thought of as the number of members
(e.g. people, firms, orders) of a queue; the number of units in a stock or inventory; or
the outcome of a birth and death process, or a branching process with immigration, may be
modelled by the INAR class. The class is thus sufficiently broad to warrant the development
of a bespoke forecasting strategy for use by practitioners. Recent applications of the INAR
model include: Franke and Seligmann (1993), Pickands and Stine (1997) and Cardinal et al.
(1999) (medicine); Bockenholt (1999) (marketing); Thyregod et al. (1999) and Pavlopoulos
and Karlis (2008) (environmental studies); Brännäs and Hellstrom (2001) and Rudholm

Being an observation-driven class, the INAR model admits a closed-form representation
of the likelihood function, with the nonparametric estimator of the forecast distribution
produced via the maximization of an empirical likelihood. Hence, from a practical point
of view, the method is relatively straightforward to implement. In particular, the more
computationally demanding simulation methods typically needed to produce probabilistic
forecasts in a discrete (non-Gaussian) state space setting are completely avoided.

The paper is organized as follows. In Section 2 we outline the structure of the INAR
model for count time series and discuss the application of a nonparametric maximum like-
lihood estimator (NPMLE) in that setting. The asymptotic efficiency of the NPMLE-
based estimator of the forecast distribution is then demonstrated, with the proof of the
derifferentiability of the mapping that defines the forecast distribution given in the Appendix.
Computational details associated with the implementation of the estimation method are
also provided. (All numerical results reported in the paper have been produced using the
GAUSS software, with programs available from the corresponding author on request.) The
finite sample performance of the asymptotically efficient estimator of the forecast distribu-
tion (AEEF hereafter), within the INAR class, is documented via simulation in Section 3.
In particular, the overall superiority of the AEEF relative to an estimated forecast distrib-
ution based on a misspecified parametric maximum likelihood estimator, in large but finite
samples, is illustrated. In Section 4 the subsampling method is described and its theoretical
validity proven. The AEEF is then applied to a time series of German stock market ice-
berg orders which, constituting a record of the number of elements over time in a queue, is
suitably modelled as an $INAR$ process. Section 5 concludes.

2 Probabilistic Forecasting in the INAR Class

The $INAR$ class of models was first introduced by Al-Osh and Alzaid (1987) and McKenzie (1988). It was further investigated by, amongst others, Du and Li (1991), Dion et al. (1995), Latour (1998), Ispany et al. (2003), Freeland and McCabe (2004), Jung et al. (2005), McCabe and Martin (2005), Silva and Oliveira (2005), Jung and Tremayne (2006), Zhu and Joe (2006), Neal and Subba Rao (2007), Bu and McCabe (2008), Bu et al. (2008) and Drost et al. (2008, 2009). McKenzie (2003) provides a review of the model class. In Section 2.1 we outline the $INAR$ class and the properties of the $NPMLE$. This is followed, in Section 2.2, by demonstration of the asymptotic efficiency of the nonparametric estimator of the forecast distribution.

2.1 NPMLE in the INAR Class

In the spirit of Du and Li (1991) we define the $INAR(p)$ class to be

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

where the innovations $\{\varepsilon_t\}$ are an i.i.d process with a distribution $G$. The distribution $G = \{g_r\}$ is a discrete sequence of probabilities on the set $Z = \{0, 1, 2, \ldots\}$. Conditional on $X_{t-k}$, $k \in \{1, 2, \ldots, p\}$, the thinning operators $\alpha_k \circ X_{t-k}$, $k \in \{1, 2, \ldots, p\}$ are binomial, and defined as

$$\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} B_{i,k,t},$$

where each collection $\{B_{i,k,t}, i = 1, 2, \ldots, X_{t-k}\}$ consists of independently distributed Bernoulli random variables with thinning parameter (probability of unity) $\alpha_k$, and the collections are mutually independent. It is assumed that $\alpha_k \in [0, 1)$, for all $k \in \{1, 2, \ldots, p\}$, and that $\sum_{k=1}^{p} \alpha_k < 1$. The innovations are taken to be independent of all thinning operations. The initial values $(X_0, X_{-1}, \ldots, X_{-p})$ are independent drawings from the stationary distribution of the model and, hence, under the conditions above, $X_t$ is also a strictly stationary process. The infinite dimensional parameter of the model is $\theta = (\alpha_1, \ldots, \alpha_p, G)$.

At time $t$, each thinning operator performs one of $p$ binomial experiments, with parameters $(X_{t-k}, \alpha_k)$, $k \in \{1, 2, \ldots, p\}$, to determine the number from that time vintage that survives in the system. When $\alpha_k$ is close to zero it is expected that there are almost no survivors from the $(t-k)$ vintage and, correspondingly, there expected to be are many survivors when $\alpha_k$ is close to unity. Consider the vintage $X_t$. At $t+1$, $X_t$ is thinned by $\alpha_1$ and at time $t+2$, $X_t$ is again thinned but using $\alpha_2$. Thus, the ‘offspring’ of $X_t$ are distributed

3
generated by
may be thought of as a birth and death, or stock process, with additions (births) being survivors remaining in the queue, from
distribution of
model is known as Poisson autoregression (PAR) since, in this case, the marginal stationary distribution of $X_t$ is also Poisson.

For any set of values $i_0, i_1, ..., i_p$ in $\mathbb{Z}$ define the function

$$ f_{i_0,i_1,...,i_p}(\theta) = \sum_{(j_1,...,j_p) \in J(i_0,...,i_p)} \prod_{k=1}^{p} p_{j_k|i_k}(\alpha_k).g_{i_0-(j_1+...+j_p)}, \quad (2) $$

where

$$ p_{j_k|i_k}(\alpha_k) = \binom{i_k}{j_k} \alpha_k^{j_k} (1 - \alpha_k)^{i_k-j_k}, \quad 0 \leq j_k \leq i_k \quad (3) $$

and

$$ J(i_0, ..., i_p) = \left\{ (j_1, ..., j_p) \in \mathbb{Z}^p : j_k \leq \left(i_0 - \sum_{l=1}^{k-1} j_l\right) \land i_k, \ k = 1, 2, ..., p \right\}. $$

Empty sums are taken to be zero, so that $j_1 \leq (i_0 \land i_1)$. Expression (2) gives the probability $Pr(X_t = i_0|X_{t-1} = i_1, ..., X_{t-p} = i_p; \theta)$ under the model (1) and is the convolution of $p$ binomials and the arrivals distribution $G = \{g_r\}$. Given observed counts $x_1, x_2, ..., x_T$, the nonparametric likelihood (given the initial observations) is

$$ L(\theta|x_1, ..., x_T) = \prod_{t=p+1}^{T} P(X_t = x_t|X_{t-1} = x_{t-1}, ..., X_{t-p} = x_{t-p}; \theta), \quad (4) $$

where $P(X_t = x_t|X_{t-1} = x_{t-1}, ..., X_{t-p} = x_{t-p}; \theta) = f_{x_t|x_{t-1},...,x_{t-p}}(\theta)$. When $p = 1$, these expressions simplify considerably and

$$ L(\theta|x_1, ..., x_T) = \prod_{t=2}^{T} \sum_{j=0}^{x_t} \binom{x_{t-1}}{j} \alpha_1^j (1 - \alpha_1)^{x_{t-1}-j} g_{x_{t}-j}. \quad (5) $$

The parameter space is $\Theta = ([0,1]^p \times \mathcal{M})$, where $\mathcal{M}$ is the space of discrete probability distributions on $\mathbb{Z}$. To obtain the NPMLE, (4) is maximized over $0 \leq \alpha_k < 1; k = 1, 2, ..., p$ and $\sum_{r=g_+}^{g_-} g_r = 1$ where $g_- = 0 \lor \min_{t=p+1,...,T} (x_t - \sum_{k=1}^{p} x_{t-k})$ and $g_+ = \max_{t=p+1,...,T} x_t$. The NPMLE is denoted $\hat{\theta} = (\hat{\alpha}; \hat{G}) = (\hat{\alpha}_k; k = 1, 2, ..., p; \{\hat{g}_r\})$ and consists of a vector, $\hat{\alpha}$,
which is an estimator of $\alpha = (\alpha_1, \ldots, \alpha_p)'$ and a sequence $\{\hat{g}_r\}$, which is an estimator of the distribution $G = \{g_r\}$. (For notational simplicity we suppress the dependence of estimators, like $\hat{\theta}$, on the sample size $T$). The sequence estimator $\hat{G} = \{\hat{g}_r\}$ contains only a finite number, $(g_+ - g_-)$, of non-zero values in finite samples but this number becomes potentially infinite as $T \to \infty$. Let the $p$-dimensional Euclidean space be denoted $\mathbb{R}^p$ and let the Banach space of sequences that are absolutely summable be $\ell^1$. The parameter space $\Theta$ is a subset of the Banach space $\mathbb{H} = (\mathbb{R}^p \times \ell^1)$ and any $h \in \mathbb{H}$ is partitioned $h = (h_\alpha, h_G)$. We use the sum norm $\|h\|_\mathbb{H} = \|h_\alpha\|_{\mathbb{R}^p} + \|h_G\|_{\ell^1}$ where $\|h_\alpha\|_{\mathbb{R}^p} = \left( \sum_{j=1}^p h_{\alpha,j}^2 \right)^{1/2}$ and $\|h_G\|_{\ell^1} = \sum_{j=0}^\infty |h_{G,j}|$, and $h_{\alpha,j}$ and $h_{G,j}$ are, respectively, the $j$th elements of $h_\alpha$ and $h_G$. Thus, $\sqrt{T} \left( (\hat{\alpha}, \hat{G}) - (\alpha, G) \right)$ is considered a random element of the space $\mathbb{H}$.

Drost et al. (2009) (DvdAW hereafter) establish asymptotic normality and efficiency for the NPMLE in the INAR class. (See Drost et al., 2008, for related work). Let $\alpha^*$ and $G^* = \{g^*_r\}$ be the true values of the binomial probabilities and the arrivals distribution in (1), and $\theta^* = (\alpha^*, G^*)$. When $G^*$ has finite $p + 4$ moments and $g_0^* < 1$, DvdAW show that the NPMLE is regular (van der Vaart, 1998, Section 25) and asymptotically Gaussian; i.e.

$$\sqrt{T} \left[ \hat{\theta} - \theta^* \right] = \sqrt{T} \left[ (\hat{\alpha}, \hat{G}) - (\alpha^*, G^*) \right] \leadsto (N_\alpha, \mathcal{N}_G),$$

where $N_\alpha$ is a $p$-dimensional zero mean normal random variable, $\mathcal{N}_G$ is a centered Gaussian process that lives in $\ell^1$ and $\leadsto$ means weak convergence. In addition, DvdAW prove asymptotic efficiency in the sense of the Hajek convolution theorem (see van der Vaart, 1998, Theorem 25.20). Let $(\hat{\alpha}, \hat{G})$ be a regular estimator, then

$$\sqrt{T} \left[ (\hat{\alpha}, \hat{G}) - (\alpha^*, G^*) \right] \leadsto (N_\alpha + W, \mathcal{N}_G + \mathcal{W}),$$

where $W$ and $\mathcal{W}$ are ‘noise’ processes independent of the Gaussian process $(N_\alpha, \mathcal{N}_G)$. Thus, any other regular estimator has a covariance structure that ‘exceeds’ that of the NPMLE and the NPMLE is the best regular estimator. This is the sense in which nonparametric asymptotic efficiency is understood.

### 2.2 Efficient Forecasting in the INAR Class

In the first instance we deal with the one-step-ahead forecast and thereafter the $m$-step-ahead case. In the model (1) the one-step-ahead forecast probability, $P(X_{T+1} = i_0|X_T = x_T, \ldots, X_{T-p+1} = x_{T-p+1}; \theta)$, for any $i_0 \in \mathbb{Z}$, is, again, a convolution of $p$ binomials and the innovation distribution and this convolution is written more succinctly as

$$f_{i_0|i_1,\ldots,i_p}^{(1)}(\theta) = f_{i_0|i_1,\ldots,i_p}^{(1)}(\theta)$$

using (2). The one-step-ahead predictive distribution is therefore

$$f_{i_1,\ldots,i_p}^{(1)}(\theta) = \left\{f_{i_0|i_1,\ldots,i_p}^{(1)}(\theta), i_0 \in \mathbb{Z} \right\}$$

(7)
and $F_{i_1,\ldots,i_p}^{(1)}(\theta)$ is a mapping from the Banach space $\mathbb{H}$ to the Banach space $\ell^1$, as defined in Section 2.1. In probabilistic forecasting the objective is to estimate the one-step-ahead distribution $F_{i_1,\ldots,i_p}^{(1)}(\theta)$. In applications, $\theta$ in (7) is to be replaced by the NPMLE estimator $\hat{\theta} = \left(\hat{\alpha}, \hat{G}\right)$, which is asymptotically efficient in the sense of Section 2.1. This suggests that $F_{i_1,\ldots,i_p}^{(1)}(\hat{\theta})$ may inherit the properties of $\hat{\theta}$ and also be asymptotically efficient, if the map $F_{i_1,\ldots,i_p}^{(1)}(\theta) : \mathbb{H} \mapsto \ell^1$ is smooth enough. Smoothness requires the existence of a derivative map $\hat{F}_{i_1,\ldots,i_p}^{(1)} : H \mapsto \ell^1$ between the same two spaces. To motivate the structure of such a map consider the total differential of (2) with respect to $\alpha_k$, $k = 1, 2, \ldots, p$, and a finite number of probabilities $g_r$. This involves specifying the partial derivatives, weighting them linearly with an increment and summing. The expression (9) in Theorem 1 below performs that calculation and allows for an infinite number of probabilities. The theorem then shows that (9) is indeed a derivative map. The proof is given in the Appendix.

**Theorem 1** Defining $F_{i_1,\ldots,i_p}^{(1)}(\theta_T)$ as in (8), the map $F_{i_1,\ldots,i_p}^{(1)} : \mathbb{H} \mapsto \ell^1$ is Frechet differentiable with derivative $\hat{F}_{i_1,\ldots,i_p}^{(1)}(h)$, where $\hat{F}_{i_1,\ldots,i_p}^{(1)} : \mathbb{H} \mapsto \ell^1$ is a bounded linear operator with typical element

$$
\hat{f}_{i_0|i_1,\ldots,i_p}^{(1)}(h) = \sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)} h_{G,i_0-(j_1+\ldots+j_p)} \prod_{k=1}^p p_{jk|i_k}(\alpha_k) + \\
\sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)} g_{i_0-(j_1+\ldots+j_p)} \sum_{k=1}^p \frac{\partial p_{jk|i_k}(\alpha)}{\partial \alpha_k} h_{\alpha,k} \prod_{l=1, l\neq k}^p p_{jk|i_l}(\alpha_k). \tag{9}
$$

In particular for $\|h\|_{\mathbb{H}} < 1$ we have $\left\| F_{i_1,\ldots,i_p}^{(1)}(\theta + h) - F_{i_1,\ldots,i_p}^{(1)}(\theta) - \hat{F}_{i_1,\ldots,i_p}^{(1)}(h) \right\|_{\ell^1} = o(\|h\|_{\mathbb{H}})$.

Since the NPMLE $\hat{\theta}$ is asymptotically efficient under the DvdAW conditions specified in Section 2.1 and since Frechet differentiability implies Hadamard differentiability, Proposition 2 of van der Vaart (1995) and Theorem 1 together imply that $F_{i_1,\ldots,i_p}^{(1)}(\hat{\theta})$ is also asymptotically efficient for the one-step-ahead distribution. Thus, $F_{i_1,\ldots,i_p}^{(1)}(\hat{\theta})$ is the AEEF (for $m = 1$) in the INAR class.

We can interpret what is meant by an asymptotically efficient forecast distribution more concretely via the Hajek convolution theorem. Since, as in (6), $\sqrt{T} \left[\hat{\theta} - \theta^*\right] \sim (N_\alpha, \mathcal{N}_G)$ and since the spaces $\mathbb{H}$ and $\ell^1$ are linear spaces, it is a consequence of Theorem 20.8 of van der Vaart (1998) that

$$
\sqrt{T} \left( F_{i_1,\ldots,i_p}^{(1)}(\hat{\theta}) - F_{i_1,\ldots,i_p}^{(1)}(\theta^*) \right) \sim \hat{F}_{i_1,\ldots,i_p}^{(1)}(N_\alpha, \mathcal{N}_G). \tag{10}
$$

It follows from Theorem 1 above that $\hat{F}_{i_1,\ldots,i_p}^{(1)}(N_\alpha, \mathcal{N}_G)$ is also a Gaussian process by the linearity of $\hat{F}_{i_1,\ldots,i_p}^{(1)}$. Thus, any other suitably standardised forecast mapping, based on a
regular estimator of $\theta$ must have a limit distribution with a covariance process no smaller than that of $F_{i_1,\ldots,i_p}^{(1)}(\hat{\theta})$ by the Hajek convolution theorem.

When $p = 1$, the one-step-ahead forecast is quite simple and may be computed, for $i \in \mathbb{Z}$, as

$$P[X_{T+1} = i | X_T = x_T; \theta] = f_{i|x_T}^{(1)}(\theta) = \sum_{j=0}^{i/x_T} p_{j|x_T}(\alpha_1) g_{i-j},$$

(11)

where the binomial probabilities, $p_{j|x_T}(\alpha_1)$, are given in (3). The estimated distribution, \( \{ P[X_{T+1} = i | X_T = x_T; \hat{\theta}] \}, i \in \mathbb{Z} \), where $\hat{\theta}$ is the NPMLE, is asymptotically efficient for the distribution \( \{ P[X_{T+1} = i | X_T = x_T; \theta] \}, i \in \mathbb{Z} \) under the DvdAW conditions.

The treatment of the $m$-step-ahead case, for $m > 1$, is facilitated by the fact that the model (1) may also be considered as a Markov Chain from $\mathbb{Z}^{p+1}$ to $\mathbb{Z}^{p+1}$. This interpretation allows the $m$-step-ahead prediction distributions to be defined recursively (see, for example, Resnick, 1992, Sec 2.3, and Bu and McCabe, 2008, in addition to the computational details provided in the following section). That is,

$$f_{i_0|i_1,\ldots,i_p}^{(m)}(\theta) = \sum_{u=0}^{\infty} f_{i_0|u,i_1,\ldots,i_{p-1}}^{(m-1)}(\theta) f_{u|i_1,\ldots,i_p}^{(1)}(\theta),$$

(12)

and

$$F_{i_1,\ldots,i_p}^{(m)}(\theta) = \{ f_{i_0|i_1,\ldots,i_p}^{(m)}(\theta) : i_0 \in \mathbb{Z} \}. \quad (13)$$

It also follows, for any $m$, that $F_{i_1,\ldots,i_p}^{(m)}(\theta) : \mathbb{H} \mapsto \ell^1$ are mappings between Banach spaces. This mapping is also sufficiently smooth, as a consequence of the following theorem, with proof of the theorem given in the Appendix.

**Theorem 2** Assume $\sum_{u=0}^{\infty} (u^2 s^u)^p g_u < \infty$ for some $s > 1$. For each $i_0 \in \mathbb{Z}$, define recursively, using (7) and (12),

$$\hat{f}_{i_0|i_1,\ldots,i_p}^{(m)}(h) = \sum_{u=0}^{\infty} \hat{f}_{i_0|u,i_1,\ldots,i_{p-1}}^{(m-1)}(h) f_{u|i_1,\ldots,i_p}^{(1)}(\theta) + \sum_{u=0}^{\infty} f_{i_0|u,i_1,\ldots,i_{p-1}}^{(m-1)}(\theta) \hat{f}_{u|i_1,\ldots,i_p}^{(1)}(h)$$

and set $\hat{F}_{i_1,\ldots,i_p}^{(m)}(h) = \{ \hat{f}_{i_0|i_1,\ldots,i_p}^{(m)}(h) : i_0 \in \mathbb{Z} \}$. Then the map $F_{i_1,\ldots,i_p}^{(m)} : \mathbb{H} \mapsto \ell^1$ is Frechet differentiable. That is, $\hat{F}_{i_1,\ldots,i_p}^{(m)} : \mathbb{H} \mapsto \ell^1$ is a bounded linear operator that satisfies

$$\left\| F_{i_1,\ldots,i_p}^{(m)}(\theta + h) - F_{i_1,\ldots,i_p}^{(m)}(\theta) - \hat{F}_{i_1,\ldots,i_p}^{(m)}(h) \right\|_{\ell^1} = o(\|h\|_{\mathbb{H}})$$

for any $m > 1$.

Thus, the $m$-step-ahead forecast distribution is asymptotically efficient in the sense of the Hajek convolution theorem for any $m \geq 1$. The condition $\sum_{u=0}^{\infty} (u^2 s^u)^p g_u < \infty$ of Theorem 2 (not required in the one-step-ahead case) is satisfied, for any $p$, by many well known...
distributions (e.g. the Poisson and the negative binomial) and trivially for any distribution with finite support. For a Poisson distribution with parameter \( \lambda \) (\( Pois(\lambda) \)),

\[
\sum_{u=0}^{\infty} (u^2 s^u)^p g_u = \sum_{u=0}^{\infty} u^2 p e^{-\lambda} (s^p \lambda)^u u! = e^{p \lambda} \sum_{u=0}^{\infty} u^2 p e^{-s^p \lambda} (s^p \lambda)^u u! < \infty
\]

for any \( s \) because a \( Pois(s^p \lambda) \) distribution has finite \( 2p \) moments. For a negative binomial distribution,

\[
g_u = \frac{\Gamma(v+u)}{\Gamma(v) \Gamma(u+1)} \pi^u (1-\pi)^v, \quad v > 0, \quad 0 < \pi < 1,
\]

we have \( \sum_{u=0}^{\infty} (u^2 s^u)^p g_u = \frac{(1-\pi)^v}{\Gamma(v)} \sum_{u=0}^{\infty} u^2 p \Gamma(v+u) (s^p \pi)^u \), which is finite for any \( s < \pi^{-1/p} \), as can be seen by applying Stirling’s formula to the gamma functions in the summation.

### 2.3 Computational Details

For \( p \geq 1 \), the likelihood function (conditional on \( p \) initial values) is the product of the conditional probabilities:

\[
P[X_t = x_t | X_{t-1} = x_{t-1}, \ldots X_{t-p} = x_{t-p}; \theta] = \prod_{x_t \wedge x_{t-1}} p_{j_1|x_{t-1}}(\alpha_1) \prod_{x_{t-j_1} \wedge x_{t-2}} p_{j_2|x_{t-2}}(\alpha_2) \prod_{x_{t-(j_1+j_2)} \wedge x_{t-3}} p_{j_3|x_{t-3}}(\alpha_3) \prod_{x_{t-(j_1+j_2+\ldots+j_{p-1})} \wedge x_{t-p}} p_{j_p|x_{t-p}}(\alpha_p) g_{x_t-(j_1+j_2+\ldots+j_p)}
\]

for \( t = p + 1, ..., T \). The asymptotically efficient estimate of the one-step-ahead forecast distribution,

\[
\{P[X_{T+1} = i_0 | X_T = x_T, X_{T-1} = x_{T-1}, \ldots X_{T-p} = x_{T-p}; \theta], \quad i_0 \in \mathbb{Z}\}
\]

is produced by simply substituting the NPMLE of \( \theta = (\alpha_k; k = 1, 2, \ldots, p, \{g_r\}) \) into the expression in (15) and evaluating the conditional probabilities over the support \( i_0 = 0, 1, \ldots, K \), with \( K \) chosen to ensure that all predictive mass is estimated. However, extending Bu and McCabe (2008) to the nonparametric case, this calculation is given an alternative representation, which is particularly useful for the \( m > 1 \) step-ahead forecast. Specifically, the \( INAR(p) \) process can be viewed as a Markov chain, with \( X_t \) assuming (in practice) a finite number of values \( \{0, 1, \ldots, K\} \) at time \( t \), and the states of the system given by \( p - \text{tuples} \) of possible values. Hence, at time \( T \), the chain could be in any one of the \((K+1)^p\) states:

\[
S = \{ (0,0,\ldots,0), (0,1,\ldots,0), \ldots, (0, K, \ldots, 0), \ldots, (K, \ldots, 0), \ldots, (K, K, \ldots, K) \}
\]
as \((X_{T-(p-1)}, \ldots, X_T)\) assume values \((j_p, j_{p-1}, \ldots, j_1)\) \(\in S\). Define the \((K+1)^p \times (K+1)^p\) matrix of transition probabilities \(Q\) as having elements:

\[
P[X_{T+1} = i_0, X_T = i_1, \ldots, X_{T-(p-1)} = i_{p-1} | X_T = j_0, X_{T-1} = j_1, \ldots, X_{T-p} = j_{p-1}; \theta]
\]

\[
= P[X_{T+1} = i_0 | X_T = j_0, X_{T-1} = j_1, \ldots, X_{T-p} = j_{p-1}; \theta] \quad \text{for } i_1 = j_0, \ldots, i_{p-1} = j_{p-2}
\]

\[
= 0 \quad \text{for any } i_1 \neq j_0, \ldots, i_{p-1} \neq j_{p-2},
\]

and the \((K+1)^p \times 1\) vectors \(\pi_T\) and \(\pi_{T+1}\) as having (respectively) elements:

\[
P[X_T = j_0, X_{T-1} = j_1, \ldots, X_{T-p} = j_{p-1}; \theta] \quad \text{and}
\]

\[
P[X_{T+1} = i_0, X_T = i_1, \ldots, X_{T-(p-1)} = i_{p-1}; \theta].
\]

The conditional distribution in (16) can thus obtained by calculating \(\pi_{T+1}^T = \pi_T^TQ\) and selecting from \(\pi_{T+1}\) the probabilities attached to \(X_{T+1}\), over \(i_0 = 0, 1, \ldots, K\), conditional on the observed values \(X_T = x_T, X_{T-1} = x_{T-1}, \ldots, X_{T-p} = x_{T-p}\). For \(m > 1\) steps ahead, we exploit the theory of Markov chains to define \(\pi_T^T = \pi_T^TQ^m\), with the \(m\)-step-ahead forecast distribution:

\[
\{P[X_{T+m} = i_0 | X_T = x_T, X_{T-1} = x_{T-1}, \ldots X_{T-p} = x_{T-p}; \theta] ; i_0 = 0, 1, \ldots, K\}
\]

extracted from \(\pi_{T+m}^T\). The asymptotically efficient estimate of the \(m\)-step ahead forecast distribution is produced by simply replacing \(\theta\) by \(\hat{\theta}\) in all calculations.

In the case where the data is clearly interpretable as the outcome of a queuing (or stock, or birth and death) process, the choice of \(p = 1\) is appropriate. In the case where a branching process interpretation applies, a choice of \(p\) needs to be made, prior to the AEEF being calculated. As in the case of lag length selection in more standard time series settings, this decision can be made via informal preliminary diagnostic testing, or via some sort of more formal model selection criterion (such as the Akaike information criterion). Perhaps more appropriately, however, given the focus here on forecast performance, the AEEF could be calculated for each INAR model associated with a different value of \(p\) (within a reasonable range), with \textit{ex-post} evaluation of predictive accuracy (using realized values) then used to select one optimal forecast distribution from the set associated with the alternative values of \(p\).

3 Finite Sample Performance in the INAR Class

In Section 2.2 we proved the asymptotic efficiency of the nonparametric estimator of the \(m\)-step-ahead forecast distribution in the \(INAR(p)\) model for \(m \geq 1\). In this section we document the finite sample performance of the nonparametric estimator, in comparison with an
estimator of the forecast distribution based on a correctly and incorrectly specified maximum likelihood estimator (MLE) respectively. We consider the INAR\((p)\) data generating process in (1) with \(p = 1\) and \(\alpha_1 = 0.6\). We assume \(\varepsilon_t\) to be distributed, respectively, as Poisson, \(\text{Pois}\(\lambda = 2\)\), binomial, \(\text{Bin}(n = 4; \pi = 0.4)\), and negative binomial, \(\text{NBin}(v = 5; \pi = 0.3)\) (with mass function as defined in (14)). These distributions are representative, respectively, of equi-, under- and over-dispersed distributions for the arrivals. Given the structure of the INAR(1) model, these specifications produce, in turn, low count data that are also equi-, under- and over-dispersed respectively; see e.g. Pavlopoulos and Karlis (2008). The value of \(\alpha_1\) is selected to approximate the empirical \(\text{NPMLE}\) of \(\alpha_1\) for the data analysed in Section 4. We focus on the one-step-ahead forecast distribution (i.e. \(m = 1\)), and for notational convenience we denote \(f^{(1)}_{i|\theta_T}(\theta), i \in \mathbb{Z}\) (in (11)) by \(f_i, i \in \mathbb{Z}\), using the notation \(\{f_i\}\) to denote the full sequence of forecast probabilities over \(\mathbb{Z}\).

The performance of the AEEF is compared with that of the estimator of \(\{f_i\}\) based on the application of a parametric MLE to the INAR(1) model with Poisson arrivals; i.e. the canonical \(\text{PAR}\) model. This MLE-based estimate of the forecast distribution (denoted hereafter by MLE-P) is based on a correctly specified model when the true arrivals are Poisson, and would be expected to perform better than the AEEF in this case. In the case where the true arrivals are either binomial or negative binomial, the MLE-P is based on a misspecified model. The interest here is in ascertaining if, and to what extent, the AEEF outperforms the MLE-P. (For brevity, in what follows we refer to the MLE-P itself as being ‘correctly specified’ and ‘misspecified’ in these respective cases). All results are based on 10000 replications of \(\{f_i\}\).

Fix a value for \(i\) and let \(\widehat{E} \left[ \left( \hat{f}_i - f_i \right)^2 \right] \) be the simple average of the squared errors \(\left( \hat{f}_i - f_i \right)^2\) over the 10000 replications, where \(\hat{f}_i\) denotes the value of the AEEF at \(i\). The ‘AV. MSE’ figures recorded in the first row of results in Table 1 are estimates of the mean squared error of \(\{\hat{f}_i\}\), calculated by averaging \(\widehat{E} \left[ \left( \hat{f}_i - f_i \right)^2 \right] \) over the support \(i = 0, 1, \ldots, K\), with \(K\) chosen to ensure that all predictive mass is estimated. The figures recorded (in parentheses) in the row immediately beneath give the ratio of the AV. MSE for the AEEF to the corresponding measure for the MLE-P. Values for the AV. MSE ratio that are less than one indicate that the AEEF is superior in terms of this measure of accuracy.

The figures presented in the two lower panels in Table 1 refer to the segments of the support corresponding, respectively, to the upper 10% tail and the lower 25% tail of the true predictive, \(\{f_i\}\). The AV. MSE figures in each of these panels record \(\widehat{E} \left[ \left( \hat{f}_i - f_i \right)^2 \right] \) averaged over the relevant part of the support, and measure the accuracy with which the AEEF estimates (both in absolute terms and relative to the MLE-P) the probability of occurrence of very large (or very small) counts.
The AV. BIAS figures presented in the first row of each of the two panels in Table 2 record $\bar{E} (\hat{f}_i - f_i)$ averaged over the support for the upper 10\% tail and the lower 25\% tail, and capture the phenomenon of under- or over-estimation of the probability of very large (or very small) counts. (The estimated bias across the full support of the count variable is equal to zero due to the summation restriction satisfied by both the estimated and true forecast distributions). The figures recorded (in parentheses) in the rows immediately below the AV. BIAS measures for the AEEF give the ratio of the measure for the AEEF to the corresponding measure for the MLE-P. Again, values for the ratio that are less than one indicate that the AEEF is superior in terms of this measure of accuracy. Positive values for the AV. BIAS ratios indicate that both the AEEF and the MLE-P either under- or over-estimate the relevant tail mass.

***** TABLES 1 AND 2 HERE *****

As is indicated by all figures in the first row of Table 1, the AV. MSE for the AEEF declines monotonically with the sample size, in accordance with the theoretical consistency of the estimator. This result also obtains in the lower 25\% tail and, in all cases but one, in the upper 10\% tail. As also expected, the AV. MSE ratios in the second row of each panel (and in the far left portion of the table) indicate that the correctly specified MLE-P is more accurate in finite samples than the AEEF, according to this measure of accuracy, when the true distribution is Poisson, with all AV. MSE ratios exceeding one in this case.

When the true DGP has binomial arrivals (figures recorded in the middle portion of Table 1), the AEEF has lower AV. MSE than the misspecified MLE-P for $T = 500$ and $T = 1000$, over both the full support (top panel) and the upper and lower tails (lower two panels). The AEEF is only slightly less accurate - with AV. MSE ratios just greater than one - in two of the three cases for $T = 100$. The superiority of the AEEF over the MLE-P uniformly increases with $T$, for all cases documented under binomial arrivals.

In the case of true negative binomial arrivals (figures recorded in the far right portion of Table 1), the AEEF is superior to the misspecified MLE-P for the two larger sample sizes, both across the full support and in the lower 25\% tail. The largest sample size ($T = 1000$) is required for the AEEF to exhibit smaller AV. MSE than the MLE-P in the upper 10\% tail. The estimator is less accurate, according to this measure, than the MLE-P for $T = 100$ in all cases, with AV. MSE ratios ranging from 2.625 (in the lower 25\% tail) to 4.080 (in the upper 10\% tail).

With reference to the bias results in Table 2, the AEEF is uniformly superior to the (misspecified) MLE-P, under both binomial and negative binomial arrivals, with the MLE-P having AV. BIAS figures that range up to 670 times larger than the corresponding values for the AEEF. The numbers in the middle portion of the table can be used to deduce

11
the result that the $MLE-P$ uniformly *overestimates* both the upper 10% and lower 25% tail probabilities when the true arrivals are binomial and the data (unconditionally) under-dispersed as a consequence. When the true arrivals are negative binomial and the data thus (unconditionally) over-dispersed, the $MLE-P$ uniformly *underestimates* both the upper 10% and lower 25% tail probabilities. Interestingly, under Poisson arrivals, in which case the $MLE-P$ is correctly specified, the bias of the $AEEF$ in estimating the lower 25% tail is *smaller* than that associated with the $MLE-P$, for all sample sizes. This superiority does not obtain in the case of estimating the upper 10% tail. The $AEEF$ exhibits no systematic tendency to either under- or overestimate the tails of the forecast distribution, under any process for the arrivals.

## 4 Empirical Application

### 4.1 Data Description

In this section we apply the $AEEF$ to an empirical series of count data. The series comprises $T = 480$ counts of ‘iceberg’ sell orders (asks) in the order book (up to and including the fifth best order only) of Deutsche Telekom stock (denoted hereafter by DEUT), collected every 10 minutes on the XETRA system of the Deutsche Borse. The data is recorded over the eight hours of each of the last 10 trading days (last two trading weeks) in the first quarter of 2004.

Iceberg orders are so-called because only a portion of the volume of the order, or the ‘tip of the iceberg’, is revealed in the order book. Such orders constitute only a small proportion of the total number of limit book orders; e.g. only 8% of shares traded in the set of German stocks analysed by Frey and Sandas (2008). Nevertheless, they have been shown to exert a significant impact on trading behaviour - and the subsequent dynamic behaviour of transaction prices - as traders adjust their bid (or ask) prices in the face of the ‘hidden liquidity’ associated with the icebergs. Not only are traders unaware of the extent of the hidden volume of iceberg orders, the very existence of such orders is not made explicit by the exchange at the time of trading. Hence, traders themselves need to adopt various strategies for identifying the number and size of iceberg orders; see Frey and Sandas for further discussion.

Over any 10 minute time period $t$, the number of iceberg orders, $X_t$, is the sum of the number of orders remaining from the previous 10 minute period, waiting for execution, $\alpha_t \circ X_{t-1}$, and the number of new iceberg orders placed in the book (or ‘arrivals’), $\varepsilon_t$. All iceberg orders are deleted from the book at the end of the trading day, even if not executed. Note that although the order book is scanned every 10 minutes only to the depth of the best five trades, it is quite possible for an iceberg trade to be among the best five bids at
any instance during that 10 minute period, leading to more than five iceberg trades being recorded after any 10 minute interval. However, the DEUT data over this particular sample period assumes values of zero to five (inclusive) only, due to the infrequency with which iceberg orders occur. The sample proportions associated with the values \{0, 1, 2, 3, 4, 5\} are \{0.479, 323, 119, 0.058, 0.017, 0.004\}. The mean and standard deviation of the sample counts are 0.823 and 1.009 respectively - indicating some overdispersion in the data - with there being no evident intraday (diurnal) pattern in the data to be modelled. The sample autocorrelation function of the DEUT data displays the characteristic exponential decline of a short-memory autoregressive process, with a first-order autocorrelation coefficient of 0.576 and significant coefficients up to and including lag 12, indicating that there is indeed dependence to be modelled and predictive power in the data. Given that the data may clearly be interpreted as time series observations on a queue or stock variable, the INAR(1) specification is inherently suitable; the NPMLE of $\alpha_1$ assumes a value of 0.551.

4.2 Assessment of Sampling Error

In addition to producing the efficient point estimator of the forecast distribution at a given horizon, we propose a method for assessing the effect of sampling variation. In particular, we aim to describe variation in the full predictive distribution and to present this information in a way that is easily understood. To this end, we use a re-sampling method to allow the effect of sampling fluctuations in the estimated forecast distribution to be visualized, while retaining the non-negativity and summation to unity properties of probabilities.

We adopt the subsampling approach of Politis, Romano and Wolf (1999) (PRW hereafter). While not dissimilar to the bootstrap approaches of Carlstein (1986), Kunsch (1989) and Liu and Singh (1992) for stationary time series, the subsampling method is more generally applicable and is, indeed, much easier to validate in abstract Banach space settings such as those in Theorems 1 and 2 above. First we describe the subsampling procedure, including a data-dependent method for choosing a number $b$ which is the size of the subsamples. We then give a theorem that justifies the use of the subsampling procedure in the current setting.

Implementation of the subsampling method involves the following steps:

1. Obtain $T - b + 1$ subsamples $Y_1 = (X_1, \ldots, X_b), Y_2 = (X_2, \ldots, X_{b+1}), \ldots, Y_{T-b+1} = (X_{T-b+1}, \ldots, X_T)$.

2. Use the NPMLE of $\theta, \hat{\theta}_{b,t}$, computed from $Y_t$ and the observed values, $x_T, x_{T-1}, \ldots, x_{T-(n-1)}$, to compute the $m$-step ahead forecast distribution $F_{t_1, \ldots, t_p}^{(m)} (\hat{\theta}_{b,t})$; $m \geq 1$. 
3. Calculate the metric $d_{b,t} = \sqrt{T} \left\| F_{{\hat{\theta}_b,t}}^{(m)} - F_{{\hat{\theta}_{\text{true}}}}^{(m)} \right\|_1$, where $F_{{\hat{\theta}_b,t}}^{(m)}$ is the estimated forecast distribution based on the empirical data and $\hat{\theta}$ the NPMLE.

4. Find the 95th percentile of $\{d_{b,1}, \ldots, d_{b,T-b+1}\}$, $d_{b}^{0.95}$, and the corresponding distribution $F_{0.95}$.

Then, relative to the replicated distributions and in terms of the $\| \cdot \|_1$ distance from $F_{{\hat{\theta}_b,t}}^{(m)}$, the chances of seeing a distribution as or more ‘extreme’ than $F_{0.95}$ is 5%.

To choose $b$ in practice we follow the suggestion given in PRW (Chapter 9):

a. For each $b \in \{b_{\text{small}}, \ldots, b_{\text{big}}\}$ carry out Steps 1 to 4 above to compute $d_{b}^{0.95}$.

b. For each $b$ compute $VI_b$ as the standard deviation of the $2k + 1$ adjacent values $\{d_{b-k}^{0.95}, \ldots, d_{b+k}^{0.95}\}$ (for $k = 2$).

c. Choose $\hat{b}$ to minimise $VI_b$.

The essence of demonstrating the validity of subsampling (and bootstrap) procedures is to show that probability statements made on the basis of the replicated distribution are (asymptotically) the same as those based on the actual sampling distribution. So, for a suitable norm (like $d_{b,t}$ above) define the empirical distribution of the replications to be

$$Q_{T,b}(x) = \frac{1}{T-b+1} \sum_{t=1}^{T-b+1} I \left[ \sqrt{T} \left\| F_{{\hat{\theta}_b,t}}^{(m)} - F_{{\hat{\theta}_{\text{true}}}}^{(m)} \right\| < x \right],$$

and $Q_T(x)$ to be the law of the sampling distribution of $\sqrt{T} \left\| F_{{\hat{\theta}_b,t}}^{(m)} - F_{{\hat{\theta}_{\text{true}}}}^{(m)} \right\|: \theta^*$ being the true parameter value. In the expression in (17), $I[.]$ is the indicator function.

The validity of the subsampling method requires that $Q_{T,b}(x) - Q_T(x)$ converge to zero in a suitable sense. This convergence is the content of the following theorem, the proof of which is given in the Appendix.

**Theorem 3** Assume that the model (1) holds for a process $X_t$. When $b \to \infty$ and $T \to \infty$ with $b/T \to 0$,

$$\rho_L(Q_{T,b}, Q_T) \to^p 0,$$

where $\rho_L$ is the bounded Lipschitz metric.

As a consequence of Theorem 3, for large enough $T$, statements based on the empirical distribution of $\sqrt{T} \left\| F_{{\hat{\theta}_b,t}}^{(m)} - F_{{\hat{\theta}_{\text{true}}}}^{(m)} \right\|$ are equivalent to statements based on the (unknown) sampling distribution of $\sqrt{T} \left\| F_{{\hat{\theta}_b,t}}^{(m)} - F_{{\hat{\theta}_b,t}}^{(m)}(\theta^*) \right\|$. Hence, we can use the subsampling procedure to make statements like: “The probability of seeing an estimated distribution for which the deviation from the true forecast distribution, $F_{{\hat{\theta}_b,t}}^{(m)}(\theta^*)$ (in the metric) is as or more extreme than $F_{0.95}$ (calculated from $Q_{T,b}(x)$) is 5%.”
4.3 Empirical Forecast Results

In Figure 1a we reproduce the estimated one-step-ahead ($m = 1$) forecast distribution for the DEUT data, along with some extreme subsampled distributions estimated from $B = T - b + 1$ replications, with $b = 235$ (selected as per Steps a. to c. above). Given that extreme values of the metric can, potentially, be associated with quite different shapes in the forecast distributions, we record the (subsampled) forecast distribution at the 95th percentile and the distributions ranked on either side of the 95th percentile.

The estimated forecast distribution assigns only 32% probability to the event of no DEUT iceberg order being included in the five best bids during the last 10 minutes of the last trading day of the first quarter in 2004. This indicates that some degree of hidden liquidity was very likely to be available ($\Pr ob(X_{T+1} \geq 1 = 0.68$), and needed to be catered for in trading decisions. The ‘extreme’ distributions on either side of the 95th percentile indicate an increase in probability to the event of zero bids, and a corresponding decrease in the probability of some degree of hidden liquidity being present. However, the extreme distribution at the 95th percentile allocates less probability mass to zero bids and, correspondingly, more probability to the presence of at least one iceberg bid. These results illustrate the way in which sampling variability serves to shift probability mass about the support of the predictive distribution and thereby alter the qualitative nature of the conclusions drawn from the analysis.

In Figure 1b, we reproduce the estimated forecast distributions for $m = 5$ steps ahead, along with the three subsampled distributions centred at the 95th percentile (based on $b = 260$). The five-step-ahead distributions are estimated using the Markov chain method described in Section 2.3. Due to the stationarity of the model, the forecast distribution five days out is closer (than is the one-step-ahead forecast) to the corresponding unconditional distribution, as estimated by the sample proportions (recorded in Figure 1b also).

5 Discussion

In this paper we develop an approach to forecasting integer-valued time series data. The method involves estimating the forecast distribution of the discrete random variable, thereby producing coherent forecasts that quantify the full uncertainty associated with future counts. For the broad INAR class an asymptotically (nonparametrically) efficient estimator of the forecast distribution ($AEEF$) is produced via likelihood methods. Simulation results for the $INAR(1)$ model indicate that the $AEEF$ performs well even in moderately sized samples. Most notably, for sample sizes of 500 and 1000, the $AEEF$ is markedly less biased than misspecified parametric comparators when estimating the tails of the forecast distribution and, hence, the probability of extreme counts (both low and high) in future time periods.
Figure 1: One-step-ahead and five-step-ahead forecast distributions for DEUT iceberg counts

a. DEUT ICEBERG ORDERS
Estimated 1-Step-Ahead Forecast Distribution plus Sub-sampled Distributions
T = 480

b. DEUT ICEBERG ORDERS
Estimated 5-Step-Ahead Forecast Distribution plus Sub-sampled Distributions (and Sample Proportions)
T = 480
For such sample sizes, the AEEF is more accurate overall, in terms of both bias and mean squared error, than a misspecified parametric estimator of the forecast distribution.

We also present a valid subsampling method for assessing the effect of sampling variation in the AEEF that incorporates the non-negativity and summation properties of the probabilities involved. A data set that may be interpreted as the output of an INAR structure is analysed, with the forecast distribution estimated and sampling variation quantified via the subsampling technique.

The \textit{ex-ante} approach to forecasting adopted here differs, of course, from that typically adopted in the forecast literature, in which a forecast distribution is selected (from a set of alternatives) via \textit{ex-post} evaluation based on observed outcomes (see Dawid, 1984; Tay and Wallis, 2000; Gneiting \textit{et al}, 2007; Gneiting, 2008; and Geweke and Amisano, 2010 for examples and general discussion, and Czado \textit{et al.} for an extensive application of such evaluation techniques to discrete count data). However, our approach should not be viewed as a competitor to the fundamental principle of assessing distributional forecasts using realized outcomes. In fact, the two approaches complement each other. The existence of a suitable model class affords the advantage of optimality whilst, at the same time, empirical validation guards against unforeseen circumstances such as, for example, an unanticipated structural break in the data generating process. Indeed, even in cases where the class is not \textit{inherently} suitable for a particular count data set, producing the (nonparametrically) efficient forecast distribution within the INAR class is still a sensible first step prior to comparing - via \textit{ex-post} methods - with relevant alternatives from outside the class.

Finally, the inclusion of covariates in the general \textit{INAR} model is potentially important for some empirical applications. In particular, the incorporation of covariate affects, such as seasonality, in the nonparametric specification of the arrivals, is an open problem and is currently being explored by the authors.
Appendix

The following preliminary lemma is used in the proofs below. A proof is available from the authors on request.

Lemma 1 If $p_{ji}(\alpha)$ is a binomial probability $\binom{i}{j} \alpha^j (1-\alpha)^{i-j}$ and $h$ is a constant then

$$\sum_{j=0}^{i} \left| p_{ji}(\alpha + h) - p_{ji}(\alpha) - \frac{\partial p_{ji}(\alpha)}{\partial \alpha} h \right| \leq 3hi^2 (i-1)(1+|h|)^{i-2} \leq 3hi^2 (1+|h|)^i \quad (18)$$

and

$$\sum_{j=0}^{i} \left| p_{ji}(\alpha + h) - p_{ji}(\alpha) \right| \leq 2|h|i(1+|h|)^{i-1} + h^2i(i-1)(1+|h|)^{i-2}. \quad (19)$$

If $|h| < 1$ then this latter bound can be reduced to

$$\sum_{j=0}^{i} \left| p_{ji}(\alpha + h) - p_{ji}(\alpha) \right| \leq 3|h|i^2(1+|h|)^i.$$

We also use the well known results on binomial thinning that $\alpha \circ (x_1 + x_2) = \alpha \circ x_1 + \alpha \circ x_2$ and $\Pr \left( \bigcirc_{k \text{ times}} x = j \right) = p_{j|x}(\alpha^k)$.

Proof of Theorem 1

From (7) and (9) we obtain the expression

$$f_{i_0|1_1,..,i_p}(\theta + h) - f_{i_0|1_1,..,i_p}(\theta) - \sum_{(j_1,..,j_p) \in J(i_0,..,i_p)} g_{i_0-(j_1+...+j_p)} \left\{ \prod_{k=1}^{p} p_{j_k|i_k}(\alpha_k + h_{\alpha,k}) - \prod_{k=1}^{p} p_{j_k|i_k}(\alpha_k) \right\}$$

$$- \sum_{k=1}^{p} \frac{\partial p_{j_k|i_k}(\alpha_k)}{\partial \alpha_k} h_{\alpha,k} \prod_{l=1, l \neq k}^{p} p_{j_l|i_l}(\alpha_k)$$

$$+ \sum_{(j_1,..,j_p) \in J(i_0,..,i_p)} h_{G,i_0-(j_1+...+j_p)} \left\{ \prod_{k=1}^{p} p_{j_k|i_k}(\alpha_k + h_{\alpha,k}) - \prod_{k=1}^{p} p_{j_k|i_k}(\alpha_k) \right\}.$$

Straightforward rearrangements show that

$$\sum_{i_0=0}^{\infty} \sum_{(j_1,..,j_p) \in J(i_0,..,i_p)} = \sum_{j_1=0}^{i_1} \cdots \sum_{j_p=0}^{i_p} \sum_{i_0=0}^{\infty}.$$
and, hence, that

\[ \left\| F_{i_1, \ldots, i_p}^{(1)} (\theta + h) - F_{i_1, \ldots, i_p}^{(1)} (\theta) - \mathcal{F}_{i_1, \ldots, i_p}^{(1)} (h) \right\|_{\mathcal{F}_1} \]

\[ \leq \sum_{i_0 = 0}^{\infty} \left| f_{i_0 | i_1, \ldots, i_p}^{(1)} (\theta + h) - f_{i_0 | i_1, \ldots, i_p}^{(1)} (\theta) - f_{i_0 | i_1, \ldots, i_p}^{(1)} (h) \right| \]

\[ = \sum_{i_1 = 0}^{i_i} \cdots \sum_{i_p = 0}^{i_p} \prod_{k=1}^{p} p_{jk | i_k} (\alpha_k + h_{\alpha,k}) - \prod_{k=1}^{p} p_{jk | i_k} (\alpha_k) - \sum_{k=1}^{p} \frac{\partial p_{jk | i_k} (\alpha_k)}{\partial \alpha_k} h_{\alpha,k} \prod_{l=1}^{p} p_{jl | i_l} (\alpha_k) \]

\[ + \| h_G \| \sum_{i_1 = 0}^{i_i} \cdots \sum_{i_p = 0}^{i_p} \prod_{k=1}^{p} p_{jk | i_k} (\alpha_k + h_{\alpha,k}) - \prod_{k=1}^{p} p_{jk | i_k} (\alpha_k) \]

\[ = \sum_{i_1 = 0}^{i_i} \cdots \sum_{i_p = 0}^{i_p} \prod_{k=1}^{p} p_{jv | i_v} (\alpha_v) \times \]

\[ \left| p_{jk_1 | k_1} (\alpha_k + h_{\alpha,k_1}) - p_{jk_1 | i_1} (\alpha_k) - \frac{\partial p_{jk_1 | i_1} (\alpha_k)}{\partial \alpha_k} h_{\alpha,k_1} \right|\]

\[ + \| h_G \| \sum_{i_1 = 0}^{i_i} \cdots \sum_{i_p = 0}^{i_p} \prod_{k=1}^{p} p_{jv | i_v} (\alpha_v) \times \]

\[ \left| p_{jk_1 | k_1} (\alpha_k + h_{\alpha,k_1}) - p_{jk_1 | i_1} (\alpha_k) \right|\]

\[ + (1 + \| h_G \|) \sum_{i_1 = 0}^{i_i} \cdots \sum_{i_p = 0}^{i_p} \sum_{l=2}^{1 \leq k_1 < \ldots < k_l \leq p} \sum_{1 \leq v \leq p \atop v \neq k_1, \ldots, k_l} \prod_{v=1}^{p} p_{jv | i_v} (\alpha_v) \times \]

\[ \prod_{1 \leq u \leq l} \left| p_{jk_u | i_u} (\alpha_u + h_{\alpha,k_u}) - p_{jk_u | i_u} (\alpha_u) \right| . \]

The last step above uses the rearrangements

\[ \prod_{k=1}^{p} p_{jk_a | i_k} (\alpha_k + h_{\alpha,k}) - \prod_{k=1}^{p} p_{jk_a | i_k} (\alpha_k) \]

\[ = \sum_{l=1}^{p} \sum_{1 \leq k_1 < \ldots < k_l \leq p} \left( \prod_{1 \leq u \leq l} p_{jk_u | i_u} (\alpha_u + h_{\alpha,k_u}) - p_{jk_u | i_u} (\alpha_u) \prod_{1 \leq v \leq p \atop v \neq k_1, \ldots, k_l} p_{jv | i_v} (\alpha_v) \right) \]
We can now apply the binomial bounds of Lemma 1 in (21)–(23). Using the condition that the 
h_{\alpha,k} displacements are less than unity in absolute value and the notation \( D = \max_{1 \leq a \leq p} i_a \),
we find that (21) is bounded; that is,

\[
\sum_{k_1=1}^{p} \sum_{j_{k_1}=0}^{i_{k_1}} \left| p_{jk_1|i_{k_1}} (\alpha_{k_1} + h_{\alpha,k_1}) - p_{jk_1|i_{k_1}} (\alpha_{k_1}) - \frac{\partial p_{jk_1|i_{k_1}} (\alpha_{k_1})}{\partial \alpha_{k_1}} h_{\alpha,k_1} \right| \prod_{1 \leq v \leq p, j_v = 0}^{i_v} p_{j_v|i_v} (\alpha_v)
\leq 3 \sum_{k_1=1}^{p} h_{\alpha,k_1}^2 i_{k_1}^2 (1 + |h_{\alpha,k_1}|)^{i_{k_1}} \\
\leq 3 \|h\|_{L}^2 D^2 \left( 1 + \max_{1 \leq k \leq p} |h_{\alpha,k}| \right)^D \\
\leq 3 \|h\|_{H}^2 D^2 (1 + \|h\|_{H})^D.
\]

Similarly we find that (22) is equal to

\[
\|h_G\|_{L} \sum_{k_1=1}^{p} \sum_{j_{k_1}=0}^{i_{k_1}} \left| p_{jk_1|i_{k_1}} (\alpha_{k_1} + h_{\alpha,k_1}) - p_{jk_1|i_{k_1}} (\alpha_{k_1}) \right| \prod_{1 \leq v \leq p, j_v = 0}^{i_v} p_{j_v|i_v} (\alpha_v)
\leq 3 \|h_G\|_{L} \sum_{k_1=1}^{p} |h_{\alpha,k_1}| i_{k_1}^2 (1 + |h_{\alpha,k_1}|)^{i_{k_1}} \leq 3 \|h_G\|_{L} \|h\|_{L} D^2 (1 + \|h\|_{H})^D \\
\leq 3 \|h\|_{H}^2 D^2 (1 + \|h\|_{H})^D.
\]

and

\[
\prod_{k=1}^{p} p_{jk|i_k} (\alpha_k + h_{\alpha,k}) - \prod_{k=1}^{p} p_{jk|i_k} (\alpha_k) - \sum_{k=1}^{p} \frac{\partial p_{jk|i_k} (\alpha)}{\partial \alpha_k} h_{\alpha,k} \prod_{l=1}^{p} p_{jl|i_l} (\alpha_k)
\]

\[
= \sum_{k_1=1}^{p} \left( p_{jk_1|i_{k_1}} (\alpha_{k_1} + h_{\alpha,k_1}) - p_{jk_1|i_{k_1}} (\alpha_{k_1}) - \frac{\partial p_{jk_1|i_{k_1}} (\alpha_{k_1})}{\partial \alpha_{k_1}} h_{\alpha,k_1} \right) \prod_{1 \leq v \leq p, v \neq k_1}^{i_v} p_{j_v|i_v} (\alpha_v)
\]

\[
+ \sum_{l=2}^{p} \sum_{1 \leq k_1 < \ldots < k_l \leq p} \left( \prod_{1 \leq u \leq l} (p_{jk_u|i_{k_u}} (\alpha_{k_u} + h_{\alpha,k_u}) - p_{jk_u|i_{k_u}} (\alpha_u)) \prod_{1 \leq v \leq p, v \neq k_1, \ldots, k_l}^{i_v} p_{j_v|i_v} (\alpha_v) \right).
\]
In the same way, (23) is bounded, with

\[ (1 + \| h_G \|_{\ell^1}) \sum_{l=2}^{p} \sum_{1 \leq k_1 < \ldots < k_l \leq p} \sum_{j_1=0}^{i_1} \ldots \sum_{j_p=0}^{i_p} \prod_{v \neq k_1, \ldots, k_l} p_{j,v} (\alpha_v) \times \]

\[ \prod_{1 \leq v \leq l} |p_{j,v} (\alpha_v + h_{\alpha,k_v}) - p_{j,v} (\alpha_v)| \]

\[ = (1 + \| h_G \|_{\ell^1}) \sum_{l=2}^{p} \sum_{1 \leq k_1 < \ldots < k_l \leq p} \prod_{1 \leq u \leq l} |h_{\alpha,k_u} (1 + |h_{\alpha,k_u}|) |^{i_u} \]

\[ \leq 3 (1 + \| h_G \|_{\ell^1}) \sum_{l=2}^{p} \left( \| h \|_\Gamma^2 D^2 (1 + \| h \|_\Gamma) D \right)^l \]

\[ \leq 3p \| h \|^2_\Gamma D^2p (1 + \| h \|_\Gamma) D^{p+1} . \]

Thus,

\[ \left\| f_{i_1, \ldots, i_p}^{(1)} (\theta + h) - f_{i_0, \ldots, i_p}^{(1)} (\theta) - \hat{f}_{i_1, \ldots, i_p}^{(1)} (h) \right\|_{\ell^1} \]

\[ = \sum_{i_0=0}^{\infty} \left| f_{i_0, i_1, \ldots, i_p}^{(1)} (\theta + h) - f_{i_0, i_1, \ldots, i_p}^{(1)} (\theta) - \hat{f}_{i_0, i_1, \ldots, i_p}^{(1)} (h) \right| \]

\[ \leq 6 \| h \|^2_\Gamma D^2 (1 + \| h \|_\Gamma) D + 3p \| h \|^2_\Gamma D^2p (1 + \| h \|_\Gamma) D^{p+1} \] (24)

\[ \leq C_1^2 \| h \|^2_\Gamma \]

for a finite constant \( C_1 \).

To show that \( \hat{f}_{i_1, \ldots, i_p}^{(1)} (h) \) is bounded, we write

\[ \left\| \hat{f}_{i_1, \ldots, i_p}^{(1)} (h) \right\|_{\ell^1} = \sum_{i_0=0}^{\infty} \left| \int_{i_0, i_1, \ldots, i_p}^{(1)} (h) \right| \]

\[ \leq \sum_{i_0=0}^{\infty} \sum_{(j_1, \ldots, j_p) \in J(i_0, \ldots, i_p)} |h_{G,i_0-j_1-\ldots-j_p}| \prod_{k=1}^{p} p_{j_k} (\alpha_k) \]

\[ + \sum_{i_0=0}^{\infty} \sum_{(j_1, \ldots, j_p) \in J(i_0, \ldots, i_p)} g_{i_0-j_1-\ldots-j_p} \sum_{k=1}^{p} \left| \frac{\partial p_{j_k}}{\partial \alpha_k} \right| \left| h_{\alpha,k} \right| \prod_{l=1}^{p} p_{j_l} (\alpha_k) \]

\[ \leq \| h_G \|_{\ell^1} + \sum_{k=1}^{p} i_k |h_{\alpha,k}| \sum_{j_1=0}^{i_1} \ldots \sum_{j_p=0}^{i_p} \prod_{1 \leq l \neq k} p_{j_l} (\alpha_k) \]

\[ = \| h_G \|_{\ell^1} + \sum_{k=1}^{p} i_k^2 |h_{\alpha,k}| \]

\[ \leq (D^2 + 1) \| h \|_\Gamma , \]
as required.

**Proof of Theorem 2**

We will prove that

\[
\left\| F^{(m)}_{i_1, \ldots, i_p} (\theta + h) - F^{(m)}_{i_1, \ldots, i_p} (\theta) - \dot{F}^{(m)}_{i_1, \ldots, i_p} (h) \right\|_{\ell^1} \leq \|h\|_H^2 C_m D^{2p} (1 + \|h\|_H) D^{p},
\]

for some small enough \(\|h\|_H\) and \(D = \max_{1 \leq k \leq p} i_k\). This implies that

\[
\left\| F^{(m)}_{i_1, \ldots, i_p} (\theta + h) - F^{(m)}_{i_1, \ldots, i_p} (\theta) - \dot{F}^{(m)}_{i_1, \ldots, i_p} (h) \right\|_{\ell^1} = o (\|h\|_H)
\]

as required for the derivative. It has already been shown in Theorem 1 that (25) holds for \(m = 1\) and so we proceed by induction and suppose that it holds for \(m - 1\) for some \(m \geq 2\). Using (12) and by adding and subtracting \(\sum_{u=0}^{\infty} \dot{f}^{(m-1)}_{u|i_1, i_2, \ldots, i_{p-1}} (\theta + h) f^{(1)}_{u|i_1, \ldots, i_p} (\theta)\) we get

\[
\left\| F^{(m)}_{i_1, \ldots, i_p} (\theta + h) - F^{(m)}_{i_1, \ldots, i_p} (\theta) - \dot{F}^{(m)}_{i_1, \ldots, i_p} (h) \right\|_{\ell^1} \leq \sum_{u=0}^{\infty} \sum_{u_0=0}^{\infty} \left| \dot{f}^{(m-1)}_{u_0 u, i_1, \ldots, i_{p-1}} (\theta + h) - \dot{f}^{(m-1)}_{u_0 u, i_1, \ldots, i_{p-1}} (\theta) - \dot{f}^{(m-1)}_{u_0 u, i_1, \ldots, i_{p-1}} (h) \right| f^{(1)}_{u|i_1, \ldots, i_p} (\theta) \quad (26a)
\]

\[
+ \sum_{u=0}^{\infty} \left| \dot{f}^{(1)}_{u|i_1, \ldots, i_p} (\theta + h) - \dot{f}^{(1)}_{u|i_1, \ldots, i_p} (\theta) - \dot{f}^{(1)}_{u|i_1, \ldots, i_p} (h) \right| . \quad (26b)
\]

In Theorem 1, (20) is bounded by (21), (22) and (23) which, in turn, leads to (24). This is sufficient to bound (26b). The same sequence of steps bounds (26a) when we take into account that the subscript \(i_0| i_1, \ldots, i_p\) is replaced by \(i_0| u, i_1, \ldots, i_{p-1}\) and so \(D = \max_{1 \leq k \leq p} i_k\) is substituted by \(D \vee u\). Thus, letting \(C_{m-1}\) denote a constant depending on \(m - 1\), (26a) and (26b) are bounded by

\[
C_{m-1} \|h\|_H^2 \sum_{u=0}^{\infty} (D \vee u)^{2p} (1 + \|h\|_H)^{(D \vee u)p+1} f^{(1)}_{u|i_1, \ldots, i_p} (\theta) + C_{m-1} \|h\|_H^2 D^{2p} (1 + \|h\|_H) D^{p+1}\]

\[
= \sum_{u=0}^{D} \|h\|_H^2 D^{2p} (1 + \|h\|_H)^{Dp+1} f^{(1)}_{u|i_1, \ldots, i_p} (\theta) + C_{m-1} \|h\|_H^2 \sum_{u=D+1}^{\infty} u^{2p} (1 + \|h\|_H)^{up+1} f^{(1)}_{u|i_1, \ldots, i_p} (\theta) + C_{m-1} \|h\|_H^2 D^{2p} (1 + \|h\|_H) D^{p+1}\]

\[
\leq 2C_{m-1} \|h\|_H^2 D^{2p} (1 + \|h\|_H)^{Dp+1} + C_{m-1} \|h\|_H^2 \sum_{u=0}^{\infty} u^{2p} (1 + \|h\|_H)^{up+1} f^{(1)}_{u|i_1, \ldots, i_p} (\theta) + C_{m-1} \|h\|_H^2 D^{2p} (1 + \|h\|_H) D^{p+1}\]

\[
\leq \|h\|_H^2 D^{2p} (1 + \|h\|_H)^{Dp+1} C_{m-1} \left( 2 + \sum_{u=0}^{\infty} u^{2p} (1 + \|h\|_H)^{up+1} f^{(1)}_{u|i_1, \ldots, i_p} (\theta) \right) \leq C_m \|h\|_H^2 D^{2p} (1 + \|h\|_H) D^{p+1}, \quad (27)
\]
where

\[ C_m = C_{m-1} \left( 2 + \sum_{u=0}^{\infty} u^{2p} (1 + \|h\|_\mathbb{H})^{u+1} f^{(1)}_{u|i_1,\ldots,i_p} (\theta) \right). \]

The constant \( C_m \) is finite because

\[
\begin{align*}
\sum_{u=0}^{\infty} u^{2p} (1 + \|h\|_\mathbb{H})^{u+1} f^{(1)}_{u|i_1,\ldots,i_p} (\theta) \\
= \sum_{u=0}^{\infty} \sum_{(j_1,\ldots,j_p) \in J(\alpha_1,\ldots,\alpha_p)} g_{u-(j_1+\ldots+j_p)} u^{2p} (1 + \|h\|_\mathbb{H})^{u+1} \prod_{k=1}^{p} p_{j_k | i_k} (\alpha_k) \\
= \sum_{j_1=0}^{i_1} p_{j_1 | i_1} (\alpha_k) \ldots \sum_{j_p=0}^{i_p} p_{j_p | i_p} (\alpha_k) \sum_{u=j_1+\ldots+j_p}^{\infty} g_{u-(j_1+\ldots+j_p)} u^{2p} (1 + \|h\|_\mathbb{H})^{u+1} \\
\leq (1 + \|h_\alpha\|_\mathbb{H})^{2p} \sum_{u=0}^{\infty} g_u (u + pD)^{2p} (1 + 1 + \|h\|_\mathbb{H})^{(u+pD)p+1} \\
\leq (1 + \|h_\alpha\|_\mathbb{H})^{2p} \sum_{u=0}^{\infty} g_u (u + pD)^{2p} (1 + \|h\|_\mathbb{H})^{pu} \\
\leq (1 + \|h_\alpha\|_\mathbb{H})^{2p} \left( \frac{2p}{p} \right) (pD)^{2p} \sum_{u=0}^{\infty} g_u \left( u^2 (1 + \|h\|_\mathbb{H})^u \right)^p,
\end{align*}
\]

using \((u + pD)^{2p} = \sum_{j=0}^{2p} \binom{2p}{j} u^j (pD)^{2p-j} \leq u^{2p} (2p)^p (pD)^{2p} \). This is finite for \( \|h\|_\mathbb{H} \) small enough such that \( 1 + 1 + \|h\|_\mathbb{H} < s \) where \( s > 1 \) is the constant such that \( \sum_{u=0}^{\infty} g_u (u^2 s^u)^p < \infty \). Thus \( C_m \) is constant for small enough \( \|h\|_\mathbb{H} \), which completes the proof of (27).

The derivative \( \dot{F}^{(m)}_{i_1,\ldots,i_p} (h) \) is linear in \( h \) by induction on \( \dot{F}^{(m-1)}_{i_1,\ldots,i_p} (h) \) noting that \( \dot{F}^{(1)}_{i_1,\ldots,i_p} (h) \) is clearly linear. The map \( \dot{F}^{(m)}_{i_1,\ldots,i_p} (h) \) can also be shown to be bounded by induction. In particular we show that

\[
\left\| \dot{F}^{(m)}_{i_1,\ldots,i_p} (h) \right\|_{\ell^1} \leq B_m \|h\|_\mathbb{H} \left( D^2 + 1 \right),
\]

for some finite constant \( B_m \). As shown in the proof of Theorem 1, this holds for \( m = 1 \) with
\[ B_m = 1. \] Now suppose that \( \hat{F}_{i_1, \ldots, i_p}^{(m-1)} (h) \) satisfies this bound. It follows that

\[
\left\| \hat{F}_{i_1, \ldots, i_p}^{(m)} (h) \right\|_{\ell^1} = \sum_{i_0=0}^{\infty} \left| \hat{f}_{i_0|i_1, \ldots, i_p}^{(m)} (h) \right|
\leq \sum_{u=0}^{\infty} \sum_{i_0=0}^{\infty} \left| \hat{f}_{i_0|i_1, \ldots, i_p}^{(m-1)} (h) \right| \left| \hat{f}_{u|i_1, \ldots, i_p}^{(1)} (\theta) \right| + \sum_{u=0}^{\infty} \left| \hat{f}_{i_1, \ldots, i_p}^{(1)} (h) \right|
\leq B_m - 1 \| h \|_{\mathbb{H}} \sum_{u=0}^{\infty} \left( (u \lor i)^2 + 1 \right) \left| f_{u|i_1, \ldots, i_p}^{(1)} (\theta) \right| + (D^2 + 1) \| h \|_{\mathbb{H}}
\leq (B_m - 1 + 1) \| h \|_{\mathbb{H}} (D^2 + 1) \sum_{u=0}^{\infty} \left| f_{u|i_1, \ldots, i_p}^{(1)} (\theta) \right| + B_m - 1 \| h \|_{\mathbb{H}}
\quad + B_m - 1 \| h \|_{\mathbb{H}} \sum_{u=0}^{\infty} u^2 \left| f_{u|i_1, \ldots, i_p}^{(1)} (\theta) \right|
\leq \| h \|_{\mathbb{H}} (D^2 + 1) \left( 2B_m - 1 + B_m - 1 \sum_{u=0}^{\infty} u^2 \left| f_{u|i_1, \ldots, i_p}^{(1)} (\theta) \right| \right)
\leq \| h \|_{\mathbb{H}} (D^2 + 1) B_m,
\]

where \( B_m \) is a constant. This constant is finite because

\[
\sum_{u=0}^{\infty} u^2 \left| f_{u|i_1, \ldots, i_p}^{(1)} (\theta) \right| = \sum_{u=0}^{\infty} u^2 \sum_{(j_1, \ldots, j_p) \in J(i_0, \ldots, i_p)} g_{u-(j_1+\ldots+j_p)} \prod_{k=1}^{p} p_{j_k|i_k} (\alpha_k)
\leq \sum_{u=0}^{\infty} (u+Dp)^2 g_u
\leq \sum_{u=0}^{\infty} u^2 g_u + 2Dp \sum_{u=0}^{\infty} u g_u + (ip)^2 < \infty
\]

under the summability conditions on \( g_u \).

**Proof of Theorem 3**

The proof follows from Theorem 7.3.1 of PRW. Assumption 7.3.1 of PRW follows from the fact that \( \sqrt{T} \left( \hat{F}_{i_1, \ldots, i_p}^{(m)} (\theta) - F_{i_1, \ldots, i_p}^{(m)} (\theta^*) \right) \sim \hat{F}_{i_1, \ldots, i_p}^{(m)} (N_\alpha, \mathcal{M}_C) \) (a continuous Gaussian Process) which, in turn, is a consequence of Theorem 2 and the fact that \( \ell^1 \) is a separable metric space; this corresponds to \( J_n (P) = Q_T \) converging to \( J (P) = Q \) on a separable subset of \( S \) (in the notation of PRW). By the Markov Chain properties of the model, the process \( X_t \) is
absolutely regularly mixing ($\beta$-mixing) (see Doukhan, 1994, and DvdAW, Proposition 2.1) and this implies that $X_t$ is $\alpha$-mixing. Finally $\tau_b = h^{1/2}$ and $\tau_T = T^{1/2}$ and so all the regularity conditions of PRW are satisfied.

References


Table 1: Finite sampling performance of the $AEEF$ and $MLE-P$ in different parts of the predictive support, under various distributions for $\varepsilon_t$ ($\alpha_1 = 0.6$). The first row in each panel reports the average (over $i$) mean squared error (AV. MSE) of the $AEEF$; the figures in parentheses beneath represent the ratio of the AV. MSE of the $AEEF$ to the AV. MSE of the $MLE-P$.

<table>
<thead>
<tr>
<th></th>
<th>$\varepsilon_t \sim Poisson$</th>
<th></th>
<th>$\varepsilon_t \sim Binomial$</th>
<th></th>
<th>$\varepsilon_t \sim NegBinomial$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda = 2$</td>
<td>$n = 4; \pi = 0.4$</td>
<td>$v = 5; \pi = 0.3$</td>
<td>$n = 4; \pi = 0.4$</td>
<td>$v = 5; \pi = 0.3$</td>
<td></td>
</tr>
<tr>
<td>$T=100$</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td>$T=500$</td>
<td>$8.5 \times 10^{-5}$</td>
<td>$4.0 \times 10^{-5}$</td>
<td>$4.0 \times 10^{-5}$</td>
<td>$4.0 \times 10^{-5}$</td>
<td>$4.0 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>$T=1000$</td>
<td>$(5.614)$</td>
<td>$(1.194)$</td>
<td>$(0.313)$</td>
<td>$(0.153)$</td>
<td>$(0.457)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AV. MSE of $AEEF$ over all $i$. (Ratio to AV. MSE of $MLE-P$ in parentheses) &amp; AV. MSE of $AEEF$ in upper 10% tail. (Ratio to AV. MSE of $MLE-P$ in parentheses) &amp; AV. MSE of $AEEF$ in lower 25% tail. (Ratio to AV. MSE of $MLE-P$ in parentheses)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$3.1 \times 10^{-5}$</td>
<td>$3.4 \times 10^{-5}$</td>
<td>$3.9 \times 10^{-5}$</td>
<td>$1.7 \times 10^{-5}$</td>
<td>$1.2 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(5.231)$</td>
<td>$(10.274)$</td>
<td>$(1.396)$</td>
<td>$(0.374)$</td>
<td>$(0.188)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AV. MSE of $AEEF$ in lower 25% tail. (Ratio to AV. MSE of $MLE-P$ in parentheses) &amp; AV. MSE of $AEEF$ in upper 10% tail. (Ratio to AV. MSE of $MLE-P$ in parentheses) &amp; AV. MSE of $AEEF$ over all $i$. (Ratio to AV. MSE of $MLE-P$ in parentheses)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0006</td>
<td>0.0001</td>
<td>0.0005</td>
<td>0.0001</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$5.1 \times 10^{-5}$</td>
<td>$5.1 \times 10^{-5}$</td>
<td>$5.1 \times 10^{-5}$</td>
<td>$5.1 \times 10^{-5}$</td>
<td>$5.2 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(4.051)$</td>
<td>$(3.742)$</td>
<td>$(0.775)$</td>
<td>$(0.202)$</td>
<td>$(2.625)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Ratio to AV. MSE of $MLE-P$ in parentheses) &amp; AV. MSE of $AEEF$ in lower 25% tail. (Ratio to AV. MSE of $MLE-P$ in parentheses) &amp; AV. MSE of $AEEF$ in upper 10% tail. (Ratio to AV. MSE of $MLE-P$ in parentheses) &amp; AV. MSE of $AEEF$ over all $i$. (Ratio to AV. MSE of $MLE-P$ in parentheses)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Finite sampling performance of the AEEF and MLE-P in the tails of the predictive support, under various distributions for $\varepsilon_t$ ($\alpha_1 = 0.6$). The first row in each panel reports the average (over $i$) bias (AV. BIAS) of the AEEF; the figures in parentheses beneath represent the ratio of the average bias of the AEEF to the AV. BIAS of the MLE-P.

<table>
<thead>
<tr>
<th>$\varepsilon_t \sim Poisson$</th>
<th>$\varepsilon_t \sim Binomial$</th>
<th>$\varepsilon_t \sim NegBinomial$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 2$</td>
<td>$n = 4; \pi = 0.4$</td>
<td>$v = 5; \pi = 0.3$</td>
</tr>
<tr>
<td>$T=100$</td>
<td>$T=100$</td>
<td>$T=100$</td>
</tr>
<tr>
<td>$T=500$</td>
<td>$T=500$</td>
<td>$T=500$</td>
</tr>
<tr>
<td>$T=1000$</td>
<td>$T=1000$</td>
<td>$T=1000$</td>
</tr>
<tr>
<td>AV. BIAS of AEEF in upper 10% tail. (Ratio to AV. BIAS of MLE-P in parentheses)</td>
<td>AV. BIAS of AEEF in lower 25% tail. (Ratio to AV. BIAS of MLE-P in parentheses)</td>
<td></td>
</tr>
<tr>
<td>-0.0001</td>
<td>-0.0004</td>
<td>0.0002</td>
</tr>
<tr>
<td>(-1.494)</td>
<td>(-0.943)</td>
<td>(-0.187)</td>
</tr>
<tr>
<td>-1.2×10^{-5}</td>
<td>-0.0001</td>
<td>-2.5×10^{-5}</td>
</tr>
<tr>
<td>(-1.580)</td>
<td>(-0.197)</td>
<td>(-0.025)</td>
</tr>
<tr>
<td>1.4×10^{-5}</td>
<td>-2.2×10^{-5}</td>
<td>-1.7×10^{-5}</td>
</tr>
<tr>
<td>(-1.222)</td>
<td>(-0.077)</td>
<td>(-0.018)</td>
</tr>
<tr>
<td>-0.0003</td>
<td>-1.9×10^{-4}</td>
<td>-0.0008</td>
</tr>
<tr>
<td>(-0.278)</td>
<td>(0.022)</td>
<td>(0.150)</td>
</tr>
<tr>
<td>0.0002</td>
<td>1.8×10^{-4}</td>
<td>7.6×10^{-5}</td>
</tr>
<tr>
<td>(0.778)</td>
<td>(0.022)</td>
<td>(-0.011)</td>
</tr>
<tr>
<td>0.0001</td>
<td>-1.6×10^{-5}</td>
<td>-1.6×10^{-5}</td>
</tr>
<tr>
<td>(0.799)</td>
<td>(-0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>