x-21-158787-6
F6960

BRITISH LIBRARY OF POLITICAL AND ECONOMIC SCIENCE

NON-GAUSSIAN STRUCTURAL TIME SERIES MODELS

Cristiano Augusto Coelho Fernandes

1990

Thesis submmitted for the degree of Doctor of Philosophy
at the London School of Economics, University of London.

UMI Number: U050284

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI®
Dissertation Publishing

UMI U050284
Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest®

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

This thesis aims to develop a class of state space models for non-Gaussian time series. Our models are based on distributions of the exponential family, such as the Poisson, the negative-binomial, the binomial and the gamma. In these distributions the mean is allowed to change over time through a mechanism which mimics a random walk. By adopting a closed sampling analysis we are able to derive finite dimensional filters, similar to the Kalman filter. These are then used to construct the likelihood function and to make forecasts of future observations. In fact for all the specifications here considered we have been able to show that the predictions give rise to schemes based on an *exponentially weighted moving average* (EWMA). The models may be extended to include explanatory variables via the kind of link functions that appear in GLIM models. This enables nonstochastic slope and seasonal components to be included. The Poisson, negative binomial and bivariate Poisson models are illustrated by considering applications to real data. Monte Carlo experiments are also conducted in order to investigate properties of maximum likelihood estimators and power studies of a post sample predictive test developed for the Poisson model.

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Prof. Andrew Harvey, for his constant help and support. These have been essential for the completion of this thesis. I am also indebted to Prof. Keith Ord for showing me the right approach to the bivariate Poisson model and to Dr. Martin Knott and Dr. Neil Shephard for the valuable conversations and discussions.

Thanks are also due to my family for the encouragment received. To Claudia, whose patience and support have been invaluable during these years, my deepest gratitude.

I also would like to acknowledge the financial support from the Brazilian Government through Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq).

SUMMARY

Page

CHAPTER FOUR BINOMIAL MODELS

CHAPTER FIVE BIVARIATE COUNT DATA MODEL

CHAPTER SIX THE GAMMA-GAMMA MODEL

CHAPTER SEVEN THE RANDOM SUM MODEL

CHAPTER EIGHT MONTE CARLO EXPERIMENTS

LIST OF TABLES

LIST OF FIGURES

NOTATION

The following symbols are commonly used in the text:

' following a scalar denotes the transpose of a vector.

~ following a variable will stand for " distributed as".

˜ above a variable denotes the conditional me!n of the variable distribution.

^ above a variable denotes an artificially generated variable.

≃ following an expression means an approximated result.

(*) at the heading of a section means that the section contains too detailed material on literature review.

NBD stands for negative binomial distribution.

As a rule vectors will be typed in bold case, although in certain self evident situations this will not occur. Other symbols and acronyms are used in the text, but these are either explained or evident from the context.

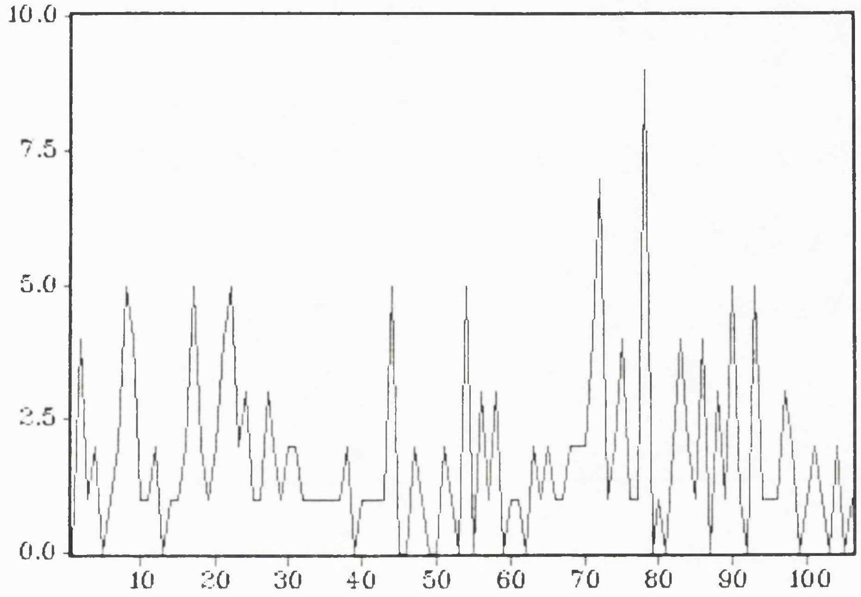
CHAPTER ONE

INTRODUCTION

In the last decade or so several techniques have been proposed to address the problem of non-linearity and/or non-Gaussianity in the modelling of time series. For example, non-linear schemes with Gaussian disturbances are discussed in Priestley (1980) under the umbrella of his state-dependent models (SDM). These may be shown to contain as special cases the bilinear models of Granger and Andersen (1978), the self exciting threshold autoregressive models (SETAR) of Tong and Lim (1981) and the exponential autoregressive models (EAR) of Ozaki (1982). Although of interest in themselves, these formulations will not be considered in our review since our main concern is with explicit non-Gaussian formulations.

Besides the empirical evidence provided elsewhere (see,e.g., Granger and Nelson 1979) sometimes the very nature of the data may suggest, in an obvious way, the inadequacy of a Gaussian scheme. Perhaps the most extreme case of this occurs when the series consists of binary data, i.e., a sequence of zeros and ones in time (see section 4.2.2). A less extreme case is exemplified by count data, in particular when the number of events in each time period is relatively small (see figure 1.1 and Chapter 3).

Figure 1.1 An example of time series of count data: goals scored by England against Scotland in international football matches.



The fitting of Gaussian models to non-Gaussian data meanwhile, should not be viewed as a practice to be avoided at any cost. A good guide is to reconcile pragmatism and good sense. It may be the case that the analysis of this type of data in an appropriate setting could become highly complex, and that the Gaussian framework provides a reasonable approximation for the data in study. Shephard and Harvey (1989), for example, use a continuous Gaussian structural model to track the share of the main political parties in recent British elections. In a more dramatic scenario, as in the case of the purse snatching data analysed in section 3.5, the fitting of a Gaussian model on the untransformed data may produce negative forecasts and prediction intervals straying into the region of negative values. For situations like this the Gaussian approximation is clearly non-admissible. A different approach must be sought.

Transforming data to normality may be a good strategy in a variety of situations. For example, the square root transformation when used in the aforementioned data set works quite well, while the logarithmic transformation is more appropriate for relatively large counts, e.g., casualties in Harvey and Durbin (1986). One must, however, be aware about certain limitations when using such transformations. Firstly, as remarked by McCullagh and Nelder (1983, p.16) very rarely will a single transformation jointly produce symmetry, constant variance and additivity of systematic effects. Furthermore when the model is set up in terms of components of interest such a strategy could jeopardize the interpretation of the modelling process. To avoid these undesirable results a model should 'be unable to predict values which violate definitional constraints", i.e., a model should be *data admissible* (Harvey 1989, p.13).

In order to properly address some situations of intrinsic non normality the *structural approach* (see Harvey 1989) has been extended into a class of non-Gaussian models (see,e.g, Harvey and Fernandes 1989a). This new class of models differs from the *conventional* state space models in that, for most of the specifications considered, neither the measurement nor the transition equations are expressed in external noise form. Instead they are replaced by two conditional probabilistic statements which are shown to capture the essential features necessary to derive *forecasting models*. J.Q. Smith (1979, 1988a-b) (henceforth Smith) discusses this modelling approach from a Bayesian perspective and uses the nomenclature of *partially specified models* for them. A similar approach has also been developed by R. Smith and Miller (1986) (henceforth S&M), but theirs is restricted to an application based on the exponential distribution, a special case of our Gamma-Gamma model, as we shall see.

The particular choice for the state predictive mechanism together with the use of conjugate distributions have allowed us to derive a broad class of simple models with high potential for applications. In particular we have considered situations in which non-normality is present in the form of:

(i) *count data* - as in the Poisson-Gamma model (Ch.3, section 2), the Negative Binomial-Beta model (Ch.3, section 3) and the Bivariate Poisson model (Ch. 6).

(ii) *binomial data* - as in the Binomial-Beta model (Ch. 4).

(iii) *gamma data*- as in the Gamma-Gamma model (Ch. 6).

(iv) *random sum*- as in the Random Sum model (Ch. 7).

Our methodology may be viewed as belonging to the class of *partially specified models*, but, amongst other things, our heuristics differs from the one adopted by Smith. We prefer to look at our framework as naturally motivated by the *structural* paradigm in time series. In particular we advocate a classical treatment for the modelling process, where we make use of *non-informative* priors and estimate the hyperparameters using the maximum likelihood method (ML). Where possible, we try to derive classical tests for model checking, as in the post-sample predictive test for the Poisson-Gamma model.

In addition our framework allows for the introduction of explanatory variables and deterministic structural components (time trend and seasonals). These effects are introduced via suitably chosen link functions, which appear in GLIM models (McCullagh and Nelder 1983). Before embarking on a full description of our class of models, which is left to the next chapter, we would like to introduce some results of interest and consider a review of some of the state space models for time series.

Amongst the already vast literature in the area we will be covering in some detail only those formulations which are closer in spirit to our methodology. We start with the *Gaussian-linear Structural Models* (henceforth SM) of Harvey (1984,1989), since these provide a natural basis of comparison when considering non-linear/non-Gaussian extensions of state space models. It is widely known that an exact and optimal filter (in the MMSE sense) is only obtainable for the Gaussian-linear specification (see,e.g, Anderson and Moore 1979, ch.5). Therefore, by examining the singularities of this formulation we will be better equipped to understand the natural difficulties one will be faced with when considering more general formulations. Our review also covers the *Dynamic Generalized Linear Models* (DGLM) by West, Harrison and Migon (1985) (henceforth WHM) and the Poisson filter by Figliuoli (1988). These formulations will be illustrated with respect to one of the simplest structures describing the level of a series, *the random walk plus noise* or *local level* model, (see Harvey 1989, p.19) since this forms the basis for the construction of our class of non-Gaussian models.

Rather than presenting the filtering and prediction equations for these models in a standard way we would like here to take a different perspective. We have chosen to use the *probabilistic* or *Bayesian* approach to carry out our analyses since this framework is naturally fitted to study state space formulations, in particular, our own class of non-Gaussian structural models. In doing so we will be providing an appropriate setting within which these formulations can be easily analysed and the differences amongst them understood against a common background.

1.1 THE PROBABILISTIC APPROACH IN STATE SPACE MODELS:

The probabilistic approach (see also Kitagawa 1987 and Pole and West 1988) for the presentation and solution of a univariate state space model is given by the following set of equations, where we have defined $\mathbf{Y}_t = \{y_1, y_2, \ldots, y_t\}$ and $\boldsymbol{\theta}_t = \{\theta_1, \theta_2, \ldots, \theta_t\}$. In our notation $p(\cdot \mid \cdot)$ denotes a generic symbol for a probability distribution/density labelled by its argument.

- *measurement equation:* this gives the probabilistic structure of the observable y_t in terms of a dynamic parameter θ_t, the *state* of the process, possibly with a physical interpretation, and a fixed parameter υ, which we call the *secondary* parameter. Their identification will become clear from the context. This equation can either be obtained via the proper manipulation of an external noise measurement equation (as in the Gaussian case; see also Figliuoli 1988) or by direct or internal specification (see Ch.3 onwards and also WHM 1985 and S&M 1986). In particular this model may be chosen from a class of the exponential family, as in WHM (1985). Symbolically we thus have

$$p(y_t \mid \theta_t, \upsilon, \mathbf{Y}_{t-1}) \qquad (1.1.1)$$

where $y_t \in \Upsilon \subset \mathbf{R}$, $\theta_t \in \Theta \subset \mathbf{R}$, $\upsilon \in \mathbf{U} \subset \mathbf{R}$.

– *transition equation:* this gives the dynamics of the state and is generally stated in external noise form, from which, hopefully, the transition density can be obtained

$$p(\theta_t | \theta_{t-1}, \psi, Y_{t-1}) \qquad (1.1.2)$$

where ψ is any existing parameter associated with the dynamics of the state. The parameters υ and ψ are put together in a single vector Ψ, the vector of *hyperparameters*, which, in the classical approach, is estimated *via* ML. In certain cases mechanisms may be derived to make these parameters dynamic, when they will no longer be considered hyperparameters. The most common example is given by the adoption of the ARCH structures for the variances of Gaussian models, as in Figliuoli (1988). See also Chs. 3 and 6. For ease of notation, in the following, we will drop the hyperparameter vector Ψ from the above distributions/densities.

- Filtering and forecasting:

(i) *state prediction (prior):*

$$p(\theta_t | Y_{t-1}) = \int_\Theta p(\theta_t, \theta_{t-1} | Y_{t-1}) \, d\theta_{t-1}$$

$$= \int_\Theta p(\theta_t | \theta_{t-1}, Y_{t-1}) \, p(\theta_{t-1} | Y_{t-1}) \, d\theta_{t-1} \, . \qquad (1.1.3)$$

(ii) *state updating (posterior):* directly obtained by use of Bayes theorem

$$p(\theta_t|Y_t) \propto p(y_t|\theta_t,Y_{t-1})\ p(\theta_t|Y_{t-1}) \qquad (1.1.4)$$

where the constant of proportionality is given by the reciprocal of the subsequent equation.

(iii) *conditional distribution (predictive distribution):*

$$p(y_t|Y_{t-1}) = \int_\Theta p(y_t|\theta_t,Y_{t-1})\ \ p(\theta_t|Y_{t-1})d\theta_t. \qquad (1.1.5)$$

(iv) *multi-step ahead forecasting distribution:* here we want to derive the distribution of y_{T+k} ($k > 2$) conditional on the data observed up to time T. This can be accomplished via two equivalent schemes. The first uses an explicit expression for the multi-step ahead distribution for the state, and is obtained through the following manipulation:

$$p(y_{T+k}|Y_T\ ) = \int_\Theta p(y_{T+k},\theta_{T+k}|Y_T\ )\ d\theta_{T+k}$$

$$= \int_\Theta p(y_{T+k}|\theta_{T+k},Y_T\ )p(\theta_{T+k}|Y_T\ )\ d\theta_{T+k}. \qquad (1.1.6)$$

where the *k-steps ahead density* for the state is given by

$$p(\theta_{T+k}|Y_T)= \int_{\Theta} p(\theta_{T+k}|\theta_{T+k-1},Y_T\ )\ p(\theta_{T+k-1}|Y_T)\ d\theta_{T+k-1}. \quad (1.1.7)$$

Forecasts may also be obtained by integrating out the variables up to time T+k-1 from the joint forecasting distribution. This is as follows:

$$p(y_{T+k}|Y_T\ )= \int_{\Upsilon} \cdots \int_{\Upsilon} p(y_{T+1},\ldots,y_{T+k-1},y_{T+k})\ dy_{T+1}\ldots dy_{T+k-1}$$

$$= \int_{\Upsilon} \cdots \int_{\Upsilon} \prod_{T+1}^{T+k} p(y_t|Y_{t-1})\ dy_{T+1}\ \cdots\ dy_{T+k-1}. \quad (1.1.8)$$

Even when the full forecasting distribution is difficult to evaluate forecasting moments may be derived by use of one of the following properties of conditional expectations:

(i) the chain rule for conditional expectations (see,e.g., Shiryayev 1984, ch.2 §7)

$$E^n(y_{T+k}|Y_T)= E_T\ E_{T+1}\ \cdots\ E_{T+k-2}\ E_{T+k-1}\ (y_{T+k})^n \quad (1.1.9a)$$

where E_{T+j} j=0,1,2,... denotes taking expectations wrt $\mathbf{Y}_{T+j}$ and n= 1,2,... .

(ii) when an explicit transition equation is available then the following standard result may prove to be useful

$$E\,(y^{n}_{T+k} \mid Y_T) = E(\,E(y^{n}_{T+k} \mid \theta_{T+k}) \mid Y_T). \qquad (1.1.9b)$$

1.2 GAUSSIAN-LINEAR STRUCTURAL MODELS:

The *structural* approach in time series modelling (see Harvey 1984,1989) was originally developed for the analysis of Gaussian time series and it is based on the representation of a series in terms of components of interest such as trend, seasonals, cycles, etc. Similar formulations have also been introduced by Harrison and Stevens (1976), Akaike (1980) and Kitagawa (1981). The framework used is that of a linear state space model (see,e.g., Anderson and Moore 1979) and as such the *Kalman filtering* equations (Kalman 1960) can be used for state estimation and likelihood construction. Note that model identification (in the time series sense) is achieved by direct specification of those components one may find relevant to describe the series. Another interesting feature of this approach is that the components are *local* rather than *global*, so that the model has an intrinsic dynamic structure. It can be shown that under specific differencing operations the structural models may be reduced to an equivalent ARIMA structure.

The SM in which the level of the process is described by a random walk structure has the standard form:

$$\text{measurement eqn: } y_t = \theta_t + \epsilon_t \qquad \epsilon_t \sim \text{NID}(0,\sigma^2) \qquad (1.2.1a)$$

$$\text{transition eqn: } \theta_t = \theta_{t-1} + \eta_t \qquad \eta_t \sim \text{NID}(0,q\sigma^2) \qquad (1.2.1b)$$

where $0 < q < \infty$ and $t= 1,\dots,T$. In the above ϵ_t and η_t are independent of each other in all time periods.

Most of the results and proofs presented here for the above specification are well known in the literature (see,e.g., Harvey 1984 and Anderson and Moore 1979). Our contribution lies in the stress we put on the properties which make the filtering and forecasting problem so unique for the Gaussian-linear case, namely, the Markovian evolution for the state and the Gaussian reproducing property under linear combinations. In fact by use of the aforementioned properties the integrals in section 1.1 are trivially solved. We then briefly comment on the potential problems one is likely to meet when trying to redefine the above framework across distributions other than the Gaussian.

- *measurement equation*: the Markovian structure of the transition equation (1.2.1a), implies that the measurement y_t is conditionally independent of the past data, i.e., given θ_t, for all t, the measurements y_t are independent variables (see,e.g, Jazwinski 1970, ch.3). This is formalized by rewriting (1.1.1) as

$$p(y_t | \theta_t, Y_{t-1}) = p(y_t | \theta_t). \qquad (1.2.2)$$

This conditional probabilistic statement is just a different way of restating the measurement equation, traditionally given in external noise form (1.2.1a). Making use of the linear property for Gaussian variables we find that

$$p(y_t | \theta_t) \sim N(\theta_t, \sigma^2) \qquad (1.2.3)$$

where the first element of the vector of hyperparamters is $\Psi_1 = \sigma^2$.

(i) *state prediction*- as a result of the Markovian property we have that

$$p(\theta_t | \theta_{t-1}, Y_{t-1}) = p(\theta_t | \theta_{t-1}). \qquad (1.2.4)$$

Using (1.2.1b) it is straightforward to show that the *transition densities* may be obtained through

$$p(\theta_t | \theta_{t-1}) = p_{\eta_t}(\theta_t - \theta_{t-1}) \qquad (1.2.5)$$

where the density on the rhs is the density of η_t evaluated at $\theta_t - \theta_{t-1}$. The evaluation of the above density is trivial in the Gaussian case given the additive structure of the transition equation. This is given by $N(\theta_{t-1}, \sigma^2)$. This alone is all one needs to derive the joint distribution of the states $p(\theta_1, \theta_2, \ldots, \theta_t)$, since as a corollary of the Markovian evolution

$$p(\theta_t, \theta_{t-1}, \ldots, \theta_1) = \prod_{i=1}^{t} p(\theta_i | \theta_{i-1})$$

with $p(\theta_1 | \theta_0)$ given. Observe than, in view of (1.2.2-3) and the above factorization the joint distribution of the observations $\{y_t\}$ and states $\{\theta_t\}$ may be fully characterized. In fact one may easily show that

$$p(Y_t, \theta_t) = \prod_{i=1}^{t} p(y_i | \theta_i) \; p(\theta_i | \theta_{i-1})$$

where the densities in the rhs are given, respectively by (1.2.3) and

(1.2.5). In the sense of the above, Gaussian state space models are known as *fully* specified state space models.

Apart from very special cases, considered by Bather (1965), the construction of transition equations appropriate for distributions other than the Gaussian is not usually possible. Fortunately, as we will see later, the existence of an explicit transition equation is not a necessary condition for the construction of forecasting models, altough it may facilitate the derivation of multi-step ahead forecasts and smoothing algorithms in certain cases.

Using (1.2.4) the state prediction equation (or prior) is simplified to:

$$p(\theta_t|Y_{t-1}) = \int_\Theta p(\theta_t|\theta_{t-1})\, p(\theta_{t-1}|Y_{t-1})\, d\theta_{t-1}. \qquad (1.2.6)$$

This is known in the literature as the *Chapman-Kolmogorov* equation (see,e.g., Jazwinski 1970, ch.3). Finally by making use of (1.2.5) the above integral may be written as

$$p(\theta_t|Y_{t-1}) = \int_\Theta p_\eta\,(\theta_t - \theta_{t-1})\; p_\theta\,(\theta_{t-1}|Y_{t-1})d\theta_{t-1}. \qquad (1.2.7)$$

It then follows that for the Gaussian-linear case the Chapman-Kolmogorov equation (1.2.6) may be written as a convolution. Again this is a direct consequence of the additive structure of the transition equation (1.2.1b). The second density in the rhs of (1.2.7), is by assumption, $N(m_{t-1},\sigma^2 p_{t-1})$. Using the reproducing property for Gaussian variables, this integral may be easily solved resulting in the predictive step for the Gaussian case:

$$p(\theta_t|Y_{t-1}) \sim N(m_{t/t-1}, p_{t/t-1}) \qquad (1.2.8)$$

where

$$m_{t/t-1} = m_{t-1} \qquad (1.2.9a)$$

$$p_{t/t-1} = \sigma^2(q + p_{t-1}). \qquad (1.2.9b)$$

Observe that in the predictive step, the location parameter of the state distribution remains unchanged while the scale parameter increases. This reflects the fact that immediately after the transition, the component θ_t will be less precise than its previous value, since no observation is used to validate this step.

(ii) *state updating* - it is a well known fact in the literature that Gaussian variables are closed under sampling, i.e., the posterior of a Gaussian variable with fixed variance will also be Gaussian (see,e.g., Berger 1985, ch.4). Using this result one can establish that the *posterior* or *filtering* density is given by

$$p(\theta_t|Y_t) \sim N\ (m_t, p_t) \qquad (1.2.10)$$

where

$$m_t = m_{t/t-1} + K_t(y_t - m_{t/t-1}) \qquad (1.2.11a)$$

$$p_t = p_{t/t-1}\ (1 - K_t) \qquad (1.2.11b)$$

and K_t, the gain, is given by, $K_t = p_{t/t-1}/(1+p_{t/t-1})$.

The above equations, together with the prediction equations (1.2.9) form the so called *Kalman filtering* equations. In a more general environment, where non-Gaussian and/or non-linear equations are considered, the computation of the filtering density may become vastly more difficult. While in the Gaussian-linear case this density is fully characterized by a finite set of dynamic equations, for general distributional assumptions we would need the whole distribution, and it is in this sense that the resulting filter is known as having infinite dimension. This problem is known in the control literature as *the non-linear filter problem* (see,e.g., Jazwinski 1970, ch.9).

(iii) *predictive density*- first, by using the general result of conditional independence for the measurement equation (see 1.2.2), equation (1.1.4) may be simplified to

$$p(y_t | Y_{t-1}) = \int_{\Theta} p(y_t | \theta_t) p(\theta_t | Y_{t-1}) d\theta_t \qquad (1.2.12a)$$

or symbolically,

$$p(y_t | Y_{t-1}) = p(y_t | \theta_t, \Psi) \underset{\theta_t}{\wedge} p(\theta_t | Y_{t-1}). \qquad (1.2.12b)$$

The operation $\wedge$ is known as *mixing* (or *compounding* for discrete distributions) and is a natural consequence of constructing state space models. The resulting distributions contain all the information that is verifiable from the data, and we sometimes refer to them as the *operational model*. They form the basis for constructing the likelihood for hyperparameter estimation and mispecification tests (see,e.g., Harvey 1989, ch.5 and section 3.4). In particular, it is a well known result that the joint distribution of the observations, or the likelihood for the vector of hyperparameters Ψ may be obtained through the following factorization (see,e.g., Harvey 1984, pp.125-147)

$$p(y_1,y_2,\ldots,y_T;\Psi) = \prod_{t=1}^{T} p(y_t|Y_{t-1}). \qquad (1.2.13)$$

Observe that to obtain the *operational model* no explicit use of the transition density has been made (see 1.2.12a-b); the key density in this operation being the state predictive density (see 1.2.7). As we shall see later this important feature will be at the core of our class of non-Gaussian models.

In a general non Gaussian/non-linear environment the resultant *compounding or mixing distribution*, might not have a closed form and therefore be computationally difficult to solve. Hence numerical methods will be required not only to maximize the likelihood function (1.2.13), but also to construct it. For the Gaussian case, this distribution is easily obtained given the additive structure of the measurement equation. Using equations (1.2.3) and (1.2.8) it may be shown that the solution of the compounding operation is given by (see, e.g., Johnson & Kotz 1969, vol.2)

$$\begin{aligned} p(y_t|Y_{t-1}) &= N(\theta_t,\sigma^2) \underset{\theta_t}{\bigtriangleup} N(m_{t/t-1},P_{t/t-1}) \\ &\sim N(\tilde{y}_{t/t-1},f_{t/t-1}) \qquad (1.2.14) \end{aligned}$$

where the moments are given by

$$\tilde{y}_{t/t-1} = m_{t/t-1} \tag{1.2.15a}$$

$$f_{t/t-1} = \sigma^2 + p_{t/t-1} \tag{1.2.15b}$$

with $m_{t/t-1}$ and $p_{t/t-1}$ as in (1.2.9a) and (1.2.9b) , respectively.

A particular feature of the Gaussian-linear model is that both the measurement and predictive densities belong to the same family of distribution. For alternative specifications this will not be true, in general. In addition, the resulting *compounding* or *mixing* distribution may or may not have a direct and simple interpretation in terms of the physical situation being investigated. Given that, by assumption, the measurement equation is data admissible, the virtual arbitrariness of the predictive density may be circumvented by noting that under the Bayesian perspective this distribution may be viewed as the expectation of the measurement distribution over the prior distribution of the state. Therefore for squared error loss, the predictive distribution (or *operational* model) is the optimal estimator of the measurement distribution taken with respect to the density which encapsulates the *modus operandi* imposed in the state evolution. If one feels too uneasy about this interpretation, a more general principle is given by the *predictive point of view* which simply judges a model on the merits of its predictive performance. This has been advocated,e.g., by Akaike (1984).

(iv) *k-steps ahead predictive distribution* - we first notice that by equation (1.2.2) the following simplification occurs

$$p(y_{T+k} \mid \theta_{T+k}, Y_T) = p(y_{T+k} \mid \theta_{T+k})$$

so that equation (1.1.6) may be rewritten as

$$p(y_{T+k} \mid Y_T) = \int_\Theta p(y_{T+k} \mid \theta_{T+k}) p(\theta_{T+k} \mid Y_T)\ d\theta_{T+k} \qquad (1.2.16a)$$

or symbolically,

$$p(y_{T+k} \mid Y_T) = p(y_{T+k} \mid \theta_{T+k}) \underset{\theta_{T+k}}{\wedge} p(\theta_{T+k} \mid Y_T). \qquad (1.2.16b)$$

Now, the first distribution on the rhs of the above equation is readily obtained by direct extrapolation of the measurement equation (1.2.3) and has the form $N(\theta_{T+k}, \sigma^2)$. By use of the Markov property the second density in this *mixing* operation may be written as

$$p(\theta_{T+k} \mid Y_T) = \int_\Theta p(\theta_{T+k} \mid \theta_{T+k-1})\ p(\theta_{T+k-1} \mid Y_T)\ d\theta_{T+k-1}$$

for k=2,3,... . It is easy to show that from (1.2.1b) the state equation projected k steps ahead has the form

$$\theta_{T+k} = \theta_T + \sum_{i=1}^{k} \eta_{T+i}$$

Once more making use of the property of Gaussianity invariance under linear combinations, the repeated application of the *convolution* operation in the previous integral results in the following expression for the state multi-step ahead density:

$$(\theta_{T+k}|Y_T) \sim N(m_{T+k}, p_{T+k}) \qquad (1.2.17)$$

where $m_{T+k/T} = m_T$ (1.2.18a)

$p_{T+k/T} = (p_T + kq)\sigma^2.$ (1.2.18b))

From this the multi-step ahead forecast distribution may be easily obtained by a last application of the convolution property, resulting in

$$p(y_{T+k}|Y_T\ ) = N(\theta_{T+k}, \sigma^2) \underset{\theta_{T+k}}{\vartriangle} N(m_{T+k}, p_{T+k})$$

$$\sim N(\tilde{y}_{T+k/T},\ f_{T+k/T}) \qquad (1.2.19)$$

where $\tilde{y}_{T+k/T} = m_{T+k/T} = m_T$ (1.2.20a)

$f_{T+k} = \sigma^2\ (\ 1+p_T + kq).$ (1.2.20b)

Note that for the above derivation one has first to compute the k-steps ahead density for the state (1.2.17), which is simple to obtain in the Gaussian linear case given the remarkable property of Gaussian invariance under sucessive convolutions.

1.3 NON-GAUSSIAN/NON-LINEAR MODELS:

One natural way to generalize the state space framework is to conserve the *skeleton* of equations (1.2.1a-b) and allow for specifications where the system and the measurement noises are not necessarily Gaussian distributed. Given the possibility of restrictions on the parameter space this linear structure may not be appropriate, and further generalizations in terms of a non-linear transition equation and/or non-linear measurement equation may be called for. Based on some of the previous results the general feeling should be that this strategy is bound to produce intractable models, since, in general, some or all of the distributions/densities of interest may not be given in an analytical form. This problem may be circumnvented by considering a number of techniques which we now briefly review.

1.3.1 *Numerical filters:*

With the rapid development of computer technology, approximation techniques based on the use of numerical methods to reconstruct densities and/or moments have become more attractive. Filters based on these principles we designate by the name of *numerical filters*. A good discussion on the practical problems of implementing such algorithms is provided by Sorenson (1988). Applications of these techniques to real data are given in Kitagawa (1987) who considers non-Gaussian state space models where both the measurement and transition equations have non-Gaussian noises. By specifying some heavy tailed distributions he was able to model general irregularities in time series, such as structural change (abrupt or gradual change in the level and/or slope of a linear model) and outliers. In his approach a rather straightforward method is taken: the probability densities are approximated by a piecewise linear function (or first order spline), and the necessary operations on the densities are realized by numerical computation. Model selection (say,e.g., when competing models have different noise distributions) is done using the AIC criterion and hyperparameter estimation using the likelihood function. One drawback of his method is that it can be extremely computationally demanding, especially when the state is multidimensional (see,e.g., H.K.Tan 1990, ch.2 and Pole and West 1988).

A more recent numerical approach for state space models has been proposed by Pole and West (1988), who consider general non-Gaussian models under the Bayesian perspective. They use Gaussian quadrature as the numerical technique to approximation. The essence of this technique is to approximate integrals by a discrete sum where the weights and grid points involve the Hermite polynomials. For details see Pole and West (1988). The authors claim theirs are more numerically efficient than Kitagawa's, but doubt still remains about the overall effect brought by these successive approximations. It is also legitimate to ask about the improvement in forecasting perfomance and overall CPU time required by these filters when compared either with standard Gaussian filters, like the SM or with the non-Gaussian *analytical filters*, which are considered in the next topic.

1.3.2 *Analytical filters:*

The central idea behind this general technique is to approximate the unknown optimal non-Gaussian filter by a finite dimensional, hence sub-optimal filter. We use the generic term *analytical filters* for this approach. Amongst the existing analytical techniques in the control literature (see, e.g.,Anderson and Moore 1979, ch.8) the extended Kalman filter (EKF) and the Gaussian sum approximations (GSA) are the most widely known. The EKF is based on the adaptation of linear algorithms to non-linear environments, while the GSA, as the name suggests, approximates the state prediction densities by a sum of Gaussian densities. These two techniques will not be considered in this review since they are not directly relevant to the discussion of our class of models. Note that it is also possible to find procedures in the literature which are the result of a mixture of both strategies, and in those cases the prevailing method will dictate the appropriate classification.

Altough interesting in themselves the review that follows might be considered too detailed and may be omitted without affecting the overall understanding of our own material.

1.3.2.1 The DGLM (*)

The Dynamic Generalized Linear Models (DGLM) were introduced by WHM in 1985, and may be viewed as a dynamic extension of the Generalized Linear Models of McCullagh and Nelder (1983) constructed under Bayesian principles. It is characterized by an internal measurement equation and a closed sampling analysis. As such the 'natural parameter' densities are exact, but a number of approximations are used to construct a stochastic mechanism for their parameters. Using the equations of the probabilistic approach we may look at the DGLM as follows:

- *measurement equation*: is chosen from the exponential family,i.e., $p(y_t|\varphi_t,\psi)= \exp(\psi(y_t\varphi_t-a(\varphi_t)))b(y_t,\psi)$ where φ_t, the *natural parameter*, satisfies

$$E(y_t|\varphi_t,\psi)= \mu_t= d/dt\ (a(\varphi_t))$$
$$Var(y_t|\varphi_t,\psi)= d^2/dt^2\ (a(\varphi_t))/\psi.$$

- *transition equation*= no explicit mechanism is assumed for the evolution of the *natural parameter* φ_t, i.e., $p(\varphi_t|\varphi_{t-1})$ is left unstated (see 1.1.2).

(i) *natural parameter prediction* - the non-existence of an explicit transition equation is circumvented by the specification of a proper mechanism describing the evolution of the natural parameter density in time. In this set up a prior distribution for φ_t (see 1.1.3), denoted

$CP(\alpha_{t/t-1},\beta_{t/t-1})$, is assumed at the outset. This is chosen to be the natural conjugate for the particular member of the exponential class taken as the measurement model. The stochastic mechanism responsible for the evolution of its two parameters is as yet unexplained.

(ii) *natural parameter updating*= given the choice of a closed sampling analysis, the posterior for φ_t, will have the general form $CP(\alpha_t,\beta_t)$ where

$$\alpha_t = \alpha_{t/t-1}+\psi y_t \text{ and } \beta_t = \beta_{t/t-1}+\psi.$$

(iii) *conditional distribution*= for a measurement equation selected from the exponential family, under a conjugate analysis, the resulting compounding integral (1.1.5 and 1.2.12a-b) will always produce an analytical distribution, symbolically,

$$p(y_t|Y_{t-1}) \sim P(\tilde{y}_{t/t-1}(\alpha_t),f_{t/t-1}(\beta_t))$$

where $\tilde{y}_{t/t-1}(\cdot)$ and $f_{t/t-1}(\cdot)$ are functions depending on the measurement model.

It now remains to be explained how they motivate the evolution of the parameters α_t and β_t. This is accomplished by a rather *ingenious* procedure. They introduce a GLIM-like link function, connecting the natural parameter φ_t with a latent process θ_t (the *state* in their notation), as yet unexplained. This relation is formally expressed by the equation $g(\varphi_t)\simeq h'\theta_t$, where h' is generally fixed ($h'=1$ for the random walk structure). Using a suitable specification for $g(\cdot)$ and by fixing its first two moments it is possible to solve numerically for α_t and β_t , and therefore, fully specify the prior-posterior analysis together with the predictive moments for y_t. It is obvious that for this to be achieved, a transition mechanism for the state itself should be motivated. They assume that the state follows the same transition equation of the Gaussian linear case (see 1.2.1b) (the state is the carrier of the structural components), but unlike in the linear case, its distribution form is left unspecified. They concentrate only on its first two moments, in order to avoid analytical intractability and also because of the limitations a complete specification would bring to the natural parameter prior choice. This explains why the link function equality should not be taken in a strict sense but rather as a *guiding relationship* to form the prior for φ_t. In the predictive step for the state the increase in uncertainty is achieved through the replacement of the state noise variance by a discount factor multiplying the previous value of the state variance, i.e., instead of (1.2.9b) now we have $p_{t/t-1}= (1/b)p_{t-1}$, where $0<b\leqslant 1$ is set by the user. Using a linear Bayesian approach they are able to derive a sub-optimal linear filter for the state, on the lines of the standard KF (see 1.2.11a-b). For the random walk structure this is given by

$$m_t= m_{t/t-1}+(g_t-m_{t/t-1})$$
$$p_t= v_t$$

with $g_t= E(g(\varphi_t)|Y_t)$ and $v_t= Var(g(\varphi_t)|Y_t)$.

This means that besides the predictive and updating steps for the natural parameter φ_t, we have extra expressions, of similar meaning, for the latent process θ_t. This completes the definition of the DGLM.

(iv) *k-steps ahead forecasting*- this is achieved by postulating the following form for the natural parameter forecasting distribution

$$p(\varphi_{T+k}|Y_T) \sim CP(\alpha_T(k),\beta_T(k))$$

which obviously doesn't follow from the solution of equation (1.1.7). These *projected* parameters are then solved by equating the moments of the above distribution with the projected moments of the latent process θ_t (see 1.2.20a-b) via the link function. Once these are obtained ,the k steps ahead forecasting distribution is postulated as having the same form as the predictive density with the *projected parameters* $\alpha_T(k)$ and $\beta_T(k)$.

We feel some concern about the validity of some of the assumptions and approximations used by the authors in their DGLM. In particular

- The state θ_t is not fully characterized, but only represented in terms of its two moments. A proper Bayesian analysis would require the full distribution.
- The filter for this process is linearly approximated.
- The k-steps ahead forecasting distributions are inexact, and have been forced in order to obtain a tractable expression for forecasting. Altough this may be plausible, the authors have not made their assumptions suficiently explicit in this important step.

For a more extensive critique of this methodology the reader is referred to the discussion paper following the article by WHM (1985) and Figliuoli (1989, ch.7).

1.3.2.2 Figliuoli's filter (*)

We now briefly review a class of *analytical filters* developed by Figliuoli (1988) in his PhD thesis, and one which the author was involved with at the begining of his research. Figliuoli's approach is developed on the lines of classical control theory, this meaning, amongst other things, that both the measurement and transition equation are given in external noise form, and that for the derivation of his sub-optimal non-linear filter, explicit optimality statements are considered. To illustrate his general methodology we have chosen to briefly describe his Poisson-lognormal model for counting data. The analysis follows in the lines of our probabilistic approach.

- *measurement equation*:the *implicit* representation of the DGP takes the form

$$p(y_t | \lambda(\theta_t)) \sim \mathrm{Poi}(\lambda(\theta_t))$$

where the intensity $\lambda(\cdot)$ is related to the unobserved state θ_t via a suitably chosen link function. Since, by assumption, the state follows a Gauss-Markov process (see 1.2.1.b) this function must ensure that $\lambda(\cdot) > 0$. In particular, the use of exponential and polynomial functions, under the Gaussian approximation to be introduced, may be shown to produce closed formula for the conditional moments (see Figliuoili 1988). For an exponential link function, it is easy to show that an equivalent way of expressing the measurement equation is given by

$$y_t = \lambda(\theta_t) + \epsilon_t$$
$$\lambda(\theta_t) = \exp(-h\theta_t)$$

where ϵ_t follows a non-centered Poisson distribution, with $E(\epsilon_t)=0$ and $E(\epsilon_t^2) = (\lambda(\theta_t))$ and h is a scalar with fixed value. Observe that now the measurement equation is non-linear on the state and also has a noise whose variance is state dependent.

- *transition equation*: has the same form adopted for the Gaussian case, i.e., $p(\theta_t | \theta_{t-1}) \sim N(\theta_{t-1}, \sigma^2)$.

(i) *state prediction*: in evaluating the predictive step (see 1.1.3 and 1.2.6) the resulting distribution will very easily become intractable. The fact is that, although the transition density $p(\theta_t | \theta_{t-1})$ is the same as in the Gaussian case, after the first interaction the posterior will lose conjugacy (since the measurement distribution is no longer Gausssian) and this will make the integral in (1.2.6) not analytical. In order to avoid the need for numerical integration, Figliuoli has assumed a Gaussian approximation for the posterior, i.e.,

$$p(\theta_t | Y_t) \sim N(m_t, p_t)$$

where m_t and p_t are as yet unexplained.

The above approximation is all one needs to guarantee that the predictive step will follow the same equations of the Gaussian linear case (see 1.2.9a-b). As pointed out earlier, for certain forms of link functions, the Gaussian approximation produces closed form conditional moments. In the control literature this is known as a *global filter*, given that no truncations of the non-linearities are required in order to implement the filter.

(ii) *state updating* - the fact that his model has not been set up in terms of a closed sampling analysis makes the establishment of the updating equations a rather elaborate issue. All one knows is that the posterior form is approximated by a Gaussian density, but the updating mechanism for its two moments is left to be determined. This asks for the introduction of some structure in order to properly frame the problem. Figliuoli assumes that the optimal algorithm describing the evolution of the posterior mean is given by the following difference equation, which is non-linear on the previous filtered state and linear on the observations:

$$m_t = f(m_{t-1/t-1}, t-1) + K_t e_t$$

where e_t is the one step-ahead prediction error, $f(\cdot)$ is a function to be determined and K_t is the filter gain. The two last quantities are determined by imposing the following criteria in the above estimator:

- optimality: m_t should be chosen such that it is a minimum variance estimator,i.e., min $\mathrm{tr}\{E(\theta_t - m_t)^2\}$.
- unbiasedness: m_t also should satisfy $E(m_t) = \theta_t$.

Using the above requirements, one can demonstrate that the mean and variance of the updating step assume the form:

$$m_t = m_{t/t-1} + K_t\{ y_t - E_{t-1}(\lambda(\theta_t)\}$$

$$p_t = p_{t/t-1} - K_t E_{t-1}\{[\lambda(\theta_t) - E_{t-1}(\lambda(\theta_t)].(\hat{\theta}_{t/t-1})\}$$, and the gain

$$K_t = E_{t-1}[\hat{\theta}_{t/t-1}[\lambda(\theta_t) - E_{t-1}(\lambda(\theta_t)].2E_{t-1}[\lambda(\theta_t) - E_{t-1}(\lambda^2(\theta_t))]$$

where $\hat{\theta}_{t/t-1} = \theta_t - \tilde{\theta}_{t/t-1}$, $\lambda(\theta_t) = e^{-h\theta}{}_t$ is the exponential link function and $E_{t-1} \equiv E(\ |Y_{t-1})$. Using the Gaussian approximation, the above formulae may be shown to take the following final form:

$$m_t = m_{t/t-1} + K_t(y_t - \tilde{\lambda}_{t/t-1})$$

$$p_t = p_{t/t-1} + K_t.h.p_{t/t-1}.\tilde{\lambda}_{t/t-1}$$

$$K_t = -p_{t/t-1}.h.\tilde{\lambda}_{t/t-1}/(s_t + \tilde{\lambda}_{t/t-1})$$

where $\tilde{\lambda}_{t/t-1} = E_{t-1}(e^{-h\theta}{}_t) = \exp(-hm_{t/t-1} + (1/2)h^2\ p_{t/t-1}\ )$ and

$s_t = \exp(-2h(m_{t/t-1} + h.p_{t/t-1})) - \exp(-h(m_{t/t-1} + h.p_{t/t-1}))$.

Observe that now the mean updating equation is coupled with the variance equation.

(iii) *conditional distribution*= given the specifications of this model one can show that the associated compounding operation is given by (see 1.1.5)

$$p(y_t|Y_{t-1}) = \mathrm{Poi}(\lambda_t(\theta_t)) \underset{\lambda_t(\theta_t))}{\triangle} \mathrm{logNormal}(-hm_{t/t-1}, h^2 p_{t/t-1})$$

The resulting compound distribution is known in the literature as the discrete lognormal (see,e.g., Johnson and Kotz 1969, ch.8) and may be shown not to possess an analytical form. However it is possible to obtain expressions for its two first moments, which are given below

$$\tilde{y}_{t/t-1}= \exp[-h(m_{t/t-1}+(1/2)hp_{t/t-1})]$$

$$f_{t/t-1}= \exp[2h(-m_{t/t-1}+h\, p_{t/t-1})] + \exp[h(-m_{t/t-1}+(1/2)hp_{t/t-1})]-$$
$$- \exp[h(-2m_{t/t-1}+hp_{t/t-1})]$$

where $m_{t/t-1}$ and $p_{t/t-1}$ are as in (1.2.9a-b).

(iv) *k-steps ahead predictive distribution*= as a by-product of the Gaussian approximation it is easy to see that the multi-step predictive distribution for the intensity will have the same form of its one step ahead distribution, i.e.,

$$(\lambda_{T+k}(\theta_{T+k})\,|\,Y_T)\sim \text{lognormal}(-hm_{T+k/T},h^2\, p_{T+k/T})$$

where $m_{t+k/T}$ and $p_{T+K/T}$ are as in (1.2.21a-b). From this it is not difficult to establish that the k-steps ahead compound distribution will also be discrete lognormal with its first two moments given by:

$$\tilde{y}_{T+k/T} = \exp[h(-m_{T+k/T} + (1/2)h.p_{T+k/T})$$
$$=\exp\{h[-m_T+(1/2)h(p_T+k\sigma^2)]\}$$

$$f_{T+K/T}= \exp[2h(-m_{T+k/T} + hp_{T+k/T}] + \exp[h(-m_{T+K/T}+ (1/2)hp_{T+k/T}] -$$
$$-\exp[h(-2m_{T+k/T}+ hp_{T+k/T})].$$

We would like here to produce some comments on Figliuoli's approach:

- It seems that there exists an inherent difficulty in his approach with regard to the *structural* interpretation of the state. In the Gaussian formulation the state is linearly associated with the mean of the process, so that whatever physical interpretation is given to it, this is easily assimilated in terms of the observed features of the process, and vice-versa. For a non-linear measurement equation this relation is rather obscured. In the above procedure the mean or intensity is not the state in itself, but is linked to it via the exponential function, and as a result the *structural* interpretation will not be as clear as in the Gaussian case. This is easily seen by considering a random walk structure for the state in the Poisson formulation. In the Gaussian case the forecasting function will be a horizontal straight line with value m_T (see 1.2.20a), while for the Figliuoli-Poisson formulation, as the first of the above equations show, the logarithm of this function will grow linearly with the state noise variance.

- I have carried out some preliminary investigation which has shown that, due to the existence of several exponential functions on his filter equations, an ad hoc prefixing of the scalar h at an untypical value (i.e., different from 1) is necessary in order to avoid floating point overflow on the VAX computer. This would further complicate, the already problematic *structural* interpretation of the model, since at the end we would be considering non-integer values of the *structural* components describing the movements on the series!.

- In the case of non-analytical predictive distributions we will be faced with potential problems for hyperparameter estimation. If the chosen criterion function is the likelihood, this would require reliable and easy to implement approximations to this distribution, which in itself is already an approximation to the 'true' likelihood function. For a minimum distance type estimator, doubt still remains about the existence of firm asymptotic results.

In short, although this seems to provide an elegant and statistically sound class of non-linear/non-Gaussian state space models, we feel that further work is still needed in order to overcome these potential drawbacks, and to investigate the empirical usefulness of the proposed methodology.

CHAPTER TWO

NON-GAUSSIAN STRUCTURAL MODELS

2.1 INTRODUCTION:

The objective of our class of models is to provide a state space forecasting technique for certain types of non-stationary/non-Gaussian data, with particular emphasis on count data. The data structures dealt with in our study are listed below.

(i) count data- as in the Poisson-Gamma (Ch.3), Negative Binomial-Beta (Ch.3), Binomial-Beta (Ch.4) and the Bivariate Poisson (Ch.5) models.
(ii) binary data - as in the Bernoulli-Beta model (Ch.4).
(iii) positive continous data - as in the Gamma-Gamma model (Ch.6).
(iv) random sums - as in the Random Sum model (Ch.7).

The structural models for non-Gaussian data are defined in terms of two equations like the *conventional* state space models: these are the measurement equation and the transition equation (see 1.1.1-2). The measurement equation is *internally* defined, i.e., it is not written in external noise form as in the *conventional* state space models and econometric equations. WHM (1985) and S&M (1986) have used a similar approach. Regarding the transition equation (1.1.2), here we also take a different view from the *conventional* models in that for most of the models considered this equation is *implicitly* defined. In the present context an *implicitly* defined state equation means that the state evolution is only set in terms of an updating mechanism which transforms its conditional densities. For forecasting purposes this will suffice to produce operational models. See,e.g., S&M (1986, p.83) and Smith (1988a-b). Since our modelling strategy is guided by the *structural* paradigm, the class of models we produce is formulated in terms of components of interest, such as level, slope and seasonals. Explanatory variables and dummy variables may also be introduced. Using our system of classification for state space models, we could say that our procedure belongs to the class of *analytical filters*, albeit for certain distributional assumptions, k-steps ahead moments of order higher than one and/or distributions may need simulations in order to be computed.

In what follows we present a more formal treatment of our modelling strategy. This does not intend to be complete, since many features in our approach (e.g., the introduction of explanatory variables/structural components, the linear character of the predictor, etc) may be only fully understood with reference to explicit distributional assumptions. As before, we make use of the probabilistic approach in state space models (see section 1.1).

2.2 AN OVERVIEW OF THE NON-GAUSSIAN STRUCTURAL MODELS

- *measurement equation*: we have chosen to work with the standard text-book parameterization of the exponential family, since, amongst other things, in this form the state, or stochastic parameter, has an easily identified conjugated prior (see,e.g., Aitchison and Dunsmore 1975). This makes the formalism simpler to be worked with, if compared with the alternative *canonical* parameterization (see,e.g., Stuart and Ord 1987, vol.1 ch.5, and WHM 1985) or *mean-value* parameterization (see,e.eg., Dalal and Hall 1983). This particular choice of setting up the problem, although operationally convenient, makes the establishment of general results less unlikely. Our measurement equation can be symbolically represented by an internal equation as in (1.1.1) and (1.2.3)

$$p(y_t | \theta_t, \upsilon_t) = \exp\{y_t A(\theta_t) + C(y_t, \upsilon_t) + D(\theta_t)\} \qquad (2.2.1)$$

with $\theta_t \in \Theta \subset \mathbb{R}$ and $\upsilon_t \in \mathbb{U} \subset \mathbb{R}$. Here A,C and D are arbitrary functions of their arguments. The mean and variance are given respectively by

$$E(y_t | \theta_t, \upsilon_t) = \mu_t = f(\theta_t, \upsilon_t) \qquad (2.2.2a)$$

$$Var(y_t | \theta_t, \upsilon_t) = g(\theta_t, \upsilon_t) \qquad (2.2.2b)$$

where $f(\cdot)$ and $g(\cdot)$ are some specific functions. Note that the secondary parameter υ_t can be either static or exogenously dynamic.

A secondary parameter is purposely made dynamic in order to introduce explanatory variables and/or structural components, either because the state does not offer the appropriate setting (as in the Negative-Binomial model, section 3.3.2) or because it gives a second option for this mechanism along with the state (as in the Gamma-Gamma model, section 6.2.2). This will be made clearer in the forthcoming chapters.

If we look back at (2.2.1), it is easy to notice that an implicit independence statement has been used in writing this equation. In fact, in analogy to its Gaussian counterpart (see 1.2.2), we have required that conditional on the state θ_t, the present observation y_t is independent from the past of the process Y_{t-1}. By doing so we will be inducing some desirable characteristics in the stochastic parameter θ_t ,e.g., the role of a sufficient statistic for forecasting purposes. See, e.g., Smith (1988a-b).

- *transition equation*: as stated previously, for most of the specifications considered, this is not given in external form, but rather as an *implicit* mechanism which is defined in the next item. Note however, that the existence of such an equation is, in principle, desirable, but not always easy to obtain. Bather (1965) showed that only for measurement models drawn from the exponential family is such a specification feasible. A step further in the generalization of this class of models was given by J.Q. Smith in his 1979 paper. He acknowledged that, in order to obtain state space models for forecasting purposes, the transition equation (see 1.1.2) could be left unstated, the crucial point being the establishment of the updating rule transforming the conditional densities of the state (see 1.1.3), and this could be motivated by analogy with the Gaussian random-walk when it reaches the steady state. Using Bayesian decision theory, Smith developed arguments in this direction, culminating in his *Power Steady* model (Smith 1979) which adopts the following transformation

$$p(\theta_t|Y_{t-1}) \propto [p(\theta_{t-1}|Y_{t-1})]^{\omega}$$

with $0 < \omega \leqslant 1$. It may be shown that such a transition rule will keep the *mode* of the densities unchanged, while increasing the expected loss (see Smith 1979). As we shall see our model strategy is constructed on the lines of Smith's design and therefore may also be classified as belonging to the class of *partially specified* models.

(i) *state prediction-* since in our formalism we lack the equivalent of a Chapman-Kolmogorov equation (see 1.2.6), the state prediction step may be arrived at by the adoption of a transformation rule between $p_1(\theta_{t-1}|Y_{t-1})$, the posterior at time t-1 and $p_2(\theta_t|Y_{t-1})$, the prior at time t. This has to be done in such a way that the framework is kept analytically tractable and the adopted rule has a meaningful interpretation in terms of the *structural* paradigm. By construction, this rule should induce, in a consistent way, a random walk evolution on the mean of the process, μ_t (2.2.2a). Observe that this desired effect has to be imposed on the mean of the measurement equation, rather than directly on the state itself. They only coincide when the function $f(\cdot)$ in (2.2.2a) is the identity, as in the Poisson specification, but for most of the cases considered $f(\cdot)$ will be a trivial function of θ_t, so that mathematical manageability will follow. The natural guideline in the construction of this transition rule is given by the original equations for the random walk evolution under the Gaussian-linear assumption (see 1.2.9a-b). In this set up, the distribution at time t-1 for the state (which is also the mean of the process) is $\mu_{t-1}/Y_{t-1} \sim N(m_{t-1}, p_{t-1})$. The predictive distribution for μ_t is also normal and is given by $N(m_{t-1},\sigma^2(q+p_{t-1}))$. The essential features in this transition are :

(i) the distributional form of the state is kept invariant during the transition.
(ii) the mean of μ_t/Y_{t-1} is the same as that of μ_{t-1}/Y_{t-1} but the variance increases.

This same effect can be induced on the non-Gaussian set up by imposing the following requirements:

1. Both $p_1(\theta_{t-1}|Y_{t-1})$ and $p_2(\theta_t|Y_{t-1})$ belong to the same class of distributions, in particular to the class of the natural conjugate distributions to the chosen measurement equation. We can therefore adopt the following general representation for these distributions:

$$p_1(\theta_{t-1}|Y_{t-1}) = p(a_{t-1},b_{t-1}) \qquad (2.2.3)$$

$$p_2(\theta_t|Y_{t-1}) = p(a_{t/t-1},b_{t/t-1}) \qquad (2.2.4)$$

where $p(\cdot)$ is the natural conjugate distribution and a_{t-1} and b_{t-1} are computed from the first t-1 observations. This structure guarantees a closed sampling analysis and all the simplicity associated with it. Note that the members of the exponential family considered in our study have either the gamma or beta densities as natural conjugates.

2. The mean and variance of the process level evolves according to the following rules:

$$E(\mu_t | Y_{t-1}) = E(\mu_{t-1} | Y_{t-1}) \qquad (2.2.5a)$$

$$Var(\mu_t | Y_{t-1}) > Var(\mu_{t-1} | Y_{t-1}). \qquad (2.2.5b)$$

Since $\mu_t = f(\theta_t, \upsilon_t)$ (see 2.2.2a) it is not difficult to see that the implementation of the condition in (2.2.5a) will result in a deterministic link between the parameters of the state densities in (2.2.3) and (2.2.4). For the second condition to be obeyed, in general, extra relations will have to be found involving ω and υ. With the benefit of hindsight the *prediction equations* in (2.2.5a-b) may be written in a general form as

$$a_{t/t-1} = \omega\, a_{t-1} + i(1 - \omega) \qquad (2.2.6a)$$

$$b_{t/t-1} = \omega\, b_{t-1} \qquad (2.2.6b)$$

where i=0 for the Poisson-Gamma and Binomial-Beta models and i=1 otherwise.

Apparently our prediction equations for the Poisson-Gamma and Binomial-Beta models are the same as those produced in the examples of Smith (1979). This should not be the case since in his formulation it is the mode of the state density that is kept constant during the transition. The explanation for this apparent contradiction lies in the fact that in his examples Smith uses a different parametrization for the state densities, namely a gamma ($a_{t/t-1}+1$, $b_{t/t-1}$) and a beta($a_{t/t-1}+1$, $b_{t/t-1}+1$) instead of the parameterizations which we have adopted.

Given that our modelling strategy is developed in terms of a desirable forecasting behaviour for the mean of the process, it is more appropriate to look at it as describing the patterns of the *data forecasting mechanism* (hencefort DFM) rather than attempting to represent the components of interest of the actual process (as in the Gaussian *structural* approach) or approximating observations (as in the ARIMA models).

(ii) *state updating*- once more the closed sampling analysis comes to our rescue. The filtering or updating distribution, by construction, will have the same form of the state predictive distribution (2.2.3). As in the Gaussian case, the filtering equations are standard results, and are given by a set of non-coupled equations linear in the predictive parameters $a_{t/t-1}$ and $b_{t/t-1}$, and on the newly observed value of the process, y_t. These may be conveniently expressed as

$$a_t = a_{t/t-1} + x_t \qquad (2.2.7a)$$

$$b_t = b_{t/t-1} + z_t \qquad (2.2.7b)$$

where $a_{t/t-1}$ and $b_{t/t-1}$ are as in (2.2.6a) and (2.2.6b), respectively.

For the Poisson-Gamma model: $x_t=y_t$, $z_t=1$;

Negative Binomial-Beta model: $x_t=\upsilon_t$, $z_t=y_t$;

Binomial-Beta model: $x_t=y_t$, $z_t= n_t-y_t$;

Gamma-Gamma model: $x_t=\upsilon_t$, $z_t=y_t$.

(iii)*conditional distribution*= given the fact that we are working under closed sampling in the exponential family, all the predictive distributions have an analytical form, and may be expressed symbolically as

$$p(y_t|Y_{t-1}) \sim P(a_{t/t-1},b_{t/t-1}) \qquad (2.2.8)$$

where $a_{t/t-1}$ and $b_{t/t-1}$ are as in (2.2.6a) and (2.2.6b), respectively. In our models we adopt the conditional mean of the predictive distribution as our measure of location. It is well known that this is the optimal forecast under the quadratic loss function and that for different loss specifications, other measures of location could be obtained,e.g., a step loss gives the mode. Note however that such a measure is not suitable for discrete distributions. To obtain the mean of the predictive distribution one makes use of the following relation

$$\begin{aligned}\tilde{y}_{t/t-1}=E(y_t|Y_{t-1})&= E\ (E(y_t|\theta_t,\upsilon_t)|Y_{t-1})\\ &= E\ (f(\theta_t)|Y_{t-1}) \qquad (2.2.9a)\end{aligned}$$

where we have used (2.2.2a) and dropped υ_t for ease of notation. As we shall see for all our models the asymptotic form of this predictor is given by an *exponentially weighted moving average* (EWMA) scheme.

The predictive variance may also be obtained by a similar formula which is given by

$$\begin{aligned}\text{Var } \tilde{y}_{t/t-1} &= E\ (\text{Var}(y_t|\theta_t)|Y_{t-1}) + \text{Var}\ (E(y_t|\theta_t)|Y_{t-1}) \\ &= E\ (g(\theta_t)|Y_{t-1}) + \text{Var}\ (f(\theta_t)|Y_{t-1}). \qquad (2.2.9b)\end{aligned}$$

(iv) *k-steps ahead forecasting*- as we shall see, the choice of having the mean of the level kept invariant during the transition will produce a constant forecast function, which, asymptoticaly has an EWMA form. This will be a general characteristic of all our models . When structural components and explanatory variables are introduced it is also possible to show that the forecasts will be a combination of EWMA schemes. Not surprisingly the form of the forecast function in Smith's framework will be different from ours since in his models it is the mode of the state density that is kept constant. In this sense Smith's models cannot be considered direct generalizations of the Gaussian steady model, where the projected mean is constant. Quoting Key and Goldophin (1980, p.93) : '... Smith's proposal of steady evolution of the system parameter is not equivalent to the concept of a steady forecasting model which necessarily pursues a constant forecast function'.

Analytical expressions for k-steps ahead moments of order higher than two are not easy to obtain, with exception of the Gamma-Gamma model. One has, therefore, to resort to non-analytical methods for the computation of these quantities if needed. The full forecasting distribution could be evaluated, in principle, by use of equations (1.1.6-8), where integrals should be replaced by sums whenever discrete data models are considered. Unfortunately it is also true that the integrals or sums involved in these evaluations will, in general, become difficult to manipulate producing further intractability. Contrary to the Gaussian case the form of the forecasting distribution will change with time in an unpredictable way. A number of strategies can be used in order to tackle these problems, and we, following S&M (1986), advocate the use of Monte Carlo simulations whenever probability forecasts and high order moments are needed for more than one step ahead. We are still not convinced about the feasibility of computationally intensive numerical methods described in section 1.3.1. to solve these problems. Approximation schemes for the forecasting distribution of a more *ad hoc* nature are adopted by WHM (1985), but this needs careful investigation. In Chapter 3 we provide details of how to implement such techniques in our framework.

2.2.1 Explanatory Variables and Structural Components

In a non-Gaussian structural model explanatory variables are introduced *via* the kind of link functions used in the GLIM framework of McCullagh and Nelder (1983). Bringing stochastic slope and seasonal effects into our class of models is not an easy task. They can, however enter in a deterministic fashion and this is done by treating them as explanatory variables. We believe this is not a serious restriction. Even with data on continuous variables, it is not unusual to find that the slope and seasonal effects are close to being deterministic; see, for example, Harvey and Durbin (1986) and Harrison (1988). In particular with count and qualitative data it seems even less likely that the observations will provide enough information to pick up changes in the slope and seasonal effects over time. In what follows we briefly discuss the way this mechanism is brought in our models letting the more technical details be presented when the specific distributional assumptions are introduced.

In our framework the effects of explanatory variables, trend and seasonal variations are brought together into a single component, which is known in the GLIM literature as the *systematic component*, and this is given by

$$\eta_t = z_t{}'\delta \qquad , \quad -\infty < \eta_t < \infty \qquad (2.2.10)$$

where $z_t = (R_t, T_t, S_t)$ is a pxl vector, with R_t being the usual vector of regressors,i.e., $R_t = \beta_t{}'x_t$, T_t the trend, S_t the seasonal component, to be defined latter and δ is a pxl vector of unknown parameters to be estimated using ML. The way to relate the *systematic* *component* with the level of the process is by adopting a suitable function which relates the mean of the measurement equation (2.2.2a) to the above component. Formally we have

$$\begin{aligned} \mu_t^{+} = f(\theta_t^{+}, \upsilon_t^{+}) &= h(\mu_t, \eta_t) \\ &= h[f(\theta_t, \upsilon_t), \eta_t] \qquad (2.2.11) \end{aligned}$$

where the symbol + denotes the presence of explanatory variables/ structural components and h(·) is the *link function* or the *inverse link function* in the GLIM notation. Observe that it may be the case that the secondary parameter offers the appropriate setting to introduce these effects as in the Negative Binomial- Beta (NBD) model of section 3.2 and the Gamma-Gamma model of Chapter 6. In selecting an appropriate link function the important factor is to produce a consistent mapping from the real line $(-\infty,\infty)$ (the space of the *systematic component*) onto the state space Θ or the secondary parameter space $\mathbf{U}$. For example the rate of a Poisson model is always positive while the proportions on a multinomial model take values between zero and one and should add up to one. The link function has to be consistent with these constraints. It is also important for the chosen transformation not to destroy the conjugacy property of the model. With the exception of the Binomial-Beta model we have been able to avoid such undesirable effect. Regardless of the mechanism selected to introduce these effects in our models, they share some common characteristics which are worth stressing at the present stage:

(i) as we shall see, the *systematic component* η_t always enters into the link function *via* an exponential function (see section 3.3.2 and eqs. 3.2.25,, 4.2.10b and 6.2.30), i.e.

$$h[f(\theta_t,\upsilon_t),\eta_t]= h[f(\theta_t,\upsilon_t),\exp(\eta_t)]$$

(ii) in view of (2.2.5a), the prediction and updating equations will have to be duly modified to take account of the presence of η_t.

We leave a more formal treatment of (ii) to the forthcoming chapters and concentrate now on the definitions of the structural components which are handled in our models.

Structural Components:

Time trend: a slope is introduced by setting one of the elements of x_t equal to time, t, so that the time component takes the form

$$T_t= \delta t \qquad , t=1,2,...,T. \qquad (2.2.12)$$

Seasonals: the seasonal component is modelled by s-1 explanatory variables constructed so that the seasonal effects sum to zero over the period s in question. In our framework this can be done by two different ways. The first option is to consider the seasonals as dummy variables and this has the form

$$S_t= \gamma_j \; z_{j,t} \qquad (2.2.13)$$

with j=0,1,2,...,s-1, where s is the seasonal period and the dummy

$z_{j,t}$ is such that

$$z_{j,t}= \begin{cases} 1 & j= \mathrm{mod}(t,s) \\ 0 & \text{otherwise.} \end{cases}$$

with the constraint that the seasonal coefficients γ_j sum to zero over the seasonal period s. One may also model seasonality by a set of trigonometric terms at the seasonal frequencies, $\lambda_j= 2\pi j/s$, $j=1,2,\ldots,[s/2]$, where

$$\begin{aligned} [s/2]= &\ s/2 && \text{for s even} \\ &(s-1)/2 && \text{for s odd.} \end{aligned}$$

The seasonal effect at time t is then given by (see,e.g.,Harvey 1989, ch.1)

$$S_t= \sum_{j=1}^{[s/2]} [a_j \cos(\lambda_j t)+ b_j \sin(\lambda_j t)] \qquad (2.2.14)$$

where a_j and b_j are estimated by the likelihood method. Note that if the full set of dummies and seasonals are included then the two formulations produce the same number of coefficients.

CHAPTER THREE

UNIVARIATE COUNT DATA MODELS

3.1 INTRODUCTION:

It is not unusual to find time series consisting of *count data*. Such series record the number of events occurring in a given interval and are necessarily non-negative integers. For instance, the monthly number of homicides in Canada, the monthly number of U.S. cases of poliomyelitis (Zeger 1988), the number of goals scored by England against Scotland in international football matches (Harvey and Fernandes 1989a) and so on. Count data models are usually based on distributions such as the Poisson or negative binomial (NBD). If the means of these distributions are constant, or can be modelled in terms of exogoneous variables, then the GLM framework of McCullagh and Nelder (1983) offer the appropriate set up. For the Poisson specification the assumed link function is the *log-linear* link (see section 2.2.1) which is written as $\ln \mu_t = x_t{}'\delta$, where δ is a pxl vector of unknown coefficients, estimated by weighted least-squares. Dispersion relative to the Poisson model may be obtained by specifying

$\mathrm{var}(y_t) = \sigma\ \mu_t$, where $0 < \sigma < \infty$, is assumed constant. If further flexibility is desired to tackle overdispersion then Cameron and Trivedi (1986) and Lawless (1987b) provide the appropriate setting by considering NBD regression models. Gourieroux, Monfort and Trognon (1984) discuss pseudo maximum likelihood methods for these specifications.

For situations where counting data shows serial correlation a number of techniques is available. Zeger (1988) develops an extension of log-linear models where the mean of the process, μ_t is assumed to depend on an unobservable noise process ϵ_t. This, by assumption is taken as a second order stationary AR process. His modelling is then developed by specifying the first two moments of the observed process y_t, which has the form $\mu_t = \exp(x_t{}'\delta)\epsilon_t$, where $E(\epsilon_t) = 1$ and $\mathrm{cov}(\epsilon_t, \epsilon_{t+\tau}) = \sigma^2\rho_\epsilon(\tau)\cdot$ It is then easy to see that the process y_t inherits both overdispersion and autocorrelation from ϵ_t. Estimation of the regression parameters in Zeger's model is accomplished by extending quasilikelihood techniques to dependent data, while for the nuisance parameters a method of moments is proposed.

The nature of count data makes standard ARIMA models inappropriate both for fitting real data and generating synthetic observations. Only when the values of the observations are large enough to justify the assumption of normally distributed disturbance, may ARIMA models be used as a reasonable approximation. A discrete version of these models has been developed by McKenzie (1985,1986) which considers, among other specifications, Poisson and NBD models with linear correlation structure. Since no inferential technique has yet been developed for those models, its use is mainly restricted to the generation of synthetic data.

In considering state space models for count data, we refer the reader to Chapter 1 where Figliuoli's Poisson model is discussed with some detail. Other references are the Poisson models of WHM (1985) and the state dependent observation variance of Zehnwirth (1988), whose filter may be shown to be a particular case of Figliuoli's filter.

In what follows we will present our class of *structural models* for univariate count data, which is based on the Poisson and NBD distributions. The development is analogous to the Gaussian random walk plus noise model (see section 1.2) in that they allow the underlying mean of the process to change over time. By introducing a hyperparameter, ω, into these local level models, past observations are discounted in making forecasts of future observations. Indeed it transpires that in all cases the predictions for all future periods can be constructed by an EWMA procedure. This is exactly what happens in (1.2.1a-b) under the normality assumption. Muth (1960) showed that such a predictor is optimal (in the MSE sense) for an ARIMA(0,1,1) process, $(1-L)y_t = (1+\omega L)\epsilon_t$.

3.2 THE POISSON-GAMMA MODEL:

The discussion of this and the subsequent models will follow the general archetype for state space models introduced in Chapter 1, section 1.1.

- *measurement equation*: suppose that the observation at time t is drawn from a Poisson distribution,

$$p(y_t | \theta_t) = \theta_t^{y_t} \, \bar{e}^{\theta_t} / y_t! \qquad , \quad y_t = 0,1,2,\ldots \qquad (3.2.1)$$

where $\theta_t > 0$. This is a particular case of (1.1.1) and corresponds to the measurement equation in (1.2.1a). The mean and variance for (3.2.1) are given respectively by:

$$E(y_t | \theta_t) = \mu_t = \theta_t$$
$$Var(y_t | \theta_t) = \mu_t.$$

It then follows that for this specification both f(·) and g(·) in (2.2.2a-b) are equal to the identity function and that the secondary parameter $\upsilon_t \equiv 1$.

(i) *state prediction*- the conjugate prior for a Poisson distribution is the gamma distribution. Let $p(\theta_{t-1} | Y_{t-1})$ denote the p.d.f. of θ_{t-1} conditional on the information at time t-1. Suppose that this distribution is gamma, that it is given by

$$p(\theta_{t-1}|Y_{t-1}) = \frac{e^{-b\,\theta_{t-1}}\,\theta_{t-1}^{a-1}\,b^{a}}{\Gamma(a_{t-1})} \qquad (3.2.2)$$

with $a=a_{t-1} > 0$ and $b=b_{t-1} > 0$ and $\Gamma(\cdot)$ is the gamma function. It is standard that

$$\mathrm{Mode}(\theta)= (a-1)/b \qquad (3.2.3a)$$

$$E(\theta^{n})= \frac{\Gamma(a+ n)\ b^{n}}{\Gamma(a)} \qquad (3.2.3b)$$

so that

$$E(\theta_{t-1}|Y_{t-1})= a_{t-1} / b_{t-1} \qquad (3.2.4a)$$

$$\mathrm{Var}(\theta_{t-1}|Y_{t-1})= a_{t-1} / b_{t-1}^{2} \qquad (3.2.4b)$$

where a_{t-1} and b_{t-1} are computed from the first t-1 observations, Y_{t-1}. In order to satisfy conditions (2.2.5a-b) which guarantee the induction of a *random walk* evolution on the mean, the state predictive density $p(\theta_t|Y_{t-1})$, which is also gamma by construction, should have its parameters $a_{t/t-1}$ and $b_{t/t-1}$ linked to the posterior parameters through the following deterministic equations (*the prediction equations*)

$$a_{t/t-1} = \omega\, a_{t-1} \qquad (3.2.5a)$$

$$b_{t/t-1} = \omega\, b_{t-1} \qquad (3.2.5b)$$

with $0 < \omega \leq 1$. It then follows that

$$
\begin{aligned}
E(\theta_t | Y_{t-1}) &= a_{t/t-1} \,/\, b_{t/t-1} \\
&= a_{t-1} \,/\, b_{t-1} = E\,(\theta_{t-1} | Y_{t-1})
\end{aligned}
\tag{3.2.6a}
$$

while

$$
\begin{aligned}
\mathrm{Var}(\theta_t | Y_{t-1}) &= a_{t/t-1} \,/\, b_{t/t-1}^{2} \\
&= \omega^{-1}\, \mathrm{Var}\,(\theta_{t-1} | Y_{t-1})
\end{aligned}
\tag{3.2.6b}
$$

so that the aforementioned conditions are satisfied. The stochastic mechanism governing the transition of θ_{t-1} to θ_t is therefore defined *implicitly* rather than explicitly. However, using known properties of gamma and beta variates, it is possible to show that it is formally equivalent to a *multiplicative equation*, originally developed by S&M (1986) in their exponential-gamma model. For our Poisson-Gamma specification this equation takes the form

$$
\theta_t = \omega^{-1}\, \theta_{t-1}\, \eta_t \tag{3.2.7}
$$

where $\eta_t \sim$ beta $(\omega a_{t-1}, (1-\omega) a_{t-1})$. As recently demonstrated by Shephard (1990b), in the context of a Gaussian local scale model, where θ_t^{-1} is the precision at time t, such a transition equation may be problematic, since if $\omega < 1$, $\theta_t \rightarrow 0$ almost surely, as $t \rightarrow \infty$. The easiest way to understand this effect is by looking at the expression for the expected value for log θ_t/θ_{t-1} which has the form

$$
\begin{aligned}
E[\log(\theta_t/\theta_{t-1}) | Y_{t-1}] &= E[\log \eta_t | Y_{t-1}] - \log \omega \\
&\cong \frac{-\,(1-\omega)}{2\,\omega} \; \frac{1}{(a_{t-1}+1)} \quad , \; 0 < \omega < 1,
\end{aligned}
$$

where the approximation for the expectation on the rhs was obtained by a Taylor series expansion about the mean of η_t, ω. The growth rate of the state is therefore negative on average, the rate of decay being slower for values of ω close to unity and inversely proportional to the value of the gamma shape parameter at time t-1, a_{t-1}. In Shephard's model this parameter is data independent, so that, eventually, this will settle to a constant value. By redefining the transition equation (3.2.7) in an appropriate way Shephard is able to remove this problem. Note, however, that in our Poisson-Gamma model a_{t-1} depends on past observations (see 3.2.8a-9a), so that the rate of convergence of the Poisson parameter may be partially counterbalanced by the weighted sum of past observations. Monte Carlo experiments conducted in our models (see Ch.8) have shown that replications based on values of $\omega < 0.8$ combined with sample sizes larger than 100 will eventually produce such effect. Since values of interest for ω are usually greater than 0.8, this will not have much relevance in our setting. In fact simulated NBD series of 700 observations are commonly obtained for values of ω greater than 0.90 (see Ch.8). Note that S&M (1986) model is based on a linear drift so that this effect does not take place.

Explicit transition equations as given in (3.2.7) may also prove useful in the derivation of multi-step ahead moments. This has been used, e.g., by S&M (1986) in their exponential-gamma model for the uncensored case. Since in our case the updating equation for the gamma shape parameter involve the observables y_t (see 3.2.8a), this strategy will be of limited use, given that the projected raw moments of order superior to one will inevitably depend on future values of y_t. As we shall see, by use of the chain rule for conditional expectations given in (1.2.9a) we will be able to derive an expressions for the k-steps ahead variance.

(ii) *state updating-* once the observation y_t becomes available, Bayes theorem is used to update knowledge about the state. It is standard that the posterior distribution, $p(\theta_t|Y_t)$, is given by a gamma distribution with parameters (*the updating equations*)

$$a_t = a_{t/t-1} + y_t \qquad (3.2.8a)$$

$$b_t = b_{t/t-1} + 1. \qquad (3.2.8b).$$

These equations together with (3.2.5a-b) complete the definition of our filter. Note that by repeated substitution from (3.2.5a-b) and (3.2.8a-b) we obtain

$$a_{t/t-1} = \sum_{j=1}^{t-1} \omega^j \, y_{t-j} \qquad (3.2.9a)$$

$$b_{t/t-1} = \sum_{j=1}^{t-1} \omega^j \, . \qquad (3.2.9b)$$

Note that the equivalent of a 'steady state' filter is obtainable for t sufficiently large when $b_{t/t-1}$ is approximately equal to $\omega/(1-\omega)$, $\omega < 1$. It is not difficult to see that convergence to the steady state solution will depend on the value of ω itself, being faster the closer ω is to its lower boundary value zero.

(iii) *conditional distribution* - predictive p.d.f.'s are given by the solution of the compounding operation (see 1.1.5 and 1.2.12a-b) and for Poisson observations and a gamma prior this yields a NBD distribution

$$p(y_t | Y_{t-1}) = \begin{bmatrix} a+ y_t - 1 \\ y_t \end{bmatrix} b^a \quad (1 + b)^{-(a + y_t)} \qquad (3.2.10)$$

where $a = a_{t/t-1}$ and $b = b_{t/t-1}$ and

$$\begin{bmatrix} a + y_t - 1 \\ y_t \end{bmatrix} = \frac{\Gamma(a + y_t)}{\Gamma(y_t + 1)\ \Gamma(a)}$$

Note that since y_t is an integer, $\Gamma(y_t + 1) = y_t!$.

It follows from the properties of the NBD that the mean and variance of the predictive distribution of y_t given Y_{t-1} are respectively

$$\tilde{y}_{t/t-1} = E(y_t | Y_{t-1}) = a_{t/t-1} / b_{t/t-1} = a_{t-1}/b_{t-1} \qquad (3.2.11a)$$

and

$$\begin{aligned} Var(y_t | Y_{t-1}) &= a_{t/t-1}\ (1+b_{t/t-1})/b^2_{t/t-1} \\ &= E(y_t | Y_{t-1})(1+\omega b_{t-1})/\omega b_{t-1} \qquad (3.2.11b) \end{aligned}$$

which shows overdispersion compared to the Poisson model. Substituting (3.2.9a-b) in the expression for the predictor in (3.2.11a) one finds that

$$\tilde{y}_{t/t-1} = \sum_{j=1}^{t-1} \omega^j\ y_{t-j} \Big/ \sum_{j=1}^{t-1} \omega^j\ . \qquad (3.2.12)$$

This is a weighted mean in which the weights decline exponentially. It has exactly the same form as the discounted least squares estimate of a mean. Using the large samples value for the denominator of (3.2.12), one may show that the forecasts can be obtained recursively by the EWMA scheme

$$\begin{aligned} \tilde{y}_{t/t-1} &= \lambda y_{t-1} + (1-\lambda)\, \tilde{y}_{t-1/t-2} \quad , \quad t = 2 ,\ldots, T \\ &= \mathrm{EWMA}(y) \qquad (3.2.13) \end{aligned}$$

where $y_{1/0} = 0$ and $\lambda = 1-\omega$ is the smoothing constant. When ω=1, the right hand side of (3.2.13), is equal to the sample mean. Regarding this as an estimate of μ, the choice of zeros as initial values for a and b in the filter is seen to be justified insofar as it yields the classical solution (see also section 3.2.1). It is also worth noting that, unlike the Gaussian case, no approximations are involved in the use of a diffuse prior in this model.

Finally it is worth investigating the existence of conditions under which our state transition rule becomes similar to that of Smith. From (3.2.3a-b) one can show that

$$[\ E(\theta_{t/t-1}) - \mathrm{Mode}(\theta_{t/t-1})] = (1 \,/\, b_{t/t-1}).$$

Substituting the asymptotic value of $b_{t/t-1}$ the above difference becomes approximately equal to $(1-\omega)/\omega$ so that when the estimated value of ω is close to one our updating rule and Smith's become very close.

(iv) *k-steps ahead forecasting* - for $k=1$ this is trivial. One has only to substitute t for T+1 in (3.2.10) to obtain the forecasting distribution with first two moments given in (3.2.11a-b). For $k > 2$ it is possible to work out analytical expressions for the first two moments of the forecasting distribution using the chain rule for conditional expectations given in (1.2.9b). One may show that these are given respectively by (see Appendix A1)

$$E(y_{T+k} | Y_T) = a_T / b_T \qquad (3.2.14a)$$

and

$$Var(y_{T+k} | Y_T) = (a_T / \omega b_T^2) \; (1 + \omega b_T + b_T \, S_{k-1} \, (1-\omega)) \qquad (3.2.14b)$$

where

$$S_{k-1} = \sum_{j=1}^{k-1} (1/b_{T+j}) \; , \; k > 2 \qquad (3.2.15)$$

with $S_0=0$. Observe that for T sufficiently large $b_T \simeq 1/(1-\omega)$ so that the k-steps ahead mean and variance become

$$E(y_{T+k} | Y_T) \simeq EWMA(y) = \mu_T \qquad (3.2.15a)$$

and

$$Var(y_{T+k} | Y_T) \simeq (\mu_T / \omega) [1 + (k-1)(1-\omega)^2] \qquad (3.2.16b)$$

which grows linearly with k the lead time.

The predictive distribution conditional on the observations up to time T, in theory, could be obtained through (1.1.6). The first distribution which appears in this integral is just the projection of the measurement equation k-steps ahead, and this is trivial to obtain. To compute the second density in (1.1.6), the state density projected k-steps ahead (1.1.7), one could use the explicit transition equation in (3.2.7), as it has been done on the Gaussian case. Unfortunately the successive applications of the compounding operation in (1.1.7), necessary to derive this density, will very easily produce analytical intractability (see Chapter 6, eqs 6.2.6-7). In the present context it is more appropriate to use (1.1.8) instead, where the integrals are replaced by summations producing the following expression

$$p(y_{T+k}|Y_T) = \sum_{y_{T+k-1}} \cdots\cdots \sum_{y_{T+1}} \prod_{j=1}^{k} p(y_{T+j}|Y_{T+j-1}) \qquad (3.2.17)$$

where the summations are to be evaluated from 0 to ∞. As we shall see it is difficult to derive a closed form expression for $p(y_{T+k}|Y_T)$ from (3.2.17) for $k>2$, but it can, in principle, be evaluated numerically. In order to shed some light on the nature of these calculations we derive its expression for two steps ahead. If we let $a=a_T$, $b=b_T$, $y_1=y_{T+1}$, $y_2=y_{T+2}$ then the appropriate substitution of (3.2.10) in the above expression leads to

$$p(y_{T+2}|Y_T)= k(y_2) \sum_{y_1=0}^{\infty} \frac{\Gamma(\omega a+y_1)\Gamma(\omega^2 a+y_2+\omega y_1)\, z^{y_1}}{\Gamma(\omega^2 a+\omega y_1) \quad y_1!} \qquad (3.2.18)$$

where

$$k(y_2)= \left[\frac{\omega b}{1+\omega b}\right]^{\omega a} \left[\frac{\omega b + \omega}{1+\omega^2 b+\omega}\right]^{\omega^2 a} \frac{1}{y_2!\ (1+\omega^2 b+\omega)^{y_2}\ \Gamma(\omega a)}$$

$$\simeq \frac{\omega^{\omega a(1+\omega)}(2-\omega)^{\omega^2 a}(1-\omega)^{y_2}}{y_2! \quad \Gamma(\omega a)} \tag{3.2.19}$$

and

$$z= \left[\frac{(\omega^2 b+\omega)^{\omega}}{(1+\omega b)\ (1+\omega+\omega^2 b)^{\omega}} \right] \simeq \omega^{\omega} (1-\omega) \qquad , 0 < z < 1. \tag{3.2.20}$$

The approximate values of the above expressions have been obtained by use of the asymptotic value of b_T for $0 < \omega < 1$. The infinite sum in (3.2.18) can be evaluated numerically along with the constant $k(y_2)$ producing probability values for y_{T+2}.

It is obvious that for $k \geqslant 3$ the computations involved in (3.2.17) will become very tedious. In these situations a number of strategies may be adopted and these are discussed below.

(i) *Monte Carlo simulation*- here we make use of the fitted model and a NBD random number generator (henceforth NBRG) (see Ch. 8) in order to simulate future values of the process y_t, say $\hat{y}_{T+i}$, $i=1,2,\dots,k$. The replications at each time are then used to evaluate the unconditional distribution of y_{T+i} and its moments. This approach has been advocated by S&M (1986), which use percentiles to predict records of some athletics data. In the absence of justifiable analytical approximations, this seems to be the most promising technique to tackle the forecasting problem in our setting. In what follows we briefly schematize how this procedure should be implemented in our framework, having the Poisson-Gamma as an example.

First note that as a by-product of the fitting of the Poisson-Gamma model to a certain data set one has readily available the ML estimate of the vector of hyperparameters $\hat{\Psi} = (\hat{\omega}, \hat{\delta}_1, \ldots, \hat{\delta}_p)$ and the updating values of the state prior parameters at time t=T, a_T and b_T. Once these quantities are made available one should proceed as follows:

1. For i=1 evaluate the *prediction equations*

$$a_{T+i/T+i-1} = \hat{\omega}\, a_{T+i-1}$$
$$b_{T+i/T+i-1} = \hat{\omega}\, b_{T+i-1}.$$

2. With the above values and an appropriate NBRG obtain one deviate

$$\hat{y}_{T+i} \sim \mathrm{NBD}(a_{T+i/T+i-1}, b_{T+i/T+i-1}).$$

3. Using the generated deviate $\hat{y}_{T+i}$ calculate the *updating equations*

$$a_{T+i} = a_{T+i/T+i-1} + \hat{y}_{T+i}$$
$$b_{T+i} = b_{T+i/T+i-1} + 1.$$

4. Set i=i+1 and repeat steps 1-3 until i=k.

5. Repeat steps 1-4 Nrep times, using a new seed at each time.

After the completion of the above process one will have available for each step ahead i=1,2,...,k a vector of generated values with dimension Nrep, i.e., { $\hat{y}_{T+i}(j)$ } j-1,2,...,Nrep. These may then be used to evaluate sampling moments, skewness and kurtosis, and the probability mass functions for each step ahead. In the case of continuous variates, as in the Gamma-Gamma model, percentiles could be evaluated as in S&M (1986).

(ii) *ad hoc approximation*- here we use the exact expressions for the k-steps ahead moments in (3.2.14-15) to construct an analytical approximation for the actual and unknown forecasting distribution $p(y_{T+k}|Y_T)$ for k>3. The 'natural' approximating distribution is the NBD($a_{T+k/T}$,$b_{T+k/T}$), but further investigation is still needed to justify the basis of this procedure. A similar approximation has been used by WHM (1985) in their DGLM (see section 1.3.2.1). In order to determine the projected parameters $a_{T+k/T}$ and $b_{T+k/T}$ one has only to match the first two moments of the above distribution (3.2.16a-b) with the correspondent moments of the NBD (see 3.2.11a-b) and solve for $a_{T+k/T}$ and $b_{T+k/T}$. After some algebra one may show that this yields

$$a_{T+k/T}= \omega\ a_T\ \ P_{k-1} \tag{3.2.21a }$$

$$b_{T+k/T}= \omega\ b_T\ \ P_{k-1} \tag{3.2.21b }$$

where $P_{k-1}= 1/[1+(1-\omega)b_T\ S_{k-1}]$ with S_{k-1} given by (3.2.16). Note that approximations of this kind may also prove useful in handling missing observations.

3.2.1 Likelihood

ML estimates for the unknown hyperparamter ω may be evaluated by substituting the formula for the predictive density (3.2.10) into (1.2.13). To initialize the gamma prior, that is the distribution of θ_t at time $t = 0$, we set $a_0=b_0 = 0$. Obviously this is an improper density. However, none of this prevents the recursions (3.2.3a-b) being initialized at $t=0$ with $a_0=b_0 = 0$. A proper distribution for θ_t is then obtained at time $t= \tau$ where τ is the index of the first non-zero observation. In this context this is also known as an 'unbiased' gamma prior (Hartigan 1983). The Jeffreys density is obtained by setting $a_0=1/2$, $b_0=0$. Note also that, it is possible to set b_t to its 'steady value' $\omega/(1-\omega)$ right from the beginning.

The log-likelihood function for the unknown hyperparameter ω is then given by

$$\log L(\omega) = \sum_{t=\tau+1}^{T} \{\log[\, \Gamma(a_{t/t-1} + y_t)/\, \Gamma(a_{t/t-1})] + a_{t/t-1} \log b_{t/t-1} - \\ - (a_{t/t-1} + y_t) \log (1+ b_{t/t-1})\} \qquad (3.2.22)$$

Maximization of the likelihood is accomplished via a quasi-Newton method based in the Gill-Murray-Pitfield algorithm provided by the NAG library (routine E04JBF). This routine is naturally fitted to handle constrained optimization so that no transformation of the hyperparameter ω is made necessary. Also given the fact that in E04JBF derivatives of the objective function are numerically calculated, no explicit formula for the derivatives of (3.2.15) are required, although we have worked them out in case of need. To calculate the terms involving the gamma function in (3.2.15) we have used the fact that for n integer and 'a' positive

$$\log[\ \Gamma(a+n)/\Gamma(a\ )] = \sum_{j=1}^{n} \log(a+j-1). \qquad (3.2.23)$$

At the present stage of our research we have been unable to find a satisfactory result which could be used to prove consistency and asymptotic normality of ML estimators. This would then enable us to construct confidence intervals for our point estimates. Sweeting's (1980) paper offers some general conditions for the establishment of these conditions for nonergodic stochastic processes, but we have not found them particularly useful in our context. Evidence of normality and consistency are nevertheless, provided in Chapter 8 where properties of ML estimators are investigated using Monte Carlo simulations. The overall message conveyed in our study is that our ML estimators are asymptotically unbiased, consistent and normally distributed for large sample sizes. For small and moderate sample sizes we have detected an unusual behaviour of the ML estimator of ω when it is set close to its upper bound value one. A significant proportion of its estimates will be exactly one even when its actual value is fixed at a different value. Similar behaviour of ML estimators are reported by Shephard and Harvey (1990) in the study of Gaussian local level models. Anyhow since values of interest for ω lie close to the boundary value the absence of confidence intervals should not be considered a serious drawback. Inference of parameters associated with structural components and regressors, to be introduced later, are made using the χ^2 approximation for the likelihood ratio test. This has also been advocated by Lawless (1987b) in the context of regression models for count data. For a more detailed discussion on this topic the reader is referred to Chapter 8, section 8.2.

3.2.2 The Discount Factor

Here we exploit in some detail the meaning conveyed by the discount parameter in our framework. We start by rewriting (3.2.13) as a geometric series. This has the following asymptotic form

$$\begin{aligned}\tilde{y}_{t+1/t} &= (1-\omega)y_t + (1-\omega)\omega y_{t-1} + (1-\omega)\omega^2\, y_{t-2} + \dots \\ &= c_0 y_t + c_1 y_{t-1} + c_2 y_{t-2} + \dots \qquad (3.2.24)\end{aligned}$$

where $c_i=(1-\omega)\omega^i$, $i=0,1,2,\dots$; $0< \omega < 1$. Given that the weights are normalized, then the above predictor may be thought of as splitting out the *information content* of the series into the current value, with weight c_0 and past values, with overall weight $1-c_0$. It is then obvious that ω represents a tradeoff between tracking ability and smoothing on the one step-ahead forecasting function (3.2.13). The more discounting is done (i.e. when $\omega\to1$), the more past terms will contribute for (3.2.24) and as a result the less the forecasts will track model changes. When the reverse occurs, i.e. when $\omega\to0$, then the current value dominates the weighted sum in (3.2.24). In this situation the system tracks very rapidly, inclusive of random changes.

In the context of state space models discounting factors have yet another suggestive interpretation which cannot be grasped in standard EWMA forecasting schemes. For example, in a Bayesian framework, Ameen and Harrison (1984-5) use discounting factors to tackle the increase in the predictive variance for components present in Gaussian DLM 's (see section 1.3.2.1). These then may be viewed as indicators of the ability of stochastic components in describing the movements of a series. Non-Gaussian structural models use discount factors in a similar fashion, but only in association with the *local level* component. If one looks at the role played by the discount factor in the transition equation for the variance of this component (3.2.6b), then it becomes clear that estimated low values of the discount parameter indicates a rather volatile level. Note that by considering the extreme value $\omega=1$ in (3.2.3a-b) we reduce our model to the static case. In this sense the deterministic components, such as time trend and seasonals, present in our formulation may be thought of as having discount factors fixed at unity.

Further insight into the discount factor may be obtained by considering its relation with the signal-to-noise ratio (SNR) of the Gaussian random walk plus noise model (see 1.2.1a-b). If the reader refers back to these equations then the SNR is defined as the ratio between the system noise variance and the measurement noise variance,i.e., SNR= Var η_t/ Var ϵ_t= q, $0 < q < \infty$. Now it can be shown that the *steady state* solution of the Kalman filter for the above specification produces a forecast function which is also equal to the EWMA scheme in (3.2.13). See Harvey (1989, p.175). It is then straightforward to establish the following relationship between the smoothing constant λ and q, the SNR

$$\lambda= (q+ \sqrt{q^2 + 4q})/(2+q + \sqrt{q^2 + 4q}).$$

Finally using that for the Poisson-Gamma model $\lambda=1-\omega$, it follows that

$$q= (1+\omega^2-2\omega)/\omega \ . \qquad (3.2.25)$$

As expected, when the level of the series behaves in a haphazard fashion , i.e., when $\omega \rightarrow 0$ then $q \rightarrow \infty$. For a deterministic level, $\omega \rightarrow 1$ and so $q \rightarrow 0$. The above formula may also be used to establish a link between the Poisson-Gamma model and the Gaussian model when the numbers on the series are not too small.

3.2.3 Explanatory Variables and Structural Components

For the Poisson-Gamma model the inverse of the log link function or the *exponential* link function will ensure that the contributions of the *systematic component* η_t (see 2.2.10) keeps the mean positive

$$h(\eta_t) = \exp(z_t' \delta). \qquad (3.2.26)$$

In our framework the way to proceed is by combining multiplicatively the *standard* mean μ_t, which is dependent on the actual and past values of the endogenous variable, with the exponential link function for the *systematic component* so that the distribution of y_t conditional on μ_t, is Poisson with mean

$$\begin{aligned} \mu_t^+ &= \mu_t \, h(\eta_t) \\ &= \mu_t \exp\,(z_t' \delta). \qquad (3.2.27) \end{aligned}$$

Using that conditional on Y_{t-1}, $\mu_t \sim \text{gamma}(\omega a_{t-1}, \omega b_{t-1})$, it follows from the properties of the gamma distribution, that, conditional on Y_{t-1}, $\mu_t^+ \sim \text{gamma}(a_{t/t-1}^+, b_{t/t-1}^+)$ where

$$a_{t/t-1}^{+} = \omega\, a_{t-1} \qquad (3.2.28a)$$

$$b_{t/t-1}^{+} = \omega\, b_{t-1} \exp(-z_t'\delta). \qquad (3.2.28b)$$

As regards updating, $\mu_t^+ \sim$ gamma(a_t^+,b_t^+) where a_t^+ and b_t^+ are obtained from $a_{t/t-1}^+$ and $b_{t/t-1}^+$ *via* updating equations of the form (3.2.8). Therefore the posterior distribution of μ_t is gamma(a_t,b_t), where a_t and b_t are given by

$$a_t = \omega\, a_{t-1} + y_t \qquad (3.2.29a)$$

$$b_t = \omega b_{t-1} + \exp(z_t'\delta), \qquad t = \tau + 1 ,\ldots, T. \qquad (3.2.29b)$$

Thus the only amendment as compared with the recursions of the standard case (3.2.8) is the replacement of unity by $\exp(z_t'\delta)$ in the equation for b_t. The log-likelihood of the observations is therefore as in (3.2.22) with $a_{t/t-1}$ and $b_{t/t-1}$ replaced by $a_{t/t-1}^+$ and $b_{t/t-1}^+$. This must be maximized with respect to ω and δ.

From (3.2.27) it follows that in the presence of exogenous variables/structural components the Poisson mean is modelled by an expression of the form

$$\mu_t^+ = \mu_t \exp(\, R_t + T_t + S_t\,) \qquad (3.2.30)$$

where R_t, T_t and S_t are as previously defined (see 2.2.12-14).The form of the above *linking* mechanism means that the trend and seasonals combine multiplicatively, just as in a logarithmic Gaussian model. As in such a model, the coefficient of the slope is to be interpreted as a growth rate, while the seasonal coefficients are multiplicative seasonal factors.

The extension of count data models to include explanatory variables opens up the possibility of carrying out *intervention analysis*. The explanatory variables x_t are replaced or augmented by a variable w_t which is designed to pick up the effect of some event or policy change.

Forecasting in the presence of explanatory variables/structural components:

For a given value of δ, we can proceed as in (3.2.8) to show that the mean of the predictive distribution of y_{T+k} is

$$E(y_{T+k} | Y_T) = \exp(z_{T+k}'\delta)\ (a_T/b_T^+)$$

$$= \exp(z_{T+k}'\delta) \sum_{j=0}^{T-1} \omega^j\ y_{T-j}\ \Big/ \sum_{j=0}^{T-1} \omega^j \exp(z_{T-j}'\ \delta)$$

$$= \exp(z_{T+k}'\delta)\ \mathrm{EWMA}(y)/\ \mathrm{EWMA}\{\exp(z'\delta)\} \qquad (3.2.31a)$$

where EWMA(y) is given by (3.2.13) and EWMA{exp(z'δ)} is defined similarly. Note that if structural effects are present in the systematic component then multi-step forecasts are obtained by direct projection of these components, i.e. by substituing t for t=T+k in

(3.2.12) and (2.2.13-14) respectively. Using an argument similar to that employed to derive the projected variance in the standard case (3.2.16) it is also possible to show that

$$\mathrm{Var}(y_{T+k/T}) = (a_T/\omega \overset{2}{b}_{T+k}{}^{*}) \; (1+ \omega b_{T+k}{}^{*} + (1-\omega) \; b_{T+k}{}^{*} \; S^{*}{}_{k-1}) \tag{3.2.31b}$$

where $b_{T+k}{}^{*} = \exp(x_{T+k}{}'\delta)b_{T+k}$ and

$$S^{*}{}_{k-1} = \sum_{j=1}^{k-1} (1/b_{T+j})\exp[(z_{T+k} - z_{T+j})'\delta]. \tag{3.2.32}$$

As before Monte-Carlo simulation and scenario projection might be used to construct forecasts. The k-steps ahead distribution may again be approximated by a NBD$(a_{T+k/T}, b^{+}{}_{T+k/T})$ where $a_{T+k/T}$ is as in (3.2.18a) and $b_{T+k/T}$ is

$$b_{T+k/T} = \omega \; b_T \; \exp(-z_{T+k}{}') \; P^{*}{}_{k-1} \tag{3.2.33}$$

with

$$P^{*}{}_{k-1} = 1/[1+b_T+(1-\omega)S^{*}{}_{k-1}]. \tag{3.2.34}$$

In the absence of explanatory variables, $\delta=0$ and as a result (3.2.32) collapses to the standard case (3.2.16) as expected.

It is interesting to compare (3.2.31a) with the result obtained from the Gaussian model (1.2.1a-b) for a given discount factor, ω. Since the level and explanatory variables are combined multiplicatively in this model, it seems sensible to make the comparision with a Gaussian model in which logarithms have been taken. The optimal estimator of μ_t is obtained by applying the EWMA operation to $\log y_t - x_t{}'\delta$. The optimal estimate of $\log y_{T+k}$ can then be expressed as

$$E(\log y_{T+k} | Y_T) = x_{T+k}'\delta + \text{EWMA}(\log y) - \text{EWMA}(x'\delta). \quad (3.2.35)$$

The other point of comparision with the Gaussian model is in the maximization of the respective likelihood functions. In the Gaussian case, the computational burden is eased considerably by the fact that δ may be concentrated out of the likelihood function by estimating it by generalized least squares; see Kohn and Ansley (1985). This suggests that it may be possible to use estimates from the Gaussian model as starting values; the difficulty lies in how to handle zero observations when logarithms are being taken.

3.3 THE NEGATIVE BINOMIAL-BETA MODEL:

- *measurement equation*: let the observations at time t be drawn from a NBD distribution,

$$p(y_t|\pi_t,\upsilon) = \begin{bmatrix} \upsilon + y_t - 1 \\ y_t \end{bmatrix} \pi_t^{\upsilon} (1-\pi_t)^{y_t} \quad , y_t= 0,1,2,... \qquad (3.3.1)$$

where $0 < \pi_t < 1$ and $\upsilon > 0$. This is known as the Pascal distribution if υ is an integer and υ=1 corresponds to the geometric distribution. In terms of our notation we have that the state $\theta_t= \pi_t$. The mean and variance are

$$E(y_t|\pi_t,\upsilon) = \upsilon(1-\pi_t)/\pi_t \qquad (3.3.2a)$$

$$Var\ (y_t|\pi_t,\upsilon) = E(y_t/\pi_t)\ [1 + \upsilon^{-1}\ E(y_t/\pi_t)\ ] \qquad (3.3.2b).$$

The distribution therefore exhibits overdispersion compared with the Poisson distribution, that is the variance exceeds the mean. However, if the mean is kept constant, the NBD tends towards the Poisson distribution as $\upsilon \to \infty$.

(i) *state prediction*- the conjugate prior distribution for the NBD is the beta density, so that $p(\pi_{t-1}|Y_{t-1})$ has the form

$$p(\pi_{t-1}|Y_{t-1}\ )=\frac{\pi_{t-1}^{a\ -1}\ \ (1-\pi_{t-1})^{b\ -1}}{B(a,b)} \qquad (3.3.3)$$

where $a=a_{t-1} > 0$, $b=b_{t-1} > 0$ and $B(\cdot)$ is the beta function given by $B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$. It is a standard result for this density that

$$E(\pi^n)=\frac{\Gamma(a+b)\ \Gamma(a+n)}{\Gamma(a)\ \Gamma(a+b+n)} \qquad n=1,2,3,\ldots \qquad (3.3.4a)$$

and

$$\mathrm{Mode}(\pi)=\frac{a-1}{a+b-2} \qquad a > 1, \quad b > 1. \qquad (3.3.4b)$$

Our approach led us to assume that $p(\pi_t|Y_{t-1})$ is also beta. At first sight it might appear that the recursions in (3.2.3) are again appropriate to express the predictive equations. However, in view of (2.2.5a), it is the expected value of $(1-\pi)/\pi$, rather than π, which needs to be kept constant while the variance increases. For a beta distribution, (3.3.3), one may easily show that

$$E((1-\pi)/\pi)=\frac{B(a-1,\ b+1)}{B(a,b)}=\frac{b}{a-1} \qquad (3.3.5)$$

provided that $a > 1$. Hence , by using (2.2.5a) we require that

$$\frac{b_{t/t-1}}{a_{t/t-1}-1}=\frac{b_{t-1}}{a_{t-1}-1} \qquad (3.3.6)$$

This can be achieved by multiplying the numerator and denominator in the expression on the right hand side of (3.3.6) by ω. The prediction equations will therefore take the form

$$a_{t/t-1} = \omega\, a_{t-1} + (1 - \omega) \qquad (3.3.7a)$$

$$b_{t/t-1} = \omega\, b_{t-1} \qquad (3.3.7b)$$

with $0 < \omega < 1$. In order to have the variance of $\pi/(1-\pi)$ increased during the transition one may show that the following condition has to be satisfied

$$a_{t-1} > (1+\omega)/(1-\omega). \qquad (3.3.8)$$

We leave a more detailed investigation of this inequality after we have established the updating equations.

Note that a multiplicative transition equation similar to that derived for the Poisson-Gamma model (3.2.7) could also, in principle, be worked out for the beta parameter π_t, by use of the multiplicative property of beta variates (see, e.g.,McKenzie 1985) given by

$$Be(a,b)\ .\ Be(a+b,c) = Be(a,b+c)$$

where $Be(\cdot)$ is the beta density. It is not difficult to see that such a specification would require transition equations different from (3.3.7a-b) and that as a result the forecast function would no longer be expressed by an EWMA scheme. Furthermore this would be of no help in finding an useful expression for the multi-step moments, given that these would be dependent on future values of the observable.

(ii) *state updating* - the posterior distribution for the state, $p(\pi_t | Y_t)$ is also a beta distribution with parameters

$$a_t = a_{t/t-1} + \upsilon \qquad (3.3.9a)$$

$$b_t = b_{t/t-1} + y_t. \qquad (3.3.9b)$$

Repeatedly substituting from (3.3.7a) and (3.3.9a) gives

$$a_{t/t-1} = \upsilon \sum_{j=1}^{t-1} \omega^j + (1 - \omega) \sum_{j=1}^{t-1} \omega^j + (1 - \omega). \qquad (3.3.10)$$

As before when t → ∞, $a_{t/t-1}$ converges to the 'steady solution' $[\upsilon(1-\omega)+1]/(1-\omega)$ so that $a_{t-1} \to (\upsilon+1)/(1-\omega)$. Using this result one may easily prove that the condition in (3.3.8), which guarantees the increase on the variance of the level during the transition, is satisfied for $\upsilon > \omega$.

(iii) *conditional distribution* - the predictive distribution is obtained by solving the compounding operation in (1.1.4). The resulting distribution is the beta-Pascal (see,e.g., Raiffa and Schlaifer 1961, p.238)

$$p(y_t | Y_{t-1}) = \frac{1}{\upsilon + y_t} \; \frac{B(\upsilon + a_{t/t-1}, y_t + b_{t/t-1})}{B(\upsilon, y_t + 1)\, B(a_{t/t-1}, b_{t/t-1})} \qquad (3.3.11)$$

The mean and variance of the above distribution are given respectively by

$$\tilde{y}_{t/t-1} = E(y_t | Y_{t-1}) = \frac{\upsilon\, b_{t/t-1}}{a_{t/t-1} - 1} = \frac{\upsilon\, b_{t-1}}{a_{t-1} - 1} \qquad , a_{t-1} > 1 \qquad (3.3.12a)$$

and

$$\mathrm{Var}(y_t | Y_{t-1}) = \frac{\upsilon\, b_{t/t-1} \, [(a_{t/t-1} + b_{t/t-1} - 1)(a_{t/t-1} + \upsilon - 1)]}{(a_{t/t-1} - 1)^2 \, (a_{t/t-1} - 2)}$$

$$= \frac{\upsilon\, b\, (a + b + 1)\, [\upsilon + \omega(a - 1)]}{(a - 1)^2 \, [\omega\, (a - 1) - 1]}$$

$$= E(y_t | Y_{t-1}) \, \frac{(a + b + 1)[\upsilon + \omega(a-1)]}{(a - 1)\, [\omega(a-1) - 1]} \quad , a > (1+\omega)/\omega \qquad (3.3.12b)$$

where for ease of notation we have made $a=a_{t-1}$ and $b=b_{t-1}$. Note that this shows overdispersion with respect to the Poisson model. It is straightforward to show that, when $t \to \infty$, $b_{t/t-1}$ can be written as an exponentially weighted average of past observations, and using that $(a_{t/t-1} - 1) \to \omega\upsilon/(1-\omega)$, the predictor $\tilde{y}_{t/t-1}$ in (3.3.12a) may again be shown to have the EWMA form (3.2.13).

As regard the relation between our updating rule and Smith's, in the present context they differ radically. This can be understood by noting that while our rule keeps invariant the quantity $b/(a-1)$ (see 3.3.5b), Smith's keeps unchanged $(a-1)/(a+b-2)$ (see 3.3.4b).

(iv) *k-steps ahead forecasting*- using an argument similar to that employed to show (3.2.14b), it is possible to verify that the forecasts k steps ahead, $k \geq 1$ are also given by an EWMA scheme.

However we have not succeeded in deriving a general expression for the variance k steps ahead. The same holds true for the forecasting distribution. In what follows we present the expressions for the first factorial moment and the distribution two steps ahead.

As in the Poisson-Gamma model the way to proceed is by first evaluating the factorial moment of order one using the chain rule for conditional expectations (1.1.9a). The variance is then obtained by substituting the derived expression, together with the appropriate forecast function (3.2.13) into (A2.2). After some tedious manipulation it is possible to show that the two steps ahead factorial moment of order one is given by

$$E(y_{T+2}{}^{(2)} | Y_T) = E(y_{T+1}{}^{(2)} | Y_T) + \frac{\upsilon(\upsilon+1)\omega b\{h(1-\omega)+\omega[(b-\omega)-\omega h(1+b)]\}}{h(h-1)[\omega(h+\upsilon)-1]} \tag{3.3.13}$$

where $b = b_T$, $E(y_{T+1}{}^{(2)} | Y_T) = \upsilon(\upsilon+1)\omega b(1+\omega b)/[h(h-1)]$ and $h = \omega(a_T - 1)$. As regard the two step ahead forecast distribution, the appropriate substitution of (3.3.9) into (3.2.17) leads to

$$p(y_{T+2} | Y_T) = k(y_2) \sum_{y_1=0}^{\infty} \frac{\Gamma(A+y_2)\,\Gamma(B)\,\Gamma(\upsilon+y_1)\,\Gamma(A+\omega y_1)}{\Gamma(B+y_2)\,\Gamma(A)\,y_1!} \tag{3.3.14}$$

where $y_1 = y_{T+1}$, $y_2 = y_{T+2}$, $A = (\omega^2 b_T + y_1)$, $B = \omega^2(a_T+b_T)+(1-\omega)(1+\omega)+\omega(y_1+\upsilon)$ and

$$k(y_2) = \frac{\Gamma(\upsilon+y_2)\,\Gamma(D+\omega B_t)\,\Gamma(C+\upsilon)\,\Gamma(D+\upsilon)}{\Gamma^2(\upsilon)\,\Gamma(D)\,\Gamma(\omega B_t)\,\Gamma(C)\,\,y_2!}$$

with

$$C= \omega^2\, a_T+ (1-\omega)(1+\omega)+ \upsilon\omega \;\hat{=}\; 1+ \upsilon/(1-\omega)$$

$$D= \omega a_T + (1-\omega) \;\hat{=}\; 1+ \upsilon\omega/(1-\omega)= a_T.$$

The approximated values are obtained for large sample size when $a_T= a_T^* \simeq 1+ \upsilon\omega/(1-\omega)$. Under this condition it is not difficult to see that $k(y_2)$ becomes

$$k(y_2)= \frac{\Gamma(\upsilon+y_2)\ \Gamma(a_T^*+\omega b_T)\ \Gamma(a_T^*+\upsilon)}{\Gamma^2(\upsilon)\ \Gamma(1+\omega(a_T^*-1))} \quad .$$

3.3.1 <u>Likelihood</u>

The parameter υ can be estimated by ML along with ω. Alternatively it may be pre-set. Using (3.3.8) one can write the log-likelihood function for the hyperparameters υ and ω as

$$\log L(\omega,\upsilon)= \sum_{t=\tau+1}^{T} \{ \log[\Gamma(\upsilon+a_{t/t-1})/\Gamma(a_{t/t-1})]+ \log[\Gamma(\upsilon+y_t+1)/\Gamma(\upsilon)]+$$

$$+ \log[\Gamma(y_t+b_{t/t-1})/\Gamma(b_{t/t-1})]-\log[\Gamma(y_t+\upsilon+ d_{t/t-1})/d_{t/t-1}] \} \qquad (3.3.15)$$

where $d_{t/t-1}= a_{t/t-1} + b_{t/t-1}$. We start the recursions (3.3.7a-b) and (3.3.8a-b) at t=0 with an 'unbiased' beta prior ,i.e., by setting $a_0=b_0=0$. In order to ensure that b_t is strictly positive we require τ to be the first value of t for which y_t is non-zero; a_t will be always positive. It is clear this is not a uniform beta prior, since this is obtained by setting a_0=b_0=1; the Jeffreys prior is obtained by setting a_0=b_0=1/2. When υ is set to the integer value which maximizes (3.3.15) then (3.2.23) may be used in order to evaluate the logarithm of the ratio of gamma functions. For non-integers values of υ one has to evaluate the gamma function directly and for this we have used the routine GAMMLN in Press *et al* (1986, p.157).

3.3.2 Explanatory Variables and Structural Components

The appropriate way of proceeding with the NBD-Beta model is to introduce the explanatory variables/structural components directly into the distribution of y_t/π_t *via* an exponential link function. This may be done by replacing υ by $\upsilon_t^+ = \upsilon \exp(\eta_t'\delta)$. Such a NBD distribution has, for a constant π, a constant variance-mean ratio; see the discussion in Cameron and Trivedi (1986, p.33). Proceeding in this way leads to the updating equation (3.3.7a) being modified to

$$a_t = a_{t/t-1} + \upsilon \exp(z_t'\delta) \qquad (3.3.16)$$

while (3.3.8b) remains unchanged. It is then possible to show that the mean of the predictive distribution of y_{T+k} is

$$E(y_{T+k}|Y_T) = \upsilon \exp(z_{T+k}'\delta)\, b_T/(a_T-1) \qquad (3.3.17)$$

and it is not difficult to deduce that it can be expressed in terms of an equation identical to (3.2.32a).

3.4 MODEL SELECTION:

In this section we introduce our model selection methodology, and with few exceptions most of the techniques discussed here are also suitable to the non-count data models considered in this dissertation.

Many of the issues which arise in the selection of GLIM models are also relevant here. However there is the additional problem of testing for randomness. Given the non-Gaussian nature of our model one has to rely on non-parametric tests which may be used both for the residuals and raw data. In particular we have implemented the following randomness/trend tests in our program:

Table 3.4.1 List of the tests for randomness/trend implemented.

TEST	STATISTICS
Runs above and below the median (or Runs for short)	Standard Normal
Runs up and down (or Rud for short)	Standard Normal
Kendall's tau test for trend	Standard Normal
Daniel's rank test for trend	Standard Normal
Rank version of Von Neumman ratio	VNR

The reference for the first four tests is Farnum and Stanton (1989, ch.2) while for the last test the reader is referred to Bartels (1982) who shows that the rank version of Von Neumman ratio has far greater power than the turning point test when used in AR(1) models with different distributional assumptions. The critical values for the VNR statistic may be evaluated through the formula

$$f_{\alpha}(T) = a + b\, T^{c}\, (\log T)^{d} \qquad (3.4.1)$$

where α is the significance level of the test, T is the sample size and the parameters a,b,c, and d are given in the table below for different significance levels.

Table 3.4.2 Values of coefficients of the VNR formula for different significance levels.

α	.005	.01	.025	.05	.1
a	-.040	-.023	-.004	.119	-.465
b	.200	.261	.381	.440	1.184
c	-.400	-.345	-.266	-.230	.088
d	2.540	2.212	1.748	1.520	.674

The null hypothesis of randomness should be rejected when VNR < $f_{\alpha}(T)$.

Once randomness of the residuals is checked one can proceed by evaluating the standardised (Pearson) residuals which are defined by

$$\nu_t = \frac{y_t - E(y_t \mid Y_{t-1})}{SD(y_t \mid Y_{t-1})} \qquad (3.4.2)$$

If the parameters in the model are known, it follows from the decomposition of the likelihood in (1.2.13) that these residuals are independently distributed with mean zero and unit variance. However, they are not, in general, identically distributed.

The following diagnostic checks are suggested in order to check model adequacy:

a) An examination of the plot of the residuals against time and against an estimate of the level.

b) A check on whether the sample variance of the residuals is close to one. A value greater than one indicates overdispersion relative to the model which is being fitted. (Note that since the mean of the residuals is not necessarily zero, the sample variance and raw second moment will not usually be the same).

c) When discriminating between alternative models one must select the model which produces the smallest of the following goodness of fit criteria:

- Akaike information criterion:

$$AIC= -2\ ML(\Psi)+ 2\ p\ . \tag{3.4.3a}$$

When no likelihood is explicitly available then one may adopt as the objective function the sum of squared residuals (SSR), so that the above formula becomes

$$AIC= \log SSR+ 2\ p\ . \tag{3.4.3b}$$

- Bayesian information criterion:

$$BIC= -2\ ML(\Psi)+ p\ \log T \tag{3.4.4.a}$$

or

$$\text{BIC} = \log \text{SSR} + p \log T. \qquad (3.4.4.b)$$

- Theil's U statistic:

$$U = \sqrt{\text{SSR(model)}} \;/\; \sqrt{\text{SSR('naive' model)}} \qquad (3.4.5)$$

where $ML(\Psi)$is the maximized log-likelihood, p is the number of independent hyperparameters, T is the number of observations used to fit the model, and the 'naive' model is the model for which the one-step ahead prediction is set equal to the last observation. The AIC and BIC version used in most of our comparisons are the ones with the likelihood function (3.4.3a and 3.4.4a) unless otherwise stated. For further discussion on the topic of model selection using information criteria the reader is referred to Priestley (1981, ch. 5).

Post-sample predictive testing may also be carried out. For the model with Poisson observations, the post-sample predictive test statistic is

$$\xi(\ell) = 2\sum_{t=T+1}^{T+\ell} a_{t/t-1} \log(a_{t/t-1} / y_t\, b_{t/t-1}) \qquad (3.4.6)$$

$$- 2\sum_{t= T+1}^{T+\ell} (a_{t/t-1} + y_t) \log (y_t + a_{t/t-1} / (1+b_{t/t-1})y_t)$$

where $a_{t/t-1}$ and $b_{t/t-1}$ are computed from the recursions (3.3.7a-b). In the special case when y_t is zero, the term in $\xi(\ell)$ at time t is

$$-2\, a_{t/t-1} \log ((1 + b_{t/t-1})/ b_{t/t-1}).$$

Under the null hypothesis that the model is correctly specified, $\xi(\ell)$

is asymptotically χ^2 with ℓ degrees of freedom. The test is analogous to the test developed by Chow (1960) for a Gaussian regression model. The derivation in the Appendix is based on the introduction of a dummy variable into the model for each of the observations in the post sample period.

3.5 APPLICATIONS:

This section illustrates the use of our classes of count data models by considering some applications to real data. When possible we compare the performance of our models with the equivalent structural Gaussian model and also with alternative classes of count data models.

3.5.1 Goals Scored by England against Scotland

Here we analyse the series of the number of goals scored by England in international football matches played against Scotland at Hampden Park in Glasgow (see figure 3.5.1). The source for this data is 'The Official Football Association Yearbook 1985/1986' (Pelham Books). Apart from the war years these matches were played in Glasgow every other year (the year 1985 is also an exception; the match should have been played at Wembley). Treating the observations as though they were evenly spaced, estimation of the Poisson-Gamma model gave:

Table 3.5.1 Poisson-Gamma model fitted to the England series of goals for matches played at Hampden Park.

estimate	goodness-of-fit				
$\tilde{\omega}$	AIC	BIC	SSR	ML	U
0.844	102.65	104.58	91.937	-50.323	0.751

The variance of the standardized residuals is 1.24. When subjected to the randomness tests of section 3.4, the residuals showed no signal of structure. A post-sample predictive test carried out over the last five and ten observations gave no hint of model breakdown with $\xi(5)=0.377$ and $\xi(10)= 8.478$. The forecasted value for the mean of future observations is 0.82. This corresponds to the forecast that would have been obtained from the Gaussian random walk plus noise model (1.2.1) by setting q=0.029 (see 3.2.25).

Figure 3.5.1 Goals scored by England against Scotland at Hampden Park and estimated level using the Poisson-Gamma model.

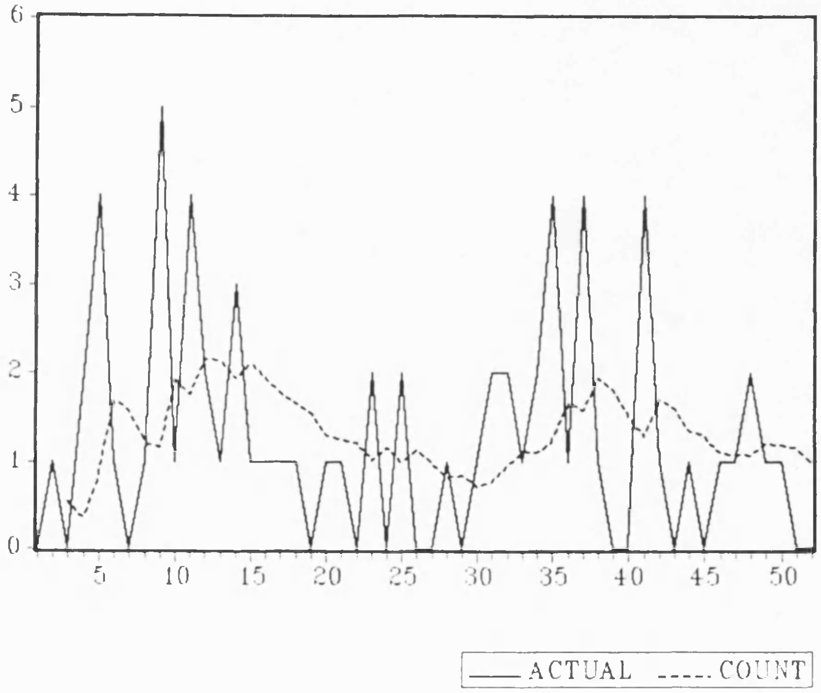


Table 3.5.2 Predictive probability distribution of goals in next match.

Number of goals					
0	1	2	3	4	>4
.471	.326	.138	.046	.013	.005

Fitting the NBD-Beta model when υ is estimated by ML yields

Table 3.5.3 NBD-Beta model fitted to the England series of goals for matches played at Hampden Park.

estimates		goodness-of-fit				
$\tilde{\omega}$	$\tilde{\upsilon}$	AIC	BIC	SSR	ML	U
0.965	4.819	103.71	107.53	90.86	-49.856	0.747

Thus the introduction of an adjustable scale parameter has resulted in less movement in the level. The variance of the standardised residuals is 1.0467 and the prediction is 1.1931. The likelihood function is relatively insensitive with respect to changes in υ. Furthermore its value at the maximum is only marginally greater than the maximised likelihood for the Poisson-Gamma model. If an allowance is made for the extra parameter *via* the AIC or BIC, the Poisson-Gamma model gives a better fit.

We now consider the full set of results of England-Scotland matches (see figure 1.1), with the model extended by the introduction of a dummy variable which takes a value of unity when England are at home. Playing at home tends to confer an advantage, and so we extend the model by introducing a dummy explanatory variable which takes a value of unity when England are at home, and is zero when they are away. Here we only report the fitting using the Poisson-Gamma model, since no specification of the NBD-Beta model has produced a better result. For discriminatory purposes we estimate the Poisson-Gamma model with (Mod1) and without the dummy variable (Mod2). The results are as follows:

Table 3.5.4 Poisson-Gamma model with dummy variable fitted to the England series of goals for matches played either at England or Hampden Park.

	estimates		goodness-of-fit				
	$\tilde{\omega}$	$\tilde{\delta}$	AIC	BIC	SSR	ML	U
Mod1	0.892	0.496	158.53	163.82	257.50	-77.263	0.679
Mod2	0.893	-	165.18	167.82	289.17	-81.590	0.699

As expected, the estimate of δ is positive. The likelihood ratio test statistic is 8.66: this statistic is asymptotically $\chi_1{}^2$ under the null hypothesis that δ is zero, and so is clearly highly significant. Since $\exp(0.496) \approx 1.64$, the results can be interpreted as saying that the expected number of goals scored by England rises approximately by 64% when they are playing at home.

3.5.2 Purse Snatching in Chicago

In their textbook, McCleary and Hay (1980) list a time series of reported purse snatchings in the Hyde Park neighbourhood of Chicago. The observations were collected by Reed (1978), and are twenty-eight days apart, running from January 1968 to September 1973. McCleary and Hay decided that the series was stationary and on the basis of the correlogram and sample partial autocorrelation function they fitted an AR(2) model.

The assumption of stationarity for this series implies that the level of purse snatchings remained constant throughout the period in question, and that the variations observed were simply fluctuations around this constant level. This in turn implies that purse snatching is in some kind of equilibrium. While this may be true, a more plausible working hypothesis is that the level of this crime is gradually changing over time. This suggests a Gaussian random walk plus noise model, (1.2.1). Estimating such a model under the time domain gives a signal noise ratio of q= 0.208. The residuals give no indication of serial correlation. For example, the Box-Ljung statistic (see,e.g., Harvey 1989, p. 259), Q(8) is equal to 7.88, and this should be tested against a chi-square distribution with 7 degrees of freedom. The prediction error variance is estimated to be 38.94, and this is only slightly above the figure reported by McCleary and Hay for their AR(2) model which, of course, contains one more parameter.

In summary, basic *a priori* considerations give rise to a structural time series model which not only has a clearer interpretation than the ARIMA model fitted by McCleary and Hay, but is more parsimonious as well. However the model is not, strictly speaking, *data admissible*. The forecast function is horizontal and cannot be negative, but a prediction interval of one RMSE on either side rapidly strays into the region of negative values of y. A logarithmic formulation, on the other hand, is not satisfactory as it fails the Bowman-Shenton normality test. A much better model is obtained by carrying out a square root transformation before fitting the model. Note that the square root transformation is the variance stabilizing transformation for a Poisson distribution; see McCullagh and Nelder (1983, pp. 129-130). The resulting fitting yields a time-domain estimated signal-noise ratio of q=0.1465, while squaring the forecasted values gives predictions of 7.39 and a much narrower prediction interval.

Of course, the purse snatchings are an example of count data, but since the numbers are not too small fitting various Gaussian models is a useful preliminary exercise. (For example, extending the model to include a stochastic slope indicates that such a component is unneccesary).

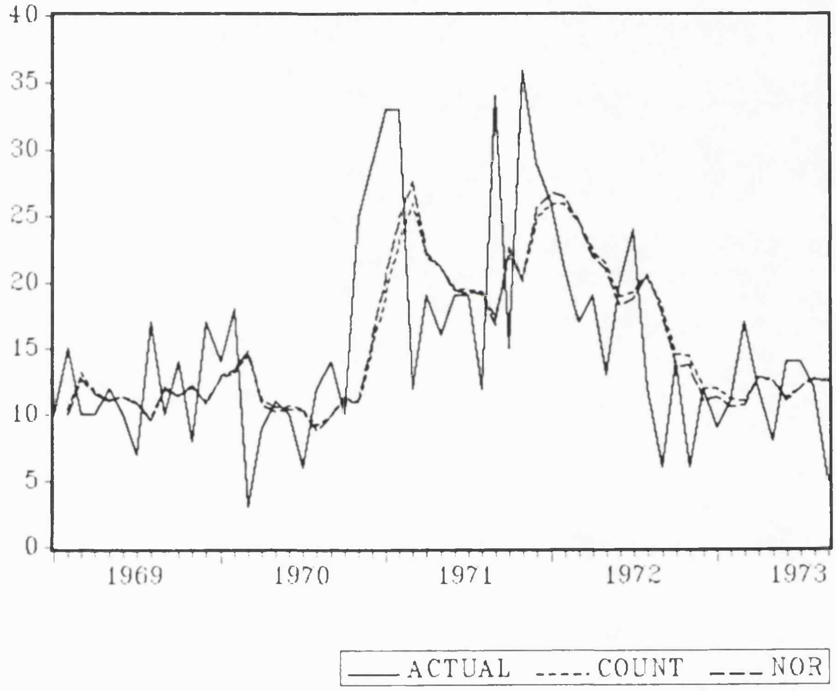
When the data are treated explicitly as count data, a NBD-Beta model seems to produce the best fit, whose summary follows

Table 3.5.5 NBD-Beta model fitted to the purse-snatching data.

estimates		goodness-of-fit				
$\tilde{\omega}$	$\tilde{\upsilon}$	AIC	BIC	SSR	ML	U
0.707	18.026	-3323.97	-3319.19	2748.80	1663.84	0.834

The predicted level is 7.66, corresponding to predictions from the Gaussian model with q=0.131. A plot of the residuals shows no evidence of heteroscedasticity, while the sample variance of the standardised residuals is 0.965. As an exercise of comparison we display in figure (3.5.2) the one step ahead predictions produced by our NBD model and these obtained through the fit of a Gaussian local level model. It is no surprise that the Gaussian model performs quite satisfactorily since the data values are not too small.

Figure 3.5.2 Purse snatchings in Hyde Park, Chicago and estimated levels using a NBD-Beta model (count) and a Gaussian model (nor).



3.5.3 Effect of the Seat Belt Law on Van Drivers in Great Britain

The effect of the seat belt law of January 1983 on various classes of road users in Great Britain was analysed in Harvey and Durbin (1986). For certain categories the numbers involved were relatively small with the result that a Gaussian model could not be regarded as a reasonable approximation. One such series is the monthly totals of drivers of light goods vehicles (LGV) killed. Here the numbers, from January 1969 to December 1984, range from two to seventeen. Since the series contains no zero observations, a Gaussian model can be fitted to the logarithms of the observations. This gives preliminary estimates for the seasonal and intervention effects which can be used as starting values in the iterative procedure used to calculate the ML estimators in a count data model. However, it is clear from doing this that a Gaussian model is not at all satisfactory in these circumstances and the results are very different for different specifications. In particular, fitting a model with fixed seasonals and no slope gives an estimate of the intervention effect which implies a 45% fall in fatalities as a result of the seat belt law. This is quite out of line with estimates obtained for other series, and indeed with the results obtained when a slope is included.

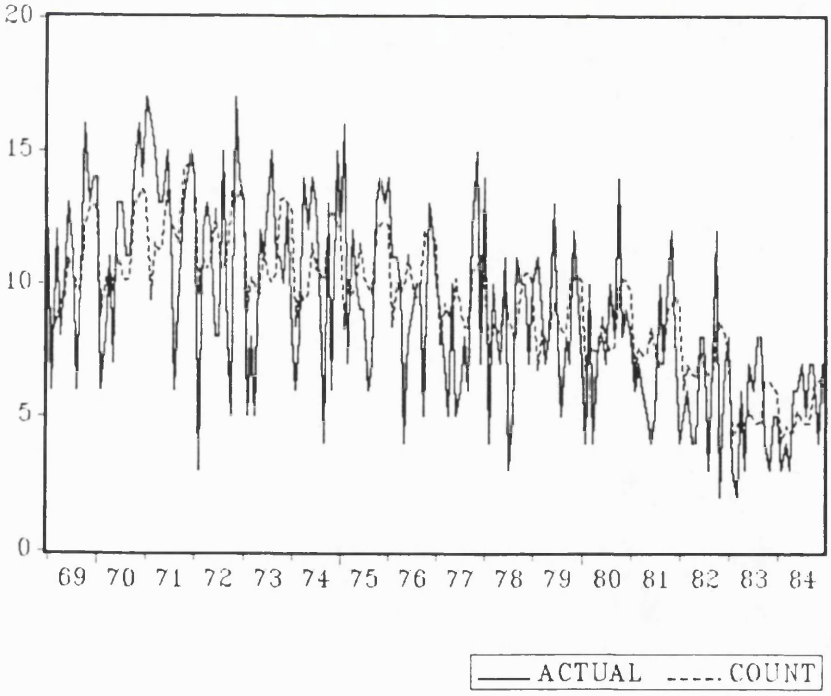
For the Poisson model it is reassuring to note that the conclusions regarding the effect of the seat belt law are affected very little by the inclusion or otherwise of a slope term. In fact the preferred specification does not have a slope. The explanatory variables are therefore an intervention and seasonals, and fitting the model gives the following estimates of ω and the intervention effect:

Table 3.5.6 Poisson-Gamma model with seasonals and intervention variable fitted to the series of LGV drivers killed.

estimates		goodness-of-fit				
$\tilde{\omega}$	$\tilde{\delta}$	AIC	BIC	SSR	ML	U
0.934	-0.2764	-4239.24	-4196.96	1480.7	2132.62	0.702

The estimate of δ implies a 24.1% reduction in fatalities which is quite close to the figures reported earlier for car drivers by Harvey and Durbin (1986). The likelihood ratio test statistic for the inclusion of the intervention variable is 25.96 and this is clearly significant when set against a $\chi_1{}^2$ distribution. Figure 3.3 shows the plot of the LGV series and the fitted Poisson-Gamma model.

Figure 3.5.3 LGV drivers killed in Great Britain and fitted Poisson-Gamma model with seasonals and intervention.



Finally the estimated seasonal factors, given by exponentiating the estimated seasonal coefficients, are very reasonable and not dissimilar to the seasonal factors reported by Harvey and Durbin (1986) for car drivers killed and seriously injured.

Table 3.5.7 Estimated seasonal factors for LGV drivers killed.

J	F	M	A	M	J	J	A	S	O	N	D
1.16	.79	.94	.89	.91	1.06	.97	.92	.92	1.16	1.19	1.19

3.5.4 U.S. Polio Incidences

Zeger (1988) lists a time series of the monthly number of cases of poliomyelitis reported by the U.S. Centers for Disease Control. The data runs from 1970.01 to 1984.12 and has been used as an illustration of his model. The specification chosen by Zeger includes a linear trend with seasonal harmonics, and this is based both on the evidence of seasonality on the spells of the disease and on a desire to investigate claims about a long-term decrease in the rate of U.S. polio infection. In considering the seasonal pattern Zeger has used only the annual and semi-annual frequencies in (2.2.14), i.e., $s=12$ and $j=1,2$. It is also worth noting that the November 1972 observation has been considered an outlier, but in his analysis, Zeger has not removed this observation 'since it had a minor effect on the findings'. In order to compare our class of count data models with Zeger's formulation we have specified a model with similar components, but treating the outlier explicitly, for which a dummy variable is defined. The table below presents the result of our best model, a NBD-Beta, against Zeger's parameter driven model.

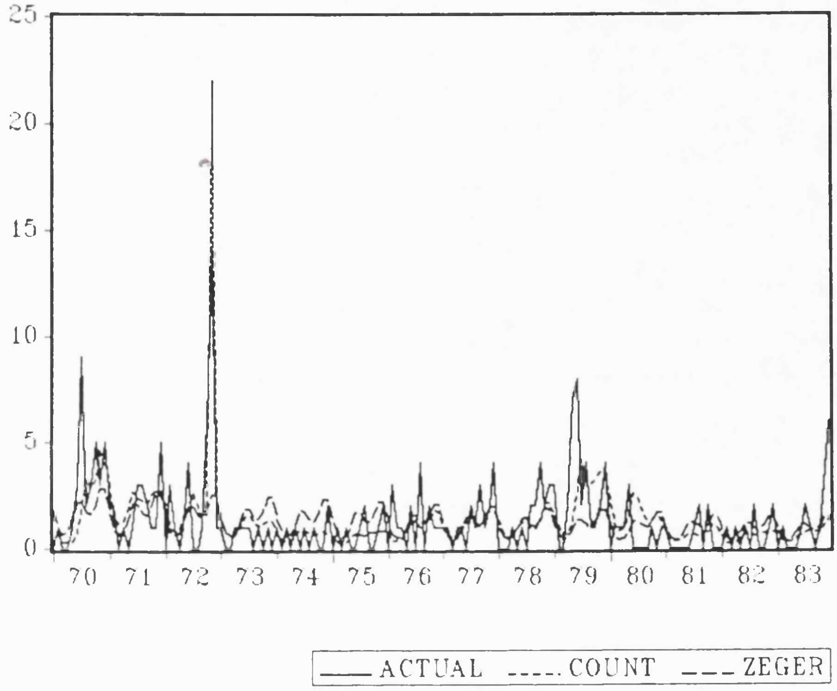
Table 3.5.8 NBD-Beta and Zeger's model applied to U.S. polio data.

	estimates				goodness-of-fit		
	trendx10^{-3}	$\tilde{\omega}$	$\tilde{\upsilon}$	$\tilde{\delta}$	SSR	AIC	BIC
NBD	-5.03	0.862	7.287	2.04	419.47	22.04	46.94
Zeger	-4.35	—	—	—	507.89	22.23	47.22

In the above table $\tilde{\delta}$ denotes the estimator of the outlier dummy. The likelihood ratio test for the inclusion of the trend variable is 0.28

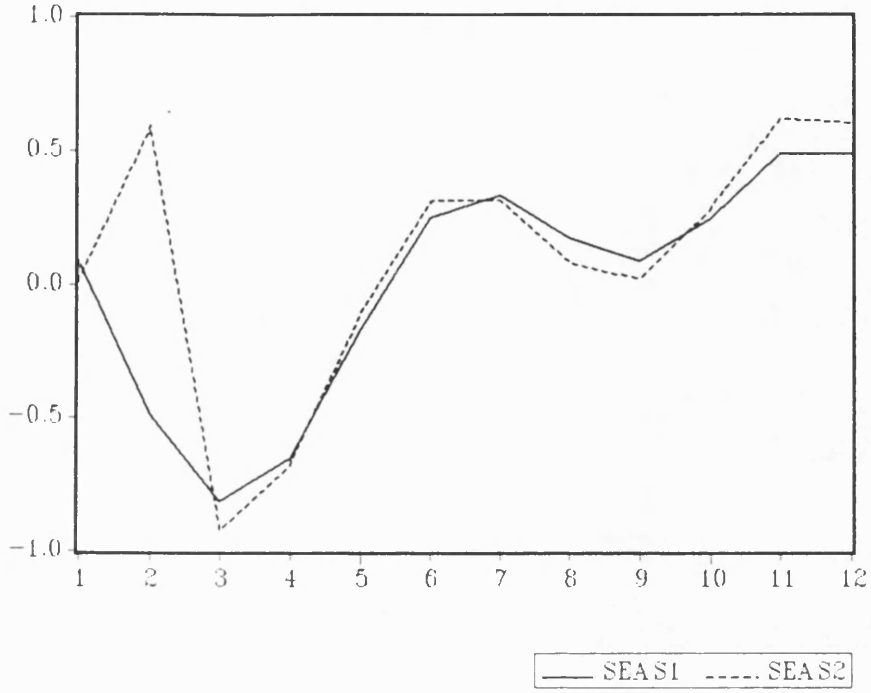
and this is obviously not significant when set against a χ_1^2 distribution, so that according to our model, there is not sufficient evidence in the data to support a long term decrease on the rate of U.S. polio cases. A similar finding has also been reported by Zeger.

Figure 3.5.4 U.S. number of cases of polio and fitted trend using a NBD-Beta model and Zeger's model.



The seasonal component for both models is depicted in figure 3.5.5. From this and the above table one may conclude that both models produce similar fit, although Zeger's model seems to be computationally more expensive since it involves simulation studies in order to determine an appropriate autocorrelation structure needed at the stage of parameter estimation.

Figure 3.5.5 Seasonal component of the NBD-Beta model (seas1) and Zeger's model (seas2) fitted to the U.S. polio series.



APPENDIX

A1- The Post-Sample Predictive Test for the Poisson-Gamma Model

The post-sample predictive test statistic for the Poisson observation model is obtained by introducing ℓ dummy variables into the model at times T+1 to T+ℓ. The statistic $\xi(\ell)$ is obtained by subtracting the log-likelihood obtained without these variables from the log-likelihood with these variables and multiplying by a factor of two.

To find the log-likelihood function for the model with dummy variables in the post sample period, first consider the case of ℓ=1. The log-likelihood function is of the form (3.2.22) with T replaced by T+1. However, the dummy variable parameter, δ, only enters the likelihood via $b_{t/t-1}{}^{+}$, which from (3.2.28b) is

$$b_{T+1/T}{}^{+} = \omega\, b_T\, e^{-\delta}. \qquad \text{(A1.1)}$$

Thus the log-likelihood can be written as

$$\begin{aligned} \log L_{T+1} = \log L^{+}{}_{T+1} + a_{T+1/T} \log b_{T+1/T}{}^{+} \\ - (a_{T+1/T} + y_{T+1}) \log (1 + b_{T+1/T}{}^{+}) \end{aligned} \qquad \text{(A1.2)}$$

where $L^{*}{}_{T+1}$ does not depend on δ. Differentiating (A1.2) with respect to δ yields

$$\exp(-\tilde{\delta}) = a_{T+1/T} / y_{T+1}\, b_{T+1/T}{}^{+} = a_T / y_{T+1}\, b_T \qquad \text{(A1.3)}$$

and so, from (A1.1)

$$\tilde{b}_{T+1/T} = a_{T+1/T} / y_{T+1}.$$

Substituting into (A1.2) gives the log-likelihood concentrated wrt δ. Note that, in the special case when $y_{T+1} = 0$, the last two terms on the right hand side of (A1.2) are zero when taken together and so $\log L_{T+1} = \log L^*_{T+1}$.

Now consider $\ell > 1$. The log-likelihood function with ℓ dummy variables, $\delta_1, \ldots, \delta_\ell$, in the post sample period is

$$\log L_{T+\ell} = \log L^*_{T+\ell} + \sum_{t=T+1}^{T+\ell} a_{t/t-1} \log b_{t/t-1}^{+}$$

$$- \sum_{t=T+1}^{T+\ell} (a_{t/t-1} + y_t) \log(1 + b_{t/t-1}^{+}) \qquad \text{(A1.5)}$$

where $b_{t/t-1}^{+}$ obeys the recursion (3.2.28b) and (3.2.29b). This implies that $b_{T+j/T+j-1}^{+}$ depends on $\delta_1, \ldots, \delta_{j-1}$ for $j = 2, \ldots, \ell$, thereby making differentiation of $\log L_{T+\ell}$ wrt $\delta_1, \ldots, \delta_\ell$ rather tedious. However if we differentiate wrt δ_ℓ first, we obtain a result analogous to (A1.4) namely

$$\tilde{b}_{T+\ell/T+\ell-1} = a_{T+\ell/T+\ell-1} / y_{T+\ell}$$

This is independent of previous values of $b_{t/t-1}^{+}$ and hence of $\delta_1, \ldots, \delta_{\ell-1}$.

Concentrating the log-likelihood with respect to δ_ℓ and proceeding to treat $\delta_1, \ldots, \delta_{\ell-1}$ in the same way, gives the following concentrated log-likelihood function

$$\log L_{T+\ell} = \log L^*_{T+\ell} + \sum_{t=T+1}^{T+\ell} a_{t/t-1} \log (a_{t/t-1} / y_t) -$$

$$- \sum_{t=T+1}^{T+\ell} (a_{t/t-1} + y_t) \log (1+ a_{t/t-1} / y_t). \qquad (A1.7)$$

The log-likelihood function under the null hypothesis, that is without dummy variables, is

$$\log L_{T+\ell} = \log L^*_{T+\ell} + \sum_{t=T+1}^{T+\ell} a_{t/t-1} \log b_{t/t-1}$$

$$- \sum_{t=T+1}^{T+\ell} (a_{t/t-1} + y_t) \log (1+ b_{t/t-1}) \qquad (A1.8)$$

where $b_{t/t-1}$ is computed via the recursion in (3.2.5b) and (3.2.8b). Subtracting (A1.8) from (A1.7) and multiplying by two gives the LR test statistic, (3.4.6). When the model includes the systematic component η_t, the only amendment on the above formula will be the substitution of $b_{t/t-1}$ and b_t by the expressions in (3.2.28b) and (3.2.29b) respectively.

A2- The k-steps ahead mean and variance for the Poisson-Gamma model

To prove (3.2.14a) we apply the chain rule for conditional expectations given in (1.1.9a). Taking the conditional expectation of y_{T+k} at time T+k-1 gives, from (3.2.11a),

$$E_{T+k-1}\,(y_{T+k}) = a_{T+k-1}\ /\ b_{T+k-1}\ .$$

Using (3.2.8a-b), and taking conditional expectations at time T+k-2 gives

$$E_{T+k-2}\,E_{T+k-1}\,(y_{T+k}) = E_{T+k-2}\left[\ \frac{\omega\, a_{T+k-2} + y_{T+k-1}}{\omega\, b_{T+k-2}+1}\ \right]$$
$$= \frac{a_{T+k-2}}{b_{T+k-2}}\qquad ,\ k>2$$

Repeating this procedure by taking conditional expectations at time T+k-3 and so on gives (3.2.14a).

To obtain the k-steps ahead variance, the appropriate way to proceed is to derive the expression for the first factorial moment, which is also obtained by use of the chain rule in (1.1.9a). Once this has been derived we use the standard result that

$$\mathrm{Var}(y_{T+k}|Y_T)= E(y_{T+k}^{(2)}|Y_T)+ E(y_{T+k}|Y_T) - E^{2}(y_{T+k}|Y_T) \qquad \text{(A2.2)}$$

where the first term on the rhs is the second factorial moment given by

$$E(y_{T+k}^{(2)}|Y_T)=E(y_{T+k}(y_{T+k}-1)|Y_T) \qquad \text{(A2.3)}$$
$$=\zeta_k$$

We now derive a recursion for ζ_k from which a closed expression for the variance is derived. It is straightforward to show that

$$E(y_{T+k}^{(2)} | Y_{T+k-1}) = a_{T+k-1} (\omega\, a_{T+k-1} + 1)/ \omega\, b^2_{T+k-1}$$
$$\equiv \zeta_{k-1},$$

from which it follows that

$$E(y_{T+k}^{(2)} | Y_{T+k-2}) = \zeta_{k-2} = \zeta_{k-1} + \frac{a_{T+k-2}\ (1-\omega)}{\omega\, b_{T+k-1}\, b_{T+k-2}} \qquad \text{(A2.4)}$$

Applying (A2.4) recursively and evaluating the expectations of the second term by (3.2.14a) yields, for (A2.3)

$$\zeta_k = \omega\ a_T(\omega a_T+1)/(\omega b_T)^2 + a_T\ ((1-\ \omega)/\omega b_T)\ S_{k-1}$$

where $S_{k-1} = \sum_{j=1}^{k-1} (1/b_{T+j})$. Substituting this back in (A2.3) and using (A2.1) the expression in (3.2.14b) follows.

CHAPTER FOUR

BINOMIAL MODELS

4.1 INTRODUCTION:

In this chapter we focus our attention on formulating models for count data where the total number of counts n_t is assumed fixed and known. When n_t is one, the data are *binary* or *dichotomous*. For n_t different from one the binomial or multinomial distribution can be used. If the probabilities of a positive occurence in these distributions can be considered fixed or dependent on exogenous variables then the GLM of McCullagh and Nelder (1983, chs.4-5) offers the appropriate framework. For those situations which are believed to be changing with time, the state space models of Kitagawa (1987) and WHM (1985) are some of the proposed solutions in the literature (see section 1.3). It is not difficult to find data suited for these models. For example, the daily occurrences of rainfall over 1mm in Tokyo (Kitagawa 1987), the weekly counts of the number of people who provide a positive response to advertising of a popular chocolate bar (WHM 1985), the number of wins for Cambridge in the university boat race between Oxford and Cambridge and so forth.

Our class of structural 'Binomial' models is based on the binomial, Bernoulli, and multinomial distributions. The development follows the line of the Poisson-Gamma model of Chapter 3, although it is mathematically more elaborate at the stage of explanatory variables introduction.

4.2 THE BINOMIAL-BETA MODEL:

- *measurement equation*: if the observations at time t are generated from a binomial distribution then

$$p(y_t|\pi_t,n_t) = \begin{bmatrix} n_t \\ y_t \end{bmatrix} \pi_t^{y_t} (1-\pi_t)^{n_t-y_t}, \qquad y_t=0,1,2,\dots,n_t \tag{4.2.1}$$

where π_t is the probability that y_t is unity when n_t is one. The value of n_t is assumed to be fixed and known. Thus observations from the binomial can be regarded as a special case of count data where there is a fixed number of opportunities for the event in question to occur. It is a standard result that

$$E(y_t|\pi_t,n_t) = n_t\,\pi_t$$
$$Var(y_t|\pi_t,n_t)= (1-\pi_t)E(y_t|\pi_t,n_t).$$

Note that here the state θ_t is the probability of a positive occurence, π_t.

(i) *state prediction*- the conjugate prior for the binomial distribution is the beta distribution as in (3.3.3). Let $p(\pi_{t-1}|Y_{t-1})$ have a beta distribution with parameters a_{t-1} and b_{t-1}. Assume that $p(\pi_t|Y_{t-1})$ is also beta with parameters given by equations exactly the same as those in (3.2.5a-b). This again ensures that the mean of $\pi_t|Y_{t-1}$ is the same as that of $\pi_{t-1}|Y_{t-1}$ but the variance increases. Specifically

$$E(\pi_t|Y_{t-1}) = a_{t/t-1} / (a_{t/t-1} + b_{t/t-1}) = a_{t-1} / (a_{t-1} + b_{t-1})$$

and

$$\begin{aligned} Var(\pi_t|Y_{t-1}) &= \frac{a_{t/t-1}\, b_{t/t-1}}{(a_{t/t-1} + b_{t/t-1})^2\,(a_{t/t-1} + b_{t/t-1} + 1)} \\ &= \frac{a\, b}{(a + b)^2(\omega\, a + \omega\, b + 1)} \end{aligned}$$

where $a=a_{t-1}$ and $b=b_{t-1}$.

(ii) *state updating*- once the t-*th* observation becomes available, the distribution of $\pi_t|Y_t$ is beta with parameters

$$a_t = a_{t/t-1} + y_t \qquad (4.2.2a)$$

$$b_t = b_{t/t-1} + (n_t - y_t) \qquad (4.2.2b).$$

(iii) *conditional distribution*- the predictive distribution, $p(y_t|n_t,Y_{t-1})$ is obtained by solving the compounding operation in

(1.1.4). This produces a beta-binomial distribution given by

$$p(y_t | n_t, Y_{t-1}) = \begin{bmatrix} n_t \\ y_t \end{bmatrix} \frac{B(a_{t/t-1} + y_t, b_{t/t-1} + n_t - y_t)}{B(a_{t/t-1}, b_{t/t-1})} \qquad (4.2.3)$$

where B(·) is the beta function. The mean and the variance of this distribution are:

$$\begin{aligned} \tilde{y}_{t/t-1} &= E(y_t | n_t, Y_{t-1}) = n_t \, a_{t/t-1}/(a_{t/t-1} + b_{t/t-1}) \\ &= n_t \, a_{t-1}/\, (a_{t-1}+b_{t-1}) \qquad (4.2.4a) \end{aligned}$$

$$\begin{aligned} \text{Var}\,(y_t | n_t, Y_{t-1}) &= \frac{n_t \, a_{t/t-1} \, b_{t/t-1} \, (a_{t/t-1} + b_{t/t-1} + n_t)}{(a_{t/t-1} + b_{t/t-1})^2 (a_{t/t-1} + b_{t/t-1} + 1)} \\ &= \frac{n_t \, a_{t-1} \, b_{t-1} \, (a_{t-1}+b_{t-1}+ \omega^{-1}\}}{(a_{t-1}+b_{t-1})^2 (a_{t-1}+b_{t-1}+ \omega^{-1})} \qquad (4.2.4b) \end{aligned}$$

(iv) *k-steps ahead forecasting*- by substituting repeatedly from the recursive equations (3.2.5a-b) and (4.2.2a-b) it can be seen that, for n_{T+k} constant, $\tilde{y}_{T+k/T}$, the predictor k-steps ahead, may be expressed as an EWMA scheme having the form

$$\tilde{y}_{T+k/T} = n_{T+k} \, \text{EWMA}(y)/\, \text{EWMA}(n). \qquad (4.2.5)$$

As before we face problems in computing variances and the forecasting distribution for $k \geq 3$. Here we only derive the first order factorial moment along with the distribution for two steps ahead. If we let

$n_1 = n_{T+1}$, $n_2 = n_{T+2}$, $a = a_T$, $b = b_T$, $h = \omega(a+b)$ then by use of the chain rule in (1.1.9a) it is possible to show that

$$E(y_{T+2}{}^{(2)} \mid Y_T) = \frac{n_2(n_2 - 1)}{n_1(n_1 - 1)} E(y_{T+1}{}^{(2)} \mid Y_T) + k(a,\omega,n_1) \qquad (4.2.6)$$

where

$$E(y_{T+1}{}^{(2)} \mid Y_T) = \frac{n_1(n_1-1)\ a\omega\ (a\omega+1)}{h\ (h+1)}$$

and

$$k(a,\omega,n) = \frac{n_2\ (n_2-1)}{h\ (\ h+1)} \left[h(1+\omega^2-\omega) + \frac{\omega^2(1+n_1 a)}{[\omega(h+n_1)][\omega(h+n_1)+1]} \right].$$

By use of (3.2.17) and the expression for the predictive density in (4.2.3) one may also show that

$$p(y_{T+2} \mid Y_T) = k_1(y_2) \sum_{y_1=0}^{n_1} \frac{\Gamma(\omega A_1+y_2)\ \Gamma(A_1)\ \Gamma(\omega B_1+n_2-y_2)\ \Gamma(B_1)}{y_1!\ (n_1-y_1)!} \qquad (4.2.7)$$

where

$$k_1(y_2) = \frac{n_2!\ n_1!\ \ \Gamma(h)\ \ \Gamma[\omega(h+n_1)]}{y_2!\ (n_2-y_2)!\ \Gamma(\omega a)\ \Gamma(\omega b)\ \Gamma[\omega(h+n_1)+n_2]}$$

$A_1 = \omega a + y_1$ and $B_1 = \omega b + (n_1 - y_1)$.

4.2.1 Likelihood

The likelihood function is again obtained by (1.2.13) with τ defined as the first time period for which

$$0 < \sum_{t=1}^{\tau} y_t < \sum_{t=1}^{\tau} n_t.$$

This condition ensures that a_τ and b_τ are strictly positive, although again there is nothing to prevent us starting the recursions (3.2.5a-b) and (4.2.2a-b) at $t = 1$ with $a_0 = b_0 = 0$; see the comments in section 3.3.1. Using (4.2.3) it may be shown that the kernel of the conditional log-likelihood has the form

$$\log L(\omega|n_t) = \sum_{t=\tau+1}^{T} \{\log[\,\Gamma(a_{t/t-1}+y_t)/\Gamma(a_{t/t-1})]+$$

$$+\log[\,\Gamma(b_{t/t-1}+v_t)/\Gamma(b_{t/t-1})]-\log[\,\Gamma(d_{t/t-1}+n_t)/\Gamma(d_{t/t-1})]$$

where $v_t = n_t - y_t$ and $d_{t/t-1} = b_{t/t-1} + a_{t/t-1}$. As before the logarithms of the ratios of gamma functions may be easily evaluated using (3.2.23).

4.2.2 Binary Data

Binary data is easily handled by the Binomial-Beta model by setting $n_t=1$ where appropriate. It is then easy to show that the predictive distribution in (4.2.3) reduces to a binomial $(1,a_{t/t-1}/(a_{t/t-1}+b_{t/t-1}))$. From this the one-step ahead predictions may be established by considering the conditional probability of a positive ocurrence,i.e.,

$$\Pr(y_t=1|Y_{t-1})= E(\pi_t|Y_{t-1})= a_{t/t-1}/ (a_{t/t-1} +b_{t/t-1})$$

with

$$\Pr(y_t=0|Y_{t-1})= 1- \Pr(y_t=1|Y_{t-1})$$

where $y_t=1$ is the event being predicted. The log-likelihood in (4.2.5) may be then simplified to

$$\log L(\omega)= \sum_{t=\tau+1}^{T} y_t \log(a_{t/t-1}/b_{t/t-1})+ \log[b_{t/t-1} / (a_{t/t-1}+b_{t/t-1})]$$

As regards multi-step ahead forecasts, it can be shown, by evaluating (3.2.17), that

$$p(y_{T+k}|Y_T)= a_T/(a_T+b_T) \text{ for } k=1,2,...$$

This should be no surprise, given that in the present context, the distribution of y_{T+k} conditional on $\mathbf{Y_T}$ coincides with the forecast function, and this by construction, is time invariant in the absence of explanatory variables/structural components (see 4.2.5).

4.2.3 Polytomous Data

When there are more than two categories, the observations are said to be *polytomous* and the multinomial distribution is appropriate. Let there be m possible categories, and suppose that the probability that, at time t, an object belongs to the i-th category is π_{it}. If there are n_t trials and the number of objects in the i-th category is y_{it}, then the *measurement equation* is given by

$$p(y_{1t},\dots,y_{mt}) = \begin{bmatrix} n_t \\ y_{it},\dots,y_{mt} \end{bmatrix} \prod_{i=1}^{m} \pi_{it}^{y_{it}} \qquad (4.2.8)$$

with

$$\sum_{i=1}^{m} y_{it} = n_t \qquad \text{and} \qquad \sum_{i=1}^{m} \pi_{it} = 1.$$

The conjugate prior for the multinomial distribution is the multivariate beta or *Dirichlet* distribution

$$p(\pi_1,\dots, \pi_m;\ a_1,\dots, a_m) = \frac{\Gamma\left(\sum_{i=1}^{m} a_i\right)}{\prod_{i=1}^{m} \Gamma(a_i)} \prod_{i=1}^{m} \pi_i^{a_i - 1}$$

where we have dropped the time index for ease of notation. When $m = 2$ this collapses to the beta distribution with $a_1 = a$ and $a_2 = b$. Proceeding as in the previous section, it is not difficult to show that the recursive equations corresponding to (3.2.3) and (4.2.2) become

$$a_{i,t/t-1} = \omega\, a_{i,t-1} \quad , \qquad (4.2.9a)$$

$$a_{i,t} = a_{i,t/t-1} + y_{it} \quad , \quad i = 1,\dots,m \qquad (4.2.9b).$$

The likelihood for ω is as in (4.2.5) with τ the first value of t which yields $a_{i,t} > 0$ for all $i = 1,..,m$. The predictive distribution in this case is known as the Dirichlet-multinomial. The forecasts can again be expressed in terms of EWMA's.

4.2.4 Explanatory Variables and Structural Components

Here we restrict our discussion to the binomial distribution measurement model (m=2 in 4.2.8). Multinomial models may be handled by extending the techniques here discussed. See Ord, Fernandes and Harvey (1989).

In order to investigate the relationship between explanatory variables/structural components affecting the probability π_t we define π_t^+ as the probability of a positive occurence when such effects are present. Following the line of the Poisson specification one must choose a suitable link function $g(\cdot)$ such that $\pi_t^+ = g(\pi_t, z_t'\delta)\cdot$ It is then obvious that a consistent link function should map the unit interval (0,1) onto the real line $(-\infty,+\infty)$. Here we chose the *logit* link function , which takes the form

$$\text{logit}(\pi_t^+)= \log[\pi_t^+/(1-\pi_t^+)]= \text{logit}(\pi_t)+ z_t'\delta \qquad (4.2.10a)$$

or

$$\pi^+ = \pi\, u/(1-\pi+\pi u) \qquad (4.2.10b)$$

where $u= \exp(z_t'\delta)$ and the subscripts are to be understood from the context. We note that

$$\begin{array}{lll} \pi= \pi^+ & \text{for } u= 1\ , & \\ \pi < \pi^+ < 1 & \text{for } u > 1 & \text{and} \\ 0 < \pi^+ < \pi & \text{for } u < 1. & \end{array} \qquad (4.2.11)$$

Using (4.2.11) one may express the *measurement* equation (4.2.1) as

$$p^+(y_t|\pi_t,n_t)= \begin{bmatrix} n \\ y \end{bmatrix} \pi^{+y}\ (1-\pi^+)^{\ n-y}$$

$$= p(y_t|\pi_t,n_t) \frac{u^y}{[1-\pi(1-u)]^n} \qquad (4.2.12)$$

where $p(y_t|\pi_t,n_t)$ is the standard binomial distribution as in (4.2.1). Unlike the specifications so far studied, the binomial model with explanatory variables produces a non-standard measurement equation, which may be looked at as a 'perturbed' version of its standard form. They obviously coincide when $u_t=1$.

We now look in detail at the solutions adopted in order to solve the problems created by loss of conjugacy. In Ord, Fernandes and Harvey (1989) we make a brief introduction to the adopted techniques which are either based on the hypergeometric series or on a modal approximation. They are looked at in detail in the next section.

Series approximation:

First we consider the predictive distribution and its first moments. The former is obtained by compounding the distribution in (4.2.12) with the beta density in (3.3.3) and this results in

$$p^+(y_t|n_t,Y_{t-1})= \frac{u^y}{B(a,b)} \int_0^1 \begin{bmatrix} n \\ y \end{bmatrix} \frac{\pi^{y+a-1}(1-\pi)^{b+n-y-1}}{[1-\pi(1-u)]^n} d\pi \qquad (4.2.13a)$$

where $a=a_{t/t-1}$ and $b=b_{t/t-1}$. Now the expansion of the term in brackets in the denominator through a binomial series produces

$$\frac{1}{[1-\pi(1-u)]^{n}} = \sum_{j=0}^{\infty} \begin{bmatrix} n+j-1 \\ j \end{bmatrix} \pi^{j} (1-u)^{j} \qquad (4.2.14)$$

which is a convergent series for $0 < u < 2$. If $u > 2$ one can redefine the logit link in terms of $(1-\pi)$ instead, and convergence will be guaranteed. Substituting (4.2.14) back into (4.2.13a) and integrating one obtains

$$p^{+}(y_t \mid n_t, Y_{t-1}) = \begin{bmatrix} n \\ y \end{bmatrix} \frac{u^{y}}{B(a,b)} \sum_{j=0}^{\infty} \begin{bmatrix} n+j-1 \\ j \end{bmatrix} (1-u)^{j} B(a+y+j, b+n-y)$$

$$(4.2.13b)$$

Constants apart the sum in the above expression is the hypergeometric series ${}_2F_1(n,a+y;a+b+n;1-u)$ (see,e.g., Abramowitz and Stegun 1965, p.556). The ratio of the (n+1)st term to the nth term on this expansion may be shown to be

$$A_{j+1} \,/\, A_j = \frac{(a+y+j)\,(n+j)\,(1-u)}{(a+b+j+n)\,(j+1)} \qquad (4.2.15)$$

so that the predictive density takes the final form

$$p^{+}(y_t \mid n_t, Y_{t-1}) = \begin{bmatrix} n_t \\ y_t \end{bmatrix} \frac{u^{y}\; B(a+y, n+b-y)}{B(a,b)} \sum_{j=0}^{\infty} A_j$$

$$= p(y_t \mid n_t, Y_{t-1})^{y} u \sum_{j=0}^{\infty} A_j \qquad (4.2.13c)$$

with $A_0=1$ and $p(y_t \mid n_t, Y_{t-1})$is the standard predictive distribution as in (4.2.3). Convergence criteria for the above series can be established by choosing ϵ, $\epsilon > 0$ for which $|A_{j+1} - A_j| < \epsilon$. As expected this distribution reduces to the beta-binomial when $u_t=1$.

Our next step will be the evaluation of its first two conditional moments, which may then be used for predictive purposes.

Using the definition of the mean for a discrete random variable and the following recursive relation for binomial coefficients

$$\begin{bmatrix} n \\ y \end{bmatrix} = \frac{n\,(n-1)\ldots(n-k+1)}{y\,(y-1)\ldots(y-k+1)} \begin{bmatrix} n-k \\ y-k \end{bmatrix} \qquad (4.2.16)$$

it is not difficult to see that

$$E^{+}(y_t \mid Y_{t-1}) = \frac{n\,u}{B(a,b)} \sum_{y=1}^{n} \begin{bmatrix} n-1 \\ y-1 \end{bmatrix} \frac{\pi^{\,a+y-1}\,(1-\pi)^{b+n-y-1}\,u^{y-1} d\pi}{[1-\pi(1-u)]^{\,n}} \,. \qquad (4.2.17a)$$

Now using (4.2.14a) this may be shown to take the form

$$E^{+}(y_t \mid Y_{t-1}) = \frac{n\,u}{B(a,b)} \; E\,[\;\pi \,/\, [1-\pi(1-u)] \mid Y_{t-1}\;] \qquad (4.2.17b)$$

where condition in n should be understood. Finally by use of the binomial theorem the argument on the above expectation can be expressed as an convergent series and after integrating wrt π we obtain the following expression

$$E^{+}(y_t \mid Y_{t-1}) = \frac{n\quad u}{B(a,b)} \sum_{j=0}^{\infty} (1-u)^{j} \;\; B(a+j+1,b).$$

This infinite sum may be recognized as the hypergeometric series ${}_2F_1(a+1,1;a+b+1;1-u)$ which follows the recursion

$$B_{j+1} / B_j = \frac{(a+j+1)\ (1-u)}{(a+b+j+1)} \quad , \; B_0=1$$

so that the predictive mean has the final form

$$E^+(y_t \mid Y_{t-1}) = \frac{n\ a\ u}{(a+b)} \quad \sum_{j=0}^{\infty} B_{j+1} \ . \qquad (4.2.17c)$$

In the present context the natural way of evaluating the variance is by first computing the first factorial moment, since its calculation will be made easier by use of the relation in (4.2.16). It is not difficult to show that on the lines of the previous derivation one may find that

$$E^+[y_t(y_t-1) \mid Y_{t-1}] = \frac{n(n-1)\ u^2}{B(a,b)} \quad \sum_{j=0}^{\infty} (j+1)\ (1-u)^j\ B(a+j+2,b). \qquad (4.2.18a)$$

As expected the above sum, constants apart is also an hypergeometric series given by ${}_2F_1(a+2,2,a+b+2,1-u)$, where the ratio of the $(n+1)^{st}$ term to the n^{th} is given by

$$C_{j+1} / C_j = \frac{(j+2)\ (j+a+2)\ (1-u)}{(j+1)\ (a+b+j+2)} \quad , \; C_0=1.$$

Using the above one may write the final expression for the first factorial moment as

$$E^+[y_t(y_t-1)] = \frac{n(n-1)\; u^2\; a(a+1)}{(a+b)\;(a+b+1)} \sum_{j=0}^{\infty} C_{j+1} \qquad (4.2.19b)$$

with n>1. From this one may then easily evaluate the predictive variance.

In order not to disturb the property of conjugacy we have assumed that the exact posterior is approximated by a *virtual* beta density, whose parameters are obtained by matching its moments with those of the exact posterior density, not as yet derived. This artifice will enable us to obtain the updating equations in a straightforward manner, keeping the model tractable as before.

The exact posterior is obtained by making use of Bayes' theorem (1.1.4) and this produces

$$p^+(\pi_t|Y_t) = \begin{bmatrix} n \\ y \end{bmatrix} \frac{k\, u^{y}\; \pi^{y+a-1}\; (1-\pi)^{n-y+b-1}}{B(a,b)\;\; [1-\pi(1-u)]^{n}} \; . \qquad (4.2.20a)$$

This also may be put in a 'perturbed' version having the form

$$p^+(\pi_t|Y_t) = \begin{bmatrix} n \\ y \end{bmatrix} \frac{k\, u^{y}\;\; B(a+y-1,n-y+b-1)}{B(a,b)\;\; [1-\pi(1-u)]^{n}}\; p(\pi_t|Y_t) \qquad (4.2.20b)$$

where k, the normalization constant, is given by 1 / $p+(y_t|Y_{t-1})$ (see 4.2.14c) and $p(\pi_t|Y_t)$ is the standard beta posterior with parameters as in (4.2.3). On the lines of the previous derivations it is possible to demonstrate that the posterior mean may be written as

$$E^+(\pi_t|Y_t) = \mu_t = \begin{bmatrix} n \\ y \end{bmatrix} \frac{k\, u^{y}}{B(a,b)} \sum_{j=0}^{\infty} \begin{bmatrix} n+j-1 \\ j \end{bmatrix} (1-u)^{j} B(a+y+j+1,n-y+b)$$

$$= \begin{bmatrix} n \\ y \end{bmatrix} \frac{k\, u^{y}\;\; B(a+y+1,n-b+y)}{B(a,b)} \sum_{j=0}^{\infty} D_{j+1} \qquad (4.2.21a)$$

where

$$D_{j+1}/D_j = \frac{(a+y+j+1)\ (n+j)\ (1-u)}{(a+b+j+1+n)\ \cdot(1+j)} , \quad D_0=1.$$

Finally subsitituting k by (4.2.14c) in (4.2.21a) the posterior mean simplifies to

$$E^+(\pi_t|Y_t) = \frac{(a+y)}{(a+b+n)} \sum_{j=0}^{\infty} D_{j+1} \Big/ \sum_{j=0}^{\infty} A_{j+1} \qquad (4.2.21b)$$

where the A's are as in (4.2.16).

Under the lines of the previous derivation one may also show that the posterior second raw moment is given by

$$E^+(\pi_t^2|Y_t) = m_t = \begin{bmatrix} n \\ y \end{bmatrix} \frac{k\ u^y}{B(a,b)} \sum_{j=0}^{\infty} \begin{bmatrix} n+j-1 \\ j \end{bmatrix} (1-u)^j B(y+a+j+2, n-y+b). \qquad (4.2.22a)$$

where again the sum may be expressed as an hypergeometric series apart from constants. This is given by ${}_2F_1(s,a+2,a+b+2;1-u)$ with ratio between consecutive terms having the form

$$F_{j+1} / F_j = \frac{(a+y+j+2)\ (n+j)\ (1-u)}{(a+b+j+2+n)\ (j+1)} , \quad F_0=1.$$

Finally by substituing k in (4.2.22a) one obtains that

$$E^+(\pi_t^2|Y_t) = \frac{(a+y)\ (a+y+1)}{(a+b+n)\ (a+b+n+1)} \sum_{j=0}^{\infty} F_{j+1} \Big/ \sum_{j=0}^{\infty} A_{j+1} \qquad (4.2.22b)$$

where the A's are as in (4.2.16). By equating the first two central moments of the *virtual* beta posterior with the correspondent moments of the true posterior, the following expressions are obtained for the *updating* equations of our model

$$a_t = \mu_t\ (\mu_t - m_t)\ /\ (m_t - \mu^2) \qquad (4.2.23a)$$

$$b_t = a_t\ (1 - \mu_t)\ /\ \mu_t \qquad (4.2.23b)$$

where μ_t and m_t are given by (4.2.21b) and (4.2.22b) respectively.

Modal approximation:

The series expansion approach may become tedious for polytomous data or u near zero. A more parsimonious approach may be given by a *modal* approximation. This technique is based on a very simple idea: the replacement of y_t in (4.2.1) by a *virtual* variable w_t, which, by construction, should increase when effects that also increase the probability of a positive event occur. If the overall sum is kept constant during this process, the net effect will be a 'reallocation' of cases to the positive event. It is obvious that the link between these effects and the virtual variable should be made by the logit function (4.2.10b). The question that remains is how to select a mechanism that implements this allocation in a proper manner.

We start by rewriting the observation model in terms of w_t

$$p(w_t | \pi_t, n_t) \propto \pi_t^{\,w_t}\ (1 - \pi_t)^{\,n_t - w_t} \qquad (4.2.24)$$

Now consider both this distribution and the observation equation in π_t^+ (4.2.14a) as functions of π_t and π_t^+ respectively. The idea is to select w_t so that the mode of (4.2.24), $\pi_m = w_t/n_t$ agrees with the mode of (4.2.14a), $\pi_m^+ = y_t/n_t$ wrt to the logit link function (4.2.10b). Using this criterion one can easily show that the *virtual* variable w_t is linked to the explanatory variables and the actual variable y_t through the relation

$$w_t = \frac{n_t\, y_t}{u_t\, n_t + y_t\, (1-u_t)} \tag{4.2.25}$$

Note that

$$\begin{array}{ll} w_t = y_t & \text{if } u = 1, \\ 0 < w_t < y_t & \text{if } u < 1 \quad \text{and} \\ y_t < w_t < n_t & \text{if } u > 1. \end{array} \tag{4.2.26}$$

The net effect of this approximation is the reallocation of 'observations' to the event of interest, with the overall sum kept constant.

Given that the structure of π_t in the measurement equation (4.2.24) is left intact, conjugacy is preserved, regardless of the fact of the non-integer nature of w_t. The predictive equations remain unaltered and the only change on the updating equations is the replacement of y_t by w_t in (4.2.2). With regard to the predictive distribution a word of caution is necessary. Although the formulae for the predictive moments are only affected by the replacement of y_t by w_t (see 4.2.4), since the observation model has been defined in a kernel form, in order to obtain a proper predictive distribution, one has to evaluate a normalization constant. After taking this into account one obtains

$$p^{+}(y_t \mid n_t, Y_{t-1}) = \begin{bmatrix} n_t \\ y_t \end{bmatrix} \quad k^{-1} \quad B(w_t + a_{t/t-1},\ b_{t/t-1} + n_t - w_t) \tag{4.2.27}$$

where

$$k = \sum_{y_t=0}^{n_t} \begin{bmatrix} n_t \\ y_t \end{bmatrix} \quad B(w_t + a_{t/t-1}, b_{t/t-1} + n_t - w_t).$$

The distribution in (4.2.27) obviously reduces to the standard case (4.2.3) when u_t=1. Given that w_t is non-integer the routine for the logarithm of the gamma function in Press *et al* (1986, p.157) is used.

CHAPTER FIVE

BIVARIATE COUNT DATA MODEL

5.1 INTRODUCTION:

In this chapter we consider a bivariate extension of our class of count data models. The model we set up is based on the total number of events recorded in each period which is assumed to follow a Poisson distribution. The split into the individual series is then determined by a binomial distribution. Both of these mechanisms may be made dynamic in the way suggested in the previous chapters. Combining the predictive distributions for each mechanism leads to a joint predictive distribution for the series, from which predictions may be made and a likelihood function constructed. The development makes use of the results previously derived for the Poisson-Gamma (section 3.2) and the Binomial-Beta models (section 4.2). Alternative methods of forming bivariate distributions for count data are shown in Stein & Juritz (1987) and Johnson and Kotz (1969, pp. 297-300). Of importance in these developments is the resulting correlation structure. It seems that a restricted range for the correlation parameter is the rule rather than the exception for bivariate models. Our approach offers some advantage in this particular point, given that it allows for the existence of both negative and positive correlation. The multivariate extension of this class of models is considered in Ord, Fernandes and Harvey (1989).

To illustrate the bivariate model we consider its application to the goals series introduced in Chapter 3, but now the goals of both teams, England and Scotland, are jointly modelled.

5.2 THE BIVARIATE COUNT DATA MODEL:

- *measurement equation*- suppose we have two series of count data observations y_{1t} and y_{2t}, $t = 1,\dots,T$. Assume that each of the individual series follow a Poisson distribution as in the univariate case (see 3.2.1), i.e.,

$$p(y_{it}|\theta_{it}) = \theta_{it}^{y_{it}} e^{-\theta_{it}} / y_{it}! \quad , \quad i=1,2 \qquad (5.2.1)$$

with the individual rates θ_{it} obeying the relation

$$\theta_{it} = \pi_{it}\,\theta_t \qquad , \text{ where} \qquad (5.2.2a)$$

$$\pi_{1t} + \pi_{2t} = 1 \qquad 0 < \pi_{it} < 1 \quad i=1,2 \qquad (5.2.2b)$$

where θ_t is the *overall* rate, not as yet explained. First some notation; define the bivariate vector $w_t = (y_{1t}, y_{2t})$. Given that the series are independent of each other, conditional on the knowledge of its respective rates, the bivariate *measurement* equation may be factorized as follows

$$p(w_t|\theta_{1t},\theta_{2t}) = \prod_{i=1}^{2} p_i(y_{it}|\theta_{it}) \qquad (5.2.3)$$

where $p_i(\cdot)$ denotes the Poisson probabilities for y_{it} given in (5.2.1). We now consider the aggregate over the series, S_t,

$$S_t = y_{1t} + y_{2t} \qquad t = 1,\ldots,T \qquad (5.2.4)$$

From a standard property of the Poisson distribution the overall sum in (5.2.1) is Poisson distributed with *overall* rate θ_t, i.e.,

$$p(S_t|\theta_t)= \theta_t^{S_t} e^{-\theta_t} / S_t! \qquad (5.2.5)$$

Hence the π's may be interpreted as the individual shares associated with each of the series. Using (5.2.1-3) with another standard statistical result one can show that conditional on the overall sum S_t, y_{1t} is binomially distributed, i.e.,

$$p(y_{1t}|S_t,\pi_{1t}) = \begin{bmatrix} S_t \\ y_{1t} \end{bmatrix} \pi_{1t}^{y_{1t}} (1-\pi_{1t})^{S_t-y_{1t}}. \qquad (5.2.6)$$

We are now able to derive the form for our bivariate *measurement* equation. Using (5.2.1-3) one may easily show that this distribution may be expressed by the product

$$p(w_t|S_t,\theta_t,\pi_{1t}) = p(S_t|\theta_t)\; p(y_{1t}|S_t,\pi_{1t}) \qquad (5.2.7)$$

where the distributions in the rhs are given respectively by (5.2.5) and (5.2.6). In order to formulate a dynamic model we now have to consider stochastic mechanisms for the evolution of the overall rate θ_t and the individual share $\pi_{1,t}$, the *states* of our bivariate model. Observe that each of the distributions in (5.2.7) are *measurement equations* for which a stochastic mechanism has already been established. See sections 3.2 and 4.2. Our strategy will then be heavily based on the results derived for these univariate formulations.

(i) *state prediction*- let $\mathbf{Y_t} = \{\mathbf{Y_{1t}}, \mathbf{Y_{2t}}\}$ where $\mathbf{Y_{it}} = \{y_{i1}, y_{i2}, \ldots, y_{it}\}$ i=1,2, and $\Theta \subset \mathbf{R}$ and $\Pi \subset \mathbf{R}$ be the parameter space for the overall rate and proportions respectively.

i-a. <u>the overall rate</u>: the obvious way to proceed is to assume a gamma prior as in (3.2.2), i.e.,

$$p(\theta_{t-1}|\mathbf{Y_{t-1}}) \sim \text{gamma}(a_{t-1}, b_{t-1}) \qquad (5.2.8)$$

given that S_t is Poisson distributed. Using the same argument developed for the univariate case, $\theta_t|\mathbf{Y_{t-1}} \sim \text{gamma}(a_{t/t-1}, b_{t/t-1})$ where

$$a_{t/t-1} = \omega_1 a_{t-1} \qquad (5.2.9a)$$

$$b_{t/t-1} = \omega_1 b_{t-1} \qquad (5.2.9b)$$

with $0< \omega_1 <1$.

i-b. the individual shares: given the constraint in (5.2.2b) we only need to consider the dynamics for one of the shares, say $\pi_t \equiv \pi_{1t}$. Since y_{1t}, conditional on S_t, is binomially distributed, we adopt the usual beta prior

$$p(\pi_{t-1} | Y_{t-1}) \sim \text{beta}(c_{t-1}, d_{t-1}) \qquad (5.2.10)$$

Following the lines of the univariate Binomial-Beta model (section 4.2) one has that $\pi_t | Y_{t-1} \sim \text{beta}(c_{t/t-1}, d_{t/t-1})$ with

$$c_{t/t-1} = \omega_2 \, c_{t-1} \qquad (5.2.11a)$$

$$d_{t/t-1} = \omega_2 \, d_{t-1} \qquad (5.2.11b)$$

where $0< \omega_2 < 1$.

(ii) *state updating*- as before the results that follow are based on the Poisson and Binomial univariate models.

ii-a. the overall rate: based on results of section 3.2 we know that the posterior for θ_t, $p(\theta_t | Y_t)$, will also be gamma with parameters

$$a_t = a_{t/t-1} + S_t \qquad (5.2.12a)$$

$$b_t = b_{t/t-1} + 1 \qquad (5.2.12b)$$

ii-b. the individual shares: given conjugacy of the pair binomial-beta the posterior for π_t, $p(\pi_t | Y_t)$ will be beta distributed with parameters

$$c_t = c_{t/t-1} + y_{1,t} \qquad (5.2.13a)$$

$$d_t = d_{t/t-1} + (S_t - y_{1t}) \qquad (5.2.13b)$$

where for the bivariate case $S_t - y_{1t} = y_{2t}$.

(iii) *conditional distribution*- in order to derive the joint predictive distribution, conditional on the overall sum, we first need to establish an equivalent bivariate version of the *compounding* operation (see 1.1.6). This gives

$$p(w_t | Y_{t-1}) = \int_\Theta \int_\Pi p(w_t | S_t, \pi_t, \theta_t) \; p(\pi_t, \theta_t | Y_{t-1}) d\theta_t \, d\pi_t. \qquad (5.2.14a)$$

The second density on the rhs of the above expression is the state joint density. Assuming that the overall rate and the individual share are independent processes, the following decomposition holds

$$p(\pi_t, \theta_t | Y_{t-1}) = p(\pi_t | Y_{t-1}) \; p(\theta_t | S_{t-1}) \qquad (5.2.15)$$

where the individual densities are respectively the beta prior and the gamma prior given in (5.2.9) and (5.2.11). Using the above factorization together with (5.2.7) one may easily show that

$$p(w_t|Y_{t-1}) = \int_{\Theta} p(S_t|\theta_t)p(\theta_t|S_{t-1})d\theta_t \int_{\Pi} p(y_{1t}|\pi_t)p(\pi_t|Y_{t-1})d\pi_t$$

$$= p(S_t|S_{t-1})\ p(y_{1t}|S_t,Y_{t-1}) \qquad (5.2.14b)$$

which is the product of the NBD (see 3.2.7) and the Binomial Beta (see 4.2.23) predictive distributions. Note that an equally valid way of deriving the above equation would be by expressing the *measurement* equation in terms of the individual rates $\{\theta_{it}\}$ instead of the *overall* rate θ_t. The joint predictive distribution is then obtained by *compounding* this distribution with the joint density for the rates, not as yet derived, giving

$$p(w_t|Y_{t-1}) = \int_{\Theta}\int_{\Theta} p(w_t|\theta_{1t},\theta_{2t})\ p(\theta_{1t},\theta_{2t}|Y_{t-1})d\theta_{1t}\ d\theta_{2t} \qquad (5.2.16)$$

Our derivation for the joint predictive distribution will be based on (5.2.14b). Substituting the necessary probability functions in (5.2.14b) we arrive at the following expression

$$p(w_t|Y_{t-1};\omega_1,\omega_2) = \frac{\Gamma(a+S_t)}{\Gamma(a)}\ \frac{\Gamma(c+y_{1t})}{\Gamma(c)}\ \frac{\Gamma(d+y_{2t})}{\Gamma(d)}\ .$$

$$.\ \left[\frac{\Gamma(c+d+S_t)}{\Gamma(c+d)}\right]^{-1} \frac{b^a \quad (1+b)^{-(a+S_t)}}{y_{1t}!\ \ y_{2t}!} \qquad (5.2.17)$$

where $a=a_{t/t-1}$, $b=b_{t/t-1}$, $c=c_{t/t-1}$ and $d=d_{t/t-1}$. As before the terms involving ratios of gamma functions are easily evaluated using (3.2.23).

Before evaluating the moments for the joint predictive distribution we investigate in some detail the joint density of the rates. It is not difficult to see that by making use of the (5.2.2a-b) and (5.2.15) one arrives at

$$p(\theta_{1t},\theta_{2t}|Y_{t-1}) = p(\mu_t|Y_{t-1})\Big|_{\mu_t=\mu_t^*} \; p(\pi_t|Y_{t-1})\Big|_{\pi_t=\pi_t^*} \; |J| \qquad (5.2.18)$$

where the first density on the rhs is the gamma prior for μ_t given in (5.2.9), the second density is the beta prior for π_t given in (5.2.11), $\mu_t^* = \theta_{1t}+\theta_{2t}$, $\pi_t^* = \theta_{1t}/(\theta_{1t}+\theta_{2t})$ and $|J|$ is the determinant of the Jacobian of the transformation. Proceeding with the necessary evaluations the bivariate density may be shown to have the form

$$p(\theta_{1t},\theta_{2t}|Y_{t-1}) = \frac{b^a e^{-b(\theta_{1t}+\theta_{2t})}}{\Gamma(a)\, B(c,d)} \theta_{1t}^{c-1}\, \theta_{2t}^{d-1} \; (\theta_{1t} + \theta_{2t})^{a-(c+d)} \qquad (5.2.19)$$

This may be considered a bivariate gamma density for which the sum is always gamma distributed. We now evaluate its moments. Using (5.2.2a) and (5.2.15) it is easy to show that the mean of $\{\theta_{it}\}$ given Y_{t-1} is

$$\begin{aligned} E(\theta_{it}|Y_{t-1}) &= E_i \\ &= E(\theta_t \, \pi_{it}|Y_{t-1}) \\ &= E(\theta_t|S_{t-1}) \, E(\pi_{it}|Y_{t-1}) \\ &= \frac{a}{b} \, \frac{c_i}{(c + d)} \qquad (5.2.20) \end{aligned}$$

where $c_i = c_{t/t-1}$ for $i=1$ and equal to $d_{t/t-1}$ for $i=2$. The variances are readily shown to be

$$\mathrm{Var}(\theta_{it}|Y_{t-1}) = E(\pi_{it}^{2}|Y_{t-1})\ E(\theta_t^{2}|Y_{t-1}) - [\ E(\pi_{it}|Y_{t-1})\ E(\theta_t|Y_{t-1})]^{2}$$

$$= \frac{a\ c_i\ [\ (c+1)(c+d) + ad\ ]}{b^{2}\ (c+d)^{2}\ (c+d+1)} \qquad (5.2.21)$$

while the covariance is given by

$$\mathrm{Cov}(\theta_{1t}\theta_{2t}|Y_{t-1}) = E(\pi_t(1-\pi_t)\theta_t|Y_{t-1}) - E(\pi_t\theta_t|Y_{t-1})E((1-\pi_t)\theta_t|Y_{t-1})$$

$$= \frac{E_1E_2\ [\ (c_{t/t-1} + d_{t/t-1})^{-}\ a_{t/t-1}\ ]}{(1 + c_{t/t-1} + d_{t/t-1})\ a_{t/t-1}} \qquad (5.2.22)$$

where the E_i's are given in (5.2.20). Given the intrinsic non-stationary character of our model care should be exercised when interpreting this measure of linear association. With the above results we are now in a position of deriving the standard moments for the predictive distribution in (5.2.17). Using (5.2.2a-b) it is easy to see that

$$E(y_{it}|Y_{t-1}) = E(\theta_{it}|Y_{t-1}) = E_i \qquad (5.2.23)$$

$$\mathrm{Var}(y_{it}|Y_{t-1}) = E(\theta_{it}|Y_{t-1}) + \mathrm{Var}(\theta_{it}|Y_{t-1})$$

$$= \frac{E_i + E_1\ E_2}{(c+d)} + \frac{(c_i+1)\ E_i}{b\ (1+c+d)} \qquad (5.2.24)$$

$$\mathrm{Cov}(y_{1t}, y_{2t}|Y_{t-1}) = \mathrm{Cov}(\theta_{1t}\ \theta_{2t}|Y_{t-1}) \qquad (5.2.25)$$

One may also verify that the univariate means may also be expressed through the general formula

$$E(y_{it}|Y_{t-1}) = E(S_t|S_{t-1})\ E(y_{i,t}|S_t,Y_{t-1}) /\ S_t \qquad i=1,2.$$

From (3.2.13), the conditional expectation of S_t is an EWMA with weights determined by the hyperparameter ω_1. Let this be denoted as $EWMA_1(S)$. Furthermore if one refers to the Binomial-Beta model (section 4.2) the second moment on the above expression may be shown to be equal to the ratio of an EWMA of the y_{it}'s with hyperparameter ω_2, denoted $EWMA_2(y_i)$,to a similar EWMA for the sum S_t. Thus

$$E(y_{it}|Y_{t-1}) = \frac{EWMA_2(y_i)\ EWMA_1(S)}{EWMA_2(S)} \qquad (5.2.26)$$

with i=1,2. In the special case when $\omega_1 = \omega_2$ this reduces to an EWMA of the observations in the i-th series. As we shall see when this equality holds true independence between the two processes follows.

5.2.1 Dependence Structure Implied by the Model

Correlation: if the recursions involving the prior parameters are initialized with 'unbiased' priors, i.e., by setting $a_0=b_0=c_0= d_0=0$, then by use of the *predictive* and *updating* equations one may easily show that the term between square brackets in the numerator of (5.2.22) reduces to

$$[(c_{t/t-1}+ d_{t/t-1}) - a_{t/t-1}]= \sum_{i=0}^{t-1} (\omega_2 -\omega_1)^i S_{t-i} \qquad (5.2.27)$$

From the above expression one may derive a guide to the covariance structure of our bivariate model

(i) if $\omega_1= \omega_2$ then there is no linear association between the two series.

(ii) if $c_{t/t-1}+ d_{t/t-1} > a_{t/t-1}$ a positive correlation is observed·

(iii)if $c_{t/t-1}+ d_{t/t-1} < a_{t/t-1}$ a negative correlation is observed.

Note that the derived correlation structure is rather rich when compared with similar bivariate count data models in the statistical literature. See,e.g., Stein and Juritz (1987).

Independence: when a=c+d (5.2.19) splits into two distinct factors and the $\{\theta_{it}\}$ are independent gamma states. Under this condition it is straightforward to show that the joint predictive distribution (5.2.17) also splits into two univariate NBD models, i.e,

$$p(z_t|Y_{t-1})= NBD(y_{1t};c,b).NBD(y_{2t};d,b) \qquad (5.2.28)$$

so that independence between the series follows. Note that, by (5.2.27) the independence condition occcurs only if ω_1=ω_2, and that as a result zero correlation implies independence. An interesting corollary of the independence of the series is that the likelihood function for $\omega_1 = \omega_2$ is given by the product of the likelihood functions for the individual series. A likelihood ratio test of the hypothesis that $\omega_1 = \omega_2$ can be carried out for $0 < \omega_1, \omega_2 < 1$. If the null hypothesis is accepted, the series should be forecast separately.

5.2.2 Likelihood

The log-likelihood function is obtained by summing the logarithms of the joint predictive distributions (5.2.16) from τ+1 to T, where τ is defined as the first value of t for which all the series have had at least one non-zero observation. By use of (5.2.14b) it is straightforward to show that the overall log-likelihood function may be factorized as

$$\log L(\omega_1,\omega_2) = \log L_1(\omega_1) + \log L_2(\omega_2) \qquad (5.2.29)$$

where $L_1()$ and $L_2()$ are the likelihood functions associated with the sum S_t and the series y_{1t} respectively. Hence the optimization problem may be split into the maximization of two separate log-likelihood functions, one with respect to ω_1 and the other with respect to ω_2.

5.2.3 Marginal Predictive Distributions

Although we have been be able to derive the univariate predictive moments in a straightforward fashion the establishment of the correspondent univariate distributions is a more elaborate issue. When a=c+d this is trivial. See (5.2.28). When a ≠ c+d, i.e., when the series are dependent, the joint predictive density is given in (5.2.14b). To obtain the marginals one has to sum this whole expression wrt to each of the variables. Clearly in this case the marginals will not be NBD's. Proceeding with the necessary operations one may easily show that the marginal for the first variable is given by

$$p(y_{1t}|Y_{t-1}) = k(y_{1t}) \sum_{y_{2t}=0}^{\infty} \frac{\Gamma(a+y_{1t}+y_{2t})\ \Gamma(d+y_{2t})}{\Gamma(c+d+y_{1t}+y_{2t})\ y_{2t}!} \ \frac{1}{(1+b)^{y_{2t}}} \qquad (5.2.30a)$$

where

$$k(y_{1t}) = \frac{\Gamma(y_{1t}+c)\ b^{a}}{B(c,d)\ \Gamma(a\ )\ (1+b)^{(a+y_{1t})}} .$$

Constants apart the infinite sum in (5.2.30a) is the hypergeometric series ${}_2F_1(a+y_{1t},d;c+d+y_{1t};1/(1+b))$. Following the lines of section (4.2.4) one may show that the ratio of the $(j+1)^{st}$ term to the j^{th} term in this expansion may be expressed as

$$A_{j+1} / A_j = \frac{(a+y_{1t}+j)\ (d+j)}{(c+d+y_{1t}+j)\ (j+1)} \quad 1 / (1+b) \qquad (5.2.31)$$

with $A_0=1$. The final form for the univariate predictive density is then

$$p(y_{1t}|Y_{t-1}) = \frac{b^a\ \Gamma(c+y_{1t})\ \Gamma(c+d)\ \Gamma(a+y_{1t})}{y_{1t}!\ \Gamma(a)\ \Gamma(c)\ \Gamma(c+d+y_{1t})\ (1+b)^{(a+y_{1t})}} \sum_{j=0}^{\infty} A_j(y_{1t}) \qquad (5.2.30b)$$

When $a=c+d$, i.e., under the independence assumption, one has that

$$\sum_{j=0}^{\infty} A_j = [(1+b)/b]^d. \qquad (5.2.32)$$

From this one may easily show that, as expected, this marginal reduces to the univariate NBD(y_{1t};c,b). For the second series it is equally possible to demonstrate that

$$p(y_{2t}|Y_{t-1}) = \frac{b^a\ \Gamma(d+y_{2t})\ \Gamma(c+d)\ \Gamma(a+y_{2t})}{y_{2t}!\ \Gamma(a)\Gamma(d)\ \Gamma(c+d+y_{2t})\ (1+b)^{(a+y_{2t})}} \sum_{j=0}^{\infty} B_j(y_{2t}) \qquad (5.2.33)$$

where the B's follow the recursive relation

$$B_{j+1} / B_j = \frac{(a+y_{2t}+j)\ (c+j)}{(c_+d+y_{2t}+j)\ (j+1)} \quad 1/ (1+b) \qquad (5.2.34)$$

with $B_0=1$. As before when $a=c+d$ this reduces to the NBD(y_{2t};d,b).

5.2.4 <u>Conditional Distributions</u>:

In this section we derive the conditional distribution of y_{1t} given y_{2t} and Y_{t-1}. Using a standard result of probability calculus one can show that

$$p(y_{1t}|y_{2t},Y_{t-1})= \frac{\Gamma(a+S_t)\ \Gamma(y_{1t}+c)\ \Gamma(c+d+y_{2t})\ \ (1+b)^{-y_{1t}}}{\Gamma(c)\ \Gamma(c+d+S_t)\ \Gamma(a+y_{2t})\ y_{1t}!\ \sum_{j=0}^{\infty} B_j(y_{2t})} \tag{5.2.35}$$

where the B's follow the recursion given in (5.2.33). Note that when the variables are independent, i.e., when a=c+d, this reduces to the NBD(y_{1t};c,b), as expected.

We now derive the first two conditional moments for y_{1t}. Using the above distribution one can show that the conditional mean has the form

$$E(y_{1t}|y_{2t},Y_{t-1})= k(y_{2t}) \sum_{y_{1t}=1}^{\infty} \frac{y_{1t}\ \Gamma(a+y_{1t}+y_{2t})\ \Gamma(c+y_{1t})}{y_{1t}!\ \Gamma(c+d+y_{1t}+y_{2t})\ (1+b)^{y_{1t}}} \tag{5.2.36a}$$

where

$$k(y_{2t})= \frac{\Gamma(c+d+y_{2t})}{\Gamma(c)\ \Gamma(a+y_{2t})\ \sum_{j=0}^{\infty} B_j(y_{2t})}. \tag{5.2.37}$$

Using that y_{1t} / y_{1t}! = $1/(y_{1t}-1)!$ one may show that the infinite sum in (5.2.36a)is the series ${}_2F_1(a+y_{2t}+1,c+1;c+d+y_{2t}+1,1/(1+b))$, apart from a constant term. It is then straightforward to show that the conditional mean may be put into the following final form

$$E(y_{1t}|y_{2t},Y_{t-1}) = \frac{c\ (a+y_{2t})}{(1+b)\ (c+d+y_{2t})} \quad \frac{\sum_{j=0}^{\infty} C_j(y_{2t})}{\sum_{j=0}^{\infty} B_j(y_{2t})} \qquad (5.2.36b)$$

where the B's are as in (5.2.33) and the C's follow the recursion

$$C_{j+1}\ /\ C_j = \frac{(a+y_{2t}+j+1)(c+j+1)}{(c+d+y_{2t}+j+1)\ (j+1)} \quad 1/(1+b). \qquad (5.2.38)$$

Observe that the conditional mean is a non-linear function in y_{2t}. One may easily prove that under the independence assumption this mean collapses to the mean of the NBD(y_{1t};c,b), namely c/b.

In order to evaluate the conditional variance we first determine the expression for the first factorial moment, since this, as will be shown, can also be expressed as a hypergeometric series. Using (5.2.35) it is straightforward to show that

$$E(y_{1t}(y_{1t}-1)|y_{2t},Y_{t-1}) = k(y_{2t}) \sum_{y_{1t}=2}^{\infty} \frac{y_{1t}(y_{1t}-1)\ \Gamma(a+y_{1t}+y_{2t})\Gamma(c+y_{1t})}{y_{1t}!\ \Gamma(c+d+y_{1t}+y_{2t})\ (1+b)^{y_{1t}}} \qquad (5.2.39a)$$

where $k(y_{2t})$ is given by (5.2.37). Using that $y_{1t}(y_{1t}-1)/y_{1t}!$

$= 1/(y_{1t}-2)!$ one may easily express the infinite sum in (5.2.39) as the hypergeometric series ${}_2F_1(a+y_{2t}+2, c+2; c+d+y_{2t}+2, 1/(1+b))$, apart from a constant term. One then arrives at the following expression for the first factorial moment

$$E(y_{1t}(y_{1t}-1) \mid y_{2t}, Y_{t-1}) = \frac{(a+y_{2t})(a+y_{2t}+1)(c+1)}{(1+b)^2 (c+d+y_{2t})(c+d+y_{2t}+1)} \frac{\sum_{j=0}^{\infty} D_j(y_{2t})}{\sum_{j=0}^{\infty} B_j(y_{2t})}$$

(5.2.39b)

where the D's are given by the recursion

$$D_{j+1}/\, D_j = \frac{(a+y_{2t}+j+2)(c+j+2)}{(c+d+y_{2t}+j+2)\ (j+1)} \quad 1/(1+b) \qquad (5.2.40)$$

and B's are as in (5.2.33). As before when $a=c+d$ this, the first factorial moment collapses to the equivalent expression for a $NBD(y_{1t};c,b)$, i.e., $c(c+1)/b^2$. The conditional variance may then be evaluated by direct manipulation of (5.2.39b) and (5.2.36b).

5.2.5 Explanatory Variables and Structural Components:

The introduction of these effects in our bivariate scheme may be accomplished *via* two different sources. If one wants to investigate the impact of explanatory variables/structural components on the overall sum then, following the developments of the univariate Poisson-Gamma model (section 3.2.3), the exponential link function is the appropriate link. As a result now the rate *prediction* and *updating* equations are as in (3.2.28) and (3.2.29) respectively, with y_t replaced by S_t. Note that the corresponding 'regression' hyperparameters are optimized jointly with ω_1 via the NBD component in (5.2.20). With regard to the one step ahead prediction for the series, only the term associated with the overall rate in (5.2.20), namely $E(\theta_{T+1}|Y_T)$, has to be duly modified.

It is also possible to introduce effects which influence the relative share associated with the first variable, π_t. Note, however, that since the two analyses proceed independently, the two sets of explanatory variables/structural components may be overlapping. From the univariate Binomial-Beta model in section 4.2.4 we know that the natural link function for the relative share π_t is the logit function (4.2.10) and that one may consider two different alternatives to introduce these effects, either by a *series expansion* or *modal approximation*. If the reader refers back to this section the necessary equations for both treatments are displayed. In what follows we comment on the modifications one has to introduce in order to adapt these equations to the bivariate case. Given the difference in notation between the two formulations, the general rule is to substitute a by c, b by d, y_t by y_{1t} and n_t by S_t where appropriate.

series expansion- the *prediction equations* (5.2.11) remain unchanged, but the *updating equations* are now as in (4.2.23) which make use of (4.2.21b) and (4.2.22b). The univariate moments (5.2.20) are only affected on the term associated with the relative share.

The joint predictive distribution (see 5.2.14b) is obviously only affected at the level of $p(y_{1t}|S_t,Y_{t-1})$, which is now given by (4.2.14c). Substituting this term in the expression for the bivariate distribution one obtains that

$$p^+(z_t|Y_{t-1})= p(z_t|Y_{t-1})\, u_t^{y_{1t}} \sum_{j=0}^{\infty} A_j(u_t) \qquad (5.2.41)$$

where p() is the standard bivariate distribution (5.2.16), $u_t= \exp(z_t'\delta)$ and the A's are as in (4.2.15). Observe that in the absence of explanatory variables this reduces to the standard case.

One may equally show that the marginal and conditional distributions for y_{1t} follow similar expressions and these are given, respectively, by

$$p^+(y_{1t}|Y_{t-1})= p(y_{1t}|Y_{t-1})\, u_t^{y_{1t}} \sum_{j=0}^{\infty} A_j(u_t) \qquad (5.2.42)$$

and

$$p^+(y_{1t}|y_{2t},Y_{t-1})= p(y_{1t}|y_{2t},Y_{t-1})\, u_t^{y_{1t}} \sum_{j=0}^{\infty} A_j(u_t) \qquad (5.2.43)$$

where the p()'s are the standard distributions given in (5.2.30b) and (5.2.35) respectively.

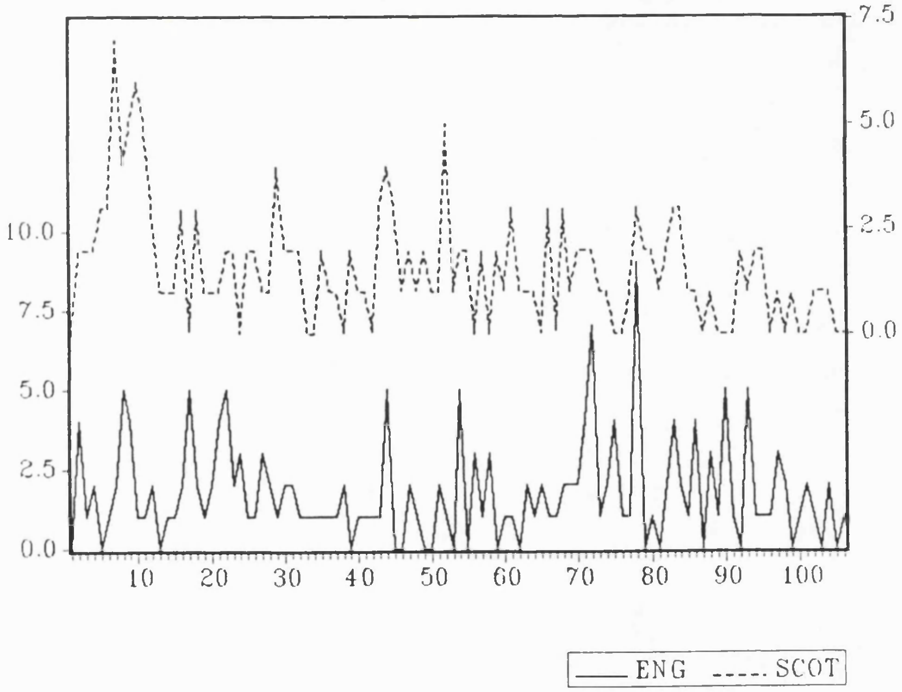
modal approximation- the *prediction* and *updating* equations remain essentially the same, the only modification being the replacement of y_{1t} by w_t which is given by (4.2.25). In the joint predictive distribution the term $p(y_{1t}|S_t,Y_{t-1})$ is replaced by the *modal* distribution given in (4.2.27). Marginal and conditional distributions may also be made available by the proper manipulation of the formulas.

5.3 APPLICATION:

In Chapter 3, section 3.5.1 we fitted the Poisson-Gamma model to the number of goals scored by England in international football matches played against Scotland at Hampden Park in Glasgow.

The bivariate model developed in this chapter can be used to formulate a model in which the goals scored by England are modelled jointly with those scored by Scotland. The series of goals for both teams are depicted in figure 5.3.1 below.

Figure 5.3.1 Series of goals by England and Scotland in football international matches



Given that the football matches have been played either in England (mostly at Wembley) or Scotland (at Hampden Park), the match venue is the natural explanatory variable for the proportion of goals scored by the teams. Since we are interested in predicting the goals scored by England we investigate how this dummy affects England's proportion. The dummy variable x_t is defined such that

$$x_t = \begin{array}{l} +1 \text{ for matches played in England.} \\ -1 \text{ '' '' '' '' Scotland.} \end{array}$$

The above dummy has also been used for the total of goals. We have found that according to standard goodness of fit criteria defined in Chapter 3, the best specifications were given by

- M_1 : model in which we have assumed at the outset the constraint $\omega_1 = \omega_2$, i.e., independence between the two series of goals. The dummy is used both for the overall sum and England's relative share.

- M_2 : unconstrained model where the dummy is used in both mechanisms.

Table 5.3.1 Bivariate model fitted to series of goals by England and Scotland.

	estimates				goodness-of-fit			
	$\tilde{\omega}_1$	$\tilde{\omega}_2$	$\tilde{\delta}_1$	$\tilde{\delta}_2$	ML	AIC	BIC	U
M_1	0.885	$=\tilde{\omega}_1$	0.136	0.203	-157.95	323.35	331.29	0.677
M_2	0.844	0.930	0.139	0.203	-158.67	323.907	334.485	0.678

where δ_1 and δ_2 are the dummy hyperparameters associated with μ and π respectively. Note that for both specifications the series expansion has been the best technique to introduce the dummy for the relative share π.

The selected specifications indicate that the venue is a relevant factor in explaining both the total number of goals and the share of England in this total. Model M_2 seems to suggest some sort of dependence between the two series, altough the improvement on the fit is barely affected if independence is assumed at the outset, by setting ω_1=ω_2. In fact the likelihood ratio test statistic is 1.44, so that the null hypothesis that this restriction is valid seems to be supported by the data. Hence we are led to believe that the goals scored by the two teams are independent and as a result they should be forecasted independently.

CHAPTER SIX

THE GAMMA-GAMMA MODEL

6.1 INTRODUCTION:

In this chapter we set up a dynamic model for time series of gamma observations. Typical situations where this type of distribution occur are given by data in the form of wind speeds (Lawrance and Lewis 1985), monthly insurance claims, daily flows of a river, etc. Our standard gamma model is characterized by a constant shape parameter and a time varying scale parameter, which evolves according to the *discounting mechanism*. A formally equivalent model was introduced by Bather (1965) but in his framework there is no consideration of maximum likelihood estimation or the introduction of explanatory variables/structural components. In our framework two different ways of implementing these effects are available: either *via* the shape or the scale parameters of the gamma measurement equation. Both are based on the use of the exponential link function. In the special case when the shape parameter is set equal to one our model reduces to the dynamic model for exponential observations given in

S&M (1986). As usual, when the measurements are independent random variables the GLM (McCullagh and Nelder 1983) should be used.

If the series being modelled is stationary or can be made stationary then some ARIMA based models are available. For example the models of Lawrance and Lewis (1985) and McKenzie (1982). The former is developed for time series with exponential marginal distribution with second order autoregressive structure. Besides its rather 'unnatural' noise structure, mostly imposed to guarantee the required marginal distribution, their NEAR(2) model presents potential difficulties at the stage of parameter estimation, given that the equivalent of our conditional density is discontinuous. Note also that, from the perspective of forecasting, the fixing of a certain marginal distribution is irrelevant. For a more detailed view on their model see the aforementioned reference and the discussion that follows. McKenzie (1985) presents a *product autoregressive* model with gamma marginal. This has an AR(1) correlation structure, with a *measurement equation* in which the noise enters multiplicatively. Given that in this setup the noise distribution doesn't need to be made explicit, applications seem to be restricted to data simulation, for example, in synthetic hydrology.

6.2 THE GAMMA-GAMMA MODEL:

- *measurement equation*: let the observations at time t be generated from a gamma density with fixed shape υ and time varying scale parameter θ_t (the *state*)

$$p(y_t|\upsilon,\theta_t)= \frac{\theta_t^{\upsilon}\; y_t^{\upsilon-1}\; e^{-\theta_t y_t}}{\Gamma(\upsilon)} \quad , \quad 0 < y_t \leqslant \infty \qquad (6.2.1)$$

where $\upsilon > 0$ and $\theta_t > 0$. The n^th^ order moment is given by the expression

$$E(y_t^n|\upsilon,\theta_t)= \frac{\Gamma(\upsilon+n)\quad \theta_t^{-n}}{\Gamma(\upsilon)} \qquad (6.2.2)$$

from which follows the standard results

$$E(y_t|\upsilon,\theta_t)= \upsilon \;/\; \theta_t \qquad (6.2.3a)$$

$$Var(y_t|\upsilon,\theta_t)= E(y_t|\upsilon,\theta_t)/\; \theta_t. \qquad (6.2.3b)$$

One can also show that

$$Mode(y_t|\upsilon,\theta_t)\begin{cases} [(\upsilon-1)/\upsilon]\; E(y_t|\upsilon,\theta_t) & , \upsilon > 1 \\ 0 & , 0 < \upsilon < 1. \end{cases} \qquad (6.2.4)$$

Observe that when the *secondary* parameter υ is set equal to one this density reduces to the exponential model, considered by S&M (1986).

(i) *state prediction*- for a gamma model specified as in (6.2.1) the natural conjugate prior is also a gamma density, so that one would think that the equations (3.2.2-3) which appear in the Poisson-Gamma model are also valid here. Although the state predictive density $p(\theta_t|Y_{t-1})$ remains gamma, the predictive equations are not the same. In view of (2.2.5a), it is the expected value of θ_t^{-1} (6.2.3a), rather than θ_t, which needs to be kept constant while the variance increases. For a gamma distribution (3.2.2), it may be shown that

$$E(\theta_t^{-1}) = b_{t-1}/(a_{t-1} - 1) \qquad (6.2.5)$$

so that, the appropriate prediction equations are given by

$$a_{t/t-1} = \omega\, a_{t-1} + (1-\omega) \qquad (6.2.6a)$$

$$b_{t/t-1} = \omega\, b_{t-1}. \qquad (6.2.6b)$$

It is possible to show that in order to have the variance of $\mu_t = \theta_t^{-1}$ increased in the transition one must have $a_{t-1} > (1+\omega)/(1-\omega)$. As before we leave a more systematic investigation of this condition until after we have established the updating equations.

(ii) *state updating-* by direct use of Bayes theorem one may easily show that $p(\theta_t|Y_t) \sim \text{gamma}(a_t,b_t)$ with

$$a_t = a_{t/t-1} + \nu \qquad (6.2.7a)$$

$$b_t = b_{t/t-1} + y_t \qquad (6.2.7b)$$

Repeated substitution from (6.2.6a) and (6.2.7a) shows that

$$a_{t/t-1} = (1-\omega) \sum_{j=1}^{t-1} \omega^{j} + (1-\omega) + \upsilon \sum_{j=1}^{t-1} \omega^{j} \qquad (6.2.8)$$

As before the equivalent of a 'steady state filter' may be obtained for large samples, when $a_{t/t-1} \rightarrow [(\upsilon-1)\omega+1]/(1-\omega)$, so that $a_{t-1} \rightarrow [\upsilon(1-\omega])/(1-\omega)$. Using this latter result one may easily establish that the condition for the variance increase becomes, asymptotically, $\upsilon > 2\omega$.

(iii) *conditional distribution*- the *mixing* operation in (1.1.5) produces a density known as the inverted-beta-2 (Raiffa and Schlaifer 1961, p.221) or the inverse beta (Aitchison and Dunmore 1975, p.24) which has the form

$$p(y_t|Y_{t-1}) = \frac{y_t^{\upsilon-1} \; b^{a}}{(b + y_t)^{\upsilon + a} \; B(\upsilon,a)} \qquad (6.2.9)$$

where $a=a_{t/t-1}$, $b=b_{t/t-1}$ and $B(\cdot)$ is the beta function. It can be shown that the Mellin transform for this density is given by

$$E(y_t^{n}|Y_{t-1}) = \frac{b_{t/t-1}^{n} \; \Gamma(n+\upsilon) \; \Gamma(a_{t/t-1}-n)}{\Gamma(a_{t/t-1}) \; \Gamma(\upsilon)} \qquad (6.2.10)$$

with $a_{t/t-1} > n$. From the above one may easily arrive at the expressions for the mean and variance which are given respectively by

$$\tilde{y}_{t/t-1} = \frac{\upsilon \; b_{t/t-1}}{(a_{t/t-1} - 1)} = \frac{\upsilon \; b_{t-1}}{(a_{t-1} - 1)} \; , \quad a_{t-1} > 1 \qquad (6.2.11a)$$

and

$$\mathrm{Var}(y_t \mid Y_{t-1}) = \frac{\nu \quad (\nu + a_{t/t-1} - 1)\; b^2_{t/t-1}}{(a_{t/t-1} - 1)^2 \quad (a_{t/t-1} - 2)}$$

$$= \frac{\nu\; [\nu + \omega\; (a_{t-1} - 1)]^2 b_{t-1}}{(a_{t-1} - 1)^2 (a_{t-1} - 1 - 1/\omega\;)} \qquad , \quad a_{t-1} > 1 + 1/\omega. \qquad (6.2.11b)$$

Following the previous models we also adopt here the 'unbiased' or non-informative gamma prior to initialize the above recursions. For $0 < \omega < 1$ one may again show that the forecast is expressed as an EWMA scheme as in (3.2.13).

transition equation- as in the Poisson-Gamma model here it is also possible to show that the *implicit* transition mechanism adopted for the state evolution is formally equivalent to the multiplicative transition equation suggested in S&M (1986). In the present context given that the distribution of the 'noise' η_t is data independent, this equation will be of importance in establishing multi-step ahead forecasting. The transition equation in (3.2.7) is given by

$$\theta_t = \omega^{-1}\; \theta_{t-1}\; \eta_t \qquad (6.2.12)$$

with $\eta_t \sim \mathrm{beta}(c_t, d_t)$. In view of (6.2.5) and (6.2.6a-b) we shall have that

$$c_t = \omega a_{t-1} + (1-\omega) \qquad (6.2.13a)$$

$$d_t = (1-\omega)(a_{t-1} - 1) \qquad (6.2.13b)$$

From (6.2.7a) we can see that the updating equation for the gamma prior shape parameter a_t is free from the endogenous variables so that η_t will also be data independent. In fact by use of (6.2.8) asymptotic values for the above parameters can be easily established and these are given by

$$c_t \cong 1 + \omega\upsilon/(1-\omega) \qquad (6.2.14a)$$

$$d_t \cong \upsilon \qquad (6.2.14b)$$

with $0 < \omega < 1$. So that eventually the 'noise' η_t attains stationarity.

Using (6.2.12) one may derive the state transition density and an expression for its nth order moment. The form of the transition density of our model coincides with that of Bather's gamma model (1965, p.838) and this is given by

$$p(\theta_t | \theta_{t-1}) = \frac{\Gamma(c_t+d_t)}{\Gamma(c_t)\ \Gamma(d_t)} \left[\frac{\theta_t\ \omega}{\theta_{t-1}}\right]^{c_t-1} \left[1 - \frac{\theta_t\ \omega}{\theta_{t-1}}\right]^{d_t-1} \frac{\omega}{\theta_{t-1}} \qquad (6.2.15)$$

with $0 < \theta_t < \omega^{-1}\ \theta_{t-1}$. The nth order moments are given by

$$E(\theta_t^n | \theta_{t-1}) = \frac{\omega^{-n}\ \theta_{t-1}^{n}\ \Gamma(c_t +n)\ \Gamma(c_t+d_t)}{\Gamma(c_t)\ \Gamma(c_t+d_t+n)} .$$

(iv) *k-steps ahead forecasting-* following S&M (1986, p.82) we use the explicit state transition equation in (6.2.12) to derive closed expressions for the nth order moments of the forecasting distribution, $p(y_{T+k}|Y_T)$. Using (1.1.9b) and (6.2.2) one may show that

$$E(y_{T+k}^{n}|Y_T)=\frac{\Gamma(\upsilon+n)}{\Gamma(\upsilon)}E(\theta_{T+k}^{-n}|Y_T). \qquad (6.2.16)$$

The expectation on the rhs of (6.2.16) can be evaluated by use of the state transition equation (6.2.12). When raised to the nth power and projected k steps ahead the reciprocal of the state equation has the form

$$\theta_{T+k}^{-n}=\omega^{nk}\theta_T^{-n}\prod_{i=1}^{k}\eta_{T+i}^{-n}. \qquad (6.2.17)$$

It follows from the above expression and the independence of $\mathbf{Y}_T$ and the η_{T+i}'s that

$$E(\theta_{T+k}^{-n}|Y_T)=\omega^{nk}E(\theta_T^{-n}|Y_T)\prod_{i=1}^{k}E(\eta_{T+i}^{-n}). \qquad (6.2.18)$$

Using the expressions for the nth order moments of the gamma and beta densities, given in (3.2.3) and (3.3.4) respectively, the above expression takes the form

$$E(\theta_{T+k}^{-n}|Y_T)=\frac{\omega^{nk}\,b_T^{n}\,\Gamma(a_T-n)}{\Gamma(a_T)}\;P(k,n) \qquad (6.2.19)$$

where

$$P(k,n) = \prod_{i=1}^{k} \frac{\Gamma[\omega a_i+(1-\omega)-n]\ \Gamma(a_i)}{\Gamma[\omega a_i+(1-\omega)]\ \Gamma(a_i-n)} \qquad .(6.2.20)$$

with $a_i = a_{T+i-1}$ obtained through the recursion

$$a_{T+i-1} = \omega^{i-1} a_T + [\upsilon + (1-\omega)] \sum_{j=0}^{i-2} \omega^{j} \qquad , \ i = 2,3,\ldots,k \qquad (6.2.21)$$

Finally subsituting (6.2.19) in (6.2.16b) we obtain the Mellin transform for the multi-step forecasting density

$$E(y^{n}_{T+k} | Y_T) = \frac{\omega^{nk}\ b_T\ \Gamma(\upsilon+n)\ \Gamma(a_T-n)}{\Gamma(a_T)\ \Gamma(\upsilon)}\ P(k,n) \qquad (6.2.22)$$

where we have assumed that $a_i > 1 + (n/\omega)$ for $T < t < T+k$. Given that when $k \rightarrow \infty$ $a_i \rightarrow 1 + \upsilon/(1-\omega)$ the previous condition may be used to establish a lower bound for the existence of multi-step moments for any k. In fact it is easy to see that this is given by the inequality

$$\upsilon > \frac{(1-\omega)}{\omega}\ n$$

so that the existence of forecasting moments depends on the following factors:

(i) on the data via υ.

(ii) on having a value of ω close to one.

If we define $E(y^{n}{}_{T+k}|Y_T) \equiv m(k,n)$ then the following recursive formula may be used to obtain the multi-steps moments

$$m\,(k+1,n+1)= \frac{m\,(k,n)\ \omega^{n+k+1}\ (\nu+n)\ \Gamma[\omega(a_{k+1}-1)-n]\ \Gamma(a_{k+1})\ S(k,n)}{(a_T-n-1)\ \Gamma[\omega(a_{k+1}-1)-1]\ \Gamma(a_{k+1}-n-1)} \qquad (6.2.23)$$

where $n=0,1,\ldots;$ $k=1,2,\ldots$ and a_{k+1} is obtained from (6.2.21) and

$$S(k,n)= \prod_{i=1}^{k} \frac{(a_i-n-1)}{[\omega(a_i-1)-n]}\,. \qquad (6.2.24)$$

Initial values for $m(k,n)$ can be read off directly from (6.2.9). Observe that, as expected, the forecast function is a constant given by the value it takes at the last observation (6.2.9b).

The appropriate way to evaluate the multi-step ahead forecasting density should be, in principle, by solving the integral given in (1.1.6), since this involves the state multi-step density $p(\theta_{T+k}|Y_T)$, and this can be computed using the integral in (1.1.7). It is clear that an analytical solution for this problem is obtainable only if the latter distribution is of the gamma type. Unfortunately this is only true for the one step ahead density. For higher steps ahead the densities produced are not even analytical, as we shall see. The reason for this behaviour is that a product of gamma and beta variables produces a gamma variable only under the condition that the second shape parameter of the beta density is equal to the difference between the shape parameter of the gamma density and the first shape parameter of the beta. This condition, which we shall call 'the gamma condition', is violated for steps higher than one, so that equation 6.2.12 will be of limited use in establishing an analytical expression for the forecasting distribution.

We now derive the full expression for the two steps ahead state density in order to understand the inherent difficulties in multi-step prediction. The appropriate way to proceed is by solving the integral given by (1.1.7). The first density on the integrand $p(\theta_{T+2}|\theta_{T+1})$ has parameters $c_{T+2}= \alpha$ and $d_{T+2}= \beta$ and is obtained by making t=T+2 in (6.2.15). The second density $p(\theta_{T+1}|Y_T)$ is gamma with parameters $a_{T+1/T}= \lambda$ and $b_{T+1/T}= \beta$ where

$$\lambda= \omega\, a_T + (1-\omega) \quad , \qquad \varphi= \omega\, b_T \qquad (6.2.25a)$$

$$\alpha= \omega^2 a_T + (1-\omega)(1+\omega) + \omega\upsilon, \quad \beta= (1-\omega)[\omega\,(a_T-1)+\upsilon] \qquad (6.2.25b)$$

One can easily check that the gamma condition is not satisfied for this equation, given that $\beta \neq (\lambda-\alpha)$. In fact, using (6.2.21a-b) it is easy to see that $\lambda-(\alpha+\beta)= -\upsilon$, $\upsilon > 0$, so that the condition will never be satisfied. After some simple manipulation one may show that the exact density is given by

$$p(\theta_{T+2}|Y_T)=\frac{\theta_{T+2}^{\alpha-1}\ \varphi^{\alpha}\ e^{-\varphi\omega\theta_{T+2}}\ g(\omega\theta_{T+2};\alpha,\beta,\lambda,\varphi)}{\Gamma(\alpha)} \qquad (6.2.26)$$

where $g(\cdot)$ is given by

$$g(\omega\theta_{T+2};\alpha,\beta,\lambda,\varphi)= \frac{\varphi^{\lambda-\alpha}\ \Gamma(\alpha+\beta\ )}{\Gamma(\beta)\Gamma(\lambda)} \int_0^{\infty} (\omega\ \theta_{T+2}+ z)^{-\upsilon}\ z^{\beta-1}\ e^{-\varphi z}\ dz \qquad (6.2.27)$$

The density in (6.2.26) is easily recognized as the product of a gamma(α,φ)density by a 'perturbing' term g(·).

Given the inherent intractability of the multi-step density one could follow some of the strategies adopted for the Poisson-Gamma case (section 3.2.2). In particular the use of Monte Carlo simulations is recommended, where predictions, following S&M (1986), could be given in terms of sample percentiles, modes or means. An approximation based on an inverted beta ($a_{T+k/T}, b_{T+k/T}$) also might be used, although with some care. As before the projected parameters $a_{T+k/T}$ and $b_{T+k/T}$ are found by equating the first two moments of the actual forecasting density (see 6.2.9-10) with the correspondent moments of the approximating inverted beta. After some algebra one may show that

$$a_{T+k/T} = \frac{2\rho_k(a_T - 1) - b(\upsilon+1)(a_T - 2)}{[\rho_k(a_T - 1) - b(\upsilon+1)(a_T - 2)]} \qquad (6.2.28a)$$

$$b_{T+k/T} = \frac{b_T\ \rho_k}{[\rho_k(a_T - 1) - b(\upsilon+1)(a_T - 2)]} \qquad (6.2.28b)$$

where $\rho_k > [b(\upsilon+1)(a_T-2)]/(a_T-1)$ and $\rho_k = \omega\ P(k,2)$ with P(k,2) obtained by setting n=2 in (6.2.20).

6.2.1 Likelihood:

Estimation of the shape and the discount parameters proceed as usual, *via* the log-likelihood function. Using (6.2.8) it is easy to see that

$$\log L(\omega,\upsilon) = \sum_{t=\tau}^{\infty} [(\upsilon-1)\log y_t + a_{t/t-1}\log b_{t/t-1} - (a_{t/t-1}+\upsilon).$$
$$\log(b_{t/t-1}+y_t) - \log B(\upsilon,a_{t/t-1})\ ] \ . \qquad (6.2.29)$$

As before the parameters recursions are initialized with an 'unbiased' prior. In view of (6.2.6a) a_t will be always positive. In order to ensure that $b_{t/t-1}$ is strictly positive we require τ to be the first value of t for which y_t is non zero (cf. the NBD-Beta model, section 3.3.1).

6.2.2 Explanatory Variables and Structural Components:

Given that the observables y_t are positive variables a sensible candidate for the link function is the exponential link (3.2.19) and this should be introduced under the requirement that in the presence of these effects the mean has the form

$$E^{+}(y_t | \upsilon, \theta_t) = (\upsilon/\theta_t) \exp(z_t' \delta). \qquad (6.2.30)$$

It is obvious from the above that either the shape υ or the state θ_t may be used for this purpose. Hence different estimates of the hyperparameter δ will be produced, according to the chosen linking mechanism.

The use of the shape parameter as a linking mechanism is appropriate when the situation suggests that exogenous variables are associated with changes in the form of the distribution. This being the case the way to proceed is by defining

$$\upsilon_t^{+} = \upsilon \exp(z_t' \delta)_{.} \qquad (6.2.31)$$

Under such an effect the equations for the prior scale parameter (6.2.6a) and (6.2.7a) become

$$a_{t/t-1}^{+} = \omega\, a_{t-1}^{+} + (1-\omega) \qquad (6.2.31a)$$

$$a_t^{+} = a_{t/t-1}^{+} + \upsilon \exp(z_t' \delta). \qquad (6.2.32b)$$

Since the prior shape parameter equations do not include υ, they remain unaltered (see 6.2.6b-6.2.7b). Regarding the projected moments

(6.2.22) these will have υ substituted by $\upsilon_{T+k} = \upsilon \exp(z_{T+k}'\delta)$ and a_T by a_T^+. It is then straightforward to show that the forecast function has the same asymptotic form as that for the Poisson-Gamma model (3.2.23a), i.e.,

$$\tilde{y}_{T+k/T} = \exp(z_{T+k}'\delta)\ \mathrm{EWMA}(y)/\ \mathrm{EWMA}[\exp(z'\delta)]. \qquad (6.2.33)$$

When the effects of explanatory variables/structural components are believed not to alter the basic form of the gamma density the state should be chosen as the linking mechanism. It is then clear that, from (6.2.30) it follows that

$$\theta_t^+ = \theta_t \exp(-z_t'\delta) \qquad (6.2.34)$$

Observe that a similar mechanism has been used for the Poisson-Gamma model the only difference being that here we have to work with the reciprocal of the state instead. At the outset (6.2.34) implies the substitution of θ_{T+k} by $\theta_{T+k}^+ = \theta_{T+k} \exp(-z_{T+k}'\delta)$ in (6.2.14). From the properties of the gamma density, conditional on Y_{t-1}, $\theta_t^+ \sim \mathrm{gamma}(a_{t/t-1}, b_{t/t-1}^+)$, where $a_{t/t-1}$ is as in (6.2.6a) and

$$b_{t/t-1}^+ = \omega\, b_{t-1} \exp(z_t'\delta) \qquad (6.2.35a)$$

which yields the updating equation

$$b_t = \omega\, b_{t-1} + y_t \exp(-z_t'\delta). \qquad (6.2.35b)$$

In face of (6.2.30-31) the appropriate substitution in the forecasting moments (6.2.19) is now given by $b_T^* = b_T \exp(z_{T+k}'\delta)$. The asymptotic form of the forecast function may be shown to be

$$\tilde{y}_{T+k/T} = \exp(z_{T+k}'\delta)\ \mathrm{EWMA}[y \exp(-z'\delta)]. \qquad (6.2.36)$$

If it is not clear from the context which mechanism to use, for predictive purposes, the selected method should be the one that produces the best fit/forecast.

CHAPTER SEVEN

THE RANDOM SUM MODEL

7.1 INTRODUCTION:

In this chapter we set up a time series model for observations which are obtained by the aggregation in time of positive and continous random variables. Since the number of such occurences at a time period t is a random variable itself, the generating mechanism is described by a *random sum*. The framework here developed is particularly useful in the context of insurance claims, where the total value of claims in a given time period is usually regarded as a random sum. Similar problems arise elsewhere. For example, one may be interested in the total expenditure on some category of consumer durables, such as videos, for a given group of the population. In order to aid the exposition we will assume that we are working in an insurance context. The model we set up is based on the adoption of the Gamma-Gamma model (see section 6.2) for the claims size which is then combined with the Poisson—Gamma model (see section 3.2) for the number of claims. In Harvey and Fernandes (1989b) this problem is investigated under the same distributional assumption for the number of claims, but with the lognormal distribution for the claims size. The framework adopted for the latter case is that of the Gaussian structural models which is also extended to cope with several groups.

7.2 THE RANDOM SUM MODEL:

measurement equation- let y_{it} be the amount of the i^{th} claim at time t, where $i=1,\dots,N_t$ and $t=1,\dots,T$. The total value of claims is the random sum

$$S_t = \sum_{i=1}^{N_t} y_{it}, \qquad t=1,2,\dots,T. \qquad (7.2.1)$$

We assume that the individual claims size at time t, y_{it} follow a gamma distribution with fixed shape parameter υ and stochastic scale parameter θ_{1t},i.e. the y_{it} follow exactly the same measurement equation given in (6.2.1). It is then standard that, conditional on the number of claims at time t, N_t ,and the state θ_{1t}, the random sum is also gamma distributed,i.e.,

$$p(S_t|N_t,\upsilon,\theta_{1t}) = \frac{(N_t\theta_{1t})^{\upsilon} \; S_t^{\upsilon-1} \; e^{-N_t\theta_{1t}S_t}}{\Gamma(\upsilon)} \qquad (7.2.2)$$

where $0< S_t <\infty$. Observe that, like y_{it} , N_t, the number of claims at time t is also a random variable. As a probability model for N_t we propose the Poisson-Gamma model of section 3.2 so that conditional on θ_{2t}, a second state parameter, we have

$$p(N_t|\theta_{2t}) = \theta_{2t}^{N_t} e^{-\theta_{2t}} / N_t! \quad , N_t = 0,1,\ldots \tag{7.2.3}$$

The measurement equation is given by the joint distribution of S_t and N_t, conditional on the states θ_{it} i=1,2. Defining $\theta_t = \{\theta_{1t}, \theta_{2t}\}$ this may be expressed by the product

$$P(S_t, N_t|\theta_t) = p(S_t|N_t, \theta_{1t})\ p(N_t|\theta_{2t}). \tag{7.2.4}$$

Bearing in mind that in our methodology the states follow a random walk structure, the simplifying assumptions used to derive (7.2.4) have been based on the following conditional probability statements

(i) that the aggregate claim at time t, S_t depends only on its past values through θ_{1t} and that it is only affected by the current value of the number of claims N_t .
(ii) the distribution of the number of claims at time t is independent of the actual and past values of the aggregate claim, depending only on its past values through θ_{2t}.

While the second assumption is perfectly reasonable within the context of insurance one might find more realistic the aggregate claims to depend also on past values of the number of claims. Contrary to initial expectations the dependence structure for the aggregate claims is not as restrictive as one might think. As we shall see later the updating mechanism for the state θ_{1t} will involve N_t so that, past values of the number of claims will also be fed into the mechanism of S_t. In view of (7.2.2) and (7.2.3) our measurement equation (7.2.4) is given by the product of a gamma and Poisson distribution. The stochastic mechanisms for the states θ_{it} i=1,2 are now briefly described, given that they have already been depicted in the previous chapters.

(i-ii) *state prediction* and *updating*- first some notation. Let $D_t = \{ N_t, S_t \}$, with $N_t = \{N_1, N_2, \ldots, N_t\}$ and $S_t = \{ S_1, S_2, \ldots, S_t\}$.

- θ_{1t}: since $p(\theta_{1t-1} | D_{t-1}) \sim \text{gamma}(c_{t-1}, d_{t-1})$ then $p(\theta_{1t} | D_{t-1}) \sim$ gamma $(c_{t/t-1}, d_{t/t-1})$ where

$$c_{t/t-1} = \omega_1 \, c_{t-1} + (1-\omega_1) \qquad (7.2.5a)$$

$$d_{t/t-1} = \omega_2 \, d_{t-1} \qquad (7.2.5b)$$

with $0 < \omega_1 < 1$. The posterior will also be gamma with parameters given by

$$c_t = c_{t/t-1} + N_t \upsilon \qquad (7.2.6a)$$

$$d_t = d_{t/t-1} + S_t . \qquad (7.2.6b)$$

- θ_{2t}: $p(\theta_{2t-1} | N_{t-1}) \sim \text{gamma}(a_{t-1}, b_{t-1})$ and then $p(\theta_{2t} | N_{t-1})$ will be also gamma with parameters

$$a_{t/t-1} = \omega_2 \, a_{t-1} \qquad (7.2.7a)$$

$$b_{t/t-1} = \omega_2 \, b_{t-1} \qquad (7.2.7b)$$

with $0 < \omega_2 < 1$. The updating equations are then given by

$$a_t = a_{t/t-1} + N_t \qquad (7.2.8a)$$

$$b_t = b_{t/t-1} + 1. \qquad (7.2.8b)$$

(iii) *conditional distribution and likelihood*- a proper time series modelling approach would be based on the joint predictive distribution of N_t and S_t and this may be evaluated by the convolution

$$p(S_t,N_t|D_{t-1}) = p(S_t,N_t|\theta_t) \underset{\theta_t}{\triangle} p(\theta_t|D_{t-1}). \qquad (7.2.9a)$$

We further assume that the states are independent processes, so that in view of (7.2.5-8) the second density on the rhs of (7.2.9a) may be expressed as

$$p(\theta_t|D_{t-1}) = p(\theta_{1t}|D_{t-1})\ p(\theta_{2t}|N_{t-1}) \qquad (7.2.10)$$

where the marginal densities in the above expression are given in the previous item. Observe that in view of (7.2.6a) past values of N_t are fed back into the distribution of θ_{1t}, so that assuming its independence from θ_{2t} is not crucial. Using (7.2.4) and the above expression one may easily deduce that the convolution in (7.2.9) results in a product of known univariate predictive distributions, i.e.,

$$p(S_t,N_t|D_{t-1}) = p(S_t|N_t,D_{t-1})\ p(N_t|N_{t-1}) \qquad (7.2.9b)$$

where the first density in the inverted-beta-2 given in (6.2.8) and the second is the NBD in (3.2.7). Substituting them in the above expression one obtains

$$P(S_t,N_t|D_{t-1})= \frac{\Gamma(a+N_t)\; b^a \; d^c \; S_t^{N_t\nu-1} \; (1+b)^{-(N_t+a)}}{\Gamma(a) \; (S_t+d)^{N_t\nu+c} \; B(N_t\nu,c)} \qquad (7.2.9c)$$

with $0 < S_t < \infty$, $N_t= 0,1,2,\ldots$, $a=a_{t/t-1}$, $b=b_{t/t-1}$, $c=c_{t/t-1}$ and $d=d_{t/t-1}$. It is simple to show that the logarithm of the bivariate likelihood function may be factorized as in (5.2.24), so that the optimization problem is split into the maximization of two separate likelihoods, one with respect to ω_1, using a inverted-beta distribution and the other with respect ω_2, using a NBD distribution.

Predictive moments for N_t and for S_t, conditional on N_t, are readily available by appropriate use of the corresponding expressions in sections (3.2) and (6.2), namely equations (3.2.11a-b) and (6.2.10a-b). With respect to the forecast function for S_t, it may be shown that the recursions (7.2.5a-b) and (7.2.6.a-b) produce, for $0 < \omega_1 < 1$, an EWMA scheme having the form

$$\tilde{S}_{t/t-1}= E(S_t|N_t,D_{t-1})= \mathrm{EWMA}(S)/\ \mathrm{EWMA}(N).$$

7.2.1 Unconditional distribution of the Random Sum

It may be of interest to predict the overall claims when the future value of the number of claims is unknown. This being the case one has to evaluate the unconditional predictive distribution for S_t, which is obtained by summing out the joint distribution (7.2.9c) wrt N_t. It is not difficult to see that this operation yields

$$P(S_t | D_{t-1}) = \frac{d^c \, (b/1+b)^a}{S_t \, (S_t+d)^c \, \Gamma(c)} \sum_{N_t=0}^{\infty} \frac{W_t^{N_t} \, \Gamma(N_t \upsilon+c)\Gamma(N_t+a)}{\Gamma(N_t \upsilon) \, \Gamma(a) \, N_t!} \tag{7.2.11}$$

with $W_t = (S_t/S_t+d)^{\upsilon}(1/1+b)$. Unconditional moments for S_t can be obtained by using that

$$E(S_t | D_{t-1}) = \underset{N_t/D_{t-1}}{E} (E (S_t | N_t, D_{t-1})) \tag{7.2.12a}$$

and

$$Var(S_t | D_{t-1}) = \underset{N_t/D_{t-1}}{E} (Var(S_t | N_t, D_{t-1})) + \underset{N_t/D_{t-1}}{Var} (E (S_t | N_t, D_{t-1})). \tag{7.2.12b}$$

In view of assumption (ii) in section 7.2 it follows that

$$E(N_t | D_{t-1}) = E(N_t | N_{t-1}) \tag{7.2.13}$$

so that using (7.2.10a-b) one may compute the above moments which are given, repectively by

$$E(S_t | D_{t-1}) = \frac{\upsilon \, d}{(c \ \ -1)} \ \frac{a}{b} \quad , \qquad c > 1 \tag{7.2.14a}$$

and

$$\mathrm{Var}(S_t|D_{t-1}) = E(S_t|D_{t-1}) \frac{\{\ [\ (c-1)\ a\ b + \upsilon\ (a+b+1\ )]}{(c\ -1)\ (c\ \ -2)} \quad +$$

$$+ \quad \frac{\upsilon\ d\ a\ (1+b)\ \ \}}{b\ (c\ -\ 1)} \qquad c > 2. \qquad (7.2.14b)$$

One may also consider the introduction of explanatory variables /structural components at the forecasting mechanisms for S_t and N_t. See sections (6.2.2) and (3.2.3) respectively.

7.2.2 Crosscovariance

The crosscovariance between S_t and N_t may also be obtained by evaluating the terms on the formula below

$$\mathrm{Cov}(S_t,N_t|D_{t-1}) = E(S_tN_t|D_{t-1}) - E(S_t|D_{t-1})E(N_t|D_{t-1}) \qquad (7.2.15)$$

The conditional means on the second term on the rhs of the above expression are given respectively in (7.2.14a) and (3.2.10). In order to evaluate the cross product term we use that

$$\begin{aligned} E(S_tN_t|D_{t-1}) &= \underset{N_t/D_{t-1}}{E} [\ E\ (S_tN_t|N_t,D_{t-1})] \\ &= \underset{N_t/D_{t-1}}{E} [\ N_t\ E(S_t|N_t,D_{t-1})] \\ &= \upsilon\ d\ (c-1)^{-1}\ E\ (N_t^2|N_{t-1}) \end{aligned}$$

$$= \frac{\upsilon\, d\, a\, (1+a+b)}{(c-1)\, b^2} \quad , c > 1 \tag{7.2.16}$$

where we have used (6.2.10a) and (7.2.13). With the above result it is straightforward to show that the crosscovariance between the number of claims and the aggregate claims has the final expression

$$\mathrm{Cov}(S_t, N_t) = \frac{\upsilon\, d\, a\, (1+b)}{(c-1)\, b^2} \quad c > 1 \tag{7.2.17}$$

Observe that this will always be positive as one might expect. Forecasts may be obtained by combining the results derived for the Poisson-Gamma model (see Ch.3, section 3.2) with those of the Gamma-Gamma model (see Ch.6, section 6.1).

CHAPTER EIGHT

MONTE CARLO EXPERIMENTS

8.1 INTRODUCTION:

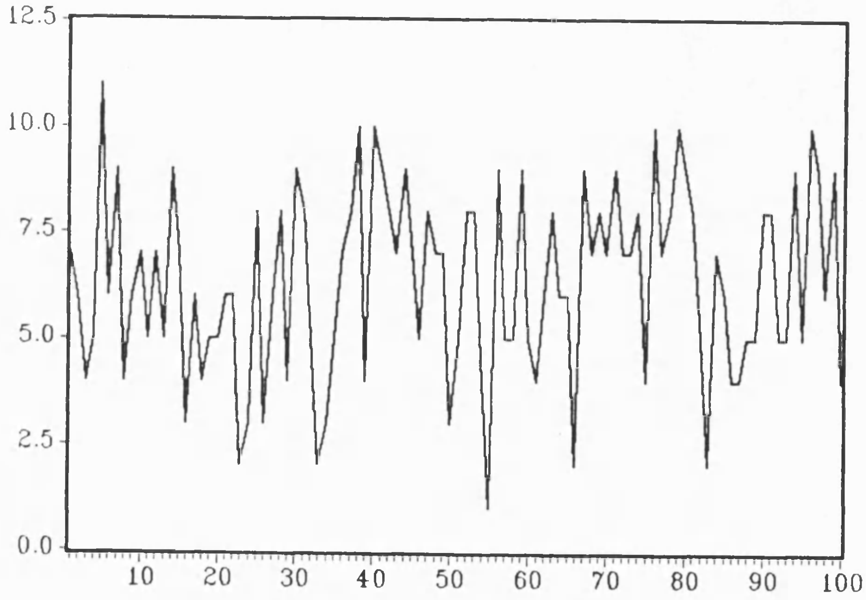
This chapter presents the results of Monte Carlo experiments conducted on the Poisson-Gamma model of Chapter 3. The purpose of our experiments is to obtain approximate answers for certain problems involving our model for which analytical and/or numerical solutions are difficult to obtain. In particular we have considered the following topics in our investigation:

(i) the small sample and asymptotic properties of the ML estimators of hyperparameters, namely, the discount and the regression parameters.

(ii) the size and power of the post-sample predictive test derived in the appendix A1.

The first part of our study was conducted on data generated from a discount only or standard Poisson-Gamma model, where the data forecasting mechanism (DFM) is asymptotically equivalent to the usual EWMA scheme as given in equation (3.2.13). Here it is of interest to consider values of ω one is most likely to obtain when fitting the model to real data, that is $\omega \in \{0.85, 0.90, 0.95, 0.98\}$. As an illustration we reproduce a simulated series of length 100 when ω is fixed at 0.98. The details of the generating process will be described later.

Figure 8.1.1 Simulated series of a Poisson-Gamma model with ω=0.98.



The same topics are investigated for specifications of our model in which a regressor variable is included, and this has been taken as Gaussian white noise. In these experiments we have fixed the regressor parameter at δ=0.04 and the discount value at ω=0.90 and ω=0.95. The other features of our Monte Carlo experiments are:

- sample size: we consider small, moderate and large sample sizes, where T ϵ {30, 50, 80, 100, 300, 500, 700}. The two largest sample sizes values are only used in simulations in which $\omega > 0.95$, i.e., when the discount parameter is sufficiently close to its upper bound value one. The reason for this choice will become clear later.

- number of replications: the number of replications, Nrep, has been set at 1000, for sample sizes $T < 100$, and to 300 otherwise.

- filter start-up: in order to start the generation process one has to provide values for the gamma prior at time t=0. We have found that appropriate values are given by a_0= 10 and b_0=1. A 'warming up' period was considered, by dropping the first 50 observations.

Generation of NBD deviates:

In the present context, the natural way of generating NBD deviates is to generate a gamma($a_{t/t-1}$,$b_{t/t-1}$) deviate θ_t and then generate a Poisson deviate with parameter θ_t. The resulting variable is by definition NBD($a_{t/t-1}$,$b_{t/t-1}$) as in (3.2.10). We have initially adopted such a procedure in our experiments. The difficulty with this technique is the non availability of ready-to-use routines for gamma deviates with a non-integer scale parameter, e.g.,the NAG library (1984) does not allow for such a possibilty. This problem was circumvented by adopting the algorithm GBH of Cheng and Feast (1980) which is appropriate for gamma deviates with shape parameter greater than 0.25. The subsequent generation of Poisson deviates presents no difficulty, and at this stage we had employed the NAG routine G05ECF. This procedure was later on abandoned in favour of a more direct and time efficient routine for NBD deviates provided by the IMSL library (1987), the routine RNBBN. This routine is based both on the above technique and the inverse CDF method, the chosen method depending on the range of the parameters involved. The generation of the Gaussian white noise variables used in the version of the model which includes explanatory variables was conducted using the IMSL routine DRRNOA. Note that the random number generator provided in the IMSL library is of the type multiplicative congruential with multiplier a= 950.706.376 and modulus m= 2^{31}-1.

8.2 HYPERPARAMETERS ESTIMATES EMPIRICAL DISTRIBUTION:

Small sample and asymptotic properties of the ML estimators of Poisson–Gamma hyperparameters are investigated using both formal and graphical techniques. The formal techniques adopted in our study are based in some of the best known tests designed to check the adequacy of the normal distribution as a model for a data set. Besides the commonly used Bowman-Shenton test (see, e.g., Jarque and Bera 1987) we have implemented some of the omnibus tests suggested in D'Agostino and Stephens (1986, ch.9). In what follows we provide a brief description of these techniques.

- Anderson-Darling A^2 (AD): this is one of the most powerful tests based on the empirical distribution function (EDF), and, according to D'Agostino and Stephens(1986, p.406), has far better power than the popular Kolmogorov-Smirnov test. The AD statistics is a member of the Cramer-von Mises family of goodness-of-fit statistics based on the EDF, which has the following general expression:

$$W= \int_{-\infty}^{\infty} [\ F_n(x) - F(x;\theta)\ ]^2\ \psi(x)\ dx$$

where $F_n(x)$ is the EDF, $F(x;\theta)$ is the continuous theoretical distribution to be tested, and $\psi(x)$ is a weighting function. The AD statistics is obtained by making $\psi(x)= [\ F(x;\theta)\ (\ 1 - F(x,\theta)\ )\ ]^{-1}$. This weight function gives greater importance to observations in the tail than do other EDF tests, counterbalancing the fact that $F_n(x) - F(x;\theta)$ approaches zero in these regions. The numerical implementation of this test is well documented in D'Agostino and Stephens(1986, pp.372-374). The null hyphothesis of normality is rejected whenever the calculated test satistic exceeds the critical values which are reproduced in the table below.

Table 8.2.1 Critical values for the A-D statistics.

size	critical value
1 %	1.035
5 %	0.752
10 %	0.631

- Bowman-Shenton (BS): this is a test based on the distributions of the skewness ($\surd b_1$) and kurtosis (b_2) coefficients. It has been particularly used to check residuals normality in several econometrics/time series packages commonly available as the PC-GIVE, DATA-FIT and STAMP. The test statistics of BS is given by

$$BS= (\surd b_1)^2 / \sigma_1^2 + (b_2-3)^2/ \sigma_2^2$$

where $\sigma_1^2= 6/Nrep$ and $\sigma_2^2= 24/Nrep$. Asymptotically BS ~ χ^2 (2) if the null hyphothesis of normality is true. As remarked by D'Agostino and Stephens (1986, p. 389) the normal approximation for the kurtosis distribution is only valid for extremely large sample sizes (Nrep in our context) well over 1000. Jarque and Bera (1987, p.169) provide a table of the significance points for the 'true' distribution of the BS test based in simulation studies. This is reproduced below for the sample sizes of our interest.

Table 8.2.2 Critical values for the BS statistic.

T	5 %	10 %
30	3.71	2.49
50	4.26	2.90
75	4.27	3.09
100	4.29	3.14
300	4.60	3.68
500	4.82	3.91
∞	5.99	4.61

- D'Agostino-Pearson (AP): this test is claimed to produce a more accurate approximation for the distributions of the skewness and kurtosis coefficients and these are based, respectively on the Johnson S_u curve and on the Anscombe and Glynn approximation (see D'Agostino and Stephens 1986 ch.9). The resulting test statistics, by construction, is also $\chi^2(2)$. As we shall see, our simulation results seem to indicate that this test is only superior to the BS test for moderate and large sample si(s. In other words its supposed higher sensitivity to non-normality is only displayed in situations in which the BS test already works quite reasonably. For reference purposes we reproduce the critical values of the $\chi^2(2)$ below.

Table 8.2.3 Critical values for $\chi^2(2)$.

size	critical value
1 %	9.209
5 %	5.991
10 %	4.605

In conjunction with the normality tests previously described we have used graphical techniques based on the histogram and the normal probability plot. As one knows if the normal distribution is the true distribution of the estimators, then, to within sampling error, the plot of the ML estimates-quantiles versus normal-quantiles will be a straight line with non-zero location and slope different from one (see, e.g., Wilk and Gnanadesikan 1968).

One should observe that given that the values of interest for the discount hyperparameter are close to its upper bound of unity, a certain proportion of the ML estimates obtained in the simulation study will inevitably coincide with this boundary value, even when the true ω is set to a different value. This is the equivalent of having the SNR, q, estimated as zero in the Gaussian local level model (see 3.2.25). Shephard and Harvey (1990) (henceforth S&H) derived for this specification an approximated expression for the probability of estimating q to be zero when its true value is set to a range of values. The main conclusions drawn from their study were:

(i) that the ML sampling distribution for q is highly sensitive to the filter initial conditions.

(ii) that the probability of obtaining estimates equal to zero decreases both as the sample size grows and as the true value of q is set further away from the boundary value.

Since in our framework the filter initial conditions are unambiguously defined the first of these conclusions will not be relevant here. However, when the true value of ω is less than one, given consistency of ML estimators, one would expect to observe a decrease in the proportion of boundary estimates as the sample size grows. Clearly this same proportion is expected to increase the closer the true value of ω is set to one. As a result of the 'boundary effect' the discount distribution may be looked at as being split in two parts: one discrete, giving the probability of $\omega=1$ estimates and the other continuous, for estimates less than one. We have excluded the boundary ML estimates from the summary statistics and normality tests and denoted its proportion by p(1) in the tables that follow.

Table 8.2.4 Descriptive statistics and normality tests for empirical ML estimates of a Poisson-Gamma model with ω=0.85.

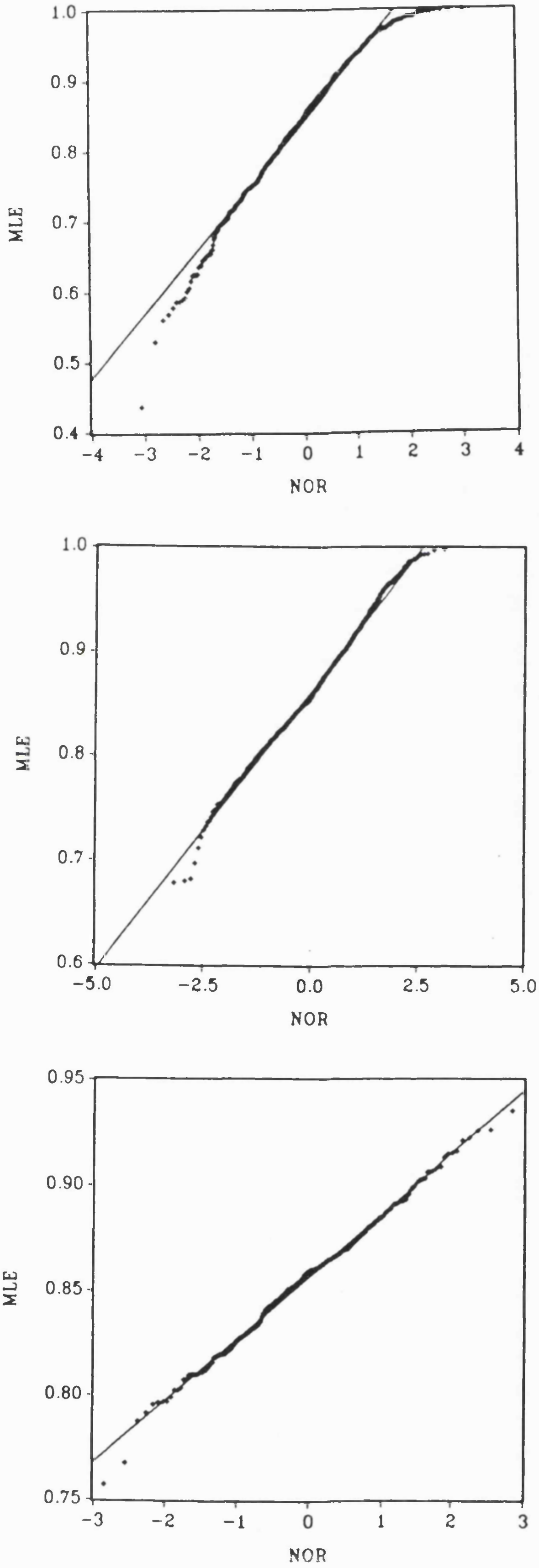
Descriptive Statistics

T	mean	bias	std.dev	skew	kurto	p(1)
30	0.840	0.010	0.092	-0.583	3.339	0.332
50	0.854	-0.004	0.071	-0.261	3.215	0.199
80	0.860	-0.010	0.060	-0.031	2.744	0.078
100	0.860	-0.010	0.053	0.068	3.103	0.049
300	0.856	-0.006	0.029	-0.137	3.111	0.000

Normality Tests

T	AD	BS	AP
30	2.629	40.996	36.443
50	0.773	10.609	10.548
80	0.511	2.661	3.1626
100	1.204	1.162	1.306
300	0.439	1.097	1.325

Figure 8.2.1 Normal probability plots for empirical ML estimates of a Poisson-Gamma model with ω=0.85.



(a) T=30

(b) T=100

(c) T=300

Figure 8.2.2 Sampling distributions for empirical ML estimates of a Poisson-Gamma model with ω=0.85.

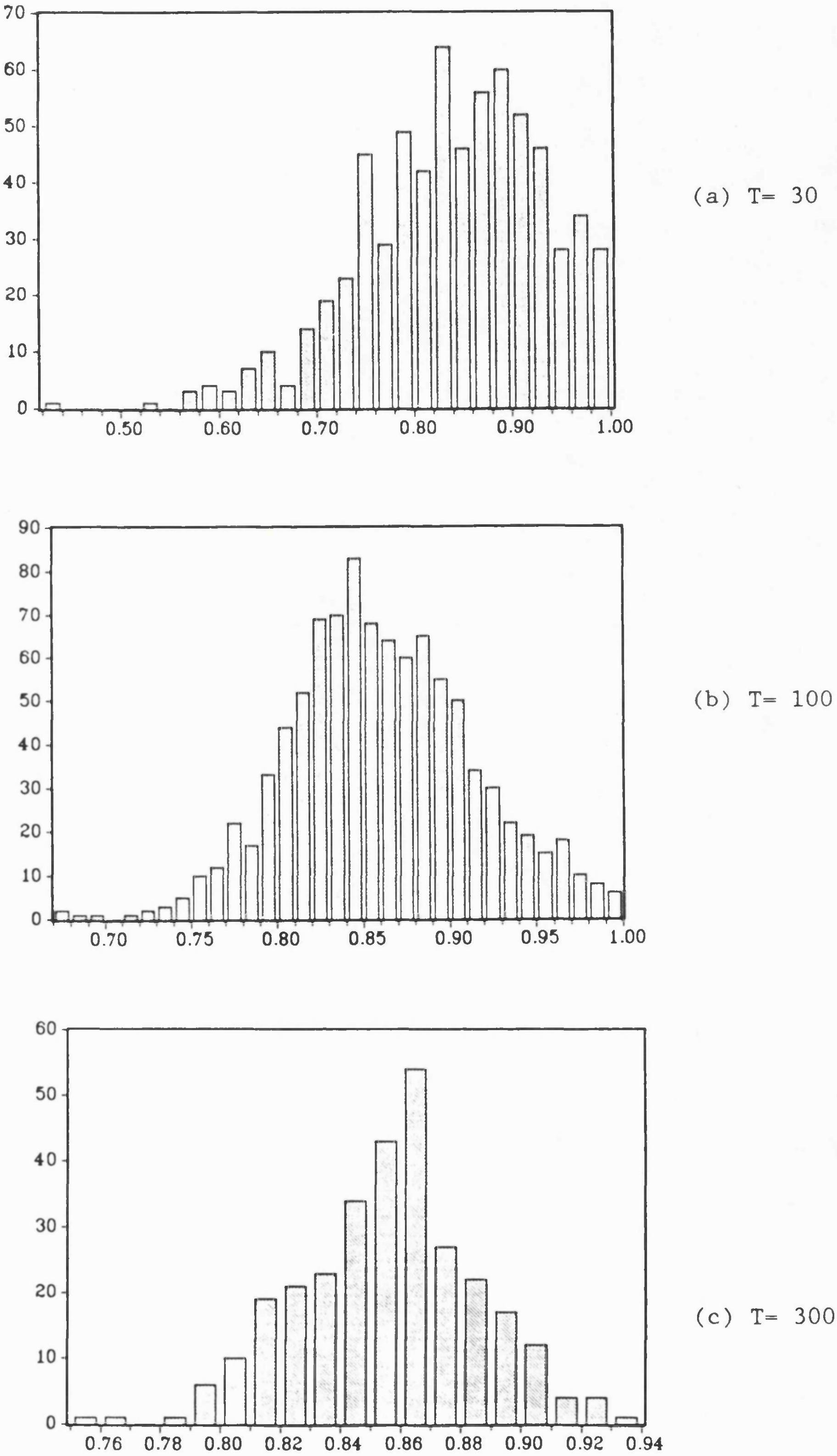


Table 8.2.5 Descriptive statistics and normality tests for empirical ML estimates of a Poisson-Gamma model with ω=0.90.

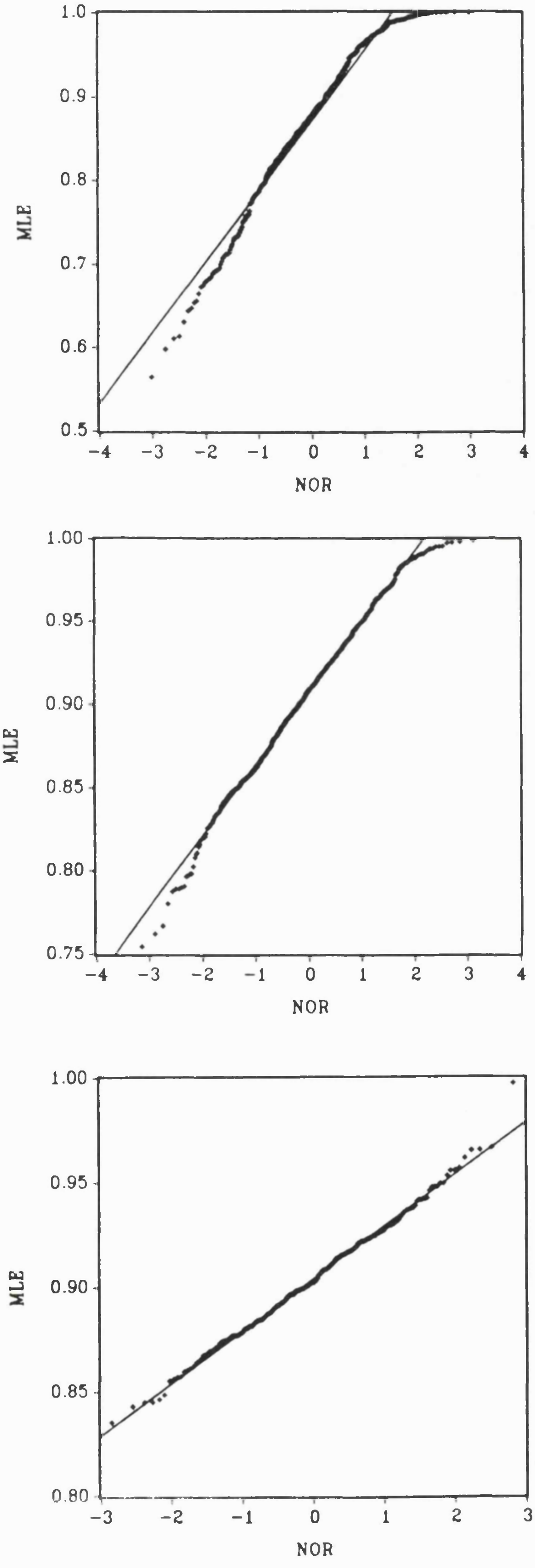
Descriptive Statistics

T	mean	bias	std.dev	skew	kurto	p(1)
30	0.869	0.031	0.085	-0.633	3.083	0.448
50	0.894	0.006	0.063	-0.648	3.664	0.320
80	0.905	-0.005	0.050	-0.321	2.815	0.177
100	0.906	-0.006	0.042	-0.231	2.983	0.122
300	0.903	-0.003	0.025	0.157	3.315	0.000

Normality Tests

T	AD	BS	AP
30	4.028	37.072	32.543
50	2.636	60.130	49.236
80	1.609	15.313	14.902
100	0.619	7.822	7.745
300	0.225	2.478	2.738

Figure 8.2.3 Normal probability plots for empirical ML estimates of a Poisson-Gamma model with ω=0.90.

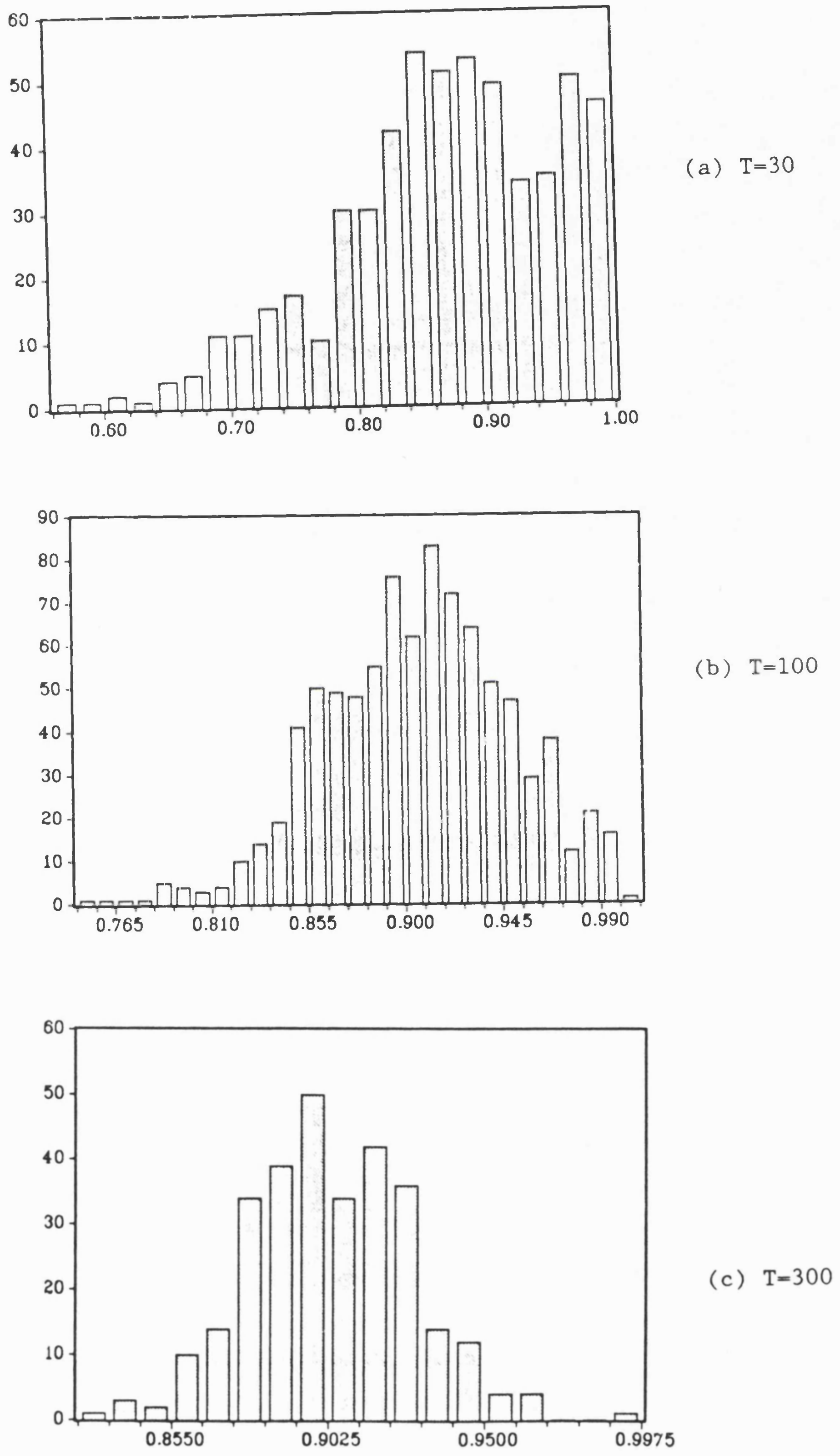


(a) T=30

(b) T=100

(c) T=300

Figure 8.2.4 Sampling distributions for empirical ML estimates of a Poisson-Gamma model with ω=0.90.



(a) T=30

(b) T=100

(c) T=300

Table 8.2.6 Descriptive statistics and normality tests for empirical ML estimates of a Poisson-Gamma model with ω=0.95.

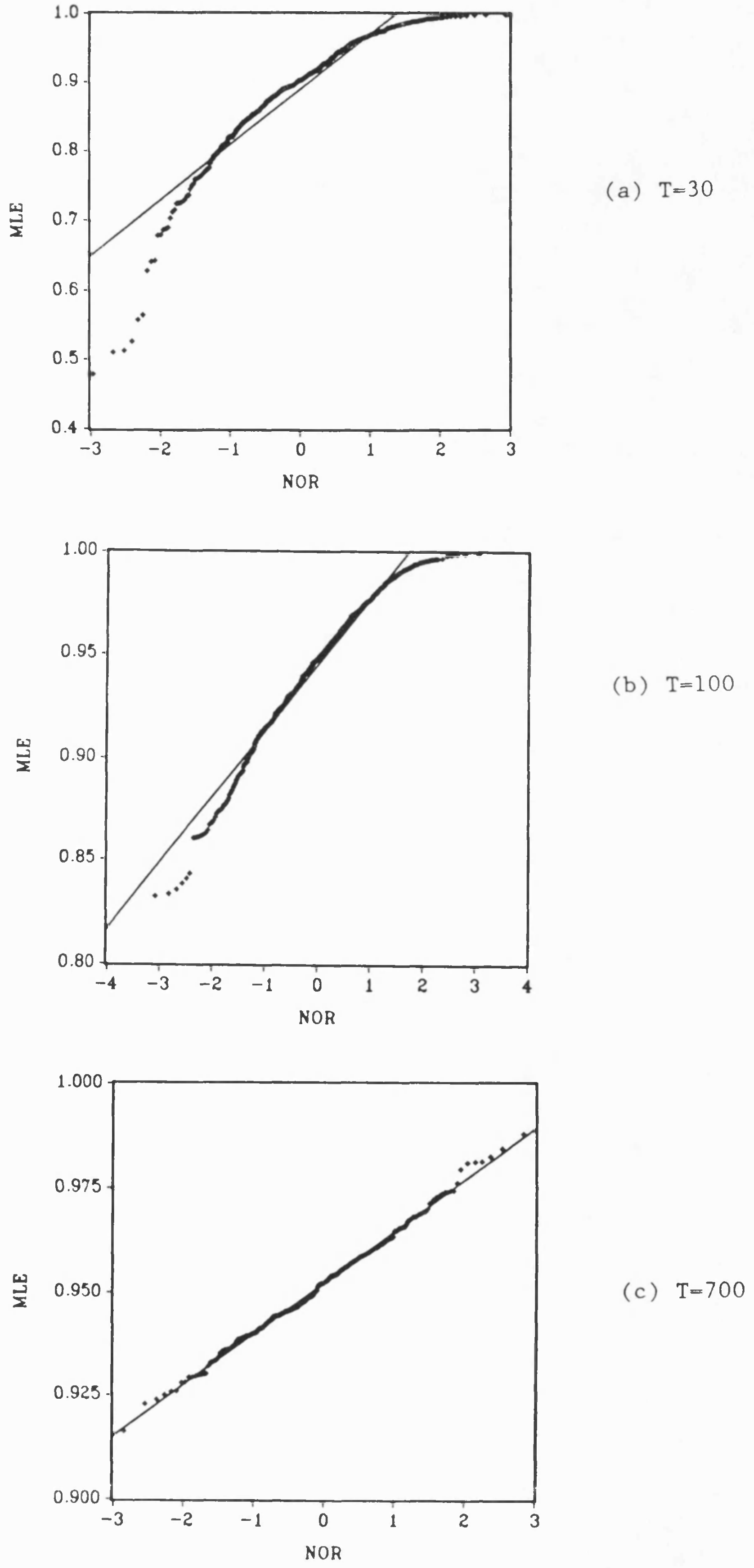
Descriptive Statistics

T	mean	bias	std.dev	skew	kurto	p(1)
30	0.888	0.062	0.084	-1.555	6.765	0.559
50	0.926	0.024	0.056	-1.151	4.504	0.499
80	0.941	0.008	0.038	-0.755	3.330	0.426
100	0.944	0.005	0.032	-0.671	3.320	0.310
500	0.952	-0.002	0.015	-0.025	3.028	0.006
700	0.952	-0.002	0.012	0.113	3.028	0.000

Normality Tests

T	AD	BS	AP
30	9.406	438.346	143.888
50	9.628	157.947	94.524
80	5.689	57.149	47.514
100	4.202	54.693	46.892
500	0.135	0.042	0.125
700	0.329	0.651	0.760

Figure 8.2.5 Normal probability plots for empirical ML estimates of a Poisson-Gamma model with ω=0.95.

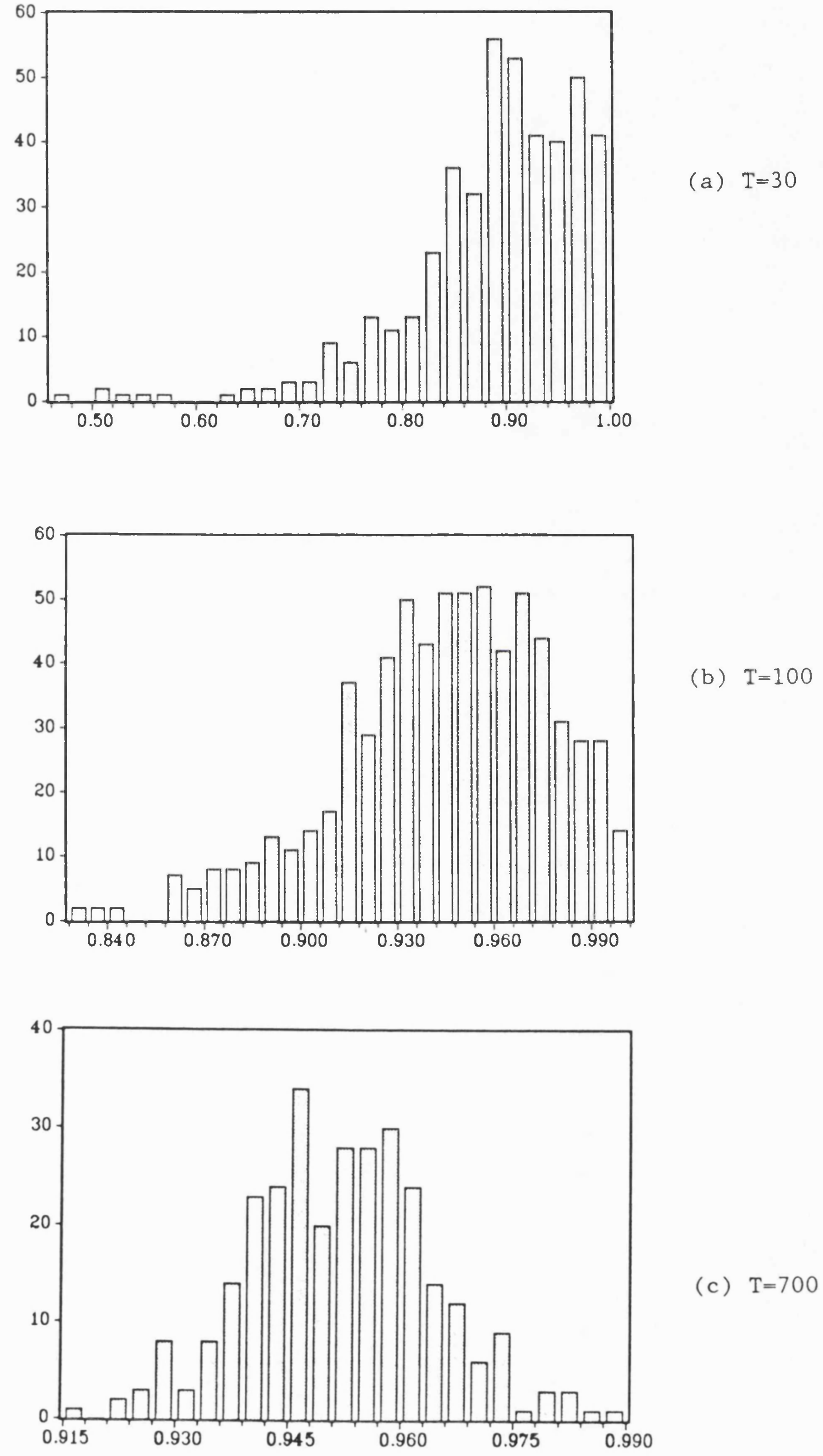


(a) T=30

(b) T=100

(c) T=700

Figure 8.2.6 Sampling distributions for empirical ML estimates of a Poisson-Gamma model with ω=0.95.



(a) T=30

(b) T=100

(c) T=700

Table 8.2.7 Descriptive statistics and normality tests for empirical ML estimates of a Poisson-Gamma model with 0.98.

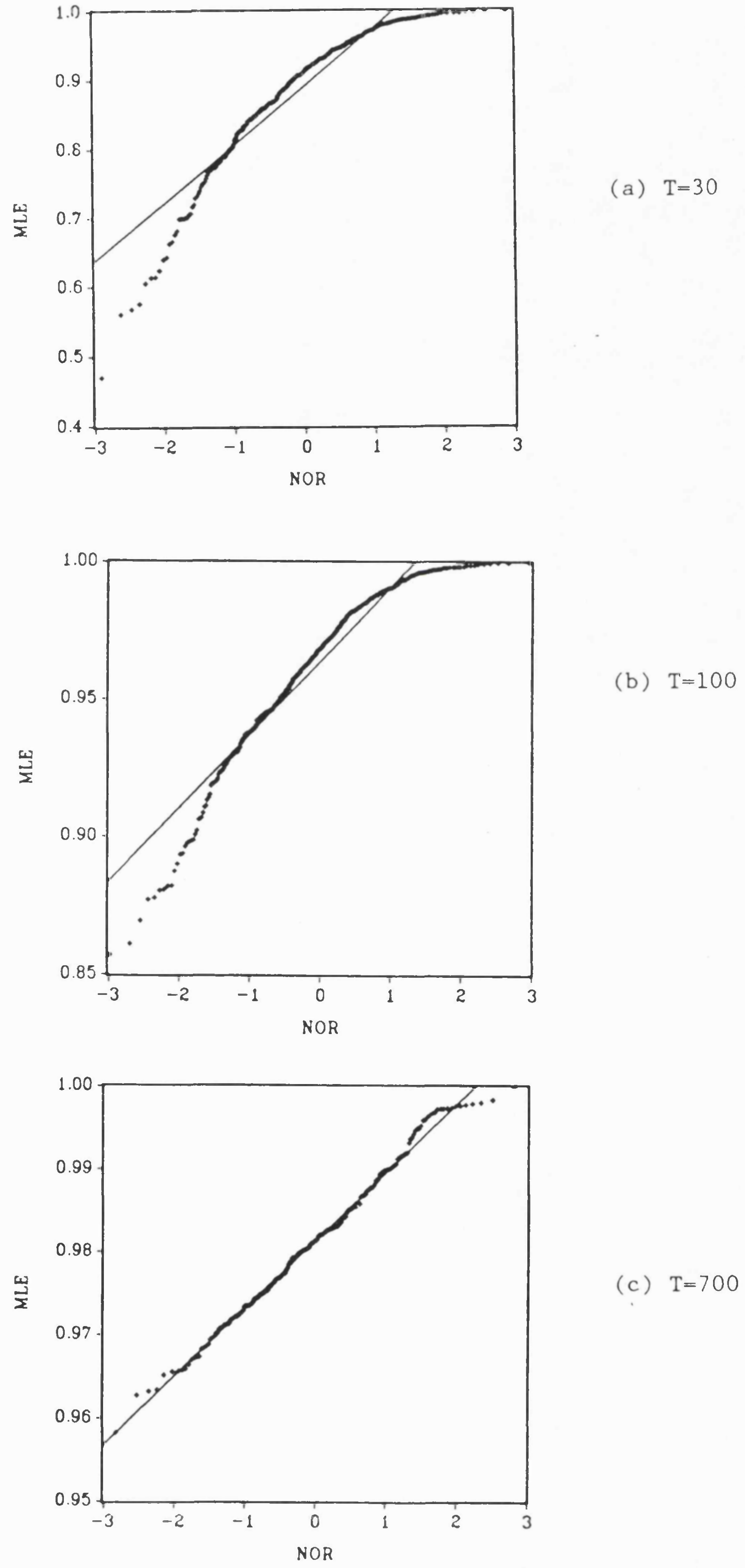
Descriptive Statistics

T	mean	bias	std.dev	skew	kurto	p(1)
30	0.889	0.091	0.089	-1.402	5.383	0.622
50	0.932	0.047	0.051	-1.254	4.697	0.586
80	0.959	0.021	0.033	-1.224	4.553	0.580
100	0.963	0.016	0.027	-1.040	4.032	0.532
500	0.980	-0.000	0.010	-0.293	2.765	0.150
700	0.981	-0.001	0.008	0.003	2.676	0.007

Normality Tests

T	AD	BS	AP
30	10.047	213.375	100.925
50	9.642	158.153	89.703
80	9.792	147.124	86.706
100	8.103	105.181	72.254
500	0.546	4.236	4.167
700	0.418	1.251	1.337

Figure 8.2.7 Normal probability plots for empirical ML estimates of a Poisson-Gamma mode with ω=0.98.

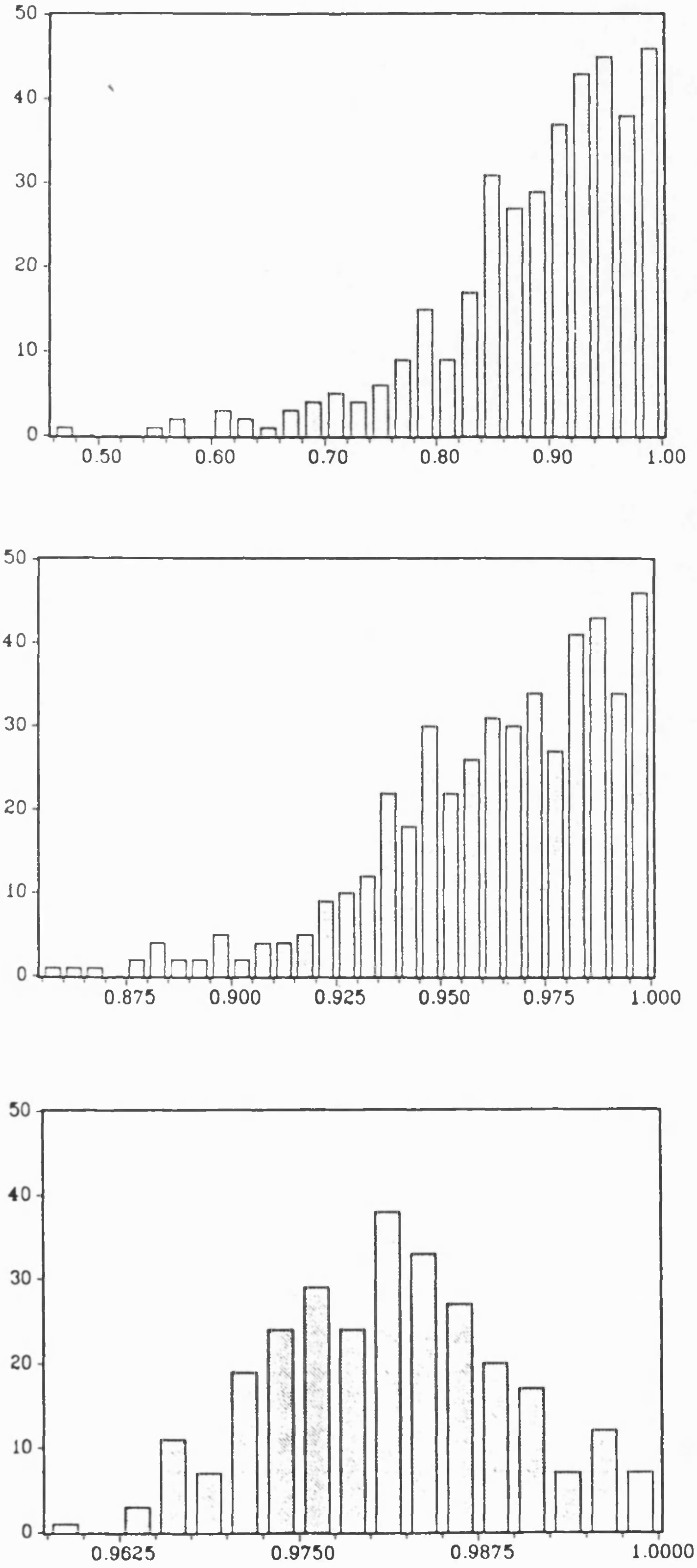


(a) T=30

(b) T=100

(c) T=700

Figure 8.2.8 Sampling distributions for empirical ML estimates of a Poisson-Gamma model with ω=0.98.



(a) T=30

(b) T=100

(c) T=700

Table 8.2.8 Descriptive statistics and normality tests for empirical ML estimates of a Poisson-Gamma model with ω=0.90 and δ=0.04.

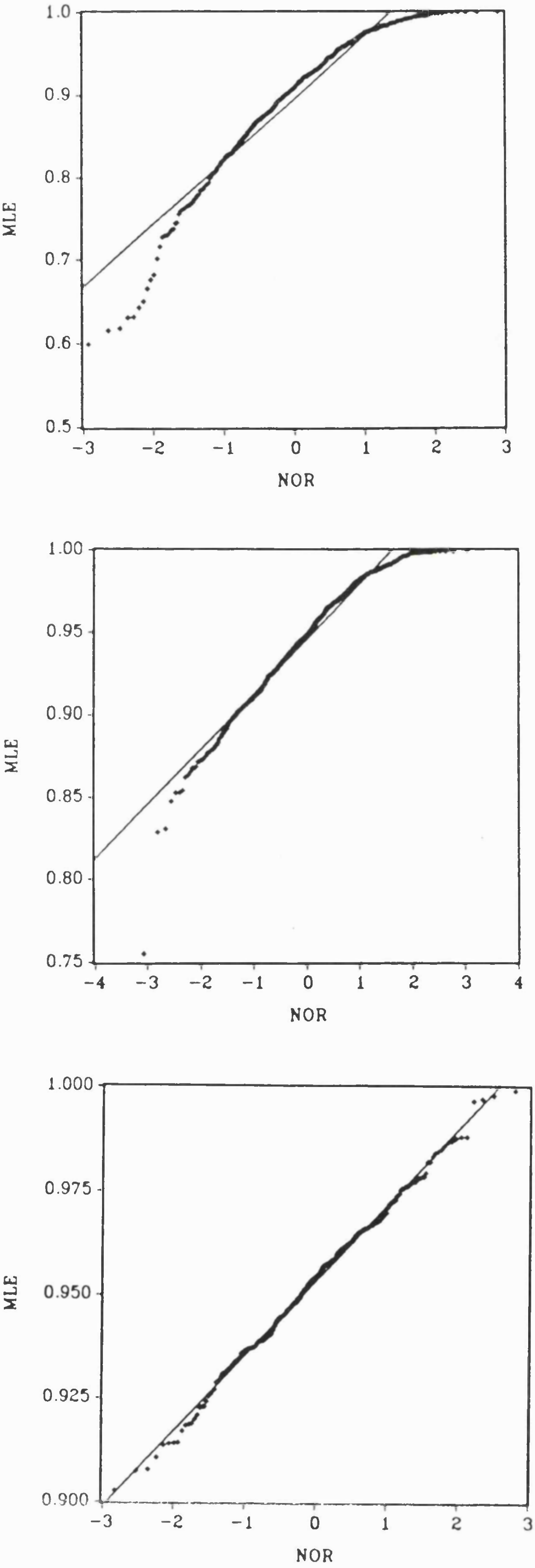
Descriptive Statistics

T		mean	bias	std.dev	skew	kurto	p(1)
30	ω	0.879	0.021	0.078	-0.935	4.425	0.488
	δ	0.041	-0.001	0.072	0.347	5.818	—
50	ω	0.893	0.006	0.062	-0.827	4.651	0.343
	δ	0.040	0.000	0.055	-0.018	5.146	—
80	ω	0.906	-0.006	0.051	-0.448	3.158	0.186
	δ	0.042	-0.002	0.042	-0.118	3.497	—
100	ω	0.909	-0.009	0.043	-0.266	3.173	0.133
	δ	0.004	-0.000	0.037	-0.026	3.866	—
300	ω	0.907	-0.007	0.025	0.347	3.526	0.066
	δ	0.004	0.000	0.021	-0.155	4.020	—

Normality Tests

T		AD	BS	AP
30	ω	4.096	117.926	74.553
	δ	3.275	351.034	85.034
50	ω	2.756	149.521	85.658
	δ	3.365	191.978	50.150
80	ω	1.673	28.128	26.407
	δ	1.355	12.606	9.663
100	ω	0.671	11.283	11.236
	δ	1.256	31.337	16.647
300	ω	0.637	9.433	9.010
	δ	1.107	14.211	8.882

Figure 8.2.11 Normal probability plots for empirical ML estimates of a Poisson-Gamma model with ω=0.95 and δ=0.04.



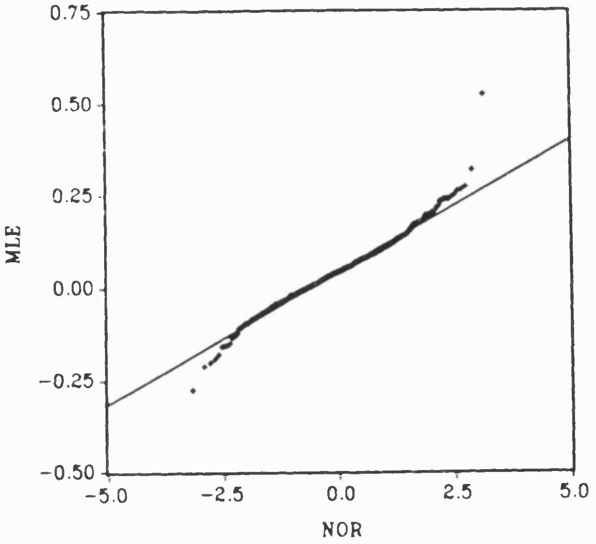
(a) ω

(i) T=30

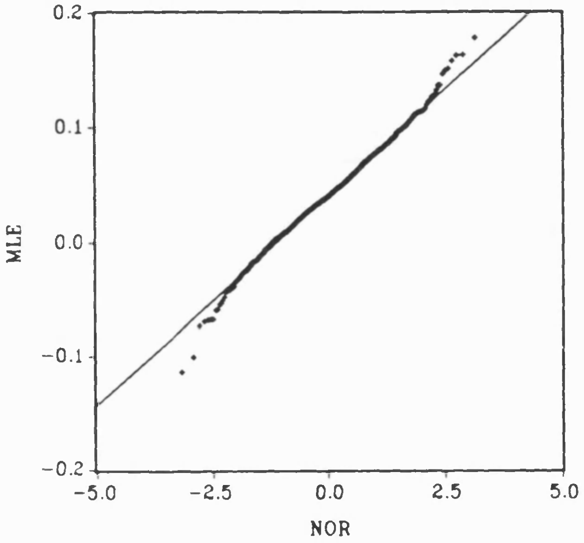
(ii) T=100

(iii) T=300

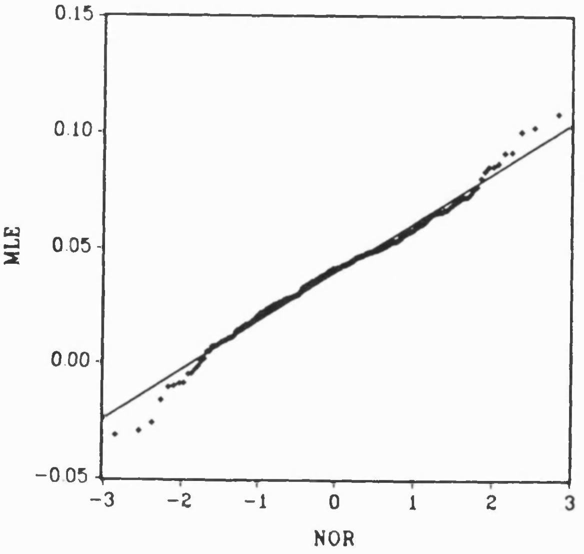
(b) δ



(i) T=30

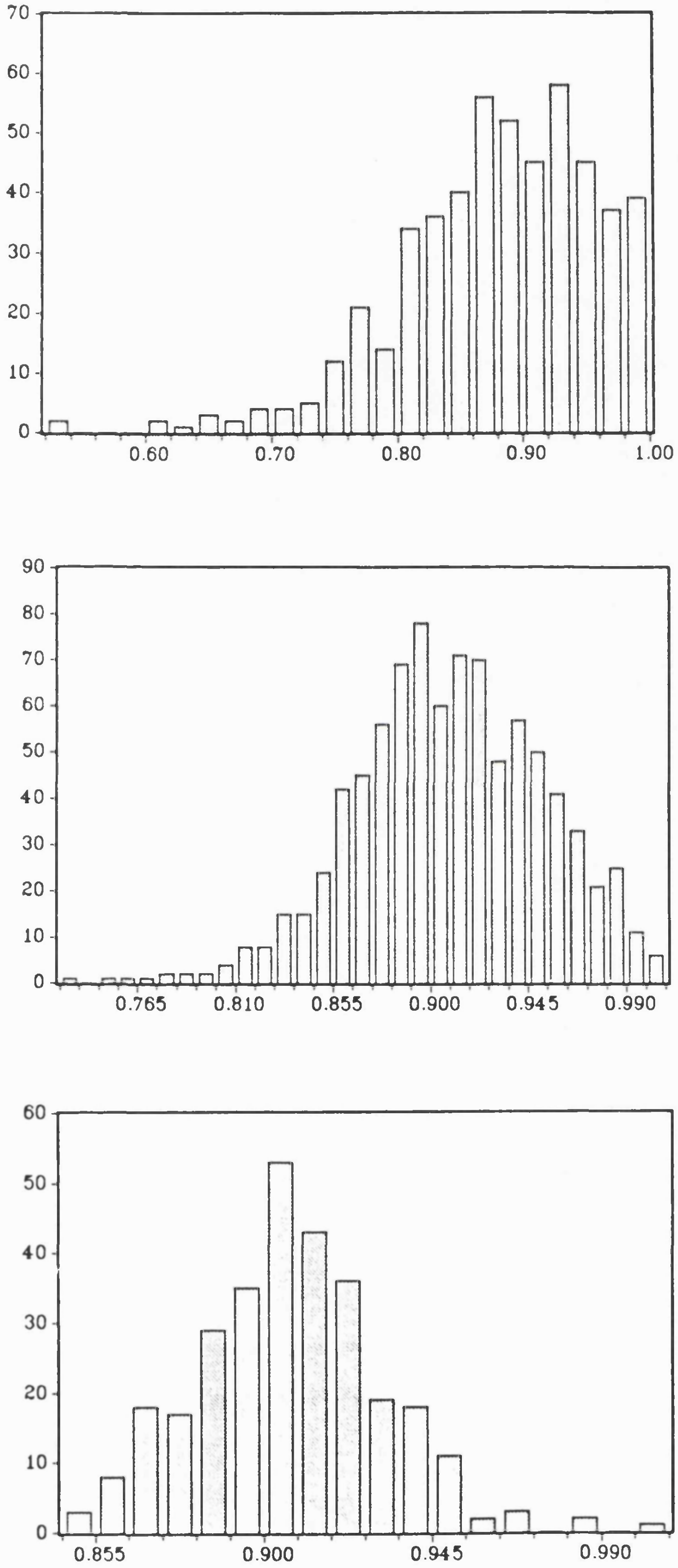


(ii) T=100



(iii) T=300

Figure 8.2.10 Sampling distributions for empirical ML estimates of a Poisson-Gamma model with ω=0.90 and δ=0.04.



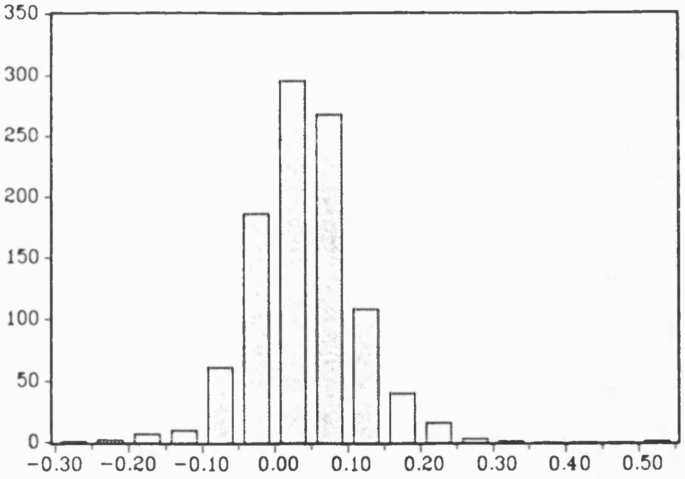
(a) ω

(i) T=30

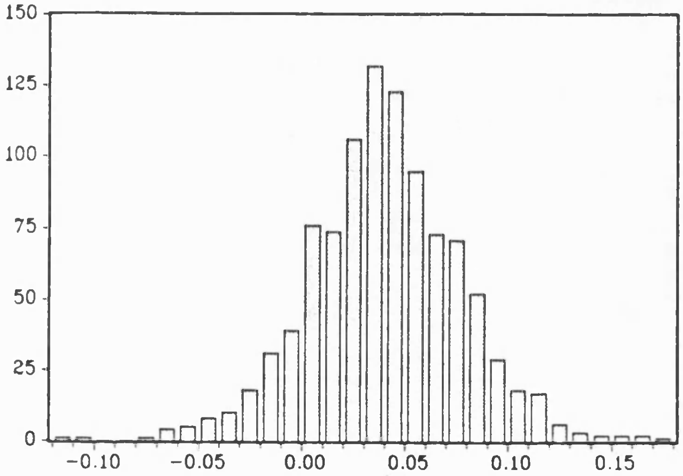
(ii) T=100

(iii) T=300

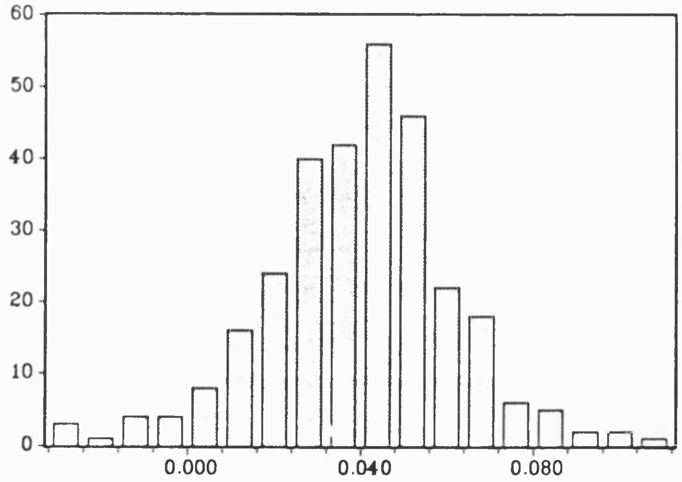
(b) δ



(i) T=30



(ii) T=100



(iii) T=300

Table 8.2.9 Descriptive statistics and normality tests for the empirical ML estimates of a Poisson-Gamma model with ω=0.95 and δ=0.04.

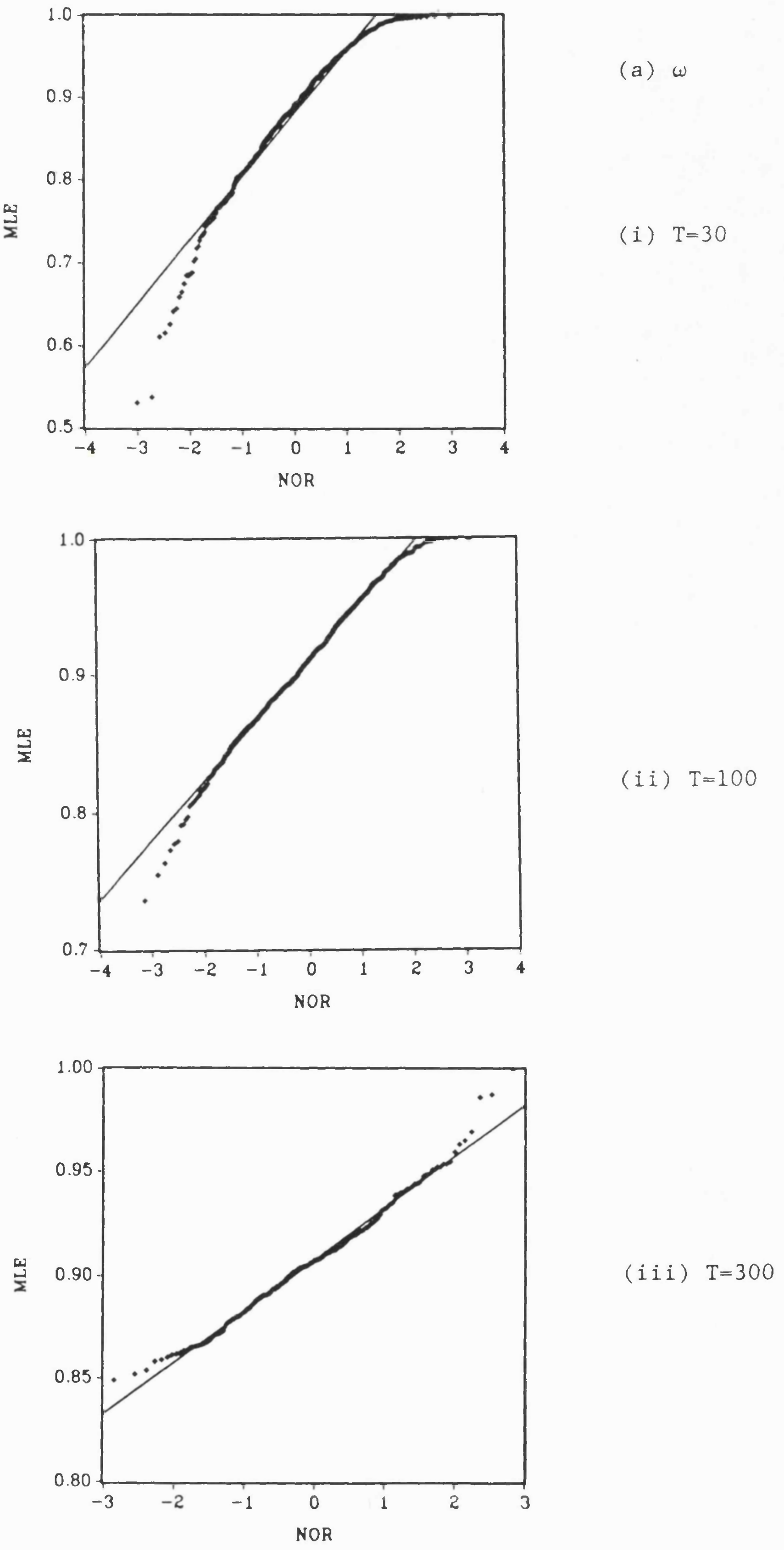
Descriptive Statistics

T		mean	bias	std.dev	skew	kurto	p(1)
30	ω	0.895	0.055	0.079	-1.079	4.263	0.605
	δ	0.043	-0.003	0.066	-0.060	3.605	—
50	ω	0.927	0.022	0.053	-0.956	4.194	0.532
	δ	0.043	-0.003	0.051	-0.024	3.092	—
80	ω	0.939	0.010	0.038	-0.768	3.367	0.396
	δ	0.039	0.001	0.039	-0.163	3.538	—
100	ω	0.945	0.004	0.034	-0.742	4.058	0.339
	δ	0.040	-0.000	0.035	-0.017	3.298	—
300	ω	0.953	-0.003	0.019	-0.135	2.894	0.09
	δ	0.004	0.000	0.021	0.106	3.245	—

Normality Tests

T		AD	BS	AP
30	ω	6.366	102.898	67.424
	δ	1.269	15.889	10.506
50	ω	5.602	99.121	67.181
	δ	0.325	0.450	0.580
80	ω	6.489	62.577	51.567
	δ	1.120	16.464	12.696
100	ω	4.235	91.493	65.342
	δ	0.440	3.750	3.312
300	ω	0.380	0.961	0.901
	δ	0.346	1.309	1.594

Figure 8.2.9 Normal probability plots for empirical ML estimates of a Poisson-Gamma model with ω=0.90 and δ=0.04.



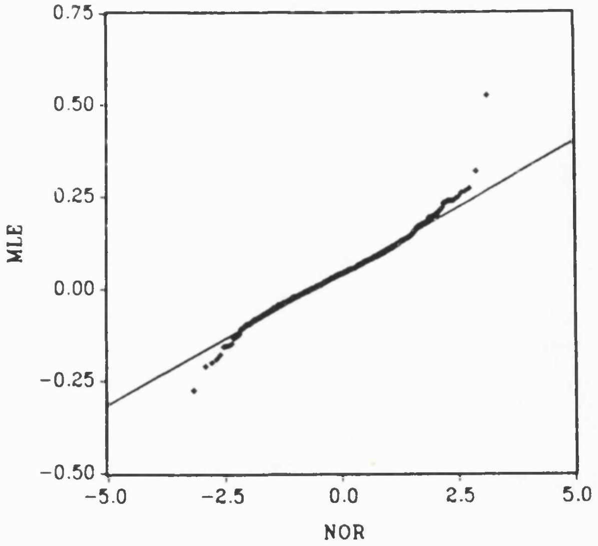
(a) ω

(i) T=30

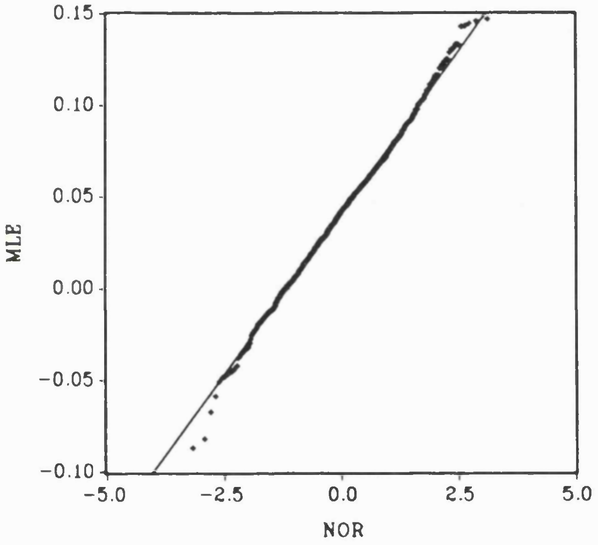
(ii) T=100

(iii) T=300

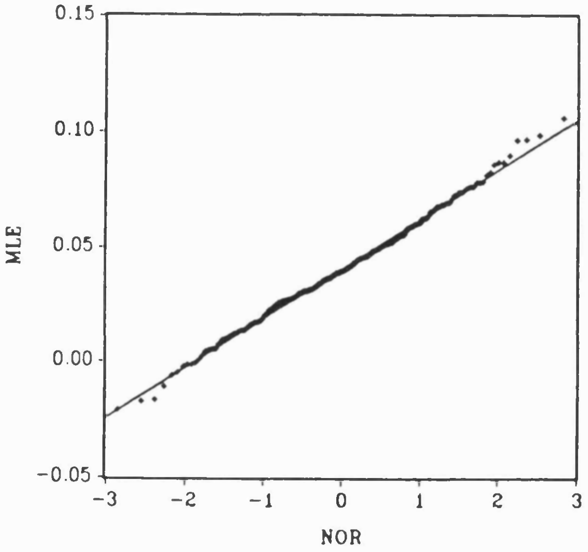
(b) δ



(i) T=30

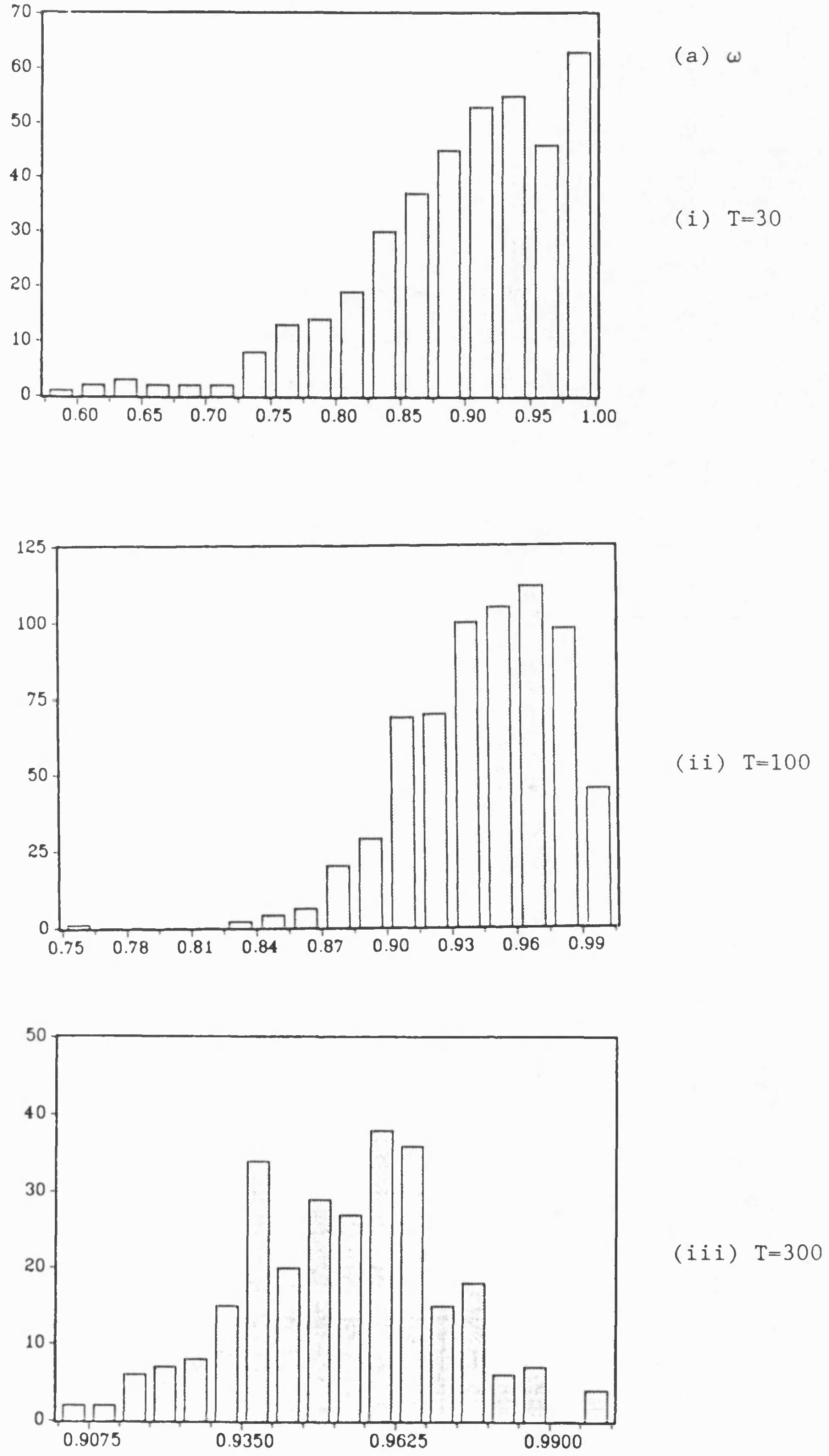


(ii) T=100



(iii) T=300

Figure 8.2.12 Sampling distributions for empirical ML estimates of a Poisson-Gamma model with ω=0.95 and δ=0.04.

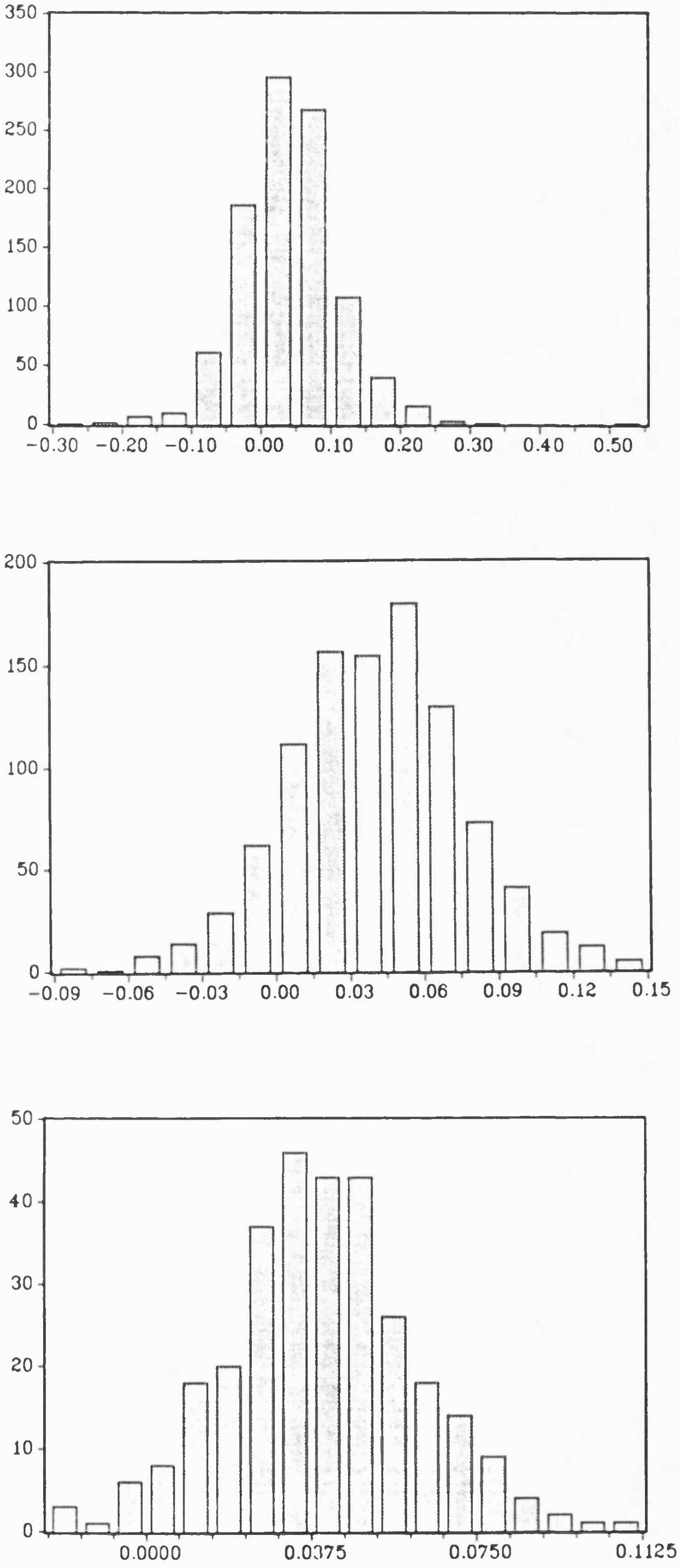


(a) ω

(i) T=30

(ii) T=100

(iii) T=300



(b) δ

(i) T=30

(ii) T=100

(iii) T=300

The main conclusions which may be derived from the above tables and plots are summarized below.

1. As expected, the occurence of boundary ML estimates decreases with increasing sample size, the rate of decrease being faster the further away the fixed value of ω is from its upper bound. While for ω=0.90 this effect vanishes altogether when T=300, for ω=0.98 samples as large as 700 observations will still produce a neglegible fraction of extreme cases. As a curiosity we compare our estimated probability of the boundary value, p(1), with the results obtained in a similar study conducted by S&H for a Gaussian local level model. In table 8.2.10 we reproduce the theoretical probabilities of a local maximum occuring at q=0, for this specification when the actual value of q is set at q=0.01 and a diffuse prior is used. This is to be compared with the empirical probabilities estimated for the equivalent Poisson-Gamma model, which is the standard specification with $\omega \triangleq 0.9$.

Table 8.2.10 Tabulation of the probability of a local maximum occurring at q=0 for the Gaussian model versus the estimated probability of a local maximum occurring at ω=1 for the Poisson-Gamma model.

	Hyperparameter true value	
T	Poisson-Gamma for ω=0.9	Gaussian for q=0.01
30	0.448	0.487
50	0.320	0.349

2. As expected, as the sample size grows both the bias and standard deviation decrease, approaching zero.

3. For $\omega \in \{0.85, 0.90\}$ and $T < 100$, and for ω-$\{0.95, 0.98\}$ and $T < 100$, the empirical distribution of the ML estimators are strongly non-normal, as may be seen from the high values returned by the normality tests and the shapes displayed in the graphs. In fact it is reasonable to suggest that these distributions may be approximated by a truncated normal. This means that the right-hand end of confidence interval constructed under the normal approximation will be shorter than it should be.

4. For $\omega \in \{0.85,0.90\}$ and $T > 100$ and for $\omega \in \{0.85, 0.98\}$ and $T > 500$ the normal approximation seems to be quite satisfactory as an inspection of the respective probability plots and histograms shows. Further evidence of this behaviour is supplied by the low values assumed by the three different normality tests used.

5. The frequency of boundary estimates seems to be unaffected by the presence of the Gaussian white noise regressor. This corroborates similar findings reported by Shephard (1990a, p.8) when studying the behaviour of these probabilities for the Gaussian local level model with fixed regressor parameter. As a corollary, on average, all the findings concerning the ML estimator properties of ω in a discount only model may also be carried out for a model with a Gaussian regressor. Note however, that preliminary studies conducted by the author have demonstrated that by considering a deterministic time trend, instead of a Gaussian regressor, the probabilities of obtaining

boundary values are considerably increased. These findings are also consistent with analytical results derived by Shephard (1990a) for the Gaussian case. As in the Gaussian local level model the ocurrence of boundary estimates has a decisive influence in forecasting since when ω is estimated as one there is no discounting. The problem of estimating parameters on the boundary of the parameter space may be reduced by considering alternative techniques to full ML, such as the profile and marginal likelihood functions. The later has been advocated, e.g., by Shephard (1990a). A introductory discussion on these techniques may be found in the second edition of McCullagh and Nelder's book on Generalized Linear Models (1989).

6. The normal approximation for the estimator of the regressor parameter seems to work quite satisfactorily even for sample sizes around 100 observations. The quality of the approximation is improved when the discount parameter is set closer to its upper bound value. For example, when ω=0.95, the normal approximation is already quite good for samples as small as 50, as can be seen from the graphics and the result of the normality tests (see table 8.2.9). This is to be expected since in this situation much of the data variability is explained by the regressor's presence. More formally, our Poisson-Gamma model approaches a NBD regression model, whose properties have been considered by a number of authors. E.g., Lawless (1987b), proves asymptotic normality of ML estimators in this specification. Corroborating our findings, he also finds that for samples as small as 25 or 30 the normal approximation is already quite satisfactory. It is also important to note that this same study has provided evidence supporting the use of the χ^2 approximation for the

likelihood-ratio statistics. In fact this statistic has been recommended by the author for inferences about the regression coefficients, except possibly in very small samples.

8.3 POST-SAMPLE PREDICTIVE TEST: SIZE AND POWER

Here we investigate the size and power of the post-sample predictive test for two specifications of the Poisson-Gamma model, namely, the discount only model and the model with regressor. The derivation of the test for both situations is in Appendix A1. Under the null hypothesis that there is no structural change in the post sample period t=T-1 to T+ℓ, the test statistic $\xi(\ell)$, is claimed to be asymptotically χ^2 with ℓ df. We have considered post-sample periods equal to ℓ=5 and ℓ=10 in our study, whose outline is produced below.

i. define and implement a mechanism able to produce a structural change in the post sample period. This will be the alternative hypothesis.

ii. generate data with structural change affecting only the observations between t=T+1 and t=T+ℓ.

iii. estimate the hyperparameters involved in the model using observations from t=1 to t=T.

iv. run the filter from t=1 to t=T+ℓ, and compute the test statistic $\xi(\ell)$. This is then compared with the critical values of a $\chi^2(\ell)$ at 1%, 5% and 10% levels (see table 8.3.2), producing the empirical rejection frequencies. A natural way of introducing structural change in the post sample period is by considering a step change affecting the DFM from t=T+1 to t=T+ℓ.

By defining a dummy variable which takes the value 1 in this period and 0 otherwise, this effect is easily introduced in our framework. It is obvious that, in view of (3.2.27), the structural change parameter δ_s will be equal to the logarithm of the ratio of the level of the post sample period to the level of the sampling period. By varying the value of δ_s one will obviously increase the power of our post sample predictive test. With this in mind the following values for δ_s have been set in our experiments.

Table 8.3.1 Correspondence between δ_s and the ratio of levels.

δ_s	ratio of levels
0.0	1.0
0.7	2.0
1.01	3.0
1.39	4.0
1.61	5.0

Obviously when δ_s=0 the empirical rejection frequencies will coincide with the empirical size of the test.

The probability of rejecting the null hypothesis of 'no structural change', say 'p', is estimated by evaluating the ratio between the number of favourable cases in which the value of the test statistics $\zeta(\ell)$ exceeds the critical values of the $\chi^2(\ell)$ to the number of replications Nrep in the Monte Carlo experiment. When Nrep is sufficiently large, confidence intervals for these probabilities are given by

$(\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/Nrep},\ \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/Nrep})$.

Since our simulations are based either in Nrep=1000 or Nrep=300 replications, the following table will be of help in comparing the values of the nominal and empirical sizes (type I error) of our predictive test.

Table 8.3.2 Type I errors approximate confidence intervals.

	Test size		
Nrep	1%	5%	10%
300	(0.0, 2.1)	(2.5, 7.5)	(6.6, 13.4)
1000	(0.4, 1.6)	(3.6, 6.4)	(8.1, 11.8)

As remarked by Kiviet (1987, p.58) for moderate number of replications care must be exercised in using the above results, given that relatively large confidence intervals will be produced. This is particularly relevant when $\alpha=1\%$ and Nrep=300.

In what follows we display the tables summarizing the findings of our simulations for the post-sample predictive test.

Table 8.3.3 Rejection percentages of the Post-sample Predictive Test of the Poisson-Gamma model with ω=0.85.

		5 df			10 df		
T		1%	5%	10%	1%	5%	10%
30	δ_s= 0.0	1.4	6.3	11.3	2.1	9.0	16.2
	0.7	63.5	74.2	80.3	64.0	73.0	79.3
	1.01	87.4	91.7	93.8	86.2	92.1	94.0
	1.39	95.9	98.3	98.8	95.8	98.2	98.9
	1.61	98.5	99.3	99.6	98.4	99.2	99.5
50	δ_s= 0.0	2.0	6.3	12.5	2.9	8.4	15.4
	0.7	66.2	76.3	80.3	66.4	75.3	80.2
	1.01	87.2	91.7	93.5	85.6	90.5	92.5
	1.39	95.5	96.7	97.6	95.5	96.7	97.2
	1.61	97.2	98.1	98.3	97.1	97.7	98.3
80	δ_s= 0.0	1.5	5.3	10.8	2.9	8.1	14.8
	0.7	61.1	71.7	77.1	59.8	69.4	76.1
	1.01	82.7	88.0	91.3	81.7	87.5	90.8
	1.39	93.1	95.2	95.8	92.7	95.2	96.6
	1.61	95.5	96.6	97.2	95.5	96.7	97.3
100	δ_s= 0.0	1.3	6.3	11.1	2.0	7.9	13.7
	0.7	62.7	72.5	77.8	63.0	72.5	77.8
	1.01	83.7	87.7	89.7	81.8	87.5	90.1
	1.39	91.4	93.2	94.7	91.2	93.4	94.9
	1.61	93.9	95.4	95.9	93.5	95.4	96.4
300	δ_s= 0.0	1.3	7.3	13.0	2.7	9.0	16.3
	0.7	83.3	89.0	92.0	82.0	87.3	90.0
	1.01	93.3	94.3	95.7	92.7	94.3	95.3
	1.39	96.7	97.0	97.0	96.7	96.7	97.0
	1.61	97.0	97.7	98.0	96.7	97.0	97.7

Table 8.3.4 Rejection percentages of the Post-sample Predictive Test of the Poisson-Gamma model with ω=0.90.

		5 df			10 df		
T		1%	5%	10%	1%	5%	10%
30	δ_s= 0.0	1.3	5.9	11.6	2.7	8.2	14.3
	0.7	76.8	85.5	89.9	78.5	86.5	89.8
	1.01	95.3	97.4	97.9	94.7	97.3	98.0
	1.39	99.2	99.7	99.7	99.2	99.7	99.8
	1.61	99.9	99.9	99.9	99.9	100.0	100.0
50	δ_s= 0.0	1.1	6.3	12.3	2.0	9.1	16.7
	0.7	75.5	85.5	90.2	79.9	86.7	90.2
	1.01	95.1	97.7	98.9	94.7	97.6	98.3
	1.39	99.4	99.8	99.9	99.5	99.9	99.9
	1.61	99.8	99.9	99.9	99.9	99.9	99.9
80	δ_s= 0.0	2.0	6.8	11.3	2.6	8.6	15.4
	0.7	78.6	87.4	90.7	79.7	88.0	91.3
	1.01	95.9	97.9	98.5	96.2	97.7	98.4
	1.39	98.9	99.5	99.6	98.9	99.5	99.6
	1.61	99.7	99.8	99.9	99.4	99.7	99.7
100	δ_s= 0.0	1.5	6.2	11.5	2.2	8.5	15.0
	0.7	75.0	84.3	88.4	75.7	85.6	89.3
	1.01	92.6	96.0	96.8	93.7	95.6	96.6
	1.39	97.9	98.9	99.2	98.0	98.7	98.9
	1.61	98.9	99.3	99.5	98.7	99.4	99.5
300	δ_s= 0.0	1.0	3.0	7.3	1.3	5.7	11.7
	0.7	67.3	76.7	83.7	67.7	78.0	82.3
	1.01	88.3	91.0	92.3	86.0	89.7	90.3
	1.39	94.0	95.0	95.0	93.3	94.3	95.7
	1.61	95.0	95.7	96.3	95.0	95.7	96.3

Table 8.3.5 Rejection percentages of the Post-sample Predictive Test of the Poisson-Gamma model with ω=0.95.

		5 df			10 df		
T		1%	5%	10%	1%	5%	10%
30	δ_s= 0.0	0.6	3.9	9.3	1.8	6.0	11.2
	0.7	84.7	92.1	95.0	87.5	92.6	95.0
	1.01	99.1	99.7	99.8	98.5	99.3	99.7
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
50	δ_s= 0.0	1.0	4.6	9.5	0.9	5.9	11.9
	0.7	87.5	94.1	95.8	90.9	94.9	97.1
	1.01	98.7	99.5	99.8	99.3	99.6	99.8
	1.39	99.9	99.9	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
80	δ_s= 0.0	0.8	5.9	12.2	1.9	7.3	12.3
	0.7	90.1	96.1	97.9	93.7	97.8	98.0
	1.01	99.2	99.6	99.7	99.7	99.7	99.7
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
100	δ_s= 0.0	1.1	5.8	11.1	1.5	7.5	12.9
	0.7	87.6	94.4	96.5	93.9	97.7	98.6
	1.01	99.5	99.8	99.9	99.7	100.0	100.0
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
500	δ_s= 0.0	2.3	5.0	10.3	2.7	6.7	10.3
	0.7	84.3	90.7	92.7	88.0	94.0	95.7
	1.01	97.0	98.3	98.7	84.3	90.7	92.7
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
700	δ_s= 0.0	1.6	4.3	9.0	2.0	6.3	10.7
	0.7	83.3	87.3	90.0	88.0	93.3	94.7
	1.01	94.0	96.7	98.0	97.3	98.0	98.7
	1.39	99.0	99.0	99.0	99.0	99.3	99.3
	1.61	99.0	99.3	99.7	99.3	99.3	99.7

Table 8.3.6 Rejection percentages of the Post-sample Predictive Test of the Poisson-Gamma model with ω=0.98.

		5 df			10 df		
T		1%	5%	10%	1%	5%	10%
30	δ_S= 0.0	2.0	5.5	11.1	1.4	6.7	12.0
	0.7	87.7	94.4	97.0	90.6	95.1	96.5
	1.01	98.8	99.4	99.7	98.5	99.2	99.5
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
50	δ_S= 0.0	0.9	4.3	10.1	0.8	5.5	11.2
	0.7	90.4	96.1	97.4	95.4	97.1	98.2
	1.01	99.4	99.9	99.9	99.5	99.8	99.8
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
80	δ_S= 0.0	1.4	5.8	11.0	1.7	6.2	11.5
	0.7	94.2	98.3	99.0	98.0	99.3	99.6
	1.01	99.9	99.9	100.0	99.9	99.9	100.0
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
100	δ_S= 0.0	1.3	5.7	11.0	1.6	7.4	12.6
	0.7	94.3	97.8	98.8	98.5	99.4	99.5
	1.01	100.0	100.0	100.0	100.0	100.0	100.0
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
500	δ_S= 0.0	1.3	4.0	8.3	1.7	6.3	9.7
	0.7	94.0	96.7	97.7	97.7	99.7	99.7
	1.01	100.0	100.0	100.0	100.0	100.0	100.0
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
700	δ_S= 0.0	0.7	3.7	9.0	1.0	6.7	11.0
	0.7	91.0	95.0	95.7	97.3	98.3	99.0
	1.01	99.7	100.0	100.0	100.0	100.0	100.0
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0

Table 8.3.7 Rejection percentages of the Post-sample Predictive Test of the Poisson-Gamma model with ω=0.90 and δ=0.04.

			5 df			10 df	
T		1%	5%	10%	1%	5%	10%
30	δ_S= 0.0	1.5	6.2	15.0	2.2	8.5	15.0
	0.7	75.0	85.6	89.3	75.7	85.6	89.3
	1.01	92.6	96.0	96.8	93.7	95.6	96.6
	1.39	97.9	98.9	99.2	98.0	98.7	98.9
	1.61	98.9	99.3	99.5	98.7	99.4	99.5
50	δ_S= 0.0	1.2	7.7	13.5	2.3	9.5	17.8
	0.7	78.5	88.6	91.3	80.6	89.0	92.6
	1.01	95.9	97.2	97.6	95.8	97.4	98.0
	1.39	98.8	99.3	99.4	98.9	99.2	99.4
	1.61	99.4	99.6	99.7	99.3	99.5	99.6
80	δ_S= 0.0	1.8	6.7	12.0	2.4	8.1	15.2
	0.7	76.1	85.8	89.9	79.3	87.3	90.7
	1.01	95.1	97.3	98.0	95.8	97.7	98.9
	1.39	99.2	99.5	99.7	99.4	99.5	99.8
	1.61	99.6	99.8	99.8	99.7	99.9	100.0
100	δ_S= 0.0	1.4	6.8	11.7	2.0	8.4	14.8
	0.7	75.2	85.6	90.3	78.5	87.2	91.9
	1.01	94.2	96.1	97.3	95.0	97.3	97.7
	1.39	98.4	99.0	99.4	98.3	99.2	99.3
	1.61	99.2	99.7	99.8	99.1	99.5	99.8
300	δ_S= 0.0	0.3	3.7	7.3	1.3	5.7	10.7
	0.7	72.3	78.7	81.3	73.3	82.0	84.7
	1.01	86.7	90.7	92.7	88.0	91.3	93.0
	1.39	95.0	95.3	96.3	95.0	96.3	97.3
	1.61	96.7	97.0	97.7	96.3	97.7	97.7

Table 8.3.8 Rejection percentages of the Post-sample Predictive Test of the Poisson-Gamma model with ω=0.95 and δ=0.04.

		5 df			10 df		
T		1%	5%	10%	1%	5%	10%
30	δ_s= 0.0	1.1	5.6	10.8	1.8	7.4	12.6
	0.7	84.7	92.3	95.7	88.0	93.1	95.6
	1.01	98.9	99.5	100.0	99.1	99.6	99.7
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
50	δ_s= 0.0	1.0	4.3	9.0	1.2	7.7	12.7
	0.7	92.1	96.2	97.8	92.1	96.2	97.8
	1.01	99.2	99.6	99.8	99.5	99.7	99.8
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
80	δ_s= 0.0	1.2	5.0	11.4	1.8	7.0	13.6
	0.7	89.3	94.9	96.1	94.0	96.9	98.3
	1.01	98.9	99.5	99.8	99.2	99.7	99.8
	1.39	99.9	99.9	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
100	δ_s= 0.0	1.6	6.8	11.9	1.8	7.4	14.4
	0.7	89.6	94.3	98.1	93.4	97.1	98.1
	1.01	99.4	99.8	99.8	99.8	99.9	99.9
	1.39	99.8	99.8	100.0	99.9	99.9	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0
300	δ_s= 0.0	0.3	3.0	8.0	1.0	7.3	11.3
	0.7	90.3	96.0	96.7	93.0	97.0	97.3
	1.01	98.3	99.3	100.0	98.3	100.0	100.0
	1.39	100.0	100.0	100.0	100.0	100.0	100.0
	1.61	100.0	100.0	100.0	100.0	100.0	100.0

We now summarize the main conclusions about the size and power of our post sample predictive test which may be drawn from the above tables.

Generally speaking, rejection percentages, or the power of the test, is usually quite attractive, irrespective of such factors like the discount value, the post sample-to-within sample level ratio, the post sample period, the sample size, etc. A more detailed analysis meanwhile reveals a number of features which we now report.

1a. As expected, when $\omega \rightarrow 1$ the power of the test increases, since the generated sample data will tend to follow a more smooth path, therefore producing higher sensibility to the test in the post sample period when the 'jump effect' is introduced. For example, at a 5% level, when $T=50$ and $\delta s=0.7$, the rejection frequencies span from 76.3% 96.1%.

2a. Obviously the power of the test increases as the ratio of the level post sample to the level within sample, δ_S increases.

3a. In general the rejection frequencies experience a slight decrease for larger sample sizes, except possibly when $\omega=0.98$. The same rather curious phenomenon has also been reported in a study carried out by Kiviet (1987, p.64) in the context of dynamic regression models.

4a. The power of the test seems to be insensitive to the length of the post sample period ℓ.

5a. The same conclusions drawn for the standard model also holds true for the regression specifications, i.e., the presence of a regressor variable seems not to affect the relationship among the power of the test and the aforementioned factors.

We now examine the results associated with the significance levels of our asymptotic test. Observe that, as quoted by Kiviet (1987, p.58), 'ideally, the observed significance level should be close to the nominal level, regardless of the values of the parameters'. Using the confidence intervals for the sizes displayed in table (8.3.2) the overall conclusion is that the estimated level is quite close to its nominal values to within sampling error, but a number of features have to be considered, and these are reported below.

1b. For moderate and small sample sizes ($T \leqslant 100$), the degree of approximation is particularly good for the shortest post sample period, i.e., for ℓ-5. This is to be understood as meaning that the nominal size falls inside the correspondent confidence intervals as displayed in table (8.3.2). In considering a larger post sample period, when ℓ-10, the actual size, generally speaking, overestimates the nominal values. This will lead to occasional rejections of the null hypothesis of 'non structural change' when actually the DFM does not display such a feature. This behaviour is particularly sensitive to the values of ω, being more pronounced for those values further away from the upper bound.

2b. For large sample size, i.e., when T-300 for $\omega \epsilon$ {0.85, 0.90} and T-500, 700 for ω-{0.95,0.98}, the actual sizes exhibit even better behaviour, irrespective of the post sample period considered. The reported frequencies also seem to be invariant to the parameter values, this being true for both the standard and regression specifications.

In summary, our Monte Carlo study has provided evidence that the χ^2 approximation for our post sample predictive test is quite acceptable for large sample sizes, irrespctive of the hyperparameter values. Meanwhile care must be exercised when using it for small and moderate series, and when the post sample period is larger than five since the actual size of $\xi(\ell)$ may over estimate the nominal size of the $\chi^2(\ell)$.

REFERENCES

Ansley, C.F. and R. Kohn (1985). Estimation, Filtering and Smoothing in State Space Models with Incompletely Specified Initial Conditions, Annals of Statistics, 13:1286-1316.

Abramowitz, M. and L.A. Stegun (eds.) (1965). Handbook of Mathematical Functions, vol. 55 in Applied Mathematical Series, Washington D.C. : National Bureau of Standards.

Aitchison, J. and I.R. Dunmore (1975). Statistical Prediction Analysis, Cambridge University Press.

Akaike, H. (1980). Seasonal Adjustment by a Bayesian Modeling, Journal of Time Series Analysis, 1: 1-13.
(1984). On the Use of Bayesian Models in Time Series Analysis, in Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics, Springer Verlag.

Ameen, J.R.M. and P.J. Harrison (1985). Normal Discount Bayesian Models, in Bayesian Statistics vol. 2, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Amsterdam: North Holland.

Anderson, B.D.O., and J.B. Moore (1979). Optimal Filtering, Englewood Cliffs: Prentice Hall.

Bartels, R. (1982). The Rank Version of Von Neumann's Ratio Test for Randomness, Journal of the American Statistical Association, 77: 40-46.

Bather, J.A. (1965). Invariant Conditional Distributions, Annals of Mathematical Statistics, 36: 829-846.

Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis, (2nd edition). Springer Verlag.

Box, G.E.P., and G.M. Jenkins (1976). Time Series Analysis: Forecasting and Control, revised edn. San Francisco: Holden-Day.

Brown, R.G. (1959). Statistical Forecsting in Inventory Control, New York: McGraw-Hill.

Cameron, C.C. and P.K. Trivedi (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests, Journal of Applied Econometrics, 1: 29-53.

Cheng, R.C.H., and G.M. Feast (1980). Gamma Variate Generators with Increased Shape Parameter Range, Communications of the ACM, Vol 23, n^{0} 7, 389-394.

Chow, G.C. (1960). Tests of Equality between Sets of Coefficients on Two Linear Regressions, Econometrica, 28: 591-605.

D'Agostino, R.B. and M.A. Stephens (1986) (eds). Goodness-of-Fit Techniques, New York: Marcel Dekker, Inc.

Farnum, N.R. and L.W. Stanton (1989). Quantitative Forecasting Methods, Boston: PWS-KENT Publishing Co.

Fernandez, F.J. and A.C. Harvey (1989). Seemingly Unrelated Time Series Equations and a Test for Homogeneity, Journal of Business and Economic Statistics, to appear.

Figliuoli, L. (1988). A State Space Approach to Non Linear and Non Gaussian Time Series Models, unpublished Ph.D. thesis, University of London.

Gourieroux, C., A. Monfort and A. Trognon. Pseudo Maximum Likelihood Methods: Applications to Poisson Models, Econometrica, 52: 701-720.

Granger, C.W.J. and A.P Andersen (1978). An Introduction to Bilinear Time Series Models, Gottingen: Vandenhoek and Ruprecht.

Harrison, P.J. (1988). Bayesian Forecasting in Operational Research, In Developments in Operational Research 1988, N.B. Cook and A.M. Jones (Eds), Oxford: Pergamon Press.

Harrison, P.J. and C.F. Stevens (1976). Bayesian Forecasting (with discussion), Journal of the Royal Statistical Society, B , 38: 205-247.

Hartigan, J.A. (1983). Bayes Theory, Springer Verlag.

Harvey, A.C. (1984). A Unified View of Statistical Forecasting Procedures, Journal of Forecasting, 3: 245-275.

(1989). Forecasting, Structural Time Series Models and the Kalman Filtering, Cambridge University Press.

Harvey, A.C. and P. Collier (1977). Testing for Functional Misspecification. In Regression Analysis, Journal of Econometrics, 6: 103-119.

Harvey, A.C. and J. Durbin (1986), The Effects of Seat Belt Legislation on British Road Casualties: A Case Study in Structural Time Series Modelling. Journal of the Royal Statistical Society, A 149: 187-227.

Harvey, A.C. and C. Fernandes (1989a). Time Series Models for Count or Qualitative Observations, Journal of Business and Economic Statistics, 7: 407-422.

(1989b). Time Series Models for Insurance Claims, The Journal of the Institute of Actuaries, 116: 513-528.

Hausman, J.A., Hall B.H. and Z. Griliches (1984). Econometric Models for Count Data with an Application to the Patents- R&D Relationship, Econometrica, 52: 909-938.

Hinich, K.H. and D.M. Patterson (1985). Evidence of Nonlinearity in Daily Stock Returns, Journal of Business & Economic Statistics, 3: 69-77.

IMLSL Stat/Library (1987). IMSL Stationary: User Manual, Version 1.0, Houston: IMSL, Inc.

Jarque, C.M. and A.K. Bera (1987). A Test for Normality of Observations and Regression Residuals, International Statistical Review, 55:163-172

Jazwinski, A.H. (1970). Stochastic Processes and Filtering Theory. New York: Academic Press.

Johnson, N.L. and S. Kotz (1969). Distributions in Statistics: Discrete Distributions. New York: Houghton Mifflin.
(1972). Distributions in Statistics: Continuous Multivariate Distributions. New York: Houghton Mifflin.

Kalman, R.E. (1960). A New Approach to Linear Filtering and Prediction problems, Journal of Basic Engineering, Transactions ASME, Series D , 82: 35-45.

Key, P. and E.J. Godolphin (1981). On the Bayesian Steady Forecasting Model, Journal of the Royal Statistical Society, Series B, 43: 92-96.

Kitagawa, G. (1981). A Nonstationary Time Series Model and its Fitting by a Recursive Filter, Journal of Time Series Analysis, 2:103-116.
(1987). Non-Gaussian State-Space Modeling of Nonstationary Time Series (with discussion), Journal of the American Statistical Association, 82:1032-1063.

Kiviet, J.F. (1987). Testing Linear Econometric Models, Amsterdam: Faculty of Actuarial Science & Economics, University of Amsterdam.

Lawless, J.F. (1987a). Regression Methods for Poisson Process Data, Journal of the American Statistical Association, 82: 808-815.
(1987b). Negative Binomial and Mixed Poisson Regression, The Canadian Journal of Statistics, 15: 209-225.

Lawrance, A.J., and P.A.W. Lewis (1985). Modelling and Residual Analysis of NonLinear Autoregressive Time Series in Exponential Variables, Journal of the Royal Statistical Society, Series B, 47: 165-202.

Lehmann, E.L. (1983). Theory of Point Estimation, New York: John Wiley and Sons, Inc.

McCleary, R. and R.A. Hay, Jr (1980). Applied Time Series Analysis for the Social Sciences, Beverly Hills, Calif: Sage Publications.

McCullagh, P. and J.A. Nelder (1983). Generalised Linear Models, London: Chapman and Hall.
(1989) 2nd edition.

McKenzie, Ed. (1982). Product Autoregression: A Time-Series Characterization of the Gamma Distribution, Journal of Applied Probability, 19: 463-468.
(1985). An Autoregressive Process for Beta Random Variables, Management Science, 31: 988-997·
(1986). Autoregressive Moving–Average Processes with Negative Binomial and Geometric Marginal Distributions, Advances in Applied Probability, 18: 679-705.

Muth, J.F. (1960). Optimal Properties of Exponentially Weighted Forecasts, Journal of the American Statistical Association, 55: 299-305.

NAG Library (1984). Fortran Library Manual, Mark 11, London: Numerical Algorithms Group.

Nelson, H.L. and C.W.J. Granger (1979). Experience with Using the Box-Cox Transformation when Forecasting Economic Time Series, Journal of Econometrics, 10: 57-69.

Ord, J.K. (1972), Families of Frequency Distributions. London: Griffin.

Ord, J.K., C. Fernandes and A.C. Harvey (1989). Time Series Models for Multivariate Series of Count Data, unpublished paper, LSE.

Ozaki, T. (1982). The Statistical Analysis of Perturbed Limit Cycle Processes using Nonlinear Time Series Models, Journal of Time Series Models, 3: 29-40.

Patil, G.P., M.T. Boswell, M.V. Ratnaparkhi and J.J.J. Roux (1985). Dictionary and Classified Bibliography of Statistical Distributions in Scientific Work, Vol.3: Multivariate Models. Fairland, Md: ICPH (International Cooperative Publishing House).

Pole, A. and M. West (1988). Efficient Numerical Integration in Dynamic Models, Research Report 136, Dept. of Statistics, University of Warwick.

Priestley, M.B. (1980). State Dependent Models: A General Approach to Nonlinear Time Series Analysis, Journal of Time Series Analysis, 1:47-71.

(1981). Spectral Analysis and Time Series, New York: Academic Press.

Press, W.H. *et al* (1987). Numerical Recipes; The Art of Scientific Computing, Cambridge University Press.

Raiffa, H. and R. Schlaifer (1961). Applied Statistical Decision Theory, Boston: Harvard University.

Rao, T. S. and M.M. Gabr (1980). A Test for Linearity of Stationary Time Series, Journal of Time Series Analysis, 1: 145-157.

Reed, D. (1978). Whistlestop: A Community Alternative for Crime Prevention, unpublished Ph.D. thesis, Dept. of Sociology, Northwestern University.

Shephard, N.G. (1990a). Maximum Likelihood Estimation of Regression Models with Stochastic Trend Components, unpublished paper, Dept. of Statistics, LSE.
(1990b). A Local Level Model, unpublished paper, Dept. of Statistics, LSE.

Shephard, N.G. and A.C. Harvey (1989). Tracking the Level of Party Support During General Election Campaigns, unpublished paper, Dept. of Statistics, LSE.
(1990). On the Probability of Estimating a Deterministic Component in the Local Level Model, to appear Journal of Time Series Analysis.

Shiryayev, A.N. (1984). Probability, Springer-Verlag.

Sinclair, C.D. and B.D. Spurr (1988). Approximations to the Distribution of the Anderson-Darling Test Statistics, Journal of the Americal Statistical Association, 83: 1190-1191.

Smith, J.Q. (1979). A Generalization of the Bayesian Steady Forecasting Model, Journal of the Royal Statistical Society, Series B, 41: 375-387.

(1988a). Non Linear State Space Models with Partially Specified Distributions on States, Research Report 150, Dept of Statistics, University of Warwick.

(1988b). A Comparison of the Characteristics of some Bayesian Forecasting Models, Research Report 140, Dept. of Statistics, University of Warwick.

Smith, R.L. and J.E. Miller (1986). A Non-Gaussian State Space Model and Application to Prediction of Records, Journal of the Royal Statistical Society, Series B, 48: 79-88.

Sorenson, H.W. (1988). Recursive Estimation for Nonlinear Dynamic Systems, in Bayesian Analysis of Time Series and Dynamic Models, edn. J.C. Spall, pp:127-165, New York: Marcel Dekker, Inc.

Stein, G.Z. and J.M. Juritz (1987). Bivariate Compound Poisson Distributions, Communications in Statistics: Theory and Methods, 16(12): 3591-3607.

Stuart, A. and J.K. Ord (1987). Kendall's Advanced Theory of Statistics, Volume I, London: Griffin (5th edition).

Sweeting, T.J. (1980). Uniform Asymptotic Normality of the Maximum Likelihood Estimator, The Annals of Statistics, 8: 1375-1381.

Tan, H.K. (1990). Robust Estimation for Structural Time Series Models, unpublished Ph.D. thesis, University of London.

Tong, H. and K.S. Lim (1981). Threshold Autoregression, Limit Cycles and Cycclical Data (with discussion). Journal of the Royal Statistical Society, Series B, 42:245-292.

Tong, H. and R. Moeanaddin (1988). On Multi-step Non-Linear Least Square Prediction, The Statistician, 37: 101-110.

Tsay, R.S. (1986). Nonlinearity Tests for Time Series, Biometrika, 73: 461-6.

West, M., P.J. Harrison and H.S. Migon (1985). Dynamic Generalized Linear Models and Bayesian Forecasting (with discussion), Journal of the American Statistical Association, 80: 73-97.

Wilk, M.B. and R. Gnanadesikan (1968). Probabilty Plotting Methods for the Analysis of Data, Biometrika, 55: 1-17.

Zeger, S.L. (1988). A Regression Model for Time Series of Counts, Biometrika, 75: 621-629.

Zehnwirt, B. (1988). A Generalization of the Kalman Filter for Models with State-Dependent Observation Variance, Journal of the American Statistical Association, 83: 164-167.