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 (2001) (economics); and Gourieroux and Jasiak (2004) (insurance).

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 likelihood 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 differentiability 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 distribution ($AEEF$ 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 distribution 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 iceberg 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

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

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 $\mathbb{Z} = \{0, 1, 2, ...\}$. Conditional on X_{t-k}, $k \in \{1, 2, ..., p\}$, the thinning operators $\alpha_k \circ X_{t-k}$, $k \in \{1, 2, ..., p\}$ are binomial, and defined as

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

where each collection $\{\mathrm{B}_{i,k,t}, i = 1, 2, ..., X_{t-k}\}$ consists of independently distributed Bernoulli random variables with thinning parameter (probability of unity) α_k, and the collections are mutually independent. It is assumed that $\alpha_k \in [0, 1)$, for all $k \in \{1, 2, ..., p\}$, and that $\sum_{k=1}^{p} \alpha_k < 1$. The innovations are taken to be independent of all thinning operations. The initial values $(X_0, X_{-1}, ..., X_{-p})$ 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 $\theta = (\alpha_1, ..., \alpha_p, G)$.

At time t, each thinning operator performs one of p binomial experiments, with parameters (X_{t-k}, α_k), $k \in \{1, 2, ..., p\}$, to determine the number from that time vintage that survives in the system. When α_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 α_k is close to unity. Consider the vintage X_t. At $t + 1$, X_t is thinned by α_1 and at time $t + 2$, X_t is again thinned but using α_2. Thus, the 'offspring' of X_t are distributed across future times $t+1$, $t+2$, ... according to the number of lags and the sizes of the thinning parameters. This allows for the effect of X_t to be propagated across multiple time periods. More formally, when $p > 1$, Dion *et al.* (1995) show that the $INAR(p)$ process may be generally viewed as a special multitype branching process with immigration.

When $p = 1$, X_t behaves like a queue, with arrivals at time t represented by ε_t and survivors remaining in the queue, from $t-1$ to t, by $\alpha_1 \circ X_{t-1}$. Alternatively the model may be thought of as a birth and death, or stock process, with additions (births) being generated by ε_t and losses (deaths) by $(X_{t-1} - \alpha_1 \circ X_{t-1})$. When G is Poisson and $p = 1$, the 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

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

where

$$p_{j_k|i_k}(\alpha_k) = \binom{i_k}{j_k} \alpha_k^{j_k} (1-\alpha_k)^{i_k-j_k}, \; 0 \le j_k \le i_k \tag{3}$$

and

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

Empty sums are taken to be zero, so that $j_1 \le (i_0 \wedge i_1)$. Expression (2) gives the probability $\Pr(X_t = i_0 | X_{t-1} = i_1, \ldots, X_{t-p} = i_p; \theta)$ 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

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

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

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

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 $0 \le \alpha_k < 1; k = 1,2,...,p$ and $\sum_{r=g_-}^{g_+} g_r = 1$ where $g_- = 0 \vee \min_{t=p+1,\ldots,T} (x_t - \sum_{k=1}^{p} x_{t-k})$ and $g_+ = \max_{t=p+1,\ldots,T} x_t$. The $NPMLE$ is denoted $\hat{\theta} = \left(\hat{\alpha}, \hat{G}\right) = (\hat{\alpha}_k; k = 1,2,...,p, \{\hat{g}_r\})$ and consists of a vector, $\hat{\alpha}$, which is an estimator of $\alpha = (\alpha_1, ..., \alpha_p)'$ 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 ℓ^1. The parameter space Θ 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 jth elements of h_α and h_G. Thus, $\sqrt{T}\left(\left(\hat{\alpha}, \hat{G}\right) - (\alpha, G)\right)$ 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 α^* 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.

$$\sqrt{T}\left[\hat{\theta} - \theta^*\right] = \sqrt{T}\left[\left(\hat{\alpha}, \hat{G}\right) - (\alpha^*, G^*)\right] \rightsquigarrow (N_\alpha, \mathfrak{N}_G), \tag{6}$$

where N_α is a p-dimensional zero mean normal random variable, $\mathfrak{N}_G$ is a centered Gaussian process that lives in ℓ^1 and $\rightsquigarrow$ 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 $\left(\tilde{\alpha}, \tilde{G}\right)$ be a regular estimator, then

$$\sqrt{T}\left[\left(\tilde{\alpha}, \tilde{G}\right) - (\alpha^*, G^*)\right] \rightsquigarrow (N_\alpha + W, \mathfrak{N}_G + \mathfrak{W}),$$

where W and $\mathfrak{W}$ are 'noise' processes independent of the Gaussian process $(N_\alpha, \mathfrak{N}_G)$. 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

$$f_{i_0|i_1,\ldots,i_p}^{(1)}(\theta) = f_{i_0|i_1,\ldots,i_p}(\theta) \tag{7}$$

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

$$F_{i_1,\ldots,i_p}^{(1)}(\theta) = \left\{f_{i_0|i_1,\ldots,i_p}^{(1)}(\theta), i_0 \in \mathbb{Z}\right\} \tag{8}$$

and $F^{(1)}_{i_1,\ldots,i_p}(\theta)$ is a mapping from the Banach space $\mathbb{H}$ to the Banach space ℓ^1, as defined in Section 2.1. In probabilistic forecasting the objective is to estimate the one-step-ahead distribution $F^{(1)}_{i_1,\ldots,i_p}(\theta)$. In applications, θ 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 $F^{(1)}_{i_1,\ldots,i_p}(\hat{\theta})$ may inherit the properties of $\hat{\theta}$ and also be asymptotically efficient, if the map $F^{(1)}_{i_1,\ldots,i_p}(\theta) : \mathbb{H} \mapsto \ell^1$ is smooth enough. Smoothness requires the existence of a derivative map $\dot{F}^{(1)}_{i_1,\ldots,i_p} : H \mapsto \ell^1$ between the same two spaces. To motivate the structure of such a map consider the total differential of (2) with respect to α_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 $F^{(1)}_{i_1,\ldots,i_p}(\hat{\theta}_T)$ as in (8), the map $F^{(1)}_{i_1,\ldots,i_p} : \mathbb{H} \mapsto \ell^1$ is Frechet differentiable with derivative $\dot{F}^{(1)}_{i_1,\ldots,i_p}(h)$, where $\dot{F}^{(1)}_{i_1,\ldots,i_p} : \mathbb{H} \mapsto \ell^1$ is a bounded linear operator with typical element*

$$\begin{aligned}\dot{f}^{(1)}_{i_0|i_1,\ldots,i_p}(h) \; = \; & \sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)} h_{G,i_0-(j_1+\ldots+j_p)} \prod_{k=1}^{p} p_{j_k|i_k}(\alpha_k) + \\ & \sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)} g_{i_0-(j_1+\ldots+j_p)} \sum_{k=1}^{p} \frac{\partial p_{j_k|i_k}(\alpha)}{\partial \alpha_k} h_{\alpha,k} \prod_{\substack{l=1 \\ l\neq k}}^{p} p_{j_k|i_k}(\alpha_k) \, . \end{aligned} \tag{9}$$

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

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 $F^{(1)}_{i_1,\ldots,i_p}(\hat{\theta})$ is also asymptotically efficient for the one-step-ahead distribution. Thus, $F^{(1)}_{i_1,\ldots,i_p}(\hat{\theta})$ 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), $\sqrt{T}\left[\hat{\theta} - \theta^*\right] \rightsquigarrow (N_\alpha, \mathfrak{N}_G)$ and since the spaces $\mathbb{H}$ and ℓ^1 are linear spaces, it is a consequence of Theorem 20.8 of van der Vaart (1998) that

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

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

$$P\left[X_{T+1} = i | X_T = x_T; \theta\right] = f^{(1)}_{i|x_T}(\theta) = \sum_{j=0}^{i \wedge x_T} p_{j|x_T}(\alpha_1) g_{i-j}, \tag{11}$$

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

$$f^{(m)}_{i_0|i_1,\ldots,i_p}(\theta) = \sum_{u=0}^{\infty} f^{(m-1)}_{i_0|u,i_1,\ldots,i_{p-1}}(\theta)\, f^{(1)}_{u|i_1,\ldots,i_p}(\theta) \tag{12}$$

and

$$F^{(m)}_{i_1,\ldots,i_p}(\theta) = \left\{f^{(m)}_{i_0|i_1,\ldots,i_p}(\theta) : i_0 \in \mathbb{Z}\right\}. \tag{13}$$

It also follows, for any m, that $F^{(m)}_{i_1,\ldots,i_p}(\theta) : \mathbb{H} \mapsto \ell^1$ 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),*

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

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

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

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 λ ($Pois\left(\lambda\right)$),

$$\sum_{u=0}^{\infty}\left(u^{2}s^{u}\right)^{p}g_{u}=\sum_{u=0}^{\infty}u^{2p}\frac{e^{-\lambda}\left(s^{p}\lambda\right)^{u}}{u!}=\frac{e^{s^{p}\lambda}}{e^{\lambda}}\sum_{u=0}^{\infty}u^{2p}\frac{e^{-s^{p}\lambda}\left(s^{p}\lambda\right)^{u}}{u!}<\infty$$

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

$$g_{u}=\frac{\Gamma\left(v+u\right)}{\Gamma\left(v\right)\Gamma\left(u+1\right)}\pi^{u}\left(1-\pi\right)^{v},\ v>0,\ 0<\pi<1, \tag{14}$$

we have $\sum_{u=0}^{\infty}\left(u^{2}s^{u}\right)^{p}g_{u}=\frac{(1-\pi)^{v}}{\Gamma(v)}\sum_{u=0}^{\infty}u^{2p}\frac{\Gamma(v+u)}{\Gamma(u+1)}\left(s^{p}\pi\right)^{u}$, 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:

$$\begin{aligned}&P\left[X_{t}=x_{t}|X_{t-1}=x_{t-1},...X_{t-p}=x_{t-p};\theta\right]\\=&\sum_{j_{1}=0}^{x_{t}\wedge x_{t-1}}p_{j_{1}|x_{t-1}}(\alpha_{1})\sum_{j_{2}=0}^{x_{t}-j_{1}\wedge x_{t-2}}p_{j_{2}|x_{t-2}}(\alpha_{2})\sum_{j_{3}=0}^{x_{t}-(j_{1}+j_{2})\wedge x_{t-3}}p_{j_{3}|x_{t-3}}(\alpha_{3})\\&\cdots\sum_{j_{p}=0}^{x_{t}-(j_{1}+j_{2}+...+j_{p-1})\wedge x_{t-p}}p_{j_{p}|x_{t-p}}(\alpha_{p})g_{x_{t}-(j_{1}+j_{2}+...+j_{p})}\end{aligned} \tag{15}$$

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

$$\left\{P\left[X_{T+1}=i_{0}|X_{T}=x_{T},X_{T-1}=x_{T-1},...X_{T-p}=x_{T-p};\theta\right],\ i_{0}\in\mathbb{Z}\right\} \tag{16}$$

is produced by simply substituting the $NPMLE$ of $\theta=(\alpha_{k};\ k=1,2,...,p,\{g_{r}\})$ 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 $\{0,1,...,K\}$ at time t, and the states of the system given by $p-tuples$ of possible values. Hence, at time T, the chain could be in any one of the $(K+1)^{p}$ states:

$$S=\{\underbrace{(0,0,...,0)}_{p\text{ terms}},(0,1,...,0),...,(0,K,...,0),...,(K,...,0),...,(K,K,...,K)\}$$

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

$$\begin{aligned} & P\left[X_{T+1} = i_0, X_T = i_1, ..., X_{T-(p-1)} = i_{p-1} | X_T = j_0, X_{T-1} = j_1, ..., X_{T-p} = j_{p-1}; \theta\right] \\ = \; & P\left[X_{T+1} = i_0 | X_T = j_0, X_{T-1} = j_1, ..., X_{T-p} = j_{p-1}; \theta\right] \text{ for } i_1 = j_0, ..., i_{p-1} = j_{p-2} \\ = \; & 0 \text{ for any } i_1 \neq j_0, ..., i_{p-1} \neq j_{p-2}, \end{aligned}$$

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

$$\begin{gathered} P\left[X_T = j_0, X_{T-1} = j_1, ..., X_{T-p} = j_{p-1}; \theta\right] \text{ and} \\ P\left[X_{T+1} = i_0, X_T = i_1, ..., X_{T-(p-1)} = i_{p-1}; \theta\right]. \end{gathered}$$

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

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

extracted from π_{T+m}^{T}. The asymptotically efficient estimate of the m-step ahead forecast distribution is produced by simply replacing θ by $\widehat{\theta}$ 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 *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 ε_t to be distributed, respectively, as Poisson, $Pois(\lambda = 2)$, binomial, $Bin(n = 4; \pi = 0.4)$, and negative binomial, $NBin(v = 5; \pi = 0.3)$ (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 α_1 is selected to approximate the empirical $NPMLE$ of α_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 $f^{(1)}_{i|x_T}(\theta)$, $i \in \mathbb{Z}$ (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 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 $\widehat{E}\left[\left(\widehat{f}_i - f_i\right)^2\right]$ be the simple average of the squared errors $\left(\widehat{f}_i - f_i\right)^2$ over the 10000 replications, where $\widehat{f}_i$ 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 $\left\{\widehat{f}_i\right\}$, calculated by averaging $\widehat{E}\left[\left(\widehat{f}_i - f_i\right)^2\right]$ 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 $\widehat{E}\left[\left(\widehat{f}_i - f_i\right)^2\right]$ 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 $\widehat{E}\left(\widehat{f}_i - f_i\right)$ 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 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, $\alpha_1 \circ X_{t-1}$, and the number of new iceberg orders placed in the book (or 'arrivals'), ε_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 α_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 $Y_1 = (X_1, \ldots, X_b), Y_2 = (X_2, ..., X_{b+1}), ..., Y_{T-b+1} = (X_{T-b+1}, ..., X_T)$.

2. Use the $NPMLE$ of θ, $\hat{\theta}_{b,t}$, computed from Y_t and the *observed* values, $x_T, x_{T-1}, \ldots, x_{T-(p-1)}$, to compute the $m-$step ahead forecast distribution $F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}_{b,t}\right)$; $m \geq 1$.

3. Calculate the metric $d_{b,t} = \sqrt{T}\left\|F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}_{b,t}\right) - F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}\right)\right\|_1$, where $F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}\right)$ 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 $||.||_1$ distance from $F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}\right)$, 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 $b \in \{b_{small}, \ldots, b_{big}\}$ 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 $\left\{d^{0.95}_{b-k}, \ldots, d^{0.95}_{b+k}\right\}$ (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

$$Q_{T,b}(x) = \frac{1}{T-b+1}\sum_{t=1}^{T-b+1} I\left[\sqrt{T}\left\|F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}_{b,t}\right) - F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}\right)\right\| < x\right], \tag{17}$$

and $Q_T(x)$ to be the law of the sampling distribution of $\sqrt{T}\left\|F^{(m)}_{i_1,\ldots,i_p}(\hat{\theta}) - F^{(m)}_{i_1,\ldots,i_p}(\theta^*)\right\|$; θ^* 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 $b/T \to 0$,*

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

where ρ_L is the bounded Lipschitz metric.

As a consequence of Theorem 3, for large enough T, statements based on the empirical distribution of $\sqrt{T}\left\|F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}_{b,t}\right) - F^{(m)}_{i_1,\ldots,i_p}\left(\hat{\theta}\right)\right\|$ are equivalent to statements based on the (unknown) sampling distribution of $\sqrt{T}\left\|F^{(m)}_{i_1,\ldots,i_p}(\hat{\theta}) - F^{(m)}_{i_1,\ldots,i_p}(\theta^*)\right\|$. 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, $F^{(m)}_{i_1,\ldots,i_p}(\theta^*)$ (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.

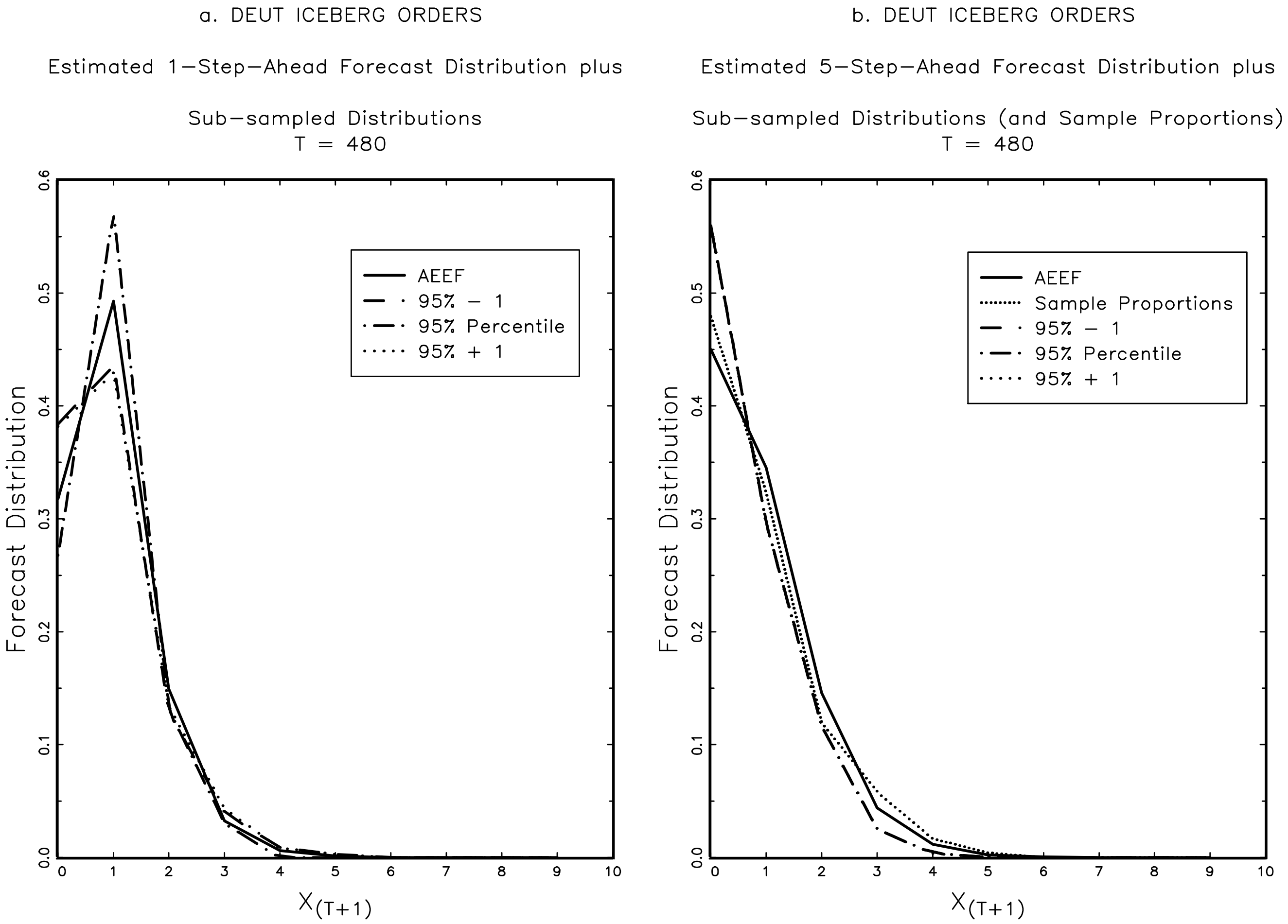
The estimated forecast distribution assigns only 32% probability to the event of *no* DEUT iceberg order being included in the five best bids during the last 10 minutes of the last trading day of the first quarter in 2004. This indicates that *some* degree of hidden liquidity was very likely to be available ($\Pr ob(X_{T+1} \geq 1 = 0.68)$, 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



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 *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 *ex-post* evaluation based on observed outcomes (see Dawid, 1984; Tay and Wallis, 2000; Gneiting *et al,* 2007; Gneiting, 2008; and Geweke and Amisano, 2010 for examples and general discussion, and Czado *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 *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 *ex-post* methods - with relevant alternatives from outside the class.

Finally, the inclusion of covariates in the general $INAR$ 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 $p_{j|i}(\alpha)$ is a binomial probability $\binom{i}{j}\alpha^j(1-\alpha)^{i-j}$ and h is a constant then*

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

and

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

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

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

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

Proof of Theorem 1

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

$$\begin{aligned}
&f^{(1)}_{i_0|i_1,\ldots,i_p}(\theta+h)-f^{(1)}_{i_0|i_1,\ldots,i_p}(\theta)-\dot{f}^{(1)}_{i_0|i_1,\ldots,i_p}(h)\\
=&\sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)} g_{i_0-(j_1+\ldots+j_p)}\left\{\prod_{k=1}^{p}p_{j_k|i_k}(\alpha_k+h_{\alpha,k})-\prod_{k=1}^{p}p_{j_k|i_k}(\alpha_k)\right.\\
&\left.\qquad-\sum_{k=1}^{p}\frac{\partial p_{j_k|i_k}(\alpha)}{\partial\alpha_k}h_{\alpha,k}\prod_{\substack{l=1\\l\neq k}}^{p}p_{j_l|i_l}(\alpha_k)\right\}\\
&+\sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)} h_{G,i_0-(j_1+\ldots+j_p)}\left\{\prod_{k=1}^{p}p_{j_k|i_k}(\alpha_k+h_{\alpha,k})-\prod_{k=1}^{p}p_{j_k|i_k}(\alpha_k)\right\}.
\end{aligned}$$

Straightforward rearrangements show that

$$\sum_{i_0=0}^{\infty}\sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)}=\sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p}\sum_{i_0=j_1+\ldots+j_p}^{\infty}$$

and, hence, that

$$
\begin{aligned}
& \left\| F^{(1)}_{i_1,\ldots,i_p}\left(\theta+h\right) - F^{(1)}_{i_1,\ldots,i_p}\left(\theta\right) - \dot{F}^{(1)}_{i_1,\ldots,i_p}\left(h\right) \right\|_{\ell^1} \\
= \; & \sum_{i_0=0}^{\infty} \left| f^{(1)}_{i_0|i_1,\ldots,i_p}\left(\theta+h\right) - f^{(1)}_{i_0|i_1,\ldots,i_p}\left(\theta\right) - \dot{f}^{(1)}_{i_0|i_1,\ldots,i_p}\left(h\right) \right| && (20) \\
\leq \; & \sum_{j_1=0}^{i_1} \cdots \sum_{j_p=0}^{i_p} \left| \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k + h_{\alpha,k}\right) - \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k\right) - \sum_{k=1}^{p} \frac{\partial p_{j_k|i_k}\left(\alpha\right)}{\partial \alpha_k} h_{\alpha,k} \prod_{\substack{l=1\\ l\neq k}}^{p} p_{j_l|i_l}\left(\alpha_k\right) \right| \\
& + \left\| h_G \right\| \sum_{j_1=0}^{i_1} \cdots \sum_{j_p=0}^{i_p} \left| \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k + h_{\alpha,k}\right) - \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k\right) \right| \\
= \; & \sum_{j_1=0}^{i_1} \cdots \sum_{j_p=0}^{i_p} \sum_{k_1=1}^{p} \prod_{\substack{1\leq v\leq p\\ v\neq k_1}} p_{j_v|i_v}\left(\alpha_v\right) \times \\
& \qquad \left| p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1} + h_{\alpha,k_1}\right) - p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right) - \frac{\partial p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right)}{\partial \alpha_{k_1}} h_{\alpha,k_1} \right| && (21) \\
& + \left\| h_G \right\| \sum_{j_1=0}^{i_1} \cdots \sum_{j_p=0}^{i_p} \sum_{k_1=1}^{p} \prod_{\substack{1\leq v\leq p\\ v\neq k_1}} p_{j_v|i_v}\left(\alpha_v\right) \times \\
& \qquad \left| p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1} + h_{\alpha,k_1}\right) - p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right) \right| && (22) \\
& + \left(1 + \left\| h_G \right\|\right) \sum_{j_1=0}^{i_1} \cdots \sum_{j_p=0}^{i_p} \sum_{l=2}^{p} \sum_{1\leq k_1<\ldots<k_l\leq p} \prod_{\substack{1\leq v\leq p\\ v\neq k_1,\ldots,k_l}} p_{j_v|i_v}\left(\alpha_v\right) \times \\
& \qquad \prod_{1\leq u\leq l} \left| p_{j_{k_u}|i_{k_u}}\left(\alpha_{k_u} + h_{\alpha,k_u}\right) - p_{j_{k_u}|i_{k_u}}\left(\alpha_u\right) \right| . && (23)
\end{aligned}
$$

The last step above uses the rearrangements

$$
\begin{aligned}
& \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k + h_{\alpha,k}\right) - \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k\right) \\
= \; & \sum_{l=1}^{p} \sum_{1\leq k_1<\ldots<k_l\leq p} \left(\prod_{1\leq u\leq l} \left(p_{j_{k_u}|i_{k_u}}\left(\alpha_{k_u} + h_{\alpha,k_u}\right) - p_{j_{k_u}|i_{k_u}}\left(\alpha_u\right) \right) \prod_{\substack{1\leq v\leq p\\ v\neq k_1,\ldots,k_l}} p_{j_v|i_v}\left(\alpha_v\right) \right)
\end{aligned}
$$

and

$$
\begin{aligned}
& \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k + h_{\alpha,k}\right) - \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k\right) - \sum_{k=1}^{p} \frac{\partial p_{j_k|i_k}\left(\alpha\right)}{\partial \alpha_k} h_{\alpha,k} \prod_{\substack{l=1 \\ l\neq k}}^{p} p_{j_l|i_l}\left(\alpha_k\right) \\
= & \sum_{k_1=1}^{p} \left(p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1} + h_{\alpha,k_1}\right) - p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right) - \frac{\partial p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right)}{\partial \alpha_{k_1}} h_{\alpha,k_1} \right) \prod_{\substack{1\leq v\leq p \\ v\neq k_1}} p_{j_v|i_v}\left(\alpha_v\right) \\
& + \sum_{l=2}^{p} \sum_{1\leq k_1<\ldots<k_l\leq p} \left(\prod_{1\leq u\leq l} \left(p_{j_{k_u}|i_{k_u}}\left(\alpha_{k_u} + h_{\alpha,k_u}\right) - p_{j_{k_u}|i_{k_u}}\left(\alpha_u\right) \right) \prod_{\substack{1\leq v\leq p \\ v\neq k_1,\ldots,k_l}} p_{j_v|i_v}\left(\alpha_v\right) \right).
\end{aligned}
$$

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 $D = \max_{1\leq u\leq p} i_u$, we find that (21) is bounded; that is,

$$
\begin{aligned}
& \sum_{k_1=1}^{p} \sum_{j_{k_1}=0}^{i_{k_1}} \left| p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1} + h_{\alpha,k_1}\right) - p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right) - \frac{\partial p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right)}{\partial \alpha_{k_1}} h_{\alpha,k_1} \right| \prod_{\substack{1\leq v\leq p \\ v\neq k_1}} \sum_{j_v=0}^{i_v} p_{j_v|i_v}\left(\alpha_v\right) \\
\leq\ & 3 \sum_{k_1=1}^{p} h_{\alpha,k_1}^2 i_{k_1}^2 \left(1 + \left|h_{\alpha,k_1}\right|\right)^{i_{k_1}} \\
\leq\ & 3 \left\|h_\alpha\right\|_{\mathbb{R}^p}^2 D^2 \left(1 + \max_{1\leq k\leq p} \left|h_{\alpha,k}\right|\right)^D \\
\leq\ & 3 \left\|h\right\|_{\mathbb{H}}^2 D^2 \left(1 + \left\|h\right\|_{\mathbb{H}}\right)^D.
\end{aligned}
$$

Similarly we find that (22) is equal to

$$
\begin{aligned}
& \left\|h_G\right\|_{\ell^1} \sum_{k_1=1}^{p} \sum_{j_{k_1}=0}^{i_{k_1}} \left| p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1} + h_{\alpha,k_1}\right) - p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right) \right| \prod_{\substack{1\leq v\leq p \\ v\neq k_1}} \sum_{j_v=0}^{i_v} p_{j_v|i_v}\left(\alpha_v\right) \\
\leq\ & 3 \left\|h_G\right\|_{\ell^1} \sum_{k_1=1}^{p} \left|h_{\alpha,k_1}\right| i_{k_1}^2 \left(1 + \left|h_{\alpha,k_1}\right|\right)^{i_{k_1}} \leq 3 \left\|h_G\right\|_{\ell^1} \left\|h_\alpha\right\|_{\mathbb{R}^p} D^2 \left(1 + \left\|h\right\|_{\mathbb{H}}\right)^D \\
\leq\ & 3 \left\|h\right\|_{\mathbb{H}}^2 D^2 \left(1 + \left\|h\right\|_{\mathbb{H}}\right)^D.
\end{aligned}
$$

In the same way, (23) is bounded, with

$$
\begin{aligned}
& (1+\|h_G\|_{\ell^1})\sum_{l=2}^{p}\sum_{1\le k_1<\ldots<k_l\le p}\sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p}\prod_{\substack{1\le v\le p\\ v\neq k_1,\ldots,k_l}} p_{j_v|i_v}(\alpha_v)\times \\
& \qquad \prod_{1\le u\le l}\left|p_{j_{k_u}|i_{k_u}}(\alpha_{k_u}+h_{\alpha,k_u})-p_{j_{k_u}|i_{k_u}}(\alpha_u)\right| \\
= \; & (1+\|h_G\|_{\ell^1})\sum_{l=2}^{p}\sum_{1\le k_1<\ldots<k_l\le p}\prod_{1\le u\le l}\sum_{j_{k_u}=0}^{i_{k_u}}\left|p_{j_{k_u}|i_{k_u}}(\alpha_{k_u}+h_{\alpha,k_u})-p_{j_{k_u}|i_{k_u}}(\alpha_u)\right| \\
\le \; & 3\,(1+\|h_G\|_{\ell^1})\sum_{l=2}^{p}\sum_{1\le k_1<\ldots<k_l\le p}\prod_{1\le u\le l}|h_{\alpha,k_u}|\, i_{k_u}^2\,(1+|h_{\alpha,k_u}|)^{i_{k_u}} \\
\le \; & 3\,(1+\|h_G\|_{\ell^1})\sum_{l=2}^{p}\left(\|h\|_{\mathbb{H}}\,D^2\,(1+\|h\|_{\mathbb{H}})^D\right)^l \\
\le \; & 3p\,\|h\|_{\mathbb{H}}^2\,D^{2p}\,(1+\|h\|_{\mathbb{H}})^{Dp+1}\,.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& \left\|F^{(1)}_{i_1,\ldots,i_p}(\theta+h)-F^{(1)}_{i_1,\ldots,i_p}(\theta)-\dot{F}^{(1)}_{i_1,\ldots,i_p}(h)\right\|_{\ell^1} \\
= \; & \sum_{i_0=0}^{\infty}\left|f^{(1)}_{i_0|i_1,\ldots,i_p}(\theta+h)-f^{(1)}_{i_0|i_1,\ldots,i_p}(\theta)-\dot{f}^{(1)}_{i_0|i_1,\ldots,i_p}(h)\right| \\
\le \; & 6\,\|h\|_{\mathbb{H}}^2\,D^2\,(1+\|h\|_{\mathbb{H}})^D+3p\,\|h\|_{\mathbb{H}}^2\,D^{2p}\,(1+\|h\|_{\mathbb{H}})^{Dp+1} \qquad (24) \\
\le \; & C_1^2\,\|h\|_{\mathbb{H}}^2
\end{aligned}
$$

for a finite constant C_1.

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

$$
\begin{aligned}
\left\|\dot{F}^{(1)}_{i_1,\ldots,i_p}(h)\right\|_{\ell^1} = \; & \sum_{i_0=0}^{\infty}\left|\dot{f}^{(1)}_{i_0|i_1,\ldots,i_p}(h)\right| \\
\le \; & \sum_{i_0=0}^{\infty}\sum_{(j_1,\ldots,j_p)\;\in J(i_0,\ldots,i_p)}\left|h_{G,i_0-(j_1+\ldots+j_p)}\right|\prod_{k=1}^{p}p_{j_k|i_k}(\alpha_k) \\
& +\sum_{i_0=0}^{\infty}\sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)}g_{i_0-(j_1+\ldots+j_p)}\sum_{k=1}^{p}\left|\frac{\partial p_{j_k|i_k}(\alpha)}{\partial\alpha_k}\right||h_{\alpha,k}|\prod_{\substack{l=1\\ l\neq k}}^{p}p_{j_l|i_l}(\alpha_k) \\
\le \; & \|h_G\|_{\ell^1}+\sum_{k=1}^{p}i_k\,|h_{\alpha,k}|\sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p}\prod_{\substack{l=1\\ l\neq k}}^{p}p_{j_l|i_l}(\alpha_k) \\
= \; & \|h_G\|_{\ell^1}+\sum_{k=1}^{p}i_k^2\,|h_{\alpha,k}| \\
\le \; & \left(D^2+1\right)\|h\|_{\mathbb{H}}\,,
\end{aligned}
$$

as required.

Proof of Theorem 2

We will prove that

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

for some small enough $\|h\|_{\mathbb{H}}$ and $D = \max_{1\leq u\leq p} i_u$. This implies that

$$\left\| F^{(m)}_{i_1,\ldots,i_p}(\theta+h) - F^{(m)}_{i_1,\ldots,i_p}(\theta) - \dot{F}^{(m)}_{i_1,\ldots,i_p}(h) \right\|_{\ell^1} = o\left(\|h\|_{\mathbb{H}}\right) \tag{25}$$

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

$$\begin{aligned}
&\left\| F^{(m)}_{i_1\ldots i_p}(\theta+h) - F^{(m)}_{i_1\ldots i_p}(\theta) - \dot{F}^{(m)}_{i_1\ldots i_p}(h) \right\|_{\ell^1} \\
\leq\ & \sum_{u=0}^{\infty}\sum_{i_0=0}^{\infty} \left| f^{(m-1)}_{i_0|u,i_1,\ldots,i_{p-1}}(\theta+h) - f^{(m-1)}_{i_0|u,i_1,\ldots,i_{p-1}}(\theta) - \dot{f}^{(m-1)}_{i_0|u,i_1,\ldots,i_{p-1}}(h) \right| f^{(1)}_{u|i_1,\ldots,i_p}(\theta) \quad \text{(26a)} \\
&+\sum_{u=0}^{\infty} \left| f^{(1)}_{u|i_1,\ldots,i_p}(\theta+h) - f^{(1)}_{u|i_1,\ldots,i_p}(\theta) - \dot{f}^{(1)}_{u|i_1,\ldots,i_p}(h) \right|. \quad \text{(26b)}
\end{aligned}$$

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

$$\begin{aligned}
& C_{m-1}\|h\|^2_{\mathbb{H}} \sum_{u=0}^{\infty} (D\vee u)^{2p} (1+\|h\|_{\mathbb{H}})^{(D\vee u)p+1} f^{(1)}_{u|i_1,\ldots,i_p}(\theta) \\
& +C_{m-1}\|h\|^2_{\mathbb{H}} D^{2p} (1+\|h\|_{\mathbb{H}})^{Dp+1} \\
=\ & C_{m-1}\|h\|^2_{\mathbb{H}} \sum_{u=0}^{D} D^{2p} (1+\|h\|_{\mathbb{H}})^{Dp+1} f^{(1)}_{u|i_1,\ldots,i_p}(\theta) \\
& +C_{m-1}\|h\|^2_{\mathbb{H}} \sum_{u=D+1}^{\infty} u^{2p} (1+\|h\|_{\mathbb{H}})^{up+1} f^{(1)}_{u|i_1,\ldots,i_p}(\theta) \\
& +C_{m-1}\|h\|^2_{\mathbb{H}} D^{2p} (1+\|h\|_{\mathbb{H}})^{Dp+1} \\
\leq\ & 2C_{m-1}\|h\|^2_{\mathbb{H}} D^{2p} (1+\|h\|_{\mathbb{H}})^{Dp+1} + C_{m-1}\|h\|^2_{\mathbb{H}} \sum_{u=0}^{\infty} u^{2p} (1+\|h\|_{\mathbb{H}})^{up+1} f^{(1)}_{u|i_1,\ldots,i_p}(\theta) \\
\leq\ & \|h\|^2_{\mathbb{H}} D^{2p} (1+\|h\|_{\mathbb{H}})^{Dp+1} C_{m-1} \left(2 + \sum_{u=0}^{\infty} u^{2p} (1+\|h\|_{\mathbb{H}})^{up+1} f^{(1)}_{u|i_1,\ldots,i_p}(\theta) \right) \\
\leq\ & C_m \|h\|^2_{\mathbb{H}} D^{2p} (1+\|h\|_{\mathbb{H}})^{Dp+1}, \quad \text{(27)}
\end{aligned}$$

where

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

The constant C_m is finite because

$$\begin{aligned}
&\sum_{u=0}^{\infty} u^{2p}\left(1+\|h\|_{\mathbb{H}}\right)^{up+1} f^{(1)}_{u|i_1,\ldots,i_p}(\theta)\\
=\ & \sum_{u=0}^{\infty}\sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)} g_{u-(j_1+\ldots+j_p)}u^{2p}\left(1+\|h\|_{\mathbb{H}}\right)^{up+1}\prod_{k=1}^{p} p_{j_k|i_k}(\alpha_k)\\
=\ & \sum_{j_1=0}^{i_1} p_{j_1|i_1}(\alpha_k)\ldots\sum_{j_p=0}^{i_p} p_{j_p|i_p}(\alpha_k)\sum_{u=j_1+\ldots+j_p}^{\infty} g_{u-(j_1+\ldots+j_p)}u^{2p}\left(1+\|h\|_{\mathbb{H}}\right)^{up+1}\\
\leq\ & \left(1+\|h_\alpha\|_{\mathbb{R}^p}\right)^{p^2D}\sum_{u=0}^{\infty} g_u\left(u+pD\right)^{2p}\left(1+\|h\|_{\mathbb{H}}\right)^{(u+pD)p+1}\\
\leq\ & \left(1+\|h_\alpha\|_{\mathbb{R}^p}\right)^{p^2D}\sum_{u=0}^{\infty} g_u\left(u+pD\right)^{2p}\left(1+\|h\|_{\mathbb{H}}\right)^{pu}\\
\leq\ & \left(1+\|h_\alpha\|_{\mathbb{R}^p}\right)^{p^2D}\binom{2p}{p}\left(pD\right)^{2p}\sum_{u=0}^{\infty} g_u\left(u^2\left(1+\|h\|_{\mathbb{H}}\right)^u\right)^p,
\end{aligned}$$

using $(u+pD)^{2p} = \sum_{j=0}^{2p}\binom{2p}{j}u^j(pD)^{2p-j} \leq u^{2p}\binom{2p}{p}(pD)^{2p}$. This is finite for $\|h\|_{\mathbb{H}}$ small enough such that $1+\|h\|_{\mathbb{H}} < s$ where $s>1$ is the constant such that $\sum_{u=0}^{\infty} g_u\left(u^2 s^u\right)^p < \infty$. 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 $\dot{F}^{(m-1)}_{i_1,\ldots,i_p}(h)$ satisfies this bound. It follows that

$$\begin{aligned}
\left\|\dot{F}^{(m)}_{i_1,\ldots,i_p}(h)\right\|_{\ell^1} &= \sum_{i_0=0}^{\infty}\left|\dot{f}^{(m)}_{i_0|i_1,\ldots,i_p}(h)\right| \\
&\leq \sum_{u=0}^{\infty}\sum_{i_0=0}^{\infty}\left|\dot{f}^{(m-1)}_{i_0|u,i_1,\ldots,i_{p-1}}(h)\right| f^{(1)}_{u|i_1,\ldots,i_p}(\theta) + \sum_{u=0}^{\infty}\left|\dot{f}^{(1)}_{u|i_1,\ldots,i_p}(h)\right| \\
&= \sum_{u=0}^{\infty}\left\|\dot{F}^{(m-1)}_{u,i_1,\ldots,i_{p-1}}(h)\right\|_{\ell^1} f^{(1)}_{u|i_1,\ldots,i_p}(\theta) + \left\|\dot{F}^{(1)}_{i_1,\ldots,i_p}(h)\right\|_{\ell^1} \\
&\leq B_{m-1}\|h\|_{\mathbb{H}}\sum_{u=0}^{\infty}\left((u\vee i)^2+1\right) f^{(1)}_{u|i_1,\ldots,i_p}(\theta) + \left(D^2+1\right)\|h\|_{\mathbb{H}} \\
&\leq \left(B_{m-1}+1\right)\|h\|_{\mathbb{H}}\left(D^2+1\right)\sum_{u=0}^{i} f^{(1)}_{u|i_1,\ldots,i_p}(\theta) + B_{m-1}\|h\|_{\mathbb{H}} \\
&\qquad + B_{m-1}\|h\|_{\mathbb{H}}\sum_{u=0}^{\infty}u^2 f^{(1)}_{u|i_1,\ldots,i_p}(\theta) \\
&\leq \|h\|_{\mathbb{H}}\left(D^2+1\right)\left(2B_{m-1}+1+B_{m-1}\sum_{u=0}^{\infty}u^2 f^{(1)}_{u|i_1,\ldots,i_p}(\theta)\right) \\
&= \|h\|_{\mathbb{H}}\left(D^2+1\right)B_m,
\end{aligned}$$

where B_m is a constant. This constant is finite because

$$\begin{aligned}
\sum_{u=0}^{\infty}u^2 f^{(1)}_{u|i_1,\ldots,i_p}(\theta) &= \sum_{u=0}^{\infty}u^2\sum_{(j_1,\ldots,j_p)\in J(i_0,\ldots,i_p)} g_{u-(j_1+\ldots+j_p)}\prod_{k=1}^{p}p_{j_k|i_k}(\alpha_k) \\
&= \sum_{j_1=0}^{i_1}p_{j_1|i_1}(\alpha_1)\ldots\sum_{j_k=0}^{i_k}p_{j_k|i_k}(\alpha_k)\sum_{u=j_1+\ldots+j_p}^{\infty}u^2 g_{u-(j_1+\ldots+j_p)} \\
&\leq \sum_{u=0}^{\infty}(u+Dp)^2 g_u \\
&= \sum_{u=0}^{\infty}u^2 g_u + 2Dp\sum_{u=0}^{\infty}u g_u + (ip)^2 < \infty
\end{aligned}$$

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 $\sqrt{T}\left(F^{(m)}_{i_1,\ldots,i_p}(\hat{\theta}) - F^{(m)}_{i_1,\ldots,i_p}(\theta^*)\right) \rightsquigarrow \dot{F}^{(m)}_{i_1,\ldots,i_p}(N_\alpha, \mathfrak{N}_G)$ (a continuous Gaussian Process) which, in turn, is a consequence of Theorem 2 and the fact that ℓ^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 (β-mixing) (see Doukhan, 1994, and DvdAW, Proposition 2.1) and this implies that X_t is α-mixing. Finally $\tau_b = b^{1/2}$ and $\tau_T = T^{1/2}$ and so all the regularity conditions of PRW are satisfied.

References

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

[2] Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests, *Journal of Business and Economic Statistics*, 25, 177-190.

[3] Bockenholt, U. (1999). Mixed INAR(1) Poisson regression models: analyzing heterogeneity and serial dependencies in longitudinal count data, *Journal of Econometrics*, 89, 317-338.

[4] Brännäs, K. and Hellstrom, J. (2001). Generalized integer valued autoregression, *Econometric Reviews,* 20, 425-443.

[5] Bu, R. and McCabe, B.P.M. (2008). Model selection, estimation and forecasting in INAR(p) models: A likelihood based Markov chain approach, *International Journal of Forecasting*, 24, 151-162.

[6] Bu, R, Hadri, K. and McCabe, B.P.M. (2008). Maximum likelihood estimation of higher-order integer valued autoregressive processes, *Journal of Time Series Analysis*, 29, 973-994.

[7] Cardinal, M., Roy, R. and Lambert, J. (1999). On the application of integer-valued time series models for the analysis of disease incidence, *Statistics in Medicine,* 18, 2025-2039.

[8] Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary time series, *Annals of Statistics* 14, 1171-1179.

[9] Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data, In press, *Biometrics.*

[10] Dawid, A.P. (1984). Present position and potential developments: some personal views. Statistical theory. The prequential approach, *Journal of the Royal Statistical Society (A)*, 147, 278–292.

[11] Dion, J-P., Gauthier, G. and Latour, A. (1995). Branching processes with immigration and integer-valued time series, *Serdica Mathematics Journal,* 21, 123-136.

[12] Doukhan, P. (1994). *Mixing: Properties and Examples* (1st ed.). Springer-Verlag: Lecture Notes in Statistics.

[13] Drost, F.C., Van den Akker, R. and Werker, B.J.M. (2008). Local asymptotic normality and efficient estimation for INAR(p) models, *Journal of Time Series Analysis,* 29, 783–801.

[14] Drost, F.C., Van den Akker, R. and Werker, B.J.M. (2009). Efficient estimation of autoregression parameters and innovation distributions for semiparametric integer-valued AR(p) models, *Journal of the Royal Statistical Society (B),* 71, 467-485.

[15] Du, J.D. and Li, Y. (2001). The integer-valued autoregressive (INAR(p)) model, *Journal of Time Series Analysis,* 12, 129–142.

[16] Franke, J. and Seligmann, T. (1993). Conditional maximum-likelihood estimates for INAR(1) processes and their applications to modelling epileptic seizure counts. In *Developments in Time Series*, 310-330. Subba Rao, T. (ed.). Chapman & Hall, London.

[17] Freeland, R. and McCabe, B.P.M. (2004). Analysis of low count time series data by Poisson autoregression, *Journal of Time Series Analysis*, 25, 701-722.

[18] Frey, S. and Sandas, P. (2008). Iceberg Orders and the Compensation for Liquidity Provision, *Draft Paper,* University of Tubingen.

[19] Geweke, J. and Amisano, G. (2010). Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns, Special Issue on Bayesian Forecasting in Economics, *International Journal of Forecasting*, 26, 216-230.

[20] Gourieroux, C. and Jasiak, J. (2004). Heterogeneous INAR(1) model with application to car insurance, *Insurance Mathematics and Economics,* 34, 177-192.

[21] Gneiting, T. (2008). Editorial: Probabilistic forecasting , *Journal of the Royal Statistical Society (A),* 171, 319-321.

[22] Gneiting, T., Balabdaoui, F. and Raftery, A. (2007). Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society (B)*, 69, 243–268.

[23] Ispany, M., Pap, G. and van Zuijlen, M. (2003). Asymptotic inference for nearly unstable INAR(1) models, *Journal of Applied Probability,* 40, 750-765.

[24] Jung, R., Ronning, G. and Tremayne, A. (2005). Estimation in conditional first order autoregression with discrete support, *Statistical Papers,* 46, 195-224.

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

[26] Kunsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations, *Annals of Statistics*, 17, 1217-1241.

[27] Latour, A. (1998). Existence and stochastic structure of a nonnegative integer-valued autoregressive process, *Journal of Time Series Analysis*, 19, 439-455.

[28] Liu, R.Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*, 225-248. LePage, R. and Billard, L. (eds.). John Wiley, New York.

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

[30] McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts, *Advances in Applied Probability,* 20, 822-835.

[31] McKenzie, E. (2003). Discrete variate time series. In *Handbook of Statistics,* 21, 573–606. Shanbhag, D.N. and Rao, C.R. (eds.) Elsevier, Amsterdam.

[32] Neal, P. and Subba Rao, T. (2007). MCMC for integer-valued ARMA processes, *Journal of Time Series Analysis*, 28, 92-110.

[33] Pavlopoulos, H. and Karlis, D. (2008). INAR(1) modelling of overdispersed count series with an environmental application. *Environmetrics*, 19, 369-393.

[34] Pickands, J. and Stine, R. (1997). Estimation for an M/G/1 queue with incomplete information, *Biometrika*, 84, 295-308.

[35] Politis, D.N., Romano, J.P. and Wolf, M. (1999). *Subsampling,* Springer, New York.

[36] Resnick, S.I. (1992). *Adventures in Stochastic Processes*, Birkhauser, Boston.

[37] Rudholm, N. (2001). Entry and the number of firms in the Swedish pharmaceuticals market, *Review of Industrial Organization,* 19, 351-364.

[38] Silva, M. and Oliveira, V. (2005). Difference equations for the higher order moments and cumulants of the INAR(p) model, *Journal of Time Series Analysis,* 26, 17-36.

[39] Tay, A.S. and Wallis, K. (2000). Density forecasting: A survey, *Journal of Forecasting*, 19, 235-254.

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

[41] van der Vaart, A.W. (1995). Efficiency of infinite dimensional M-estimators, *Statistica Neerlandica,* 49, 9–30.

[42] van der Vaart, A.W. (1998). *Asymptotic Statistics*, Cambridge University Press, Cambridge.

[43] Zhu, R. and Joe, H. (2006). Modelling count data time series with Markov processes based on binomial thinning, *Journal of Time Series Analysis,* 725–738.

Table 1: Finite sampling performance of the $AEEF$ and MLE-P in different parts of the predictive support, under various distributions for ε_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.

$\varepsilon_t \sim Poisson$			$\varepsilon_t \sim Binomial$			$\varepsilon_t \sim NegBinomial$		
$\lambda = 2$			$n = 4$; $\pi = 0.4$			$v = 5$; $\pi = 0.3$		
T=100	T=500	T=1000	T=100	T=500	T=1000	T=100	T=500	T=1000
$AV.\ MSE$ of $AEEF$ over all i. (Ratio to $AV.\ MSE$ of MLE-P in parentheses)								
0.0005	8.5×10^{-5}	5.1×10^{-5}	0.0005	0.0001	4.0×10^{-5}	0.0005	0.0001	4.0×10^{-5}
(4.947)	(4.425)	(5.614)	(1.194)	(0.313)	(0.153)	(3.329)	(0.883)	(0.457)
$AV.\ MSE$ of $AEEF$ in upper 10% tail. (Ratio to $AV.\ MSE$ of MLE-P in parentheses)								
0.0002	3.1×10^{-5}	3.4×10^{-5}	0.0002	3.9×10^{-5}	1.7×10^{-5}	0.0002	2.9×10^{-5}	1.2×10^{-5}
(5.231)	(4.343)	(10.274)	(1.396)	(0.374)	(0.188)	(4.080)	(1.385)	(0.737)
$AV.\ MSE$ of $AEEF$ in lower 25% tail. (Ratio to $AV.\ MSE$ of MLE-P in parentheses)								
0.0006	0.0001	5.1×10^{-5}	0.0005	0.0001	5.1×10^{-5}	0.0007	0.0001	5.2×10^{-5}
(4.051)	(3.647)	(3.742)	(0.775)	(0.202)	(0.101)	(2.625)	(0.530)	(0.2514)

Table 2: Finite sampling performance of the *AEEF* and *MLE-P* in the tails of the predictive support, under various distributions for ε_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*.

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