

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/139956>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

A BAYESIAN ENTROPY APPROACH TO FORECASTING

By

Reinaldo Castro Souza

A thesis submitted for the degree of Doctor of Philosophy

STATISTICS DEPARTMENT
UNIVERSITY OF WARWICK

November 1978

SUMMARY

This thesis describes a new approach to steady-state forecasting models based on Bayesian principles and Information Theory. Shannon's entropy function and Jaynes' principle of maximum entropy are the essential results borrowed from Information Theory and are extensively used in the model formulation. The Bayesian Entropy Forecasting (BEF) models obtained in this way extend beyond the constraints of normality and linearity required in all existing forecasting methods. In this sense, it reduces in the normal case to the well known Harrison and Stevens steady-state model. Examples of such models are presented, including the Poisson-gamma process, the Binomial-Beta process and the Truncated Normal process. For all of these, numerical applications using real and simulated data are shown, including further analyses of epidemic data of Cliff et al, (1975).

ACKNOWLEDGEMENTS

To Professor P.J. Harrison, who gave unsparingly of his time and energy to provide the general guidance and suggestions that made this thesis possible, I wish to express my sincere appreciation.

To the Department of Statistics, University of Warwick, and in particular to Dr. J.K. Ord for supplying the data for the example in chapters 6 and 7, and to Dr. T. Leonard and Mr. I. Liddell for many interesting and helpful suggestions.

To Virginia A. Aguilar for typing the difficult manuscript with great skill and patience.

To CAPES-Brazil and PUC/RJ-Brazil for their supporting grant.

FOR

Edson, Maria Alice and Maria Carmen

"Time changes, but certain things are timeless"

TABLE OF CONTENTS

1	INTRODUCTION	
1.1	Scope of the Thesis	8
1.2	Organization of the Thesis	10
1.3	Thesis Terminology and Notations	11
2	ENTROPY FUNCTION	
2.1	Historical Remarks	13
2.2	The Notion of Entropy	14
2.3	Definition of Entropy-Discrete Case	16
2.4	Basic Properties of Discrete Entropy	17
2.5	The Extension to the Continuous Case	22
2.6	Other Approaches to the Continuous Case	26
3	JAYNES' PRINCIPLE OF MAXIMUM ENTROPY	
3.1	Introduction	33
3.2	Jaynes' Principle	39
3.3	Properties of the Maximum Entropy Density	42
3.4	Applications	48
3.5	Examples of Maximum Entropy Distributions	51
4	BAYESIAN ENTROPY FORECASTING (BEF)- GENERAL MODEL FORMULATION	
4.1	Historical Development of Time Series	52
4.2	Bayesian Forecasting	56
4.3	Bayesian Forecasting Limitations and Proposed Extension	59
4.4	Normal Additive Model; Entropy Results	60
4.4.1	Posterior-Prior Transition	60

4.4.2	Uncertainty Function	61
4.5	Bayesian Entropy Forecasting System	64
4.5.1	Model Foundations	64
4.5.2	General Assumptions	66
4.5.3	System Evolution	67
4.5.4	Exponential Approximation	72
4.5.5	Model Formulation	75
4.6	BEF - Properties	79
4.6.1	Normal Additive Model	79
4.6.2	Non-Additive Normal Model	81
4.6.3	Parameter Prediction	83
4.6.4	System Evolution	84
4.6.5	Steady State Model-Definition	86
4.6.6	Goodness of Fit-Relative Entropy Criterion	88
4.6.7	Aggregate likelihood for Estimation of "c"	92
4.7	Sufficient Statistic Specification	93
5	STEADY STATE POISSON-GAMMA MODEL	
5.1	Introduction	97
5.2	Entropy of the Gamma Variate	98
5.3	BEF Poisson-Gamma System ; Model Description	100
5.4	Limiting Form of the BEF Poisson-Gamma Model	103
5.5	Applications and Discussions	105

6	POISSON-GAMMA MULTI STATE MODEL	
6.1	Introduction	109
6.2	The Model	110
6.2.1	Information a Priori	111
6.2.2	Updating System	113
6.3	Collapsing Procedure	115
6.4	Case Study	120
6.4.1	Preliminary Data Analysis	121
6.4.2	Results	123
6.4.3	SM and MM Comparison	125
7	STEADY STATE BINOMIAL-BETA MODEL	
7.1	Introduction	126
7.2	Beta Variate Characteristics	127
7.3	Entropy of the Beta Variate	130
7.4	BEF Binomial-Beta System ; Model Description	131
7.5	Limiting Form of the Binomial-Beta BEF	134
7.6	Applications	135
8	STEADY STATE TRUNCATED NORMAL MODEL	
8.1	Introduction	141
8.2	Truncated Normal Distribution	143
8.2.1	Definitions and Characterizations	143
8.2.2	Properties	147
8.3	Entropy of the Truncated Normal Variate	150

8.4	Bayesian Analysis for the Truncated Normal Distribution	155
8.4.1	Parameter Posterior Distribution	156
8.4.2	Predictive Distribution	159
8.5	BEF Truncated Normal System ; Model Description	161
8.6	Applications	165
Appendix A		169
Appendix B		173
Appendix C		179
Appendix D		182
Appendix E		190
Appendix F		196
Appendix G		212
References		220

CHAPTER 1 : INTRODUCTION

1.1) Scope of the Thesis:

The past eight years have witnessed an unprecedented growth in the field of forecasting. The first major advance, of course, was Box and Jenkins' very clear formulation of forecasting models in 1970. However, their solution of the least square prediction problem was still shackled to the fundamental ideas of Wiener and Kolmogorov. Undoubtedly this was one of the most important contributions to the subject.

At almost the same time, Harrison and Stevens developed an important approach to forecasting using important results of Kalman and Bucy, already extensively used in Control Theory problems, together with Bayesian statistical theory. This approach gave rise to the so called "Bayesian Forecasting Methods" which offered something quite different from the Wiener and Kolmogorov theory. It is well known that the three basic assumptions on which all the previous forecasting methods are based are:

- stationarity of the underlying process,
- mean square prediction error as a forecasting criterion,
- predictor as a linear function of past observations.

These were partially overcome by the advent of the Bayesian approach. For instance, the stationarity of the underlying process is not required and also, by its distributional predictive nature a criterion of optimality other than the mean square error is possible.

Despite the above improvements and its simple, elegant formulation, the Bayesian approach as it stands still has its limitations. For instance, the models are still linear, where the observation noise and parameter disturbance

are additively related to the observation and system equations respectively, and (from the linear least square property of the Kalman filter) it is efficient only for the Normal process.

These two restrictions constitute the prime motivation for this dissertation. Our principal aim in this thesis is to develop an extension of Harrison and Stevens' approach in which the constraints of linearity and normality are not required. With this extension we are not merely satisfying the four essential basic foundations of the Bayesian Approach, namely:

- (i) Parametric formulation.
- (ii) Probabilistic information on the parameters at any given time.
- (iii) Sequential model definition.
- (iv) Uncertainty as to the underlying model.

but furthermore, we include the following two properties:

- (v) Non-linear general formulation.
- (vi) Unrestricted to any sort of distribution.

However, the original target of an unconditional formulation applicable to any kind of model has not been entirely reached. In this thesis we discuss only steady state models: a particular but important subclass of all models. On the other hand, we feel that this work has gone an appreciable way towards the original goal and further extensions, which might include a broader class of models such as the linear growth, seems quite feasible following the same argument.

The extension was made possible by the use of Shannon's entropy,

a crucially important measure of uncertainty and Jaynes' principle of maximum entropy. By the incorporation of Shannon's entropy into a Bayesian framework, the steady state linear normal model can be redefined in terms of the entropy function and, using the fact that entropy is an unrestricted measure of uncertainty, the extension follows naturally.

1.2) Organization of the Thesis

The thesis could be classified into three main parts. Part I (Chapter 2 and 3) is devoted to the definition and characterizations of the entropy function, as well as its main properties. In chapter 3 we show the mathematical formulation of Jaynes' principle of maximum entropy to assign the least prejudiced probability distribution for a random variable and some of its most important properties.

In Part II (Chapter 4) the theoretical Bayesian Entropy Forecasting (BEF) model for a steady state system is defined and described, starting from the steady state linear normal model. It also includes a brief survey of time series modelling and a summary of some of the most important forecasting methods.

Part III (Chapters 5 to 8) deals with some applications of the model to different processes such as:

- Poisson-Gamma single state process (Chapter 5).
- Poisson-Gamma multistate process (Chapter 6).
- Binomial-Beta single state process (Chapter 7).
- Truncated normal process (Chapter 8).

For each of these we show the relevant numerical results concerning their application to simulated and real data. Of particular interest is the analysis of the measles epidemic data in chapters 5,6 and 7.

Finally, the thesis is complemented by 7 appendices (A to G) containing mainly tables and figures related to the numerical results of the applications in Part III.

1.3) Thesis Terminology and Notations.

Throughout the thesis we use several notations, some of them standard and some others newly defined for the particular topic under consideration. However, in order to avoid confusion we try to clarify any unfamiliar notation on its first appearance and thereafter where necessary. On the other hand, we make use of some standard abbreviations such as: r.v. (random variable), pdf (probability density function), \mathbb{R} (real numbers), \mathbb{R}^+ (positive real numbers), \mathbb{Z} (integers).

All the probability distributions that we shall use in the thesis are defined in terms of density functions over Euclidean spaces with respect to Lebesgue measures. We adopt either "p" or "f" as a generic symbol for a probability density function. Also, we use the conventional distinction between a random variable and its realisation as a value, i.e. capital letters X,Y, etc. representing random variables and lower case letters x,y etc. representing their realised values.

To conclude, the term "parameter" is extensively used in the thesis to mean the random variable representing the "level" of the steady state process. The Greek letter θ , sometimes suffixed θ_t , is the generic

symbol we use to represent for the level to avoid misunderstanding with parameters of a probability distribution, which are usually represented by the conventional Greek letters α , β , γ , μ , σ etc. We reserve the term Y_t for the random variable representing the process observation in the model formulation.

CHAPTER 2: ENTROPY FUNCTION

2.1) Historical Remarks

The word *entropy* has had a long and controversial evolution in science. In the original greek its literal meaning is *transformation* and it was with this literal sense that in 1850 Clausius [see Tribus 1961a and 1969] introduced the word *entropy* in his work as a quantity associated with transformations from work effects to heat effects in thermodynamics. It was only at the beginning of this century that it was used again, this time in a completely different subject, in the works of S. Boltzmann and M. Planck in Statistical Mechanics. They proposed a general procedure for determining the distribution of the total energy of a system among its elemental single components, when the assumption is made that all such elemental single components are independent and identically distributed. The *Boltzmann H-functions* which originated from their work, is used a great deal in statistical mechanics [Planck. 1950; Mackey 1957].

It was, however, only in 1948 that it became universally known due to the work of C.E. Shannon in the context of communication theory [Shannon & Weaver, 1949]. In his work Shannon developed thoroughly a new and useful axiomatic quantitative study of the acquisition, production and transmission of information, named afterwards *Shannon's Information Theory* . This work produced again another definition of an *entropy function*; in this case, a quantitative measure of the missing information in a message or in a probability distribution. As remarked by Shannon, *Information Theory* is very broadly based, in the sense that

it applies to all kind of systems for which the given information is incomplete, that is, for those systems where uncertainty is involved. More generally, information theoretic concepts are relevant to any field in which inductive probabilities are useful, for inductive probabilities arise whenever the given information is not sufficient to permit deductive inferences. Although ever since Shannon, information theory had grown into a broad, highly developed body of knowledge, only in 1957 did E.T.Jaynes show that Shannon's entropy function had a deeper meaning and in fact, as a disciple of statistical mechanics, he demonstrated that both *entropies* were in fact the same thing and therefore not mere analogies. [Jaynes, 1957 & 1958; Tribus, 1961a]

2.2) The notion of Entropy

Let $S = (\zeta_1, \zeta_2, \dots, \zeta_n)$ be the set of possible outcomes ζ 's in some physical experiment. Suppose also that at first we do not know anything more about the experiment and the occurrence of any of the possible outcomes. Then, suppose we are told that the outcome ζ_i is more likely to occur. Provided the given information is reliable, our previous state of knowledge must change and it would be useful to have a quantitative measure for the information newly acquired. Putting the problem in a quantitative form, suppose that our original state of knowledge and our state of knowledge after receiving the information are represented by probability assignments P^0 and P respectively; in other words, we have two probability schemes:

$$(S, F, P^0) \text{ and } (S, F, P)$$

where: S is the sample space (assumed finite)

F is the field of events

$$P^0 = (p_1^0, p_2^0, \dots, p_n^0) ; p_i^0 = \text{Prob}(\zeta_i)$$

$$P = (p_1, p_2, \dots, p_n) ; p_i = \text{Prob}(\zeta_i | \text{Inform.})$$

The above set up for the problem allows us to introduce the concepts of information and entropy. Firstly if we are interested in a quantitative measure for the information provided by the new data relative to our prior knowledge, we have to take into account the two probability distributions P^0 and P , representing respectively our state of uncertainty before and after gaining the information. We finish up with a quantity $I(P, P^0)$ known as *information in P relative to P^0* or simply *information*. Secondly, the problem could be formulated in a slightly different way, where we could only be interested in an absolute quantitative measure of the information. The quantity proposed by Shannon, known as *Shannon's Entropy*, is a measure of the missing information or the amount of uncertainty in a single probability assignment. Put in this way, we can clearly see the basic conceptual difference between Information and Entropy. In the first we measure quantitatively information in a probability assignment relative to a prior assignment, while in the second we have the same sort of measure in an absolute way. We shall point out later that Shannon's entropy, although simpler and easier to work with, suffers from the defect that it can not be consistently generalised from discrete to continuous probability spaces. On the other hand $I(P, P^0)$, being a relative measure of information does not suffer from this defect. Attempts have been made to

formulate a clear, simple and consistent measure of information or even to develop a general theory in terms of information rather than entropy. Among the various works in this particular area we cite: Vincze, (1972); Hobson, (1971); Kolmogorov, (1956); Kullback, (1959); Jaynes, (1968), Vincze, (1959 & 1965) and Pérez, (1957).

2.3 Definition of Entropy-Discrete Case

Let S_n denote the set of all finite discrete probability distributions $\{P=(p_1, p_2, \dots, p_n); p_i \geq 0; i=1, 2, \dots, n; \sum p_i=1\}$.

In other words, P may be regarded as an experiment having n possible outcomes x_1, x_2, \dots, x_n with probabilities $p(x_1)=p_1, p(x_2)=p_2, \dots, p(x_n)=p_n$. Then, *the entropy of the distribution P* , or a measure of how uncertain we are about the outcome of the experiment is given by:

$$H(P) = H(p_1, p_2, \dots, p_n) = - \sum_i p_i \ln p_i \quad (2.1)$$

for $P \in S_n$ and all $n=1, 2, \dots$ and also, with the usual convention that whenever $p_i=0$ we set $p_i \ln p_i=0$.

Theorem: (Fundamental Theorem of Information Theory).

Up to a constant of proportionality, the function $H(P)$ given in equation (2.1) is the only function satisfying the three requirements for being a measure of uncertainty of an assignment of probability P :

- i) Continuity on p_i
- ii) Monotonic increasing function of "n" if all the p_i are

equal ($p_i=1/n$). That is, with equally likely events, there is more choice, or uncertainty when there are more possible events.

iii) Consistency:

$$H(p_1, p_2, \dots, p_n) = H(p_1+p_2, p_3, \dots, p_n) + (p_1+p_2) \cdot H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$$

or, if a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H .

Proof: The original proof of the theorem is found in Shannon and Weaver, (1949-Appendix 2), and some elaborated proofs can be found in Mathai & Rathie, (1975); Feinstein, (1958) and Akaike, (1971).

2.4 Basic Properties of Discrete Entropy

Apart from the properties (i) to (iii) above, Shannon's entropy has many other properties and characterisations, some of which we show below. For a thorough treatment of these properties, see for instance:

Shannon & Weaver, (1949); Mathai & Rathie, (1975) and Kullback, (1959) .

Using the index n in $H(P)$ to denote the entropy of $P=(p_1, p_2, \dots, p_n)$ i.e., $H_n(P)=H(p_1, p_2, \dots, p_n)$; we enumerate the following further properties of $H_n(P)$:

1) Non-Negativity:

$$H_n(P) \geq 0 \quad (H_n(P)=0 \text{ if and only if } p_i=1 \text{ for some } i=1,2,\dots,n)$$

2) Expansibility:

$$H_{n+1}(P,0) = H_n(P)$$

i.e., the entropy remains the same if we add possibilities with zero probability.

3) Inequality and Maximum Value:

$H_n(p_1, p_2, \dots, p_n) \leq H_n(1/n, 1/n, \dots, 1/n)$ with equality if and only if $p_i = 1/n$ for all $i=1, 2, \dots, n$.

Also, by substitution in (2.1), the maximum H_n exists and is equal to $\ln n$, when all the p_i are equal to $1/n$.

For instance, when $n=2$ let: $p_1=p$ and $p_2=1-p$

Thus $H_2(p_1, p_2) = -p \ln p - (1-p) \ln(1-p)$ and $\max H_2(p_1, p_2) = \ln 2 = H_2(1/2; 1/2)$ as shown below in the graph of $H_2(p_1, p_2)$ against p :

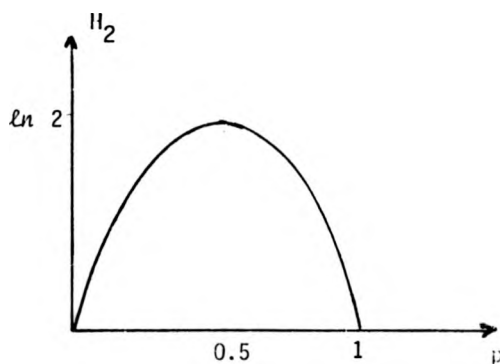


Figure 2.1 : $H_2(p, 1-p) \times p$

4) Symmetry:

$$H_n(p_1, p_2, \dots, p_n) = H_n(p_{\alpha_1}, p_{\alpha_2}, \dots, p_{\alpha_n})$$

where $(\alpha_1, \alpha_2, \dots, \alpha_n)$ is any arbitrary permutation of the indices $(1, 2, \dots, n)$. From the above, we can state that the entropy is the same whatever the order in which the possible outcomes are labelled.

5) Joint Events:

Let:

$$P^1 = (p_1^1, p_2^1, \dots, p_n^1) \in S_n$$

$$P^2 = (p_1^2, p_2^2, \dots, p_m^2) \in S_m$$

where S_n and S_m are classes of all finite discrete probability distributions P^1 and P^2 respectively.

$$P = (P^1, P^2) = (p_{11}, \dots, p_{1m}, \dots, p_{n1}, \dots, p_{nm}) \in S_{nm} \quad p_{ij} \text{ is}$$

the probability of joint occurrence of i with probability p_i^1 and j with probability p_j^2 .

S_{nm} as above.

Then:

$$H_{nm}(P) \leq H_n(P^1) + H_m(P^2)$$

Alternatively, the entropy or uncertainty of a joint experiment is less than or equal to the sum of the entropies of the individual experiments. It is equal if and only if the individual experiments are independent.

6) Coherence:

This property is in fact a direct consequence of property 3),

but it is worth mentioning in its own right. As a measure of uncertainty in a probability assignment, for any change toward equalisation of the p_i (loss of information or increase of the uncertainty), the entropy increases.

Formally, if we have:

$$P=(p_1, p_2, \dots, p_n) \text{ and } P^*=(p_1^*, p_2^*, \dots, p_n^*)$$

and $\sum_i |p_i - 1/n| \geq \sum_i |p_i^* - 1/n|$, then:

$$H_n(P) \leq H_n(P^*)$$

7) Conditional Entropy

Let:

P^1, P^2, P, p_{ij} be as defined in property (5).

$(x_1^1, x_2^1, \dots, x_n^1)$ and $(x_1^2, x_2^2, \dots, x_m^2)$ the possible outcomes of experiments P^1 and P^2 respectively.

$p(j|i)$ the conditional probability of the outcome x_j^2 given that the outcome of experiment with distribution P^1 is x_i^1 ; $i=1, 2, \dots, n$ and $j=1, 2, \dots, m$.

Then, the conditional entropy of P^2 given P^1 is:

$$H_m(P^2|P^1) = - \sum_{i,j} p_{ij} \ln p(j|i)$$

From the above and the results of property (5), we obtain:

$$H_{nm}(P) = H_n(P^1) + H_m(P^2|P^1) \text{ and } H_m(P^2) \geq H_m(P^2|P^1)$$

Verbally, the sum of the amount of uncertainty in the probability assignment P^1 for the first experiment and the amount of uncertainty for the conditional experiment is the entropy of the joint experiment. Also, the above inequality states that if there is any dependence between two experiments, there is always a gain of information (or a decrease of the degree of uncertainty) of one of the experiments, given the knowledge about the outcome of the other.

8) Invariability:

Let:

X be a discrete random variable which can assume values

x_1, x_2, \dots, x_n with probabilities $p_i = P(X=x_i)$, $i=1, 2, \dots, n$.

H_X represents the entropy of the experiment under consideration (instead of using the $H(P)$ notation of (2.1)).

$Y = t(X)$ a one-to-one transformation of the random variable X and H_Y its associated entropy.

Then, this property states that:

$$H_Y = H_X = H(P)$$

That is, the formula (2.1) for the entropy of an experiment is invariant with respect to any bijective transformation of the variable; it is not dependent on the domain of the variable, but depends only on the probability distribution.

The properties just presented in no sense exhaust the properties

- and characterisations of Shannon's entropy function. The prime objective of describing these few properties was to clarify the ideas behind the entropy function as an absolute measure of the amount of uncertainty in a single assignment of a probability distribution for an experiment. For a detailed mathematical and probabilistic study of all the properties and characterisation theorems of Shannon entropy, we refer mainly to Mathai & Rathie, (1975) .

2.5 The Extension to the Continuous Case:

If in the definition of section 2.3 we let the number of possible outcomes n for a given experiment increase indefinitely so that P tends to a continuous probability density function $p(x)$ of a continuous random variable $X \in \mathcal{X}$, it would be natural to try to define the entropy as a limiting case of the entropy for discrete distributions (2.1).

However, if we do so, we obtain:

$$H [p(x)] = - \int_{\mathcal{X}} p(x) \ln p(x) \cdot dx - \lim_{\Delta x_i \rightarrow 0} \sum_i \Delta x_i \cdot p(x_i) \cdot \ln \Delta x_i$$

Accordingly, the expression for $H [p(x)]$ diverges as $\Delta x_i \rightarrow 0$ whatever the value of the first term. Instead of defining $H[p(x)]$ as a limiting case, Shannon suggests that we should simply define the entropy for a continuous random variable $X \in \mathcal{X}$ with probability density function $p(x)$ purely by analogy as follows:

$$H [p(x)] = - E_{p(x)} \{ \ln p(x) \} = - \int_{\mathcal{X}} p(x) \cdot \ln p(x) \cdot dx \quad (2.2)$$

and for a random vector $\underline{x} = (x_1, x_2, \dots, x_n)^T \in \mathcal{X}^n$ and associated $p(\underline{x})$:

$$H [p(\underline{x})] = - E_{p(\underline{x})} \{ \ln p(\underline{x}) \} = - \int_{\mathcal{X}_1, \dots, \mathcal{X}_n} p(\underline{x}) \cdot \ln p(\underline{x}) \cdot dx_1, \dots, dx_n \quad (2.3)$$

The entropy as defined in (2.2) or (2.3) has nearly all the important properties described in the last section and as such, is a measure of the amount of uncertainty in the probability assignment $p(x)$ for a continuous random variable X . However, as remarked by Shannon, the continuous entropy function (2.2) or (2.3), is not general in the sense that for some particular cases, properties (1) and (8) are not attained. Let us first consider the lack of invariance under a monotonic change of variable.

Let:

X be a continuous random variable, $X \in \mathcal{X}$, with pdf $p_X(X)$.

$Y = g(X)$ be a monotonic transformation of X .

Thus, Y is also a continuous random variable, $Y \in \mathcal{Y}$, with pdf $p_Y(Y)$.

Then, by (2.2):

$$H(X) = - \int_{\mathcal{X}} p_X(x) \cdot \ln p_X(x) \cdot dx \quad \text{and} \quad H_Y = - \int_{\mathcal{Y}} p_Y(y) \cdot \ln p_Y(y) \cdot dy$$

since, by definition $p_Y(Y) = p_X [g^{-1}(Y)] \cdot |J|$, where $|J| = |dx/dy|$ is the jacobian of the transformation, substitution in the above equations gives:

$$H_Y = - \int_Y p_X[g^{-1}(Y)] \cdot |J| \cdot \ln \{p_X[g^{-1}(Y)] \cdot |J|\} \cdot dy$$

or, after expanding the logarithm:

$$H_Y = - \int_X p_X(x) \cdot [\ln p_X(x) + \ln |J|] \cdot dx$$

and finally:

$$H_Y = H_X - E_{p_X(x)} \{ \ln |J| \} \quad \text{-----(2.4)}$$

Equation (2.4) clearly shows the dependence of the entropy of Y on the Jacobian on the transformation, confirming the lack of invariance under the change of variable $x \rightarrow g(x)$. This restriction led Shannon to give an extra interpretation to entropy. For both, the discrete and the continuous case (2.1) and (2.2) measure the randomness or the amount of uncertainty involved in the assignment P or $p(x)$ to a discrete or a continuous random variable X respectively. However, the measurement in (2.1) is completely absolute in the sense that no matter what random variable is describing the experiment, the entropy is always the same. On the other hand, the entropy in (2.2) or (2.3), measures the uncertainty relative to the coordinate system (sample space) adopted, i.e., relative to the random variable used. It is however important to remark that, in most of the applications, we in fact are interested in the increase or decrease of the amount of uncertainty of systems whose randomness is changing continuously in time. In this case the Jacobian term of (2.4) would appear in both entropies, cancelling out eventually. This means

that the lack of invariance of the measure (2.2) is not a restriction to its use.

With respect to the possible situations in which the entropy is negative, the problem can be easily circumvented by adopting a scale of measurement for the entropy for each kind of distribution under consideration.

Let us consider, for example, the normal distribution:

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then: } H_X = \ln \sqrt{2\pi e \sigma^2}$$

It is quite clear from the above that H_X can assume any value in \mathbb{R} and also that zero entropy does not mean perfect information or a degenerate distribution. In fact, $H_X = 0$ for $\sigma^2 = \sigma_0^2 = (2\pi e)^{-1}$ means that there is still some uncertainty (though small), about the outcome of the experiment. We could for instance, adopt this state of uncertainty as the standard one and then compare subsequent values of H_X with this standard. Any positive H_X would indicate that we have a broader distribution than σ_0^2 and a negative H_X would indicate a still narrower distribution than σ_0^2 , that eventually tends to $-\infty$ as σ^2 approaches zero.

2.6) Other Approaches to the Continuous Extension

Although we shall use the simplified Shannon's entropy (2.2) or (2.3) in our model formulation later on, it is worth mentioning some other attempts towards a general definition of entropy. A lot of different approaches to the problem have been put forward after Shannon and in all of them, a slightly different interpretation of

a *measure of uncertainty* is made in order that a unique function is obtained for both the discrete and continuous cases. We briefly describe a few of these approaches and point out their similarities.

We start with the work by Hobson [Hobson, 1971; Reza, 1961 and Pinsker, 1964]. He sets up the problem by first defining a relative measure of information for discrete distribution and then, extending it to the continuous case.

Let:

$S = \{ \zeta_1, \zeta_2, \dots, \zeta_n \}$ be a finite sample space.

P^0, P be a pair of probability distribution assignments in S before and after gaining some evidence about the outcome of the experiment respectively, where:

$$P^0 = \{ p_1^0, p_2^0, \dots, p_n^0 \} ; \quad p_i^0 = \text{Prob} \{ \zeta_i \} ; \quad i=1,2,\dots,n$$

$$P = \{ p_1, p_2, \dots, p_n \} ; \quad p_i = \text{Prob} \{ \zeta_i | \text{Inform.} \}; \quad i=1,2,\dots,n$$

Then, instead of defining a measure of the information missing in a single probability assignment as Shannon did, Hobson defines a quantitative measure for the information provided by the new data which he called *Information in P relative to P⁰* or simply *Information* as:

$$I(P, P^0) = E_P \{ \ln(p_i/p_i^0) \} = \sum_{i=1}^n p_i \cdot \ln(p_i/p_i^0) \quad \text{--- (2.5)}$$

Hobson shows that the above quantity, while measuring the gain of information instead of the missing information, satisfies all the

main properties of Shannon's entropy and that it is easily extended to the continuous case, preserving the properties.

The extension from discrete to continuous variables is first made by extending the measure in (2.5) from a finite discrete to an infinite discrete sample space. For this case $I(P, P^0)$ becomes:

$$I(P, P^0) = \sum_{i=1}^{\infty} p_i \ln(p_i/p_i^0) \quad \text{--- (2.6)}$$

Assuming S to be a segment of the real line ($a \leq x \leq b$) and $P^0 \& P$ a pair of continuous probability assignments with densities $f^0(x)$ and $f(x)$; the information in P relative to P^0 , or the information in $f(x)$ relative to $f^0(x)$ is easily obtained using (2.6) and, taking limits of discrete partitions in $[a, b]$, we obtain:

$$\begin{aligned} I(P, P^0) &= I[f(x), f^0(x)] = E \{ \ln [f(x)/f^0(x)] \} = \\ &= \int_a^b f(x) \cdot \ln \frac{f(x)}{f^0(x)} \cdot dx \quad \text{--- (2.7)} \end{aligned}$$

The relative measure of information for a continuous distribution in (2.7), as opposed to the absolute measure of missing information in (2.2), is non negative and invariant under a one-to-one transformation $X \rightarrow Y = g(x)$.

Hobson then proceeds by introducing a concept similar to Shannon's entropy, defining a measure of missing information or uncertainty in the probability assignment P , by considering the prior

assignment P^0 and the assignment P^m , corresponding to the maximum knowledge about the outcomes ζ 's.

Since $I(P^m, P^0)$ is the maximum information possible relative to P^0 and $I(P, P^0)$ is the actual information relative to P^0 , the missing information necessary to attain the maximum knowledge state P^m (missing information or *uncertainty in P*), is:

$$U(P; P^m, P^0) = I(P^m, P^0) - I(P, P^0) \quad - - - (2.8)$$

Again, the above quantity has all the properties required for a measure of uncertainty in P ; it is applicable to either the discrete or the continuous case but has the disadvantage of requiring the knowledge of two extra probability assignments namely the prior P^0 and the maximum state of knowledge P^m .

Another interesting approach towards a general definition of entropy is that of Vincze (1959), (1965) and (1972). He starts by giving a rather different interpretation to Shannon's entropy in discrete finite space. Vincze interprets entropy as a measure related not to the probability distribution, but to a decomposition of the space of the elementary events.

If $D_N = (A_1, A_2, \dots, A_N)$ is a decomposition of $S = \{\zeta_1, \zeta_2, \dots, \zeta_n\}$ and $P_N = \{p_i = P(A_i); i = 1, 2, \dots, N; \sum_{i=1}^N p_i = 1\}$, then the entropy associated with the particular decomposition D_N is given by:

$$H_N = -E_{P_N} \{\ln p_i\} = - \sum_{i=1}^N p_i \cdot \ln p_i \quad - - - (2.9)$$

where $H_N \in [0, \ln N]$. The above measure of uncertainty is in fact Shannon's entropy (2.1). However, instead of considering H_N for measuring the uncertainty associated with the decomposition D_N , Vincze suggests an equivalent measure called information denoted by I_N that has the property of measuring uncertainty by means of information, defined by:

$$I_N = E_{p_N} \{ \ln N \cdot p_i \} = \ln N - H_N \quad (2.10)$$

where $I_N \in [0, \ln N]$.

As remarked by Vincze, one of the main advantages of using (2.10) instead of (2.9) is that under mild conditions concerning the continuous distribution, although H_N tends to infinity, the remaining information I_N will have a finite limit. In fact, when we pass from the discrete to the continuous case, the above information I_N (also known as complementary entropy), tends to a limit called I-divergence in the literature but interpreted in this context as the information of a continuous random variable $X \in \mathcal{X}$ and given by:

$$I(X) = E_{f(x)} \left\{ \ln \frac{f(x)}{\phi(x)} \right\} = \int_{\mathcal{X}} f(x) \cdot \ln \frac{f(x)}{\phi(x)} \cdot dx \quad (2.11)$$

where $f(x)$ is the probability density function of X and $\phi(x)$ is the *distribution of our interest*, defined by a reasonable partition of \mathcal{X} .

Some interesting applications of the use of the I-divergence in finding confidence intervals for unknown parameters of various density functions, by a suitable choice of the distribution of interest are shown in Vincze,(1965).

Finally, we briefly mention Jaynes' set up for the same problem [Jaynes, 1958 & 1968].

In his work, Jaynes is only interested in finding an absolute measure of uncertainty for a continuous distribution. In fact, he departs from (2.1) for the entropy of a discrete distribution. He then points out the restrictions of (2.2) for measuring the same thing for the continuous case by emphasizing once more that (2.2) is not a result of any derivation. He proceeds with his argument by taking the entropy of a discrete distribution to the limit obtaining:

$$H[p(x)] = - E \left\{ \ln \frac{p(x)}{m(x)} \right\} = - \int_X p(x) \cdot \ln \frac{p(x)}{m(x)} \cdot dx \quad - \quad - \quad (2.12)$$

where $m(x)$ is an invariant measure, proportional to the limiting density of discrete points. In this case, both $p(x)$ and $m(x)$ transform in the same way under a change of variable and so, $H[p(x)]$ of (2.12) is an invariant measure. In fact, an extra interpretation given to $m(x)$ by Jaynes is that: *apart from a normalising constant, $m(x)$ is a prior distribution describing complete ignorance about X .*

We conclude this section by remarking that whether we use Hobson's information (2.7), Vincze's I-divergence (2.11) or Jaynes'

$H [p(x)]$ (2.12) for measuring the randomness in the probability density function assigned to a continuous random variable a subjective prior assignment $f^0(s)$, $\phi(x)$ or $m(x)$ is required. However, all three approaches are general, in the sense that all desirable properties are preserved.

CHAPTER 3: JAYNES' PRINCIPLE OF MAXIMUM ENTROPY

3.1) Introduction

Let us consider the simple form of Bayes' theorem for a discrete random variable X_i , written as:

$$p(x_i|DK) \propto p(D|x_i,K) \cdot p(x_i|K)$$

One of the main controversies in using the above theorem has been the question of how to assign prior probabilities $p(x_i|K)$, based only on the information K prior to any observation. We could for instance, break the situation up into mutually exclusive and exhaustive possibilities and use the *principle of insufficient reason* in such a way that no one of them is preferred to any other, i.e., assigning a *uniform prior*. However, situations occur in which we are given some other relevant evidence that increases our state of knowledge in such a way that the uniform prior assignment turns out to be inappropriate. In this case, with this extra prior information, we have some reason to prefer some possibilities to others. Our aim is to assign a probability which is, in some sense, as uniform as it can be subject to the available information. It should *spread out* all over the sample space, not assigning zero probability to any situation, unless the available information really leads to this conclusion.

So, the aim of avoiding unwarranted conclusions leads us to search for a reasonable function that measures *the uniformity* of a probability distribution which could be maximised subject to the constraints which represent the available information. In fact, this function which we seek

measures the *uncertainty* or *ignorance* about a situation whose maximisation, subject to the constraints, would give us the minimally prejudiced assignment of a probability distribution.

In this chapter we will show that the only function that gives the minimally prejudiced distribution required in the above set up of our problem is the Shannon entropy developed in chapter 2. Before we proceed with the mathematical formulation of this problem, we show first through some simple examples that other functions, such as the variance or $E\{p_i\}$ (or $E\{p(x)\}$ for the continuous case) which also measure the p_i spread, uniformity or uncertainty of a probability distribution do not give the minimally prejudiced distribution we want.

Let us first consider a die throwing experiment in which we are given the information:

- i) The die has six sides with " $f_i=i$ " spots on the i^{th} side.
- ii) The average number of spots obtained in a previous long series of throws was 4.5 (instead of 3.5 for a fair die).

Based on these two pieces of information, we want to assign a minimally prejudiced probability distribution to this experiment;

$$P\{f_i = i\} = p_i, \quad i = 1, 2, \dots, 6$$

and let us suppose first that we choose the variance of the required distribution as the objective function, that is:

$$\text{Max } \sum_{i=1}^6 (f_i - 4.5)^2 \cdot p_i \quad \text{--- (3.1)}$$

subject to:

$$\sum_{i=1}^6 f_i p_i = 4.5 \quad \text{----- (3.2)}$$

$$\sum_{i=1}^6 p_i = 1 ; \quad p_i \geq 0, \quad i=1, \dots, 6 \quad \text{----- (3.3)}$$

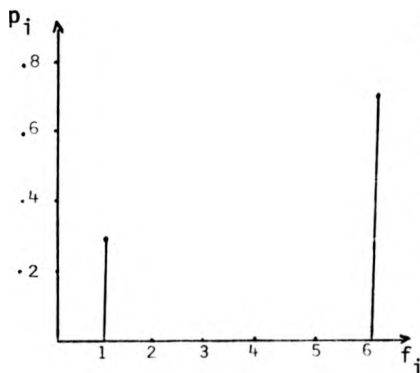
The solution to this maximisation procedure is:

$$P\{f_1\} = 0.3 ; \quad P\{f_6\} = 0.7 ; \quad P\{f_2\} = P\{f_3\} = P\{f_4\} = P\{f_5\} = 0$$

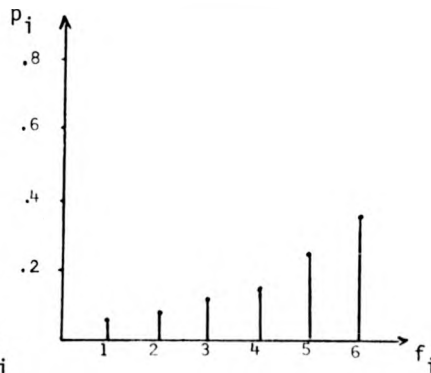
On the other hand, if we use Shannon entropy (2.1) in place of (3.1) above as the objective function we would obtain by its maximisation subject to the constraints (3.2) and (3.3):

$$P\{f_1\} = 0.055 \quad P\{f_2\} = 0.079 \quad P\{f_3\} = 0.114 \quad P\{f_4\} = 0.165$$

$$P\{f_5\} = 0.240 \quad P\{f_6\} = 0.347$$



Max. Variance distribution



Max. Entropy distribution

Comparison between the two distributions shows clearly the inadequacy of the variance as the uncertainty function. Accordingly, it arbitrarily assigns zero probabilities whereas the given information does not imply this. In contrast, the maximum entropy distribution takes full account of the provided information by spreading out the distribution over the sample points without jumping to conclusions not explicitly stated.

As a second example, let us consider a simple version of the die experiment. Consider an experiment that admits only three possible outcomes and let X_i be a discrete random variable that can only take the values 1, 2 and 3. Suppose also that we are given the extra information about \bar{X} , the mean of X_i .

As in the last example, we want to assign the least prejudiced distribution for X_i . Let us consider first the function $-E\{p_i\} = -\sum p_i^2$ as the uncertainty function to be maximised. So, we are to find:

$$P = \{(p_1, p_2, p_3) ; p_i = \text{Prob}(X_i = i) ; i = 1, 2, 3\}$$

so that:

$$F(p_i) = -E\{p_i\} = -\sum_{i=1}^3 p_i^2 \text{ is maximised} \quad (3.4)$$

subject to the constraints:

$$\sum_{i=1}^3 p_i = 1 \quad (3.5)$$

$$\sum_{i=1}^3 X_i p_i = \bar{X} \quad (3.6)$$

It is easy to show that the solution to the above problem (using for example the Lagrange multipliers) as a function of \bar{X} is:

$$p_1 = (8-3\bar{X})/6 \quad p_2 = 1/3 \quad p_3 = (3\bar{X}-4)/6$$

Plotting these probabilities against \bar{X} we get:

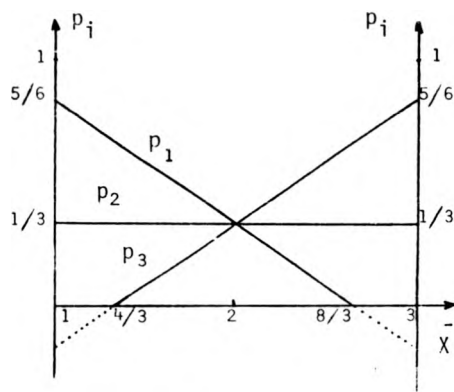


Figure 3.1 :
Max $-E\{p_i\}$ before adjustment
for negative probabilities.

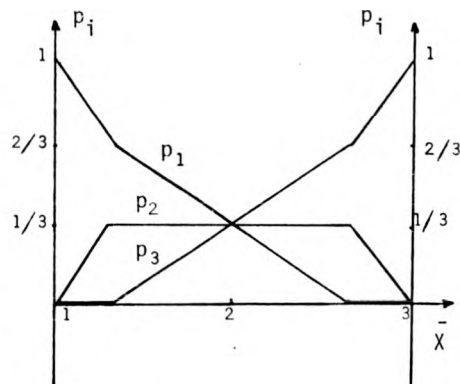


Figure 3.2 :
Max $-E\{p_i\}$ after adjustment for
negative probabilities.

In figure 3.1 above the curves for p_1 and p_3 clearly show that for $1 \leq \bar{X} \leq 4/3$ and $8/3 \leq \bar{X} \leq 3$ respectively, the probabilities are negative. To replace this impossibility we introduce the extra constraint that $p_i \geq 0$; $i=1,2,3$ and we obtain the final result as plotted in figure 3.2.

As a matter of comparison, let us solve the same problem by using Shannon entropy $H(p_i)$ instead of $F(p_i)$ in (3.4). Using again the same argument, the following distribution is obtained:

$$p_i = \exp\{(2-i)\alpha\} / (1+2 \cosh \alpha) ; \bar{x} = (e^{2\alpha} + 2e^{\alpha} + 3) / (e^{2\alpha} + e^{\alpha} + 1)$$

or, after simplifying:

$$p_2 = \sqrt{[4-3(\bar{x}-2)^2]}/9 - 1/3 ; p_1 = (3 - \bar{x} - p_2) / 2 \quad \text{and}$$

$$p_3 = (\bar{x} - 1 - p_2) / 2$$

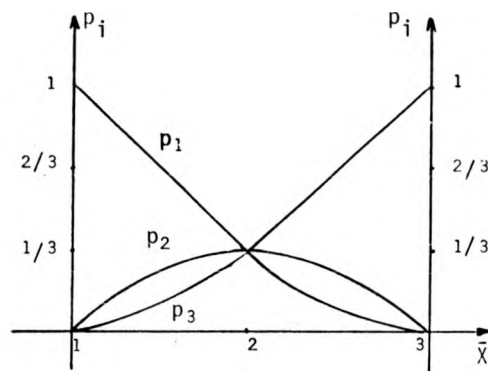


Figure 3.3 :

Maximum entropy distribution.

Although the $\text{Max} - \sum p_i^2$ shows a big improvement over the maximum variance distribution (see the die experiment of the previous example), for certain values of \bar{x} it assigns zero probabilities and that is again jumping to conclusions not present on the given information. On the other hand, the maximum entropy distribution (figure 3.3) represents in fact the least prejudiced probability distribution for X_i that meets the objectives of our problem. Another point in favour of the entropy is that the extra constraint $p_i \geq 0$, which must be introduced in the first case, is automatically included in the entropy formulation.

The two simple examples discussed, illustrates how the entropy function is in fact a consistent measure of uncertainty, and that it leads to least assignment of probability distribution for a random variable.

In the next section we show the mathematical set up of the problem by postulating the principle and the general solution.

3.2) Jaynes Principle of Maximum Entropy:

We now formalise the procedure to find the least prejudiced probability assignment introduced in the last section. Originated in 1957 by E.T. Jaynes, the rationale behind the proposed principle of maximum entropy is that the probability distribution desired has maximum uncertainty (minimum information content) while representing some explicitly stated known information.

The principle is general, in the sense that it always gives a minimally prejudiced probability distribution, although, as stated by Jaynes, (1958) and (1968), the information given concerning the random variable in question, should be a *testable piece of information*, defined as follows:

A piece of information concerning a random variable X is called testable if for any proposed probability assignment $p(x)$ for X , there is a procedure which will determine unambiguously whether $p(x)$ does or does not agree with the given information.

Before we state the principle, we would like to point out that

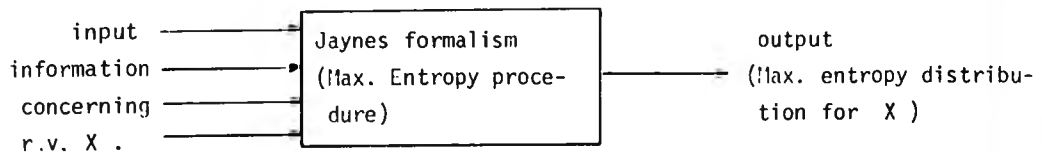
among all the possible testable information, Jaynes considers in his formalism only those concerned with averages of functions of the random variable being studied, since this class of information is the most common one we find in practical problems. But the principle as a whole, is applicable to any kind of testable information.

We now formulate the principle and its mathematical set up mainly for the continuous case. The discrete development is similar and has been extensively explored in the literature. For comprehensive developments and illustrative examples see: Jaynes, (1958,1963 and 1968); Hobson, (1971); Tribus, (1961a & 1969) and Goldman,(1953).

The principle:

The minimally prejudiced probability distribution is that which maximises the entropy subject to constraints supplied by the given testable information.

Put this way, Jaynes' principle encompasses the well known principle of insufficient reason as a special case. However, there is no way of proving Jaynes formalism. As pointed out by Tribus,(1961a) it should rather be interpreted as an axiom for a system of inductive logic. To see this point more clearly, let us consider the schematic representation for the principle as shown below:



Accordingly, if the output conclusions agree with posterior observations of the experiment, we conclude that the input information is coherent and sufficient for our purpose. On the other hand, an output not agreeing with the observations, forces us to admit that the input information is not correct and finally, a vague output corresponds to insufficient input information.

Bearing in mind this rationality behind the principle, let us now proceed with the calculations in order to obtain the maximum entropy distribution.

We are faced with the so-called isoperimetric problem of the calculus of variations that could be formulated generally as:

Find p as a function of $x \in X$ such that the function $I(p)$ defined as:

$$I(p) = \int_X F(X,p) \cdot dx \quad \text{--- (3.7)}$$

is maximised, subject to the conditions:

$$\int_X \phi_i(X,p) \cdot dx = K_i ; \quad i=1,2,\dots, n \quad \text{--- (3.8)}$$

where $\phi_i(X,p)$ and K_i are preassigned functions of X,p and constants respectively. From the calculus of variations, the $p(x)$ which maximises $I(p)$ is obtained by solving the equation:

$$\frac{\partial F}{\partial p} + \lambda_1 \frac{\partial \phi_1}{\partial p} + \dots + \lambda_n \frac{\partial \phi_n}{\partial p} = 0 \quad \text{--- (3.9)}$$

Where $\lambda_i, i=1,2,\dots, n$ are adjustable constants (Lagrange multipliers), calculated by direct substitution of $p(x)$ into constraint equations (3.8).

We can now easily adapt our problem to the above set up as follows:

$X \in \mathcal{X}$ is a continuous variable

$p(x)$ is the probability density of X , to be obtained by maximising the entropy (2.2), i.e., by setting $F(X,p) = -p(x) \cdot \ln p(x)$ in (3.7).

$\phi_i(X,p) = g_i(x) \cdot p(x); i=1,2,\dots, n$; where $g_i(x)$ are known functions of X , whose expectations with respect to $p(x)$ are known and equal to K_i - constraint equations.

$$\int_{\mathcal{X}} p(x) \cdot dx = 1 \quad \text{is the normalising constraint.}$$

Taking these quantities into the general solution (3.9) (with an additional adjustable constant λ_0 due to the normalising constraint) we obtain after simplifications the maximum entropy density $p(x)$:

$$p(x) = z \cdot \exp \left\{ - \sum_{i=1}^n \lambda_i g_i(x) \right\}; \quad z = \exp \{ -\lambda_0 \} \quad \text{--- (3.10)}$$

(The discrete case is similarly set by substituting summations for integrals).

3.3) Properties of the Maximum Entropy Density:

We now state and prove some of the statistical properties of $p(x)$ (equation 3.10). Though many properties and mathematical relations

Where $\lambda_i, i=1,2,\dots, n$ are adjustable constants (Lagrange multipliers), calculated by direct substitution of $p(x)$ into constraint equations (3.8).

We can now easily adapt our problem to the above set up as follows:

$X \in \mathcal{X}$ is a continuous variable

$p(x)$ is the probability density of X , to be obtained by maximising the entropy (2.2), i.e., by setting $F(X,p) = -p(x) \cdot \ln p(x)$ in (3.7).

$\phi_i(X,p) = g_i(x) \cdot p(x); i=1,2,\dots, n$; where $g_i(x)$ are known functions of X , whose expectations with respect to $p(x)$ are known and equal to K_i - constraint equations.

$\int_{\mathcal{X}} p(x) \cdot dx = 1$ is the normalising constraint.

Taking these quantities into the general solution (3.9) (with an additional adjustable constant λ_0 due to the normalising constraint) we obtain after simplifications the maximum entropy density $p(x)$:

$$p(x) = z \cdot \exp \left\{ - \sum_{i=1}^n \lambda_i g_i(x) \right\}; z = \exp \{ -\lambda_0 \} \quad \text{--- (3.10)}$$

(The discrete case is similarly set by substituting summations for integrals).

3.3) Properties of the Maximum Entropy Density:

We now state and prove some of the statistical properties of $p(x)$ (equation 3.10). Though many properties and mathematical relations

can be derived from the maximum entropy approach, we only show those that specifically concern our work.

We conclude the section by stating and proving theorem and a corollary, important for our model formulation. A parallel development for the discrete case can be found in chapter 5 of Tribus, (1969).

i) Partition Function Properties:

"The mean, variance and covariance of the random variables $g_i(x)$; $i=1,2,\dots,n$ are related to the Lagrange multipliers $\lambda_1, \lambda_2, \dots, \lambda_n$ and the Partition Function (zeroth Lagrange multiplier λ_0 ; also known as Potential Function) by:

$$E_{p(x)} \{g_i(x)\} = - \frac{\partial \lambda_0}{\partial \lambda_i} \text{ ----- (3.11)}$$

$$\text{Var}_{p(x)} \{g_i(x)\} = \frac{\partial^2 \lambda_0}{\partial \lambda_i^2} \text{ ----- (3.12)}$$

$$\text{Cov}_{p(x)} \{g_i(x) \cdot g_j(x)\} = \frac{\partial^2 \lambda_0}{\partial \lambda_i \cdot \partial \lambda_j} \text{ ----- (3.13)}$$

$$i, j = 1, 2, \dots, n "$$

Proof:

Taking $p(x)$ of (3.10) into the normalising constraint, we get:

$$\int_{\mathcal{X}} e^{-\lambda_0} e^{-\sum_k \lambda_k g_k(x)} \cdot dx = 1$$

or:
$$e^{\lambda_0} = \int_{\mathcal{X}} e^{-\sum_k \lambda_k g_k(x)} \cdot dx \text{ ----- (3.14)}$$

Differentiating (3.14) with respect to λ_i , we obtain:

$$e^{\lambda_0} \frac{\partial \lambda_0}{\partial \lambda_i} = - \int_{\mathcal{X}} e^{-\sum_k \lambda_k g_k(x)} g_i(x) \cdot dx$$

$$\text{or: } \frac{\partial \lambda_0}{\partial \lambda_i} = - \int_{\mathcal{X}} e^{-\lambda_0} \cdot e^{-\sum_k \lambda_k g_k(x)} g_i(x) \cdot dx$$

using again (3.10):

$$\frac{\partial \lambda_0}{\partial \lambda_i} = - \int_{\mathcal{X}} g_i(x) \cdot p(x) \cdot dx = - E_{p(x)} \{ g_i(x) \} = -K_i$$

To prove (3.12) we follow the same argument by differentiating (3.14) twice with respect to λ_i . We obtain, after simplification:

$$\left(\frac{\partial \lambda_0}{\partial \lambda_i} \right)^2 + \frac{\partial^2 \lambda_0}{\partial \lambda_i^2} = \int_{\mathcal{X}} e^{-\lambda_0} e^{-\sum_k \lambda_k g_k(x)} \cdot g_i^2(x) \cdot dx$$

Using (3.10) & (3.11) we obtain:

$$E_{p(x)}^2 \{ g_i(x) \} + \frac{\partial^2 \lambda_0}{\partial \lambda_i^2} = E_{p(x)} \{ g_i^2(x) \} \quad \text{and (3.12) follows}$$

Finally, differentiating (3.14) with respect to λ_i and λ_j and taking into account (3.11) and the fact that:

$$\text{cov}_{p(x)} \{ g_i(x), g_j(x) \} = E_{p(x)} \{ g_i(x) g_j(x) \} - E_{p(x)} \{ g_i(x) \} \cdot E_{p(x)} \{ g_j(x) \}$$

expression (3.13) follows immediately.

ii) Maximum Entropy Properties:

"The maximum entropy value is related to the Lagrange multipliers λ_i and the expectations K_i ; $i=1,2,\dots, n$ by:

$$H_m = H_m(\lambda_1, \lambda_2, \dots, \lambda_n) = H_m(K_1, K_2, \dots, K_n) \text{ --- (3.15)}$$

$$\frac{\partial H_m}{\partial K_i} = \lambda_i \quad ; \quad i=1,2,\dots, n \text{ --- (3.16)}$$

where H_m is the maximum entropy value.

Proof:

Taking $p(x)$ (equation 3.10) into $H(p)$ (equation 2.2) we obtain:

$$H_m = H[p(x)] = - \int_{\mathcal{X}} \left[-\lambda_0 - \sum_i g_i(x) \cdot \lambda_i \right] \cdot p(x) \cdot dx$$

by expanding the terms within brackets:

$$H_m = \lambda_0 + \sum_i \lambda_i \cdot E_{p(x)} \{ g_i(x) \} = \lambda_0 + \sum_i \lambda_i \cdot K_i$$

Since the potential function as given in (3.14) can be expressed as a function of the λ_i 's alone and consequently the K_i 's in (3.11), H_m can be expressed as a function of the λ_i 's only; $i=1,2,\dots, n$.

Conversely, regarding the K_i 's as the independent variables, the λ_i 's could be solved for K_i 's and an expression for H_m as a function of the K_i 's alone is obtained.

To prove (3.15) let us consider the differential element dH_m from the above:

$$dH_m = d\lambda_0 + \sum_{i=1}^n K_i \cdot d\lambda_i + \sum_{i=1}^n \lambda_i \cdot dk_i$$

Using the fact that $\lambda_0 = \lambda_0(\lambda_1, \lambda_2, \dots, \lambda_n)$, $d\lambda_0$ can be written as:

$$d\lambda_0 = \frac{\partial \lambda_0}{\partial \lambda_1} d\lambda_1 + \frac{\partial \lambda_0}{\partial \lambda_2} d\lambda_2 + \dots + \frac{\partial \lambda_0}{\partial \lambda_n} d\lambda_n = \sum_{i=1}^n \frac{\partial \lambda_0}{\partial \lambda_i} \cdot d\lambda_i$$

and from (3.11) :
$$d\lambda_0 = - \sum_{i=1}^n K_i \cdot d\lambda_i$$

Therefore:
$$dH_m = - \sum_{i=1}^n K_i \cdot d\lambda_i + \sum_{i=1}^n K_i \cdot d\lambda_i + \sum_{i=1}^n \lambda_i \cdot dK_i = \sum_{i=1}^n \lambda_i \cdot dK_i$$

and (3.16) follows.

iii) Theorem:

"The maximum entropy distribution (3.10) is a member of the regular case exponential family of distributions"

Proof:

If the random variable X has a probability density function which is a member of a regular case exponential family of distributions indexed by parameters $\underline{\theta} = (\theta_1, \dots, \theta_n)$, then its pdf can be written as:

$$p(x, \underline{\theta}) = A(\underline{\theta}) \cdot \exp \left\{ \sum_{i=1}^n Q_i(\underline{\theta}) \cdot R_i(x) \right\}$$

where, for $i=1, 2, \dots, n$:

$R_i(x)$ are functions of X alone and not of $\underline{\theta}$

$A(\underline{\theta})$, $Q_i(\underline{\theta})$ are functions of $\underline{\theta}$ alone and not of X .

Let $p(x|\underline{\theta})$ be the parametrised probability density function corresponding to $p(x)$ of (3.10). Taking $p(x|\underline{\theta})$ into (2.2) it is clear that after integrating out x , we are left with a function of $\underline{\theta}$ alone ; i.e.

$$H_m = - \int_{\mathcal{X}} p(x|\underline{\theta}) \lambda \ln p(x|\underline{\theta}) \cdot dx = H_m(\underline{\theta})$$

or, using (3.15), $H_m = H_m(\underline{\theta}, \underline{K}, \underline{\lambda})$

where: $\underline{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_n)$ and $\underline{K} = (K_1, K_2, \dots, K_n)$

using (3.16) we now obtain:

$$\frac{\partial H_m(\underline{\theta}, \underline{K}, \underline{\lambda})}{\partial K_i} = \lambda_i \quad \therefore \quad \lambda_i = \lambda_i(\underline{\theta}, \underline{K}) = \lambda_i(\underline{\theta}) \quad \dots \quad (3.17)$$

$i = 1, 2, \dots, n .$

That is to say, the Lagrange multipliers λ_i , $i = 1, 2, \dots, n$ are functions of $\underline{\theta}$ alone (since \underline{K} are specified constants independent of x) and not of x .

Also, from (3.14) and taking into account (3.17) we can write for the partition function λ_0 :

$$\lambda_0 = \lambda_0(\underline{\theta}) \quad \dots \dots \dots (3.18)$$

Then, using the fact that $g_i(x)$; $i=1, 2, \dots, n$ are by assumption functions of x alone and not of $\underline{\theta}$ and the results (3.17) and (3.18) the maximum entropy density has the form of $p(x, \underline{\theta})$ above and the theorem follows.

iv) Corollary

The specified functions $g_i(x)$; $i=1,2,\dots, n$ are such that for a given random sample $\underline{x} = (x_1, x_2, \dots, x_n)$ from this distribution $[\sum_j g_1(x_j), \sum_j g_2(x_j), \dots, \sum_j g_n(x_j)]$; $j= 1,2,\dots, n$, comprise a set of joint sufficient statistics which is minimal if none of them is redundant.

3.4) Applications:

In this section we give a brief survey of the most recent and important applications of entropy and Jaynes Principle of Maximum Entropy to various subjects. Particularly in the statistical context, although not yet completely organised as a statistical method, the cited principle has proved to be of great help in many situations, mainly in Bayesian Statistics, where it provides a constructive criterion for setting up prior probabilities distributions on the basis of partial knowledge where conventional methods do not apply.

If it had been our aim to describe a complete survey of these applications we would have to start by giving an extensive list of its various uses in the fields of Communication Theory and later in Statistical Mechanics. We however interpret these subjects as the *Entropy Parents* and as such we are only concerned with the use of entropy in other fields.

i) Mathematical Ecology:

In the subject of Ecology Shannon's entropy has provided an entirely different way of measuring diversity in populations, assumed to contain an indefinitely large number of individuals that could be classified into a finite number of species.

Assuming also that each individual belongs to one and only one class and that p_i is the probability of an individual being in the species group C_i , $i=1,2,\dots, n$; then $H(p_1, p_2, \dots, p_n)$ provides a measure of the diversity of the population [Pielov, 1966, 1967 and 1969; Brown & Disk, 1975] .

ii) Reliability Studies:

In reliability studies of equipment which is maintained over a long period of time through replacement of components, the lifetime behavior associated with these models ranges from complete determinacy to complete uncertainty. The associated probability of survival, hazard and number of replacements can be obtained by maximising the entropy associated with the randomness [Tribus, 1962 ; Flehinger & Lewis, 1959].

iii) Thermodynamics:

Using entropy it is possible to show that the general maximum entropy formalism is intrinsically related to the experimentally measured quantities of a system in thermodynamic equilibrium. For instance, if H_e is the experimentally measured entropy of a system and H_s the corresponding Shannon's entropy then $H_s \leq H_e$, with equality if and only if the probability distribution in H_s is that one which gives maximum H_s . [Jaynes, 1963 a ; Tribus, 1961a, 1961b] .

iv) Statistical Inference:

The problem of decision making in the face of uncertainty can, by its very nature , be formulated and solved by using the notion of entropy as a criterion for setting up prior probability assignments.

Once the loss function has been specified, our uncertainty as to the best decision arises solely from our uncertainty as to the state of nature and so, the entropy. We refer mainly to : Jaynes, (1963b); Dutta(1966); Edwards.(1972); Vasicek,(1974) and Barnard,(1951).

v) Stock Market Prices:

A very general probability distribution of future stock price in a market can be obtained by use of Jaynes formalism. The maximum entropy distribution of future stock price for an investor having specified prior information is general and agrees with past observations of the market prices. [Mandelbrot & Taylor, 1967 ; Cozzolino & Zahner, 1973] .

vi) Econometrics:

In the field of Economics, Shannon's entropy has also been used a great deal. In Econometrics for instance, certain estimation methods such as least square, weighted regression, maximum likelihood are used and can be shown to be optimal in the Information Theoretical sense. We refer specially to: Tintner,(1960); Tintner & Sastry,(1969) and Theil,(1967).

vii) Model Identification-Time Series:

The application of entropy in the time series context is due to Akaike,(1971, 1972, 1974, 1977 a, 1977b, 1977c and 1978) and Tong,(1975a and 1975b). Akaike succeeded in deriving a 1-dimensional statistic for selecting an optimal model from a class of competing models by using

the generalized entropy of a distribution with respect to another (or the Kullback-Leibler mean information for discrimination between two distributions ; Kullback, 1969). Akaike's criterion, (also known as A.I.C. - Akaike's information criterion), is particularly important in estimating the order of auto regressive and/or moving average models.

3.5) Examples of Maximum Entropy Distributions

We conclude this chapter with some illustrative examples of maximum entropy distributions, obtained by the use of Jayne's formalism techniques developed in the previous sections.

$g_i(x); i=1, \dots, n$	$E \{g_i(x)\}$	X	$X \sim$
$g_1 = X$ $g_j = 0; j=2, \dots, n$	$E \{g_1\} = \lambda$	\mathbb{R}^+	Exponential (λ)
$g_1 = X$ $g_2 = \ln X$ $g_j = 0; j=3, \dots, n$	$E \{g_1\} = \alpha$ $E \{g_2\} = \beta$	\mathbb{R}^+	Gamma (α, β)
$g_1 = \ln X$ $g_2 = \ln(1-X)$ $g_j = 0; j=3, \dots, n$	$E \{g_1\} = \alpha$ $E \{g_2\} = \gamma$	$[0,1]$	Beta (α, γ)
$g_1 = X$ $g_2 = X^2$ $g_j = 0; j=3, \dots, n$	$E \{g_1\} = \mu$ $E \{g_2\} = \mu^2 + \sigma^2$	\mathbb{R}	Normal (μ, σ^2)
		\mathbb{R}^+	Single Truncated Normal (μ, σ^2)
		$[a,b]; a,b$ finite	Double Truncated Normal (μ, σ^2)

CHAPTER 4 : BAYESIAN ENTROPY FORECASTING (BEF)-
GENERAL MODEL FORMULATION

4.1) Historical Development of Time Series

Throughout this section we shall consider Khintchine's and Kolmogorov's interpretation of time series [Khintchine, 1932 ; Kolmogorov, 1933]. According to them, if we accept the broad view of a times series Y_t as a set of observations ordered sequentially in time, then, it is also possible to interpret it as:

- i) A stochastic process whose variables Y_1, Y_2, \dots, Y_n are observed at equispaced time intervals t_1, t_2, \dots, t_n .
- ii) An n-dimensional probability distribution Y_t . It is with that interpretation of time series in mind that we start our brief historical development of time series.

The first attempt towards an explanation of the functional form of a time series, dates from the very beginning of the last century. This was due to Joseph Fourier who claimed the approximation of any time series by a combination of sine and cosine curvers.

It was only at the beginning of this century that Fourier's idea was used again by Schuster, (1906). He succeeded in estimating periodicities in time series by introducing periodogram analysis. However, the limitations of use of the periodogram analysis [Beveridge, 1922], together with the great advances in probability theory and statistics experienced at the beginning of the twentieth century, provoked substantial developments in time series analysis. Starting in 1927

with Yule and complemented in 1938 by Wold [Yule, 1927 ; Wold, 1938; Walker, 1931 and Slutsky, 1937], the concepts of autoregressive and/or moving average (AR, MA, ARMA) schemes were introduced, which proved to be the most general linear representation for a stationary time series. Wold did not give much attention to the parametric estimation of this new scheme. The first methods for estimating the parameters of an AR, MA, ARMA model are due to Kolmogorov, (1941) and Man & Wold, (1943).

In order to follow our chronological description, it is worthwhile considering now the important work by Wiener in estimation theory. Around 1940 Wiener working in the field of communication theory, developed new techniques for filtering a signal at the receiver whose transmission has been distorted by a white noise process [Wiener, 1940]. In other words, if Y_t^* is a signal transmitted at time t and v_t is the random disturbance in the transmission of Y_t^* , Wiener assumed that the signal received is additively related to Y_t^* and v_t , i.e. :

$$Y_t = Y_t^* + v_t \quad \text{for all } t = 1, 2, \dots$$

where the v_t are assumed to be independent identically distributed Gaussian random variables, with $E\{v_t\} = 0$ and $E\{v_t^2\} = \sigma^2$. Wiener developed an estimation procedure for the white noise in the frequency domain for a continuous process so that an optimal filter was obtained (The analytical solution to the Wiener-Hopf integral equation). The discrete version of Wiener's work was independently developed by Kolmogorov by assuming that a stationary time series has a representation as above, thus the reconstruction of the real process Y_t^* could be obtained.

From that point, both Wold's autoregressive and/or moving average scheme in the time series context and Wiener's filter theory in the engineering context were developed a great deal, but it was only with the advent of computational facilities that a real boom occurred. The first major step forward was the work by Kalman and Bucy in 1960 [Kalman, 1960 and Kalman & Bucy, 1961] which proposed a solution to the Wiener-Hopf integral equation by transforming it into its equivalent differential equation, but working in the time domain. The recurrence relations and updating equations obtained - the *Kalman Filter*, as it is nowadays known could easily be solved by use of digital computers. Ever since Kalman, the new filter theory was developed and applied to different areas of engineering, particularly, in Control Theory [De Russo et al, 1967 ; Sage & Melsa, 1971 and Meditch, 1969].

Wold's scheme however, had its real great boom ten years later with the important work by Box and Jenkins [Box & Jenkins, 1970]. Box and Jenkins' contribution, undoubtedly has started a new era in time series and forecasting. Using the facilities of digital computers mentioned above, they proposed a new strategy for the construction of a set of linear stochastic equations, describing the behavior of a time series, whether stationary or not. Briefly, they assume that the given series Y_t can be reduced to stationarity by differencing a finite number of times, i.e. by determining the stationary series w_t by:

$$w_t = (1 - B)^d Y_t$$

where:

d is a positive integer.

B is a backward shift operator on the index of Y_t , such that:

$$BY_t = Y_{t-1}, B^2Y_t = Y_{t-2}, \text{ etc....}$$

It is then assumed that the stationary series w_t can be represented by an ARMA model of the form:

$$\left(1 - \sum_{i=1}^p \phi_i B^i\right) w_t = \left(1 - \sum_{j=1}^q \theta_j B^j\right) a_t$$

where:

ϕ_i are the autoregressive parameters ($i=1,2,\dots,p$)

θ_j are the moving average parameters ($j=1,2,\dots,q$)

a_t is a white noise sequence, with constant variance σ_a^2

or, in terms of Y_t :

$$\left(1 - \sum_{i=1}^p \phi_i B^i\right)(1-B)^d Y_t = \left(1 - \sum_{j=1}^q \theta_j B^j\right) a_t ;$$

known as an ARIMA (p,d,q) model.

Finally, the well known Box and Jenkins procedure to fit a model of the above form to a given set of data, consists of a three-steps iterative cycle procedure: identification (p,d,q values), estimation (ϕ_i, θ_j and σ_a^2), diagnostic checking (validity of the identified model) and then the forecasting stage. A lot of applications and further developments of the method have been extensively published. We only refer to some of them. [Makridakis, 1974; Gilchrist, 1976; Souza, 1974; D'Araujo, 1974; Brubacher, 1976 and Cleveland, 1972].

Almost at the same time as Box and Jenkins, a new and important approach for forecasting was put forward by Harrison and Stevens [Harrison & Stevens 1971, 1976a and 1976b]. They were in fact pioneers of the use of the Kalman filter results in a time series forecasting context. The so-called *Bayesian Forecasting System* or *Adaptive Forecasting* based on a joint use of Kalman results and Bayesian Statistics, offered a great improvement over the existing methods. Instead of considering a simple fit to past data in order to predict the future in a purely automatic way, they are mainly concerned in their method with the actual present information and its effects on the future. Since our model formulation is an extension of the above cited method, we dedicate the next section to a brief summary of Harrison and Steven's method, as well as the justification of our proposed extension.

We conclude this section by mentioning the recent *State Space Forecasting* proposed by Mehra,(1976, 1977a, 1977b, 1977c). He used only the Kalman Filter results for forecasting single and/or multiple time series, in other words using only the past data in order to get the model identification and the parametric estimation in a very automatic way. Although the method is very general and easy to use, it has the great disadvantage that the past history of the process is an essential requirement due to its non-Bayesian nature.

4.2) Bayesian Forecasting

In this section we give a brief description of the Kalman Filter-Bayesian approach for forecasting as proposed by Harrison and Stevens, pointing out the main advantages accruing to this new approach.

The model formulation is based on a complete parametric description of the process, which is incorporated into a dynamic linear set of equations describing:

- i) process observation
- ii) parameter evolution

In its general form, the Dynamic Linear Model (DLM) is:

$$\text{Observation equation: } Y_t = F_t \theta_t + v_t \quad \text{--- (4.1)}$$

$$\text{Parameter evolution equation: } \theta_t = G \theta_{t-1} + w_t \quad \text{--- (4.2)}$$

where:

Y_t is an $(m \times 1)$ vector of observations

θ_t is an $(n \times 1)$ vector of unknown parameters

F_t is an $(m \times n)$ matrix of independent variable (known at time t)

G is an $(n \times n)$ system matrix

v_t is an $(m \times 1)$ vector representing the observation noise;
 $v_t \sim N(0, V_t)$

w_t is an $(n \times 1)$ vector representing the parameter noise;
 $w_t \sim N(0, W_t)$

t is the time index ($t=1,2,\dots$)

The parameters are easily updated from time to time by use of the Kalman Filter updating equations, in other words, if:

$$(\theta_{t-1} | D_{t-1}) \sim N(m_{t-1}; C_{t-1}), \quad D_{t-1} = (y_1, y_2, \dots, y_{t-1})$$

then, once we observe $Y_t=y_t$, the parameter distribution at time t is:

$$(\theta_t | D_t) \sim N(m_t, C_t) ;$$

where m_t and C_t are obtained by use of the Kalman Filter recurrence equations as follows:

$$m_t = G \cdot m_{t-1} + Ae$$

$$C_t = R - A \hat{Y} A^T$$

where:

$$e = y_t - \hat{y}$$

$$\hat{y} = F_t G m_{t-1}$$

$$R = G C_{t-1} G^T + W_t$$

$$A = R F_t^T (\hat{Y})^{-1}$$

$$\hat{Y} = F_t R F_t^T + V_t$$

See Harrison & Stevens, (1976) for details.

The DLM formulation (4.1) and (4.2) offers something quite different from the conventional linear forecasting models. In fact, nearly all linear forecasting models can be framed in the DLM form. It is basically characterised by:

- i) Easy interpretation and easy model construction.
- ii) Its parametric formulation as opposed to the functional form of nearly all the models.
- iii) Its probabilistic information on the parameters at any time-
- iv) A sequential model formulation that permits a description of the systematic changes in the parameters of a system.
- v) A mixed model formulation to cope with sudden model changes or even uncertainty as to the underlying model at any given time.

To conclude, it is worth pointing out that by its very nature, the DLM (4.1) and (4.2) has the important properties that, the stationarity of the underlying process is not required and that its distributional predictive nature, allows us to have a different criterion of optimality other than the mean square errors.

4.3 Bayesian Forecasting Limitations and Proposed Extension

Although the Bayesian Forecasting method described in the last section has provided a simple and elegant model formulation, it has not fully extended the traditional forecasting system. It has still limitations, such as:

- i) The models are still linear in the sense that, the observation noise and parameter disturbance are additively related to the observation and parameter equations respectively.
- ii) From the linear least squares property of the Kalman filter, it is efficient only for a normal process.

In fact i) and ii) are closely related since the normality assumptions do not merely affect the distributions involved. They are also key concepts for the sufficiency and linearity of the Kalman Filter.

The restrictions i) and ii) are our main motivations towards an extension of the Bayesian Forecasting method. It is our prime objective in this extension, to set up a forecasting model whose efficiency is achieved for distributions other than the normal.

It is in fact in the system equation (4.4) that our problem lies. At first, it seems impossible to get hold of the prior at any time given the last posterior, in the absence of (4.4). For the normal additive model above we know that the transition from the parameter posterior at time t ; $(\theta_t | D_t)$ to the parameter prior at time $t+1$; $(\theta_{t+1} | D_t)$, is nicely obtained by straight use of (4.4). However, without the linear relationship between the parameter and the error component (4.4), such transition can not be easily obtained.

Denoting $(\theta_t | D_t) \rightarrow (\theta_{t+1} | D_t)$ the *Posterior-Prior Transition*, our problem can be summarized as finding this transition without using an additive formulation like (4.4). Although we have illustrated this problem with the Normal DLM formulation, it is quite clear that this *Posterior-Prior Transition* problem is general, i.e., provided we have a parametric model formulation, whatever conditional distribution is assumed for the observation $(Y_t | \theta_t)$, the $(\theta_t | D_t) \rightarrow (\theta_{t+1} | D_t)$ problem will be present.

4.4.2) Uncertainty Function

The problem just described can be tackled by the use of an entropy argument. However, the straight forward use of Shannon's entropy as a measure of uncertainty would not be recommended (this was pointed out in chapter 2 with reference to a continuous distribution). Referring to section 2.4, we can see that if $X \sim N(\mu, \sigma^2)$, then $H_X \propto \ln \sigma$ and consequently $H_X \in \mathbb{R}$. In fact, as we shall see later, for all the continuous distributions included in this work, we have $H_X \in \mathbb{R}$.

In order to avoid a negative measure of uncertainty we define a transformation on H_X such that the new measure is entirely defined on \mathbb{R}^+ .

Moreover, such measure should be a monotonic increasing function of the amount of uncertainty of the distribution of X (in the normal case, the variance), assuming a zero value in the total absence of uncertainty (where the distribution is concentrated at a point) and assuming a maximum value for the maximum uncertainty distribution. We shall denote this positive measure of uncertainty as S_X throughout.

Definition:

The *e-transform function* S_X is the positive measure of uncertainty defined by:

$$S_X = \exp [H_X] ; H_X \text{ Shannon's entropy of } X$$

As an example, if $X \sim N(\mu, \sigma^2)$, then:

$$H_X = \ln(\sqrt{2\pi e} \cdot \sigma) \Rightarrow S_X = \sqrt{2\pi e} \cdot \sigma$$

$$S_X \in \mathbb{R}^+ \quad \text{and:} \quad \begin{array}{l} S_X = 0 \quad \text{distribution concentrated at a point} \\ S_X \rightarrow \infty \quad \text{maximum uncertainty distribution monotonic} \\ S_X : \quad \text{increasing function of } \sigma . \end{array}$$

Not only is S_X entirely defined on \mathbb{R}^+ as we have just seen, but this function possesses a one-to-one relationship with the *predictability per observation* of a probability distribution, as we show below:

Let $X \in \Omega$; $\Omega = \{1, 2, \dots, N\}$ be a discrete r.v. with probability distribution $p_i = p(X=i)$; $i=1, 2, \dots, N$.

If x_1, \dots, x_n is a set of independent observations of X , it is then clear that the *predictability* of this sample is measured by its corresponding likelihood, i.e., we define:

$$\text{Pred.} = \prod_{i=1}^n p_i ; \quad \text{where Pred. stands for the predictability of } x_1, x_2, \dots, x_n .$$

From the above, the *predictability per observation* or the average predictability can be defined as the geometric mean of the sample predictability:

$$\text{Pred./Obs.} = \sqrt[n]{\prod_{i=1}^n p_i}$$

Or, assuming that for the N possible sample values the observation $x_i = i ; i=1,2,\dots, N$ occurs n_i times, where $\sum_{i=1}^N n_i = n$ (sample size), we have:

$$\text{Pred./Obs.} = \sqrt[n]{\prod_{i=1}^N p_i^{n_i}} = \prod_{i=1}^N p_i^{f_i} ; f_i = n_i/n$$

From the above, it is clear that if H_X is the Shannon's entropy of X , then:

$$\lim_{n \rightarrow \infty} \text{Pred./Obs.} = \exp \left[- H_X \right] = S_X^{-1}$$

Alternatively, the S_X function is a measure of the uncertainty per observation in a probability distribution. Recall that since $H_X = - \sum_{i=1}^N p_i \ln p_i$ then:

$$S_X = \exp \left[H_X \right] = \prod_{i=1}^N p_i^{-1}$$

From what we have seen it is quite clear that S_X possesses all the desirable interpretive properties of a measure of uncertainty in the formulation of a forecasting procedure.

4.5) Bayesian Entropy Forecasting System.

We now describe in detail our *Bayesian Entropy Forecasting Model* (BEF) proposed in the previous sections. We shall first give an outline of the model foundations and general assumptions, and then proceed with its analytical description.

4.5.1) Model Foundations

As already mentioned, the model we are proposing is an extension towards a generalization of the Harrison and Stevens Bayesian Forecasting system. We would like to start by remarking that we are also putting forward a *Statistical Forecasting System*, as opposed to a *Statistical Forecasting Method*. The simple reason for calling our approach a system, instead of a method, is that we are not simply producing the *best fit* on a given set of past data and then use this fitted curve to get an account of the future behaviour of the process. We are in fact proposing a forecasting system that not only takes into account the past history as the unique source of information, but also includes in the model building, qualitative or subjective information that is provided by the people involved with the system being modelled. As remarked by Harrison and Stevens (1976a), these people often have information quite beyond the mere past data history, that once incorporated into a model, would produce a more realistic forecasting system, responding quickly to major changes in the process and remaining stable during *quiet periods*.

The basic characteristics or foundations of the BEF system are:

- i) Parametric Structural Representation, allowing a simple model construction, as well as facilitating the communication between the forecaster and the method itself.

- ii) Probabilistic Parameter Description. This means that we have a random variable for the unknown parameter of the system whose distribution is inferred from the data and other information available at each time-point.
- iii) Sequential Model Description. By that we mean the flexibility of our model in offering at any time an updated parameter distribution, by incorporating into the least prejudiced prior, the information contained in the observed data.
- iv) Model Uncertainty. Instead of being concerned only with the uncertainty on the parameters of the model itself, our model formulation also offers us alternatives in order to select an appropriate model (or models) at each time, i.e., the uncertainty as to the model itself is also considered. Following Harrison and Stevens (1976a) classification, we could either be faced with:
 - Multi-Process Models Class I: where, out of a discrete set of model alternatives, a unique unknown model from this set obtains at all time.
 - Multi-Process Models Class II: where, at any given time, the model representing the underlying process is a random choice from a set of discrete alternative models.
- v) Non-Linear General Formulation. This is in fact the first generalization introduced by our BEF over the DLM Bayesian forecasting. As we shall see later, we substitute the observation and parameter additive equations of the DLM formulation

by a distributional specification, and a non-linear version of the normal model is obtained. Apart from that, such a broad model definition offers no difficulty for a non-normal generalization.

- vi) Valid for a Broad Class of Distributions. This is due to the use of entropy function as a measure of uncertainty in a probability distribution. Since entropy is a general measure of uncertainty for any distribution, any model definition based on it, can achieve maximum efficiency for distributions other than the normal.

4.5.2) General Assumptions.

With the considerations of the previous sections, we are now ready to describe our BEF system. Although the model we are putting is general, we are mainly concerned in this thesis with the steady state BEF model. We start by stating the two basic assumptions on which our model is based:

i) Information Loss:

The information (in Shannon's sense; the amount of uncertainty), decays with time. The greater the current information the greater the decay.

ii) Parametric Family of Distribution:

The form of the probability distribution (beliefs) about a future state of the process, belongs to a parameterised family of distributions whose e-transform uncertainty function S_e exists and is such that; $S_e = \exp(\Pi_e)$; Π_e where Π_e is the Shannon's entropy for the family.

4.5.3) System Evolution.

Before we present the formulation of our model, in this section we explore in detail the general assumptions (i) and (ii) of section 4.5.2. We shall see that by assuming an *Information decay* as in (i), the system evolution can be completely specified in terms of the parameter uncertainty function S_t ; provided the conditions established in (ii) are satisfied.

Let $Y_t \in Y$ and $\theta_t \in \Omega$ be the two r.v.'s representing respectively the process observation and the process parameter of a steady state model, where t is the time index; $t=1,2,\dots$.

Assume also that the conditional pdf of $(Y_t|\theta_t)$ is known for all $t=1,2,\dots$, and that the parameter posterior at time t ; $(\theta_t|D_t)$ has been obtained, where $D_t = (y_1, y_2, \dots, y_t)$. If $(\theta_{t+1}|D_t)$ represents the prior at time $t+1$ our task is to specify completely the pdf of $(\theta_{t+1}|D_t)$ on the basis of the available information, for all $t=1,2,\dots$. In other words, we want to establish a functional form for the parameter evolution i.e., the *posterior-prior transition* $(\theta_t|D_t) \rightarrow (\theta_{t+1}|D_t)$ mentioned before.

On the assumption that the process parameter belongs to the family of distributions (ii) of section 4.5.2, let:

$p_{t,t}$: represents the posterior parameter pdf of $(\theta_t|D_t)$ and
 $S_{t,t}$ its associated uncertainty (both known at time t).

$p_{t+1,t}$: represents the prior parameter pdf of $(\theta_{t+1}|D_t)$ and
 $S_{t+1,t}$ its associated uncertainty (both unknown at time t).

From the fundamental assumptions of section 4.5.2, it is quite obvious that the next prior level of uncertainty; $S_{t+1,t}$ is always greater than our present level of uncertainty; $S_{t,t}$ (for $S_{t,t}$ finite), and that this increase in the system uncertainty ($S_{t+1,t} - S_{t,t}$), naturally depends on the current value for $S_{t,t}$.

In terms of the pdf's involved, this implies that $p_{t+1,t}$ depends directly on $S_{t,t}$ and $p_{t,t}$, i.e., $p_{t+1,t} = \psi(p_{t,t}; S_{t,t})$. We show next that by elaborating the idea of information decay of section 4.5.2, we can establish a functional form for $\psi(.,.)$.

Without loss of generality, let us assume for the moment (for the sake of illustration) that the system parameter θ_t is a discrete r.v.; $\theta_t \in [\theta_1, \theta_2, \dots, \theta_n]$ for all $t=1, 2, \dots$

Furthermore, let us also assume that the posterior at time t , i.e., $p_{t,t}$ may be represented by:

$$p_{t,t} = \{p_{t,i} ; p_{t,i} = \text{Prob.}(\theta_t = \theta_i) ; i=1, 2, \dots, n\}$$

If we denote the unknown prior at time t ; $p_{t+1,t}$ in a similar way, i.e.:

$$p_{t+1,t} = \{p_{t+1,i} ; p_{t+1,i} = \text{Prob}(\theta_{t+1} = \theta_i); i=1, 2, \dots, n\};$$

The information decay assumption could be equivalently stated as:

The greater $p_{t,i} ; i=1, 2, \dots, n$ is from its average, the faster it declines .

Clearly the message in the above statement is that: if the information (or predictability) of the posterior distribution of the parameter at time t is high, then we expect a decrease in information (equivalently, an increase in the uncertainty) of the parameter distribution as we move ahead into the future, until the maximum level of uncertainty (uniform distribution; $p_{.,i} = p$; $i=1,2,\dots, n$) is reached as illustrated in Figure 4.1 .

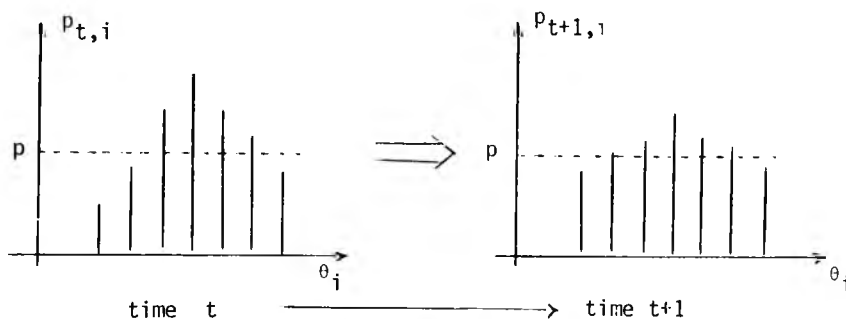


Figure 4.1 : Illustration of $p_{t,i} \rightarrow p_{t+1,i}$; for $\theta_i = \theta_i$; $i=1,2,\dots,n$.

From what we have seen, it is quite clear that given the last posterior level of uncertainty $S_{t,t}$, there exists a mapping $S_{t,t} \in \mathbb{R}^+ \rightarrow [0,1]$, such that $p_{t+1,t}$ could be directly obtained from it by raising $p_{t,t}$ to a power, whose value is the realisation of the function corresponding to the above mapping.

It is also clear from the assumption that such a function is an increasing function of $S_{t,t} \in \mathbb{R}^+$.

The argument as detailed above for the discrete case is clearly reproducible for the continuous case and, consequently, the $(\theta_t | D_t) \rightarrow (\theta_{t+1} | D_t)$ transition for the steady state model could be formally written as:

$$p_{t+1,t} \propto p_{t,t}^{h(S_{t,t})} \quad - \quad - \quad - \quad - \quad - \quad (4.6)$$

Definition

The *posterior-prior transition function* $h(S_{t,t})$ is defined as:
 (see illustration in figure 4.2) $h(S_{t,t}): \mathbb{R}^+ \rightarrow [0,1]$, and has the properties:

- i) Monotonic increasing function of the actual uncertainty.
- ii) $\lim_{S_{t,t} \rightarrow \infty} h(S_{t,t}) = 1$
- iii) $h(S_{t,t}=0) = 0$

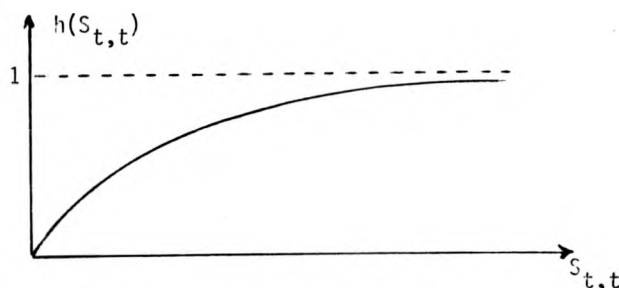


Figure 4.2: Illustrative plot of $h(S_{t,t})$ against $S_{t,t}$

If we happened to know $h(S_{t,t})$ or even an approximation to it then, the only problem left would be the case when we have no uncertainty at time t ($S_{t,t}=0$). In this particular case, the prior $p_{t+1,t}$ can not be obtained from (4.6). However, from the same information decay property of the system, it is intuitive that the assumptions of section 4.5.2, when interpreted in terms of the information (or uncertainty) contents of a distribution (e.g., S), could be restated as:

The greater the information (or, the less the uncertainty) of the distribution, the faster it declines (or, its uncertainty increases).

The above, interpreted in terms of S , is as follows:

If $S_{t,t}$ is close to zero (i.e., $p_{t,t}$ is highly predictable) then,

the increase in the system uncertainty; $(S_{t+1,t} - S_{t,t})$ is higher than the corresponding increase for bigger $S_{t,t}$. It is also true that for the two extremes ($S_{t,t}=0$ or $S_{t,t} \rightarrow \infty$), we should have a maximum value, c^* say, for $(S_{t+1,t} - S_{t,t})$ for any $t=1,2,\dots$ and $(S_{t+1,t} - S_{t,t}) \rightarrow 0$, respectively.

Although we do not know the exact evolutionary form of the system uncertainty function $S_{t,t} \rightarrow S_{t+1,t}$, from the information decay assumption of the model, we can formalise some of its properties:

- i) $S_{t+1,t}$ is a monotonic increasing function of $S_{t,t}$
- ii) $\lim_{S_{t,t} \rightarrow \infty} \frac{S_{t+1,t}}{S_{t,t}} = 1$
- iii) $\lim_{S_{t,t} \rightarrow 0} S_{t+1,t} = c^*$; where c^* is a positive constant.

We are now left with the problem of finding a functional specification for $S_{t+1,t}(S_{t,t})$.

As we have already mentioned, the exact form of this function is unknown; all we can say is that $S_{t+1,t}(S_{t,t})$ possesses the properties (i) to (iii) above. Moreover, this function is obviously related to the posterior-prior transition function: $h(S_{t,t})$, since both give an account of the system parameter evolution in time. In view of this evidence we assume that the *uncertainty ratio function* of a steady state model $(S_{t+1,t}/S_{t,t})$ is related to $h(S_{t,t})$ by a function of the form:

$$S_{t+1,t}/S_{t,t} = 1 / [h(S_{t,t})]^K \quad - \quad - \quad - \quad - \quad (4.7)$$

where K is a real constant.

From the definition of $h(S_{t,t})$, it is clear that properties (i) and (ii) of $S_{t+1,t}(S_{t,t})$ are trivially satisfied, and by a suitable choice of K we can make $\lim_{S_{t,t} \rightarrow 0} S_{t+1,t} = c^*$; c^* a positive constant.

In our model we shall adopt $K = \frac{1}{2}$ in equation (4.7). As we will show later, such a value for K matches exactly the posterior-prior transition of the normal additive model. In figure 4.3 we illustrate this uncertainty evolution function for a particular c^* .

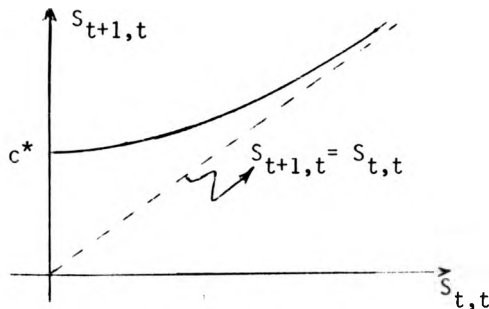


Figure 4.3: Illustrative plot of $S_{t+1,t}(S_{t,t}) \times S_{t,t}$ for a particular c^* .

4.5.4) Exponential Approximation.

From what we have shown in the previous section, the knowledge of the function $h(S_{t,t})$ at all time-points $t=1,2,\dots$ would enable us to obtain the transition $(\theta_t | D_t) \rightarrow (\theta_{t+1} | D_t)$ exactly. On the other hand, given the knowledge of properties (i) to (iii) of $h(S_{t,t})$, it seems quite obvious that we could set an exponential function to approximate the original function satisfying all the required properties.

Let $g(S_{t,t})$ denote such a function:

Theorem 1:

The function $g(S_{t,t}) = [1 - \exp(-c S_{t,t})]^2$; where c is a positive real constant, satisfies all the properties required to represent the posterior-prior transition function for the steady state model.

Proof:

First of all, $g(S_{t,t}) \in [0,1]$ for $c, S_{t,t} \in \mathbb{R}^+$

Also:

(i) $g(S_{t,t})$ is a monotonic increasing function of the actual uncertainty $S_{t,t}$.

(ii) $\lim_{S_{t,t} \rightarrow \infty} g(S_{t,t}) = \lim_{S_{t,t} \rightarrow \infty} [1 - e^{-c S_{t,t}}]^2 = 1$

(iii) $g(S_{t,t}=0) = 0$

Now, using (4.7) with $K=1/2$ and the fact that $g(S_{t,t})$ is an approximation to $h(S_{t,t})$, we can write:

$$S_{t+1,t} \approx \frac{S_{t,t}}{\sqrt{g(S_{t,t})}} = \frac{S_{t,t}}{[1 - \exp(-c S_{t,t})]} \quad - \quad - \quad - \quad (4.8)$$

and consequently:

Theorem 2 :

The uncertainty evolution function (4.8), with $g(S_{t,t})$ as defined in the theorem 1, has the same properties as the corresponding theoretical uncertainty evolution function as defined in section 4.5.3 .

Proof:

- (i) From (4.8), the first derivative of $S_{t+1,t}$ with respect to $S_{t,t}$ is given by:

$$\frac{\partial S_{t+1,t}}{\partial S_{t,t}} = \frac{a}{\partial S_{t,t}} \left[\frac{S_{t,t}}{\sqrt{g(S_{t,t})}} \right] = \frac{1 - e^{-cS_{t,t}}(1 - cS_{t,t})}{[1 - e^{-cS_{t,t}}]^2}$$

Since $e^{-cS_{t,t}} < 1$ for all $c, S_{t,t} \in \mathbb{R}^+$ we have:

$$\text{for } c \cdot S_{t,t} > 1 : \left[\frac{1 - c S_{t,t}}{e^{cS_{t,t}}} \right] < 0 \Rightarrow \frac{\partial S_{t+1,t}}{\partial S_{t,t}} > 0$$

$$\text{for } 0 < c S_{t,t} : \left| \frac{1 - c S_{t,t}}{e^{cS_{t,t}}} \right| < 1 \Rightarrow \frac{\partial S_{t+1,t}}{\partial S_{t,t}} > 0$$

Consequently, $S_{t+1,t}$ is an increasing function of $S_{t,t}$.

$$(ii) \lim_{S_{t,t} \rightarrow \infty} \frac{S_{t+1,t}}{S_{t,t}} = \lim_{S_{t,t} \rightarrow \infty} \frac{1}{[1 - \exp(-c S_{t,t})]} = 1$$

$$(iii) \lim_{S_{t,t} \rightarrow 0} S_{t+1,t} = \lim_{S_{t,t} \rightarrow 0} \frac{S_{t,t}}{[1 - e^{-cS_{t,t}}]} = \frac{1}{c} = \text{Constant} > 0 \quad (c \neq 0).$$

4.5.5) Model Formulation.

Let:

Y_t be the random variable defined on a sample space Ψ (process observation).

θ_t be the random variable defined on a parameter space Ω (process parameter) .

t be the time index ; $t=1,2,\dots$

A) Information

Assume that at time $t-1$ the following information is available:

i) $p(Y_{t-1} | \theta_{t-1})$: the conditional pdf of the r.v. $(Y_{t-1} | \theta_{t-1})$ supposed to be known for all $t=1,2,\dots$

ii) $p_{t-1,t-1}$: the posterior pdf of the r.v. $(\theta_{t-1} | D_{t-1})$, $D_{t-1}=(y_1, y_2, \dots, y_{t-1})$, and its associated entropy $H_{t-1,t-1}$ $S_{t-1,t-1}$.

iii) Posterior-Prior Transition Function ;

$g(S_{i,i}) = [1 - \exp(-c S_{i,i})]^2$; $i=0,1,2,\dots$, where:
 c is a positive constant

$S_{i,i}$ is the positive measure of uncertainty of the posterior $(\theta_i | D_i)$; $S_{i,i} \in \mathbb{R}^+$.

B) Parameter Updating Procedure.

B.1) Prior Distribution: $(\theta_t | D_{t-1})$

The prior pdf for $(\theta_t | D_{t-1})$, i.e., $p_{t,t-1}$ is the distribution obtained through the transition function $g(S_{t-1,t-1})$ by the system equation:

$$p_{t,t-1} \propto p_{t-1,t-1}^{g(S_{t-1,t-1})} \quad \text{if } S_{t-1,t-1} > 0$$

and $p_{t,t-1}$ such that $S_{t,t-1} = c^{-1}$ if $S_{t-1,t-1} = 0$

B.2) Posterior Distribution: $(\theta_t | D_t)$

The posterior parameter pdf, i.e., $p_{t,t}$ is easily computed by the simple operation of Bayes rule:

$$p_{t,t} \propto p_{t,t-1} \cdot p(Y_t | \theta_t)$$

where: $p_{t,t-1}$ is known from B.1

$p(Y_t | \theta_t)$ is known by assumption A-ii for all $t=1,2,\dots$

and then simple relationships for updating the parameters after observing y_t are obtained. It is important to mention that the procedure as stated is very general, in the sense that no restriction is imposed for any distribution involved. The procedure is made rather elegant if

$p_{t,t-1}$ is a member of the conjugate family to the distribution for $(Y_t | \theta_t)$. Note however, that the entropy approach here means that even if the distributions are not conjugate, the updating procedure is extremely easy; the perhaps unwieldy posterior does not affect the future computations involved in the method.

C) Prediction:

With the posterior as obtained in B.2 above, the next step consists of the prediction of future values of the observation Y_{t+j} ; $j=1,2,\dots$ standing at time t , that is, given D_t . The steps are as follow:

C.1) Parameter Prediction Distribution $(\theta_{t+j} | D_t)$

The parameter predictive pdf for $(\theta_{t+j} | D_t)$ is the distribution obtained by a sequential use of the transition function, as shown below:

$$P_{t+j,t} \propto P_{t+j-1,t}^{g(S_{t+j-1,t})} ; \quad j=1,2,\dots$$

where: $S_{t+j-1,t}$ is the uncertainty of $(\theta_{t+j-1} | D_t)$

In words, we assume that the same function $g(\cdot)$, that controls the posterior-to-prior transition through the system equation (B.1), gives the parameter predictive distribution for time $t+j$, $j=1,2,\dots$, standing at time t . For that, we interpret the last prior $(\theta_{t+j-1} | D_t)$ as the posterior at time $t+j-1$, in order to get the next prior (time $t+j$). In order to make the above specification general, we should consider the possible but unlikely case in which $S_{t,t} = 0$, i.e., the distribution of $(\theta_t | D_t)$ is concentrated in a point. In this case the next predictive for $(\theta_{t+1} | D_t)$ is such that its uncertainty is constant, that is:

$$P_{t+1,t} \text{ is such that } S_{t+1,t} = c^{-1} \text{ if } S_{t,t} = 0$$

C.2) Observation Prediction Distribution $(Y_{t+j} | D_t)$

We obtain the desired forecast pdf for $(Y_{t+j} | D_t)$, i.e., $p(Y_{t+j} | D_t)$, directly by integrating out θ_{t+j} in the joint pdf of $(Y_{t+j}, \theta_{t+j} | D_t)$:

$$p(Y_{t+j} | D_t) = \int_{\Omega} p(Y_{t+j}, \theta_{t+j} | D_t) \cdot d\theta_{t+j}$$

where:

$$p(Y_{t+j}, \theta_{t+j} | D_t) = p(Y_{t+j} | \theta_{t+j}, D_t) \cdot p_{t+j,t}$$

and

$$p(Y_{t+j} | \theta_{t+j}, D_t) = p(Y_{t+j} | \theta_{t+j}) \text{ is known by assumption A-ii}$$

$$p_{t+j,t} \text{ is known from C.1}$$

4.6) BEF - Properties.

4.6.1) Normal Additive Model

The first property of the BEF model is that it includes as a particular case the steady state normal model of Harrison & Stevens (1976a). In fact, by defining the normal additive model in terms of the uncertainty function $S_{t,t}$, we obtain the exact functions $h(S_{t,t})$ and $S_{t+1,t}/S_{t,t}$ defined in section 4.5.3. In a sense, this important property backs up all the assumptions we made in order to define the general steady state model, such as, the choice $K=1/2$ in equation 4.7.

Referring to section 4.4.1, the $(\theta_t | D_t) \rightarrow (\theta_{t+1} | D_t)$ transition for the normal additive model is given by:

$$\text{If: } (\theta_t | D_t) \sim N(m_t, C_t), \quad - \quad - \quad - \quad - \quad - \quad - \quad (4.9)$$

$$\text{then: } (\theta_{t+1} | D_t) \sim N(m_t, C_t + W) \quad - \quad - \quad - \quad - \quad - \quad - \quad (4.10)$$

and also, the particular but important case:

$$(\theta_t | D_t) \sim N(m_t, 0) \Rightarrow (\theta_{t+1} | D_t) \sim N(m_t, W) \quad - \quad - \quad - \quad (4.11)$$

The corresponding uncertainty values $S_{t,t}$ and $S_{t+1,t}$ are respectively (see section 4.4.2):

$$S_{t,t} = \sqrt{2\pi e C_t}$$

$$S_{t+1,t} = \sqrt{2\pi e (C_t + W)} = \sqrt{S_{t,t}^2 + W_k^2} \quad - \quad - \quad - \quad - \quad (4.12)$$

$$\text{where } W_k = \sqrt{2\pi e W}$$

From (4.12) we can clearly see that:

- (i) $S_{t+1,t}$ is a monotonic increasing function of $S_{t,t}$
- (ii) $\lim_{S_{t,t} \rightarrow \infty} \frac{S_{t+1,t}}{S_{t,t}} = 1$
- (iii) $S_{t+1,t} \Big|_{S_{t,t}=0} = W_k = \text{constant} > 0$

i.e., the function (4.12) satisfies all the properties of the uncertainty evolution function of section 4.5.3 .

Let us now study the $(\theta_t | D_t) \rightarrow (\theta_{t+1} | D_t)$ transition for the normal additive model in terms of the corresponding pdf's . Denoting:

$$p_{t,t} : \text{pdf} (\theta_t | D_t)$$

$$p_{t+1,t} : \text{pdf} (\theta_{t+1} | D_t) ,$$

we obtain from equations (4.9) & (4.10):

$$p_{t+1,t} \propto p_{t,t}^{\frac{C_t}{C_t+W}} ; C_t > 0$$

and, from 4.12:

$$p_{t+1,t} \propto p_{t,t}^{h(S_{t,t})}$$

where $h(S_{t,t}) = \frac{S_t^2}{S_t^2 + W_k^2} \text{ --- (4.13)}$

From (4.13) we can clearly see that $h(S_{t,t}) : S_{t,t} \in \mathbb{R}^+ \rightarrow [0,1]$, satisfies all the required properties of the posterior prior transition function of the steady state model introduced in 4.5.3 .

Finally, from (4.12) we can write:

$$\frac{S_{t+1,t}^2}{S_{t,t}^2} = \frac{S_{t,t}^2 + W_k^2}{S_{t,t}^2}$$

and consequently, from (4.13) we obtain:

$$S_{t+1,t} = \frac{S_{t,t}}{\sqrt{h(S_{t,t})}} \quad - \quad - \quad - \quad - \quad - \quad (4.14)$$

If we take the limit as $S_{t,t}$ goes to zero we obtain:

$$\lim_{S_{t,t} \rightarrow 0} S_{t+1,t} = \lim_{S_{t,t} \rightarrow 0} \frac{S_{t,t}}{\sqrt{h(S_{t,t})}} = W_k = \text{constant} > 0 \quad \text{for } W > 0$$

As we can see, the normal additive model defined in terms of $S_{t,t}$, exhibits all the assumed properties of our BEF steady state formulation. Moreover, the exact functions we obtained here perfectly match the theoretical assumptions of section 4.5.3 .

4.6.2) Non-Additive Normal Model

Let us now consider the BEF model as formulated in 4.5.5 applied to normal observations as shown below:

Observation Equation:

$$(Y_t | \theta_t) \sim N(\theta_t, V) \quad t=1,2,\dots$$

System Equation:

$$p_{t+1,t} \propto p_{t,t} g(S_{t,t})$$

where:

$$p_{t,t} : \text{pdf of } (\theta_t | D_t)$$

$$p_{t+1,t} : \text{pdf of } (\theta_{t+1} | D_t)$$

$$g(S_{t,t}) ; S_{t,t} \text{ as defined before.}$$

From the above set up we obtain for the posterior prior transition:

$$\text{given that: } (\theta_t | D_t) \sim N(m_t, C_t)$$

$$\text{then: } (\theta_{t+1} | D_t) \sim N(m_{t+1}^*, C_{t+1}^*)$$

$$\text{where: } m_{t+1}^* = m_t$$

$$C_{t+1}^* = \begin{cases} C_t/g(S_{t,t}) & \text{if } C_t > 0 \text{ (or } S_{t,t} > 0) \\ 1/(2\pi e c^2) & \text{if } C_t = 0 \text{ (i.e., } S_{t+1,t} = 1/c) \end{cases}$$

From the above and the corresponding additive normal model, where the exact transition function $h(S_{t,t})$ is used in place of the approximation $g(S_{t,t})$, we can clearly see that the constant "c" of $g(S_{t,t}) = [1 - \exp(-c S_{t,t})]^2$ is the only parameter of the model that needs a specification before hand. In a sense, it functions as the noise variance W of the DLM formulation, since either "c" or "W" gives an account of the system's uncertainty variation.

To conclude, we show a simple numerical simulation, comparing the DLM with the entropy approach just described. Let the DLM with $W=10$ and $V=400$ ($V/W=40$) then, the limiting posterior variance [Harrison & Stevens, 1976a] and the corresponding S . value are:

$$C_\ell = \frac{W}{2} \left| \sqrt{1 + 4 \frac{V}{W}} - 1 \right| \sim 58.4 ; S_\ell = \sqrt{2 \cdot \pi \cdot e \cdot C_\ell} \sim 31.6$$

Choosing "c" of $g(S_{t,t})$ such that $g(31.6) = h(31.6)$, we obtain $c \sim 0.082$.

In table C.1 (appendix C) we show the values of $g(S_{t,t})$ against $h(S_{t,t})$ for $S_{t,t} \in [22.6 ; 39.2]$ or $C_t \in [30;90]$. It is clear that within this most likely range of variation for C_t , $g(S_{t,t})$ is responding satisfactorily to the true variation $h(S_{t,t})$. Values of C_t outside this range, though unlikely to occur, will be eventually brought into this interval, as a consequence of the limiting property of the steady state model.

In table C.2 (appendix C), we can see the comparison of the prior uncertainty for many values of the posterior uncertainty $S_{t,t}$ using (4.14) with $h(S_{t,t})$ and $g(S_{t,t})$ respectively.

Finally, in table C.3 (appendix C) the results of the maximum support estimator for the constant "c" are shown, using the data generated by the DLH model with $W=10$ and $V=400$.

The increasing sample size is to emphasize the convergence to the limiting value of c.

4.6.3) Parameter Prediction

The "l" steps ahead parameter prediction is sequentially obtained by:

$$P_{t+j,t} \propto P_{t+j-1,t}^{g(S_{t+j-1,t})} ; \quad j=1,2,\dots, l$$

where:

$p_{t,t}$ is the parameter posterior pdf at time t and $S_{t,t}$ its corresponding uncertainty.

$p_{t+j,t}$ is the parameter prior pdf at time $t+j$ and $S_{t+j,t}$ its corresponding uncertainty, $j=1,2,\dots,l$

In terms of the uncertainty functions, the above parameter prediction scheme is as illustrated below in figure 4.4 :

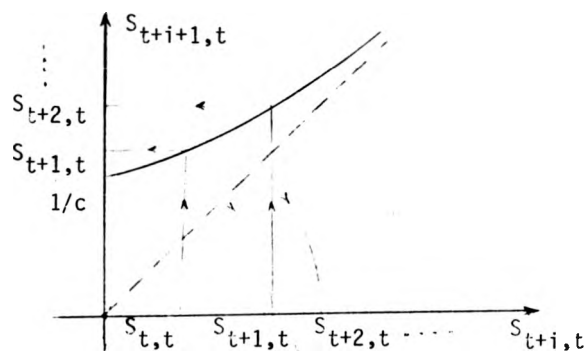


Figure 4.4: l -steps ahead parameter prediction scheme; $i=0,1,\dots,l-1$.

4.6.4) System Evolution

In the general model we just described, it was assumed that the parameter evolution (or system equation), was given by the posterior-prior pdf relationship:

$$p_{t+1,t} \propto p_{t,t}^{g(S_{t,t})}$$

In fact, this is the key concept in our model formulation and enabled us to formulate models for a broader class of distributions.

One of the motivations for the use of such a relationship as the system equation, comes from the normal model results. As we showed in section 4.6.1

the normal model formulated in terms of a positive measure of

uncertainty, leads automatically to this kind of parameter evolution (see equation 4.13 in special). The extension for distributions other than the normal seems quite reasonable if we consider the system evolution specified only in terms of its entropy. In other words we assume that, whatever distribution is attributed for θ_t , the process information prior depends only on the last posterior state of uncertainty and not on the distribution itself.

Provided the system parameter belongs to the family as specified in ii) of section 4.5.2, we then define a steady state model, as the system that admits a unique posterior-prior exponent transition function $h(S_{t,t}) : \mathbb{R}^+ \rightarrow [0,1]$, with the properties:

- i) Monotonic increasing function of $S_{t,t}$
- ii) $\lim_{S_{t,t} \rightarrow \infty} h(S_{t,t}) = 1$
- iii) $h(S_{t,t}=0) = 0$

Accepting the existence of this unique $h(S_{t,t})$ as a general function for the steady model, the results of section 4.6.2 for the normal model using the approximating function $g(S_{t,t})$ are obviously generalised to non-normal distributions. The approximation seems reasonable if we recall the limiting properties of a steady state model. We know very well that, given the nature of the steady model, the system uncertainty will always lie in a finite interval and within this interval a linear approximation could even be assumed.

As a matter of illustration suppose that for a generic steady model, $I_S = (S_{t_1, t_1} ; S_{t_2, t_2})$ is the most likely interval for $S_{t,t}$ to lie in, as shown

figure 4.5 . It is then obvious that by setting an approximation $g(S_{t,t})$ to $h(S_{t,t})$, we really want $g(S_{t,t})$ as close as possible to $h(S_{t,t})$ within I_S . In fact, we do not need to bother about the occurrence of $S_{t,t}$ outside I_S . Whether using the true function $h(S_{t,t})$ or the approximation $g(S_{t,t})$ they will be eventually brought into the interval, unless some permanent change has happened in the system pattern, in which case, there would be another most likely interval for $S_{t,t}$

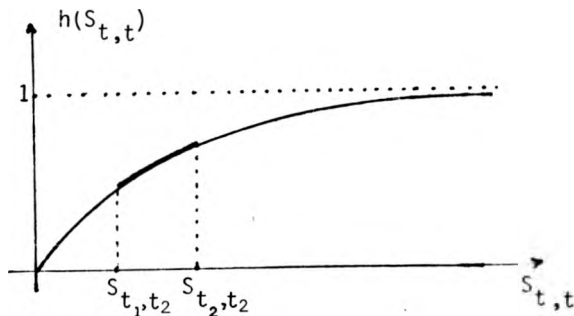


FIGURE 4.5 :
 $h(S_{t,t})$ and a generic most likely interval I_S .

4.6.5) Steady State Model-Definition

If we consider in our model formulation the parameter θ_t as representing the level of the process, we then have, according to Harrison and Stevens notation, a steady model. Assuming this particular model within our BEF framework, the following result can be obtained:

Theorem 3

If the parameter distribution is differentiable and unimodal, then

a steady state model is the one in which the mode remains constant in the posterior-to-prior transition.

Proof:

Let

$p_t(\theta)$ denote the posterior pdf at time t , i.e., $p_{t,t}$

$p_{t+1}^*(\theta)$ denote the prior pdf at time $t+1$ given D_t ,

i.e., $p_{t+1,t}$

m_t the mode of $p_t(\theta)$

m_{t+1}^* the mode of $p_{t+1}^*(\theta)$

Since m_t is the mode of $(\theta_t | D_t)$ and by assumption $p_t^*(\theta)$ is differentiable, we can write:

$$\left. \frac{\partial}{\partial \theta} p_t(\theta) \right|_{\theta = m_t} = 0$$

From the system equation (4.14) we can write for $p_{t+1}^*(\theta)$:

$$p_{t+1}^*(\theta) \propto [p_t(\theta)]^g$$

and, by differentiating with respect to θ :

$$\frac{\partial}{\partial \theta} p_{t+1}^*(\theta) \propto [p_t(\theta)]^{g-1} \frac{\partial}{\partial \theta} p_t(\theta)$$

For $\theta = m_t$, we get:

$$\left. \frac{\partial}{\partial \theta} p_{t+1}^*(\theta) \right|_{\theta = m_t} = 0 \quad \therefore \quad m_{t+1}^* = m_t$$

that is one of the most important differences between our BEF steady state model and other formulations for the same steady model. While in other models the mean is kept constant in the posterior-to-prior transition, in our method the mode remains constant.

A similar conclusion was obtained by Smith,(1978) by redefining the steady state model in a decision space. In doing so, he obtains an expression like (4.6) but with a constant in place of $g(S_{t,t})$ for all $t=1,2,\dots$. This seems to be a very strong assumption, in the sense that, he is forced to assume the *steady state* of the steady model from the very beginning.

4.6.6) Goodness of Fit-Relative Entropy Criterion.

In our model formulation, we adopt as our forecasting pdf the distribution for $(Y_{t+j}|D_t)$; $j=1,2,\dots$, obtained by integrating out the parameter in the joint observation-parameter distribution. By the use of an entropy argument, we show in this section the goodness of fit of this predictive distribution.

Let: $A = \{p(Y_{t+1} | \theta_{t+1}) ; \theta_{t+1} \in \Theta\}$ be a class of density functions for parameters models defined on a sample space Ψ and parametric space Θ , and

$D_t = \{y_1, y_2, \dots, y_t\}$ as defined before.

The goodness of fit problem could then be stated as:

"Fit a model for $p(Y_{t+1} | \theta_{t+1})$ on the basis of D_t and the fact that the true $\theta_{t+1} \in \Theta$ is unknown for all $t=0,1,\dots$."

It is clear that the possible fitting models to $p(Y_{t+1} | \theta_{t+1})$, are basically classified into the categories:

i) Estimative Density Function Class β_1 (EDF)

$$\beta_1 = \{ p_1(Y_{t+1} | D_t) = p[Y_{t+1} | \theta_{t+1} = \theta_{t+1}(D_t)] ; \beta_1 \cong A \}$$

where $\theta_{t+1}(D_t)$ is some efficient point estimate for θ_{t+1} based on D_t .

ii) Predictive Density Function Class β_2 (PDF)

$$\beta_2 = \{ p_2(Y_{t+1} | D_t) = \int_{\Theta} [p(Y_{t+1}, \theta_{t+1} | D_t) \cdot d\theta_{t+1}] ; \beta_2 \cong A \}$$

i.e., the predictive distribution as used in our model formulation (see section 4.5.5-C.2).

As we said at the beginning of this section, we use an entropy argument as the discrimination criterion between the two classes. In our present case we use the *Relative Entropy* or the *Discriminating Measure between two pdf's*, defined as:

If $p(x)$ is the true pdf of a continuous rv $X \in X$ (discrete case is similar), and $f(x)$ an approximation to $p(x)$, then, the entropy of $p(x)$ with respect to $f(x)$ is:

$$H[p, f] = - \int_X \ln \left[\frac{p(x)}{f(x)} \right] \cdot p(x) \cdot dx$$

It is clear that (refer back to chapter 2) $H[p, f]$ is an invariant, non-positive quantity ($H=0$ if $p=f$) and is a measure of overall closeness between $p(x)$ and $f(x)$. $-H[p, f]$ is the Kullback and Leibler direct measure of divergence. Consequently, the greater the relative entropy, the higher is the degree of approximation between $p(x)$ and $f(x)$. In this case, the maximisation of $H[p, f]$, or its expectation provides a criterion of goodness of fit of the pdf $f(x)$ as an approximation to $p(x)$.

For details of the properties and the use of this discriminating measure see, for instance : Akaike, (1977-b, 1977-c); Aitchison (1975) and Aitchison & Dunsmore (1975).

Theorem:

"The predictive distribution (PDF) is optimal in the sense of the relative entropy criterion".

Proof:

Let $q(Y_{t+1} | D_t)$ and $r(Y_{t+1} | D_t)$ be two contenders for the role of estimating $p(Y_{t+1} | \theta_{t+1})$.

Then, the measure of discrepancy between $q(Y_{t+1} | D_t)$ & $p(Y_{t+1} | \theta_{t+1})$ and $r(Y_{t+1} | D_t)$ & $p(Y_{t+1} | \theta_{t+1})$ is, respectively:

$$H_1 [p, q] = - \int_Y \ln \left[\frac{p(Y_{t+1} | \theta_{t+1})}{q(Y_{t+1} | D_t)} \right] \cdot p(Y_{t+1} | \theta_{t+1}) \cdot dY_{t+1}$$

$$H_2 [p, r] = - \int_Y \ln \left[\frac{p(Y_{t+1} | \theta_{t+1})}{r(Y_{t+1} | D_t)} \right] \cdot p(Y_{t+1} | \theta_{t+1}) \cdot dY_{t+1}$$

By the definition of H , we can say that:

" q is closer to p than r if:

$$H [p; q, r] = H_2 [p, r] - H_1 [p, q] = - \int_{\mathcal{Y}} \ln \left[\frac{q(Y_{t+1}|D_t)}{r(Y_{t+1}|D_t)} \right] \cdot p(Y_{t+1}|\theta_{t+1}) \cdot dY_{t+1}$$

is non positive".

The above measure depends on θ_{t+1} (and D_t , which is supposed to be known). On the other hand, given the knowledge of the prior pdf for $(\theta_{t+1}|D_t)$, the natural measure of relative closeness, would then be its expected value with respect to $p_{t+1,t}$, that is:

$$E_{(\theta_{t+1}|D_t)} \{ H [p; q, r] \} = \int_{\Theta} H [p; q, r] \cdot p_{t+1,t} \cdot d\theta_{t+1}$$

or, taking into account the expression for $H [p; q, r]$:

$$E_{(\theta_{t+1}|D_t)} \{ H [p; q, r] \} = - \int_{\Theta} p_{t+1,t} \int_{\mathcal{Y}} p(Y_{t+1}|\theta_{t+1}) \cdot \ln \left[\frac{q(Y_{t+1}|D_t)}{r(Y_{t+1}|D_t)} \right] \cdot dY_{t+1}$$

By changing the order of the integrals:

$$E_{(\theta_{t+1}|D_t)} \{ H [p; q, r] \} = - \int_{\mathcal{Y}} \left[\int_{\Theta} p(Y_{t+1}|\theta_{t+1}) \cdot p_{t+1,t} \cdot d\theta_{t+1} \right] \cdot \ln \left[\frac{q(Y_{t+1}|D_t)}{r(Y_{t+1}|D_t)} \right] \cdot dY_{t+1}$$

By the definition of H, we can say that:

"q is closer to p than r if:

$$H [p; q, r] = H_2 [p, r] - H_1 [p, q] = - \int_Y \ln \left[\frac{q(Y_{t+1}|D_t)}{r(Y_{t+1}|D_t)} \right] \cdot p(Y_{t+1}|\theta_{t+1}) \cdot dY_{t+1}$$

is non positive".

The above measure depends on θ_{t+1} (and D_t , which is supposed to be known). On the other hand, given the knowledge of the prior pdf for $(\theta_{t+1}|D_t)$, the natural measure of relative closeness, would then be its expected value with respect to $p_{t+1,t}$, that is:

$$E_{(\theta_{t+1}|D_t)} \{ H [p; q, r] \} = \int_{\Theta} H [p; q, r] \cdot p_{t+1,t} \cdot d\theta_{t+1}$$

or, taking into account the expression for $H [p; q, r]$:

$$E_{(\theta_{t+1}|D_t)} \{ H [p; q, r] \} = - \int_{\Theta} p_{t+1,t} \int_Y p(Y_{t+1}|\theta_{t+1}) \cdot \ln \left[\frac{q(Y_{t+1}|D_t)}{r(Y_{t+1}|D_t)} \right] \cdot dY_{t+1} \cdot d\theta_{t+1}$$

By changing the order of the integrals:

$$E_{(\theta_{t+1}|D_t)} \{ H [p; q, r] \} = - \int_Y \left[\int_{\Theta} p(Y_{t+1}|\theta_{t+1}) \cdot p_{t+1,t} \cdot d\theta_{t+1} \right] \cdot \ln \left[\frac{q(Y_{t+1}|D_t)}{r(Y_{t+1}|D_t)} \right] \cdot dY_{t+1}$$

But from (ii), the inner integral is the $p_2(Y_{t+1}|D_t)$ of the class β_2 . Consequently, the above can be written as:

$$E_{(\theta_{t+1}|D_t)} \{ H[p;q,r] \} = - \int_{\mathcal{Y}} p_2(Y_{t+1}|D_t) \cdot \ln \left[\frac{q(Y_{t+1}|D_t)}{r(Y_{t+1}|D_t)} \right] \cdot dY_{t+1} \quad \text{--- (4.15)}$$

By making $q(Y_{t+1}|D_t) = p_2(Y_{t+1}|D_t)$, the expression (4.15) becomes the relative entropy $H[p_2,r]$, which is by definition non-positive for all $r(Y_{t+1}|D_t)$ different from $p_2(Y_{t+1}|D_t)$ (unless $r=p_2$, when $H[p_2,r] = 0$), and in particular for $r(Y_{t+1}|D_t) = p_1(Y_{t+1}|D_t)$ of the class β_1 .

Consequently, the predictive distribution of our model formulation is unrivalled in its closeness to the true distribution $p(Y_{t+1}|\theta_{t+1})$.

4.6.7) Aggregate Likelihood for Estimation of "c"

According to our model formulation, the prior distribution for the parameter at any time depends only upon an unknown parameter "c", i.e., the constant that appears in the function $g(S.)$. In this section, we show how this constant can be estimated sequentially through the available data. We use mainly the idea of aggregate likelihood of a Bayesian model, suggested by Akaike (1977b) and adapted to our BEF models.

Let us start by assuming that our prior distribution belongs to a parameterized family G , where:

$$G = \{ q(\theta_{i+1}|c) = p(\theta_{i+1}|c, D_i) \quad \theta_{i+1} \in \Theta ; c \text{ unknown positive constant ; } i=1,2,\dots, t \}$$

Then, using the fact that $p(Y_t|\theta_t)$ is known for every $t=1,2,\dots$ by assumption, we could get $r(Y_i|c) = p(Y_i|c, D_{i-1})$ by straight integration, as shown below:

$$r(Y_i|c) = \int_{\Theta} p(Y_i|\theta_i) \cdot q(\theta_i|c) d\theta_i ; \quad i = 1, 2, \dots, t$$

If we now let:

$$r(D_t|c) = \prod_{i=1}^t r(Y_i|c) ,$$

we define $L(c) = \ln r(D_t|c)$ the Aggregate Likelihood of the Bayesian model, specified by the data distribution $p(Y_i|\theta_i)$ and prior $q(\theta_i|c)$. We can then obtain an estimate for c by maximising $L(c)$, i.e.:

$$\hat{c} = \max_{c \in \mathbb{R}^+} L(c)$$

As shown by Akaike, this estimate obtained by direct maximisation of the aggregate likelihood, will at least asymptotically, approximate the optimum choice within the parametric family G .

4.7) Sufficient Statistic Specification

We finish this chapter with an interesting alternative model formulation using mainly the material covered in chapter 3. If we concentrate only on the concept of sufficiency, we can reformulate our model by using the intrinsic relationship between the Maximum Entropy Distribution and sufficiency, described in the theorem and corollary at section 3.3 .

This straight link and the properties of the maximum entropy distribution, suggests to us that a general Bayesian formulation, applicable to distributions not only normal, is possible.

The general model formulation would be similar to what we have described in section 4.5.3, the only difference lying in the posterior-to-prior parameter specification.

Referring to the steady state linear model as our usual starting point towards a non-normal extension, instead of exploring the posterior prior exponent transition function and relating it to the posterior entropy, we should now examine the sufficient statistics specification for the parameter prior distribution. In other words, from equation (4.10) we see that:

$$E \{ \theta_{t+1} | D_t \} = m_t \quad \text{and} \quad E \{ \theta_{t+1}^2 | D_t \} = C_t + W + m_t^2$$

The above average equations, when put into the Jaynes' formalism, functioning as constraints, would result in a normal maximum entropy distribution for the prior. It is then clear that for the Kalman filter models, the distribution assumed for $(\theta_{t+1} | D_t)$; $t=0,1,\dots$ is the least prejudiced one, constrained on the given sufficient statistics. Put this way, there would be no need for the additive formulation of equations (4.3) and (4.4). Finally, we can achieve the desired non-normal extension, if we consider that the process parameter distribution is such that, the results of the theorem and corollary of section 3.3 are applicable. In this case we should have to change the general assumption (ii) of section 4.5.2, by constraining the process parameter to a parameterised

class of the exponential family. In doing so, we are able to use all the results of chapter 3 concerning Maximum Entropy and Sufficiency and the general model formulation would not differ from what we have described in section 4.5.5, apart from the prior parameter pdf obtained as follow:

Instead of A-(iii) of section 4.5.5, we should have the expected system evolution as a known information:

$$E \{g_i(\theta_t | D_{t-1})\} = G^{(i)} \left[(\theta_{t-1} | D_{t-1}); \phi^{(i)}(t-1, t) \right]; i=1,2,\dots,n$$

where $\phi^{(i)}(t-1, t)$ dictates the evolution of the system parameters from $t-1$ to t assumed known and g_i ; $i=1,2,\dots,n$ are the known functions, specifying the minimally sufficient statistics for the distribution of $(\theta_t | D_{t-1})$.

The prior pdf for $(\theta_t | D_{t-1})$, i.e., $p_{t,t-1}$ is then given by the Jaynes' principle as the least prejudiced distribution, obtained by maximizing the entropy $H_{t,t-1}$, subject to the constraints described above. For a detailed description of this general formulation, we refer to Souza & Harrison (1977), chapter 2.

As a final remark we would like to point out that, in using this formulation for distributions other than the normal, we are likely to come across difficulties in the implementation of the system parameter evolution functions $\phi^{(i)}(t-1, t)$. This is due to the difficult interpretation of some of the sufficient statistics of the parameter distribution related to the model itself. As shown in our previous work, we could

avoid this problem by specifying the evolution of functions related to the sufficient statistics which are easier to interpret. For instance, for the Poisson-Gamma model, instead of working with $\ln(\theta_t | D_{t-1})$ itself, we could formulate this evolution in terms of the coefficient of variation of $(\theta_t | D_{t-1})$, which is for a gamma distribution well defined in the interval $[0,1]$.

Another disadvantage of this formulation is related to the steady state model. In adopting the sufficient statistics formulation, we are forced to accept that the mean of the parameter distribution is held constant for the steady state model, whatever the distribution is. This seems for us quite strong, specially when dealing with skewed distributions. In such cases, the mode of the distribution seems more appropriate to be kept constant in the posterior-to-prior transition.

CHAPTER 5 :

STEADY STATE POISSON-GAMMA MODEL

5.1) Introduction:

In this chapter we apply our BEF formulation to the case where the process level $\theta_t \in \Theta$ is assumed to be a gamma distributed r.v. for $t=1,2,\dots$. For the process observation r.v. $Y_t \in \mathcal{Y}$ we assume the usual conjugate form, i.e., a Poisson distribution. As we have mentioned before, the use of this conjugate form is not compulsory for the method; we use it merely for the sake of simplicity and tractability of the posterior.

This model was first proposed in a recent paper by Leonard and Harrison, (1977). They use a Bayesian technique which enables them to extend the Harrison & Stevens method for Poisson observations. The first stage equation of the steady state DLM formulation (observation equation) is substituted by an assumption that the observations Y_1, Y_2, \dots are independent and Poisson distributed given their respective means $\theta_1, \theta_2, \dots$, and the second stage (system equation) remains the same i.e., $\theta_i = \theta_{i-1} + w_i$; $i=1,2,\dots,n$, for which the first two moments of the error term are required to be specified. A further extension of their method was proposed by Souza & Harrison, (1977) by the use of the least prejudiced assignment of pdf for the parameter evolution as opposed to the additive parameter equation assumed by Leonard & Harrison.

Finally, Smith, (1978) treats the same Poisson-Gamma process. As we have commented in section 4.6.5, Smith's formulation, although obtained through a decision theoretic argument has a similar updating system to ours.

However, as we shall see later, there is a fundamental difference between the BEF and Smith's model, related to the limiting properties of the steady state model.

This chapter deals with the theoretical description of the model and its applications to simulated and real data. The various tables containing the numerical results are shown in appendix D.

5.2) Entropy of the Gamma Variate

Before proceeding with the description of the model, a preliminary study concerning the parameter distribution is required. In fact, we need to show first that Shannon's entropy and the e-transform uncertainty function for a gamma variate satisfies the basic assumption (ii) of section 4.5.2 .

Let $X \in \mathbb{R}^+$ be a continuous r.v. gamma distributed with parameters α and β , i.e.:

$X \sim G(\alpha, \beta)$, where:

$X \in \mathbb{R}^+$ is a continuous r.v.

α is the shape parameter ($\alpha > 0$)

β is the scale parameter ($\beta > 0$)

Denoting the pdf of X by $f = f(X|\alpha, \beta)$

$$f = f(X|\alpha, \beta) = \beta^\alpha \cdot X^{\alpha-1} \cdot e^{-\beta X} / \Gamma(\alpha) \quad \text{--- (5.1)}$$

To obtain the expression for the entropy of X , we first write (5.1) in the equivalent form:

$$f = \exp \left[(\alpha-1) \ln x - \beta x + \ln (\beta^\alpha / \Gamma(\alpha)) \right]$$

From which we can write:

$$\theta_1 = \alpha ; \quad \theta_2 = \beta ; \quad K_1(x) = \ln x ; \quad K_2(x) = x$$

$$A_1(\theta_1) = \alpha - 1 ; \quad A_2(\theta_2) = -\beta ; \quad Q(\alpha, \beta) = \ln \left[\beta^\alpha / \Gamma(\alpha) \right]$$

$$\text{and } S(x) = 0.$$

Since $A_1(\theta_1)$ and $A_2(\theta_2)$ are differentiable, we can take the above functions into the results of appendix A, giving the following expression for the entropy of X :

$$H_X = \ln \Gamma(\alpha) + \alpha [1 - \psi(\alpha)] + \psi(\alpha) - \ln \beta \quad \text{--- (5.2)}$$

where $\Gamma(u) = \int_0^{\infty} t^{u-1} \cdot e^{-t} \cdot dt$ is the gamma function of $u > 0$

and $\psi(u) = \frac{d[\ln \Gamma(u)]}{du} = \frac{\Gamma'(u)}{\Gamma(u)}$ is the Digamma function of

$u > 0$. [Abramowitz & Stegun, 1965].

Now, to obtain the range of variation for H_X in (5.2), we first need to check the range of definition of α and β . From the considerations made in chapter 4, we assume in our model that the mode of the distribution exists. This means that $\alpha > 1$ since $\text{Mode}(X) = (\alpha-1)/\beta$. Also, since $\text{Var}(X) = \alpha/\beta^2$ and $\text{Coeff. Var.}(X) = 1/\sqrt{\alpha}$ it is clear that we have:

(i) For the maximum uncertainty distribution for X when $(\alpha \rightarrow 1)$ and $(\beta \rightarrow 0)$ and, from (5.2), $\lim_{\substack{\alpha \rightarrow 1 \\ \beta \rightarrow 0}} H_X = +\infty$

(ii) For the minimum uncertainty distribution for X when $\alpha, \beta \rightarrow +\infty$ and again, from (5.2):

$\lim_{\alpha, \beta \rightarrow +\infty} H_X = -\infty$, because [see Abramowitz & Stegun, 1965]:

$\lim_{\alpha \rightarrow +\infty} \{ \ln \Gamma(\alpha) + \alpha [1 - \Psi(\alpha)] \} = 0$ and $\lim_{\alpha \rightarrow +\infty} \Psi(\alpha) = \text{const.} (\sim 4)$

From (i) & (ii): $H_X \in \mathbb{R}$ and consequently the e-transform uncertainty function for X , satisfies the basic assumption (ii) of section 4.6.2 and is given by:

$$S_X = \exp \{ H_X \} = \exp \{ \ln \Gamma(\alpha) + \alpha [1 - \Psi(\alpha)] + \Psi(\alpha) - \ln \beta \} \quad \text{---(5.3)}$$

5.3) BEF Poisson-Gamma System; Model Description

With S_t as defined in (5.3), we are now ready to apply our BEF as described in the previous chapter to the Poisson-Gamma process.

Notation:

At any given time $t=1,2,\dots$ let:

Y_t be the process observation.

θ_t be the process parameter (unknown) ;

(i) For the maximum uncertainty distribution for X when $(\alpha \rightarrow 1)$ and $(\beta \rightarrow 0)$ and, from (5.2), $\lim_{\substack{\alpha \rightarrow 1 \\ \beta \rightarrow 0}} H_X = +\infty$

(ii) For the minimum uncertainty distribution for X when $\alpha, \beta \rightarrow +\infty$ and again, from (5.2):

$\lim_{\alpha, \beta \rightarrow +\infty} H_X = -\infty$, because [see Abramowitz & Stegun, 1965]:

$\lim_{\alpha \rightarrow +\infty} \{ \ln \Gamma(\alpha) + \alpha [1 - \Psi(\alpha)] \} = 0$ and $\lim_{\alpha \rightarrow +\infty} \Psi(\alpha) = \text{const.} (\sim 4)$

From (i) & (ii): $H_X \in \mathbb{R}$ and consequently the e-transform uncertainty function for X , satisfies the basic assumption (ii) of section 4.6.2 and is given by:

$$S_X = \exp \{ H_X \} = \exp \{ \ln \Gamma(\alpha) + \alpha [1 - \Psi(\alpha)] + \Psi(\alpha) - \ln \beta \} \quad \text{---(5.3)}$$

5.3) BEF Poisson-Gamma System; Model Description

With S_X as defined in (5.3), we are now ready to apply our BEF as described in the previous chapter to the Poisson-Gamma process.

Notation:

At any given time $t=1,2,\dots$ let:

Y_t be the process observation.

θ_t be the process parameter (unknown);

$(\theta_{t-1}|D_{t-1})$: process parameter posterior at time $t-1$
with pdf $p_{t-1,t-1}$ (known)

$(\theta_t|D_{t-1})$: process parameter prior at time t with
pdf $p_{t,t-1}$ (unknown)

$$S_{t-1,t-1} = S [(\theta_{t-1}|D_{t-1})] \quad \text{given by 5.3}$$

$$g(S_{t-1,t-1}) = [1 - \exp(-c S_{t-1,t-1})]^2 ; c \in \mathbb{R}^+$$

Then:

THE MODEL	
Observation equation:	$(Y_t \theta_t) \sim \text{Poisson}(\theta_t)$
System equation:	$p_{t,t-1} \propto [p_{t-1,t-1}]^{g(S_{t-1,t-1})}$

with the model specified as above, the following step shows how the process parameter is sequentially updated in time.

Information:

- (i) The process observations are generated according to the model described above and $g(\cdot)$ is such that c is supposed known at all times.
- (ii) The posterior parameter process distribution at time $t-1$ is assumed to be:

$$(\theta_{t-1}|D_{t-1}) \sim \text{Gamma}(\alpha_{t-1}; \beta_{t-1})$$

where $\alpha_{t-1} > 1$ and $\beta_{t-1} > 0$ for all $t=1,2,\dots$

$(\theta_{t-1}|D_{t-1})$: process parameter posterior at time $t-1$
with pdf $p_{t-1,t-1}$ (known)

$(\theta_t|D_{t-1})$: process parameter prior at time t with
pdf $p_{t,t-1}$ (unknown)

$$S_{t-1,t-1} = S [(\theta_{t-1}|D_{t-1})] \quad \text{given by 5.3}$$

$$g(S_{t-1,t-1}) = \left[1 - \exp(-c S_{t-1,t-1}) \right]^2 ; c \in \mathbb{R}^+$$

Then:

THE MODEL	
Observation equation:	$(Y_t \theta_t) \sim \text{Poisson}(\theta_t)$
System equation:	$p_{t,t-1} \propto [p_{t-1,t-1}]^{g(S_{t-1,t-1})}$

with the model specified as above, the following step shows how the process parameter is sequentially updated in time.

Information:

- (i) The process observations are generated according to the model described above and $g(\cdot)$ is such that c is supposed known at all times.
- (ii) The posterior parameter process distribution at time $t-1$ is assumed to be:

$$(\theta_{t-1}|D_{t-1}) \sim \text{Gamma}(\alpha_{t-1}; \beta_{t-1})$$

where $\alpha_{t-1} > 1$ and $\beta_{t-1} > 0$ for all $t=1,2,\dots$

UPDATING	PROCEDURE
<p><u>Prior time t :</u></p> $(\theta_t D_{t-1}) \sim \text{Gamma} (\alpha_t^* ; \beta_t^*)$ $\alpha_t^* = g(S_{t-1,t-1}) (\alpha_{t-1} - 1) + 1 \quad - - - - - (5.4)$ $\beta_t^* = g(S_{t-1,t-1}) \beta_{t-1} \quad - - - - - (5.5)$ <p><u>Updating:</u></p> <p>Observing $Y_t = y_t$, $(\theta_t D_t)$ is updated as:</p> $(\theta_t D_t) \sim \text{Gamma} (\alpha_t, \beta_t)$ $\alpha_t = \alpha_t^* + y_t \quad - - - - - (5.6)$ $\beta_t = \beta_t^* + 1 \quad - - - - - (5.7)$	

Finally, the prediction of future observations is obtained as summarized below:

PREDICTION	ℓ -STEPS	AHEAD
<p><u>Parameter:</u> $(\theta_{t+j} D_t) ; j=1,2,\dots,\ell$</p> $(\theta_{t+j} D_t) \sim \text{Gamma} (\alpha_{t+j}^* ; \beta_{t+j}^*)$ <p>where, for $j=2,3,\dots,\ell$</p> $\alpha_{t+j}^* = g(S_{t+j-1,t}) \alpha_{t+j-1}^* - 1) + 1 \quad - - - - - (5.8)$ $\beta_{t+j}^* = g(S_{t+j-1,t}) \beta_{t+j-1}^* \quad - - - - - (5.9)$ $S_{t+j-1,t} = S[(\theta_{t+j-1} D_t)]$ <p>and for $j=1$ as in equations (5.4) & (5.5) with $t \rightarrow t+1$</p>		

Observation $(Y_{t+j}|D_t); j=1,2,\dots, \ell$

$$(Y_{t+j}|D_t) \sim \text{Neg. Bin.} (p_{t+j}^{(1)}, p_{t+j}^{(2)})$$

$$p(Y_{t+j}|D_t) = \binom{Y_{t+j} + p_{t+j}^{(2)} - 1}{p_{t+j}^{(2)} - 1} \cdot [p_{t+j}^{(1)}]^{Y_{t+j}} \cdot [1 - p_{t+j}^{(1)}]^{p_{t+j}^{(2)}}$$

$$p_{t+j}^{(1)} = 1 / (1 + \beta_{t+j}^*) \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (5.10)$$

$$p_{t+j}^{(2)} = \alpha_{t+j}^* \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (5.11)$$

5.4) Limiting Form of the BEF Poissor.-Gamma Model

Referring to the Harrison & Stevens steady state normal model, it is not difficult to see that it reaches its limiting form with a constant positive value for the posterior variance C_t . This is due to the fact that C_t does not depend on the observations (y_1, y_2, \dots, y_t) , but just on the value of t . Following the argument, it is shown by Harrison & Stevens, (1976 a) that in this steady state of the model the limiting form for the posterior mean (or mode) tends to:

$$m_t = A \sum_{i=0}^{\infty} (1-A)^i y_{t-i} \quad \text{as } t \rightarrow \infty \quad - \quad - \quad - \quad - \quad (5.12)$$

where $A = C/V$; C : limiting posterior variance C_t . Of course this limiting process with constant A is the established "Exponentially Weighted Moving Average" (EWMA).

If we now concentrate on our Poisson-Gamma BEF model, we can clearly see that the above argument does not follow. This is due to the fact that in the present case, the system uncertainty is not independent of the observations y_t , as we can see from equations (5.3), (5.6) and (5.7). In other words, while in the normal model $C_t \rightarrow C$ automatically implies $S_{t,t} \rightarrow S$ and consequently $g(S_{t,t}) \rightarrow g$, in the Poisson-Gamma case, neither, $S_{t,t}$ nor $g(S_{t,t})$ will have a fixed limiting value but instead, will vary according to the amount of information brought to the system by the most recent observation.

As we have mentioned before, this limiting property of our BEF is the key difference between our formulation and Smith's model. The constant value for the exponent $g(S_{t,t})$ (posterior-to-prior transition-system equation) at all times assumed by Smith is never reached in our formulation, even in the limiting state, since in this case, we have a most likely limiting interval for $g(S_{t,t})$ ($S_{t,t}$ or $H_{t,t}$), as opposed to a single limiting value.

5.5) Applications and Discussions:

We now show some interesting features of our BEF described in the previous sections, when applied to simulated and real Poisson data.

The method was first applied to the data shown in table D.1 of appendix D. They correspond to 500 constant mean Poisson observations generated by simulation (mean constant equal to 3). The objective of the use of the method to this set of observations is not only to test the consistency of the method but also to check its validity as an estimative procedure for the constant mean of the Poisson data.

In other words, we should like the method to correct the initial wrong assumption that we have a time dependent rate for the mean of the Poisson process.

We considered initially the first 250 observations and then the whole set. For the first half of the sample, we estimate the constant "c" of the model using the aggregate likelihood procedure described in section 4.6.7. The results are shown in table D.2. As we should expect, the optimum \hat{c} ($\hat{c} \sim 39.6$) is rather big, which gives for $g(S_{t,t})$ a constant value very close to 1 for all $S_{t,t}$. However, the real confirmation of a constant mean Poisson process can be drawn when we add the other half. We should expect now a higher value for c since by adding these new observations we are giving more information to the

model and consequently, the uncertainty value $S_{t,t}$ tends to decrease with t . The calculated $\hat{c} = 49.4$ shown in table D.3 confirms the consistency of the method. As a matter of illustration, we show in table D.4 the values for the entropies (H and S) for the last five observations for both cases. From there we can clearly see the gaining of information due to the new observations added to the model, in terms of the uncertainty functions.

The initial values used in both cases ($\alpha_0=100$, $\beta_0=33$), constitute a reasonable representation of our state of knowledge about the system given the prior information available. In setting these values, we used the fact that the Poisson data have a constant mean around 3 and so, we assume the initial mode for the parameter equal to this value, i.e., $(\alpha_0-1)/\beta_0 = 3$. Consequently, the initial coefficient of variation ($1/\sqrt{\alpha_0}$) is equal to 0.1, giving an indication of the high degree of certainty we have about the parameter of the model. It should be recalled that the coefficient of variation of a Gamma variate for which $\alpha > 1$ lies in the interval $[0,1]$.

To conclude this illustration, from table D.5 we can see how steady the system is after 500 observations and also the degree of certainty about the parameter, expressed by the small variance for the parameter distribution.

As a second illustration, we show an application of our method for real data in which there exists a random fluctuation of the underlying mean, that is, the data form a sample from a Poisson process of varying rate.

The data correspond to the number of weekly deaths caused by acute respiratory infections in Greater London, covering the period from 15th February 1972 to 1st October 1976, as shown in table D.6 and illustrated in figure D.1 .

Following the sequence of section 5.3, we first estimated the constant c from the given data. The result shown in table D.7 gives $c = 0.57$ and the corresponding value for the aggregate likelihood equal to 27.6292 .

Also interesting to point out in this estimated value for c , is the indication that a Poisson process should be the true assumption and its relatively low value ($c=0.57$) indicates among other things the existence of a variation on the underlying parameter. As initial value for the parameters we chose $\alpha_0=6$ and $\beta_0=2$. These values seem to be reasonably in accordance with the data of table D.5, since they correspond to an initial mode equal to 2.5 and an initial coefficient of variation of about 0.41 .

An important feature of the method is its independence of the choice of these starting values, especially if the sample size is not small. However, a preliminary analysis on the existing information is recommended and helpful in setting fair starting values.

In two more tables we give results obtained by the model in two different sections of the series. We only show the posterior and 1-step ahead distributions for the parameter. In the first, table D.8, we can see clearly how quickly the model settles down regardless of the initial value adopted and then, in table D.9, how the model

cope with quite large fluctuations in the system.

Finally, in figure D.2 we show the plot of the posterior mode for the 199 observations. From this illustration we can see the smooth change in the system parameter mode with the observation pattern.

CHAPTER 6 : POISSON-GAMMA MULTI STATE MODEL

6.1) Introduction:

As stated in chapter 4, our BEF allows us to consider in the model formulation, uncertainty in the parameter values and in the model itself.

In this chapter, we show how the single Poisson-gamma model of chapter 5 can be extended to take the uncertainty in the generating model into consideration, at each time-point. This problem, as considered for the normal case by Harrison & Stevens (1976a), can be incorporated into classification II of the Multi Process Models.

The formulation of the Multi Process Poisson-Gamma Model which we shall present here, is in particular applied to epidemic data by considering two different possibilities (states) of the generating model at each time-point:

State I : No epidemic

State II: Epidemic

The main purpose of the extension is to allow for prompt recognition by the model of state changes within the system. From the nature of epidemic data, a single state approach would take a considerable number of observations (a long transition time) to react to changes in the system while the two-state approach reduces this transition time, yielding a more reliable forecasting system.

Although a general n-state model could be formulated, we confine ourselves in this chapter to the two-state case applied to data showing the epidemic wave pattern. Models with the same basic structure are

often appropriate to other situations.

In the next two sections we give a theoretical description of the model and its updating procedures and in the final section its numerical application to a particular set of epidemic data is shown.

6.2) The Model

We now describe briefly the steps leading to the model structure and its updating equations for the parameters and probabilities involved.

Accordingly, we observe a Poisson process Y_t whose level θ_t follows a gamma distribution. We believe that at any time t , the generating model is a random choice between two models, i.e., two states, where:

$$M_t^{(1)} : \text{Model 1: (No epidemic); } \theta_t = \theta_c \quad - \quad - \quad - \quad - \quad (6.1)$$

θ_c small positive constant

$$M_t^{(2)} : \text{Model 2: (Epidemic); } \theta_t \text{ r.v. gamma distributed} \quad - \quad - \quad - \quad (6.2)$$

Equation (6.1) states that when there is no epidemic, the observations come from a Poisson process with a constant, low-valued rate ($\theta_t = \theta_c$), implying the assignment of a high probability to the occurrence of small-valued observations (depending on the selected value for θ_c), and almost zero probability to the occurrence of high-valued observations. On the other hand, with $M_t^{(2)}$ of (6.2), θ_t is a gamma distributed random variable and the model itself corresponds to the single steady state Poisson-Gamma BEF, described in chapter 5.

Given the information up to time $t-1$ (D_{t-1}), the updating system $t-1 \rightarrow t$ is as follows:

6.2.1) Information a Priori:

Given only D_{t-1} , before data Y_t comes to hand we know the quantities described in subsection (a), and calculate the quantities of subsection (b) as shown below:

(a) Known quantities at time $t-1$:

(a1) Probability that the model j was operating at time $t-1$:

$$p_{t-1}^{(j)} = \text{Prob} \{ M_{t-1}^{(j)} | D_{t-1} \}; j=1,2 \quad - \quad - \quad - \quad - \quad (6.3)$$

(a2) Parameter distribution conditional on $M_{t-1}^{(j)}$ (model j in operation at $t-1$):

$$(\theta_{t-1} | M_{t-1}^{(1)} D_{t-1}) = \theta_c.$$

$$(\theta_{t-1} | M_{t-1}^{(2)} D_{t-1}) \sim \text{Gamma}(\alpha_{t-1}^{(2)}; \beta_{t-1}^{(2)}) \quad - \quad - \quad - \quad - \quad (6.4)$$

(a3) Model transition probabilities, i.e., probability that model j is operating at time t given that model i was operating at time $t-1$.

We use the notation:

$$\pi_{ij} = \text{Prob} \{ M_t^{(j)} | M_{t-1}^{(i)} D_{t-1} \}; i,j=1,2 \quad - \quad - \quad - \quad - \quad (6.5)$$

There are four such probabilities

t	M ⁽¹⁾	M ⁽²⁾
t-1	M ⁽¹⁾	M ⁽²⁾
M ⁽¹⁾	Π_{11}	Π_{12}
M ⁽²⁾	Π_{21}	Π_{22}

(b) Calculated quantities at time t-1:

- (b1) Probability based on D_{t-1} that model i operated at time $t-1$ and model j will operate at time t , i.e.,
 $\text{Prob} \{M_{t-1}^{(i)} M_t^{(j)} | D_{t-1}\}$.

Since:

$$\text{Prob} \{M_{t-1}^{(i)} M_t^{(j)} | D_{t-1}\} = \text{Prob} \{M_t^{(j)} | M_{t-1}^{(i)} D_{t-1}\} \times$$

$$\text{Prob} \{M_{t-1}^{(i)} | D_{t-1}\},$$

we have, using equations (6.3) and (6.5):

$$\text{Prob} \{M_{t-1}^{(i)} M_t^{(j)} | D_{t-1}\} = \Pi_{ij} \cdot p_{t-1}^{(i)} \quad - \quad - \quad - \quad (6.6)$$

$i, j=1, 2$

- (b2) Conditional one step ahead predictive distribution, i.e. the distribution for $(Y_t | M_{t-1}^{(i)} M_t^{(j)} D_{t-1})$; $i, j=1, 2$.

To calculate this distribution we first need the conditional distribution for the parameter

$$(\theta_t | M_{t-1}^{(i)} M_t^{(j)} D_{t-1}).$$

Referring to our model definition (6.1) and (6.2), we can clearly see that to calculate this parameter distribution we have to consider separately the cases $M_t^{(1)}$ and $M_t^{(2)}$, due to the definitions of our models.

For $j=2$ and $i=1,2$ it is clear that:

$$(\theta_t | M_{t-1}^{(i)}, M_t^{(2)}, D_{t-1}) \sim \text{Gamma}(\alpha_t^{*(i,2)}, \beta_t^{*(i,2)})$$

where, for the particular transition 1 to 2 (no epidemic to epidemic), a subjective assumption for the distribution is required. From the conditional parameter distribution, we use the results from chapter 5 to obtain:

$$(Y_t | M_{t-1}^{(i)}, M_t^{(2)}, D_{t-1}) \sim \text{Neg.Bin.}(p_{1,t}^{(i,2)}, p_{2,t}^{(i,2)}) \quad (6.7)$$

where $p_{1,t}^{(i,2)}$ and $p_{2,t}^{(i,2)}$ are calculated from $\alpha_t^{*(i,2)}$ and $\beta_t^{*(i,2)}$ by the use of equations (5.10) and (5.11).

For $j=1$ and $i=1,2$ we have a different situation. In this case, whatever happened at time $t-1$, we are certain about the parameter at time t as we can see from (6.1).

$$\text{Therefore: } (\theta_t | M_{t-1}^{(i)}, M_t^{(1)}, D_{t-1}) = \theta_c$$

and consequently:

$$(Y_t | M_{t-1}^{(i)}, M_t^{(1)}, D_{t-1}) \sim \text{Poisson}(\theta_c) \quad (6.8)$$

6.2.2) Updating System:

Having observed $Y_t = y_t$, the parameter and the probabilities involved in the model are updated as follows:

Referring to our model definition (6.1) and (6.2), we can clearly see that to calculate this parameter distribution we have to consider separately the cases $M_t^{(1)}$ and $M_t^{(2)}$, due to the definitions of our models.

For $j=2$ and $i=1,2$ it is clear that:

$$(\theta_t | M_{t-1}^{(i)}, M_t^{(2)}, D_{t-1}) \sim \text{Gamma}(\alpha_t^{*(i,2)}, \beta_t^{*(i,2)})$$

where, for the particular transition 1 to 2 (no epidemic to epidemic), a subjective assumption for the distribution is required. From the conditional parameter distribution, we use the results from chapter 5 to obtain:

$$(Y_t | M_{t-1}^{(i)}, M_t^{(2)}, D_{t-1}) \sim \text{Neg. Bin.}(p_{1,t}^{(i,2)}, p_{2,t}^{(i,2)}) \quad (6.7)$$

where $p_{1,t}^{(i,2)}$ and $p_{2,t}^{(i,2)}$ are calculated from $\alpha_t^{*(i,2)}$ and $\beta_t^{*(i,2)}$ by the use of equations (5.10) and (5.11).

For $j=1$ and $i=1,2$ we have a different situation. In this case, whatever happened at time $t-1$, we are certain about the parameter at time t as we can see from (6.1).

$$\text{Therefore: } (\theta_t | M_{t-1}^{(i)}, M_t^{(1)}, D_{t-1}) = \theta_c$$

and consequently:

$$(Y_t | M_{t-1}^{(i)}, M_t^{(1)}, D_{t-1}) \sim \text{Poisson}(\theta_c) \quad (6.8)$$

6.2.2) Updating System:

Having observed $Y_t = y_t$, the parameter and the probabilities involved in the model are updated as follows:

(i) Posterior parameter distribution: $(\theta_t | M_{t-1}^{(i)} M_t^{(j)} D_t); i, j=1,2$

As we have mentioned in (b2) of section 6.2.1 for the prior parameter distribution; in order to get the posterior, two distinct cases should be considered, depending on the model obtained at time t .

For $j=2$ and $i=1,2$, by straight forward use of Bayes' Law we obtain:

$$(\theta_t | M_{t-1}^{(i)} M_t^{(2)} D_t) \sim \text{Gamma}(\alpha_t^{(i,2)}, \beta_t^{(i,2)}) \quad (6.9)$$

where:

$$\alpha_t^{(i,2)} = \alpha_t^{*(i,2)} + y_t$$

$$\beta_t^{(i,2)} = \beta_t^{*(i,2)} + 1$$

For $j=1$ and $i=1,2$ we then use (6.1), giving:

$$(\theta_t | M_{t-1}^{(i)} M_t^{(1)} D_t) = \theta_c \quad (6.10)$$

(ii) Model probabilities:

Given D_t , our task now is to obtain an updated expression for the probability that model i was operating at time $t-1$ and model j is in operation at time t , i.e., we want to update $\text{Prob} \{ M_{t-1}^{(i)} M_t^{(j)} | D_t \}$ which we call $p_t^{(i,j)}$ for simplicity.

we know that:

$$\text{Prob} \{ M_{t-1}^{(i)} M_t^{(j)} | D_t \} \propto \text{Prob} \{ y_t | M_{t-1}^{(i)} M_t^{(j)} D_{t-1} \} \cdot \text{Prob} \{ M_{t-1}^{(i)} M_t^{(j)} | D_{t-1} \} .$$

However, from (6.6) we know that:

$$\text{Prob} \{ M_{t-1}^{(i)} \quad M_t^{(j)} \mid D_{t-1} \} = \pi_{ij} \cdot p_{t-1}^{(i)} ; \quad i, j=1, 2$$

and the first term on the right hand side is obtained directly from the corresponding distributions given by either equation (6.7) or (6.8). Denoting this value by $p_{t-1}^{(i,j)}(y_t)$ We then have:

$$p_t^{(i,j)} \propto \pi_{ij} \cdot p_{t-1}^{(i)} \cdot p_{t-1}^{(i,j)}(y_t)$$

or, by normalizing:

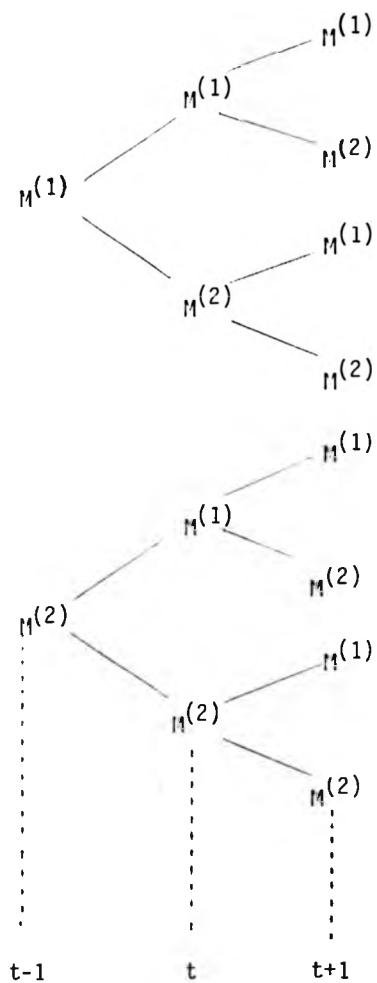
$$p_t^{(i,j)} = K \cdot \pi_{ij} \cdot p_{t-1}^{(i)} \cdot p_{t-1}^{(i,j)}(y_t) \quad - \quad - \quad - \quad - \quad (6.11)$$

where:

$$K = \left[\sum_{i=1}^2 \sum_{j=1}^2 \pi_{ij} \cdot p_{t-1}^{(i)} \cdot p_{t-1}^{(i,j)}(y_t) \right]^{-1}$$

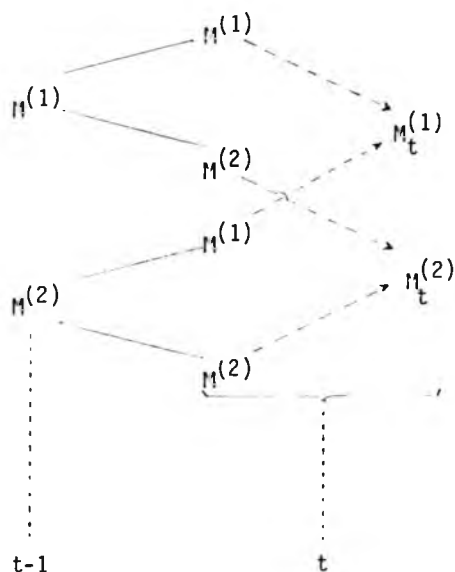
6.3) Collapsing Procedure:

The results obtained so far, although mathematically correct, present a serious practical difficulty. We started with two models $M^{(1)}$ and $M^{(2)}$ at time $t-1$ and obtained four models at time t . Repeating the procedure for the transition $t \rightarrow t+1$ we arrive at eight models, as schematically shown below:



If we proceed in this way, after a few observations the computation would become rather tedious and the computer time and storage would become intolerable.

An approximation has to be introduced and we shall adopt the following scheme:



Analytically, this collapsing procedure operates as follows:

(i) Collapsed Model $M_t^{(1)}$

In this case, both models obtained at time t show the peculiarity of $\theta_t = \theta_c$. Then $M_t^{(1)}$ is the model representing $\theta_t = \theta_c$ and therefore, the prior probability at time t (collapsed probability) is:

$$p_t^{(1)} = \text{Prob} \{ M_t^{(1)} | D_t \} = \sum_{i=1}^2 p_t^{(i,1)} \quad \text{--- (6.12)}$$

where $p_t^{(i,1)}$, $i=1,2$ is given by (6.11)

(ii) Collapsed Model $M_t^{(2)}$

If we look at equation (6.9), we can clearly see that there are two possible ways to obtain model 2 at time t : $M^{(1)}$ at $t-1$

to $M^{(2)}$ at t and $M^{(2)}$ at $t-1$ to $M^{(2)}$ at t . In both cases, the parameter θ_t has a known Gamma distribution and a corresponding updated probability $p_t^{(1,2)}$ and $p_t^{(2,2)}$ assigned for each.

In other words, we have at time t a mixture of two Gamma distributions and our aim is to approximate this mixed distribution by a single one that preserves the main characteristics of the mixed distribution. It is also clear that to enable the procedure to be carried out at future time points, this collapsed parameter distribution is required also to be Gamma distributed.

This problem, usually regarded as the dissection of a heterogeneous population into more homogeneous parts [Johnson & Kotz ; 1969 and 1970], was first faced in the time series context by Harrison and Stevens (1971), (1976a) for the normal case.

They approximated a mixture of a finite number of normal distributions by a single normal, by considering the mean and variance for the single distribution to be the same as for the mixed distribution, that is, by equating the sufficient statistics of the mixture to the corresponding sufficient statistics of the desired single distribution.

Although we have the same problem in our gamma case, our approach to the collapsed single prior gamma distribution is elegantly obtained through the same line of general thinking.

Firstly, if we refer to the results in chapter 3 it is quite clear that the Harrison & Stevens procedure to collapse the normal

mixture into a single normal can be interpreted in a different way. Indeed, by specifying the sufficient statistics of the desired single distribution they are in fact using Jayne's principle and consequently they are breaking up the discrete mixture in a single one that is the least prejudiced probability assignment satisfying the constraints given in terms of the sufficient statistics.

Following the same general train of thought, the collapsed single Gamma distribution is elegantly obtained by straight forward use of Jayne's formalism as presented in chapter 3.

In other words, if we consider the mean and the geometric mean of the mixture as the known constraints for Jayne's principle, then, satisfying this information we obtain as the least prejudiced distribution a single Gamma distribution that collapses the mixture.

Consider the distributions in (6.9) written in terms of the expected values of the sufficient statistics :

$$(\theta_t | M_{t-1}^{(i)} M_t^{(2)} D_t) \sim \text{Gamma} (m_t^{(i,2)} ; gm_t^{(i,2)})$$

where:

$$m_t^{(i,2)} = E \{ \theta_t | M_{t-1}^{(i)} M_t^{(2)} D_t \} = \alpha_t^{(i,2)} / \beta_t^{(i,2)}$$

$$gm_t^{(i,2)} = E \{ \ln \theta_t | M_{t-1}^{(i)} M_t^{(2)} D_t \} = \psi(\alpha_t^{(i,2)}) - \ln(\beta_t^{(i,2)}) ;$$

$$\psi(\cdot) \text{ Digamma function; } \psi(x) = \frac{d}{dx} \Gamma(x); \Gamma(\cdot) \text{ gamma function.}$$

Then, the collapsed distribution for the parameter at time t , i.e., the distribution for $(\theta_t | M_t^{(2)} D_t)$ is the maximum entropy distribution subject to the constraints:

$$E \{ \theta_t | M_t^{(2)} D_t \} = m_t^{(2)}$$

$$E \{ \ln \theta_t | M_t^{(2)} D_t \} = gm_t^{(2)}$$

where:

$$m_t^{(2)} = \sum_{i=1}^2 m_t^{(i,2)} \cdot p_t^{(i,2)} / p_t^{(2)} \quad - \quad - \quad - \quad - \quad (6.13)$$

$$gm_t^{(2)} = \sum_{i=1}^2 gm_t^{(i,2)} p_t^{(i,2)} / p_t^{(2)} \quad - \quad - \quad - \quad - \quad (6.14)$$

$$p_t^{(2)} = \text{Prob} \{ M_t^{(2)} | D_t \} = \sum_{i=1}^2 p_t^{(i,2)} \quad - \quad - \quad - \quad - \quad (6.15)$$

In this way, the distribution obtained for $(\theta_t | M_t^{(2)} D_t)$ is, according to Jayne's principle and the constraints (6.13) and (6.14), a single gamma distribution that collapses the mixed parameter distribution. (See section 3.5).

6.4) Case Study.

We now show the results of the two-state model described in the previous section when applied to the data given in Table E.1 and illustrated in figure E.1; 222 weekly notifications of measles cases in Truro Rural District, Cornwall, covering the period from the 40th week of 1966 to the 52nd week of 1970 [Cliff et al, 1975] (See also table F.5 of appendix F) .

The same dataset was also used as a specimen for the single state Poisson-gamma model of chapter 5 and the results obtained from the single-state model (SI) and the multi-state model (IM) are compared. We briefly explain how the initial parameters and probabilities can be better selected by use of the data.

We then show the relevant results of the IM approach applied to the measles data and finally, the comparison between the SI and the IM.

6.4.1) Preliminary Data Analysis:

The input values necessary to set the IM, as described in section 6.2.1 are:

$$p_0^{(j)} = \text{Prob} \{ H^{(j)} \mid D_0 \} ; \quad j=1,2$$

$$\theta_c \text{ for model } H_t^{(1)}$$

Π_{ij} matrix

location for the distribution of $(\theta_t \mid H_{t-1}^{(1)}, H_t^{(2)} \mid D_{t-1})$

Here the data have been used to help give plausible initial values for these quantities. In order that we may use the data, we merely have to construct definitions for "epidemic" and "non-epidemic" periods and the transitions from one period to another. It is quite obvious that an epidemic period is well characterized (as is a non-epidemic period). A period of no notifications, possible including one or two non consecutive notified cases, would roughly constitute a non-epidemic period, while a period where non-zero notifications predominate, constitute an epidemic wave. With respect to the transitions we can consider:

- (i) If we are in an epidemic period, two consecutive zero observations following a non-zero observation can approximately be considered an epidemic to non-epidemic transition.
- (ii) If we are in a non-epidemic period, two consecutive non-zero observations, one of them greater than or equal to 2, following at least two zero observations can approximately be considered a non-epidemic to epidemic transition.

In accordance with (i) and (ii), the measles data of table E.1 show 4 epidemic to non-epidemic transitions and 3 non-epidemic to epidemic transitions out of the 222 observations. These balanced occurrences suggest that a reasonable estimate for the transition probability matrix is:

$$\Pi \approx \begin{pmatrix} 0.97 & 0.03 \\ 0.04 & 0.96 \end{pmatrix}$$

In selecting θ_c for the constant mean Poisson model $\pi^{(1)}$, we can again use the data to have an idea of its value. Bearing in mind considerations (i) and (ii), we could say that out of 222 observations, model 1 (non-epidemic period) is appropriate at weeks: 41/1967 to 42/1967, 46/1967 to 21/1968; 34/1968 to 15/1970 and 44/1970 to 52/1970, making a total of 127 times. Within these intervals, the observed sum of all data is 15, and so, based only on this information, a reasonable value for θ_c would be $\theta_c \sim 0.12$.

With respect to the model probabilities $p_0^{(j)}$; $j=1,2$ the data suggest a tendency to favour model 2 and so, we use $p_0^{(1)} = 0.4$ and $p_0^{(2)} = 0.6$.

Finally, as we have mentioned before, the prior specification of the parameter at time t for the model transition $\Pi^{(1)}$ at time $t-1$ to $\Pi^{(2)}$ at time t , needs a subjective assumption for the location of the parameter distribution. That is due to the fact that for such a transition, we are facing the situation where the prior parameter uncertainty is already established by the BEF formulation, i.e.:

$$S(\theta_t | D_{t-1}) = 1/c$$

$$S(\theta_{t-1} | D_{t-1}) = 0$$

Then, the specification of a location for $(\theta_t | D_{t-1})$ and the above known uncertainty would suffice for the distribution of $(\theta_t | \Pi_{t-1}^{(1)}, \Pi_t^{(2)}, D_{t-1})$. For the particular sample of table (E.1), it seems reasonable to assume:

$$(\theta_t | \Pi_{t-1}^{(1)}, \Pi_t^{(2)}, D_{t-1}) \sim \text{Gamma} (\text{mode} \approx 3.5; S(\cdot) = 1/c)$$

6.4.2) Results:

We now present the relevant results obtained by the IM approach to the measles data. With the initial probabilities, θ_c and transition matrix as given in section 6.4.1, we first estimated the constant c for the model $\Pi^{(2)}$, following the same procedure as discussed in section 4.6.7. The results in table E.2 give $\hat{c}=1.66$ and the corresponding support equal to 107.36138.

We next show some interesting features obtained by the IM, especially the updating of the various probabilities involved in some sections of the data.

From table E.3 we can see how quickly the MM recognizes the transition $M^{(1)}$ to $M^{(2)}$ when an unexpected two notifications are observed and how this change is confirmed when three is observed at the next time-point. It is interesting to note the increase in the posterior probability of the transition $M^{(1)}$ to $M^{(2)}$ ($p_t^{(1,2)}$) from 0.02 at time 55 to 0.57 at time 56, as we should expect. Also from table E.3 we can clearly see that at time 59, although no notifications have been observed, the MM does not have enough information for a change of state. However, the change in the posterior probability of the transition $M^{(2)}$ to $M^{(1)}$ ($p_t^{(2,1)}$), from 0.001 to 0.340 is quite substantial and it is only when another zero is observed at the next time-point that the transition to $M^{(1)}$ is confirmed.

Another interesting $M^{(1)}$ to $M^{(2)}$ transition is shown in table E.4. When four is observed at time 186 after a long non epidemic period, the MM goes directly to $M^{(2)}$ with a very high probability. It is only at time-point 192 that the epidemic out-break is confirmed, because between $t=187$ and $t=191$ the few cases registered are not consistent enough to guarantee the transition. However, it is important to note that, after the unexpected four at $t=186$ the MM changes from $M^{(1)}$ to $M^{(2)}$ and there stays, even though the following observations do not strongly support this transition.

To conclude, we show in table E.5 the end of the epidemic period started in $t=186$. After observing the first zero at time 213, the MM is not sure enough of the end of the epidemic wave, though the probabilities are substantially revised. The transition $M^{(2)} \rightarrow M^{(1)}$ is

established, however, when another zero is observed at the next time-point.

6.4.3) SM and MM comparison:

In order to show the improvements achieved with the MM formulation compared with the SM formulation, the basic techniques developed in chapter 5 were applied to the same data of table E.1 .

As usual, we first estimated the constant c and the results are shown in table E.6 . From this estimation procedure, we can see the substantial improvement in the aggregate likelihood $\sum_{t=1}^{222} \ln p(Y_t | D_{t-1})$.

From tables E.2 and E.6 we have, respectively:

$$\max \sum_{t=1}^{222} \ln p(Y_t | D_{t-1}, MM) = 107.36138 \quad , \quad \text{and}$$

$$\max \sum_{t=1}^{222} \ln p(Y_t | D_{t-1}, SM) = 57.72938$$

This value for the aggregate likelihood under MM, almost twice that under the SM, is mainly caused by the speedy response of the MM when changes in the system pattern occur, as opposed to the slow reaction of the single model, i.e., the SM always takes more observations than the MM to cope with the various changes in the system behaviour over the time scale. These points are shown in tables E.7 and E.8 in terms of the characteristics of the posterior parameter distribution and illustrated in figures E.2 and E.3 where the posterior mode for the 222 data points under MM & SM respectively are plotted.

CHAPTER 7: STEADY STATE BINOMIAL-BETA MODEL

7.1) Introduction:

As a further illustration of the method, we show in this chapter how our BEF model formulation can be applied to the Binomial-Beta process. We shall assume that the process level $\theta_t \in [0,1]$ is a Beta distributed rv. for all $t=1,2,\dots$, while for the process observation we assume the conjugate form, that is, $Y_t \in Z$ is Binomial distributed; $t=1,2,\dots$.

The same formulation applies to the case where the process observation is assumed to be Negative Binomially distributed. However, we shall describe it only assuming the Binomial distribution simply because the Negative Binomial case is a straightforward extension of the Binomial case.

The problem has received scant attention in the literature and in fact Smith, (1978), mentioned above, is the only work dealing with the Binomial-Beta process. However, Smith's approach requires the steady state assumption of the model to be made at each time point.

The organization of the chapter follows the pattern of the previous ones: we give a brief summary of the main characteristics of the Beta distribution before we proceed with the theoretical model description. The last section focusses on the application of the model to real and simulated data. The numerical results of these are shown in appendix F. The real data are the same measles data as in chapter 5 and 6, now illustrating the spatial spread of the epidemic over the whole of Cornwall.

7.2) Beta Variate Characteristics

In order to obtain Shannon's entropy of the Beta distribution, required in our BEF model, we shall first describe briefly the main characteristics of the Beta variate. This summary is largely a congregation of the relevant facts which were found in: Johnson, (1970b); Raiffa & Schlaifer, (1961); Hastings & Peacock, (1974) and Tribus, (1969).

Let X be a continuous rv. defined on the interval $[0,1]$. Then, we say that X is Beta distributed with parameters α and γ ; i.e., $X \sim \text{Be}(\alpha, \gamma)$, if its pdf can be written as:

$$f = f(X | \alpha, \gamma) = [B(\alpha, \gamma)]^{-1} \cdot X^{\alpha-1} \cdot (1-X)^{\gamma-1} \quad (7.1)$$

Where:

$$X \in [0,1]$$

α, γ are the shape parameters; $\alpha, \gamma > 0$

$B(\alpha, \gamma) = [\Gamma(\alpha) \cdot \Gamma(\gamma)] / \Gamma(\alpha + \gamma)$ is the Beta function with parameters α & γ , defined by:

$$B(\alpha, \gamma) = \int_0^1 u^{\alpha-1} \cdot (1-u)^{\gamma-1} \cdot du$$

It is not difficult to show that the mean, variance and the mode of $X \sim \text{Be}(\alpha, \gamma)$ are respectively:

$$E\{X | \alpha, \gamma\} = \alpha / (\alpha + \gamma) \quad (7.2)$$

$$\text{Var}\{X | \alpha, \gamma\} = \alpha \cdot \gamma / [(\alpha + \gamma)^2 \cdot (\alpha + \gamma + 1)] \quad (7.3)$$

$$\text{Mode}\{X | \alpha, \gamma\} = (\alpha - 1) / (\alpha + \gamma - 2) \quad \text{if } \alpha, \gamma > 1 \quad (7.4)$$

We show next the possible forms that $f(x|\alpha, \gamma)$ can have as a function of the values for the parameters α and γ . They can be summarized as follows:

(i) $\alpha > 1$ and $\gamma > 1$ (Figure 7.1)

In this case, $f(x|\alpha, \gamma)$ has a single mode given by (7.4) and:

(i.1) Mode $\{x|\alpha, \gamma\} > 0.5$ if $\alpha > \gamma \Rightarrow f(x|\alpha, \gamma)$ is skewed to the right.

(i.2) Mode $\{x|\alpha, \gamma\} < 0.5$ if $\alpha < \gamma \Rightarrow f(x|\alpha, \gamma)$ is skewed to the left.

(i.3) Mode $\{x|\alpha, \gamma\} = 0.5$ if $\alpha = \gamma \Rightarrow f(x|\alpha, \gamma)$ is symmetrical.

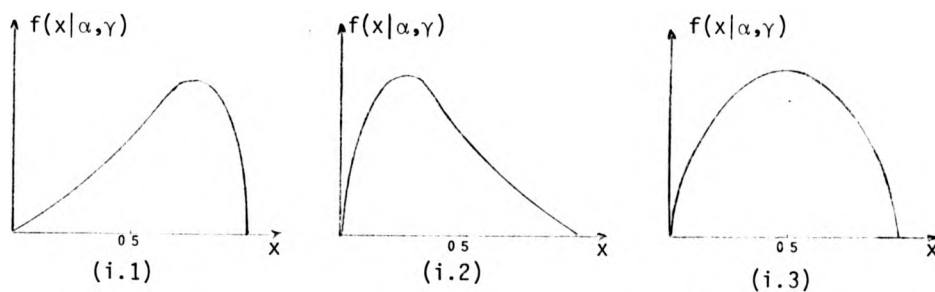


Figure 7.1 : Illustration of Beta pdf - Cases (i).

(ii) $\alpha = \gamma = 1$ (Figure 7.2)

In this case $f(x|\alpha, \gamma)$ is rectangular

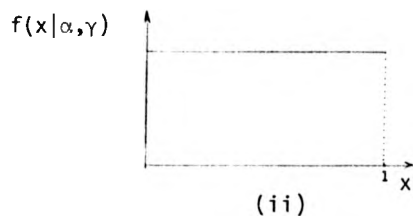


Figure 7.2: Illustration of Beta pdf - Case (ii)

(iii) $\alpha < 1$ and $\gamma < 1$ (Figure 7.3)

In this case $f(X|\alpha, \gamma)$ has an antimode, i.e.,

$f(X|\alpha, \gamma)$ is U-shaped.

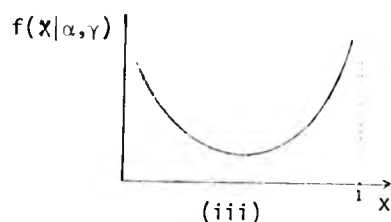


Figure 7.3 : Illustration of Beta pdf - Case (iii).

iv) $(\alpha-1) \cdot (\gamma-1) < 0$ (Figure 7.4)

In this case $f(x|\alpha, \gamma)$ has no mode, i.e., $f(x|\alpha, \gamma)$ is:

(iv.1) J-shaped to the right if $\alpha > \gamma$

(iv.2) J-shaped to the left if $\alpha < \gamma$

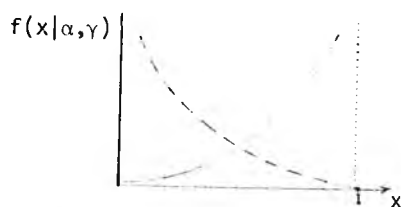


Figure 7.4 : Illustration of Beta pdf - Cases (iv)

Thinking now in terms of our BEF model, we shall consider in this work that Binomial-Beta process whose parameter distribution admits as maximum uncertainty distribution the rectangular form (ii). This means that we shall only consider the cases (i) and (ii), i.e., we assume $\alpha, \gamma \geq 1$.

7.3) Entropy of the Beta Variate.

With the assumptions of the previous section and the results of appendix A, we now proceed with the calculation of the Shannon's entropy

H_X , where $X \sim \text{Be}(\alpha, \gamma)$

From (7.1) $f = f(x | \alpha, \gamma)$ can also be written as:

$$f = \exp \left[(\alpha - 1) \cdot \ln x + (\gamma - 1) \cdot \ln (1-x) - \ln B(\alpha, \gamma) \right]$$

Using the above expression and appendix A, we can define:

$$\theta_1 = \alpha ; \theta_2 = \gamma$$

$$K_1(x) = \ln x ; K_2(x) = \ln(1-x)$$

$$A_1(\theta_1) = A_1(\alpha) = \alpha - 1 ; A_2(\theta_2) = A_2(\gamma) = \gamma - 1$$

$$Q(\alpha, \gamma) = - \ln B(\alpha, \gamma) ; S(x) = 0$$

and consequently:

$$\frac{\partial A_1(\alpha)}{\partial \alpha} = 1 ; \frac{\partial A_2(\gamma)}{\partial \gamma} = 1 ; \frac{\partial Q(\alpha, \gamma)}{\partial \alpha} = \frac{\partial Q(\alpha, \gamma)}{\partial \gamma} = - (\alpha + \gamma - 1)$$

Taking these results into expression (A.3) of appendix A, we obtain for H_X :

$$H_X = -(\alpha + \gamma - 1)(\alpha + \gamma - 2) + \ln B(\alpha, \gamma) \quad (7.5)$$

For the Beta distribution we shall consider in the present work, we only need to study the variation of H_X with α and γ , for $\alpha, \gamma \geq 1$.

It is not difficult to show that:

(i) For $\alpha = \gamma = 1$ (Maximum uncertainty distribution)

$$H_X = 0 .$$

(ii) For $\alpha, \gamma > 1$ we always have $H_X < 0$ and for the minimum uncertainty distribution ($\alpha, \gamma \rightarrow \infty$), we have:

$$\lim_{\alpha, \gamma \rightarrow \infty} H_X = -\infty$$

From (i) and (ii) it is clear that for $\alpha, \gamma > 1$, H_X is non positive and not defined on \mathbb{R} as in the previous cases. This is due to the fact that, in this case X is defined in a finite interval giving $H_X = 0$ for the maximum uncertainty assignment. As a consequence, the S_X function defined as usual, i.e., $S_X = \exp(H_X)$ maps $H_X \in \mathbb{R}^+$ onto $[0,1]$ for the kind of Beta distributions we are considering. This is however not a restriction for our BEF model. In fact, the same $g(S_{t,t})$ curve for the posterior-to-prior transition can be assumed, the only difference lying in the fact that $g(S_{t,t})$ has reached its asymptotic value at $S_{t,t} = 1$ and consequently for $S_{t,t} \in I_S$, where $I_S \subset [0,1]$, the process is in its steady state.

From (7.5) we can write for S_X :

$$S_X = B(\alpha, \gamma) \cdot \exp[-(\alpha + \gamma - 1)(\alpha + \gamma - 2)] \quad - \quad - \quad - \quad - \quad (7.6)$$

7.4) BEF Binomial-Beta system; Model Description.

The BEF for the Binomial-Beta process can now be formulated following the same sequence as in the previous applications.

Notation :

At any given time $t=1,2,\dots$ let:

Y_t be the process observation

θ_t be the process parameter (unknown)

$(\theta_{t-1} | D_{t-1})$: process parameter posterior at time $t-1$, with
with pdf $p_{t-1,t-1}$ (known)

$(\theta_t | D_{t-1})$: process parameter prior at time t ,
with pdf $p_{t,t-1}$ (unknown)

$S_{t-1,t-1} = S[(\theta_{t-1} | D_{t-1})]$ as defined in (7.6)

$g(S_{t-1,t-1}) = [1 - \exp(-c S_{t-1,t-1})]^2$; $c \in \mathbb{R}^+$

THE MODEL	
Observation equation :	$(Y_t \theta_t, n) \sim \text{Binomial}(n, \theta_t)$ n known
System equation :	$p_{t,t-1} \propto [p_{t-1,t-1}]^{g(S_{t-1,t-1})}$

The process parameter is sequentially updated in time as follows:

Information:

- (i) The process observations are generated according to the model described above and $g(\cdot)$ is such that c is supposed known at all times.

- (ii) The posterior parameter process distribution at time $t-1$ is assumed to be:

$$(\theta_{t-1} | D_{t-1}) \sim \text{Beta}(\alpha_{t-1}, \gamma_{t-1}); \text{ where } \alpha_{t-1}, \gamma_{t-1} \geq 1$$

for all $t=1,2,\dots$

UPDATING	PROCEDURE
<u>Prior time t:</u>	
	$(\theta_t D_{t-1}) \sim \text{Be}(\alpha_t^*, \gamma_t^*)$
	$\alpha_t^* = g(S_{t-1,t-1}) \cdot (\alpha_{t-1} - 1) + 1$ - - - (7.7)
	$\gamma_t^* = g(S_{t-1,t-1}) \cdot (\gamma_{t-1} - 1) + 1$ - - - (7.8)
<u>Updating:</u>	
	Observing $Y_t = y_t$ and with n known, $(\theta_t D_t)$
	is updated as:
	$(\theta_t D_t) \sim \text{Be}(\alpha_t, \gamma_t)$
	$\alpha_t = \alpha_t^* + y_t$ - - - (7.9)
	$\gamma_t = \gamma_t^* - y_t + n$ - - - (7.10)

The prediction of future observations is then obtained as:

PREDICTION	ℓ -STEPS	AHEAD
<u>Parameter</u> : $(\theta_{t+j} D_t)$; $j=1,2,\dots,\ell$		

$$(\theta_{t+j}|D_t) \sim \text{Be}(\alpha_{t+j}^*, \gamma_{t+j}^*)$$

where, for $j=2,3,\dots,\ell$

$$\alpha_{t+j}^* = g(S_{t+j-1,t}) \cdot (\alpha_{t+j-1}^* - 1) + 1 \quad (7.11)$$

$$\gamma_{t+j}^* = g(S_{t+j-1,t}) \cdot (\gamma_{t+j-1}^* - 1) + 1 \quad (7.12)$$

$$S_{t+j-1,t}^* = S[(\theta_{t+j-1}|D_t)]$$

and for $j=1$ as in equations (7.7) & (7.8) with $t \rightarrow t+1$

Observation: $(Y_{t+j}|D_t)$; $j=1,2,\dots,\ell$

$$(Y_{t+j}|D_t) \sim \text{Be-Bi}(\alpha_{t+j}^*, \gamma_{t+j}^*, n)$$

where:

$$p(Y_{t+j}|D_t) = \binom{n}{Y_{t+j}} \frac{B(\alpha_{t+j}^* + Y_{t+j}; n + \gamma_{t+j}^* - Y_{t+j})}{B(\alpha_{t+j}^*; \gamma_{t+j}^*)}$$

7.5 Limiting form of the Binomial-Beta BEF

The limiting form for the Binomial-Beta BEF model follows the same argument of the corresponding limiting form of the Poisson-Gamma BEF model described in section 5.4. Here again the system uncertainty is not independent of the observations, implying automatically that either $S_{t,t}$ or $g(S_{t,t})$ will not have a fixed limiting value but instead, depend directly on the amount of information brought into the system by the most recent observation.

This point once again emphasizes the difference between our formulation and Smith's model as we have already mentioned in chapter 6.

7.6) Applications:

We conclude this chapter by showing the performance of the Binomial-Beta BEF when applied to simulated and real data. In order to show the consistency of the method we first apply our model to the set of data shown in table F.1. They correspond to 490 Binomial observations generated by computer with constant parameter $\theta=0.375$ and $n=8$. In applying our BEF to this set of observations, we should like the model itself to correct our initial wrong assumption that we have a Binomial-Beta system, i.e., the assumption that θ_t is a time dependent Beta distributed random variable. In terms of our BEF formulation, among other things, the constant c of the function $g(S_{t,t})$ estimated from the data, should be very high to compensate for the low value of the uncertainty as time progresses.

Let us consider initially the first half of the data. Using the procedure described in chapter 4, we show in table F.2 the results of the constant c estimation from the 245 data points, which form the first half of the sample. From F.2 we can clearly see that the estimate of c is $\hat{c} = 0.12 \times 10^8$, a very high value indeed, giving a clear indication that we can be quite sure that a static assumption for θ_t would be preferable. However, the support for this model:

$$\sum_{t=1}^{245} \ln p(Y_t | D_{t-1}; \hat{c}) \sim 46.415$$

is slightly less than the corresponding support for the static model ($g(S_{t,t})$ for all $t=1,2,\dots$), i.e.,

$$\sum_{t=1}^{245} \ln p(Y_t | D_{t-1}; g(\cdot)=1) \cong 46.421 .$$

Although the static assumption for θ_t is nearly confirmed, the relatively small size of the sample (245) do not yield sufficient information to confirm the time independence of θ_t .

If we now add the other half of the sample and proceed with the estimation of c from all 490 observations, we obtain $\hat{c} = 0.46 \times 10^8$ as shown in table F.3. This increase in the value for c , practically confirms the static assumption made previously, as we should expect. It is also interesting that the support for the static model:

$$\sum_{t=1}^{490} \ln p(Y_t | D_{t-1}; g(\cdot)=1) \approx 99.10$$

is now approximately equal to the support of the BEF with $\hat{c}=0.46 \times 10^8$ (see table F.3). This clearly shows that the data in table F.1 come from a Binomial distribution with $n=8$ and $\theta=0.375$ and that in this case, our BEF formulation provides a sequential Bayesian estimation procedure for the unknown constant parameter θ . By way of illustration, in table F.4 we show the results of the prior-to-posterior analysis for θ_t , for the last eight time points.

As we can see, the posterior mode provides a very good estimate for θ_t and the corresponding low steady value for the variance (0.0006) gives an account of the time-invariance of θ_t .

We now show an interesting application of the Binomial-Beta BEF model formulation to the analysis of the notification statistics for measles outbreaks in Cornwall-England. For a better understanding of the data, we reproduce appendix I of Cliff et al, (1975) where the number of notifications distributed according to the areas in the region,

Although the static assumption for θ_t is nearly confirmed, the relatively small size of the sample (245) do not yield sufficient information to confirm the time independence of θ_t .

If we now add the other half of the sample and proceed with the estimation of c from all 490 observations, we obtain $\hat{c} = 0.46 \times 10^8$ as shown in table F.3. This increase in the value for c , practically confirms the static assumption made previously, as we should expect. It is also interesting that the support for the static model:

$$\sum_{t=1}^{490} \ln p(Y_t | D_{t-1}; g(\cdot)=1) \approx 99.10$$

is now approximately equal to the support of the BEF with $\hat{c}=0.46 \times 10^8$ (see table F.3). This clearly shows that the data in table F.1 come from a Binomial distribution with $n=8$ and $\theta=0.375$ and that in this case, our BEF formulation provides a sequential Bayesian estimation procedure for the unknown constant parameter θ . By way of illustration, in table F.4 we show the results of the prior-to-posterior analysis for θ_t , for the last eight time points.

As we can see, the posterior mode provides a very good estimate for θ_t and the corresponding low steady value for the variance (0.0006) gives an account of the time-invariance of θ_t .

We now show an interesting application of the Binomial-Beta BEF model formulation to the analysis of the notification statistics for measles outbreaks in Cornwall-England. For a better understanding of the data, we reproduce appendix I of Cliff et al, (1975) where the number of notifications distributed according to the areas in the region,

are as shown in table F.5 . For the purpose of analysis, we consider two different sets of observations; one relating to the number of rural districts (RD) affected by the epidemic week by week and the other related to the corresponding number of municipal boroughs (MB) and urban districts (UD) affected by the disease. In counting these data, we consider a unit affected if at least one case is notified for that particular unit. As a result, we obtain the two set of observations shown in tables F.6 & F.7 and illustrated in figures F.1 & F.2. Table F.6 (Figure F.1), shows the weekly number of rural districts units (RD) affected by the measles epidemic out of the 10 RD units of the area (see table F.5), and table F.7 (Figure F.2) shows the weekly number of municipal boroughs & urban districts units (MB & UD) affected by the measles epidemic out of the 17 MB & UD units of the area (see table F.5).

Assuming that the number of units affected by the disease follows a Binomial (θ_t, n) process, whose rate of units affected θ_t has a time varying Beta distribution, the two set of data of tables F.6 and F.7 are respectively:

- (i) $(Y_t | \theta_t) \sim \text{Bi}(\theta_t; 10)$; θ_t Beta distributed and Y_t is the random variable representing the number of RD affected by the measles epidemic.
- (ii) $(Y_t | \theta_t) \sim \text{Bi}(\theta_t; 17)$; θ_t Beta distributed and Y_t is the random variable representing the number of MB & UD affected by the measles epidemic.

Let us now consider the results of the application of our Binomial-Beta BEF to the data of tables F.6 and F.7. We show separately the relevant results for each case and then the relationship between them.

For the RD data of table F.6, we start by estimating the constant c following the sequence of chapter 4. The results shown in table F.8 give $\hat{c} = 2.5$ and the corresponding maximum aggregate likelihood equal to 48.233. This low value for c is a clear indication of the time variation of the rate of the RD affected by the epidemic. Indeed, if we consider the static assumption for this rate $[g(\cdot)=1]$ and calculate the aggregate likelihood, we obtain:

$$\sum_{t=1}^{222} \ln p(Y_t | D_{t-1}; g(\cdot)=1) \cong 32.1761,$$

confirming that a constant rate θ_t would be a very poor assumption.

If we now look at the data as given in table F.5, it is clear that for the period covered we have a severe outbreak of the disease, starting from approximately the 44th week of 1966 and finishing at around the 31st week of 1967, although, apart from Truro RD, only a few districts are contaminated by the disease after the 24th week of 1967. The RD are again affected, but not as badly as before, nearly a year later, between the 22nd and 36th weeks of 1968 and only in 1970, between the 16th and the 36th weeks they are again involved in an outbreak.

To show the response of our model to the above 3 outbreaks, we produce in the tables F.9, F.10 and F.11 the parameters and the mode of the distribution for the rate of the RD units affected by the epidemic. Confirming the evidence from the past data, we can see from table F.9 how the model responds satisfactorily to the critical period, especially between the 5th week of 1967 and the 17th week of 1967 when they are most affected. Another interesting facet is the speedy updating of the model parameter

as the epidemic spreads over the area. As a final illustration, we show in table F.12 part of the one-step ahead predictive distribution and the corresponding observed value for the last 13 weeks.

A similar analysis was made for the MB & UD data of table F.7. The estimation of the constant c is shown in table F.13 and as we can see, $\hat{c}=2.9$ and:

$$\sum_{t=1}^{222} \ln p(Y_t | D_{t-1}, \hat{c}) = 39.4367 .$$

We again compared the steady assumption for the rate θ_t with the corresponding static model ($g(\cdot)=1$) that gave:

$$\sum_{t=1}^{222} \ln p(Y_t | D_{t-1}; g(\cdot)=1) = 25.744 .$$

Tables F.14, F.15 and F.16 illustrates the three major outbreaks for the MB & UD units, in terms of the parameter distribution and in table F.17 the one-step ahead predictive distribution is shown for $t=210, \dots, 222$.

Finally, from the results obtained for the two areas separately, it is quite clear that in all major measles epidemics, the outbreak profiles for the RD and the MB & UD units are almost identical, although in all outbreaks, the rural areas are the first to be ravaged by the epidemic. It is also interesting to note that the peak of the epidemic is reached earlier in the rural districts than in the town, and that the high proportion of infected rural districts is retained until the (later) peaking of the urban epidemic profile, after which the two profiles decay simultaneously sharply to the non-epidemic (background) rate.

To clarify these points we show in figures F.3 and F.4 the posterior mode of the rate of units affected for the RD and the MB & UD areas respectively. From these two curves, we can also see that the rate of RD units affected by the measles epidemic is always higher than the contemporary rate for the MB & UD units and the rural epidemic profile is more ragged than the urban profile.

CHAPTER 8 : STEADY STATE TRUNCATED NORMAL MODEL

8.1) Introduction:

As a final illustration of our method, we describe in this chapter how the truncated normal process can be framed within our BEF model formulation. We shall consider throughout this chapter the process level $\theta_t \in \mathbb{R}^+$ a truncated normal variate, truncated at $\theta_t=0$, while the observation $Y_t \in \mathbb{R}^+$ is also assumed to have a truncated normal distribution, truncated at $Y_t=0$. With the above assumptions, we shall show that the posterior distribution for the parameter is not exactly truncated normal but it is made truncated normal by the use of a Taylor series, expanded as far as the quadratic term. The conjugacy thus obtained for the process is easily modelled according to our BEF formulation : this offers a simple updating system for the process parameter. The non-existence of a standard form for the posterior distribution may be the main reason for the absence of the truncated normal model in Bayesian analysis. The existing literature is only concerned with classical approaches to estimative procedures for the parameters of single and/or double truncated normal distribution. We refer particularly to Cohen, (1949, 1950, 1951, 1955 and 1959) ; Hald, (1949) ; Shah & Jaiswal, (1966) ; Halperin, (1952); Francis, (1946); Raj, (1953); Weiler, (1959); Tallis, (1961) and Regier & Hamdas, (1971). We hope that the above approximation to conjugacy will open the way to a Bayesian formulation for the truncated normal problem.

In this chapter however, we shall consider the steady state BEF model applied to a truncated normal system (process parameter and system observation assumed truncated normal distributed) and, without loss of generality, the truncation point is assumed to be zero for both, i.e., we assume $\theta_t, Y_t \in \mathbb{R}^+$. The prime objective of such a formulation is to provide a model

applied to situations where a steady state normal model would be a strong assumption. Clearly, many situations arise where the nature of the physical system being modelled constrains the observations to take values necessarily greater than some fixed value Y_{inf} (in the present case, we have $Y_{inf}=0$) and, unless this fixed value is very unlikely to occur and the observations show a high degree of concentration (or a very low variance), a purely normal model cannot be the correct assumption. Instead, we can make use of this extra piece of information ($Y_t \geq 0$) and set a truncated normal model that is certainly more in accordance with the real situation. It is also important to remember that in considering the truncated normal model we are automatically extending the normal model that we have described in chapter 4, since, as we shall see later in this chapter, the truncated normal BEF naturally tends to the normal BEF if the system pattern shows such a tendency. As we have mentioned above, the present formulation has $Y_{inf}=0$ for both the process level and system observation. However, it is worth mentioning that any other truncation point can be considered, and even a double truncated normal distribution could be put in terms of our BEF formulation, if that were the case.

The organization of the chapter is slightly different from the previous ones. In section 8.2 we define and derive some important properties of the truncated normal distribution using mainly the material covered in chapter 3. The Shannon's entropy and the corresponding S_t function are shown in section 8.3. In section 8.4 we discuss the problem related with the posterior in the truncated normal model and in section 8.5 the BEF formulation is shown. Finally, the numerical results of some applications of the model are presented in section 8.6 and appendix G.

8.2) Truncated Normal Distribution

In this section we briefly review the concepts of a truncated normal random variable, its characterizations and properties. We shall concentrate on the maximum entropy characterization of the distribution by the use of the material presented in chapter 3.

8.2.1) Definition and Characterizations.

Let X be a continuous rv. from which the following information is available:

- (i) $X \in \mathbb{R}^+$
- (ii) $E\{X\} = m$
- (iii) $E\{X^2\} = v^2 + m^2$ (or $\text{Var}\{X\} = v^2$)

Using Jayne's formulation (chapter 3) to assign the least prejudiced distribution for X , taking into account information (i), (ii) & (iii), the maximum entropy distribution obtained for X is given by:

$$f(x) = \exp(-\lambda_0 - \lambda_1 x - \lambda_2 x^2) \quad \text{--- (8.1)}$$

(8.1) is a truncated normal pdf, truncated at $x=0$, with mean "m" and variance v^2 and Lagrange multipliers λ_i ; $i=0,1,2$.

The same truncated normal distribution for X can also be characterized by the moments of the untruncated distribution. If we consider:

$$X \sim N(\mu, \sigma^2); \text{ truncated at } x=0, \text{ where } \mu \text{ \& } \sigma^2 \text{ are the mean}$$

and variance of the normal untruncated distribution for X , then, the pdf for X can be written as:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{[1-\Phi(-\mu/\sigma)]} \cdot \exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right] \quad \text{--- (8.2)}$$

$$\text{where: } \Phi(t) = \int_{-\infty}^t \phi(u) \cdot du ; \phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{u^2}{2}\right]$$

From (8.1) & (8.2), we obtain:

$$\mu = -\lambda_1 / 2\lambda_2 ; \sigma^2 = 1 / 2\lambda_2 \quad \text{--- (8.3)}$$

In order to have the complete specification for the distribution of X , two kinds of problems should be considered:

- (a) We know the λ_i 's ; $i=0,1,2$ (or μ & σ^2) and want m & v^2 .
- (b) We know m & v^2 and we want the λ_i 's ; $i=0,1,2$ (or μ & σ^2).

Problem (a) does not offer much difficulty, for once λ_1 & λ_2 are known a priori, the moments of the untruncated distribution can be obtained from (8.3) and then, m & v^2 can be easily obtained by:

$$m = \mu + \frac{\sigma}{M(-\mu/\sigma)} \quad \text{--- (8.4)}$$

$$v^2 = \sigma^2 \left[1 - \frac{\mu}{\sigma M(-\mu/\sigma)} - \frac{1}{M^2(-\mu/\sigma)} \right] \quad \text{--- (8.5)}$$

where $M(\cdot)$ is the Mill's ratio, defined by:

$$M(t) = \frac{[1-\Phi(t)]}{\phi(t)} ,$$

and m & v^2 are easily obtained by taking the expectation of x , and x^2 respectively, with respect to $f(x)$ as defined in (8.2).

The solution for (b) is not so easy. One possible way would be the solution of the system of equations (8.4) and (8.5) for μ & σ^2 . But such a system has no straight forward solution as we can see, due to the presence of the Mill's ratio function. However, if we use the properties of the maximum entropy distribution as developed in section 3.3 an easier solution can be obtained, as we show now.

From (3.11) and the information (ii) & (iii), we have:

$$\frac{\partial \lambda_0}{\partial \lambda_1} = -m \quad ; \quad \frac{\partial \lambda_0}{\partial \lambda_2} = -(m^2 + v^2) \quad - \quad - \quad - \quad - \quad (8.6)$$

Also, by solving the integral for the partition function (3.14) with $g_1(x) = x$ and $g_2(x) = x^2$, we obtain:

$$\lambda_0 = \frac{\lambda_1^2}{4\lambda_2} + \ln \left[\sqrt{\frac{\pi}{4\lambda_2}} \cdot \operatorname{erfc} \left(\frac{\lambda_1}{2\sqrt{\lambda_2}} \right) \right] \quad - \quad - \quad (8.7)$$

where $\operatorname{erfc}(\cdot)$ is the complementary error function, defined by:

$$\operatorname{erfc}(t) = 1 - \operatorname{erf}(t) \quad \text{and} \quad \operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-u^2} du$$

To proceed with the solution of the above equations, we use the procedure suggested by Tribus, (1969).

Defining:

$$t = \frac{\lambda_1}{2\sqrt{\lambda_2}} \quad ; \quad z = z(t) = 2t - \frac{2}{\sqrt{\pi}} \cdot \frac{e^{-t^2}}{[1 - \operatorname{erf}(t)]} \quad - \quad - \quad - \quad (8.8)$$

It is easy to show that:

$$\frac{\partial \lambda_0}{\partial \lambda_1} = \left(\frac{\partial t}{\partial \lambda_1} \right) \cdot z(t) \quad \text{and} \quad \frac{\partial \lambda_0}{\partial \lambda_2} = \left(\frac{\partial t}{\partial \lambda_2} \right) \cdot z(t) - \frac{1}{2\lambda_2}$$

and from (8.8) we have:

$$\frac{\partial t}{\partial \lambda_1} = \frac{1}{2\sqrt{\lambda_2}} \quad \frac{\partial t}{\partial \lambda_2} = -\frac{t}{2\lambda_2}$$

Taking the above into (8.6) we obtain:

$$\lambda_1 m = 1 - 2 \cdot \lambda_2 \cdot m^2 \cdot (Q^2 + 1) \quad (8.9)$$

where $Q = v/m$ ($m > 0$), is the coefficient of variation of X .

If we now introduce the variables α and β , defined as:

$$\alpha = \lambda_1 m \quad \text{and} \quad \beta = \sqrt{\lambda_2} m, \quad \text{we obtain:}$$

$$\text{from (8.6):} \quad \beta = -z(t)/2 \quad (8.10)$$

$$\text{from (8.8):} \quad \alpha = -z(t) \cdot t \quad (8.11)$$

$$\text{from (8.9):} \quad Q^2 = (1 - \alpha)/2\beta^2 - 1 \quad (8.12)$$

Since the coefficient of variation of X is defined on the interval $[0,1]$, and in (8.12) we have Q as a function of t , we can construct a table relating $Q(t) \times t$, instead of analytically solving the equation for t given Q . In doing so, the solution to the problem is straightforward as summarized below:

Given m & v^2 , calculate:

- . $Q = v/m$
- . t from $Q(t) \times t$
- . $z(t)$ from (8.8)
- . β and α from (8.10) and (8.11)
- . $\lambda_1 = \alpha/m$
- . $\lambda_2 = \beta^2/m^2$
- . λ_0 from (8.7)

8.2.2) Properties

If we consider for a moment the corresponding untruncated normal distribution for X and since we are only taking into account the truncation at zero, we could characterise the truncated distribution in terms of the percentage of truncation on the untruncated normal. Let us consider the three cases where the truncation is less than, equal to and greater than 50%, and study the behaviour of the functions defined in sub-section 8.2.1 .

First, from (8.3) it is clear that since $\sigma^2 \geq 0$, then:
 $\lambda_2 \geq 0$. Also, $\beta \geq 0$ because $m \geq 0$ for truncation at zero.

(i) Truncation = 50% ; $\mu_N = 0$

In this case we have:

$$\text{from (8.3) : } \lambda_1 = 0 \quad ; \quad \alpha = 0$$

$$\text{from (8.8) : } t = 0 \quad ; \quad z = -2/\sqrt{\pi}$$

$$\text{from (8.10): } \beta = 1/\sqrt{\pi}$$

$$\text{from (8.12): } Q = \sqrt{\frac{\pi}{2}} - 1 \Rightarrow Q_0 = 0.76$$

(ii) Truncation < 50% ; $\mu_N > 0$

from (8.3) : $\lambda_1 < 0$; $\alpha < 0$

from (8.8) : $t < 0$

from (8.12) : $Q < Q_0$

and as $Q \rightarrow 0$ the distribution tends to the normal
untruncated, with $\mu = m$ & $\sigma^2 = v^2$

(iii) Truncated > 50% , $\mu_N < 0$

from (8.3) : $\lambda_1 > 0$; $\alpha > 0$

from (8.8) : $t > 0$

from (8.12): $Q > Q_0$

and as $Q \rightarrow 1$ the distribution tends to the exponential
with parameter $m = v$.

In figures 8.1 and 8.2 we illustrate α and β as a function of t for reference. The corresponding table of values can be found in Tribus, (1969).

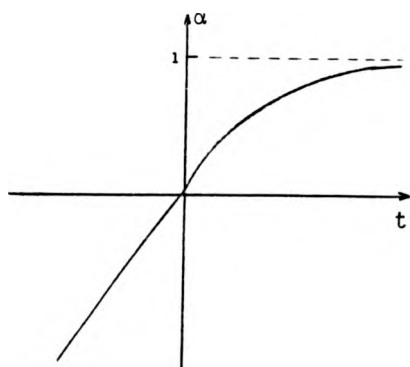


Figure 8.1 : $\alpha \times t$ curve

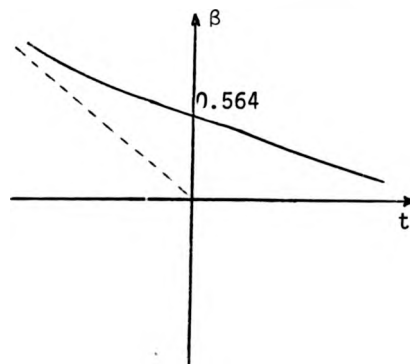


Figure 8.2 : $\beta \times t$ curve

To conclude this section, we show in figure (8.3) below the variation of the coefficient of variation Q with t , and in it the three regions (1), (2) and (3) ; meaning respectively :

Region 1 : $Q_N \leq Q \leq Q_0$

Region where exist a truncation always less than 50% and greater than or equal to $100\epsilon\%$ (eg: $\epsilon = 0.005 \Rightarrow Q_N \approx 0.39$)

Region 2 : $Q < Q_N$

Region where the maximum truncation is very small (less than $100\epsilon\%$), implying that a normal untruncated distribution is the best fit.

Region 3 : $Q > Q_0$

Region where there exists a truncation of at least 50%. As the percentage of truncation increases (or Q approaches 1), the distribution goes over to exponential.

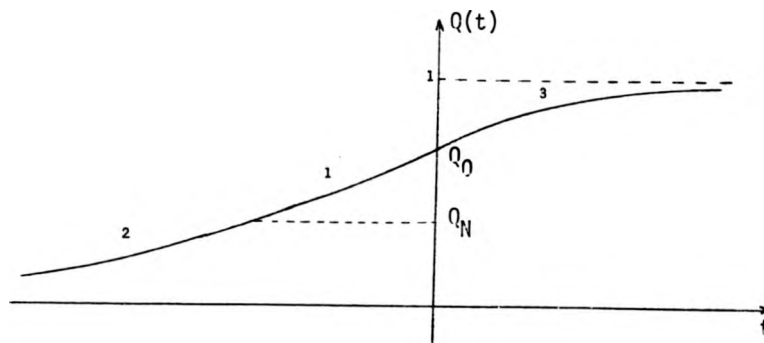


Figure 8.3 : $Q(t) \times t$.

As a final remark, it is clear that our objective in this chapter is to set our BEF for situations where the distributions involved are those lying in Region 1 mostly, that is, for the cases where neither an exponential nor a normal untruncated model is adequate (Q in Regions 2 & 3).

8.3) Entropy of the Truncated Normal Variate:

In order to be able to use our BEF for the truncated normal system, we first have to find the expression of the Shannon's entropy for a truncated normal variate (and its corresponding e-transform uncertainty function), and check whether it matches the basic assumption (ii) of section 4.5.2 .

Let us assume that $X \in \mathbb{R}^+$ is a continuous r.v. with a truncated normal distribution (truncation point at $x=0$), with parameters and pdf as described in section 8.2 . The Shannon's entropy of X can be easily obtained if we make use of the results given in appendix A.

For that, let us consider the pdf of X as given by equation B.1 . Then, if we define:

$$\theta_1 = \lambda_1 \quad \text{and} \quad \theta_2 = \lambda_2 \quad ,$$

we can write for the other functions:

$$K_1(x) = x \quad ; \quad K_2(x) = x^2$$

$$A_1(\theta_1) = A_1(\lambda_1) = -\lambda_1 \quad ; \quad A_2(\theta_2) = A_2(\lambda_2) = -\lambda_2$$

$$Q = -\lambda_0 \quad ; \quad S(X)=0$$

and the corresponding derivatives:

$$\frac{\partial A_1}{\partial \theta_1} = -1 \quad ; \quad \frac{\partial A_2}{\partial \theta_2} = -1$$

$$\frac{\partial Q}{\partial \theta_1} = -\frac{\partial \lambda_0}{\partial \lambda_1} = m \quad (\text{from 8.6})$$

$$\frac{\partial Q}{\partial \theta_2} = -\frac{\partial \lambda_0}{\partial \lambda_2} = m^2 + v^2 \quad (\text{from 8.6})$$

Taking these values into expression (A.3) of appendix A, we obtain for H_X the following expression:

$$H_X = \lambda_1 m + \lambda_2 (m^2 + v^2) + \lambda_0$$

If we now substitute $\lambda_1 m$ for its equivalent expression as given by equation 8.9 with Q substituted by v/m , we obtain:

$$H_X = 1 - \lambda_2 \cdot (m^2 + v^2) + \lambda_0 \quad - \quad - \quad - \quad - \quad - \quad (8.13)$$

An alternative expression for H_X above, in terms of the moments of the untruncated distribution (μ & σ^2), can be obtained by straight substitution of m & v^2 in (8.13) by their equivalent equations (8.4) and (8.5). We obtain:

$$H_X = \frac{1}{2} \left[1 - \frac{1}{M} \left(\frac{\mu}{\sigma} \right) - \left(\frac{\mu}{\sigma} \right)^2 \right] + \lambda_0 \quad - \quad - \quad - \quad - \quad - \quad (8.14)$$

where $M = M(-\mu/\sigma)$ is the Mill's ratio as defined in (8.5).

In order to be able to formulate our BEF model for this truncated normal model, we next have to show that H_X as defined in (8.13) or (8.14) is well defined in \mathbb{R} . However, it is not straightforward to show this, either from (8.13) or (8.14). If for instance we concentrate on (8.14) for a moment, we can clearly see that since $\sigma \in \mathbb{R}^+$, $\mu \in \mathbb{R}$ and $m \in \mathbb{R}^+$, we cannot still guarantee that $H_X \in \mathbb{R}$ because of the presence of λ_0 . On the other hand, if we could show that the entropy decreases with the truncation point then, it is quite clear that the limiting value for the H_X would be the entropy of the normal untruncated distribution which is well defined in \mathbb{R} . That is true, for, if we had a truncation less than 100ε% (ε very small), then $m \rightarrow \mu$, $v^2 \rightarrow \sigma^2$ and $H_X \rightarrow \ln \sqrt{2\pi e \sigma^2} \in \mathbb{R}$

Theorem :

The Shannon's entropy of a truncated normal variate is a decreasing function of the truncation point (see figure 8.4)

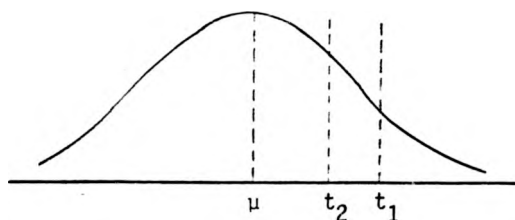


Figure 8.4 : Theorem illustration;
truncation points t_1, t_2
entropies H_1, H_2 ; $t_1 > t_2 \Rightarrow H_1 < H_2$

Proof :

The proof can be made easier if, instead of considering the truncation point variable, we consider it fixed at zero and have the untruncated mean variable. In other words, assuming σ constant and μ variable, we have to show that H_X is an increasing function of μ .

Let H_X be as given in (8.14).

The derivative of H_X with respect to μ is:

$$\frac{\partial H_X}{\partial \mu} = \frac{1}{2} \left[-\frac{\mu}{\sigma} \cdot \frac{\partial M^{-1}}{\partial \mu} - \frac{1}{M\sigma} + 2 \frac{\mu}{\sigma^2} \right] + \frac{\partial \lambda_0}{\partial \mu}$$

From the definition of $M = M(-\mu/\sigma)$ it is not difficult to show that:

$$\frac{\partial M^{-1}}{\partial \mu} = -\frac{1}{\sigma} \left[\frac{\mu}{\sigma} \cdot \frac{1}{M} + \frac{1}{M^2} \right]$$

And from (8.7), we have for $\partial \lambda_0 / \partial \mu$:

$$\frac{\partial \lambda_0}{\partial \mu} = \frac{2 \cdot \lambda_1 \cdot \partial \lambda_1 / \partial \mu}{4 \lambda_2} - \frac{2}{\Pi} \cdot e^{-\lambda_1^2 / 4 \lambda_2} \cdot \frac{1}{\operatorname{erfc}(\lambda_1 / 2\sqrt{\lambda_2})}$$

From (8.3) : $\frac{\partial \lambda_1}{\partial \mu} = -2\lambda_2 = -\frac{1}{\sigma^2}$, and substitution for λ_1 and λ_2 gives:

$$\frac{\partial \lambda_0}{\partial \mu} = \frac{\mu}{\sigma^2} + \frac{2}{\sqrt{\Pi}} \cdot e^{-\mu^2 / 2\sigma^2} \cdot \frac{1}{\sqrt{2} \cdot \sigma} \cdot \frac{1}{\operatorname{erfc}(-\mu)} = \frac{\mu}{\sigma^2} + \frac{1}{\sigma} \cdot \frac{\phi(-\mu/\sigma)}{[1 - \Phi(-\mu/\sigma)]}$$

$$\therefore \frac{\partial \lambda_0}{\partial \mu} = \frac{1}{\sigma} \cdot \left(\frac{\mu}{\sigma} + \frac{1}{M} \right)$$

Taking $\frac{\partial M^{-1}}{\partial \mu}$ and $\frac{\partial \lambda_0}{\partial \mu}$ into the expression for $\frac{\partial H_X}{\partial \mu}$ we obtain:

$$\frac{\partial H_X}{\partial \mu} = \frac{1}{2} \left[-\frac{\mu}{\sigma^2} \left(\frac{\mu}{\sigma} \cdot \frac{1}{M} + \frac{1}{M^2} \right) + \frac{1}{\sigma M} + 2 \frac{\mu}{\sigma^2} \right] + \frac{1}{\sigma} \left(\frac{\mu}{\sigma} + \frac{1}{M} \right)$$

and after simplifications we obtain:

$$\frac{\partial H_X}{\partial \mu} = \frac{1}{2\sigma M} \left[1 + \frac{\mu}{\sigma} \left(\frac{\mu}{\sigma} + \frac{1}{M} \right) \right]$$

Let us now consider the two possible cases:

$\mu > 0$ and $\mu < 0$ and study the corresponding variation on

$\partial H_X / \partial \mu$:

(i) $\mu < 0$

In this case, since σ & M ($-\mu/\sigma$) > 0 ; $\partial H(X)/\partial \mu$ is trivially positive.

(ii) $\mu > 0$

Define $y = -\mu/\sigma < 0$

Then (8.15) can be written as:

$$\frac{\partial H_X}{\partial \mu} = \frac{1}{2\sigma M} \left[|y|^2 - \frac{|y|}{M} + 1 \right]$$

since σ and M are by definition greater than or equal to zero, our only problem lies with the equation into brackets.

Defining: $F(|y|) = |y|^2 - \frac{|y|}{M} + 1$, we have:

$F(|y|) \geq 0$ and consequently:

$$M \geq \frac{|y|}{1+(y)^2}$$

The above is true if and only if $M \geq \frac{1}{2}$ for all $y < 0$ (i.e., $\mu > 0$).

However, since $1 - \Phi(z) \geq 1/2$ and $\phi(z) \leq \frac{1}{\sqrt{2\pi}}$

for $z < 0$, we can use the definition of M as given in equation 8.5 to show that:

$$M \geq \frac{1}{2} \cdot \sqrt{2\pi} \approx 1.25 > 1/2$$

and the theorem follows.

It is now clear that $H_X \in \mathbb{R}$ and consequently from equation (8.14) the e-transform uncertainty function is given by:

$$S_X = \exp \left\{ \frac{1}{2} \cdot \left[1 - \frac{1}{M} \cdot \left(\frac{\mu}{\sigma} \right) - \left(\frac{\mu}{\sigma} \right)^2 \right] + \lambda_0 \right\} \quad \text{--- (8.15)}$$

8.4) Bayesian Analysis for the Truncated Normal Distribution

Before we proceed with the description of the BEF steady state model, we dedicate this section to a brief Bayesian Analysis of a generic truncated normal model. The objective of this study is mainly related to the posterior and the predictive distributions which we obtain via an approximation procedure (to be used in our BEF model later on).

It is worth mentioning that this particular problem provides a Bayesian method for estimating the parameters of the truncated normal distribution. As we mentioned in section 8.1, the existing literature for this problem contains only classical estimation procedures of all sorts. Possibly, the difficulties in obtaining the posterior is the main reason for the lack of interest in a Bayesian approach to this problem.

8.4.1) Parameter Posterior Distribution

Consider a continuous random variable $Y_t \in \mathbb{R}^+$ such that:

$$Y_t \sim N(\theta, v^2); \text{ truncated at zero for each } t=1,2,\dots$$

Suppose that at time $t-1$ the prior information about θ is given by the distribution:

$$(\theta | D_{t-1}) \sim N(\mu_{t-1}, \sigma_{t-1}^2); \text{ truncated at zero.}$$

Observing $Y_t = y_t$ at time t , we can use Bayes' theorem to obtain for the posterior:

$$p(\theta | D_t) \propto p(\theta | D_{t-1}) \cdot p(y_t | \theta)$$

since:

$$p(\theta | D_{t-1}) \propto \exp\left[-\frac{1}{2\sigma_{t-1}^2} (\theta - \mu_{t-1})^2\right] \quad \text{and}$$

$$p(y_t | \theta) \propto \frac{1}{[1 - \Phi(-\theta/v)]} \cdot \exp\left[-\frac{1}{2v^2} (y_t - \theta)^2\right],$$

we then have for the posterior:

$$p(\theta | D_t) \propto \exp\left[-\frac{1}{2\tau_t^2} (\theta - \xi_t)^2\right] \frac{1}{[1 - \Phi(-\theta/v)]} \quad \text{-----(8.16)}$$

where:

$$\xi_t = (\mu_{t-1} v^2 + y_t \sigma_{t-1}^2) / (\sigma_{t-1}^2 + v^2)$$

$$\tau_t^2 = \sigma_{t-1}^2 v^2 / (\sigma_{t-1}^2 + v^2)$$

The above pdf for $(\theta|D_t)$, constrained for $(\theta|D_t) \in \mathbb{R}^+$, is a truncated distribution but not quite truncated normal, due to the factor $1/[1-\Phi(-\theta/v)]$. (It would be exactly truncated normal if we had $1/[1-\Phi(-\xi_t/\tau_t)]$ instead). However, for all $\theta \in \mathbb{R}^+$ and $v > 0$, $1/[1-\Phi(-\theta/v)]$ is a monotonic decreasing function of θ , entirely defined on the interval $[1,2]$, i.e.:

$$\frac{1}{[1-\Phi(-\theta/v)]} \Big|_{\theta=0} = 2 \quad \text{and} \quad \lim_{\theta \rightarrow \infty} \frac{1}{[1-\Phi(-\theta/v)]} = 1$$

From the above, we can see that the effect of $1/[1-\Phi(-\theta/v)]$ on the exponential term of (8.16) is not accentuated, suggesting that $p(\theta|D_t)$ is nearly truncated normal with truncated parameters ξ_t & τ_t^2 . In fact, we could approximate $p(\theta|D_t)$ by a truncated normal distribution if we expanded $\ln\{1/[1-\Phi(-\theta/v)]\}$ for θ around ξ_t , up to the quadratic term. The expansion thus obtained, when substituted in (8.16), gives exponential terms in θ and θ^2 and consequently, a truncated normal distribution for $(\theta|D_t)$.

The above mentioned Taylor expansion for $\ln\{1/[1-\Phi(-\theta/v)]\}$ gives:

$$\ln\{1/[1-\Phi(-\theta/v)]\} \sim F_1 \cdot (\theta - \xi_t) + \frac{F_2}{2} (\theta - \xi_t)^2$$

where:

$$F_1 = -\frac{1}{v \cdot M(-\xi_t/v)} ; F_2 = \frac{\xi_t}{v^3 \cdot M(-\xi_t/v)} + \frac{1}{v^2 \cdot M^2(-\xi_t/v)} \quad \text{--- (8.17)}$$

and $M(\cdot)$ is the Mill's ratio, as defined in (8.5).

Taking this expansion into (8.16), we obtain an approximate truncated normal distribution for the posterior $p(\theta|D_t)$.

By way of illustration, we show in table G.1 of appendix G the results of a simulation for the above problem, with the objective of comparing the true and approximate distributions. We considered three possible degrees of truncation on the prior (with $\sigma^2=1$ for all of them), and in each case we calculate the posterior mean and variance (true and approximation) for $y_t=0,1,2,3$ and $v^2=2$. The close agreement of the true posterior mean and variance to the corresponding approximated mean and variance is quite remarkable, even for the unlikely cases of high truncation on the prior and low y_t 's. For these cases, the posterior is highly truncated and as we have commented before, an exponential approximation would suit better. For example, for the 95% truncation on the prior ($\mu = -1.6452$) and $y_t=0$, the obtained posterior is approximately 98% truncated and yet the approximation is still very good. In fact, for this particular case, the coefficient of variation of the posterior is ~ 0.91 : according to the results of section 8.2, this corresponds closely to an exponential distribution, i.e., $Q \sim 0.91 \gg Q_0$ lies in Region 3 of figure 8.3. These results are indeed very encouraging and reduce tremendously the complexity involved in the Bayesian analysis for the truncated normal model.

where:

$$F_1 = -\frac{1}{v \cdot M(-\xi_t/v)} \quad ; \quad F_2 = \frac{\xi_t}{v^3 \cdot M(-\xi_t/v)} + \frac{1}{v^2 \cdot M^2(-\xi_t/v)} \quad \dots \quad (8.17)$$

and $M(\cdot)$ is the Mill's ratio, as defined in (8.5).

Taking this expansion into (8.16), we obtain an approximate truncated normal distribution for the posterior $p(\theta|D_t)$.

By way of illustration, we show in table G.1 of appendix G the results of a simulation for the above problem, with the objective of comparing the true and approximate distributions. We considered three possible degrees of truncation on the prior (with $\sigma^2=1$ for all of them), and in each case we calculate the posterior mean and variance (true and approximation) for $y_t=0,1,2,3$ and $v^2=2$. The close agreement of the true posterior mean and variance to the corresponding approximated mean and variance is quite remarkable, even for the unlikely cases of high truncation on the prior and low y_t 's. For these cases, the posterior is highly truncated and as we have commented before, an exponential approximation would suit better. For example, for the 95% truncation on the prior ($\mu = -1.6452$) and $y_t=0$, the obtained posterior is approximately 98% truncated and yet the approximation is still very good. In fact, for this particular case, the coefficient of variation of the posterior is ~ 0.91 : according to the results of section 8.2, this corresponds closely to an exponential distribution, i.e., $Q \sim 0.91 \gg Q_0$ lies in Region 3 of figure 8.3. These results are indeed very encouraging and reduce tremendously the complexity involved in the Bayesian analysis for the truncated normal model.

8.4.2) Predictive Distribution

Suppose now that, for the static model we have been considering we want to find a predictive distribution for Y_t , given the information up to time $t-1$, i.e., we want the predictive distribution $p(Y_t | D_{t-1})$. In this case, following the procedure of chapter 4, this predictive distribution can be obtained by integrating out the parameter θ in the joint distribution $p(Y_t, \theta | D_{t-1})$:

$$p(Y_t | D_{t-1}) = \int_{\mathbb{R}^+} p(Y_t, \theta | D_{t-1}) \cdot d\theta$$

where: $p(Y_t, \theta | D_{t-1}) = p(Y_t | \theta, D_{t-1}) \cdot p(\theta | D_{t-1})$

and: $(Y_t | \theta, D_{t-1}) = (Y_t | \theta) \sim N(\theta, v^2)$; truncated at $(Y_t | \theta) = 0$

$(\theta | D_{t-1}) \sim N(\mu_{t-1}, \sigma_{t-1}^2)$; truncated at $(\theta | D_{t-1}) = 0$

Taking these two pdf's into the above integral, we obtain:

$$p(Y_t | D_{t-1}) \propto \exp\left[-\frac{(Y_t - \mu_{t-1})^2}{2(\sigma_{t-1}^2 + v^2)}\right] \cdot \int_{\mathbb{R}^+} \exp\left[-\frac{1}{2\tau_t^2} (\theta - \xi_t)^2\right] \cdot \frac{1}{[1 - \Phi(-\theta/v)]} \cdot d\theta$$

The solution for the above integral is not easily obtained due to the presence of the term $1/[1 - \Phi(-\theta/v)]$. However, if we use its Taylor expansion as shown in the posterior calculation, we obtain, after rearranging the terms in θ , an integral of the form:

$$\int_{\mathbb{R}^+} e^{-b\theta - a\theta^2} d\theta = \frac{1}{2} \sqrt{\frac{\pi}{a}} \cdot \operatorname{erfc}\left(\frac{b}{2\sqrt{a}}\right) \cdot \exp(b^2/4a)$$

we then obtain for $p(Y_t|D_{t-1})$:

$$p(Y_t|D_{t-1}) \propto \exp\left[-\frac{(Y_t - \mu_{t-1})^2}{2(\sigma_{t-1}^2 + v^2)}\right] \cdot A(\xi_t) \cdot B(\xi_t) \cdot C(\xi_t) \quad (8.18)$$

where:

$$A(\xi_t) = \exp\left[\frac{F_2 \cdot \xi_t^2}{2} - F_1 \xi_t\right] \quad (8.19)$$

$$B(\xi_t) = \frac{1}{\sqrt{\lambda_2}} \cdot \exp\left[\frac{\lambda_1^2}{4 \lambda_2}\right] \cdot \operatorname{erfc}\left[\frac{\lambda_1}{2\sqrt{\lambda_2}}\right] \quad (8.20)$$

$$C(\xi_t) = [1 - \Phi(-\xi_t/v)] \quad (8.21)$$

ξ_t , τ_t^2 , F_1 , F_2 as defined in 8.16 & 8.17, and:

$$\lambda_1 = \xi_t F_2 - \xi_t / \tau_t^2 - F_1 \quad (8.22)$$

$$\lambda_2 = 1/2\tau_t^2 - F_2/2 \quad (8.23)$$

The above pdf for $(Y_t|D_{t-1})$ is again a truncated one, but is not normal and again, the same argument used in the posterior approximation can be used again here. In other words, if we consider:

$$a(\xi_t) = \ln[A(\xi_t)] ; b(\xi_t) = \ln[B(\xi_t)] \quad \text{and} \quad c(\xi_t) = \ln[C(\xi_t)] ,$$

we can expand the functions $a(\xi_t)$, $b(\xi_t)$ and $c(\xi_t)$ in a Taylor series for ξ_t around the prior mode up to the quadratic term. We end up with a quadratic function in ξ_t which is easily convertible to a quadratic exponential function in Y_t by use of 8.16. In this case, we again obtain an approximate truncated normal distribution for the predictive distribution. The above mentioned expansions for $a(\xi_t)$,

$b(\xi_t)$ and $c(\xi_t)$ are derived in appendix S.

To conclude this section, we show in table G.2 another simulation in order to check the goodness of the described approximation for the predictive distribution. We again considered the same five different degrees of truncation on the prior (with $\sigma^2=1$ for all of them), and for each case we calculate the predictive for different values of v^2 ($v^2 = 1,2,3,4$). From the results in table G.2, we can clearly see that the approximation is really satisfactory, even for the unlikely cases of high truncation on the prior and low v^2 .

As a final remark, we would like to point out that in both tables G.1 & G.2, the systematic error appearing in the mean and variance for either case is the consequence of the truncation after the quadratic term in all the Taylor expansions involved. What we call true mean and variance were calculated by use of numerical methods for integration, and for computational reasons greater accuracy proved unattainable especially in calculating the function value at each discrete point. Also, in the predictive distribution calculation, the first integral in θ was solved numerically instead of using the Taylor expansion for $1/[1-\phi(-\theta/v)]$.

8.5) BEF Truncated Normal System; Model Description

With the considerations made in the previous sections of this chapter we now use our BEF model formulation applied to a truncated normal process.

Notation :

At any given time $t=1,2,\dots$, let:

Y_t be the process observation

θ_t be the process parameter (unknown) ;

$(\theta_{t-1} | D_{t-1})$: process parameter posterior at time $t-1$ with
pdf $p_{t-1,t-1}$ (known).

$(\theta_t | D_{t-1})$: process parameter prior at time t with pdf
 $p_{t,t-1}$ (unknown)

$S_{t-1,t-1} = S[(\theta_{t-1} | D_{t-1})]$ given by 8.15

$g(S_{t-1,t-1}) = [1 - \exp(-c S_{t-1,t-1})]^2$; $c \in R^+$

Then:

THE	MODEL
Observation equation:	$(Y_t \theta_t) \sim N(\theta_t, v^2)$; truncated at zero where: $E\{Y_t \theta_t\} = \theta_t + v.M^{-1}(-\theta_t/v)$ $Var\{Y_t \theta_t\} = v^2 [1 - \theta_t v^{-1}.M^{-1}(-\theta_t/v) - M^{-2}(-\theta_t/v)]$ Mode $\{Y_t \theta_t\} = \theta_t$
System equation:	$p_{t,t-1} \propto [p_{t-1,t-1}]^g(S_{t-1,t-1})$

and the process parameter is sequentially updated in time as follows:

Information:

- (i) The process observations are generated according to the model above and $g(\cdot)$ is such that c is supposed known at all times.

(ii) The posterior parameter process distribution at time $t-1$ is assumed to be:

$$(\theta_{t-1} | D_{t-1}) \sim N(\mu_{t-1}; \sigma_{t-1}^2) ; \text{ truncated at zero.}$$

UPDATING	PROCEDURE
<u>Prior time t:</u>	
	$(\theta_t D_{t-1}) \sim N(\mu_t^*, \sigma_t^{*2}) ; \text{ truncated at zero}$
	$\mu_t^* = \mu_{t-1} \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.24)$
	$\sigma_t^{*2} = \sigma_{t-1}^2 / g(S_{t-1}, t-1) \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.25)$
<u>Updating:</u>	
Observing $Y_t=y_t$, $(\theta_t D_t)$ is updated as:	
	$(\theta_t D_t) \sim N(\mu_t, \sigma_t^2) ; \text{ truncated at zero}$
	$\mu_t = -\lambda_1 / 2 \lambda_2 \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.26)$
	$\sigma_t^2 = 1/2 \cdot \lambda_2 \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.27)$
where:	
	$\lambda_1 = \xi_t \cdot F_2 - \xi_t / \tau_t^2 - F_1 \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.28)$
	$\lambda_2 = 1/2 \cdot \tau_t^2 - F_2 / 2 \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.29)$
ξ_t , τ_t^2 , F_1 and F_2 as defined in (8.16) & (8.17), with μ_t^* and σ_t^{*2} in place of μ_{t-1} and σ_{t-1}^2 respectively.	

To obtain the last step of our BEF formulation, i.e., the prediction of future observations, we use the results of appendix B for the observation prediction distribution, as schematically described below:

PREDICTIVE j steps ahead $j=1,2,\dots,\ell$

Parameter : $(\theta_{t+j}|D_t)$; $j=1,2,\dots,\ell$

$$(\theta_{t+j}|D_t) \sim N(\mu_{t+j}^* ; \sigma_{t+j}^{*2}) ; \text{truncated at zero.}$$

where, for $j=2,3,\dots,\ell$

$$\mu_{t+j}^* = \mu_{t+j-1}^* \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.30)$$

$$\sigma_{t+j-1}^{*2} = \sigma_{t+j-1}^{*2} / g(S_{t+j-1,t}) \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.31)$$

$$S_{t+j-1,t}^* = S[(\theta_{t+j-1}|D_t)]$$

and for $j=1$ as in equations (8.24) & (8.25) with $t \rightarrow t+1$.

Observation : $(Y_{t+j}|D_t)$; $j=1,2,\dots,\ell$

$$(Y_{t+j}|D_t) \sim N(\mu_{Y_{t+j}} ; \sigma_{Y_{t+j}}^2) ; \text{truncated at zero}$$

$$\mu_{Y_{t+j}} = - \lambda P_{t+j}^{(1)} / (2 \cdot \lambda P_{t+j}^{(2)}) \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.32)$$

$$\sigma_{Y_{t+j}}^2 = 1 / (2 \cdot \lambda P_{t+j}^{(2)}) \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (8.33)$$

where $\lambda P_{t+j}^{(1)}$ and $\lambda P_{t+j}^{(2)}$ are respectively λP_1 and λP_2 of equations (B.32) and (B.33), with $\mu \rightarrow \mu_{t+j}^*$ and $\sigma^2 \rightarrow \sigma_{t+j}^{*2}$.

8.6) Applications:

We finish this chapter by showing a few practical numerical results obtained by the application of the truncated normal BEF model just described.

As a first example, we consider the 336 truncated normal observations shown in table G.3. They correspond to generated data whose truncated normal parameters are fixed and equal to:

$$(Y_t | \theta, v^2) \sim N(\theta, v^2) ; Y_t \in \mathbb{R}^+, \text{ where:}$$

- (i) Mean of the untruncated distribution $\theta = 2.4$
- (ii) Variance of the untruncated distribution $v^2 = 4$
- (iii) Truncation point at $Y_t = 0$; percentage of truncation approximately 12%.

The objectives in analysing this set of data are twofold: firstly to provide a numerical check of the approximation for the posterior and secondly, to check the model itself and its consistency. Let us assume that for the data of table G.3 we know the parameter $v^2 = 4$ and we want an estimate for the parameter θ by following a Bayesian argument. From what we have seen in section 8.4, if we assumed a truncated normal prior for θ , the posterior obtained is truncated but not normal due to the factor $1 / [1 - \Phi(-\theta/v)]$ in the posterior pdf. We are proposing in this chapter a Taylor expansion for this factor in order to bring the truncated posterior back into a normal form. We could also use the results of chapter 3 and obtain for the parameter posterior the least prejudiced distribution at each time point, by performing some numerical integration in the original posterior.

Assuming for both cases: $(\theta|D_0) \sim N(2,6)$; $\theta \in \mathbb{R}^+$ and $v^2=4$, we show in tables G.4 and G.5 the results concerning the Bayesian sequential estimation for θ , where:

- (i) In table G.4 the parameter posterior distribution corresponds to the approximation described in section 8.4 .
- (ii) In table G.5 the parameter posterior distribution is the least prejudiced distribution satisfying the constraints obtained through the true posterior by a numerical integration.

As we can see, in either case the $\text{Mode}(\theta|D_t)$ converges to the true θ ($\theta=2.4$) and the low variance (0.0171) is a clear indication of the certainty about this estimated value after 336 observations. It also emphasizes the goodness of the approximation not only in accuracy but also in processing time: on the University of Warwick's Burroughs B6700, the process time spent for processing the Bayesian analysis of the 336 observations was approximately 11 seconds under (i), and 158 seconds under (ii).

If we now assume that for the same data of table G.3 we have a steady state model instead of a static model (θ is a time dependent parameter θ_t), and used the truncated normal BEF model of section 8.5 , we should expect that if the formulation is consistent, it should give a negative response to the steady state assumption. As usual, we start by estimating the constant c of the function $g(S_{t,t})$. The results, presented in table G.6 give $\hat{c} = 11.25$ and the corresponding aggregate likelihood equal to 56.204 . It is interesting to notice

that if we had considered the static assumption from the very beginning ($g(S_{t,t})=1$ for all $t=1,2,\dots$), the aggregate likelihood obtained is:

$$\sum_{t=1}^{336} \ln \left[p(Y_t | D_{t-1}; g(\cdot)=1) \right] \sim 56.203 .$$

This evidently shows that the initial assumption of a steady state model is wrong, i.e., a static model for θ_t is the true model. As a matter of illustration, we show in table G.7 the results obtained by the BEF with $\hat{c} = 11.25$ for the last seven time points.

As a final illustration, we consider the application of our truncated normal BEF model to the data shown in table G.8 and figure G.1. They correspond to the weekly sales figures for children shoes, model S225/7, covering the period from 19/8/1966 to 28/11/1969 (157 observations), obtained from SATRO (Shoe & Allied Trades Research Association). This particular dataset is in fact an exaggeration of what we have mentioned about the misuse of a normal model. As we can see, they show a pretty unstable pattern, with short steady periods of low sales followed by unexplainable high valued observations.

It is then clear that if a steady model is to be assigned to these data, a truncated normal one should clearly be the recommended one. To show that we applied both; the steady state normal and truncated normal models to the data of table G.8. First, to have an idea of the process observation variance we made use of the simple procedure described by Harrison & Stevens (1976a) for estimation of v and W (DLM formulation see chapter 4) from the given data. We obtain $v \sim 5$ and

$v/W \sim 0.2$: from the kind of data we have, these seem to be reasonable estimates. As a matter of comparison, we adopt the same $v^2=25$ for the truncated normal model. The starting values for the process parameter adopted was: $(\theta_1|D_0) \sim N(8, 16)$ with $(\theta_1|D_0) \in R^+$ for the truncated normal model.

The results concerning the estimation for c of $g(S_{t,t})$ are shown in table G.9. The very low value for c obtained ; $\hat{c} = 0.09$ among other things, indicates a high degree of uncertainty present in the data. In table G.10 we show the results of the predictive distribution obtained through the truncated normal model and in table G.11 the corresponding predictive distribution obtained by the normal model. If we compare the two tables we can clearly see that the predictive distribution in G.11 has not only a higher variance nearly all the time, but also shows an average of 25% truncation. It is interesting to notice that the mode of $p(Y_t|D_{t-1})$ in G.10 is very close to the corresponding expectation $E(Y_t|D_{t-1})$ in G.11, indicating that if a single figure forecast were to be made we would have nearly the same value from either model. However, for decision purposes where, rather than a single figure we need the whole distribution, it is quite obvious that the truncated normal model offers better results.

Finally, these two simple examples not only illustrate the practical aspects of the implementation of the model itself, but also the importance of the steady state truncated normal BEF model as a complement to the corresponding steady state normal model.

APPENDIX A :

Shannon's Entropy for the Exponential Class of Density Functions.

We describe in this appendix a useful formula for the calculation of Shannon's entropy for distributions belonging to a sub-class of the regular case of the exponential family of pdf's. Although this result is not general as we are going to see later, it is still quite useful in our present work since all the distributions we are dealing with belong to this constrained class. We shall borrow Hogg & Craig's, (1970) notation throughout.

We define the exponential family of pdf β as:

$$\beta = \{f(x, \theta); \theta \in \Theta; \theta \text{ m-vector}; \theta \in \mathbb{R}^m; a < x < b\}, \quad \text{whose}$$

pdf's $f(x; \theta)$ or $f(x; \theta_1, \dots, \theta_m)$ is given by:

$$f(x; \theta_1, \dots, \theta_m) = \exp \left\{ \sum_{j=1}^m A_j(\theta_1, \dots, \theta_m) \cdot K_j(x) + Q(\theta_1, \dots, \theta_m) + S(x) \right\} \quad \text{--- (A.1)}$$

If in addition we have:

- i) a, b do not depend upon $\theta_i; i=1, 2, \dots, m$
- ii) $A_j(\theta_1, \dots, \theta_m)$ are non trivial, functionally independent and continuous functions of $\theta_j; j=1, 2, \dots, m$
- iii) $K_j(x); j=1, 2, \dots, m$ are continuous for $a < x < b$ and no one is a linear homogeneous function of the others.
- iv) $S(x)$ is a continuous function of $x; a < x < b$.

then (A.1) is called a regular case of the exponential family.

We now consider the family $A; A \in \beta$, where A is defined as β except for the A_j functions that are supposed to have the single form:

$$A_j(\theta_1, \dots, \theta_m) = A_j(\theta_j) \quad \text{for } j=1, 2, \dots, m \quad \text{--- (A.2)}$$

Theorem:

Shannon's entropy for the pdf's that belongs to the family A of probability densities is given by:

$$H(f) = \sum_{j=1}^m \frac{A_j(\theta_j) \cdot \partial Q / \partial \theta_j}{\partial A_j(\theta_j) / \partial \theta_j} - Q(\theta_1, \dots, \theta_m) - E_f[S(x)] \quad \text{--- (A.3)}$$

provided $Q(\theta_1, \dots, \theta_m)$ is differentiable with respect to all θ_j ; $j=1, 2, \dots, m$.

Proof:

From (A.1), (A.2) and (2.2) of chapter 2, we can write for $H(f)$:

$$H(f) = - E_f \{ \ln f(x; \theta_1, \dots, \theta_m) \} = - E_f \left\{ \sum_{j=1}^m A_j(\theta_j) \cdot K_j(x) + Q(\theta_1, \dots, \theta_m) + S(x) \right\}$$

or:

$$H(f) = - \sum_{j=1}^m A_j(\theta_j) \cdot E_f [K_j(x)] - Q(\theta_1, \dots, \theta_m) - E_f [S(x)] \quad \text{--- (A.4)}$$

In order to calculate $E_f [K_j(x)]$ of (A.4), let us consider the identity below:

$$\int_a^b \exp \left[\sum_{j=1}^m A_j(\theta_j) \cdot K_j(x) + Q(\theta_1, \dots, \theta_m) + S(x) \right] \cdot dx = 1$$

if we differentiate the above identity with respect to θ_i we obtain

$$\frac{\partial}{\partial \theta_i} \int_a^b \exp \left[\sum_{j=1}^m A_j(\theta_j) \cdot K_j(x) + Q(\theta_1, \dots, \theta_m) + S(x) \right] \cdot dx = 0$$

or, by proceeding with the differentiations we obtain after simplifications:

$$\int_a^b \left[K_i(x) \cdot \frac{\partial A_i(\theta_i)}{\partial \theta_i} + \frac{\partial Q(\theta_1, \dots, \theta_m)}{\partial \theta_i} \right] \cdot \exp \left[\sum_{j=1}^m A_j(\theta_j) \cdot K_j(x) + Q(\theta_1, \dots, \theta_m) + S(x) \right] \cdot dx = 0$$

Since the exponential term on the left hand side of the above is from (A.1) equal to $f(x; \theta_1, \dots, \theta_m)$, we can write:

$$\frac{\partial A_i(\theta_i)}{\partial \theta_i} \int_a^b K_i(x) \cdot f(x; \theta_1, \dots, \theta_m) \cdot dx = - \frac{\partial Q(\theta_1, \dots, \theta_m)}{\partial \theta_i} \int_a^b f(x; \theta_1, \dots, \theta_m) \cdot dx$$

However:

$$\int_a^b K_i(x) \cdot f(x; \theta_1, \dots, \theta_m) \cdot dx = E_f [K_i(x)]$$

and

$$\int_a^b f(x; \theta_1, \dots, \theta_m) \cdot dx = 1$$

We finally obtain:

$$E_f [K_i(x)] = - [\partial Q(\theta_1, \dots, \theta_m) / \partial \theta_i] / [\partial A_i(\theta_i) / \partial \theta_i] ;$$

$$i = 1, 2, \dots, m$$

Then, taking the above expectation into (A.4) we obtain (A.3) and the proof follows.

As a final remark, the term $E_f [S(x)]$ that appears in (A.3) could in principle be a barrier for its use. However, in many cases $S(x) = 0$ or $S(x)$ is a particular function such that $E_f [S(x)]$ is easily obtained.

APPENDIX B:

Approximation for the Predictive Distribution of the Truncated Normal - Model.

In this Appendix, we show the main calculations involved in the approximating distribution for $(Y_t | D_{t-1})$ of section 8.4.2. As we know, we can approximate $p(Y_t | D_{t-1})$ for a truncated normal distribution, by expanding $a(\xi_t)$, $b(\xi_t)$ and $c(\xi_t)$ - equations (8.19), (8.20) and (8.21) respectively in a Taylor series. These expansions are shown separately in (i), (ii) and (iii) below.

(i) Taylor expansion for $a(\xi_t)$

From (8.19) :

$$a(\xi_t) = \ln A(\xi_t) = \frac{\xi_t^2 F_2}{2} - F_1 \xi_t$$

Then, the first and second derivatives of $a(\xi_t)$ are, respectively:

$$a_1 = \left. \frac{\partial a(\xi_t)}{\partial \xi_t} \right|_{\xi_t = \mu} = \mu F_2 + \frac{\mu^2 F_2'}{2} - \mu F_1' - F_1 \quad - \quad - \quad - \quad (B.1)$$

$$a_2 = \left. \frac{\partial^2 a(\xi_t)}{\partial \xi_t^2} \right|_{\xi_t = \mu} = F_2 + 2\mu F_2' + \frac{\mu^2 F_2''}{2} - 2F_1' - \mu F_1'' \quad - \quad - \quad - \quad (B.2)$$

where, from (8.17) :

$$F_2' = \left. \frac{\partial F_2}{\partial \xi_t} \right|_{\xi_t = \mu} = [(M^{-2})' - M^{-1}/v + \mu (M^{-1})'/v] / v^2 \quad - \quad - \quad - \quad (B.3)$$

$$F_2'' = \left. \frac{\partial^2 F_2}{\partial \xi_t^2} \right|_{\xi_t = \mu} = [(M^{-2})'' + 2(M^{-1})'/v + \mu (M^{-1})''/v] / v^2 \quad - \quad - \quad - \quad (B.4)$$

$$F_1' = - (M^{-1})' / v \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.5)$$

$$F_1'' = - (M^{-1})'' / v \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.6)$$

and:

$M^{-1}(\cdot)$ is the inverse of the Mill's ratio (see equation 8.5) ;

$$M^{-1} = M^{-1}(-\xi_t/v) \Big|_{\xi_t=\mu} = \frac{\phi(-\mu/v)}{[1-\phi(-\mu/v)]} \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.7)$$

$$(M^{-1})' = \frac{\partial}{\partial \xi_t} M^{-1}(-\xi_t/v) \Big|_{\xi_t=\mu} = - \frac{\mu}{v^2} M^{-1} - \frac{1}{v} M^{-2} \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.8)$$

$$(M^{-1})'' = \frac{\partial^2}{\partial \xi_t^2} M^{-1}(-\xi_t/v) \Big|_{\xi_t=\mu} = - \frac{1}{v^2} [M^{-1+\mu}(M^{-1})'] - \frac{1}{v} (M^{-2})' \quad - \quad - \quad - \quad (B.9)$$

$$(M^{-2})' = \frac{\partial}{\partial \xi_t} M^{-2}(-\xi_t/v) \Big|_{\xi_t=\mu} = 2.M^{-1} \cdot (M^{-1})' \quad - \quad - \quad - \quad - \quad (B.10)$$

$$(M^{-2})'' = \frac{\partial^2}{\partial \xi_t^2} M^{-2}(-\xi_t/v) \Big|_{\xi_t=\mu} = 2 [M^{-1}(M^{-1})'' + (M^{-1})',^2] \quad - \quad - \quad (B.11)$$

Taking into account the expression for ξ_t given in (8.15), the final Taylor expansion for $a(\xi_t)$, in terms of Y_t is given by:

$$a(\xi_t) \approx \left[\frac{a_2 \sigma^4}{2(\sigma^2 + v^2)^2} \right] \cdot Y_t^2 + \left[\frac{a_2 \mu \sigma^2 v^2}{(\sigma^2 + v^2)^2} + \frac{(a_1 - a_2 \mu) \sigma^2}{(\sigma^2 + v^2)^2} \right] \cdot Y_t \quad - \quad - \quad - \quad (B.12)$$

Where a_1 and a_2 are as shown in (B.1) and (B.2) and μ & σ^2 are respectively μ_{t-1} & σ_{t-1}^2 (mean and variance of the untruncated prior).

ii) Taylor expansion for $b(\xi_t)$

From (8.20) we can write for $b(\xi_t)$.

$$b(\xi_t) = \ln[B(\xi_t)] = -\frac{1}{2} \ln \lambda_2 + \frac{\lambda_1^2}{4\lambda_2} + \ln \left[\operatorname{erfc} \left(\frac{\lambda_1}{2\sqrt{\lambda_2}} \right) \right]$$

Before we proceed with the derivatives of $b(\xi_t)$, let us define some auxiliary functions and their corresponding derivatives; as follows:

$$\alpha = \lambda_1 / 2\sqrt{\lambda_2} \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.13)$$

$$f(\alpha) = e^{-\alpha^2} / \operatorname{erfc}(\alpha) \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.14)$$

$$\alpha' = \frac{\partial \alpha}{\partial \xi_t} \Big|_{\xi_t = \mu} = \frac{\lambda_1'}{2\sqrt{\lambda_2}} - \frac{\lambda_1 \lambda_2'}{4\sqrt{\lambda_2}^3} \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.15)$$

$$f'(\alpha) = \frac{\partial f(\alpha)}{\partial \xi_t} \Big|_{\xi_t = \mu} = 2\alpha' \left[-2\alpha \cdot f(\alpha) + \frac{f^2(\alpha)}{\sqrt{\pi}} \right] \quad - \quad - \quad - \quad (B.16)$$

$$\beta_1 = \frac{\alpha}{\sqrt{\lambda_2}} - \frac{f(\alpha)}{\sqrt{\pi} \lambda_2} \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.17)$$

$$\beta_2 = -\frac{\alpha^2}{\lambda_2} - \frac{1}{2\lambda_2} + \frac{\alpha}{\sqrt{\pi} \lambda_2} \cdot f(\alpha) \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad (B.18)$$

$$\beta_1' = \left. \frac{\partial \beta_1}{\partial \xi_t} \right|_{\xi_t = \mu} = \frac{(\alpha' - f'(\alpha)/\sqrt{\pi})}{\sqrt{\lambda_2}} + \frac{\lambda_2'}{2\sqrt{\lambda_2^3}} (-\alpha + f(\alpha)/\sqrt{\pi}) \quad \text{--- (B.19)}$$

$$\beta_2' = \left. \frac{\partial \beta_2}{\partial \xi_t} \right|_{\xi_t = \mu} = \frac{\alpha'}{\lambda_2} \left[-2\alpha + f(\alpha)/\sqrt{\pi} \right] + \frac{\lambda_2'}{\lambda_2^2} \left[\frac{1}{2} - \frac{\alpha f(\alpha)}{\sqrt{\pi}} + \alpha^2 \right] + \frac{\alpha f'(\alpha)}{\sqrt{\pi} \lambda_2} \quad \text{(B.20)}$$

and finally, from (8.22) and (8.23)

$$\lambda_1' = \left. \frac{\partial \lambda_1}{\partial \xi_t} \right|_{\xi_t = \mu} = F_2' + \mu F_2' - 1/\tau_t^2 - F_1' \quad \text{--- (B.21)}$$

$$\lambda_1'' = \left. \frac{\partial^2 \lambda_1}{\partial \xi_t^2} \right|_{\xi_t = \mu} = 2 F_2'' + \mu F_2'' - F_1'' \quad \text{--- (B.22)}$$

$$\lambda_2' = \left. \frac{\partial \lambda_2}{\partial \xi_t} \right|_{\xi_t = \mu} = -F_2' / 2 \quad \text{--- (B.23)}$$

$$\lambda_2'' = \left. \frac{\partial^2 \lambda_2}{\partial \xi_t^2} \right|_{\xi_t = \mu} = -F_2'' / 2 \quad \text{--- (B.24)}$$

with F_1' , F_1'' , F_2' and F_2'' as given by equations (B.3) to (B.6).

Using the auxiliary functions (B.13) to (B.24) it is not difficult to show that the first two derivatives of $b(\xi_t)$ are respectively.

$$b_1' = \left. \frac{\partial b(\xi_t)}{\partial \xi_t} \right|_{\xi_t = \mu} = \beta_1 \lambda_1' + \beta_2 \lambda_2' \quad \text{--- (B.25)}$$

$$b_2 = \frac{\partial^2 b(\xi_t)}{\partial \xi_t^2} \Big|_{\xi_t = \mu} = \beta_1' \lambda_1' + \beta_1'' \lambda_1'' + \beta_2' \lambda_2' + \beta_2'' \lambda_2'' \quad \text{--- (B.26)}$$

The above equations (B.25) & (B.26) and (8.16) enable us to write for the Taylor expansion of $b(\xi_t)$:

$$b(\xi_t) \approx \left[\frac{b_2 \sigma^4}{2(\sigma^2 + v^2)^2} \right] \gamma_t^2 + \left[\frac{b_2 \mu \sigma^2 v^2}{(\sigma^2 + v^2)^2} + \frac{(b_1 - b_2 \mu) \sigma^2}{(\sigma^2 + v^2)} \right] \gamma_t \quad \text{--- (B.27)}$$

(iii) Taylor expansion for $c(\xi_t)$

From (8.21), $c(\xi_t)$ can be written as:

$$c(\xi_t) = \ln C(\xi_t) = \ln [1 - \phi(-\xi_t/v)]^{-1}$$

In this case, it is not difficult to show that:

$$c_1 = \frac{\partial c(\xi_t)}{\partial \xi_t} \Big|_{\xi_t = \mu} = -M^{-1}/v \quad \text{--- (B.28)}$$

$$c_2 = \frac{\partial^2 c(\xi_t)}{\partial \xi_t^2} \Big|_{\xi_t = \mu} = \frac{\mu M^{-1}}{v^3} + \frac{M^{-2}}{v^2} \quad \text{--- (B.29)}$$

and then:

$$c(\xi_t) \approx \left[\frac{c_2 \sigma^4}{2(\sigma^2 + v^2)^2} \right] \gamma_t^2 + \left[\frac{c_2 \mu \sigma^2 v^2}{(\sigma^2 + v^2)^2} + \frac{(c_1 - c_2 \mu) \sigma^2}{(\sigma^2 + v^2)} \right] \gamma_t \quad \text{--- (B.30)}$$

Finally, the truncated normal approximation for the predictive distribution can be obtained, by taking expansions (B.12), (B.27) and (B.30) into equation

(8.18), that gives, after simplifications:

$$p(Y_t | D_{t-1}) \sim N(\mu_{p_t}, \sigma_{p_t}^2), \text{ truncated at zero,}$$

where:

$$\mu_{p_t} = -\lambda_{p_1}/2 \cdot \lambda_{p_2} ; \sigma_{p_t}^2 = 1/2 \cdot \lambda_{p_2} \quad - \quad - \quad - \quad - \quad - \quad (B.31)$$

and:

$$\lambda_{p_1} = \frac{1}{(\sigma^2 + v^2)} \left[\frac{\mu \sigma^2 v^2}{\tau_t^2 (\sigma^2 + v^2)} - \mu - \frac{a_2 \mu \sigma^2 v^2}{(\sigma^2 + v^2)} - (a_1 - a_2 \mu) \sigma^2 - \frac{b_2 \mu \sigma^2 v^2}{(\sigma^2 + v^2)} - (b_1 - b_2 \mu) \sigma^2 - \frac{c_2 \mu \sigma^2 v^2}{(\sigma^2 + v^2)} - (c_1 - c_2 \mu) \sigma^2 \right] \quad - \quad - \quad - \quad (B.32)$$

$$\lambda_{p_2} = \frac{1}{2(\sigma^2 + v^2)} \left[1 + \frac{\sigma^4}{(\sigma^2 + v^2)} (\tau_t^{-2} - a_2 - b_2 - c_2) \right] \quad - \quad - \quad (B.33)$$

APPENDIX C :

Numerical results concerning the Non-Additive Normal model simulation of section 4.6 .

C_t	$S_{t,t}$	$h(S_{t,t})$	$g(S_{t,t})$
30.	22.64	0.75	0.71
32.	23.38	0.76	0.73
34.	24.10	0.77	0.74
36.	24.80	0.78	0.76
38.	25.48	0.79	0.77
40.	26.14	0.80	0.78
42.	26.78	0.81	0.79
44.	27.41	0.81	0.80
46.	28.03	0.82	0.81
48.	28.63	0.83	0.82
50.	29.22	0.83	0.83
52.	29.80	0.84	0.83
54.	30.37	0.84	0.84
56.	30.93	0.85	0.85
58.	31.47	0.85	0.85
60.	32.01	0.86	0.86
62.	32.54	0.86	0.87
64.	33.06	0.86	0.87
66.	33.57	0.87	0.88
68.	34.08	0.87	0.88
70.	34.58	0.88	0.89
72.	35.07	0.88	0.89
74.	35.55	0.88	0.89
76.	36.03	0.88	0.90
78.	36.50	0.89	0.90
80.	36.96	0.89	0.91
82.	37.42	0.89	0.91
84.	37.88	0.89	0.91
86.	38.33	0.90	0.92
88.	38.77	0.90	0.92
90.	39.21	0.90	0.92

TABLE C.1 : $g(S_{t,t}) \times h(S_{t,t})$ values.

$S_{t,t}$	$S_{t+1,t}$	
	true	approx.
0.	13.07	12.20
2.	13.22	13.22
4.	13.67	14.30
6.	14.38	15.44
8.	15.32	16.63
10.	16.46	17.87
12.	17.74	19.16
14.	19.15	20.51
16.	20.66	21.90
18.	22.24	23.33
20.	23.89	24.81
22.	25.59	26.34
24.	27.33	27.90
26.	29.10	29.50
28.	30.90	31.13
30.	32.72	32.80
32.	34.57	34.50
34.	36.43	36.23
36.	38.30	37.98
38.	40.18	39.76
40.	42.08	41.56
42.	43.99	43.39
44.	45.90	45.23
46.	47.82	47.08
48.	49.75	48.96
50.	51.68	50.84
52.	53.62	52.74
54.	55.56	54.65
56.	57.50	56.57
58.	59.45	58.50
60.	61.41	60.44
62.	63.36	62.39
64.	65.32	64.34
66.	67.28	66.30
68.	69.24	68.26
70.	71.21	70.23
72.	73.18	72.20
74.	75.15	74.17
76.	77.12	76.15
78.	79.09	78.13
80.	81.06	80.11

TABLE C.2 : $S_{t+1,t}$ values.

Number of Observations	c
100	0.12
200	0.10
300	0.098
450	0.094
600	0.088

TABLE C.3 : "c" estimation by simulated data.

APPENDIX D : POISSON-GAMMA BEF - NUMERICAL RESULTS

This appendix contains tables and plots illustrating the numerical results of the Poisson-Gamma BEF of chapter 5, section 5.5 .

2	7	1	2	6	1	3	1	5	3	2	4	2	1	1	3
1	2	4	0	2	5	3	1	1	3	3	3	2	6	4	1
3	5	2	5	2	2	3	7	4	3	4	7	2	2	5	2
3	3	0	2	2	1	0	2	2	2	1	2	2	3	3	7
3	5	3	2	3	3	4	2	1	3	5	2	5	3	4	1
1	3	6	4	0	9	1	1	3	2	5	3	4	1	5	1
1	3	3	2	1	2	1	5	1	3	3	4	2	3	10	4
5	3	3	4	0	2	4	3	2	4	6	4	2	6	4	0
2	4	3	3	2	3	2	2	2	2	1	4	0	6	1	3
4	3	2	3	3	4	4	1	3	3	6	2	1	4	9	1
0	4	8	6	4	2	4	4	5	5	1	4	5	4	3	3
3	1	1	6	2	0	2	3	3	1	3	2	5	4	0	7
0	1	1	2	2	3	9	2	5	5	2	2	2	11	1	3
3	6	0	3	4	3	4	5	3	1	1	1	5	4	4	6
2	7	3	5	3	1	1	6	2	4	0	4	7	5	1	3
3	4	3	1	6	3	1	3	1	4	4	5	3	4	2	2
3	4	4	3	7	2	2	3	3	6	4	4	5	1	1	3
6	3	6	4	0	2	5	1	2	7	1	2	3	3	4	4
3	1	5	2	2	5	1	2	10	4	3	5	7	5	4	0
4	5	2	3	3	3	5	5	1	1	2	0	1	3	1	2
2	1	1	1	6	2	1	1	4	4	5	3	2	1	5	0
3	4	3	9	2	5	3	3	4	5	5	4	0	4	6	5
2	0	2	2	3	6	1	5	2	4	3	5	1	6	5	2
6	3	2	3	2	2	3	10	6	5	2	4	3	6	6	3
1	2	2	2	3	4	4	2	1	2	3	4	1	5	3	3
5	4	3	0	2	2	6	2	2	4	4	1	2	5	0	2
3	4	4	4	2	5	2	5	3	1	3	4	7	3	2	0
1	4	4	3	3	1	0	7	3	3	1	2	2	5	3	1
6	1	2	4	4	3	2	5	2	6	2	3	1	4	2	8
3	5	5	3	4	1	0	2	5	3	2	2	4	5	3	1
4	2	2	4												

TABLE D.1 : 500 Constant mean Poisson Observations (mean=3)

APPENDIX D : POISSON-GAMMA BEF -- NUMERICAL RESULTS

This appendix contains tables and plots illustrating the numerical results of the Poisson-Gamma BEF of chapter 5, section 5.5

2	7	1	2	6	1	3	1	5	3	2	4	2	1	1	3
1	2	4	0	2	5	3	1	1	3	3	3	2	6	4	1
3	5	2	5	2	2	3	7	4	3	4	7	2	2	5	2
3	3	0	2	2	1	0	2	2	2	1	2	2	5	3	7
3	5	3	2	3	3	4	2	1	3	5	2	5	3	4	1
1	3	6	4	0	9	1	1	3	2	5	3	4	1	5	1
1	3	3	2	1	2	1	5	1	3	3	4	2	3	10	4
5	3	3	4	0	2	4	3	2	4	6	4	2	6	4	0
2	4	3	3	2	3	2	2	2	2	1	4	0	6	1	3
4	3	2	3	3	4	4	1	3	3	6	2	1	4	0	1
0	4	8	6	4	2	4	4	5	5	1	4	5	4	3	3
3	1	1	6	2	0	2	3	3	1	3	2	5	4	0	7
0	1	1	2	2	3	9	2	5	5	2	2	2	11	1	3
3	6	0	3	4	3	4	5	3	1	1	1	5	4	4	6
2	7	3	5	3	1	1	6	2	4	0	4	7	5	1	3
3	4	3	1	6	3	1	3	1	4	4	5	3	4	2	2
3	4	4	3	7	2	2	3	3	6	4	4	5	1	1	3
6	3	6	4	0	2	5	1	2	7	1	2	3	3	4	4
3	1	5	2	2	5	1	2	10	4	3	5	7	5	4	0
4	5	2	3	3	3	5	5	1	1	2	0	1	3	1	2
2	1	1	1	6	2	1	1	4	4	5	3	2	1	5	0
3	4	3	9	2	5	3	3	4	5	5	4	0	4	6	5
2	0	2	2	3	6	1	5	2	4	3	5	1	6	5	2
6	3	2	3	2	2	3	10	6	5	2	4	3	6	6	3
1	2	2	2	3	4	4	2	1	2	3	4	1	5	3	3
5	4	3	0	2	2	6	2	2	4	4	1	2	5	0	2
3	4	4	4	2	5	2	5	3	1	3	4	7	3	2	0
1	4	4	3	3	1	0	7	3	3	1	2	2	5	3	1
6	1	2	4	4	3	2	5	2	6	2	3	1	4	2	8
3	5	5	3	4	1	0	2	5	3	2	2	4	5	3	1
4	2	2	4												

TABLE D.1 : 500 Constant mean Poisson Observations (mean=3)

c	AGG. LIKL.
0.1	0.20076300×10^2
5.0	0.39006000×10^2
10.0	0.39218800×10^2
20.0	0.39244610×10^2
30.0	0.39244736×10^2
35.0	0.39244737×10^2
39.0	0.39244738×10^2
39.2	0.39244738×10^2
39.4	0.39244739×10^2
39.6	0.39244741×10^2
39.8	0.39244740×10^2
40.0	0.39244740×10^2
45.0	0.39244739×10^2

TABLE D.2 : c × Aggregate likelihood from first 250 obs. of table D.1 .

c	AGG. LIKL.
30.0	0.81273659×10^2
35.0	0.81273748×10^2
40.0	0.81273757×10^2
45.0	0.81273759×10^2
49.0	0.81273759×10^2
49.2	0.81273759×10^2
49.4	0.81273760×10^2
49.6	0.81273759×10^2
50.0	0.81273759×10^2
55.0	0.81273758×10^2

TABLE D.3 : c × Aggregate Likelihood from all the obs. of table D.1

Time	$H_{t,t}$	$S_{t,t}$
246	-.7995628E+00	4495255E+00
247	-.8015308E+00	4486417E+00
248	-.8047825E+00	4471852E+00
249	-.8067300E+00	4463151E+00
250	-.8099548E+00	4448782E+00
496	-.1133470E+01	3219143E+00
497	-.1135125E+01	3213819E+00
498	-.1135809E+01	3211621E+00
499	-.1137136E+01	3207364E+00
500	-.1138459E+01	3203122E+00

TABLE D.4 : Entropy values for:

(i) $t=246$ to 250 - Model $\hat{c}=39.6$

(ii) $t=496$ to 500 - Model $\hat{c}=49.4$

$$H_{t,t} = H(\theta_t | D_t) ; S_{t,t} = S[\theta_t | D_t]$$

PRIOR	A=1544.932	B= 503.994	MEAN=	3.0555	VAP.=	0.0061	MODE=	3.0635
PRED.	PI= 0.998	P2=1544.98	MEAN=	3.0555	VAP.=	3.0716		
	TIME 495	*JBS.* 3.						
POST.	A=1547.982	B= 504.994	MEAN=	3.0553	VAP.=	0.0061	MODE=	3.0634
PRIOR	A=1547.932	B= 504.994	MEAN=	3.0553	VAP.=	0.0061	MODE=	3.0534
PRED.	PI= 0.998	P2=1547.98	MEAN=	3.0553	VAP.=	3.0714		
	TIME 496	*JBS.* 1.						
POST.	A=1548.932	B= 505.994	MEAN=	3.0513	VAP.=	0.0061	MODE=	3.0593
PRIOR	A=1548.931	B= 505.994	MEAN=	3.0513	VAP.=	0.0061	MODE=	3.0593
PRED.	PI= 0.998	P2=1548.98	MEAN=	3.0513	VAP.=	3.0673		
	TIME 497	*JBS.* 4.						
POST.	A=1552.981	B= 506.994	MEAN=	3.0531	VAP.=	0.0060	MODE=	3.0611
PRIOR	A=1552.931	B= 506.994	MEAN=	3.0531	VAP.=	0.0060	MODE=	3.0611
PRED.	PI= 0.998	P2=1552.98	MEAN=	3.0531	VAP.=	3.0692		
	TIME 498	*JBS.* 2.						
POST.	A=1554.981	B= 507.994	MEAN=	3.0510	VAP.=	0.0060	MODE=	3.0591
PRIOR	A=1554.931	B= 507.994	MEAN=	3.0510	VAP.=	0.0060	MODE=	3.0591
PRED.	PI= 0.998	P2=1554.98	MEAN=	3.0510	VAP.=	3.0670		
	TIME 499	*JBS.* 2.						
POST.	A=1556.981	B= 508.994	MEAN=	3.0589	VAP.=	0.0060	MODE=	3.0570
PRIOR	A=1556.931	B= 508.994	MEAN=	3.0589	VAP.=	0.0060	MODE=	3.0570
PRED.	PI= 0.998	P2=1556.98	MEAN=	3.0589	VAP.=	3.0649		
	TIME 500	*JBS.* 4.						
POST.	A=1560.981	B= 509.993	MEAN=	3.0508	VAP.=	0.0060	MODE=	3.0588

TABLE D.5: Poisson-Gamma DEF - general output for data of table D.1.

A & B stand for parameters α & β of the gamma distribution.

10	5	10	11	8	5	8	7	7	6	5	4	7	2	5	3
1	4	4	4	3	1	4	3	4	1	1	3	3	1	4	3
1	4	3	0	2	2	4	3	2	5	3	6	0	3	0	6
3	3	5	5	6	6	8	4	7	8	2	7	5	8	8	5
6	15	8	7	5	5	6	1	3	3	3	4	1	1	2	1
0	3	1	1	0	0	3	0	2	0	1	0	0	3	4	2
6	4	2	1	3	3	4	3	2	5	0	1	8	2	7	1
4	5	4	3	5	6	4	7	5	7	8	7	1	0	0	3
3	0	3	1	4	0	1	1	1	1	2	1	2	1	1	0
1	3	6	2	3	4	2	2	3	2	3	2	3	6	5	12
9	5	1	2	9	8	17	14	8	10	4	6	5	3	2	4
2	1	1	1	1	1	1	4	0	2	4	0	2	1	0	1
0	1	3	1	2	3	3									

TABLE D.6 : 199 weekly deaths caused by acute respiratory infections
in Greater London, covering the period from 15/2/72 to
01/10/76 .

c	AGG. LIKL.
25.0	0.2327577×10^2
15.0	0.2327593×10^2
10.0	0.2328679×10^2
5.0	0.2361436×10^2
1.0	0.2631456×10^2
0.8	0.2711050×10^2
0.6	0.2761453×10^2
0.59	0.2762182×10^2
0.58	0.2762271×10^2
0.57	0.2762915×10^2
0.56	0.2762894×10^2
0.55	0.2762594×10^2
0.50	0.2756349×10^2
0.30	0.2583690×10^2
0.10	0.1601014×10^2

TABLE D.7 : c × Aggregate Likelihood 199 data of table D.6 .

Time t	Obs. Y _t	$(\theta_t D_t)$				$(\theta_{t+1} D_t)$			
		Mode	Var.	α_t	β_t	Mode	Var.	α_{t+1}^*	β_{t+1}^*
1	10	5.231	2.038	15.37	2.75	5.231	2.210	14.31	2.55
2	5	5.166	1.537	19.31	3.55	5.166	1.738	17.30	3.15
3	10	6.330	1.582	27.30	4.15	6.330	1.773	24.55	3.72
4	11	7.319	1.595	35.55	4.72	7.319	1.783	32.02	4.24
5	8	7.449	1.459	40.02	5.24	7.449	1.654	35.51	4.63
6	5	7.014	1.277	40.51	5.63	7.014	1.486	35.07	4.86
7	8	7.183	1.255	43.07	5.86	7.183	1.466	37.16	5.04
8	7	7.152	1.213	44.16	6.04	7.152	1.426	37.84	5.15
9	7	7.128	1.185	44.84	6.15	7.128	1.401	38.23	5.22
10	6	6.946	1.142	44.23	6.22	6.946	1.362	37.41	5.24

TABLE D.8 : Posterior and Prior parameter distributions;
t=1 to 10; $\hat{c}=0.57$; $\alpha_0=6$, $\beta_0=2$; 199 weekly
data of table D.6 .

Time t	Obs. Y_t	$(\theta_t D_t)$				$(\theta_{t+1} D_t)$			
		Mode	Var.	α_t	β_t	Mode	Var.	α_{t+1}^*	β_{t+1}^*
165	9	5.445	1.069	29.69	5.27	5.445	1.300	24.76	4.36
166	8	5.921	1.139	32.76	5.36	5.921	1.362	27.71	4.51
167	17	7.932	1.473	44.71	5.51	7.932	1.666	39.73	4.88
168	14	8.963	1.552	53.73	5.88	8.963	1.739	48.18	5.26
169	8	8.810	1.432	56.18	6.26	8.810	1.627	49.69	5.23
170	10	8.992	1.401	59.69	6.53	8.992	1.598	52.58	6.74
171	4	8.251	1.245	56.58	6.74	8.251	1.456	48.75	5.79
172	6	7.919	1.189	54.75	6.79	7.919	1.402	46.70	5.77
173	5	7.488	1.129	51.70	6.77	7.488	1.347	43.61	5.69
174	3	6.817	1.041	46.61	6.69	6.817	1.269	38.61	5.52
175	2	6.078	0.956	40.61	6.52	6.078	1.192	32.95	5.26
176	4	5.746	0.944	36.95	6.26	5.746	1.182	29.89	5.03

TABLE D.9 : Posterior and Prior parameter distribution:

t=165 to 176 ; $\hat{c}=0.57$; 199 weekly data of table D.6

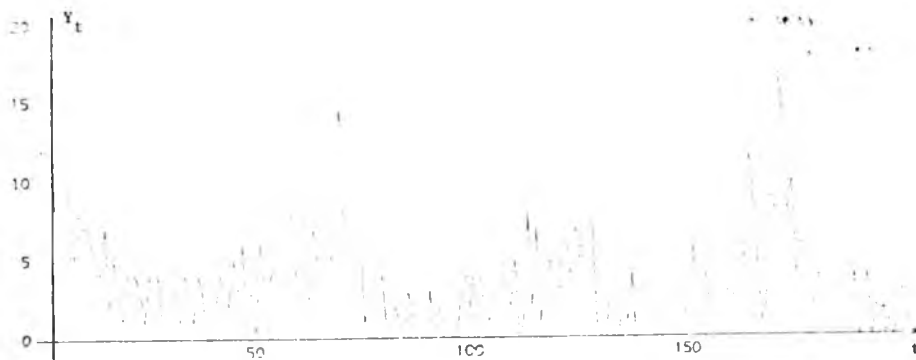


FIGURE D.1 : Plot of table D.6 data:
199 weekly deaths caused by acute respiratory infections
in Greater London - from 15th February 1972 to 1st October 1976.

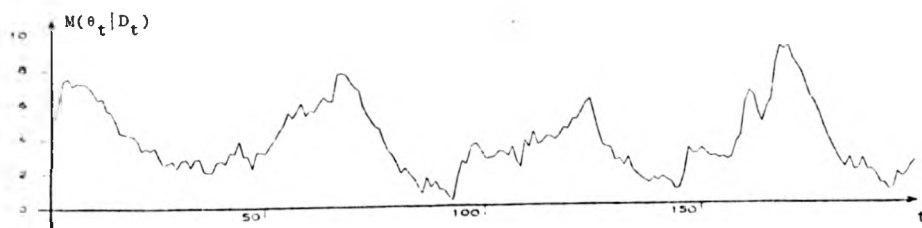


FIGURE D.2 : Plot of $M(\theta_t | D_t) \times t$; $t=1,2,\dots,199$.
Data from table D.6, where:
 $M(\theta_t | D_t) = \text{mode}(\theta_t | D_t)$.

APPENDIX E : POISSON-GAMMA BEF MULTISTATE MODEL - NUMERICAL RESULTS

This appendix contains the tables showing the relevant numerical results of the Poisson-Gamma BEF multistate model of chapter 6, section 6.4.

2	2	2	0	11	18	23	10	7	29	13	3	7	16	5	15
6	8	7	6	14	18	5	14	20	23	8	10	11	10	6	20
19	5	5	0	6	4	3	16	8	10	22	21	4	4	2	1
1	0	1	0	1	0	0	2	3	2	0	0	1	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	2	1	0	1	0	4	5	0	5	0
1	2	0	0	1	0	1	0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	4	2	1	0	1	3	3
10	4	7	1	8	11	11	18	17	33	21	21	10	7	8	4
1	0	11	2	0	0	0	1	0	0	0	0	0	0		

TABLE E.1 : 222 weekly notifications of measles cases in Truro Rural Districts, Cornwall, from the 40th week of 1966 to the 52nd week of 1970.

c	AGG. LIKL.
0.24	105.43989
0.43	105.53368
0.80	106.01669
1.10	106.54655
1.30	107.29818
1.50	107.33910
1.60	107.35877
1.63	107.36053
1.66	107.36138
1.68	107.36094
1.71	107.35840
1.73	107.35840
1.80	107.33887
2.10	107.15809
3.10	107.09938
4.00	107.09938

TABLE E.2 : $c \times$ Aggregate Likelihood 222 measless notification cases of table E.1 - MM approach.

Time	Obs.	$p_t^{(1,1)}$	$p_t^{(1,2)}$	$p_t^{(2,1)}$	$p_t^{(2,2)}$	$p_t^{(1)}$	$p_t^{(2)}$	Mode ($\theta_t D_t$)
52	0	0.991	0.002	0.005	0.005	0.993	0.007	0.12
53	1	0.934	0.055	0.0	0.011	0.934	0.066	0.12
54	0	0.991	0.002	0.002	0.004	0.993	0.007	0.12
55	0	0.997	0.002	0.0	0.0	0.997	0.003	0.12
56	2	0.391	0.570	0.0	0.039	0.391	0.609	2.99
57	3	0.0	0.026	0.0	0.973	0.001	0.999	2.99
58	2	0.0	0.0	0.001	0.999	0.001	0.999	2.96
59	0	0.010	0.0	0.340	0.651	0.349	0.651	2.86
60	0	0.844	0.002	0.049	0.105	0.893	0.107	0.12
61	1	0.794	0.046	0.003	0.157	0.797	0.203	0.12
62	0	0.971	0.002	0.008	0.019	0.979	0.021	0.12

TABLE E.3 : 222 measles notification cases; transitions illustration, from $t=52$ to $t=62$ $p_t^{(i,j)} = \text{Prob}\{M_{t-1}^{(i)} M_t^{(j)} | D_t\}$; $p_t^{(k)} = \text{Prob}\{M_t^{(k)} | D_t\}$ $i, j, k=1,2$; Mode ($\theta_t | D_t$) = Mode ($\theta_t | M_t^{(2)} D_t$) or θ_c .

Time	Obs.	$p_t^{(1,1)}$	$p_t^{(1,2)}$	$p_t^{(2,1)}$	$p_t^{(2,2)}$	$p_t^{(1)}$	$p_t^{(2)}$	Mode ($\theta_t D_t$)
185	0	0.997	0.002	0.0	0.0	0.998	0.002	0.12
186	4	0.001	0.943	0.0	0.056	0.001	0.999	3.01
187	2	0.0	0.0	0.001	0.999	0.001	0.999	2.99
188	1	0.001	0.0	0.022	0.978	0.022	0.978	2.93
189	0	0.1959	0.001	0.271	0.533	0.467	0.533	2.83
190	1	0.3464	0.020	0.012	0.621	0.359	0.641	2.76
191	3	0.001	0.023	0.0	0.976	0.001	0.999	2.78
192	3	0.0	0.0	0.0	0.999	0.0	1.0	2.79

TABLE E.4 : 222 measles notification cases; transitions illustration from $t=185$ to $t=192$ $p_t^{(i,j)}$, $p_t^{(k)}$, Mode ($\theta_t | D_t$) as explained in table E.3 .

Time	Obs.	$p_t^{(1,1)}$	$p_t^{(1,2)}$	$p_t^{(2,1)}$	$p_t^{(2,2)}$	$p_t^{(1)}$	$p_t^{(2)}$	Mode ($\theta_t D_t$)
210	0	0.983	0.002	0.014	0.0	0.998	0.002	0.12
211	11	0.0	0.608	0.0	0.392	0.0	1.0	3.51
212	2	0.0	0.0	0.001	0.999	0.001	0.999	2.98
213	0	0.013	0.0	0.320	0.667	0.333	0.667	2.21
214	0	0.758	0.002	0.048	0.192	0.806	0.194	0.12
215	0	0.947	0.002	0.007	0.043	0.954	0.046	0.12

TABLE E.5 : 222 measles notification cases; transitions illustration from $t=210$ to $t=215$ $p_t^{(i,j)}$, $p_t^{(k)}$, Mode ($\theta_t | D_t$) as explained in table E.3 .

c	AGG. LIKL.
0.50	45.35266
0.70	49.10042
0.80	54.40331
1.00	57.02862
1.12	57.63406
1.16	57.70977
1.18	57.72636
1.20	57.72938
1.22	57.71928
1.24	57.69655
1.30	57.55927
1.50	56.85070
2.00	55.19103
2.50	50.00568
3.50	32.96881
5.00	14.23188

TABLE E.6 : $c \times$ Aggregate Likelihood
222 measles notification
cases of table E.1 - SM
approach.

Time	Obs.	MM			SM	
		$p_t^{(1)}$	$p_t^{(2)}$	Mode($\theta_t D_t$)	Mode($\theta_t D_t$)	Var ($\theta_t D_t$)
46	4	0.0	1.0	8.69	10.25	0.54
47	2	0.03	0.97	8.14	9.82	0.52
48	1	0.67	0.33	0.12	9.36	0.49
49	1	0.91	0.09	0.12	8.92	0.51
50	0	0.99	0.01	0.12	8.44	0.45
51	1	0.93	0.07	0.12	8.04	0.44
52	0	0.99	0.01	0.12	7.59	0.42
53	1	0.93	0.07	0.12	7.22	0.41
54	0	0.99	0.01	0.12	6.79	0.39
55	0	1.0	0.0	0.12	6.39	0.38

TABLE E.7 : MM and SM results, from t=46 to t=55 ;

$$p_t^{(i)} = \text{Prob} \{M_t^{(i)} | D_t\} \quad , \quad i=1,2 .$$

Time	Obs.	MM			SM	
		$p_t^{(1)}$	$p_t^{(2)}$	Mode ($\theta_t D_t$)	Mode ($\theta_t D_t$)	Var ($\theta_t D_t$)
91	0	0.98	0.02	0.12	0.57	0.19
92	4	0.0	1.0	2.95	1.41	0.40
93	5	0.0	1.0	3.01	0.51	2.17
94	0	0.35	0.65	2.90	0.36	1.77
95	5	0.0	1.0	2.99	0.43	2.32
96	0	0.35	0.69	2.85	0.33	1.96
97	1	0.26	0.74	2.76	0.30	1.81
98	2	0.10	0.99	2.72	0.30	1.84
99	0	0.35	0.65	2.55	0.25	1.57
100	0	0.86	0.14	0.12	0.22	1.33

TABLE E.8 : MM and SM results, from t=91 to t=100;

$$p_t^{(i)} = \text{Prob} \{M_t^{(i)} | D_t\} \quad ; \quad i=1,2 .$$

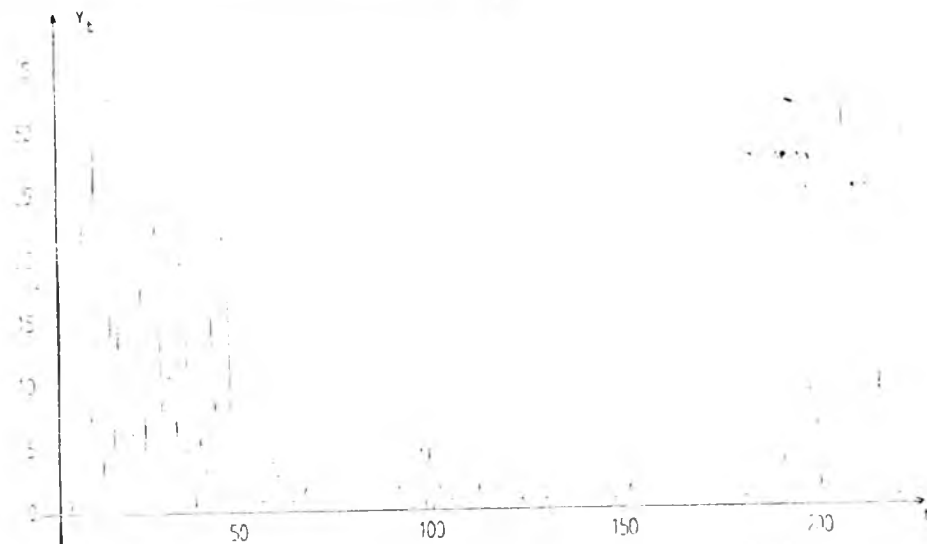


FIGURE E.1 : Plot of table E.1 data:
Weekly notifications of measles cases in Truro Rural District,
Cornwall from the 40th week of 1966 to the 52nd week of 1970.
(222 observations).

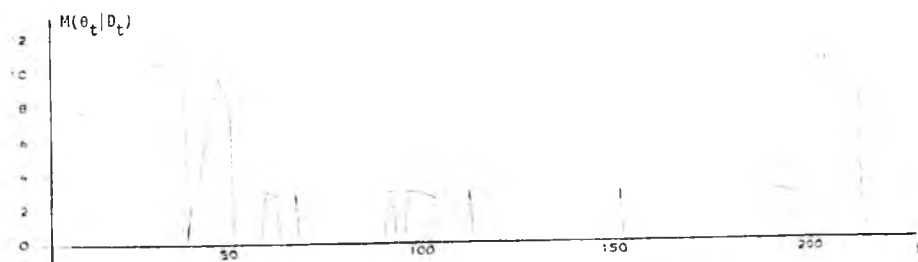


FIGURE E.2 : Plot of $M(\theta_t | D_t)$ t under ML formulation, $t=1,2,\dots,222$.
Data from table E.1, where: $M(\theta_t | D_t) = \text{Mode}(\theta_t | D_t)$



FIGURE E.3: Plot of $M(\theta_t | D_t) \times t$ under SI formulation, $t=1,2,\dots,222$.
Data from table E.1, where: $M(\theta_t | D_t) = \text{Mode}(\theta_t | D_t)$

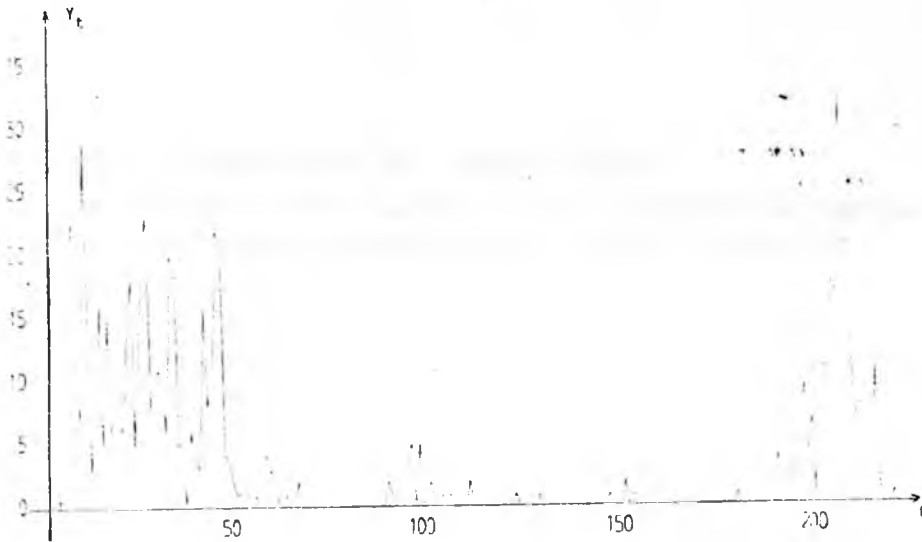


FIGURE E.1 : Plot of table E.1 data:
Weekly notifications of measles cases in Truro Rural District,
Cornwall from the 40th week of 1966 to the 52nd week of 1970.
(222 observations).

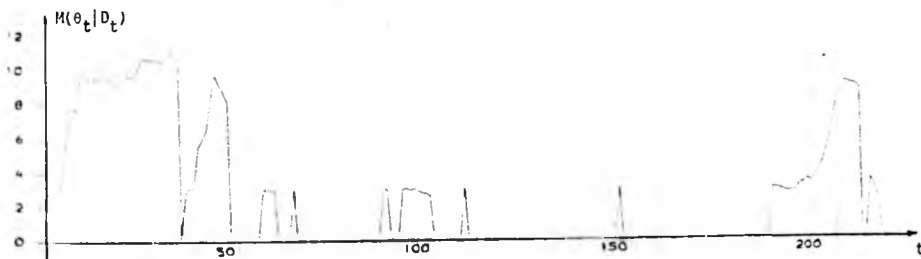


FIGURE E.2 : Plot of $M(\theta_t | D_t)$ t under IMI formulation, $t=1,2,\dots,222$.
Data from table E.1, where: $M(\theta_t | D_t) = \text{Mode}(\theta_t | D_t)$

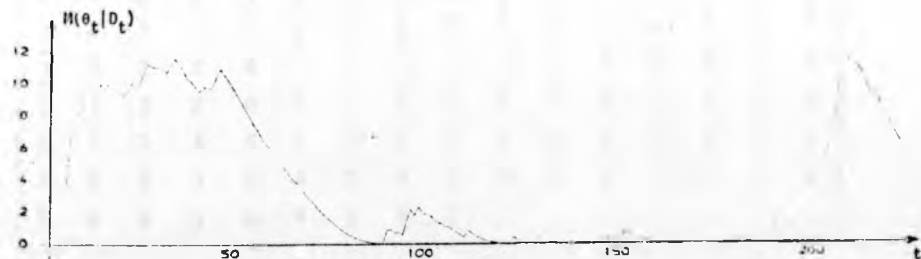


FIGURE E.3: Plot of $M(\theta_t | D_t) \times t$ under SII formulation, $t=1,2,\dots,222$.
Data from table E.1, where: $M(\theta_t | D_t) = \text{Mode}(\theta_t | D_t)$

APPENDIX F : BINOMIAL-BETA BEF -NUMERICAL RESULTS

This appendix contains tables and plots illustrating the numerical results of the Binomial-Beta BEF System of chapter 7, section 7.6 .

6	2	3	5	3	2	3	4	3	1	2	2	3	5	3	1
2	1	4	3	1	5	6	2	2	4	3	5	5	3	5	2
5	2	4	5	5	3	4	4	2	2	4	2	3	4	2	4
2	2	6	3	3	4	2	3	2	5	3	4	4	4	4	1
4	1	2	4	1	5	2	5	5	4	3	4	4	6	4	3
2	3	3	4	4	3	4	3	2	3	4	4	4	4	5	1
2	3	3	1	1	4	3	0	2	3	1	4	3	2	2	3
3	2	4	1	2	4	2	1	1	2	3	5	0	0	8	5
3	1	4	4	4	2	1	2	4	2	3	2	5	2	4	5
5	3	7	4	4	3	3	2	2	4	3	4	0	4	3	3
3	6	5	4	4	2	5	1	5	5	1	3	5	2	4	2
1	3	3	2	4	2	1	4	2	3	3	6	1	2	3	3
2	3	5	3	1	2	3	2	3	1	1	4	4	2	3	5
5	0	4	2	2	1	6	5	2	1	3	3	2	2	5	4
2	1	3	4	3	4	5	5	1	3	4	2	4	2	1	4
3	7	2	5	4	6	4	2	4	5	2	4	1	0	5	4
4	3	6	3	5	3	3	2	3	5	3	3	3	1	0	6
0	6	2	3	3	3	4	0	3	4	1	5	3	6	4	3
3	5	0	3	2	1	3	2	4	4	2	4	2	2	3	2
3	3	3	2	2	2	1	2	4	2	3	3	3	4	2	1
5	3	3	5	3	2	4	4	3	4	3	3	2	1	3	3
2	5	3	4	2	2	2	2	1	2	5	4	2	2	4	1
4	3	2	2	2	2	2	4	2	2	4	2	6	4	1	5
4	4	4	3	2	4	4	2	5	2	2	3	3	5	4	1
4	4	5	4	3	2	2	2	3	1	4	4	2	3	3	1
3	4	4	2	4	2	2	2	5	4	0	2	4	1	1	4
3	0	2	4	4	1	3	1	5	3	1	2	5	3	3	3
3	4	3	2	0	1	2	2	3	4	2	2	2	3	2	3
4	3	3	3	5	1	3	5	3	6	2	2	4	2	3	3
2	1	4	3	4	4	2	4	3	3	1	4	4	4	1	3
5	4	4	3	0	4	1	4	2							

TABLE F.1 : 490 generated Binomial (0.375;8) data.

c	Aggregate Likelihood
0.15×10^8	46.4120018
0.14×10^8	46.4126957
0.13×10^8	46.4131031
0.125×10^8	46.4133541
0.12×10^8	46.4149768
0.115×10^8	46.4134186
0.11×10^8	46.4128617
0.10×10^8	46.4136844
0.75×10^7	46.4130264
0.60×10^7	46.4117459
0.25×10^7	46.4109460
0.10×10^7	46.4055137
0.18×10^6	46.3987440
0.15×10^5	46.3834723
0.50×10^3	46.3363340
0	26.6666667

TABLE F.2 : $c \times$ Aggregate likelihood for first half of data from table F.1 .

c	Aggregate Lik.
0.50×10^8	99.1830423
0.49×10^8	99.1843888
0.48×10^8	99.1766030
0.46×10^8	99.1886967
0.44×10^8	99.1824266
0.40×10^8	99.1774883
0.35×10^8	99.1801794
0.25×10^8	99.1792754
0.15×10^8	99.1753550
0.10×10^8	99.1696431

TABLE F.3 : $c \times$ Aggregate likelihood for data of table F.1 .

t	$(0_t D_{t-1})$				Obs. Y_t	$(0_t D_t)$			
	α	γ	Mode	Var.		α	γ	Mode	Var.
483	145.44	248.80	0.3682	0.0006	4	149.44	252.80	0.3709	0.0006
484	146.89	248.47	0.3709	0.0006	4	150.89	252.47	0.3734	0.0006
485	147.01	245.97	0.3734	0.0006	3	150.01	250.97	0.3735	0.0006
486	148.46	248.36	0.3735	0.0006	0	148.46	256.36	0.3661	0.0006
487	142.31	245.71	0.3661	0.0006	4	146.31	249.71	0.3688	0.0006
488	145.19	249.51	0.3688	0.0006	1	147.29	256.51	0.3639	0.0006
489	142.95	249.10	0.3639	0.0006	4	146.95	253.10	0.3667	0.0006
490	145.94	251.35	0.3667	0.0006	2	147.94	257.35	0.3643	0.0006

TABLE F.4 : Binomial-Beta BEF model-data from table F.1
 Prior-Posterior parameter distribution ;
 t=483,484,...,490

2	2	2	2	2	4	4	4	4	4	3	4	4	5	7	6
5	6	8	10	8	7	8	8	6	8	7	6	4	3	5	5
6	4	6	4	3	6	6	8	4	3	3	4	3	2	1	1
2	0	3	1	1	0	1	1	2	1	1	0	2	1	1	2
2	2	2	0	0	1	1	0	0	2	0	0	2	0	0	1
0	1	0	1	0	1	5	3	5	3	3	5	5	4	3	3
5	2	2	2	4	2	1	1	1	2	1	3	0	1	0	0
0	1	1	0	0	2	3	0	2	2	2	3	5	2	1	1
2	2	2	2	3	1	0	0	3	2	2	2	1	2	4	1
2	2	1	3	5	2	2	1	1	0	0	1	2	0	1	1
1	1	2	1	1	1	2	1	1	1	2	1	1	1	1	3
1	3	2	2	2	1	3	0	0	5	5	6	6	7	5	7
5	7	5	6	5	5	6	5	6	4	6	3	4	3	1	1
2	1	3	4	3	1	2	1	1	1	1	2	1	1		

TABLE F. 6 : Weekly number of rural districts (RD) affected by the measles epidemic obtained from table F.5 .

1	3	2	1	1	3	1	3	5	7	2	4	4	8	7	9
11	6	10	11	10	9	9	9	9	8	9	7	8	8	8	9
8	5	9	6	8	7	9	6	6	5	6	5	6	6	3	3
2	3	2	1	3	2	0	1	2	2	1	0	1	1	1	1
1	3	3	1	2	1	3	2	2	2	0	1	1	1	0	0
2	1	0	2	1	0	3	2	1	2	2	3	3	6	5	3
6	6	5	6	2	1	3	4	3	1	2	1	1	0	0	2
2	2	3	3	0	4	2	4	3	1	3	3	3	4	3	3
2	4	1	2	2	3	3	2	1	2	0	3	1	4	5	2
5	6	3	6	4	4	7	2	2	0	1	0	1	1	2	3
2	2	1	2	1	1	1	1	1	2	1	4	1	2	3	3
1	2	2	4	2	3	2	3	2	5	3	5	5	7	8	9
6	8	8	8	9	7	7	9	9	9	7	9	8	4	2	3
6	2	3	2	2	2	2	3	2	0	0	1	0	0		

TABLE F.7 : Weekly number of municipal boroughs(MB) and urban districts (UD) affected by the measles epidemic. Obtained from table F.5

c	Aggregate Likelihood
5000	39.0851164
1000	39.7353946
500	40.0828649
100	41.1384569
40	42.0325740
10	44.3593308
4.5	46.7527091
3.0	48.0156816
2.6	48.2213393
2.5	48.2332605
2.4	48.2209074
2.0	47.8044789
1.0	40.8676841
0.5	30.8524909
0.2	22.9103144
0.1	20.9662061
0	20.1818182

TABLE F.8 : $c \times$ Aggregate likelihood
for RD data of table F.6 .

Time t	$(\theta_t D_{t-1})$				Obs. Y_t	$(\theta_t D_t)$			
	α	γ	Mode	Var.		α	γ	Mode	Var.
18	17.18	17.23	0.4994	0.0071	6	21.18	21.23	0.5231	0.0055
19	17.95	16.46	0.5231	0.0070	8	25.95	18.46	0.5884	0.0053
20	20.07	14.34	0.5884	0.0069	10	30.07	14.33	0.6854	0.0048
21	23.22	11.20	0.6854	0.0062	8	31.22	13.20	0.7124	0.0046
22	24.10	10.32	0.7124	0.0059	7	31.10	13.32	0.7095	0.0046
23	24.01	10.42	0.7095	0.0060	8	32.01	12.42	0.7308	0.0044
24	24.40	9.73	0.7308	0.0057	8	32.70	11.73	0.7471	0.0043
25	25.23	9.20	0.7471	0.0055	6	31.23	13.20	0.7125	0.0046
26	24.11	10.33	0.7125	0.0059	8	32.11	12.33	0.7331	0.0044
27	24.78	9.66	0.7331	0.0057	7	31.78	12.66	0.7253	0.0045
28	24.53	9.91	0.7253	0.0058	6	30.53	13.91	0.6958	0.0047
29	23.57	10.87	0.6958	0.0061	4	27.57	16.87	0.6261	0.0052
30	21.30	13.13	0.6261	0.0067	3	24.30	20.13	0.5492	0.0055

TABLE F.9 : Binomial-Beta BEF model-RD data from table F.6 Prior-Posterior parameter distribution; $t=18,19,\dots,30$; $n=10$.

Time t	$(\theta_t D_{t-1})$				Obs. Y_t	$(\theta_t D_t)$			
	α	γ	Mode	Var.		α	γ	Mode	Var.
88	6.26	28.23	0.1619	0.0042	3	9.26	35.23	0.1944	0.0036
89	7.31	27.16	0.1944	0.0047	5	12.31	32.16	0.2663	0.0044
90	9.65	24.82	0.2663	0.0057	3	12.65	31.82	0.2743	0.0045
91	9.90	24.55	0.2743	0.0058	3	12.90	31.55	0.2803	0.0045
92	10.10	24.35	0.2803	0.0058	5	15.10	29.35	0.3321	0.0049
93	11.77	22.67	0.3321	0.0063	5	16.77	27.67	0.3717	0.0052
94	13.05	21.38	0.3717	0.0066	4	17.05	27.38	0.3783	0.0052
95	13.27	21.16	0.3783	0.0067	3	16.27	28.16	0.3599	0.0051
96	12.67	21.76	0.3599	0.0066	3	15.67	28.76	0.3458	0.0050
97	12.21	22.21	0.3458	0.0065	5	17.21	27.21	0.3821	0.0052
98	13.39	21.03	0.3821	0.0067	2	15.39	29.03	0.3392	0.0050
99	12.00	22.43	0.3392	0.0064	2	14.00	30.43	0.3064	0.0048
100	10.94	23.49	0.3064	0.0061	2	12.94	31.49	0.2813	0.0045

TABLE F.10 : Binomial-Beta BEF model- RD data from table F.6.

Prior-Posterior parameter distribution ; $t=88,89,\dots,100$;

$n=10$.

Time t	$(\theta_t D_{t-1})$				Obs. Y_t	$(\theta_t D_t)$			
	α	γ	Mode	Var.		α	γ	Mode	Var.
189	12.43	22.01	0.3525	0.0065	6	18.43	26.01	0.4108	0.0053
190	14.32	20.11	0.4108	0.0069	7	21.32	23.11	0.4789	0.0055
191	16.53	17.90	0.4789	0.0070	5	21.53	22.90	0.4839	0.0055
192	16.69	17.73	0.4389	0.0071	7	23.69	20.73	0.5348	0.0055
193	18.34	16.08	0.5348	0.0070	5	23.34	21.08	0.5266	0.0055
194	18.07	16.35	0.5266	0.0070	7	25.07	19.35	0.5675	0.0054
195	19.40	15.02	0.5675	0.0069	5	24.40	20.02	0.5516	0.0055
196	18.88	15.54	0.5516	0.0070	6	24.88	19.54	0.5630	0.0054
197	19.25	15.17	0.5630	0.0070	5	24.25	20.17	0.5481	0.0055
198	18.77	15.75	0.5481	0.0070	5	23.77	20.65	0.5368	0.0055
199	18.40	16.02	0.5368	0.0070	6	24.40	20.02	0.5517	0.0055
200	18.88	15.53	0.5517	0.0070	5	23.88	20.53	0.5395	0.0055
201	18.49	15.93	0.5395	0.0070	6	24.49	19.93	0.5538	0.0054
202	18.95	15.47	0.5538	0.0070	4	22.95	21.47	0.5175	0.0055
203	17.78	16.64	0.5175	0.0071	6	23.78	10.64	0.5370	0.0055
204	18.41	16.01	0.5370	0.0070	3	21.41	23.01	0.4811	0.0055

TABLE F.11 : Binomial-Beta BEF model - RD data from table F.6
 Prior-Posterior parameter distribution;
 $t = 189, 190, \dots, 204; n = 10$.

Time	$Y_t n_{t-1}$	$p(Y_t n_{t-1})$	Obs. Y_t
210	1	0.160474	1
	2	0.235926	
	3	0.231845	
211	1	0.207007	3
	2	0.257663	
	3	0.216614	
212	1	0.186625	4
	2	0.249798	
	3	0.224747	
213	2	0.226195	3
	3	0.233980	
	4	0.177592	
214	2	0.222962	1
	3	0.234323	
	4	0.180568	
215	1	0.190334	2
	2	0.251421	
	3	0.223427	
216	1	0.202367	1
	2	0.256102	
	3	0.218654	
217	1	0.241158	1
	2	0.264932	
	3	0.198386	
218	0	0.147075	1
	1	0.270650	
	2	0.264520	
219	0	0.176658	1
	1	0.291954	
	2	0.259305	
220	0	0.202768	2
	1	0.306835	
	2	0.252198	
221	0	0.178433	1
	1	0.293080	
	2	0.258890	
222	0	0.204298	1
	1	0.307609	
	2	0.251728	

TABLE F.12 : Binomial-Beta BEF predictive distribution - RD data from table F.6 $t=210,211,\dots,222$; $n=10$.

c	Aggregate Likelihood
500	33.7555911
250	34.0865722
100	34.7486119
50	35.3835204
25	36.1952246
10	37.5963212
5	38.8459551
4	39.1931702
3.1	39.4263327
3.0	39.4348140
2.9	39.4366449
2.8	39.4304054
2.7	39.4143840
2.0	38.7892465
1.0	32.7262626
0.5	23.4042801
0.0	12.33

TABLE F.13 : $c \times$ Aggregate likelihood for MB & UD data of table F.7 .

Time t	$(\theta_t D_{t-1})$				Obs. Y_t	$(\theta_t D_t)$			
	α	γ	Mode	Var.		α	γ	Mode	Var.
18	21.68	26.36	0.4493	0.0050	6	27.68	37.36	0.4233	0.0037
19	20.49	27.55	0.4233	0.0050	10	30.49	34.55	0.4678	0.0038
20	22.53	25.50	0.4678	0.0051	11	33.53	31.50	0.5161	0.0038
21	24.76	23.28	0.5161	0.0051	10	34.76	30.28	0.5356	0.0038
22	25.65	22.38	0.5356	0.0051	9	34.65	30.78	0.5339	0.0038
23	25.58	22.46	0.5339	0.0051	9	34.58	30.46	0.5327	0.0038
24	25.52	22.51	0.5327	0.0051	9	35.52	30.51	0.5318	0.0038
25	25.48	22.55	0.5318	0.0051	9	34.48	30.55	0.5312	0.0038
26	25.45	22.58	0.5312	0.0051	8	33.45	31.58	0.5148	0.0038
27	24.70	23.33	0.5148	0.0051	9	33.70	31.33	0.5188	0.0038
28	24.88	23.15	0.5188	0.0051	7	31.88	33.15	0.4899	0.0038
29	23.55	24.48	0.4899	0.0051	8	31.55	33.48	0.4847	0.0038
30	23.31	24.72	0.4847	0.0051	8	31.31	33.72	0.4809	0.0038

TABLE F.14 : Binomial-Beta BEF model - MB & UD data from table F.7 , Prior-Posterior parameter distribution ; $t=18,19,\dots,30$; $n=17$.

Time t	$(\theta_t D_{t-1})$				Obs. Y_t	$(\theta_t D_t)$			
	α	γ	Mode	Var.		α	γ	Mode	Var.
88	4.72	43.44	0.0805	0.0018	2	6.72	58.44	0.0905	0.0014
89	5.18	42.97	0.0905	0.0020	1	6.18	58.97	0.0820	0.0013
90	4.78	43.37	0.0820	0.0018	2	6.78	58.37	0.0916	0.0014
91	5.23	42.92	0.0916	0.0020	2	7.23	57.92	0.0986	0.0015
92	5.55	45.59	0.0986	0.0021	3	8.55	56.59	0.1196	0.0017
93	6.52	41.62	0.1196	0.0024	3	9.52	55.62	0.1349	0.0019
94	7.22	40.90	0.1349	0.0026	6	13.22	51.90	0.1936	0.0024
95	9.23	38.18	0.1936	0.0033	5	14.93	50.18	0.2207	0.0027
96	11.17	36.92	0.2207	0.0036	3	14.17	50.92	0.2088	0.0026
97	10.62	37.46	0.2088	0.0035	6	16.62	48.46	0.2476	0.0029
98	12.41	35.67	0.2476	0.0039	6	18.41	46.67	0.2760	0.0031
99	13.72	34.35	0.2760	0.0042	5	18.72	46.35	0.2809	0.0031
100	13.94	34.13	0.2809	0.0042	6	19.94	45.13	0.3003	0.0032

TABLE F.15 : Binomial-Beta BEF model - MB & UD data from table
 F.7 Prior-Posterior parameter distribution;
 $t=88,89,\dots,100$; $n=17$.

Time t	$(\theta_t D_{t-1})$				Obs. y_t	$(\theta_t D_t)$			
	α	γ	Mode	Var.		α	γ	Mode	Var.
189	10.75	37.34	0.2116	0.0035	5	15.75	49.34	0.2339	0.0028
190	11.78	36.30	0.2339	0.0038	7	18.78	46.30	0.2818	0.0031
191	13.98	34.09	0.2818	0.0042	8	21.98	43.09	0.3327	0.0034
192	16.32	31.73	0.3327	0.0046	9	25.32	39.73	0.3857	0.0036
193	18.76	29.29	0.3857	0.0049	6	24.76	40.29	0.3769	0.0036
194	18.35	29.69	0.3769	0.0048	8	26.35	38.69	0.4022	0.0036
195	19.52	28.53	0.4022	0.0049	8	27.52	37.53	0.4206	0.0037
196	20.36	27.67	0.4206	0.0050	8	28.36	36.67	0.4341	0.0037
197	20.98	27.05	0.4341	0.0050	9	29.98	35.05	0.4598	0.0038
198	22.17	25.87	0.4598	0.0051	7	29.17	35.87	0.4468	0.0037
199	21.57	26.47	0.4468	0.0050	7	28.57	36.47	0.4374	0.0037
200	21.14	26.90	0.4374	0.0050	9	30.14	34.90	0.4622	0.0038
201	22.28	25.76	0.4622	0.0051	9	31.28	33.76	0.4803	0.0038
202	23.11	24.92	0.4803	0.0051	9	32.11	32.92	0.4936	0.0038
203	23.72	24.31	0.4936	0.0051	7	30.72	34.31	0.4715	0.0038
204	22.71	25.33	0.4715	0.0051	9	31.71	33.33	0.4871	0.0038

TABLE F.16 : Binomial-Beta BEF model - MB & UD data from table F.7
 Prior-Posterior parameter distribution; $t=189,190,\dots,204$;
 $n=17$.

Time	$Y_t D_{t-1}$	$p(Y_t D_{t-1})$	Obs. Y_t
210	4	0.158787	2
	5	0.177052	
	6	0.163405	
211	3	0.162361	3
	4	0.186864	
	5	0.173438	
212	3	0.181168	2
	4	0.191310	
	5	0.163508	
213	2	0.168978	2
	3	0.202499	
	4	0.187200	
214	2	0.196477	2
	3	0.210771	
	4	0.175581	
215	1	0.153637	2
	2	0.214962	
	3	0.211557	
216	1	0.174198	3
	2	0.226783	
	3	0.208881	
217	1	0.162727	2
	2	0.220508	
	3	0.210710	
218	1	0.181258	0
	2	0.230250	
	3	0.207360	
219	0	0.145928	0
	1	0.254143	
	2	0.247494	
220	0	0.229062	1
	1	0.302583	
	2	0.232676	
221	0	0.248501	1
	1	0.309590	
	2	0.226735	
222	0	0.335069	0
	1	0.325761	

TABLE F.17 : Binomial-Beta BEF predictive distribution - MB & UD data from table F.7
 $t=210,211,\dots,222$; $n=17$.

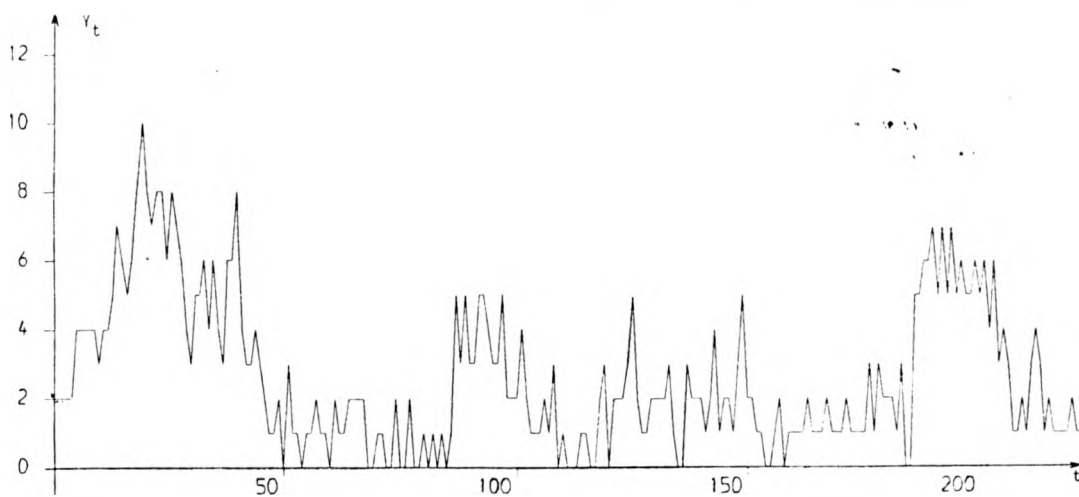


FIGURE F.1 : Plot of table F.6 data:

Weekly number of rural districts (RD) in Cornwall affected by measles epidemic from the 40th week of 1966 to the 52nd week of 1970 (222 observations).

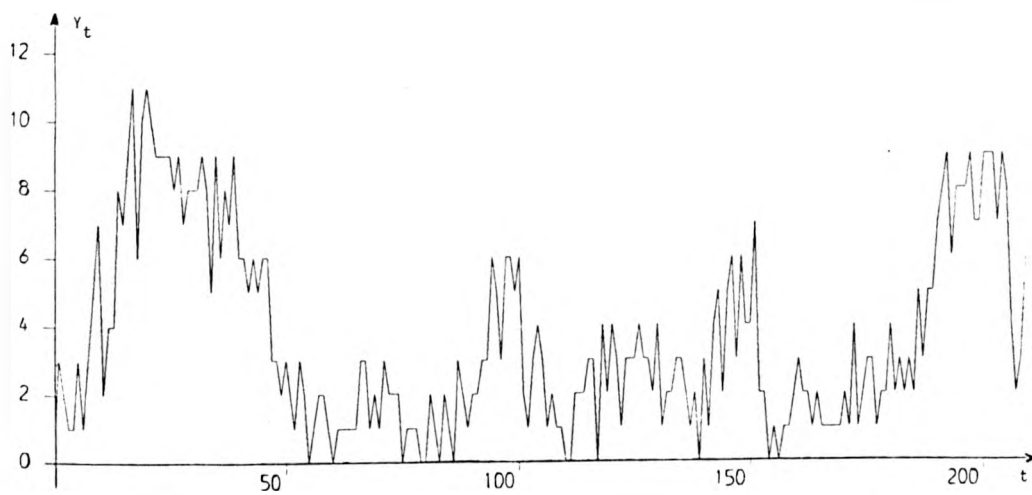


FIGURE F.2: Plot of table F.7 data:

Weekly number of municipal boroughs and urban districts (MB & UD) in Cornwall affected by measles epidemic from the 40th week of 1966 to the 52nd week of 1970 (222 observation).

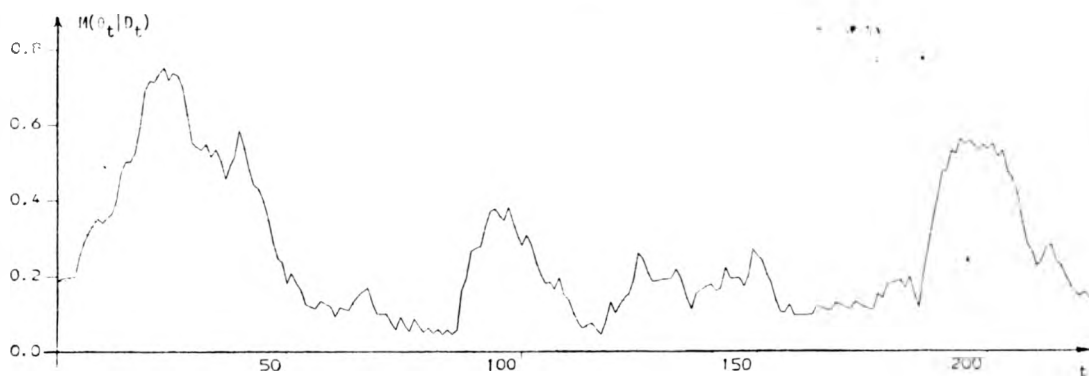


FIGURE F.3 : Plot of $M(\theta_t | D_t) \times t$ for RD data of table F.6, where:
 $M(\theta_t | D_t) = \text{Mode}(\theta_t | D_t)$; $t=1,2,\dots, 222$.

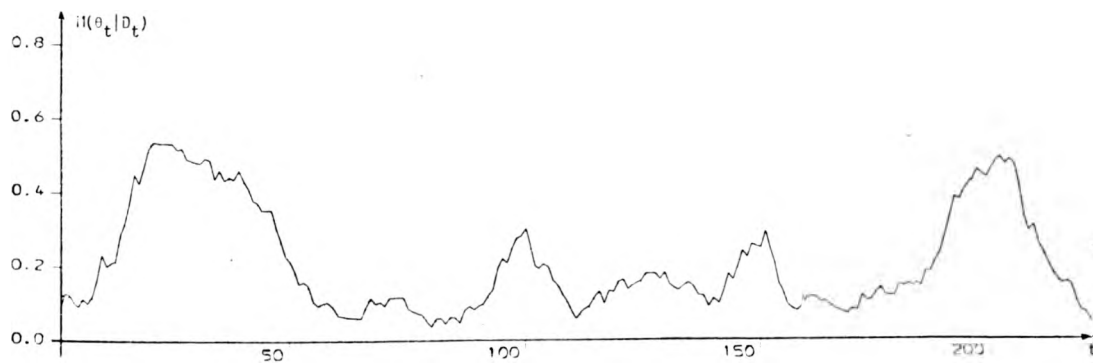


FIGURE F.4 : Plot of $M(\theta_t | D_t) \times t$ for MB & UD data of table F.7, where:
 $M(\theta_t | D_t) = \text{Mode}(\theta_t | D_t)$, $t=1,2,\dots, 222$.

APPENDIX G :

Numerical results concerning the simulations and application of the truncated normal BEF model - Chapter 8.





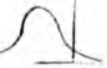

Prior ($\theta_t D_{t-1}$)		Obs. y_t $v^2=2$	Posterior ($\theta_t D_t$)						
Charact.	μ_t		σ_t^2	m_t		C_t		μ_t^*	σ_t^{*2}
 5% trunc.	1.65	1	0	1.13	1.14	0.48	0.49	0.89	0.78
	1.65	1	1	1.25	1.27	0.57	0.58	1.09	0.77
	1.65	1	2	1.70	1.71	0.64	0.65	1.66	0.74
	1.65	1	3	2.03	2.04	0.68	0.68	2.02	0.72
 25% trunc.	0.67	1	0	0.75	0.76	0.31	0.32	0.10	0.82
	0.67	1	1	0.92	0.93	0.39	0.40	0.51	0.80
	0.67	1	2	1.14	1.15	0.49	0.50	0.91	0.78
	0.67	1	3	1.41	1.42	0.57	0.58	1.30	0.76
 50% trunc.	0	1	0	0.58	0.58	0.22	0.23	-0.48	0.85
	0	1	1	0.70	0.71	0.28	0.25	-0.05	0.83
	0	1	2	0.86	0.87	0.36	0.37	0.37	0.81
	0	1	3	1.07	1.07	0.45	0.47	0.77	0.79
 75% trunc.	-0.67	1	0	0.45	0.46	0.15	0.16	-1.07	0.87
	-0.67	1	1	0.54	0.55	0.20	0.21	-0.63	0.85
	-0.67	1	2	0.65	0.65	0.26	0.27	-0.20	0.83
	-0.67	1	3	0.80	0.80	0.33	0.34	0.22	0.81
 90% trunc.	-1.28	1	0	0.38	0.38	0.11	0.11	-1.62	0.89
	-1.28	1	1	0.44	0.45	0.14	0.15	-1.17	0.87
	-1.28	1	2	0.52	0.53	0.19	0.19	-0.72	0.86
	-1.28	1	3	0.63	0.63	0.24	0.25	-0.29	0.84
 95% trunc.	-1.65	1	0	0.34	0.35	0.09	0.10	-1.95	0.90
	-1.65	1	1	0.39	0.40	0.12	0.13	-1.50	0.88
	-1.65	1	2	0.46	0.47	0.15	0.16	-1.05	0.87
	-1.65	1	3	0.55	0.55	0.20	0.21	-0.60	0.85

TABLE G.1 : Posterior distribution - true and approximated distribution ; comparison $(\theta_t | D_{t-1}) \sim N(\mu_t; \sigma_t^2)$
 $m_t = E\{\theta_t | D_t\}$; $C_t = \text{Var}\{\theta_t | D_t\}$;

μ_t^*, σ_t^{*2} parameters of the untruncated posterior (under approximation)

Prior ($\theta_t D_{t-1}$)		v^2	Predictive ($Y_t D_{t-1}$)					
μ_t	σ_t^2		m_{y_t}		c_{y_t}		μ_{y_t}	$\sigma_{y_t}^2$
			true	app.	true	app.		
1.65	1	1	1.92	1.90	1.34	1.32	1.52	2.03
1.65	1	2	2.12	2.10	1.79	1.75	1.55	2.90
1.65	1	3	2.30	2.29	2.24	2.20	1.56	3.86
1.65	1	4	2.48	2.47	2.68	2.65	1.58	4.84
0.67	1	1	1.41	1.33	0.92	0.78	0.84	1.41
0.67	1	2	1.68	1.63	1.33	1.24	0.90	2.44
0.67	1	3	1.90	1.88	1.75	1.68	0.93	3.45
0.67	1	4	2.10	2.08	2.17	2.11	0.95	4.46
0	1	1	1.21	1.13	0.72	0.60	0.59	1.21
0	1	2	1.50	1.46	1.13	1.05	0.64	2.25
0	1	3	1.74	1.71	1.54	1.48	0.67	3.27
0	1	4	1.95	1.93	1.94	1.90	0.68	4.28
-0.67	1	1	1.09	1.02	0.60	0.52	0.42	1.12
-0.67	1	2	1.40	1.36	1.00	0.95	0.47	2.15
-0.67	1	3	1.64	1.62	1.40	1.36	0.50	3.17
-0.67	1	4	1.85	1.83	1.79	1.76	0.51	4.18
-1.65	1	1	0.98	0.92	0.51	0.45	0.26	1.06
-1.65	1	2	1.30	1.27	0.90	0.86	0.31	2.08
-1.65	1	3	1.55	1.53	1.29	1.26	0.34	3.10
-1.65	1	4	1.76	1.75	1.68	1.65	0.35	4.10

TABLE G.2 : Predictive distribution - true and approximated distributions comparison.

$$(\theta_t | D_{t-1}) \sim N(\mu_t, \sigma_t^2); \theta_t \in \mathbb{R}^+$$

$$m_{y_t} = E\{Y_t | D_{t-1}\}; C_{y_t} = \text{Var}\{Y_t | D_{t-1}\}; v^2 = \text{Unt. Var}(Y_t | \theta_t)$$

$\mu_{y_t}; \sigma_{y_t}^2$ parameters of the untruncated distr. for

$(Y_t | D_{t-1})$ (approx).

1.72	5.73	0.68	5.17	3.77	5.21	6.13	1.32	0.92	1.33	2.65	1.04	3.07	2.92	5.35	1.96
0.40	0.93	2.45	2.48	3.68	4.03	3.21	3.20	5.01	6.19	3.27	3.37	2.34	1.64	0.04	1.48
3.23	2.23	5.38	1.51	4.24	3.40	2.58	1.91	2.00	2.37	1.93	1.46	2.30	2.51	0.52	3.45
2.16	6.58	4.60	4.44	3.59	1.48	2.55	2.45	1.02	3.21	3.00	3.84	0.20	3.99	1.18	2.00
2.01	2.59	3.41	3.09	0.43	3.49	2.42	1.27	1.29	4.30	5.00	2.48	2.91	2.59	1.89	1.62
3.49	7.31	2.48	0.78	8.53	0.56	1.39	0.99	2.52	0.05	4.96	5.27	0.13	5.27	4.47	1.96
1.02	0.01	2.22	2.15	4.96	3.07	3.95	0.64	3.54	0.99	3.85	4.70	0.80	1.06	3.60	0.51
1.82	4.47	7.40	0.51	2.78	2.86	1.48	0.05	1.66	1.36	0.37	2.95	5.18	3.66	0.88	3.08
2.72	3.14	2.37	5.63	3.46	1.12	3.59	3.62	2.54	2.25	0.44	2.27	2.31	1.71	4.13	1.50
1.39	2.98	4.82	6.76	3.81	2.29	4.71	4.99	2.27	1.80	0.22	5.70	3.34	4.54	3.92	3.30
5.93	0.91	1.29	4.07	4.83	0.96	2.62	1.72	1.19	2.15	0.11	3.31	2.01	2.13	2.93	3.39
3.04	3.99	3.13	3.37	1.34	1.71	1.12	1.22	2.85	3.07	2.01	2.26	5.49	1.55	3.92	1.67
1.81	3.08	3.55	2.99	5.83	2.93	0.95	3.62	0.82	1.28	5.84	0.01	5.67	3.97	3.82	4.60
6.21	3.83	4.56	3.52	1.28	5.26	2.13	1.86	3.22	1.26	3.12	3.18	5.88	1.31	1.94	3.20
6.01	3.66	2.21	2.28	6.08	1.81	0.84	3.06	2.17	3.08	3.82	1.23	2.75	1.59	3.06	4.93
4.17	5.38	2.95	4.57	4.43	2.48	1.60	3.88	4.75	3.71	5.21	3.39	3.55	4.49	2.96	3.58
0.28	2.32	0.32	2.58	2.77	2.91	0.97	4.14	5.35	0.76	4.86	3.69	3.71	2.74	3.42	5.30
3.61	4.02	1.05	4.65	3.65	0.03	1.01	2.07	0.42	2.64	1.14	2.37	2.38	4.92	0.92	1.87
3.27	0.23	0.05	4.39	6.50	4.44	1.98	0.75	1.39	3.96	4.75	0.28	1.41	3.88	2.30	1.98
1.25	0.41	4.63	1.46	0.54	4.56	2.70	1.18	3.37	4.95	5.28	3.08	2.87	3.50	0.28	4.12
0.83	1.58	3.77	1.94	4.45	3.96	0.11	4.85	7.04	0.38	0.64	1.91	2.29	0.34	3.24	0.12

TABLE G.3 : 336 simulated truncated normal observations Y_t , where: $Y_t \sim N(2.4; 4)$; $Y_t \in R^+$

Time t	$(\theta_t D_{t-1})$		Obs. Y_t	$(\theta_t D_t)$	
	μ_t^*	σ_t^{*2}		μ_t	σ_t^2
330	2.3815	0.0174	2.20	2.3788	0.0174
331	2.3788	0.0174	1.28	2.3721	0.0173
332	2.3721	0.0173	4.35	2.3787	0.0173
333	2.3787	0.0173	5.22	2.3890	0.0172
334	2.3890	0.0172	1.79	2.3845	0.0172
335	2.3845	0.0172	4.07	2.3898	0.0171
336	2.3898	0.0171	5.27	2.4002	0.0171

TABLE G.4 : Prior - Posterior parameter distribution ; data from table G.3, using approximation for the posterior $(\theta_t | D_{t-1}) \sim N(\mu_t^*, \sigma_t^{*2}) ; (\theta_t | D_{t-1}) \in \mathbb{R}^+$
 $(\theta_t | D_t) \sim N(\mu_t ; \sigma_t^2) ; (\theta_t | D_t) \in \mathbb{R}^+ ; t=330, \dots, 336.$

Time t	$(\theta_t D_{t-1})$		Obs. y_t	$(\theta_t D_t)$	
	μ_t^*	σ_t^{*2}		μ_t	σ_t^2
330	2.3756	0.0174	2.20	2.3729	0.0174
331	2.3729	0.0174	1.28	2.3662	0.0173
332	2.3662	0.0173	4.35	2.3728	0.0173
333	2.3728	0.0173	5.22	2.3832	0.0172
334	2.3832	0.0172	1.79	2.3787	0.0172
335	2.3787	0.0172	4.07	2.3840	0.0171
336	2.3840	0.0171	5.27	2.3949	0.0171

TABLE G.5 : Prior-Posterior parameter distribution ; data from table G.3 ; using numerical integration for the $(\theta_t | D_{t-1}) \sim N(\mu_t^* ; \sigma_t^{*2}) ; (\theta_t | D_{t-1}) \in \mathbb{R}^+$
 $(\theta_t | D_t) \sim N(\mu_t ; \sigma_t^2) ; (\theta_t | D_t) \in \mathbb{R}^+ ; t=330, \dots, 336 .$

c	Aggregate Likelihood
20	56.20288
19	56.20290
17	56.20298
15	56.20322
14	56.20344
13	56.20375
12.5	56.20392
12	56.20407
11.50	56.20416
11.25	56.20417
11	56.20414
10.5	56.20390
10	56.20332
9	56.20037
8	56.19378
5	56.16381
3	56.23166
1.5	56.33833
0.5	55.43699

TABLE G.6 : c x Aggregate Likelihood data from table G.3 .

Time t	$(\theta_t D_{t-1})$		$(Y_t D_{t-1})$		Obs. y_t	$(\theta_t D_t)$	
	μ_t	σ_t^{*2}	μ_{y_t}	$\sigma_{y_t}^2$		μ_t	σ_t^2
330	2.3672	0.0210	2.3649	4.0210	2.20	2.3640	0.0209
331	2.3640	0.0210	2.3616	4.0210	1.28	2.3560	0.0209
332	2.3560	0.0209	2.3536	4.0210	4.35	2.3640	0.0209
333	2.3640	0.0209	2.3617	4.0210	5.22	2.3766	0.0209
334	2.3766	0.0209	2.3742	4.0209	1.79	2.3712	0.0208
335	2.3712	0.0209	2.3688	4.0209	4.07	2.3777	0.0208
336	2.3777	0.0209	2.3754	4.0209	5.27	2.3904	0.0208

TABLE G.7 : BEF truncated normal model for data from table G.3;

$$(\theta_t | D_{t-1}) \sim N(\mu_t^*, \sigma_t^{*2}) ; (\theta_t | D_{t-1}) \in \mathbb{R}^+$$

$$(Y_t | D_{t-1}) \sim N(\mu_{y_t}, \sigma_{y_t}^2) ; (Y_t | D_{t-1}) \in \mathbb{R}^+$$

$$(\theta_t | D_t) \sim N(\mu_t, \sigma_t^2) ; (\theta_t | D_t) \in \mathbb{R}^+ ; t=330, \dots, 336.$$

12	21	9	29	17	17	20	14	23	9	11	12	2	2	2	3
8	3	8	4	10	2	1	4	2	57	4	26	10	5	2	20
8	8	8	4	5	12	7	2	26	5	6	3	4	2	7	3
7	13	9	13	8	7	22	14	15	4	14	15	13	0	11	0
10	3	4	2	4	11	9	5	4	1	7	0	6	0	7	4
7	2	3	7	7	8	0	3	9	6	8	6	8	9	11	10
3	12	13	19	24	23	0	19	14	13	9	5	7	7	6	7
4	15	8	4	15	3	2	4	3	5	4	5	3	4	1	6
3	2	5	2	5	1	3	0	4	2	2	3	6	18	10	15
8	7	7	6	6	2	5	4	11	2	5	2	2			

TABLE G.8 : Weekly sales figures of children shoes, from 19/8/1966 to 28/11/1969 (157 observations). Model S225/7 .
Source: SATRO (Shoe & Allied Trades Research Association).

c	Aggregate Likelihood
3.0	7.931130
2.5	7.977485
2.0	8.088962
1.0	8.737608
0.8	8.968159
0.5	9.489371
0.2	10.64046
0.15	10.91206
0.10	11.03273
0.09	11.03542
0.08	11.02999
0.07	11.01083
0.05	10.93819
0.04	10.90288
0.02	10.85493

TABLE G.9 : $c \times$ Aggregate Likelihood data from table G.8 .

Time t	$(Y_t D_{t-1})$		Obs. Y_t
	μ_{y_t}	$\sigma_{y_t}^2$	
147	8.0099	51.4381	7
148	6.8666	49.9115	6
149	5.8520	46.3794	6
150	5.3897	43.9871	2
151	4.1656	35.3477	5
152	4.3443	36.3925	4
153	4.1546	35.1516	11
154	7.0345	56.3738	2
155	4.5191	37.7219	5
156	4.4907	37.4694	2
157	3.8015	33.2560	2

TABLE G.10 : BEF truncated normal model; predictive distribution ; data from table G.8

$$(Y_t | D_{t-1}) \sim N(\mu_{y_t}; \sigma_{y_t}^2); (Y_t | D_{t-1}) \in \mathbb{R}^+$$

t=147, ..., 157 .

Time t	$(Y_t D_{t-1})$		Obs. Y_t
	μ_{y_t}	$\sigma_{y_t}^2$	
147	8.3230	53.3819	7
148	7.6146	53.0296	6
149	7.1243	52.6660	6
150	5.5613	52.4814	2
151	5.3963	52.1162	5
152	4.9742	51.7734	4
153	6.8157	51.6556	11
154	5.3462	51.4578	2
155	5.2431	51.1226	5
156	4.2577	50.8639	2
157	3.5738	50.5799	2

TABLE G.11 : Normal linear model ; predictive distribution;

$$\text{data from table G.8 } (Y_t | D_{t-1}) \sim N(\mu_{y_t}; \sigma_{y_t}^2);$$

$(Y_t | D_{t-1}) \in \mathbb{R} \quad t=147, \dots, 157.$

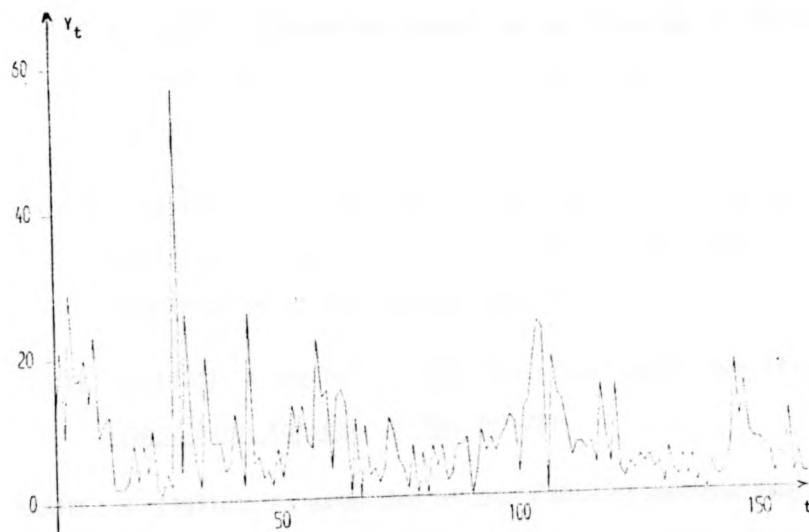


FIGURE G.1 : Plot of table G.8 data:
weekly sales figures for shoes covering the period
from 19/8/1966 to 28/11/1969 (157 observations).

REFERENCES

- ABRAMOWITZ, M. and STEGUN, I.A. (1965). Handbook of Mathematical Functions.
Dover Publication.
- AITCHISON, J. (1975). Goodness of prediction fit. Biometrika. 62, 547-554.
- AITCHISON, J. and DUNSMORE, I.R. (1975). Statistical Prediction Analysis.
Cambridge University Press.
- AKAIKE, H. (1971). Information theory and an extension of the maximum
likelihood principle. 2nd International Symposium on Information
Theory, Akademiai Kiado, Budapest, 267-281.
- AKAIKE, H. (1972). Use of an information theoretic quantity for Statistical
model identification. Proceedings of the Fifth Hawaii International
Conference on System Science, 249-250.
- AKAIKE, H. (1974). A new look at the Statistical model identification.
Trans. Auto. Control. AC-19, 267-281.
- AKAIKE, H. (1977a). An extension of the method of maximum likelihood and
the Stein's problem. Ann. Inst. Statist. Math. 29, Part A, 165-187.
- AKAIKE, H. (1977b). An objective use of Bayesian models. Ann. Inst. Statist.
Math. 29, Part A, 9-20.
- AKAIKE, H. (1977c). On entropy maximization principle. Applications of
Statistics. Ed. P.R. Krishnaiah. Amsterdam, North-Holland, 27-41.
- AKAIKE, H. (1978). A new look at the Bayes procedure. Biometrika. 65, 1,
53-59.

- BARNARD, G.A. (1951). The theory of Information. J. Roy. Statist. Soc. B 13, 46-64.
- BEVERIDGE, W.H. (1922). Wheat prices and rainfall in Western Europe. J. Roy. Statist. Soc. 85, 412-459.
- BOX, G.E.P. and JENKINS G.M. (1970). Time Series Analysis: Forecasting and Control. San Francisco: Holden Day.
- BROWN, G.R. and FISK C. (1975). A note on the entropy formulation of distribution models. Op1. Res. Q26, 4, 755-758.
- BRUBACHER, S.R. (1976). Time Series Modelling with instantaneous nonlinear transformations. Ph.D thesis. University of Lancaster.
- CLEVELAND, W.P. (1972). Analysis and Forecasting of seasonal time series. Ph.D. thesis. University of Wisconsin.
- CLIFF, A.D.; HAGGETT, P.; ORD, J.K.; BASSET, K. and DAVIES, D. (1975). Elements of Spatial Structure. Cambridge University Press.
- COHEN, A.C. (1949). On estimating the mean and standard deviation of truncated Normal distributions. Journ. Ann. Stat. Assn., 44, 518-525.
- CHOEN, A.C. (1950). Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. Ann. Math. Statist., 21, 557-569.
- COHEN, A.C. (1951). On estimating the mean and variance of singly truncated normal frequency distributions from the first three sample moments. Ann. Inst. Statist. Math., 3, 37-44.

- COHEN, A.C. (1955). Censored samples from truncated normal distributions. Biometrika, 42, 516-519.
- COHEN, A.C. (1957). On the solution of estimating equations for truncated and censored samples from normal populations. Biometrika, 44, 225-236.
- COHEN, A.C. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. Technometrics, 1, 3, 217-237.
- COZZOLINO, J.M. and ZAHNER, M.J. (1973). The maximum entropy distribution of the future market price of a stock. Opns. Res. 21, 1200-12.
- D'ARAUJO, R.P. (1974). Transformação e Estimção de Parametros para Modelos adaptados a previsão de séries temporais. Master thesis. PUC/RJ, Brazil.
- DERUSSO, P.M. ; ROY, R.J. and CLOSE, C.M. (1967). State Variables for Engineers. New York: John Wiley and Sons.
- DUTTA, M. (1966). On maximum (Information Theoretic) entropy estimation. Sankhya Ser A, 28, 319-328.
- EDWARDS, A.W.F. (1972). Likelihood. Cambridge University Press.
- FEINSTEIN, A. (1958). Foundations of Information Theory. New York: Mc. Graw-Hill.
- FLEHINGER, B.J. and LEWIS, P.A. (1959). Two-parameter lifetime distribution for Reliability studies of Renewal Process. IBM Journal, 58-63.
- FRANCIS, V.J. (1946). On the distribution of the sum of n sample values drawn from a truncated normal population. J. Roy. Statist. Soc., B, 8, 223-232.

- GILCHRIST, W. (1976). Statistical Forecasting. New York: Wiley-Interscience Publication.
- GOLDMAN, S. (1953). Information Theory. New Jersey, Prentice Hall Inc.
- HALD, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. Skandinavisk Aktuarietidskrift, 32, 119-134.
- HALPERIN, M (1952). Estimation in the truncated normal distribution. Journ. Ann. Stat. Assn., 47, 457-465.
- HARRISON, P.J. and STEVENS, C.F. (1971). A Bayesian Approach to short-term forecasting. Op1. Res. Q. 22, 341-362.
- HARRISON, P.J. and STEVENS, C.F. (1976a). Bayesian forecasting (with discussion). J. Roy, Statist. Soc. B 38, 205-247.
- HARRISON, P.J. and STEVENS, C.F. (1976b). Bayesian forecasting in action: case studies. Internal Report. Dept. of Statistics. University of Warwick.
- HASTINGS, N.A.J. and PEACOCK, J.B. (1974). Statistical Distributions. A Handbook for Students and Practitioners. London: The Butterworth Group.
- HOBSON, A. (1971). Concepts in Statistical Mechanics Gordon and Breach Science.
- HOGG, R.V. and CRAIG, A.T. (1970). Introduction to Mathematical Statistical. 3rd ed, Toronto: The Macmillan Co. of Canada.

- JAYNES, E.T. (1957). Information Theory and Statistical Mechanics.
Physical Review 106, 4, 620-630.
- JAYNES, E.T. (1958). Probability theory in Science and Engineering.
Colloquium lecture in Pure and Applied Science. Field Research
Laboratory. Socony Mobil Oil Co. Dallas, Texas.
- JAYNES, E.T. (1963). New Engineering Applications of Information Theory.
Symposium on Engineering Applications of Random Function Theory and
Probability. J.L. Bogdanoff and F. Kozin Eds. New York: Wiley.
- JAYNES, E.T. (1963a). Information Theory and Statistics Mechanics.
Statistical Physics 3 K.W. Ford, Ed. New York, Benjamin Inc.,
182-218.
- JAYNES, E.T. (1968). Prior Probabilities. IEEE Transactions on Systems,
Science and Cybernetics. SSC-4, 3, 227-241.
- JOHNSON, N.L. and KOTZ, S. (1969). Discrete Distributions. Boston:
Houghton Mifflins Co.
- JOHNSON, N.L. and KOTZ, S. (1970a). Continuous Univariate Distributions-1.
Boston: Houghton Mifflin Co.
- JOHNSON, N.L. and KOTZ, S. (1970b). Continuous Univariate Distributions-2.
Boston: Houghton Mifflin Co.
- KALMAN, R.E. (1960). A new approach to linear filtering and prediction
problem. Journal of Basic Engineering D82, 35-44.

- KALMAN, R.E. and BUCY, R.S. (1961). New results in linear filtering and prediction theory. Journal of Basic Engineering D83, 95-107.
- KHINTCHINE, A. (1932). Sulle successioni stazionarie di eventi. Giorn. Ist. Ital. Attuari 3, pp.267.
- KOLMOGOROFF, A.N. (1933). Giorn. Ist. Ital. Attuari 4, 83 .
- KOLMOGOROFF A.N. (1941). Interpolation and Extrapolation of Stationary random sequences. Bulletin of Academy of Science, 5. USSR, 3-14.
- KOLMOGOROFF, A.N. (1956). On the Shannon Theory of information transmission in the case of continuous signals. IRE Transactions on Information Theory, 102-108.
- KULLBACK, S. and LEIBLER, R.A. (1951). On information and sufficiency. Ann. Math. Statist. 22, 79-86.
- KULLBACK, S. (1959). Information and Statistics. New York : John Wiley and Sons.
- LEONARD, T. and HARRISON, P.J. (1977). Bayesian updating for the steady state Kalman Filter. (to appear in Technometrics).
- MACKAY, G.W. (1957). Quantum mechanics and Hilbert space. Amer. Math. Monthly 64, 45-57.
- MAKRIDAKIS, S. (1974). A survey of time series. INSEAD internal report.
- MANDELBROT, B. and TAYLOR, H. (1969). On the distribution of stock-price differences. Opns. Res. 15, 1057-1062.

- MANN, H.B. and WALD, A. (1943). On the statistical treatment of linear stochastic difference equations. Econometrics 11, 383, 173-220.
- MATHAI, A.M. and RATHIE, P.N. (1975). Basic Concepts on Information Theory and Statistics. New Delhi: Wiley Western Ltd.
- MEDITCH, J.S. (1969). Stochastic Optimal Linear Estimation and Control. New York: Mc Graw-Hill Book Co.
- MEHRA, R.K. (1976). A survey of time series modeling and forecasting methodology. Paper presented at the workshop on Recent Development in Real-Time Forecasting and Control of Water Resource Systems. Laxenburg, Austria.
- MEHRA, R.K. (1977a) . Unified state space forecasting for single and multiple time-series applications. To appear in Management science.
- MEHRA, R.K. (1977b) . State Space forecasting. Rapidata Publication, 1-34
- MEHRA, R.K. (1977c) . Kalman filter and their applications to forecasting To appear in Management Science.
- PÉREZ, A. (1957). Notions generalisees d'incertitude, d'entropie et d'information du point de vue de la théorie de martingales. Transactions of the first Prague Conference on Information Theory, 183-208.
- PIELOU, E.C. (1966), Shannon's formula as a measure of specific diversity: its use and misuse. Amer. Natur. 100, 463-465.
- PIELOU, E.C. (1967). The use of Information theory in the study of diversity in biological population. Proc. 5th Berkeley Symposium on Math. Stat. and Prob. 4, 163-172.

- PIELOU, E.C. (1969). An Introduction to Mathematical Ecology. New York: Wiley Interscience.
- PIINSKER, M.S. (1964). Information and Information Stability of Random Variables and Processes. Edited by A. Feinstein. San Francisco: Holden-Day Inc.
- PLANK, M. (1900). Zur Theorie des gesetzes der energieverteilung in normal spektrum. Verhandl. der Deutsch. Phys. Ges.
- RAIFFA, H. and SCHLAIFER, R. (1961). Applied Statistical Decision Theory Boston: The M.I.T. Press.
- RAJ, D. (1953). On moments estimation of the parameters of a normal population from singly and doubly truncated samples. Ganita, 4, 79-84.
- REGIER, M.H. and HAMDAN, M.A. (1971). Correlation in a Bivariate normal distribution with truncation in both variables. Aust. J. Stats., 13, 2, 77-82.
- REZZA, F.M. (1961). An Introduction to Information Theory. New York: Mc Graw-Hill.
- SAGE, A.P. and MELSA, J.L. (1971). Estimation Theory with Applications to Communications and Control. New York: Mc Graw-Hill Book Co.
- SCHUSTER, A. (1906). On the periodicities of sunspots. Philosophical Transactions A. 206, 69-81.
- SHAH, S.M. and JAISWAL, M.C. (1964). Estimation of parameters of doubly truncated normal distribution from first four sample moments. Ann. Inst. Statist. Math., 18, 107-111 .

- SHANNON, C.E. and WEAVER, W. (1949). The Mathematical Theory of Communications. Chicago: University of Illinois Press.
- SLUTSKY, E. (1937). The summation of random causes as the source of cyclic processes. Econometrica, 5, 105-146.
- SMITH, J.Q. (1978). A generalisation of the Bayesian steady state forecasting model (Part I). Internal Report. Dept. of Statistics. University of Warwick.
- SOUZA, R.C. and HARRISON, P.J. (1977). A Bayesian-Entropy forecasting approach. Internal Report. Dept. of Statistics. University of Warwick.
- SOUZA, R.C. (1974). Identificação e Aplicação de Testes para Modelos adaptados a Previsões de Séries Temporais. Master Thesis. PUC/RJ, Brazil.
- TALLIS, G.H. (1961). The moment generating function of the Truncated multi-normal distribution. J. Roy. Statist. Soc., B, 23, 233-239.
- THEIL, H. (1967). Economics and Information Theory. Rand McNally. Chicago: Illinois.
- TINTNER, G. (1960). Application of the theory of Information to the problem of weighted regression. Onore de Cornado Gini, 1. Rome: Inst. de Statist. Univ. degli studi, 29.
- TINTNER, G. and SASTRY, M.V.R. (1969). Information theory and the statistical estimation of Econometric relations. Institut für ökonometric. Tech. Hoch. Schule, Wien.

- TONG, H. (1975a). Determination of the order of a Markov chain by Akaike's Information Criterion. J. Appl. Prob. 12, 488-497.
- TONG, H. (1975b). Autoregressive model fitting with noisy data by Akaike's Information Criterion. IEEE Trans. on Information Theory, 476-480.
- TRIBUS, M. (1961a) . Thermostatic and Thermodynamics. Princeton: D. Van Nostrand Co.
- TRIBUS, E.T. (1961b) . Information theory as the basis for thermostatics and Thermodynamics. Journal of Applied Mechanics. B .
- TRIBUS, E.T. (1962). The use of Maximum Entropy estimate in the estimation of Reliability. Recent Developments in Information and Decision Process. Machol R.E. and Gray P. editors. 102-104.
- TRIBUS, M. (1969). Rational Descriptions, Decisions and Designs. New York: Pergamon Press.
- VASICEK, O. (1974). A test for normality based on sample entropy. Working paper 7430. Graduate School of Managment, University of Rochester.
- VINCZE, I. (1959). An interpretation of the I-divergence of Information Theory. Transactions of the 2nd Prague Conference on Information Theory, 681-684.
- VINCZE, I. (1965). Some questions concerning the probabilistic concept of Information. IIS and AHS Selected Translations in Math. Stat.5, 373-380.

VINCZE, I. (1972). On the maximum probability principle in Statistical Physics. Colloquia Mathematica Societatis Janos Bolyai. 9 European meeting of Statisticians, Budapest.

WALKER, A.M. (1931). On the periodicity in series of related terms. Proceedings of the Royal Society of London 131A, 518-532.

WEILER, H. (1959). Means and standard deviations of a truncated normal bivariate distribution. Aust. J. Stats., 1, 73-81.

WOLD, H.O. (1938). A study in the analysis of stationary time series. Stockholm: Almqvist and Wiksell.

YULE, G.U. (1927). On the method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Philosophical Transactions A.226, 267-298.