MONASH University

Department of Econometrics and Business Statistics

Optimal Probabilistic Forecasts for Counts

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

August 2009

Optimal Probabilistic Forecasts for Counts[*]

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

August 16, 2009

Abstract

Optimal 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 efficiency in that setting. The ideas are demonstrated within the context of the integer autoregressive class 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 optimality are supplemented by simulation results which demonstrate the overall superiority of the nonparametric method relative to a misspecified parametric maximum likelihood estimator, in large but finite samples. The method is applied to counts of wage claim benefits, stock market iceberg orders and civilian deaths in Iraq, with bootstrap methods used to quantify sampling variation in the estimated forecast distributions.

KEYWORDS: Nonparametric Inference; Asymptotic Efficiency; Count Time Series; INAR Model Class; Bootstrap Distributions; Iceberg Stock Market Orders.

JEL CODES: C14, C22, C53.

1

1 Introduction

Probabilistic forecasting involves the assignment of a probability distribution to the future values of a random variable. As such, probabilistic forecasts fit naturally with the human propensity to quantify uncertainty in terms of probabilities and to frame forecasts of an uncertain future in probabilistic terms. Probabilistic forecasts are also coherent - i.e. consistent with the sample space of the variable in question - and replete with all important distributional (in particular tail) information. In contrast, point forecasts, based on single summary measures of central location (e.g. a (conditional) mean, median or mode), convey no such distributional information and, potentially, also lack coherence as, for example, when a conditional mean forecast of an integer-valued variable assumes non-integer values.

Despite earlier attempts to draw attention to the value of probabilistic forecasts (e.g. Dawid, 1984), such forecasts have only started to gain real purchase in the literature since the mid to late 1990's (see, for example, Abramson and Clemen, 1995; Diebold *et al.*, 1998; Tay and Wallis, 2000; Berkowitz, 2001, Krysztofowicz, 2001; Gneiting and Raftery, 2005; Gneiting *et al.*, 2005; Elsner and Jagger, 2006; Egorova *et al.*, 2006; Gneiting *et al.*, 2006; Corradi and Swanston, 2006; Alkema *et al.,* 2007; Bao *et al.* 2007; Amisano and Giacomini, 2007; Gneiting *et al.*, 2007; Gneiting and Raftery, 2007; Gneiting, 2008; Czado *et al.*, 2009; Geweke and Amisano, 2009). Central to much of this literature is the *ex-post* evaluation of distributional forecasts using observed outcomes. Calibration with realized values is assessed via the probability integral transform method (e.g. Dawid; Diebold *et al.*; Geweke and Amisano), predictive accuracy tests (e.g. Corradi and Swanston; Amisano and Giacomini), or via the application of calibration criteria in combination with measures of predictive 'sharpness', including the use of various scoring rules (e.g. Gneiting *et al.*, 2007; Gneiting and Raftery, 2007; Czado *et al.*). Most notably, Czado *et al.* investigate alternative evaluation methods in the context of probabilistic forecasts for discrete count data, the data type of interest in this paper.

The methods used in the existing literature to evaluate and compare alternative forecast distributions often treat these distributions as primitives. That is, the methods are applicable no matter what formal model and inferential technique (if any) have been used to assign probabilities to the future values of the random variable (see Dawid; Corradi and Swanston; Gneiting *et al.*, 2007; Gneiting, 2008, for discussion). In particular, comparisons of the predictive accuracy of alternative distributions do not preclude the possibility that all alternatives are misspecified versions of the true dynamic process that has generated the data.

In contrast, the focus of this paper is on producing probabilistic forecasts that are *ex-ante* optimal within a given broad class of structural models deemed appropriate for a particular data type. The optimality is achieved by estimating the forecast distribution

nonparametrically over the model class and proving asymptotic efficiency in that setting. On the assumption that the broad model class is a suitable structure for the empirical data under analysis, the optimality of the nonparametric estimator of the forecast distribution provides strong motivation for its adoption. Whilst certainly not to be viewed as a competitor to the fundamental principle of assessing distributional forecasts using realized outcomes, this approach does serve to re-focus attention on the suitability of the model class used to forecast particular data types and on the production of optimal forecasts within that class. 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.[1]

The optimal probabilistic forecasts are derived within the context of a particular class of time series models for count data: the integer autoregressive ($INAR$) class. The $INAR$ class may variously be interpreted as a queue, a stock, a birth and death process or a special type of branching process (with immigration). Each of these interpretations is suggestive of certain types of count data. Any data series that may be thought of as the number of clients (e.g. people, firms, machines, computer software, stock market orders) waiting for a service in a specified time period is a queue. The number of units in an inventory at a given time is a stock variable. So too, in a given time period, is the number of firms located in a region, the number of aircraft in a specified partition of airspace and the number of people with a certain disease or characteristic. Over long time spans, the numbers of people, plants, animals, species etc. in a given environment may be thought of as a birth and death process. Branching processes are concerned with phenomena where characteristics (offspring) are transmitted between generations, for example, the propagation of surnames, the transmission of genes, the growth of bacteria and so on. The $INAR$ model is a branching process that allows for immigration but may place some restrictions on the number of offspring transmitted between generations.

The $INAR$ class is thus a behavioural/structural model of a potentially very large collection of count data time series, and a suitable class over which to optimize. Of course, the class may also be used as a versatile modelling tool for *any* kind of count time series data, even those without an inherent queue or branching interpretation. In such settings, the optimal forecast distribution over the $INAR$ class may still be produced, and compared - via ex-post methods - with forecast distributions produced from competing count data

---

[1]The current paper paper does not contribute to the ex-post evaluation literature in any way, nor demonstrate the empirical application of the evaluation techniques that have already been extensively reviewed and applied in the literature cited above. As noted in the text, such techniques would simply complement the approach developed in this paper, in particular empirical settings, if deemed necessary by the investigator.

models. The broad scope of the empirical literature in which the $INAR$ class is applied is indicative of its relevance, with recent examples including: Franke and Seligmann (1993), Pickands and Stine (1997) and Cardinal *et al.* (1999) (medicine); Bockenholt (1999) (marketing); Thyregod *et al.* (1999) (environmental studies); Brännäs and Hellstrom (2001) and Rudholm (2001) (economics); Brännäs and Shahiduzzaman (2004) (finance); Gourieroux and Jasiak (2004) (insurance) and Pavlopoulos and Karlis (2007) (environmental studies).

The focus on producing an optimal estimator of a forecast distribution for the count variable entails the need for a measure of sampling variability in any empirical application. Standard scalar methods would enable us to construct a confidence interval for the probability that the future variable assumes a particular value (or finite set of values). However, it is advantageous to be able 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 bootstrap methods to allow the effect of sampling fluctuations to be visualized whilst retaining the positivity and summation to unity properties of probabilities.

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 $NPMLE$ of the forecast distribution is demonstrated, with the proof of the differentiability of the mapping that defines the forecast distribution given in the Appendix. The finite sample performance of the $NPMLE$, within the $INAR$ class, is documented via simulation in Section 3. In particular, the overall superiority of the $NPMLE$ relative to a misspecified parametric maximum likelihood estimator, in large but finite samples, is illustrated. In Section 4 the $NPMLE$ is applied to three data series. The first two series, which enumerate, respectively, Canadian wage loss benefit claims and German stock market iceberg orders, both constitute a record of the number of elements over time in a queue and, hence, are suitably modelled as an $INAR$ process. The third series is a daily count of civilian deaths in Iraq, during 2006. As the deaths are due, at least in part, to local sectarian violence, we may consider today's deaths to be a combination of retaliation to (or 'offspring' of) other prior deaths and new deaths (immigration). The civilian deaths may thereby be viewed as a branching process and thus modelled via the $INAR$ class. Section 5 concludes.

2   Probabilistic Forecasting of Count Data in the INAR Class

Coherent forecasting for count data requires that positive probabilities be assigned only to the non-negative integers to form the forecast distribution. Within a model class, such as the $INAR$ class described below, optimal forecasting is synonymous with optimal estima-

tion of the forecast distribution, where this distribution is to be estimated nonparametrically over the model class. Thus, optimal probabilistic forecasting requires the production of the (asymptotically) efficient nonparametric estimator of the $INAR$ model, and the demonstration of the required smoothness of the transformation that defines the forecast distribution.

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), Brännäs (1994), Dion *et al.* (1995), Latour (1998), Ispany *et al.* (2003, 2005), Freeland and McCabe (2004a,b, 2005), Jung *et al.* (2005), McCabe and Martin (2005), Silva and Oliveira (2005), Jung and Tremayne (2006a,b), Silva and Silva (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. The $INAR$ class models dependence between the observations directly and thus exemplifies a class of *observation-driven* models.[2] 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 optimality of the $NPMLE$ 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 assumed to be independent of all thinning operations. The initial values $(X_0, X_{-1}, ..., X_{-p})$ are assumed to be independent drawings from the stationary distribution of the model. The infinite dimensional parameter of the model is $\theta = (\alpha_1, ..., \alpha_p, G)$.

---

[2]This is in contrast to the class of *parameter-driven* models, which introduce dynamics in the counts indirectly by specifying time-varying parameters as functions of a random latent process.

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

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

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

$$P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p}; \theta) = f_{x_t|x_{t-1},...,x_{t-p}}(\theta).$$

When $p = 1$, these expressions simplify considerably and

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

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 $g_- = 0 \vee \min_{t=p+1,...,T} (x_t - \sum_{k=1}^{p} x_{t-k})$ and $g_+ = \max_{t=p+1,...,T} x_t$. The $NPMLE$ is denoted $\hat{\theta} = \left(\hat{\alpha}, \hat{G}\right) = (\hat{\alpha}_k; k = 1, 2, ..., p, \{\hat{g}_r\})$ 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 $G = \{g_r\}$.[3] 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

$$\|h_\alpha\|_{\mathbb{R}^p} = \left(\sum_{j=1}^{p} h_{\alpha,j}^2\right)^{1/2}$$

$$\|h_G\|_{\ell^1} = \sum_{j=0}^{\infty} |h_{G,j}|$$

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.

$$\sqrt{T}\left[\hat{\theta} - \theta^*\right] = \sqrt{T}\left[\left(\hat{\alpha}, \hat{G}\right) - (\alpha^*, G^*)\right] \rightsquigarrow (N_\alpha, \mathfrak{N}_G), \tag{5}$$

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

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

---

[3]For notational simplicity we supress the dependence of estimators, like $\hat{\theta}$, on the sample size T.

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 asymptotic efficiency is understood.

2.2   Optimal 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,

$$P(X_{T+1} = i_0 | X_T = x_T, \dots, X_{T-p+1} = x_{T-p+1}; \theta),$$

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

$$f^{(1)}_{i_0|i_1,\dots,i_p}(\theta) = f_{i_0|i_1,\dots,i_p}(\theta) \tag{6}$$

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

$$F^{(1)}_{i_1,\dots,i_p}(\theta) = \left\{ f^{(1)}_{i_0|i_1,\dots,i_p}(\theta), i_0 \in \mathbb{Z} \right\} \tag{7}$$

and $F^{(1)}_{i_1,\dots,i_p}(\theta)$ 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 $F^{(1)}_{i_1,\dots,i_p}(\theta)$. In applications, θ in (6) 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 $F^{(1)}_{i_1,\dots,i_p}(\hat{\theta})$ may inherit the properties of $\hat{\theta}$ and also be asymptotically efficient, if the map $F^{(1)}_{i_1,\dots,i_p}(\theta) : \mathbb{H} \mapsto \ell^1$ is smooth enough. That the map is sufficiently smooth is a consequence of the following Theorem, proved in the Appendix.

Theorem 1 *Defining $F^{(1)}_{i_1,\dots,i_p}(\hat{\theta}_T)$ as in (7), the map $F^{(1)}_{i_1,\dots,i_p} : \mathbb{H} \mapsto \ell^1$ is Frechet differentiable with derivative $\dot{F}^{(1)}_{i_1,\dots,i_p}(h)$, where $\dot{F}^{(1)}_{i_1,\dots,i_p} : \mathbb{H} \mapsto \ell^1$ is a bounded linear operator with typical element*

$$
\dot{f}^{(1)}_{i_0|i_1,\dots,i_p}(h) = \sum_{(j_1,\dots,j_p)\in J(i_0,\dots,i_p)} h_{G,i_0-(j_1+\dots+j_p)} \prod_{k=1}^{p} p_{j_k|i_k}(\alpha_k) +
$$

$$
\sum_{(j_1,\dots,j_p)\in J(i_0,\dots,i_p)} g_{i_0-(j_1+\dots+j_p)} \sum_{k=1}^{p} \frac{\partial p_{j_k|i_k}(\alpha)}{\partial \alpha_k} h_{\alpha,k} \prod_{\substack{l=1 \\ l\neq k}}^{p} p_{j_k|i_k}(\alpha_k). \tag{8}
$$

In particular for $\|h\|_{\mathbb{H}} < 1$ we have

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

8

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 optimal probability forecast 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 (5), $\sqrt{T}\left[\hat{\theta} - \theta^*\right] \rightsquigarrow (N_\alpha, \mathfrak{N}_G)$ 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

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

It follows from Theorem 1 above that $\dot{F}_{i_1,\ldots,i_p}^{(1)}(N_\alpha, \mathfrak{N}_G)$ is also a Gaussian process by the linearity of $\dot{F}_{i_1,\ldots,i_p}^{(1)}$. 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 $F_{i_1,\ldots,i_p}^{(1)}(\hat{\theta})$ 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

$$P\left[X_{T+1} = i | X_T = x_T; \theta\right] = f_{i|x_T}^{(1)}(\theta) = \sum_{j=0}^{i \wedge x_T} p_{j|x_T}(\alpha) g_{i-j}, \tag{9}$$

where the binomial probabilities, $p_{j|x_T}(\alpha)$, 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). That is,

$$f_{i_0|i_1,\ldots,i_p}^{(m)}(\theta) = \sum_{u=0}^{\infty} f_{i_0|u,i_1,\ldots,i_{p-1}}^{(m-1)}(\theta) f_{u|i_1,\ldots,i_p}^{(1)}(\theta) \tag{10}$$

and

$$F_{i_1,\ldots,i_p}^{(m)}(\theta) = \left\{f_{i_0|i_1,\ldots,i_p}^{(m)}(\theta) : i_0 \in \mathbb{Z}\right\}. \tag{11}$$

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.

9

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 (6) and (8),*

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

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

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

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 ($Pois(\lambda)$),

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

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

$$g_u = \frac{\Gamma(v + u)}{\Gamma(v) \Gamma(u + 1)} \pi^u (1 - \pi)^v, \ v > 0, \ 0 < \pi < 1, \tag{12}$$

we have

$$\sum_{u=0}^{\infty} (u^2 s^u)^p g_u = \frac{(1 - \pi)^v}{\Gamma(v)} \sum_{u=0}^{\infty} u^{2p} \frac{\Gamma(v + u)}{\Gamma(u + 1)} (s^p \pi)^u,$$

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.

3 Finite Sample Performance in the INAR Class

In the previous section we have proven the asymptotic optimality 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 estimator, in comparison with both a correctly specified and incorrectly specified parametric estimator.[4] We focus on the one-step-ahead forecast distribution (i.e. $m = 1$), and for notational convenience we denote

---

[4] All numerical results reported in this and the following empirical section have been produced using the GAUSS software. Programs are available from the corresponding author on request.

$f_{i|x_T}^{(1)}(\theta)$, $i \in \mathbb{Z}$ by f_i, $i \in \mathbb{Z}$, using the notation $\{f_i\}$ to denote the full sequence of forecast probabilities over $\mathbb{Z}$. We consider the $INAR(p)$ data generating process in (1) with $p = 1$ and ε_t distributed, respectively, as Poisson, $Pois(\lambda = 2)$, binomial, $Bin(n = 4; \pi = 0.4)$, and negative binomial, $NBin(v = 5; \pi = 0.3)$[5]. These distributions are representative, respectively, of equi-, under- and over-dispersed distributions for the arrivals. The true value of α_1 is set to 0.2 and 0.6 respectively. These specifications produce, in turn, low count data with sample autocorrelations that are typical of those observed in practice, including for the data sets analysed in Section 4[6].

The performance of the $NPMLE$ is compared with that of the parametric estimator of $\{f_i\}$ based on the application of MLE to the $INAR(1)$ model with Poisson arrivals; i.e. the canonical PAR model. This parametric MLE (denoted MLE-P) is obviously misspecified when the arrivals are either binomial or negative binomial. In what follows we denote the $NPMLE$ of f_i by $\widehat{f}_i$ and the Poisson based MLE-P of f_i by $\widehat{f}_i^P$, where

$$\widehat{f}_i^P = \sum_{j=0}^{i \wedge x_T} p_{j|x_T}\left(\widehat{\alpha}_1^P\right) \exp\left\{-\widehat{\lambda}^P\right\} \frac{\left(\widehat{\lambda}^P\right)^{(i-j)}}{(i-j)!},$$

for $i = 0, 1, \ldots, K$, and $\widehat{\alpha}_1^P$ and $\widehat{\lambda}^P$ are produced via MLE-P. All results are based on 5000 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 5000 replications. The '$AV.\ MSE$' figures recorded in the first row of results in Tables 1 to 3 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 the rows immediately below the $AV.$ MSE measures for the $NPMLE$ give the ratio of the relevant measure for the $NPMLE$ to the corresponding measure for the MLE-P. Clearly, values for the $AV.\ MSE$ ratio that are less than one indicate that the $NPMLE$ is superior in terms of this measure of accuracy.

The figures presented in the second panel in each of Tables 1 to 3 refer only to the upper 10% tail in the predictive support. The '$AV.\ BIAS$' figures consist of $\widehat{E}\left(\widehat{f}_i - f_i\right)$ averaged over the upper 10% of the support $i = 0, 1, \ldots, K$. The $AV.\ MSE$ figures are computed analogously. Whilst the $AV.\ MSE$ figures (both in raw and ratio form) measure the accuracy with which the $NPMLE$ estimates the probability of occurrence of rare large

---

[5] The negative binomial random variable used in the simulation experiments has a mass function as defined in (12).

[6] See the sample statistics reported for several empirical count series in Feigen *et al.* (2008).

counts, the corresponding $AV.\ BIAS$ figures capture the phenomenon of under- or over-estimation of the tail probability.[7] Positive values for the $AV.\ BIAS$ ratios indicate that *both* the $NPMLE$ and the $MLE\text{-}P$ either under- or over- estimate the tail mass.

Finally, in the bottom panel of all three tables, statistics associated with estimation of α_1 are presented. The (estimated) bias and mean squared error of the $NPMLE$ of α_1 (denoted by $\widehat{\alpha}_1$), calculated as $BIAS = \widehat{E}(\widehat{\alpha}_1 - \alpha_1)$ and $MSE = \widehat{E}\left[(\widehat{\alpha}_1 - \alpha_1)^2\right]$ respectively, are reported, along with the ratios of these figures to the corresponding figures for the $MLE\text{-}P$ of α_1 ($\widehat{\alpha}_1^P$). For all three panels in each table, results that are favourable to the $NPMLE$ (i.e. ratios with magnitude less than unity) are highlighted in bold font.

As is indicated by the results in the top panel of Table 1, the $AV.\ MSE$ for the $NPMLE$ of f_i, across the full support, like the corresponding figures for the correctly specified parametric estimator, are negligible in absolute size, most notably for the larger sample sizes. In that sense the $NPMLE$ is competitive with the correctly specified $MLE\text{-}P$ in finite samples. In other words, even if the true arrivals distribution were *known* to be Poisson, use of the nonparametric approach would not lead to qualitatively different predictive conclusions than if the correctly specified parametric estimator were used. $AV.$ MSE for the $NPMLE$ declines monotonically with the sample size, in accordance with the theoretical consistency of the estimator. In comparison with its performance over the full support, the performance of the $NPMLE$ in estimating the upper 10% tail of $\{f_i\}$ (recorded in the second panel of Table 1) is more competitive, overall, with that of the correctly specified $MLE\text{-}P$, with lower $AV.\ BIAS$ actually recorded for the $NPMLE$ in two cases.

The results reported in the bottom panel of Table 1 show that both the nonparametric and parametric estimators of α_1 are slightly negatively biased, with both the $BIAS$ and MSE of the $NPMLE$ declining monotonically in T. Again, although the parametric estimator is superior to the nonparametric estimator, according to both measures, this superiority declines as T increases, with the $BIAS$ and MSE of the $NPMLE$ being zero to two decimal places for $T = 1000$, for both values of α_1. The magnitudes of the corresponding $BIAS$ and MSE values for the $NPMLE$ of α_1, across the two different true values for α_1, are very similar.

When the true DGP has binomial arrivals and the $MLE\text{-}P$ is misspecified as a consequence, the results recorded in Table 2 show the $NPMLE$ to be uniformly more accurate than the $MLE\text{-}P$ in estimating $\{f_i\}$, for $T = 500$ and $T = 1000$. This result holds both for estimation over the full support (first panel) and for estimation of the upper tail (second panel). The $NPMLE$ is only slightly less accurate, in terms of $AV.\ MSE$, in three of the

---

[7]The estimated bias across the *full* support of the count variable is equal to zero due to the summation restriction imposed on estimated and true forecast distributions.

Table 1: Finite sampling performance of the $NPMLE$ and MLE-P;

True Poisson arrivals

Figures in bold denote results that are favourable to the $NPMLE$

	$\varepsilon_t \sim Pois$			$\varepsilon_t \sim Pois$		
	$\lambda = 2$; $\alpha_1 = 0.2$			$\lambda = 2$; $\alpha_1 = 0.6$		
T :	100	500	1000	100	500	1000
	Average over all i			Average over all i		
$AV.\ MSE\ of\ \widehat{f}_i$	0.0008	0.0001	6.3e-005	0.0005	8.8e-005	4.1e-005
$\frac{AV.\ MSE\ of\ \widehat{f}_i}{AV.\ MSE\ of\ \widehat{f}_i^P}$	5.1779	4.4981	4.5802	4.7491	4.6934	4.4943
	Average in upper 10% tail			Average in upper 10% tail		
$AV.\ BIAS\ of\ \widehat{f}_i$	7.0e-005	-3.7e-005	-4.2e-005	-2.8e-005	-8.7e-006	-2.8e-005
$\frac{AV.\ BIAS\ of\ \widehat{f}_i}{AV.\ BIAS\ of\ \widehat{f}_i^P}$	**0.5899**	7.1831	1.5091	**-0.1934**	18.4619	1.8453
$AV.\ MSE\ of\ \widehat{f}_i$	0.0002	3.9e-005	1.8e-005	0.0002	2.7e-005	1.3e-005
$\frac{AV.\ MSE\ of\ \widehat{f}_i}{AV.\ MSE\ of\ \widehat{f}_i^P}$	4.4037	3.8901	3.8828	4.6177	3.8970	3.9846
$BIAS\ of\ \widehat{\alpha}_1$	-0.0506	-0.0084	-0.0032	-0.0568	-0.0067	-0.0030
$\frac{BIAS\ of\ \widehat{\alpha}_1}{BIAS\ of\ \widehat{\alpha}_1^P}$	4.1138	2.4706	3.5556	9.3115	4.7857	5.1976
$MSE\ of\ \widehat{\alpha}_1$	0.0167	0.0025	0.0011	0.0147	0.0014	0.0006
$\frac{MSE\ of\ \widehat{\alpha}_1}{MSE\ of\ \widehat{\alpha}_1^P}$	1.7579	1.3158	1.1000	3.8684	2.0000	1.9500

Table 2: Finite sampling performance of the $NPMLE$ and MLE-P;

True binomial arrivals;

Figures in bold denote results that are favourable to the $NPMLE$

	$\varepsilon_t \sim Bin$			$\varepsilon_t \sim Bin$		
	$n = 4;\ \pi = 0.4;\ \alpha_1 = 0.2$			$n = 4;\ \pi = 0.4;\ \alpha_1 = 0.6$		
$T:$	100	500	1000	100	500	1000
	Average over all i			Average over all i		
$AV.\ MSE\ of\ \widehat{f}_i$	0.0010	0.0002	7.9e-005	0.0005	8.7e-005	4.023e-005
$\frac{AV.\ MSE\ of\ \widehat{f}_i}{AV.\ MSE\ of\ \widehat{f}_i^P}$	**0.7588**	**0.1700**	**0.0841**	1.2156	**0.3128**	**0.1596**
	Average in upper 10% tail			Average in upper 10% tail		
$AV.\ BIAS\ of\ \widehat{f}_i$	-6.8e-005	-0.0002	-5.2e-005	0.0012	4.9e-005	-0.0001
$\frac{AV.\ BIAS\ of\ \widehat{f}_i}{AV.\ BIAS\ of\ \widehat{f}_i^P}$	**0.0135**	**0.0415**	**0.0112**	**-0.1705**	**-0.0065**	**0.0147**
$AV.\ MSE\ of\ \widehat{f}_i$	0.0012	0.0002	0.0001	0.0010	0.0002	8.1e-005
$\frac{AV.\ MSE\ of\ \widehat{f}_i}{AV.\ MSE\ of\ \widehat{f}_i^P}$	1.2241	**0.2652**	**0.1362**	1.8532	**0.4273**	**0.2035**
$BIAS\ of\ \widehat{\alpha}_1$	-0.0155	-0.0024	-0.0010	-0.0316	-0.0067	-0.0028
$\frac{BIAS\ of\ \widehat{\alpha}_1}{BIAS\ of\ \widehat{\alpha}_1^P}$	**-0.1158**	**-0.0170**	**-0.0073**	**-0.3802**	**-0.0790**	**-0.0330**
$MSE\ of\ \widehat{\alpha}_1$	0.0105	0.0020	0.0010	0.0087	0.0014	0.0007
$\frac{MSE\ of\ \widehat{\alpha}_1}{MSE\ of\ \widehat{\alpha}_1^P}$	**0.3519**	**0.0909**	**0.0475**	**0.9774**	**0.1867**	**0.0884**

Table 3: Finite sampling performance of the $NPMLE$ and $MLE\text{-}P$;

True negative binomial arrivals;

Figures in bold denote results that are favourable to the $NPMLE$

	$\varepsilon_t \sim NBin$			$\varepsilon_t \sim NBin$		
	$v = 5; \pi = 0.3; \alpha_1 = 0.2$			$v = 5; \pi = 0.3; \alpha_1 = 0.6$		
$T:$	100	500	1000	100	500	1000
	Average over all i			Average over all i		
$AV.\ MSE\ of\ \widehat{f_i}$	0.0006	0.0001	5.2e-005	0.0004	7.6e-005	3.9e-005
$\frac{AV.\ MSE\ of\ \widehat{f_i}}{AV.\ MSE\ of\ \widehat{f_i^P}}$	2.0570	**0.4772**	**0.2496**	2.6267	**0.7795**	**0.4400**
	Average in upper 10% tail			Average in upper 10% tail		
$AV.\ BIAS\ of\ \widehat{f_i}$	0.0001	9.0e-006	-2.5e-005	0.0001	-5.0e-005	3.2e-006
$\frac{AV.\ BIAS\ of\ \widehat{f_i}}{AV.\ BIAS\ of\ \widehat{f_i^P}}$	**-0.0639**	**-0.0049**	**0.0132**	**-0.1478**	**0.0506**	**-0.0035**
$AV.\ MSE\ of\ \widehat{f_i}$	0.0001	2.6e-005	1.2e-005	0.0001	2.2e-005	1.3e-005
$\frac{AV.\ MSE\ of\ \widehat{f_i}}{AV.\ MSE\ of\ \widehat{f_i^P}}$	3.2918	1.2727	**0.6922**	2.8432	1.0881	**0.7735**
$BIAS\ of\ \widehat{\alpha}_1$	-0.0571	-0.0070	-0.0034	-0.0630	-0.0067	-0.0026
$\frac{BIAS\ of\ \widehat{\alpha}_1}{BIAS\ of\ \widehat{\alpha}_1^P}$	1.2682	**0.1894**	**0.0929**	0.9850	**0.1110**	**0.0475**
$MSE\ of\ \widehat{\alpha}_1$	0.0180	0.0022	0.0010	0.0182	0.0013	0.0006
$\frac{MSE\ of\ \widehat{\alpha}_1}{MSE\ of\ \widehat{\alpha}_1^P}$	2.1190	**0.8298**	**0.4845**	2.2751	**0.3212**	**0.1631**

four cases for $T = 100$. Again, as noted above for the Poisson arrivals case, the absolute level of accuracy with which both methods estimate the true forecast distribution is high; that said, the overall superiority of the nonparametric estimator, for the larger sample sizes in particular, is worthy of note.

As indicated by the results in the third panel of Table 2, for all sample sizes, the $NPMLE$ has smaller $BIAS$ and MSE than the $MLE\text{-}P$ in estimating α_1. Most notably, when the true value of α_1 is quite low ($\alpha_1 = 0.2$), and for the larger sample sizes, the magnitude of the $BIAS$ of the $MLE\text{-}P$ ranges from (approximately) 13 to 137 times greater than that of the $NPMLE$. Clearly, misspecification of the arrivals process impacts on the ability of the parametric estimator to accurately estimate the dynamics in the data. Moreover, a comparison of the corresponding figures in the first two rows of the bottom panel in Table 2 shows that whilst the $BIAS$ and MSE of the $NPMLE$ of both values of α_1 decline uniformly with T, the $BIAS$ of the $MLE\text{-}P$ (for both values of α_1), and the MSE of the $MLE\text{-}P$ (in the $\alpha_1 = 0.6$ case), do not.

Whilst the results recorded in Table 3 - for the case of true negative binomial arrivals - are not as clear cut as those in Table 2, the $NPMLE$ is still the superior estimator, overall, for the larger sample sizes. The $NPMLE$ has smaller $AV. \ MSE$ values, over the full support (first panel), than does the $MLE\text{-}P$, for $T = 500$ and $T = 1000$, and in the upper tail (second panel) for $T = 1000$. Most notably, the magnitude of $AV. \ BIAS$ for the $NPMLE$ in the upper tail is uniformly (i.e. for all values of T and for both values of α_1) lower than that of the $MLE\text{-}P$, with the latter underestimating the tail probability in *all* cases. Once again, the misspecification of the arrivals process appears to impact on the ability of the $MLE\text{-}P$ to accurately estimate α_1, with the $NPMLE$ being more accurate, according to both measures, for the larger sample sizes. Both the $BIAS$ and MSE of the $NPMLE$, for both values of α_1, decline uniformly with T.

4   Empirical Applications

4.1   Data Description

In this section we apply the $NPMLE$ to three empirical series of count data. The first series constitutes $T = 120$ monthly counts of workers collecting wage loss benefits for burns injuries received whilst working in the British Columbia (Canada) logging industry from January 1984 to December 1994. This data set (denoted hereafter by BURNS) has been analysed using $INAR$-type specifications in Freeland and McCabe (2004a,b) and McCabe and Martin (2005). During any month t, the observed number of claimants, X_t, is the sum of the number of claimants from the previous period who continue to collect benefits (i.e. who remain in the claims queue), $\alpha_1 \circ X_{t-1}$, and the number of newly injured workers (i.e.

'arrivals' in the queue), ε_t. The BURNS data assumes values of 0, 1 and 2 only, due to the infrequency with which burn injuries occurred over the relevant time period.

The second data set comprises $T = 3072$ counts of 'iceberg' buy orders (bids) in the order book (up to and including the fifth best bid only) of Deutsche Telekom stock, collected every 10 minutes on the XETRA system of the Deutsche Borse (denoted hereafter by DEUT). The data is recorded over the 8 hours of each of the 64 trading days in first quarter of 2004.[8] 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, but 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.[9] 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 ten 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. The DEUT data in this sample assumes values of 0 to 5 (inclusive) only, due to the infrequency with which iceberg bids occur.[10]

The final data set, denoted by IRAQ, comprises daily counts of violent deaths suffered by Iraqi civilians during the 1 January 2006 to 31 December 2006 period. The time series was constructed from the information provided on the website http://www.iraqbodycount.org. As detailed on this website, the database documents violent incidents that have led to loss of life of Iraqi citizens, as reported either in the media or, in certain cases, by non-government organizations. For some deaths the database provides only a range of possible dates, and in these cases we choose the earliest of these dates. Two extreme counts of 18 and 19 were omitted from the data set, leaving counts ranging from 0 to 11 (inclusive).

The sample autocorrelation functions of the BURNS and DEUT data sets indicate significant first-order autocorrelation, indicating that there is indeed dependence to be modelled and predictive power in the data. Given that both of these data sets may clearly

---

[8]This data has been kindly supplied by Joachim Grammig, with the permission of the Deutsche Börse. A detailed analysis of the impact of iceberg orders on price dynamics is conducted in Frey and Sandas (2008). Jung and Tremayne (2008) also analyse the autocorrelation properties of these count time series, using the $INAR$ family of models.

[9]In the set of German stocks analysed by Frey and Sandas (2008), iceberg orders account for only 8% of shares traded. Note that 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.

[10]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. Hence, it is quite possible for there to be more than five iceberg trades recorded after any 10 minute interval. The Deutsche Telekom series, however, has no count exceeding the value of 5.

be interpreted as time series observations on the number of members of a queue, the $INAR(1)$ specification is inherently suitable for modelling these data. In the case of the IRAQ data set, the number of deaths, X_t at time t, may be thought of as those exogenously generated by the conflict ('immigrants'), ε_t, combined with those ('offspring'), $\alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p}$, which result - both directly and indirectly - from deaths at earlier periods. The impulse response function for the $INAR(p)$ model provides some insight into this process. Consider, for example, the model with $p = 2$ and $X_0 = X_{-1} = 0$. In this case,

$$
\begin{aligned}
X_1 &= \varepsilon_1 \\
X_2 &= \alpha_1 \circ X_1 + \varepsilon_2 \\
X_3 &= \alpha_1 \circ X_2 + \alpha_2 \circ X_1 + \varepsilon_3 \\
&=^d \varepsilon_3 + \alpha_1 \circ \varepsilon_2 + \alpha_1^2 \circ \varepsilon_1 + \alpha_2 \circ \varepsilon_1.
\end{aligned}
$$

The deaths in day 3 may be disaggregated into the following components. The term ε_3 represents those deaths generated exogenously in day 3 (immigrants). The deaths on day 3 that have occurred as the result of retaliation to the shock ε_2 in day 2 are $\alpha_1 \circ \varepsilon_2$ (offspring). The offspring in day 3 of day 2's retaliation to the shock ε_1 in day 1 $(\alpha_1 \circ \varepsilon_1)$ are $\alpha_1^2 \circ \varepsilon_1$. Finally, those deaths on day 3 that have occurred as the lagged response to the day 1 shock are $\alpha_2 \circ \varepsilon_1$. The branching process interpretation of the $INAR(p)$ model thus captures the effect that 'violence begets violence'.[11] The IRAQ data set has significant autocorrelation at the third lag and reasonable (although not formally significant) autocorrelation at lag two and hence an $INAR(3)$ model is a reasonable choice of specification.[12]

4.2 Bootstrap Measurement of Sampling Error

Rather than producing pointwise confidence intervals for individual forecast probabilities, we produce a representation of sampling variation in the *entire* estimated forecast distribution (for any $m \geq 1$), retaining the property that the forecast probabilities sum to unity. Given the applicability of the $INAR$ class to the data sets in question, the B bootstrap

---

[11]The binomial thinning operator imposes restrictions on the number of deaths that can occur as retaliation to the exogenous shock, ε_1, in the period immediately following; i.e. $\alpha_2 \circ \varepsilon_1 \leq \varepsilon_1$. Use of a more general operator (see Weiß, 2008, for a survey of thinning operators) would remove this restriction; however, investigation of this option is beyond the scope of this paper. Note that the binomial thinning operator does still allow the cumulated response to ε_1, $\alpha_2 \circ \varepsilon_1 + \alpha_1^2 \circ \varepsilon_1$, to be greater than ε_1.

[12]The results of all preliminary data analysis are available from the authors on request. Note that the choice of value for $p > 1$ in the $INAR(p)$ model in cases where the queue/stock interpretation is not appropriate (as in the IRAQ case) could be based on predictive performance. That is, an ex-post evaluation of predictive accuracy could be used to determine the particular $INAR(p)$ specification to which the optimality criterion is, in turn, applied to produce a forecast distribution. In order to keep the paper within reasonable bounds, we have chosen to use the simple preliminary diagnostic analysis reported in the text to select p in the case of the IRAQ data.

samples of size T are produced using the dynamic structure of (1), based on the assumed value of p. Specifically, for the jth bootstrap sample, we perform the following steps:

1. Given the estimated arrivals distribution, defined by $\hat{G}$, draw the $i.i.d.$ series, $\left\{ \varepsilon_t^{(j)} \right\}_{t=1}^{T}$;

2. Given $\widehat{\alpha}$ and p initial values for X_t, generate $\left\{ x_t^{(j)} \right\}_{t=1}^{T}$ via the $INAR(p)$ model in (1);

3. Use $\left\{ x_t^{(j)} \right\}_{t=1}^{T}$ to produce $\hat{G}^{(j)}$ and $\widehat{\alpha}^{(j)}$;

4. Use $\hat{G}^{(j)}$, $\widehat{\alpha}^{(j)}$ and the *observed* values, $x_T, x_{T-1}, \ldots, x_{T-(p-1)}$, to produce $\left\{ \hat{f}_i^{(j)} \right\}$, where $\left\{ \hat{f}_i^{(j)} \right\}$ denotes the $NPMLE$ of the m-step-ahead forecast distribution, for any $m \geq 1$;[13]

5. Repeat for $j = 1, 2, \ldots, B$.

For each $\left\{ \hat{f}_i^{(j)} \right\}$, $j = 1, 2, \ldots, B$, we calculate the 'distance' between the empirical m-step-ahead forecast distribution, $\left\{ \hat{f}_i \right\}$, and the jth bootstrap distribution using a suitable metric, namely:

$$d_B \left(\left\{ f_i^{(j)} \right\}, \left\{ \hat{f}_i \right\} \right) = \sum_{i=1}^{K} \left| \hat{f}_i^{(j)} - f_i^{(j)} \right|.$$

The B bootstrap distributions are then ranked in the metric and the distributions at various percentiles of the metric noted. For example, measuring distance by the metric, the chance of seeing a distribution that is more 'extreme' than the distribution at the $95th$ percentile is 5%. Given that extreme values of the metric can, potentially, be associated with quite different shapes in the forecast distributions, we also record the forecast distributions ranked two places either side of that at the $95th$ percentile. For example, given the choice of $B = 3000$ below, the distribution at the $95th$ percentile is that with $2850th$ largest value of $d_B \left(\left\{ f_i^{(j)} \right\}, \left\{ \hat{f}_i \right\} \right)$. We also report those distributions with the $2848th$, $2849th$, $2851st$ and $2852nd$ largest values of the metric. These distributions serve to quantify the way in which 'extreme' sampling behaviour can manifest itself over the full predictive support.[14]

---

[13]Note, we have chosen not to use an additional superscript for the forecast horizon m (as would be consistent with the notation used in Section 2.2) for notational clarity.

[14]Comparable results to those recorded in the following section were produced for a second metric:

$$c_B \left(\left\{ f_i^{(j)} \right\}, \left\{ \hat{f}_i \right\} \right) = \sum_{i=1}^{K} \frac{\left| \hat{f}_i^{(j)} - f_i^{(j)} \right|}{1 + \left| \hat{f}_i^{(j)} - f_i^{(j)} \right|} 2^{-i}.$$

Given that the results were qualitatively very similar to those based on d_B, we report the latter results only.

4.3   Empirical Forecast Results

In Figures 1a to 1c respectively, we reproduce the estimated one-step-ahead ($m = 1$) forecast distribution for the BURNS, DEUT and IRAQ data sets, along with the corresponding five bootstrap distributions centred on the $95th$ percentile of the metric, as estimated from $B = 3000$ replications. The forecasts are produced for the last time period in each data set, with the forecasts for BURNS and IRAQ based on all data up to and including the penultimate observation in the relevant samples. For the DEUT data set, in order to mimic a prediction strategy based on very recent data, we present results in which only the final two days of data in the full series ($T = 95$) are used to predict the number of iceberg orders in the last 10 minutes of the last day in the sample.

In Figure 1a the empirical estimate of the one-step-ahead forecast distribution for BURNS claims in December, 1994, allocates non-negligible predictive probability to the values of 0 to 3 only. A probability of approximately 64% is assigned to the event of zero claimants next month, with a probability of half that magnitude (approximately 32%) assigned to the event of a single claimant. The five distributions that define an extreme value of the metric $d_B(.)$ *all* do so by allocating quite a deal more weight to zero and, correspondingly, less weight to one, than does the empirical forecast distribution. The extreme distributions also all essentially reduce the support of the forecast distribution to 0 and 1.

In Figure 1b, the estimated forecast distribution assigns only 22% probability to the event of *no* DEUT iceberg order being included in the 5 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, and needed to be catered for in trading decisions. Four of the extreme distributions indicate an increase in probability to the event of zero bids, and a corresponding decrease in probability to the existence of some degree of hidden liquidity. However, the distribution below that at the $95th$ percentile allocates *less* probability mass to zero bids and, correspondingly, *more* probability to the presence of at least one iceberg bid.[15]

Finally, Figure 1c reproduces the estimated one-step-ahead forecast distribution for IRAQ civilian deaths on 31 December, 2006. While the estimated distribution assigns the bulk of probability mass to very low numbers of deaths (≤ 3), a non-negligible probability is assigned to counts larger than 6. In this case, the extreme distributions are quite variable in their shape, allocating either more *or* less probability mass to various parts of the support, than does the empirical estimate, with the corresponding adjustment made to the mass assigned to the remainder of the support.[16]

---

[15] As successively larger segments of the DEUT data series are used to estimate the forecast distribution, the five extreme bootstrap distributions become increasing similar, being visually indistinguishable from the empirical estimate when the full sample ($T = 3071$) is used for inference.

[16] As has already been noted in the paper, our focus in is on producing an optimal estimate of the forecast

Figure 1: $NPMLE$ estimation of the one-step-ahead forecast distributions for the three empirical count time series

1a. BURNS Wage Benefit Claims

Estimated One–Step–Ahead Forecast Distribution for December 1994 plus

Bootstrap Distributions Ranked at (and Near) the 95th Percentile; T = 119
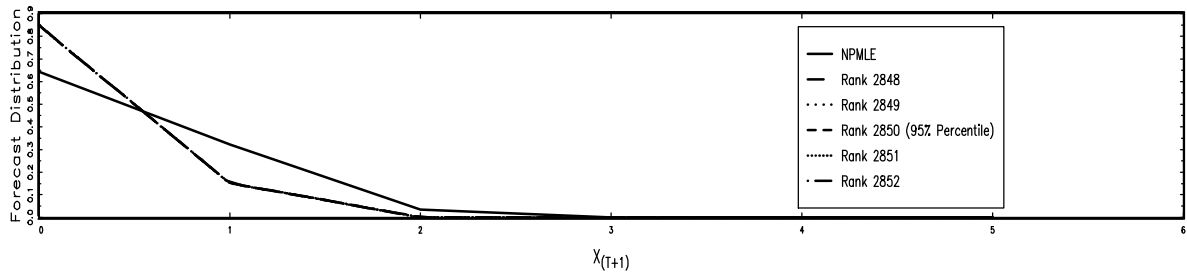
1b. DEUT Iceberg Bids

Estimated One–Step–Ahead Forecast Distribution for Last 10–Minutes of Day plus

Bootstrap Distributions Ranked at (and Near) the 95th Percentile; T = 95

1c. IRAQ Civilian Deaths

Estimated One–Step–Ahead Forecast Distribution for 31 December, 2006, plus

Bootstrap Distributions Ranked at (and Near) the 95th Percentile; T = 362

In Figure 2, we juxtapose the three estimated one-step-ahead forecast distributions with the corresponding distributions for $m = 5$ steps ahead. Along with both empirical estimates we reproduce the bootstrap distribution at the 95th percentile. The five-step-ahead distributions are estimated using the Markov chain structure described in Section 2.2 (see (11)) and first applied in an $INAR$ setting in Bu and McCabe (2008). Due to the stationarity of the models in each case, the forecast distributions five days out are closer (than are the one-step-ahead forecasts) to the corresponding unconditional distributions, as estimated by the sample proportions in each case (recorded in each graph of the right panel in Figure 2). For both the BURNS and IRAQ data, the extent of sampling variability, as measured here by the deviation of the bootstrap distribution at the 95% percentile from the empirical estimate, is less for the five-step-ahead forecast than for the one-step-ahead estimate. This reduction does not occur in the DEUT case.

To conclude the empirical analysis we note that the adoption of the $INAR(1)$ model for two of the three data sets (BURNS and DUET) allows us to estimate in those cases the mean number of time periods that an element will remain in the queue - or mean waiting time - as $w = 1/(1 - \widehat{\alpha}_1)$, where $\widehat{\alpha}_1$ denotes the $NPMLE$ estimate of α_1. We report these estimates in Table 4, along with the estimates of $\widehat{\alpha}_1$ itself. The high frequency order book data exhibits a higher degree of first-order autocorrelation, as measured by $\widehat{\alpha}_1$, than does the lower frequency claims data. Interestingly, in the case of the order data set, the most recent data ($T = 95$) produces the highest value for $\hat{\alpha}_1$, with the degree of autocorrelation declining as soon as the data set used to estimate the model extends further into the past. Associated with the higher value of $\widehat{\alpha}_1$, the expected time spent waiting in the order book is largest when measured using only recently observed data, amounting to more than three periods (half an hour). On the other hand, the smaller values of $\widehat{\alpha}_1$ for the claims data lead to estimates of less than two periods (months) for the average number of months that a worker is expected to be claiming benefits.

5 Conclusions

In this paper we demonstrate an approach to forecasting integer-valued time series data. The method involves estimating the forecast distribution of the random variable in question and, in so doing, allows for the full uncertainty associated with possible future values of the variable to be quantified. Within the $INAR$ class an optimal estimate is produced

---

distribution, within the $INAR$ class, and not on calibrating the estimated forecast distribution with the observed count in time period $T + 1$. However, we do note that the modal prediction for the BURNS data is equivalent to the realized count of 0 at time $T + 1$. The modal prediction for the DEUT data is 1, with the realized value being 0. Finally, the modal prediction for the IRAQ data is 2, whilst the realized value is 1.

Figure 2: One-step-ahead and five-step-ahead forecast distributions for the three empirical count time series

2a. BURNS Wage Benefit Claims

Estimated One–Step–Ahead Forecast Distribution for December 1994
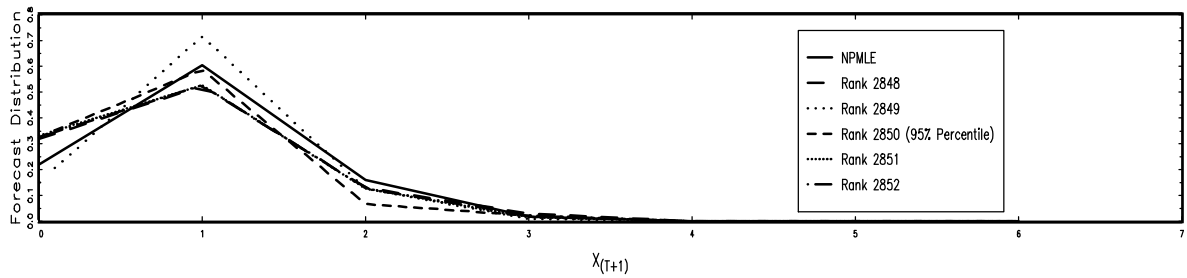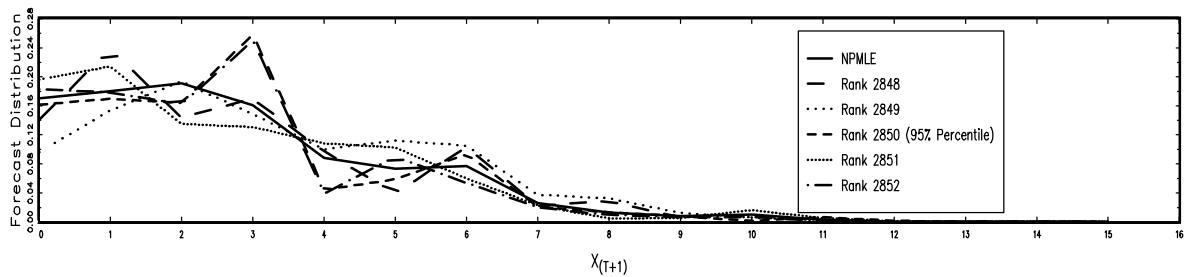
Plus the Bootstrap Distribution at the 95th Percentile; T = 119

2b. BURNS Wage Benefit Claims

Estimated Five–Step–Ahead Forecast Distribution for April 1995

Plus the Bootstrap Distribution at the 95th Percentile; T = 119

2c. DEUT Iceberg Bids

Estimated One–Step–Ahead Forecast Distribution for Last 10–Min Period of Day

Plus the Bootstrap Distribution at the 95th Percentile; T = 95

2d. DEUT Iceberg Bids

Estimated Five–Step–Ahead Forecast Distribution for Fourth 10–Min Period of Next Day

Plus the Bootstrap Distribution at the 95th Percentile; T = 95

2e. IRAQ Civilian Deaths

Estimated One–Step–Ahead Forecast Distribution for 31 December, 2006

Plus the Bootstrap Distribution at the 95th Percentile; T = 362

2f. IRAQ Civilian Deaths

Estimated Five–Step–Ahead Forecast Distribution for 4 January, 2007

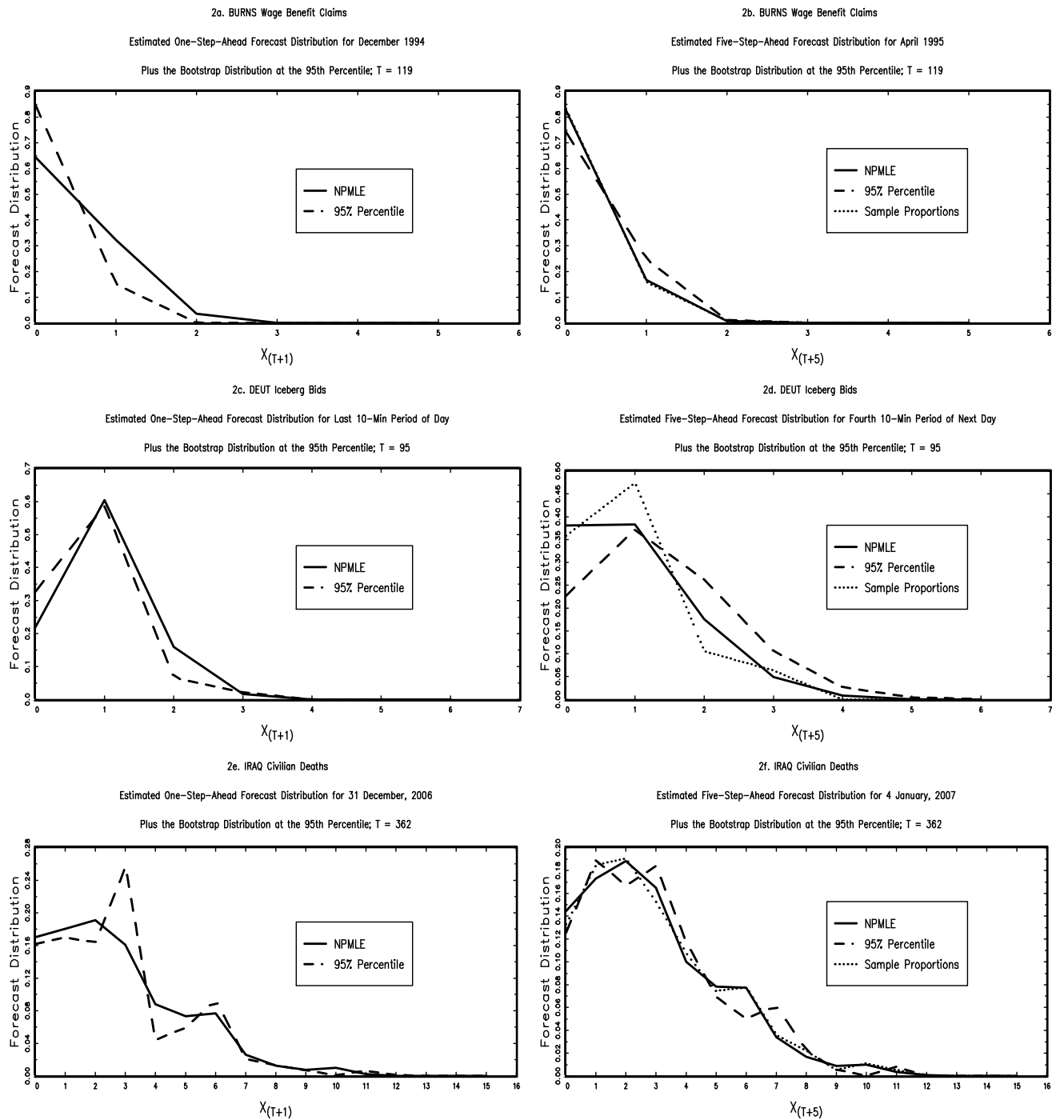Plus the Bootstrap Distribution at the 95th Percentile; T = 362

Table 4: Estimated mean waiting times for the queue ($INAR(1)$) data sets

		Estimated binomial thinning parameter	Estimated mean waiting time
		$\widehat{\alpha}_1$	$w = 1/(1 - \widehat{\alpha}_1)$
BURNS:	$T = 119$	0.2571	1.3461 months
DEUT:	$T = 95$	0.7119	3.4710 ten minute periods
	$T = 500$	0.5410	2.1786 ten minute periods
	$T = 1000$	0.5271	2.1146 ten minute periods
	$T = 3072$	0.5384	2. 1664 ten minute periods

by treating the arrivals process nonparametrically and proving the asymptotic efficiency of the estimated forecast distribution. The $INAR$ model class is applicable to a broad range of count data types which may variously be viewed as a queue, a stock, a birth and death process, or a branching process with immigration. In circumstances where the $INAR$ class applies, the optimality of the estimated distribution provides motivation for its use as a basis for making probabilistic forecasts of the count variable. Even in cases where the class is not *inherently* suitable to a particular count data set, producing the optimal forecast distribution within the $INAR$ class is still a sensible first step prior to comparing - using ex-post methods - with relevant alternatives from outside the class.

Simulation results for the $INAR(1)$ model indicate that the $NPMLE$ performs well even in moderately sized samples. Most notably, the $NPMLE$ is superior, overall, to a misspecified parametric estimator, in particular when estimating the upper tail of the forecast distribution and the dynamic parameter in the $INAR(1)$ model.

We also present a bootstrap-based method for assessing the effect of sampling variation on the $NPMLE$ of the forecast distribution which incorporates the positivity and summation properties of the probabilities involved. Three data sets, all of which may be interpreted as the output of $INAR$ structures, are analysed, with forecast distributions produced and sampling variation assessed.

Appendix

The following preliminary lemma is used in the proofs below:

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*

$$\sum_{j=0}^{i}\left|p_{j|i}(\alpha+h)-p_{j|i}(\alpha)-\frac{\partial p_{j|i}(\alpha)}{\partial\alpha}h\right|\leq 3h^2 i\,(i-1)\,(1+|h|)^{i-2}\leq 3h^2 i^2\,(1+|h|)^i \quad (13)$$

$$\sum_{j=0}^{i}\left|p_{j|i}(\alpha+h)-p_{j|i}(\alpha)\right|\leq 2\,|h|\,i\,(1+|h|)^{i-1}+h^2 i\,(i-1)\,(1+|h|)^{i-2}. \qquad (14)$$

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 (6) and (8) we obtain the expression

$$f^{(1)}_{i_0|i_1,\ldots,i_p}(\theta+h)-f^{(1)}_{i_0|i_1,\ldots,i_p}(\theta)-\dot f^{(1)}_{i_0|i_1,\ldots,i_p}(h)$$

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

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

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

Straightforward rearrangements show that

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

25

and, hence, that

$$\left\| F_{i_1,\ldots,i_p}^{(1)}\left(\theta + h\right) - F_{i_1,\ldots,i_p}^{(1)}\left(\theta\right) - \dot{F}_{i_1,\ldots,i_p}^{(1)}\left(h\right) \right\|_{\ell^1}$$

$$= \sum_{i_0=0}^{\infty} \left| f_{i_0|i_1,\ldots,i_p}^{(1)}\left(\theta + h\right) - f_{i_0|i_1,\ldots,i_p}^{(1)}\left(\theta\right) - \dot{f}_{i_0|i_1,\ldots,i_p}^{(1)}\left(h\right) \right| \tag{15}$$

$$\leq \sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p} \left| \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k + h_{\alpha,k}\right) - \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k\right) - \sum_{k=1}^{p} \frac{\partial p_{j_k|i_k}\left(\alpha\right)}{\partial \alpha_k} h_{\alpha,k} \prod_{\substack{l=1\\l\neq k}}^{p} p_{j_l|i_l}\left(\alpha_k\right) \right|$$

$$+ \|h_G\| \sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p} \left| \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k + h_{\alpha,k}\right) - \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k\right) \right|$$

$$= \sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p}\sum_{k_1=1}^{p}\prod_{\substack{1\leq v\leq p\\v\neq k_1}} p_{j_v|i_v}\left(\alpha_v\right) \times$$

$$\left| p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1} + h_{\alpha,k_1}\right) - p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right) - \frac{\partial p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right)}{\partial \alpha_{k_1}} h_{\alpha,k_1} \right| \tag{16}$$

$$+ \|h_G\| \sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p}\sum_{k_1=1}^{p}\prod_{\substack{1\leq v\leq p\\v\neq k_1}} p_{j_v|i_v}\left(\alpha_v\right) \times$$

$$\left| p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1} + h_{\alpha,k_1}\right) - p_{j_{k_1}|i_{k_1}}\left(\alpha_{k_1}\right) \right| \tag{17}$$

$$+ \left(1 + \|h_G\|\right) \sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p}\sum_{l=2}^{p}\sum_{1\leq k_1<\ldots<k_l\leq p}\prod_{\substack{1\leq v\leq p\\v\neq k_1,\ldots,k_l}} p_{j_v|i_v}\left(\alpha_v\right) \times$$

$$\prod_{1\leq u\leq l} \left| p_{j_{k_u}|i_{k_u}}\left(\alpha_{k_u} + h_{\alpha,k_u}\right) - p_{j_{k_u}|i_{k_u}}\left(\alpha_u\right) \right|. \tag{18}$$

The last step above uses the rearrangements

$$\prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k + h_{\alpha,k}\right) - \prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k\right)$$

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

and

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

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

$$+ \sum_{l=2}^{p} \sum_{1\leq k_1<...<k_l\leq p} \left(\prod_{1\leq u\leq l} \left(p_{j_{k_u}|i_{k_u}}(\alpha_{k_u} + h_{\alpha,k_u}) - p_{j_{k_u}|i_{k_u}}(\alpha_u) \right) \prod_{\substack{1\leq v\leq p \\ v\neq k_1,...,k_l}} p_{j_v|i_v}(\alpha_v) \right).$$

We can now apply the binomial bounds of Lemma 1 in (16)–(18). 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 (16) is bounded; that is,

$$\sum_{k_1=1}^{p} \sum_{j_{k_1}=0}^{i_{k_1}} \left| p_{j_{k_1}|i_{k_1}}(\alpha_{k_1} + h_{\alpha,k_1}) - p_{j_{k_1}|i_{k_1}}(\alpha_{k_1}) - \frac{\partial p_{j_{k_1}|i_{k_1}}(\alpha_{k_1})}{\partial \alpha_{k_1}} h_{\alpha,k_1} \right| \prod_{\substack{1\leq v\leq p \\ v\neq k_1}} \sum_{j_v=0}^{i_v} p_{j_v|i_v}(\alpha_v)$$

$$\leq 3 \sum_{k_1=1}^{p} h_{\alpha,k_1}^2 i_{k_1}^2 (1 + |h_{\alpha,k_1}|)^{i_{k_1}}$$

$$\leq 3 \|h_\alpha\|_{\mathbb{R}^p}^2 D^2 \left(1 + \max_{1\leq k\leq p} |h_{\alpha,k}| \right)^D$$

$$\leq 3 \|h\|_{\mathbb{H}}^2 D^2 (1 + \|h\|_{\mathbb{H}})^D.$$

Similarly we find that (17) is equal to

$$\|h_G\|_{\ell^1} \sum_{k_1=1}^{p} \sum_{j_{k_1}=0}^{i_{k_1}} \left| p_{j_{k_1}|i_{k_1}}(\alpha_{k_1} + h_{\alpha,k_1}) - p_{j_{k_1}|i_{k_1}}(\alpha_{k_1}) \right| \prod_{\substack{1\leq v\leq p \\ v\neq k_1}} \sum_{j_v=0}^{i_v} p_{j_v|i_v}(\alpha_v)$$

$$\leq 3 \|h_G\|_{\ell^1} \sum_{k_1=1}^{p} |h_{\alpha,k_1}| i_{k_1}^2 (1 + |h_{\alpha,k_1}|)^{i_{k_1}} \leq 3 \|h_G\|_{\ell^1} \|h_\alpha\|_{\mathbb{R}^p} D^2 (1 + \|h\|_{\mathbb{H}})^D$$

$$\leq 3 \|h\|_{\mathbb{H}}^2 D^2 (1 + \|h\|_{\mathbb{H}})^D.$$

In the same way, (18) is bounded, with

$$\left(1+\|h_G\|_{\ell^1}\right)\sum_{l=2}^{p}\sum_{1\leq k_1<...<k_l\leq p}\sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p}\prod_{\substack{1\leq v\leq p\\ v\neq k_1,...,k_l}}p_{j_v|i_v}\left(\alpha_v\right)\times$$

$$\prod_{1\leq u\leq l}\left|p_{j_{k_u}|i_{k_u}}\left(\alpha_{k_u}+h_{\alpha,k_u}\right)-p_{j_{k_u}|i_{k_u}}\left(\alpha_u\right)\right|$$

$$=\ \left(1+\|h_G\|_{\ell^1}\right)\sum_{l=2}^{p}\sum_{1\leq k_1<...<k_l\leq p}\prod_{1\leq u\leq l}\sum_{j_{k_u}=0}^{i_{k_u}}\left|p_{j_{k_u}|i_{k_u}}\left(\alpha_{k_u}+h_{\alpha,k_u}\right)-p_{j_{k_u}|i_{k_u}}\left(\alpha_u\right)\right|$$

$$\leq\ 3\left(1+\|h_G\|_{\ell^1}\right)\sum_{l=2}^{p}\sum_{1\leq k_1<...<k_l\leq p}\prod_{1\leq u\leq l}\left|h_{\alpha,k_u}\right|i_{k_u}^2\left(1+\left|h_{\alpha,k_u}\right|\right)^{i_{k_u}}$$

$$\leq\ 3\left(1+\|h_G\|_{\ell^1}\right)\sum_{l=2}^{p}\left(\|h\|_{\mathbb{H}}D^2\left(1+\|h\|_{\mathbb{H}}\right)^D\right)^l$$

$$\leq\ 3p\|h\|_{\mathbb{H}}^2 D^{2p}\left(1+\|h\|_{\mathbb{H}}\right)^{Dp+1}.$$

Thus,

$$\left\|F_{i_1,...,i_p}^{(1)}\left(\theta+h\right)-F_{i_1,...,i_p}^{(1)}\left(\theta\right)-\dot{F}_{i_1,...,i_p}^{(1)}\left(h\right)\right\|_{\ell^1}$$

$$=\ \sum_{i_0=0}^{\infty}\left|f_{i_0|i_1,...,i_p}^{(1)}\left(\theta+h\right)-f_{i_0|i_1,...,i_p}^{(1)}\left(\theta\right)-\dot{f}_{i_0|i_1,...,i_p}^{(1)}\left(h\right)\right|$$

$$\leq\ 6\|h\|_{\mathbb{H}}^2 D^2\left(1+\|h\|_{\mathbb{H}}\right)^D+3p\|h\|_{\mathbb{H}}^2 D^{2p}\left(1+\|h\|_{\mathbb{H}}\right)^{Dp+1}\qquad(19)$$

$$\leq\ C_1^2\|h\|_{\mathbb{H}}^2$$

for a finite constant C_1.

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

$$\left\|\dot{F}_{i_1,...,i_p}^{(1)}\left(h\right)\right\|_{\ell^1}\ =\ \sum_{i_0=0}^{\infty}\left|\dot{f}_{i_0|i_1,...,i_p}^{(1)}\left(h\right)\right|$$

$$\leq\ \sum_{i_0=0}^{\infty}\sum_{(j_1,...,j_p)\,\in J(i_0,...,i_p)}\left|h_{G,i_0-(j_1+...+j_p)}\right|\prod_{k=1}^{p}p_{j_k|i_k}\left(\alpha_k\right)$$

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

$$\leq\ \|h_G\|_{\ell^1}+\sum_{k=1}^{p}i_k\left|h_{\alpha,k}\right|\sum_{j_1=0}^{i_1}\cdots\sum_{j_p=0}^{i_p}\prod_{\substack{l=1\\ l\neq k}}^{p}p_{j_l|i_l}\left(\alpha_k\right)$$

$$=\ \|h_G\|_{\ell^1}+\sum_{k=1}^{p}i_k^2\left|h_{\alpha,k}\right|$$

$$\leq\ \left(D^2+1\right)\|h\|_{\mathbb{H}},$$

as required.

Proof of Theorem 2

We will prove that

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

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

$$\left\| F_{i_1,\ldots,i_p}^{(m)}\left(\theta+h\right) - F_{i_1,\ldots,i_p}^{(m)}\left(\theta\right) - \dot{F}_{i_1,\ldots,i_p}^{(m)}\left(h\right) \right\|_{\ell^1} = o\left(\|h\|_{\mathbb{H}}\right) \tag{20}$$

as required for the derivative. It has already been shown in Theorem 1 that (20) holds for $m = 1$ and so we proceed by induction and suppose that it holds for $m-1$ for some $m \ge 2$. Using (10) and by adding and subtracting $\sum_{u=0}^{\infty} f_{i_0|u,i_1,\ldots,i_{p-1}}^{(m-1)}\left(\theta+h\right) f_{u|i_1,\ldots,i_p}^{(1)}\left(\theta\right)$ we get

$$\left\| F_{i_1\ldots i_p}^{(m)}\left(\theta+h\right) - F_{i_1\ldots i_p}^{(m)}\left(\theta\right) - \dot{F}_{i_1\ldots i_p}^{(m)}\left(h\right) \right\|_{\ell^1}$$

$$\le \sum_{u=0}^{\infty}\sum_{i_0=0}^{\infty} \left| f_{i_0|u,i_1,\ldots,i_{p-1}}^{(m-1)}\left(\theta+h\right) - f_{i_0|u,i_1,\ldots,i_{p-1}}^{(m-1)}\left(\theta\right) - \dot{f}_{i_0|u,i_1,\ldots,i_{p-1}}^{(m-1)}\left(h\right) \right| f_{u|i_1,\ldots,i_p}^{(1)}\left(\theta\right) \tag{21a}$$

$$+ \sum_{u=0}^{\infty} \left| f_{u|i_1,\ldots,i_p}^{(1)}\left(\theta+h\right) - f_{u|i_1,\ldots,i_p}^{(1)}\left(\theta\right) - \dot{f}_{u|i_1,\ldots,i_p}^{(1)}\left(h\right) \right|. \tag{21b}$$

In Theorem 1, (15) is bounded by (16), (17) and (18) which, in turn, leads to (19). This is sufficient to bound (21b). The same sequence of steps bounds (21a) 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 $D = \max_{1 \le k \le p} i_k$ is substituted by $D \vee u$. Thus, letting C_{m-1} denote a constant depending on $m-1$, (21a) and (21b) are bounded by

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

$$+ C_{m-1} \|h\|_{\mathbb{H}}^2 D^{2p} \left(1 + \|h\|_{\mathbb{H}}\right)^{Dp+1}$$

$$= C_{m-1} \|h\|_{\mathbb{H}}^2 \sum_{u=0}^{D} D^{2p} \left(1 + \|h\|_{\mathbb{H}}\right)^{Dp+1} f_{u|i_1,\ldots,i_p}^{(1)}\left(\theta\right)$$

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

$$+ C_{m-1} \|h\|_{\mathbb{H}}^2 D^{2p} \left(1 + \|h\|_{\mathbb{H}}\right)^{Dp+1}$$

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

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

$$\le C_m \|h\|_{\mathbb{H}}^2 D^{2p} \left(1 + \|h\|_{\mathbb{H}}\right)^{Dp+1}, \tag{22}$$

29

where

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

The constant C_m is finite because

$$\sum_{u=0}^{\infty} u^{2p}\left(1 + \|h\|_{\mathbb{H}}\right)^{up+1} f_{u|i_1,\dots,i_p}^{(1)}\left(\theta\right)$$

$$= \sum_{u=0}^{\infty}\sum_{(j_1,\dots,j_p)\in J(i_0,\dots,i_p)} g_{u-(j_1+\dots+j_p)} u^{2p}\left(1 + \|h\|_{\mathbb{H}}\right)^{up+1}\prod_{k=1}^{p} p_{j_k|i_k}\left(\alpha_k\right)$$

$$= \sum_{j_1=0}^{i_1} p_{j_1|i_1}\left(\alpha_k\right)\dots\sum_{j_p=0}^{i_p} p_{j_p|i_p}\left(\alpha_k\right)\sum_{u=j_1+\dots+j_p}^{\infty} g_{u-(j_1+\dots+j_p)} u^{2p}\left(1 + \|h\|_{\mathbb{H}}\right)^{up+1}$$

$$\leq \left(1 + \|h_\alpha\|_{\mathbb{R}^p}\right)^{p^2 D}\sum_{u=0}^{\infty} g_u\left(u + pD\right)^{2p}\left(1 + \|h\|_{\mathbb{H}}\right)^{(u+pD)p+1}$$

$$\leq \left(1 + \|h_\alpha\|_{\mathbb{R}^p}\right)^{p^2 D}\sum_{u=0}^{\infty} g_u\left(u + pD\right)^{2p}\left(1 + \|h\|_{\mathbb{H}}\right)^{pu}$$

$$\leq \left(1 + \|h_\alpha\|_{\mathbb{R}^p}\right)^{p^2 D}\binom{2p}{p}\left(pD\right)^{2p}\sum_{u=0}^{\infty} g_u\left(u^2\left(1 + \|h\|_{\mathbb{H}}\right)^u\right)^p,$$

using $(u + pD)^{2p} = \sum_{j=0}^{2p}\binom{2p}{j}u^j\left(pD\right)^{2p-j} \leq u^{2p}\binom{2p}{p}\left(pD\right)^{2p}$. 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 (22).

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

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

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 $\dot{F}_{i_1,\ldots,i_p}^{(m-1)}(h)$ satisfies this bound. It follows that

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

where B_m is a constant. This constant is finite because

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

under the summability conditions on g_u.

Proof of Lemma 1

For completeness we provide details of the bounds used in Lemma 1. First we use the binomial expansions,

$$
\begin{aligned}
(\alpha + h)^j &= \sum_{k=0}^{j} \binom{j}{k} \alpha^{j-k} h^k = \alpha^j + \sum_{k=1}^{j} \binom{j}{k} \alpha^{j-k} h^k \\
(1 - \alpha - h)^{i-j} &= \sum_{l=0}^{i-j} (-1)^l \binom{i-j}{l} (1-\alpha)^{i-j-l} h^l \\
&= (1-\alpha)^{i-j} + \sum_{l=1}^{i-j} (-1)^l \binom{i-j}{l} (1-\alpha)^{i-j-l} h^l,
\end{aligned}
$$

to give

$$
\begin{aligned}
p_{j|i}(\alpha + h) - p_{j|i}(\alpha) &= \binom{i}{j} (\alpha + h)^j (1 - \alpha - h)^{i-j} - \binom{i}{j} \alpha^j (1-\alpha)^{i-j} \\
&= \sum_{k=1}^{i-j} (-1)^k \binom{i}{j} \binom{i-j}{k} \alpha^j (1-\alpha)^{i-j-k} h^k + \sum_{k=1}^{j} \binom{i}{j} \binom{j}{k} \alpha^{j-k} (1-\alpha)^{i-j} h^k \\
&\quad + \sum_{k=1}^{j} \sum_{l=1}^{i-j} (-1)^l \binom{i}{j} \binom{j}{k} \binom{i-j}{l} \alpha^{j-k} (1-\alpha)^{i-j-l} h^{k+l} \\
&= \sum_{k=1}^{i-j} (-1)^k \binom{i}{k} \binom{i-k}{j} \alpha^j (1-\alpha)^{i-j-k} h^k + \sum_{k=1}^{j} \binom{i}{k} \binom{i-k}{j-k} \alpha^{j-k} (1-\alpha)^{i-j} h^k \\
&\quad + \sum_{k=1}^{j} \sum_{l=1}^{i-j} (-1)^l \binom{i}{k} \binom{i-k}{l} \binom{i-k-l}{j-k} \alpha^{j-k} (1-\alpha)^{i-j-l} h^{k+l} \\
&= \sum_{k=1}^{i-j} (-1)^k \binom{i}{k} p_{j|i-k}(\alpha) h^k + \sum_{k=1}^{j} \binom{i}{k} p_{j-k|i-k}(\alpha) h^k \\
&\quad + \sum_{k=1}^{j} \binom{i}{k} h^k \sum_{l=1}^{i-j} (-1)^l \binom{i-k}{l} p_{j-k|i-k-l}(\alpha) h^l.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\sum_{j=0}^{i} \left| p_{j|i}(\alpha + h) - p_{j|i}(\alpha) \right| &\leq \sum_{j=0}^{i-1} \sum_{k=1}^{i-j} \binom{i}{k} p_{j|i-k}(\alpha) |h|^k + \sum_{j=1}^{i} \sum_{k=1}^{j} \binom{i}{k} p_{j-k|i-k}(\alpha) |h|^k \\
&\quad + \sum_{j=1}^{i-1} \sum_{k=1}^{j} \binom{i}{k} |h|^k \sum_{l=1}^{i-j} \binom{i-k}{l} p_{j-k|i-k-l}(\alpha) |h|^l.
\end{aligned}
$$

The three terms on the right-hand-side of the inequality are bounded by three further

inequalities: first

$$\sum_{j=0}^{i-1}\sum_{k=1}^{i-j}\binom{i}{k}p_{j|i-k}\left(\alpha\right)|h|^{k}=\sum_{j=0}^{i-1}\sum_{k=1}^{i-j}\binom{i}{k}\binom{i-k}{j}\alpha^{j}\left(1-\alpha\right)^{i-k-j}|h|^{k}$$

$$=\sum_{j=0}^{i-1}\sum_{k=1}^{i-j}\frac{i!}{k!}\frac{1}{j!\left(i-k-j\right)!}\alpha^{j}\left(1-\alpha\right)^{i-k-j}|h|^{k}$$

$$=\sum_{j=0}^{i-1}\binom{i}{j}\alpha^{j}\left(1-\alpha\right)^{i-j}\sum_{k=1}^{i-j}\binom{i-j}{k}\left(1-\alpha\right)^{-k}|h|^{k}$$

$$=|h|\sum_{j=0}^{i-1}\binom{i}{j}\alpha^{j}\left(1-\alpha\right)^{i-j}\sum_{k=0}^{i-j-1}\binom{i-j}{k+1}\left(1-\alpha\right)^{-k-1}|h|^{k}$$

$$=|h|\sum_{j=0}^{i-1}\binom{i}{j}\alpha^{j}\left(1-\alpha\right)^{i-j-1}\sum_{k=0}^{i-j-1}\frac{i-j}{k+1}\binom{i-j-1}{k}\left(1-\alpha\right)^{-k}|h|^{k}$$

$$\leq|h|\sum_{j=0}^{i-1}\left(i-j\right)\binom{i}{j}\alpha^{j}\left(1-\alpha\right)^{i-j-1}\left(1+\frac{|h|}{1-\alpha}\right)^{i-j-1}$$

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

second

$$\sum_{j=1}^{i}\sum_{k=1}^{j}\binom{i}{k}p_{j-k|i-k}\left(\alpha\right)|h|^{k}=\sum_{j=1}^{i}\sum_{k=1}^{j}\binom{i}{k}\binom{i-k}{j-k}\alpha^{j-k}\left(1-\alpha\right)^{i-j}|h|^{k}$$

$$=\sum_{j=1}^{i}\sum_{k=1}^{j}\frac{i!}{k!}\frac{1}{\left(j-k\right)!\left(i-j\right)!}\alpha^{j-k}\left(1-\alpha\right)^{i-j}|h|^{k}$$

$$=\sum_{j=1}^{i}\binom{i}{j}\alpha^{j}\left(1-\alpha\right)^{i-j}\sum_{k=1}^{j}\binom{j}{k}\alpha^{-k}|h|^{k}$$

$$=|h|\sum_{j=1}^{i}\binom{i}{j}\alpha^{j}\left(1-\alpha\right)^{i-j-1}\sum_{k=0}^{j-1}\frac{j}{k+1}\binom{j-1}{k}\alpha^{-k}|h|^{k}$$

$$\leq|h|\sum_{j=1}^{i}j\binom{i}{j}\alpha^{j-1}\left(1-\alpha\right)^{i-j}\left(1+\frac{|h|}{\alpha}\right)^{j-1}$$

$$=|h|\,i\sum_{j=0}^{i-1}\binom{i-1}{j}\left(\alpha+|h|\right)^{j}\left(1-\alpha\right)^{i-j-1}$$

$$=|h|\,i\left(1+|h|\right)^{i-1}$$

and third

$$\sum_{j=1}^{i-1}\sum_{k=1}^{j}\binom{i}{k}|h|^{k}\sum_{l=1}^{i-j}\binom{i-k}{l}p_{j-k|i-k-l}\left(\alpha\right)|h|^{l}$$

$$=\sum_{j=1}^{i-1}\sum_{k=1}^{j}\frac{i!}{k!}|h|^{k}\sum_{l=1}^{i-j}\frac{1}{l!}\frac{1}{\left(j-k\right)!\left(i-j-l\right)!}\alpha^{j-k}\left(1-\alpha\right)^{i-j-l}|h|^{l}$$

$$=\sum_{j=1}^{i-1}\binom{i}{j}\alpha^{j}\left(1-\alpha\right)^{i-j}\sum_{k=1}^{j}\binom{j}{k}\alpha^{-k}|h|^{k}\sum_{l=1}^{i-j}\binom{i-j}{l}\left(1-\alpha\right)^{-l}|h|^{l}$$

$$=h^{2}\sum_{j=1}^{i-1}\binom{i}{j}\alpha^{j}\left(1-\alpha\right)^{i-j}\sum_{k=0}^{j-1}\binom{j}{k+1}\alpha^{-k-1}|h|^{k}\sum_{l=0}^{i-j-1}\binom{i-j}{l+1}\left(1-\alpha\right)^{-l-1}|h|^{l}$$

$$=h^{2}\sum_{j=1}^{i-1}\binom{i}{j}\alpha^{j-1}\left(1-\alpha\right)^{i-j-1}\sum_{k=0}^{j-1}\frac{j}{k+1}\binom{j-1}{k}\alpha^{-k}|h|^{k}\times$$

$$\sum_{l=0}^{i-j-1}\frac{i-j}{l+1}\binom{i-j-1}{l}\left(1-\alpha\right)^{-l}|h|^{l}$$

$$\leq h^{2}\sum_{j=1}^{i-1}j\left(i-j\right)\binom{i}{j}\alpha^{j-1}\left(1-\alpha\right)^{i-j-1}\sum_{k=0}^{j-1}\binom{j-1}{k}\alpha^{-k}|h|^{k}\times$$

$$\sum_{l=0}^{i-j-1}\binom{i-j-1}{l}\left(1-\alpha\right)^{-l}|h|^{l}$$

$$\leq h^{2}i\left(i-1\right)\sum_{j=1}^{i-1}\binom{i-2}{j-1}\alpha^{j-1}\left(1-\alpha\right)^{i-j-1}\sum_{k=0}^{j-1}\binom{j-1}{k}\alpha^{-k}|h|^{k}\times$$

$$\sum_{l=0}^{i-j-1}\binom{i-j-1}{l}\left(1-\alpha\right)^{-l}|h|^{l}$$

$$=h^{2}i\left(i-1\right)\sum_{j=0}^{i-2}\binom{i-2}{j}\alpha^{j}\left(1-\alpha\right)^{i-2-j}\sum_{k=0}^{j}\binom{j}{k}\alpha^{-k}|h|^{k}\times$$

$$\sum_{l=0}^{i-2-j}\binom{i-2-j}{l}\left(1-\alpha\right)^{-l}|h|^{l}$$

$$=h^{2}i\left(i-1\right)\sum_{j=0}^{i-2}\binom{i-2}{j}\alpha^{j}\left(1-\alpha\right)^{i-2-j}\left(1+\frac{|h|}{\alpha}\right)^{j}\left(1+\frac{|h|}{1-\alpha}\right)^{i-2-j}$$

$$=h^{2}i\left(i-1\right)\sum_{j=0}^{i-2}\binom{i-2}{j}\left(\alpha+|h|\right)^{j}\left(1-\alpha+|h|\right)^{i-2-j}$$

$$=h^{2}i\left(i-1\right)\left(1+|h|\right)^{i-2}.$$

This completes the proof of (14). To prove (13) note that

$$
\sum_{j=0}^{i} \left| p_{j|i}\left(\alpha + h\right) - p_{j|i}\left(\alpha\right) - \frac{\partial p_{j|i}\left(\alpha\right)}{\partial \alpha} h \right|
$$

$$
\leq \sum_{j=0}^{i-2} \sum_{k=2}^{i-j} \binom{i}{k} p_{j|i-k}\left(\alpha\right) |h|^{k} + \sum_{j=2}^{i} \sum_{k=2}^{j} \binom{i}{k} p_{j-k|i-k}\left(\alpha\right) |h|^{k}
$$

$$
+ \sum_{j=1}^{i-1} \sum_{k=1}^{j} \binom{i}{k} |h|^{k} \sum_{l=1}^{i-j} \binom{i-k}{l} p_{j-k|i-k-l}\left(\alpha\right) |h|^{l},
$$

and we can see that the third term on the right-hand-side above, is already bounded by the development above. The first term satisfies

$$
\sum_{j=0}^{i-2} \sum_{k=2}^{i-j} \binom{i}{k} p_{j|i-k}\left(\alpha\right) |h|^{k} = \sum_{j=0}^{i-2} \sum_{k=2}^{i-j} \binom{i}{k} \binom{i-k}{j} \alpha^{j} \left(1-\alpha\right)^{i-k-j} |h|^{k}
$$

$$
= \sum_{j=0}^{i-2} \sum_{k=2}^{i-j} \frac{i!}{k!} \frac{1}{j!\left(i-k-j\right)!} \alpha^{j} \left(1-\alpha\right)^{i-k-j} |h|^{k}
$$

$$
= \sum_{j=0}^{i-2} \binom{i}{j} \alpha^{j} \left(1-\alpha\right)^{i-j} \sum_{k=2}^{i-j} \binom{i-j}{k} \left(1-\alpha\right)^{-k} |h|^{k}
$$

$$
= h^{2} \sum_{j=0}^{i-2} \binom{i}{j} \alpha^{j} \left(1-\alpha\right)^{i-j} \sum_{k=0}^{i-j-2} \binom{i-j}{k+2} \left(1-\alpha\right)^{-k-2} |h|^{k}
$$

$$
= h^{2} \sum_{j=0}^{i-2} \binom{i}{j} \alpha^{j} \left(1-\alpha\right)^{i-j-2} \sum_{k=0}^{i-j-2} \frac{\left(i-j\right)\left(i-j-1\right)}{\left(k+2\right)\left(k+1\right)} \binom{i-j-2}{k} \left(1-\alpha\right)^{-k} |h|^{k}
$$

$$
\leq h^{2} \sum_{j=0}^{i-2} \left(i-j\right)\left(i-j-1\right) \binom{i}{j} \alpha^{j} \left(1-\alpha\right)^{i-j-2} \sum_{k=0}^{i-j-2} \binom{i-j-2}{k} \left(1-\alpha\right)^{-k} |h|^{k}
$$

$$
= h^{2} \sum_{j=0}^{i-2} i\left(i-1\right) \binom{i-2}{j} \alpha^{j} \left(1-\alpha\right)^{i-j-2} \left(1 + \frac{|h|}{1-\alpha}\right)^{i-j-2}
$$

$$
= h^{2} i\left(i-1\right) \sum_{j=0}^{i-2} \binom{i-2}{j} \alpha^{j} \left(1-\alpha + |h|\right)^{i-j-2}
$$

$$
= h^{2} i\left(i-1\right) \left(1 + |h|\right)^{i-2}.
$$

and the second,

$$\sum_{j=2}^{i}\sum_{k=2}^{j}\binom{i}{k}p_{j-k|i-k}(\alpha)\,|h|^{k} = \sum_{j=2}^{i}\sum_{k=2}^{j}\binom{i}{k}\binom{i-k}{j-k}\alpha^{j-k}(1-\alpha)^{i-j}\,|h|^{k}$$

$$= \sum_{j=2}^{i}\sum_{k=2}^{j}\frac{i!}{k!}\frac{1}{(j-k)!\,(i-j)!}\alpha^{j-k}(1-\alpha)^{i-j}\,|h|^{k}$$

$$= \sum_{j=2}^{i}\binom{i}{j}\alpha^{j}(1-\alpha)^{i-j}\sum_{k=2}^{j}\binom{j}{k}\alpha^{-k}|h|^{k}$$

$$= h^{2}\sum_{j=2}^{i}\binom{i}{j}\alpha^{j-2}(1-\alpha)^{i-j}\sum_{k=0}^{j-2}\frac{j(j-1)}{(k+2)(k+1)}\binom{j-2}{k}\alpha^{-k}|h|^{k}$$

$$\leq h^{2}\sum_{j=2}^{i}i(i-1)\binom{i-2}{j-2}\alpha^{j-2}(1-\alpha)^{i-j}\left(1+\frac{|h|}{\alpha}\right)^{j-2}$$

$$= h^{2}i(i-1)\sum_{j=0}^{i-2}\binom{i-2}{j}(\alpha+|h|)^{j}(1-\alpha)^{i-2-j}$$

$$= h^{2}i(i-1)(1+|h|)^{i-2}.$$

This completes the proof of (13).

References

[1] Abramson, B. and Clemen, R. (2005). Probabilistic forecasting, *International Journal of Forecasting*, 11, 1-4.

[2] Alkema, L., Raftery, A.E. and Clark, S.J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding, *Annals of Applied Statistics*, 1, 229–248.

[3] 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.

[4] Amisano, G. and Giacomini, R. (2007). Comparing Density Forecasts via Weighted Likelihood Ratio Tests, *Journal of Business and Economic Statistics*, 25, 177-190.

[5] Bao, Y., Lee, T-H. and Saltog, B. (2007). Comparing Density Forecast Models, *Journal of Forecasting,* 26, 203–225.

[6] Berkowitz, J. (2001). Testing density forecasts with applications to risk management, *Journal of Business and Economic Statistics*, 19, 465–474.

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

[8] Brännäs, K. (1994). Estimation and testing in integer valued AR(1) models, *Umea Economic Studies No. 355.* University of Umea.

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

[10] Brännäs, K. and Shahiduzzaman, Q. (2004). Integer-valued moving average modelling of the number of transactions in stocks. *Umea Economic Studies No. 637,* University of Umea.

[11] 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.

[12] 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.

[13] 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.

[14] Corradi, V. and Swanson, N. (2006). Predictive density and conditional confidence interval accuracy tests', *Journal of Econometrics*, 135, 187-228.

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

[16] 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.

[17] Diebold, F.X., Gunther, T. and Tay, A. (1998). Evaluating density forecasts with applications to financial risk management, *International Economic Review*, 39, 863-883.

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

[19] 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.

[20] 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, Series B*, 71, 467-485.

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

[22] Egorova, A.E., Hong, Y. and Li, H. (2006). Validating forecasts of the joint probability density of bond yields: Can affine models beat random walk? *Journal of Econometrics*, 135, 255–284.

[23] Elsner, J.B. and Jagger, T.H. (2006). Prediction models for annual U.S. hurricane counts, *Journal of Climate*, 19, 2935–2952.

[24] Feigen, P.D., Gould, P., Martin, G.M. and Snyder, R.D. (2008). Feasible parameter regions for alternative discrete state space models', *Statistics and Probability Letters*, 78, 2963-2970.

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

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

[27] Freeland, R. and McCabe, B.P.M. (2004b). Forecasting discrete valued low count time series, *International Journal of Forecasting*, 20, 427-434.

[28] Freeland, R. and McCabe, B.P.M. (2005). Asymptotic properties of CLS estimators in the Poisson AR(1) model. *Statistics and Probability Letters,* 73, 147-153.

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

[30] Geweke, J. and Amisano, G. (2009). Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns, Forthcoming, Special Issue on Applied Bayesian Forecasting in Economics, *International Journal of Forecasting.*

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

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

[33] Gneiting, T., Balabdaoui, F. and Raftery, A. (2007). Probabilistic forecasts, calibration and sharpness', *Journal of the Royal Statistical Society (B)*, 69, 243–268.

[34] Gneiting, T., Larson, K.,Westrick, K., Genton, M.G. and Aldrich, E. (2006). Calibrated probabilistic forecasting at the Stateline wind energy centre: the regime-switching space-time (RST) method, *Journal of the American Statistical Association,* 101, 968–979.

[35] Gneiting, T. and Raftery, A.E. (2005). Weather forecasting with ensemble methods, *Science,* 310, 248-290.

[36] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association,* 102, 359-378.

[37] Gneiting, T., Raftery, A.E., Westveld, A.H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Monthly Weather Review,* 133, 1098-1118.

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

[39] Ispany, M., Pap, G. and van Zuijlen, M. (2005). Fluctuation limit of branching processes with immigration and estimation of the means. *Advances in Applied Probability*, 37, 523-538.

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

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

[42] Jung, R. and Tremayne, A. (2006b). Coherent forecasting in integer time series models, *International Journal of Forecasting,* 22, 223-238.

[43] Jung, R. and Tremayne, A. (2008). Count time series with overdispersed data. *Draft paper.*

[44] Krysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology, *Journal of Hydrology,* 249, 2-9.

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

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

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

[48] McKenzie, E. (2003). Discrete variate time series. In: Shanbhag, D.N., Rao, C.R. (Eds.), Handbook of Statistics, vol. 21. Elsevier, Amsterdam, 573–606.

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

[50] Pavlopoulos, H. and Karlis, D. (2007). INAR(1) modelling of overdispersed count series with an environmental application, Forthcoming in *Environmetrics*.

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

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

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

[54] Silva, I. and Silva, M. (2006). Asymptotic distribution of the Yule-Walker estimator for INAR(p) processes. *Statistics and Probability Letters*, 76, 1655-1663.

[55] 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.

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

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

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

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

[60] Weiß, C.H. (2008). Thinning operations for modeling time series of counts—a survey, *Advances in Statistical Analysis*, 92, 319-341.

[61] 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.