ASPECTS OF RECURSIVE BAYESIAN ESTIMATION.

BY

MICHAEL WEST, B.Sc.

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy, May 1982.

CONTENTS.

1. Introduction.

1.1 General introduction.

1.2 Outline of thesis.

1.3 Notation and Terminology.

2. Concepts of robust Bayesian estimation.

2.1 Introduction to basic ideas.

2.1.1 General comments.

2.1.2 Score/Influence functions.

2.2 Processing a single of observation: unknown location parameter.

2.2.1 Background.

2.2.2 Weak prior information.

2.2.3 Strong prior information.

2.3 Scale parameter case.

2.3.1 Simple scale estimation.

2.3.2 Location/scale problem.

Appendices A. 1 Lemma 2.2.1.

2 Survey of unimodal, symmetric heavy-tailed distributions.

3 Scale mixtures of normals.

3. The Dynamic Linear Model.

3.1 Introduction: Model and Outlook.

3.2 Robustifying the Kalman Filter: Scalar observations.

3.2.1 General approach and prior beliefs.

3.2.2 The work of Masreliez and Martin.

3.2.3 The Gradient algorithm.

3.2.4 The Modal algorithm.

3.2.5 Scale mixtures of normals.

BEST COPY

AVAILABLE

TEXT IN ORIGINAL IS CLOSE TO THE EDGE OF THE PAGE

3.2.6 Numerical studies.

3.3 Further analysis.

3.3.1 Prediction.

3.3.2 Smoothing.

3.4 Extension to general model: vector observations.

4. The D.L.M.: Scale problems.

4.1 Introduction.

4.2 The normal model.

4.2.1 Scale parameter case.

4.2.2 Unknown covariance structure.

4.2.3 Numerical studies.

4.3 Non-normality.

4.3.1 Robust (sequential) estimation of scale parameters in the D.L.M.

4.3.2 Prediction and smoothing.

4.3.3 Vector observations: unknown covariance structure.

Appendices 4. 1 Lemma 4.1.

2 Lemma 4.2.

5. Classical Time Series Models.

5.1 Autoregressions and outlier type.

5.1.1 Innovations outliers.

5.1.2 Additive outliers.

5.1.3 Outliers of unknown type.

5.2 Outliers in Autoregressive-Moving Average Models.

Appendix 5.

6. Asymptotic Theory of Recursive Algorithms.

6.1 Introduction: Stochastic Approximation.

6.2 Recursive estimation: location and regression models.

6.2.1 Scalar location problem.

6.2.2 General Linear Model.

6.3 Scale estimation.

6.3.1 Simple scale estimation.

6.3.2 Joint regression/scale estimation.

Appendix 6. Theorem A6.1.

7. Multi-State Modelling: A case study and theoretical extensions.

7.1 Introduction.

7.2 Renal Transplant Project.

7.2.1 Transplant data: history and details.

7.2.2 The system model: physical considerations and error structure.

7.3 The Dynamic Linear Multi-State Model.

7.3.1 Linear Growth model.

7.3.2 Learning about variation in the data.

7.4 Implementation.

7.4.1 Graphical output.

7.4.2 Inferences and Decisions.

7.5 Conclusions.

Appendix 7. Approximating the error structure with a tractible form.

ABSTRACT.

This thesis is concerned with the theoretical and practical aspects of some problems in Bayesian time series analysis and recursive estimation. In particular, we examine procedures for accommodating outliers in dynamic linear models which involve the use of heavy-tailed error distributions as alternatives to normality.

Initially we discuss the basic principles of the Bayesian approach to robust estimation in general, and develop those ideas in the context of linear time series models. Following this, the main body of the thesis attacks the problem of intractibility of analysis under outlier accommodating assumptions. For both the dynamic linear model and the classical autoregressive-moving average schemes we develop methods for parameter estimation, forecasting and smoothing with non-normal data. This involves the theoretical examination of non-linear recursive filtering algorithms as robust alternatives to the Kalman filter and numerical examples of the use of these procedures on simulated data. The asymptotic behaviour of some special recursions is also detailed in connection with the theory of stochastic approximation.

Finally, we report on an application of Bayesian time series analysis in the monitoring of medical time series, the particular problem involving kidney transplant patients.

ACKNOWLEDGEMENTS.

I should like to express my sincere gratitude to Professor A.F.M. Smith for his efforts in initiating this research, his continued suggestions, guidance and general support throughout the period of study.

I am also grateful to Mr. Ian Trimble for his contributions to the understanding and development of the problems of the final Chapter of this thesis, and my thanks are due to Miss Jean Taylor for her patient and careful typing of the original manuscript.

This research was funded by the Science and Engineering Research Council of Great Britain.

To Lauren.

Chapter 1. Introduction.

1.1. General Introduction.

Recent years have seen a continuing development and flourishing of the applications of mathematical statistics to interesting and important problems in many areas. Time series analysis in particular has received an enormous amount of interest from workers in socio-economic studies, the physical and engineering sciences, and the life sciences and medicine.

In time series studies, as in all statistical modelling, much thought and effort is required in the development of a framework for analysis, the model to be embedded in that framework, and the study of the characteristics of the model. This thesis is devoted to the examination of a wide class of flexible time series models in a Bayesian framework.

(a) Bayesian framework.

The philosophical basis of the Bayesian approach to statistics is still somewhat controversial. However, it is indisputable that this framework provides a rich and logical backgrop for mathematical modelling and has well-defined procedures for handling uncertainty and producing inferences. Furthermore, the combination of prior and experimental information is both rigorous and natural and the intimate relationship between Bayesian statistics and decision theory results in a comprehensive framework for the utilization of results.

I believe that the need for a theory satisfying these, and other, practical requirements is in part responsible for the current growing interest in Bayesian methods and that, in future, many more important applications will be seen to use, and indeed demand, a Bayesian approach.

(b) Dynamic Linear Models.

The extension of Bayesian methods for linear models to the dynamic linear models discussed by Harrison and Stevens (1976) has provided a class of models capable of imitating the behaviour of many observed time series. Although relatively new to the statistical literature, such

models have been used in engineering applications for some time with Kalman (1963) detailing the basic analysis. The flexibility of these models, and their potential as aids to understanding physical systems as well as forecasting, suggests that more interest will be centred on their application in diverse fields in the near future.

(c) Outliers and robustness.

Another growth area in statistics in recent years has been the study of robustness and outliers in statistical data. On the Bayesian approach, Box and Tiao (1962, 1968) provided early contributions to the literature, discussing the ideas more fully in Box and Tiao (1973). In general, a procedure which is based on a parametric model can be made robust against the assumptions of that model by an extension to a wider class of a priori plausible models each of which is used to analyse the data. Bayes' Theorem is then used to provide a posterior distribution for the class of models entertained.

Concerning the treatment of outliers within this framework, an observation which is outlying relative to a particular model can be accommodated in an analysis by considering a further model in which it does not outly. In simple models, inference in the presence of outliers using such an approach is well developed, with the works of Box and Tiao (1973) and Ramsay & Novick (1980) being particularly relevant. O'Hagan (1979) discusses the location estimation problem and examines outlier accommodating models relevant to the Bayesian approach and provides a starting point for the development of more complex models that are the subject of this thesis.

1.2. Outline of thesis.

Chapter 2 presents a discussion of some general concepts of Bayesian estimation and the roles of prior and likelihood assumptions. In the special robust location estimation problem, we examine the recursive updating of beliefs with reference in particular to likelihood characteristics and the consequences for posterior distributions and hence inferences. O'Hagan (1979) discusses ideas applicable to this simple model and we consider this, along with the work of Masreliez (1975), in investigating possible error densities as alternatives to normality. Other works in robust Bayesian estimation, including Box and Tiao (1973), Box (1980), and Ramsay and Novick (1980), are discussed, and parallels are drawn with the major classical approaches of Huber (1964, 1977) and Hampel (1974).

In Chapter 3 we examine approximate Bayesian methods for estimation in dynamic linear models. Masreliez (1975) and Masreliez and Martin (1977) developed useful recursive algorithms as approximations to the intractible Bayesian analysis of state-space models with heavy-tailed, non-normal error densities. We discuss several serious problems associated with these schemes and develop alternatives which, in addition to solving these problems, provide a strong framework for the calculation of approximations to the posterior and predictive distributions of interest. The resulting schemes have considerable intuitive appeal and are rather closely connected to the mixture modelling approach of Harrison and Stevens (1976). Indeed the latter can be seen to be a special case of our model.

Chapter 4 is concerned with the estimation of scale and covariance parameters. We examine briefly some schemes for sequential estimation of scale parameters and covariance matrices of multivariate time series under the usual assumption of normality. Then we turn to heavy-tailed distributions and develop methods for scale estimation which complement and extend the non-linear filtering algorithms already presented for the dynamic linear model.

In Chapter 5 we move away from dynamic Bayesian linear models to classical autoregressive-moving average schemes, thoroughly discussed by Box and Jenkins (1971). Despite the vast amount of research effort that has been devoted to the theory and applications of such models, relatively little has appeared on the robust estimation problem. Notable exceptions are the works of Fox (1972), Abraham and Box (1979), Kleiner et al (1979) and Martin (1978, 1979), all of whom concentrate on pure AR models. Fox distinguishes two types of outliers that occur and require different models, one of which is inherently non-linear in the parameters to be estimated. Our analyses provide general Bayesian methods for both types of outliers in ARMA models based on an extension of the state space representation discussed by Priestley (1978) and the techniques of Chapters 3 and 4.

Chapter 5 considers the mathematical form of some of the above mentioned recursive algorithms in the special case of constant parameters. Here we adopt and extend methods of stochastic approximation, developed and used by Robbins and Monro (1951), Kashyap, Blaydon and Fu (1970), Fabian (1978) and Martin and Masreliez (1975), and use this to examine the asymptotic consistency of filtering algorithms with the Bayesian analysis. The works of Berk (1966) and Heyde and Johnstone (1978) in asymptotic Bayesian theory are relevant here.

Finally, Chapter 7 consists of a report on an application of Bayesian time series modelling in a medical problem. Whilst studying the mainly theoretical work of earlier Chapters, the opportunity arose for participation in a project involving the Mathematics department at Nottingham University and the Renal Unit at Nottingham City Hospital. Early developments were reported by Smith and Cook (1980) and this Chapter discusses our more recent contributions to this continuing project.

1.3. Notation and terminology.

An attempt has been made to follow the style of notation of related works, choosing the simplest form where previous authors have differed.

Throughout vectors are underlined,as $\underset{\sim}{x}$, for example.Matrices appear as capital letters, both Greek and Roman. We make no distinction between random variables and their realized values since, generally, the context will be unambiguous. All probability distributions are defined via densities over Euclidean spaces with respect to Lebesque measure. Such densities are represented by the generic symbols p, f and π and the arguments follow standard notation. For example $p(x)$, $p(x|y)$ are the densities of x and of x given y respectively. The following special densities are used repeatedly;

$N_\theta[m,c]$ is the normal density of θ with mean m and variance c. The multivariate form is $N_{\underset{\sim}{\theta}}[\underset{\sim}{m},C]$.

$G_\lambda[a,b]$ is the gamma density of λ, proportional to $\lambda^{a-1}e^{-\lambda b}$ for $\lambda>0$.

For a (p×p) positive definite symmetric matrix Λ, $W_\Lambda[a,B]$ is the Wishart density, proportional to $|\Lambda|^{(a-p-1)/2}\exp\{-\frac{1}{2}\ \mathrm{trace}(B\Lambda)\}$.

In all cases the subscripts will be dropped when context allows and we shall write $x \sim p$ when x has density p. For example

$$x|y \sim N[m,c]$$

when the conditional distribution of x given y is normal with mean m and variance c.

Further notation will be defined as necessary.

CHAPTER 2. Concepts of robust Bayesian estimation.

2.1 Introduction

2.1.1 General Comments

Given a scalar parameter θ about which we have some prior knowledge formally described by a prior distribution function, how do we assess the influence of a single observation, y, say, on our beliefs about θ? For any given likelihood, the effect of y is totally described and all we need do is examine the posterior distribution of θ given y. In this Chapter we consider initially a general parameter θ and discuss the notions of the influence of an observation and its relationship with the concept of an outlier. We than concentrate on the special cases of location and scale parameters and attempt to build a framework on which to base the analysis of more complex models of later Chapters.

Our main interest lies in obtaining robust methods of estimation and thus we concentrate on the effects of likelihood assumptions, taking the prior to be unquestioned. Chapter 4 deals with a special model where the prior must be treated as suspect as well and we leave further discussion to that Chapter. For the moment we concentrate on the likelihood and the data y and we shall see that, in a particular model, the robustification of estimation procedures is achieved by separating the concepts of influential observation and outlier. A non-robust analysis is such that influence increases as consistency with the prior decreases i.e. as the datum becomes more and more aberrant. For a robust analysis, the influence of the observation reaches a peak and then begins to decay as the observation becomes more aberrant.

2.1.2 Score/influence functions.

Consider a general scalar parameter θ with prior density $\pi(\theta)$ and a single observation y related to θ through a likelihood $p(y|\theta)$. We consider now the influence y has on our beliefs about the unobservable θ.

In exploring features of the posterior distribution for θ given y an important step is to consider most likely values i.e. the posterior modes. Let θ^* be such a value. Then, if both π and p are differentiable in θ, we have

$$g_\theta(\theta^*|y) = 0 \qquad (2.1.1)$$

where

$$g_\theta(\theta|y) = -\frac{\partial}{\partial\theta} \ln p(\theta|y)$$

$$= -\frac{\partial}{\partial\theta} \ln \pi(\theta) - \frac{\partial}{\partial\theta} \ln p(y|\theta)$$

$$= g_\theta(\theta) + g_\theta(y|\theta) \text{ , say} \qquad (2.1.2)$$

in an obvious notation. We recognize $g_\theta(y|\theta)$ as the efficient score function (of $p(y|\theta)$ with respect to θ), see for example Cox and Hinkley (1974), and following this we call $g_\theta(\theta)$ and $g_\theta(\theta|y)$ the prior and posterior score functions respectively. Thus (2.1.2) rephrases the multiplicative Bayes' Theorem in the additive form

posterior score = prior score + likelihood score

and this form is particularly useful in examining the behaviour of the posterior as y varies for a given likelihood, and as prior/likelihood characteristics vary. In this context, Ramsay and Novick (1980) have termed $g_\theta(\theta|y)$ the influence function of $p(\theta|y)$ with respect to θ, and introduce the concept of P- (for prior) robustness by calling $\pi(\theta)$ P- robust if $g_\theta(\theta)$ is bounded with respect to θ.

In a different vein, Box (1980) discusses the use of the likelihood score, $g_\theta(y|\theta_0)$, as a measure of the discrepancy from a parameter value of θ_0 as indicated by the observed data y, for a given model. We can interpret this by noting that a large "discrepancy" leads to a large "difference" between prior and posterior at θ_0 as measured by the difference in score functions there. Similarly, consistency of the data y with a value θ_0 means only a small change in score.

Turning now to sensitivity of $p(\theta|y)$ to y, we define similar score functions, but now with respect to y. Subject to differentiability assumptions,

$$g_y(\theta|y) = -\frac{\partial}{\partial y} \ln p(\theta|y) \qquad (2.1.3)$$

$$= -\frac{\partial}{\partial y} \ln p(y|\theta) + \frac{\partial}{\partial y} \ln p(y)$$

$$= g_y(y|\theta) - g_y(y) \quad , \text{ say,} \qquad (2.1.4)$$

in an obvious notation. Here $g_y(y|\theta)$ is the score function of the likelihood with respect to y and $g_y(y)$ is that of the marginal (or predictive) density of y. Parelleling Box's use of $g_\theta(y|\theta)$, we can interpret $g_y(y_0|\theta)$ as a measure of the discrepancy of an observation y_0 at a parameter value θ. The posterior score $g_y(\theta|y)$ then measures the "influence" of the observation which will be large when the likelihood score is large relative to the marginal score. This latter function is given by

$$g_y(y) = E[g_y(y|\theta)|y] \qquad (2.1.5)$$

i.e. the marginal score is the posterior expectation of the likelihood score. This result is proved in Appendix A2.1, Lemma 2.1.1, and proves extremely useful in later sections.

Again, bounding the likelihood score function in y will be a primary requirement when modelling with protection against outliers in mind. This is termed L- robustness by Ramsay and Novick (1980) and provides a point of contact with the classical theory of robust estimation where the likelihood score coincides with the influence function of an M- estimator in location problems (Hampel, 1974). Now within the classical framework, the influence function provides a qualitative means of assessing the influence of particular observations on the behaviour of sampling theory procedures. In particular in studying the behaviour of estimators defined as functionals of the empirical distribution function in the case of i.i.d. random variables, the influence function at a point x essentially measures the influence of an additional observation at x when the sample size tends to infinity. This final point is important; the influence function is an asymptotic concept and is thus independent of the sample.

Within a coherent framework, we have seen how the likelihood score function determines the sensitivity of the posterior distribution to a single observation. Considering a random sample $\{y_1, \ldots, y_n\} = \underset{\sim}{y}$, we obtain an analagous relation between score functions given by

$$g_{y_j}(\theta|\underset{\sim}{y}) = g_{y_j}(y_j|\theta) - g_{y_j}(\underset{\sim}{y}) \qquad (2.1.6)$$

where, now,

$$g_{y_j}(\theta|\underset{\sim}{y}) = -\frac{\partial}{\partial y_j} \ln p(\theta|\underset{\sim}{y}),$$

$$g_{y_j}(y_j|\theta) = g_y(y_j|\theta) = -\frac{\partial}{\partial y} \ln p(y|\theta)\Big|_{y=y_j},$$

and

$$g_{y_j}(\underset{\sim}{y}) = -\frac{\partial}{\partial y_j} \ln p(\underset{\sim}{y}) \quad , \quad j=1,\ldots,n.$$

Now the marginal score $g_{y_j}(\underset{\sim}{y})$ is given by

$$g_{y_j}(\underset{\sim}{y}) = E\left[g_{y_j}(y_j|\theta)\,\middle|\,\underset{\sim}{y}\right]$$

as in Ramsay and Novick (1980). So the influence of y_j at θ, as defined by the posterior score with respect to y_j, is measured by the likelihood score (or influence function) minus its expected value given the sample. Further discussion is given in Ramsay and Novick.

Our main point is that the influence of a particular observation must be gauged relative to any other available data, (even in the simplest context of a random sample i.e. independent observations), and purely asymptotic considerations relating to the likelihood will not suffice.

This line of thought is developed extensively in later Chapters where, in a sequential processing of observations, the influence of an observation is measured naturally by reference to the "prior", which depends on past data.

We follow up this idea now for the location parameter problem.

2.2 Location parameter case

2.2.1 Introduction

The likelihood now has the form $p(y|\theta) = p(y-\theta)$, and we identify the various likelihood score functions via

$$g_y(y|\theta) = -\,g_\theta(y|\theta) = g(y-\theta), \text{ say}$$

where $g(u)$ is <u>the</u> score function of $p(u)$, $-\frac{\partial}{\partial u}\ln p(u)$, $u \in \mathbb{R}$.

Since our intention is to examine various error densities as alternatives to normality we assume that

(i) $p(u)$ is unimodal and symmetric about zero, and positive for all u.

(ii) $p(u)$ is twice (piece-wise) differentiable in u.

O'Hagan (1979) introduces the important concept of <u>outlier-proneness</u> of a distribution within this framework. Let $y_1, \ldots, y_n$ be independent, identically distributed with density $p(y-\theta)$. For $r=1,2,\ldots, n$, define the observation set D_r by $D_r = \{y_1, \ldots, y_r\}$. Then the distribution whose density is $p(\cdot)$ is said to be <u>outlier-prone of order r</u> if

$$\lim_{|y_{r+1}| \to \infty} \int_{-\infty}^{t} \left[p(\theta|D_{r+1}) - p(\theta|D_r)\right] d\theta = 0, \qquad (2.1.7)$$

for all real t and <u>any</u> prior $\pi(\theta)$.

In fact O'Hagan distinguishes left and right outlier-proneness according as $y_{n+1} \to \pm\infty$; since $p(y)$ is symmetric we do not need this distinction. Clearly, as noted by O'Hagan, outlier-proneness of order 1, relevant in the case of two observations, is the strongest property, implying outlier-proneness of order n for all n.

Now, using the earlier results of Dawid (1973), O'Hagan proves that a distribution with density $p(\cdot)$ satisfying the above conditions is outlier-prone of order 1 (or just outlier-prone) if, additionally, the following conditions are satisfied;

(iii) for all $\epsilon>0$, $h>0$, there exists A such that, when $y>A$ and $|y'-y|<h$,

$$|p(y') - p(y)| < \epsilon\, p(y) \qquad (2.1.8)$$

(iv) there exists B such that, for all $y>B$,

$$-\frac{\partial}{\partial y} \ln p(y) \text{ is decreasing in } y. \qquad (2.1.9)$$

NB the assumed symmetry imposes similar conditions on the left-hand tail of $p(y)$.

Condition (iii) requires that $p(y)$ be essentially uniform in the tails and we shall see that this restricts the rate of decay of $p(y)$ to be no greater than $\exp\{-k|y|\}$ as $|y| \to \infty$. Condition (iv) then classifies $p(y)$ according to the behaviour of the score function. We refer to likelihoods which satisfy these conditions as robust, and examine in some depth some examples in the Appendix A2.2.

Returning now to the posterior θ-score,

$$g_\theta(\theta|y) = g_\theta(\theta) + g(\theta-y),$$

we have pointwise convergence of posterior to prior θ-score whenever $p(y)$ is robust. Note however that this occurs when $g(y) \to 0$ as $|y| \to \infty$ without the monotonicity requirement of (iv), and so, from a practical point of view, robust estimation can be achieved with, for example, posterior modes converging to prior modes as $|y| \to \infty$, without the guarantee of convergence of distribution functions.

What are the implications of these ideas for our problem of updating beliefs described by a prior $\pi(\theta)$ on receiving a single observation y with likelihood $p(y-\theta)$? Clearly, in addition to assuming a robust likelihood, we require a certain strength of prior information in order that an aberrant observation be discredited. O'Hagan gives conditions which determine this, as follows;

Let $m(\theta)$ be any prior measurable function. Then, if

$$\text{(v)} \quad \int_{-\infty}^{\infty} \pi(\theta)\, p(\theta)^{-1}\, d\theta < \infty \qquad (2.1.10)$$

and

$$\text{(vi)} \quad \int_{-\infty}^{\infty} |m(\theta)|\pi(\theta)\, p(\theta)^{-1} d\theta < \infty \qquad (2.1.11)$$

we have $$\lim_{|y| \to \infty} \int_{-\infty}^{t} [p(\theta|y) - \pi(\theta)]d\theta = 0$$

for all real t, and, in addition,

$$\lim_{|y| \to \infty} E[m(\theta)|y] - E[m(\theta)] = 0.$$

This result holds when p dominates π in the sense of (v) and (vi).

We provide an example.

Example 2.2.1

Let prior and likelihood be Student t, thus providing both P- and L- robustness in the terminology of Ramsay and Novick.

So

$$\pi(\theta) \propto [h + (\theta-m)^2]^{-(h+1)/2} ,$$

and

$$p(y-\theta) \propto [k + (y-\theta)^2]^{-(k+1)/2} , \quad h, k > 0,$$

where m is the point of symmetry of the prior (the prior mean if $h > 1$). We can distinguish the following cases, depending upon the degrees of freedom parameters h and k.

Let $r = h-k$.

(a) $r > 1$.

Our prior beliefs are "stronger" than our beliefs about the likelihood in the sense that the degrees of freedom parameter is larger. Clearly (v) holds and so the posterior distribution converges to the prior as $|y| \to \infty$. For posterior moments, note that

$$E[\theta^P|y] < \infty \quad \text{for } p < h+k+1,$$

whilst

$$E[\theta^P] < \infty \quad \text{for } p < k.$$

so the result of O'Hagan ensures convergence of moments only for $p < r-1$.

b) $\underline{r < -1}$.

The symmetry between π and p means that the roles are reversed; the discussion of a) is relevant with prior and likelihood interchanged, m replacing y and y, m.

c) $\underline{|r| \leq 1}$.

Now neither π nor p is dominant as defined by (v). For h=k,

$$p(\theta|y) \propto \{[k + (\theta-y)^2]\ [k + (\theta-m)^2]\}^{-(k+1)/2}$$

Clearly, if $\bar{\theta} = (y+m)/2$, then $p(\theta|y)$ is symmetric about $\bar{\theta}$ for all y and m. As $|y|$ increases, $p(\theta|y)$ becomes bimodal with one mode tending to the fixed value m, the other following y, as can be seen from the (cubic) modal equation. Similar behaviour is evidenced with other values of r in this interval, and indeed, for other symmetric distributions as we discuss in the next section.

2.2.2. Weak prior information.

We examine now a special case of the above framework when neither π nor p dominates the other in the sense of the condition (v), (2.2.10). We consider the case of π and p having the same functional form, with π having point of symmetry m;

i.e.: $$\pi(\theta) = p(\theta-m).$$

It is clear that we cannot now distinguish an aberrant observation y from an "aberrant" prior specification since extremeness of y corresponds to $|y-m| \to \infty$. Thus outlier rejection will not be obtained. What does happen?

Symmetry considerations.

The assumed symmetry of p leads to the observation that, as in example 2.2.1, if $\bar{\theta} = (y+m)/2$ then

$$p(\bar{\theta}+x|y) \propto p\left((y-m)/2+x\right).p\left((m-y)/2+x\right)$$

$$\propto p\left((m-y)/2-x\right).p\left((y-m)/2-x\right)$$

, by symmetry of p,

$$\propto p(\bar{\theta}-x|y), \quad \text{for all real } x.$$

So $p(\theta|y)$ is symmetric about $\bar{\theta}$ and, since $E[\theta|y] < \infty$, $\bar{\theta}$ is the posterior mean. Further, $\bar{\theta}$ is always a candidate for a posterior mode since

$$g_\theta(\theta|y) = g(\theta-m) + g(\theta-y)$$

$$= 0 \text{ at } \theta = \bar{\theta} \text{ since } g(-u) = -g(u).$$

Now when $|y-m|$ is large, we expect bimodality of $p(\theta|y)$ whenever g is redescending, so $\bar{\theta}$ will be the location of a minimum of $p(\theta|y)$ with two modes symmetrically located about $\bar{\theta}$. Consider the following example.

Example 2.2.2.

Take a Cauchy density, $p(u) \times [1+u^2]^{-1}$. Then

$$g_\theta(\theta|y) = \frac{(\theta-m)}{[1+(\theta-m)^2]} + \frac{(\theta-y)}{[1+(\theta-y^2)]}$$

$$= 0$$

implies

$$\phi[1+(\phi-3)^2] + (\phi-3)\ [1+\phi^2] = 0$$

where

$$\phi = \theta-m \text{ and } z = y-m.$$

Thus $2\phi^3-3\phi^2z + \phi(2+z^2) - y = 0$

or $(2\phi-z)(\phi^2-\phi z+1) = 0,$

having solutions,

$$\phi_0 = z/2 \quad \text{or } \theta_0 = \overline{\theta} = (y+m)/2,$$

and $$\phi_{1,2} = z/2 \text{ I } \sqrt{\{z^2-4\}}\ /2\ ,$$

or $$\theta_{1,2} = \theta_0 \pm \sqrt{\{(y-m)^2-4}\ /2\ .$$

Thus θ_0 is the mode if $|y-m| \leq 2$, minimum if not. Clearly also

$$\lim_{y-m} \begin{cases} \theta_1 = m\ , \\ \theta_2 = y\ . \end{cases}$$

In general the above discussion does not carry over to the case of different scale factors in π and p since the symmetry breaks down. However we can obtain a feel for the form of behaviour by considering a particular family of densities.

Special case: the Stable distributions.

Assume that both π and p are symmetric stable of index a, $1 \leq a \leq 2$, with characteristic functions

$$\chi_\theta(t) = \exp\ \{\ imt - |ct|^a\ \}\ ,$$

and

$$\chi_\theta(t) = \exp\ \{\ i\theta t - |st|^a\ \}\ , \qquad c, s > 0.$$

When c=s the discussion above is relevant. Otherwise we can still obtain an expression for the posterior mean which is an intuitively appealing generalization of the normal theory result corresponding to a=2.

Lemma 2.2.2 In the above framework

$$E[\theta|y] = (c^a+s^a)^{-1}(c^a y+s^a m).$$

Proof. Define $\phi = \theta - m$ and $z = y - m$. Then the joint characteristic function of z and ϕ is

$$\chi_{z,\phi}(u,v) = E\left[e^{iuz+iv\phi}\right]$$

$$= E\left[E\left[e^{iuz}|\phi\right].\ e^{iv\phi}\right]$$

$$= E\left[e^{i(u+v)\phi-|su|^a}\right]$$

$$= \exp\left\{-|su|^a-|c(u+v)|^a\right\} .$$

Now it is shown by Lukacs and Laha (1971), Lemma 6.3.1, that, for any two random variables z,ϕ,

$$E[\phi|z] = \alpha z \text{ if and only if}$$

$$\frac{\partial}{\partial v}\chi_{z,\phi}(u,v)\Big|_{v=0} = \alpha.\ \frac{\partial}{\partial u}\chi_{z,\phi}(u,0), \text{ for all } u.$$

In our case

$$\frac{\partial}{\partial v}\chi_{3,\phi}(u,v) = \begin{cases} \chi_{3,\phi}(u,v).\ \left\{-c^a.a.|u+v|^{a-1}.\ \operatorname{sgn}(u+v)\right\}, & 0 \neq u+v, \\ 0, & u+v=0, \end{cases}$$

or $\frac{\partial}{\partial v}\chi_{3,\phi}(u,v) = \chi_{3,\phi}(u,0).\ \left\{-a.c^a.|u|^{a-1}.\ \operatorname{sgn}(u)\right\}$,

for all u.

Further, $\frac{\partial}{\partial u}\chi_{3,\phi}(u,0) = \chi_{3,\phi}(u,0).\ \left\{-a(c^a+s^a).\ |u|^{a-1}.\ \operatorname{sgn}(u)\right\}$.

Applying the above quoted result we see that

$$E[\phi|z] = \alpha z \text{ for } \alpha = c^a.(c^a+s^a)^{-1}.$$

Hence, transforming back to $\theta = \phi+m$ and $y = z+m$ we have

$$E[\phi|y] = m+c^a.(c^a+s^a)^{-1}.(y-m)$$

and the result follows.

So $\lim_{|y-m| \to \infty} E[\theta|y]$ does not exist. It is intuitively clear that the posterior variance should diverge too, in this case and for other densities. We can show this to be true for the special case a=1, the Cauchy density.

Example 2.2.3.

Set a=1, $\phi = c^{-1}(\theta-m)$, $z = c^{-1}(y-m)$ and $k = c^{-1}s$. Then by the above Lemma,

$$E[\phi|y] = z.(1+k)^{-1}.$$

Now $E[\phi^2|z] = p(z)^{-1} \int_{-\infty}^{\infty} k\phi^2\pi^{-2}[1+\phi^2]^{-1}[k^2+(\phi-z)^2]^{-1}\, d\phi$

$$< \infty.$$

Further $p(z)^{-1} = \pi(1+k)^{-1}[(1+k)^2+z^2]$.

Partial fractions expansion of the above integrand leads to

$$\pi^{-1}p(z)^{-1} \int_{-\infty}^{\infty} \left\{ \frac{k(a\phi+b)}{\pi[1+\phi^2]} + \frac{k[e(\phi-z)+d]}{\pi[k^2+(\phi-z)^2]} \right\} d\phi$$

where, setting $r^2 = k^2+z^2$,

$$d = [2z^2-r^2(1-r^2)].[4z^2+(1-r^2)^2]^{-1},$$

and

$$b = (1-r^2).[4z^2+(1-r^2)^2]^{-1}.$$

The integral exists with the integrands involving a and e contributing nothing. Thus

$$E[\phi^2|z] = \pi^{-1}p(z)^{-1}[kb+d].$$

On substituting for p(z), b, d we obtain

$$\text{var}[\theta|y] = c^2\, \text{var}[\phi|z] = c^2\{E[\phi^2|z] - E[\phi|z]^2\}$$

$$= c^2 \left[\frac{(r^2-k)(r^2-1)+2z^2}{4z^2+(1-r^2)^2} \right] \left[\frac{(1+k)^2+z^2}{(1+k)} \right] - \frac{c^2z^2}{(1+k)^2}$$

$$= \left[\frac{c^2(r^2-k)(r^2-1)+2(y-m)^2}{4(y-m)^2+(1-r^2)^2c^2}\right]\left[\frac{(c+s)^2+(y-m)^2}{(k+1)}\right]$$

$$- \frac{c^2(y-m)^2}{(c+s)^2} \quad .$$

This complicated expression simplifies when c=s=1 to

$$\text{var}[\theta|y] = 1+(y-m)^2/4.$$

In all cases, $\lim_{|y-m| \to \infty} \text{var}[\theta|y]$ diverges.

Now these sort of results indicate a rather conservative "robust" analysis in which the posterior density flattens out between m and y as $|y-m|$ increases with modes following m and y. The ambiguity is ever present. In order to avoid it we need a dominant prior specification as we now discuss.

2.2.3 Strong prior information.

Given a robust likelihood p, we need to satisfy (v) of equation (2.2.10) in order to achieve full outlier rejection as defined by the convergence of posterior to prior distributions. The robust likelihoods descend no faster than $e^{-k|u|}$ as $|u| \to \infty$, (for details see Appendix A2.2.(b)), and so any prior for which

$$\int_{-\infty}^{\infty} \pi(\theta).e^{-k|\theta|}\, d\theta < \infty$$

will suffice; in particular $\pi(\theta) \sim e^{-h|\theta|^{1+\epsilon}}$, $\epsilon > 0$, would be appropriate. Further, this will lead to convergence of posterior moments of all orders to prior moments. Clearly this specification implies an asymmetry of the treatment of π and p; the prior is non-robust and must be so in order to avoid the ambiguous analysis discussed in § 2.2.2.

Now the exponential power prior $e^{-h|\theta|^{1+\epsilon}}$ is not very tractible for general $\epsilon > 0$. The case $\epsilon = 1$, the normal density is, however,

and we discuss this now.

Normal Prior.

How can we justify a normal prior? The reason of tractibility is certainly important, and later results show that a normal prior provides just enough structure to enable closed form expressions for posterior moments to be derived for a wide range of symmetric likelihoods. We have also the following considerations which suggest that a normal prior will often be a not unreasonable assumption.

(i) We may actually have such prior beliefs, possibly from some previous analysis. If, for example, y is one of a set of observations y_1, y_2,... of which y_1,..., y_n are "good" i.e. can be assigned a normal likelihood, then a pragmatic approach to modelling the data might be to suppose an inhomogenous sample reserving the robust likelihood for dubious observations. O'Hagan (1979) makes some remarks along these lines.

If our prior before making any observations is normal, then the posterior given y_1,..., y_n is too and forms a normal prior for θ before observing the data y.

(ii) Arguing along similar lines to (i), but more formally, let the data generating mechanism be a mixture of the form

$$f(y) = (1-\epsilon)\ \phi(y) + \epsilon p(y)$$

Where ϕ is the standard normal p.d.f.

Then, as in Box (1980),

$$p(y_1,\ldots, y_n|\theta) = w_0 p_0(\theta) + w_1 \sum_{j=1}^{n} p_{1j}(\theta) + \ldots$$

where

$$p_0(\theta) = \prod_{k=1}^{n} \phi(y_k-\theta)$$

is the likelihood of $(y_1,\ldots, y_n|\theta)$ when all are from the normal component of f ,

$$p_{1j}(\theta) = \Big\{ \prod_{\substack{k=1 \\ k\neq j}}^{n} \phi(y_k-\theta) \Big\}.p(y_j-\theta)$$

is the likelihood of $(y_1,\ldots, y_n|\theta)$ when all but y_j are from the normal component, and

$$w_0 = (1-\epsilon)^n,\ w_1 = \epsilon(1-\epsilon)^{n-1},..\ \text{etc.}$$

Again if our prior is normal, then $p(\theta|y_1,\ldots, y_n)$ is a sum of terms the first of which is normal and, when $y_1,\ldots, y_n$ are "good" observations, this term will dominate the others as discussed by Box and Tiao (1963) and Box (1980). This then gives an approximately normal prior before observing y.

(iii) Asymptotic considerations. For large n, under rather weak conditions on the prior and likelihood, $p(\theta|y_1\ldots y_n)$ approaches normality. See Heyde and Johnstone (1979) for details of such conditions.

So for the rest of this section we take a normal prior for θ when mean m and variance c^2,

$$\pi(\theta) = c^{-1}\phi\big(c^{-1}(\theta-m)\big),\ c > 0, \text{ and}$$

proceed with an examination of various characteristics of $p(\theta|y)$.

The posterior modes.

Let θ^* be a posterior mode. Then the modal equation is

$$c^{-2}.(\theta^*-m) = g(y-\theta^*)$$

or

$$\theta^* = m + c^2.g(y-\theta^*)$$

indicating the robustifying nature of a redescending score function. This equation can be solved iteratively to find θ^*. We note that a simple plot of the posterior score

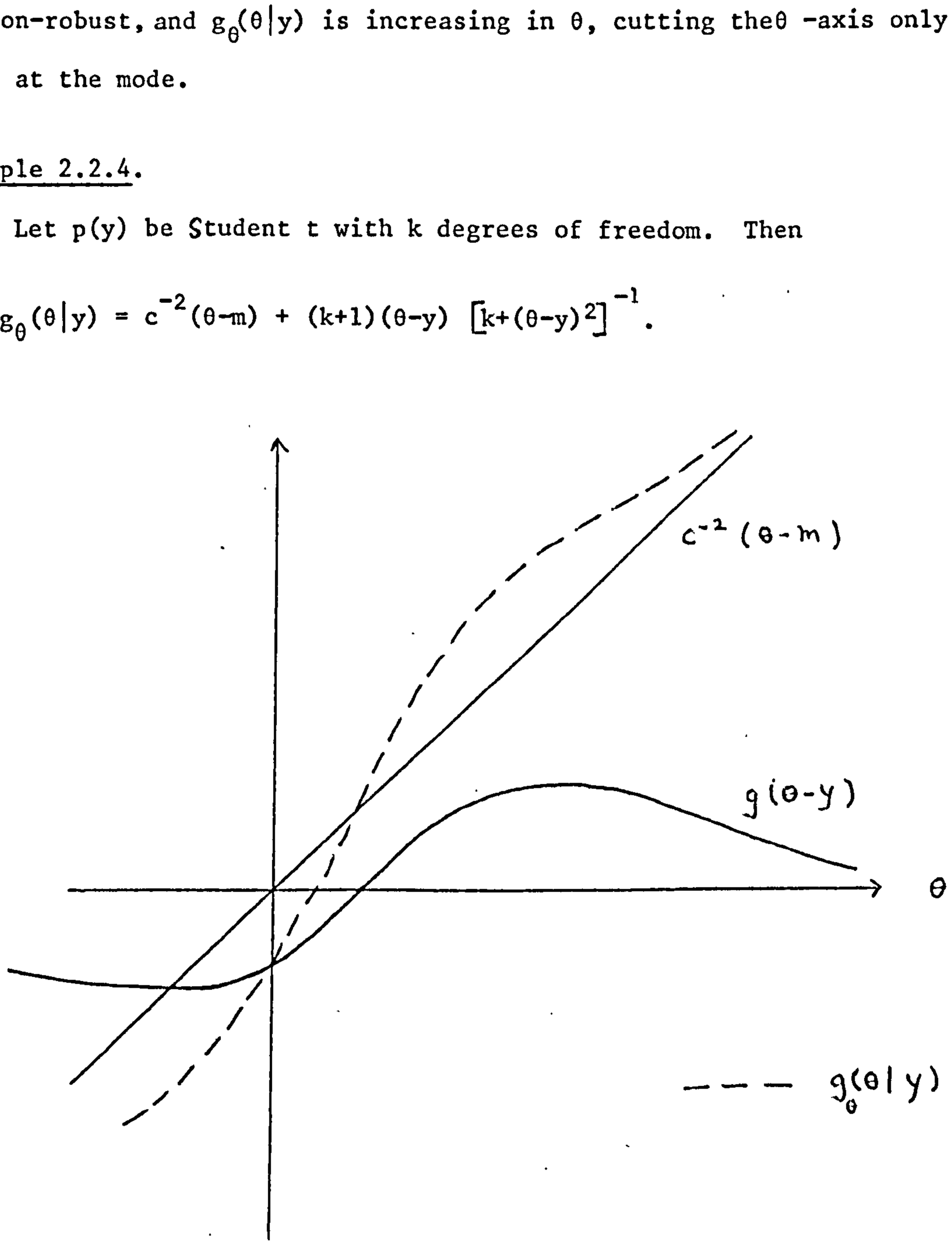
$$g_\theta(\theta|y) = c^{-2}(\theta-m) + g(\theta-y)$$

will provide a guide to the position of modes and indicate whether $p(\theta|y)$ is bimodal. If $g(y)$ is non-decreasing for $y > 0$, then $p(y)$ is non-robust, and $g_\theta(\theta|y)$ is increasing in θ, cutting theθ -axis only once at the mode.

Example 2.2.4.

Let $p(y)$ be Student t with k degrees of freedom. Then

$$g_\theta(\theta|y) = c^{-2}(\theta-m) + (k+1)(\theta-y)\left[k+(\theta-y)^2\right]^{-1}.$$



Bimodality only occurs for small k when c^{-2} is very small and $|y-m|$ large.

Example 2.2.5.

Define the Huber k density (Huber, 1964) by

$$p(y) \alpha \begin{cases} \phi(y), & |y| \leq k; \\ \exp -k|y|, & \text{otherwise.} \end{cases}$$

If p(y) has this form then $g_\theta(\theta|y)$ is piecewise linear, given by

$$\begin{cases} c^{-2}(\theta-m) + \theta-y, & |y-\theta| \leq k; \\ c^{-2}(\theta-m) + k, & \theta > y+k; \\ c^{-2}(\theta-m) - k, & \theta < y-k. \end{cases}$$

In this case the posterior mode can be written in closed form as follows:-

Let $m_0 = (1+c^2)^{-1}(c^2y+m)$,

$m_1 = m-c^2k$,

$m_2 = m+c^2k$.

Then $$\theta^* = \begin{cases} m_0, & |y-m| < k(1+c^2); \\ m_1, & y-m < -k(1+c^2); \\ m_2, & y-m > k(1+c^2). \end{cases}$$

Alternatively, $\theta^* = m + c^2 g^*(y-m)$

where

$$g^*(u) = \begin{cases} u.(1+c^2)^{-1}, & |u| < k(1+c^2); \\ k \operatorname{sgn}(u), & \text{otherwise,} \end{cases}$$

is a slightly modified version of the score function g(u).

Finally note that piecewise linearity of $g_\phi(\theta|y)$ means piecewise normality of $p(\theta|y)$.

For $\theta \in I_j$,

$$p(\theta|y) \alpha c_j^{-1}\phi\left(c_j^{-1}(\theta-m_j)\right), \quad j=0,1,2,$$

where $I_0 = \{\theta| \ |\theta-y| \leq k\}$,

$I_1 = \{\theta| \ \theta < y-k\}$,

$I_2 = \{\theta| \ \theta > y+k\}$,

and $c_1^2 = c_2^2 = c^2$, $c_0^{-2} = 1+c^{-2}$.

Posterior mean and variance.

It is possible to obtain closed form expressions for the mean and variance of $p(\theta|y)$ within this location parameter framework with a normal prior. Masreliez (1975) proves the following theorem in a more general setting. Clearly we can derive similar results for higher moments if required to investigate skewness and kurtosis of the posterior distribution, with moments of order k requiring the existence of the kth derivative of the log likelihood.

Theorem (Masreliez).

Let $g_y(y) = -\frac{\partial}{\partial y} \ln p_y(y)$ and $G_y(y) = \frac{\partial}{\partial y} g_y(y)$

where

$$p_y(y) = \int_{-\infty}^{\infty} p(y-\theta)\pi(\theta)\, d\theta \quad .$$

Then (i) $E[\theta|y] = m + c^2 g_y(y)$,

(ii) $var[\theta|y] = c^2 - c^4 G_y(y)$.

Proof (Masreliez, 1975).

By definition $E[\theta|y] = \int_{-\infty}^{\infty} \theta\pi(\theta)p(y-\theta)d\theta . p_y(y)^{-1}$

So $p_y(y)\left[E[\theta|y]-m\right] = \int_{-\infty}^{\infty} (\theta-m)\pi(\theta)p(y-\theta)d\theta.$

Now since $\pi(\theta) = c^{-1}\phi\{c^{-1}(\theta-m)\}$, then

$$(\theta-m)\pi(\theta) = -c^2 \frac{\partial\pi(\theta)}{\partial\theta} \qquad (2.2.1)$$

so

$$p_y(y)\left[E[\theta|y]-m\right] = -c^2 \int_{-\infty}^{\infty} \frac{\partial\pi(\theta)}{\partial\theta} . p(y-\theta)d\theta$$

$= -c^2 \left\{ \left[p(y-\theta)\pi(\theta)\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{\partial p(y-\theta)}{\partial\theta} \pi(\theta)d\theta \right.$, on integrating by parts,

$= -c^2 \int_{\infty}^{\infty} \pi(\theta) \frac{\partial}{\partial y} p(y-\theta)d\theta$

$= -c^2 . \frac{\partial}{\partial y} p_y(y)$, on interchanging the orders of integration and

differentation. So (i) follows. Similarly, using (2.2.1),

$$p_y(y).E\left[(\theta-m)^2|y\right] = -\,c^2 \int_{-\infty}^{\infty} \frac{\partial\pi(\theta)}{\partial\theta}\,.(\theta-m)p(y-\theta)d\theta$$

which, on integrating by parts, gives

$$= c^2 \int_{-\infty}^{\infty} \pi(\theta).\{p(y-\theta) + (\theta-m)\,\frac{\partial p}{\partial\theta}(y-\theta).\}d\theta$$

$$= c^2 p_y(y) + c^4 \int_{-\infty}^{\infty} \frac{\partial\pi(\theta)}{\partial\theta}.\,\frac{\partial p}{\partial y}(y-\theta)d\theta$$

$$= c^2 p_y(y) - c^4\, H(y),$$

where, by integration by parts,

$$H(y) = \{\, -\left[\pi(\theta)\,\frac{\partial p}{\partial y}(y-\theta)\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \pi(\theta).\,\frac{\partial^2 p}{\partial y\partial\theta}(y-\theta)d\theta\,\}$$

$$= -\,\frac{\partial^2}{\partial y^2} p_y(y),$$

again by interchanging the orders of integration and differentiation.

Thus

$$\mathrm{var}\left[\theta|y\right] = E\left[\theta-m)^2|y\right] - (E\left[\theta|y\right]-m)^2$$

$$= c^2 - c^4\,\left[H(y).p_y(y)^{-1} + g_y^2(y)\right]\,,$$

and (ii) follows by noting that

$$G_y(y) = p_y(y)^{-1}\,H(y) + g_y^2(y)\quad .$$

So the marginal score, which by Lemma 2.1.1 is the posterior expected value of the likelihood score, is used explicitly in determining the posterior mean. The marginal density p_y is just the convolution of the heavy-tailed likelihood p with a normal prior π and we might intuitively expect the tail behaviour to mirror that of p. This is indeed so and we note the following properties of the marginal density, score and information functions:-

(i) p_y is unimodal symmetric about m. This follows directly from the definition as a convolution of two unimodal symmetric densities. Thus $g_y(y)$ is skew symmetric about m and $G_y(y)$ is symmetric.

(ii) For robust likelihoods such that g(y) is bounded and redescending then

a) $g_y(y)$ is bounded. This follows since

$$g_y(y) = E[g(y-\theta)|y]$$

So $|g(y-\theta)| \leq m$ implies $|g_y(y)| \leq m$.

b) $g_y(y)$ redescends. This follows from the work of O'Hagan (1979) since we know that

$$E[\theta|y] - m \to 0 \text{ as } |y| \to \infty.$$

As an example consider the case of a stable likelihood.

Assuming without loss of generality that m=0, then

$$p_y(y) = (2\pi)^{-1} \int_{-\infty}^{\infty} \exp\{ - |t|^a - t^2c^2/2 - ity \} dt$$

$$= \pi^{-1} \int_0^{\infty} \exp\{ - t^a - t^2c^2/2 \}. \cos(ty)\, dt$$

$$= (\pi y)^{-1} \int_0^{\infty} \exp\{ - (t/y)^a - t^2c^2/2y^2 \} \cos(t)\, dt.$$

Now $\exp(-t^2c^2/2y^2) = 1 + (\phi t^2c^2/2y^2)$, where $|\phi| \leq 1$. So, in the notation of Lemma A2.2.1 of Appendix A2.2,

$$p_y(y) = (\pi y)^{-1}.\text{Re } I_0(y) + c^2(2y^2)^{-1}.O(\text{Re } I_2(y)).$$

We can follow the proof of Lemma A2.21 to show that $g_y(y)$ behaves like y^{-1} as $y \to \infty$.

(iii) The moment structure of p_y mirrors that of p since, for k>0,

$$E[y^k] = E\left[E[y^k|\theta]\right]. < \infty \text{ whenever } E[y^k|\theta] < \infty.$$

Moreover marginal moments diverge when likelihood moments do so.

(iv) $G_y(y)$ may be negative. This occurs when $g_y(y)$ redescends (for y>0) and leads to the posterior variance exceeding the prior variance. This type of behaviour is noted by O'Hagan (1981) in a similar context and is quite natural. In such cases, $G_y(y)$ is positive for "small" values of the residual y-m and therefore $\text{var}[\theta|y] < c^2$. As $|y-m|$ increases, G_y goes negative and so $\text{var}[\theta|y] > c^2$ reflecting the uncertainty about y. (Is it a good observation or not?) For larger $|y-m|$, G_y tends to zero and y is classified as an outlier, being ultimately ignored.

Here we can see how use of an outlier-prone distribution inverts the relationship between the influence and the extremeness of an observation as we mentioned in the introduction (§2.1.1).

Consider for example the usual normal (non-robust) analysis. The marginal score is just the linear function

$$g_y(y) = (1+c^2)^{-1}(y-m).$$

So as $|y-m|$ increases i.e. as y becomes aberrant, (given that the prior is unquestioned), then $E[\theta|y]$ grows with y and thus the influence of y grows.

For, say, a Student t likelihood, considering the redescending shape of $g_y(y)$ we see that the influence of y on $E[\theta|y]$ reaches a peak and then begins to decay as y becomes more and more aberrant.

(v) Turning now to non-robust likelihoods, p_y is non-robust. In particular, if g(y) is increasing then p(y) is log concave and therefore strongly unimodal, as discussed by Barndorff-Neilsen (1978, Chapter 6). Furthermore, Corollary 6.1 of this reference states that the convolution of two strongly unimidal (continuous) distributions is itself strongly unimodal hence log concave. Since the normal distribution is strongly unimodal then our marginal distribution is and so $g_y(y)$ increases.

This accords with the result of O'Hagan (1979), that strongly unimodal distributions (that are continuous and symmetric) are outlier resistant and therefore $p(\theta|y)$ cannot "reject" the outlier y. In our framework the posterior mean is increasing for $y>m$ and tends either to infinity with y (as in the exponential power case, with $1<\beta\leq 2$, Appendix A2.2 (ii)), or to a constant value (as with a logistic or double-exponential likelihood).

We illustrate the similarity between likelihood and marginal scores with three examples. In each case, $m=0, c^2=1$. The likelihoods are

(a) Contaminated normal CN[0.1;1,5] ;

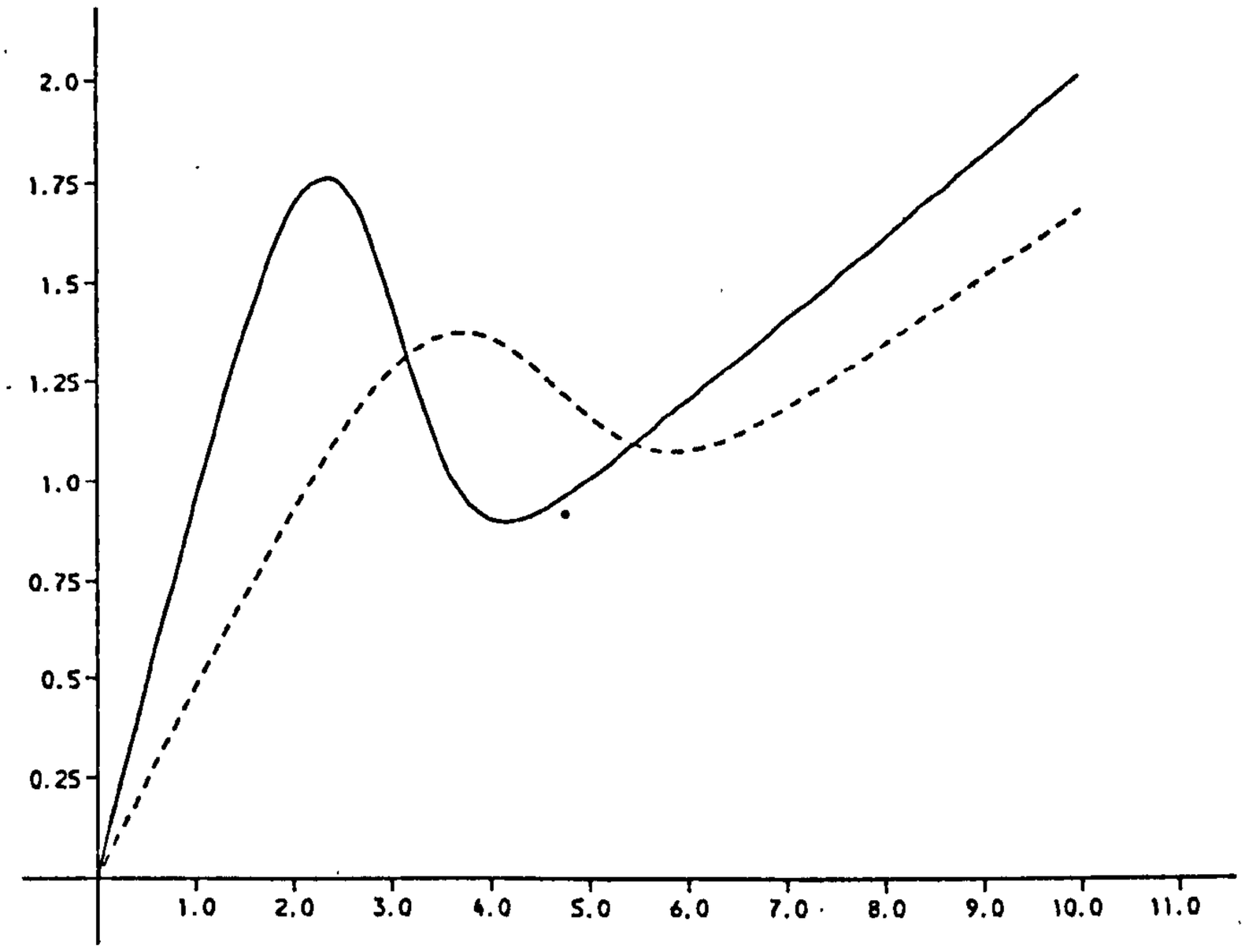
(b) Double exponential.

(c) Student t-8;

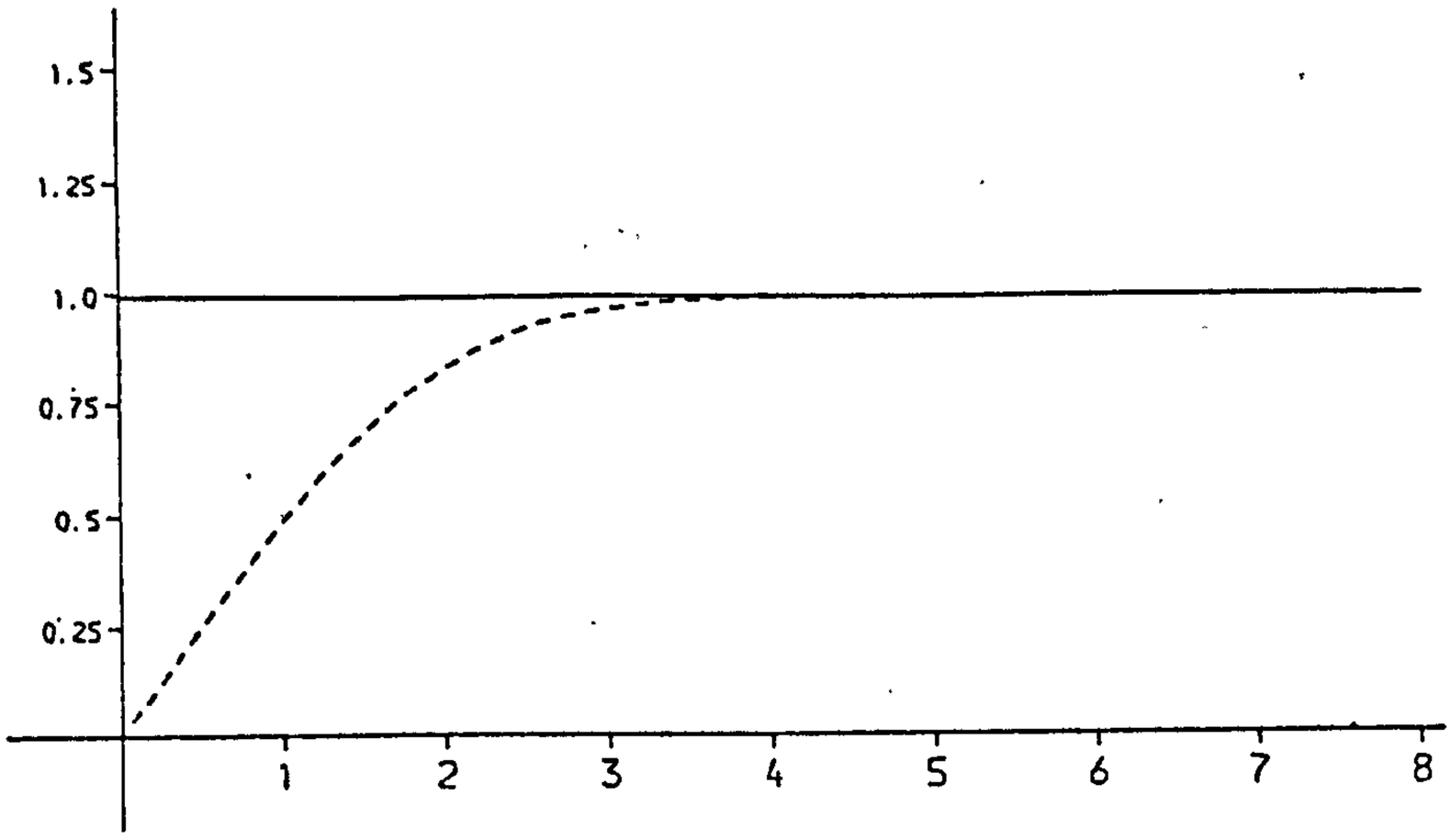
The full line is the likelihood score, the dashed line the marginal.

We then turn to the related problem of unknown scale parameters and explore ideas similar to those above for the location case.

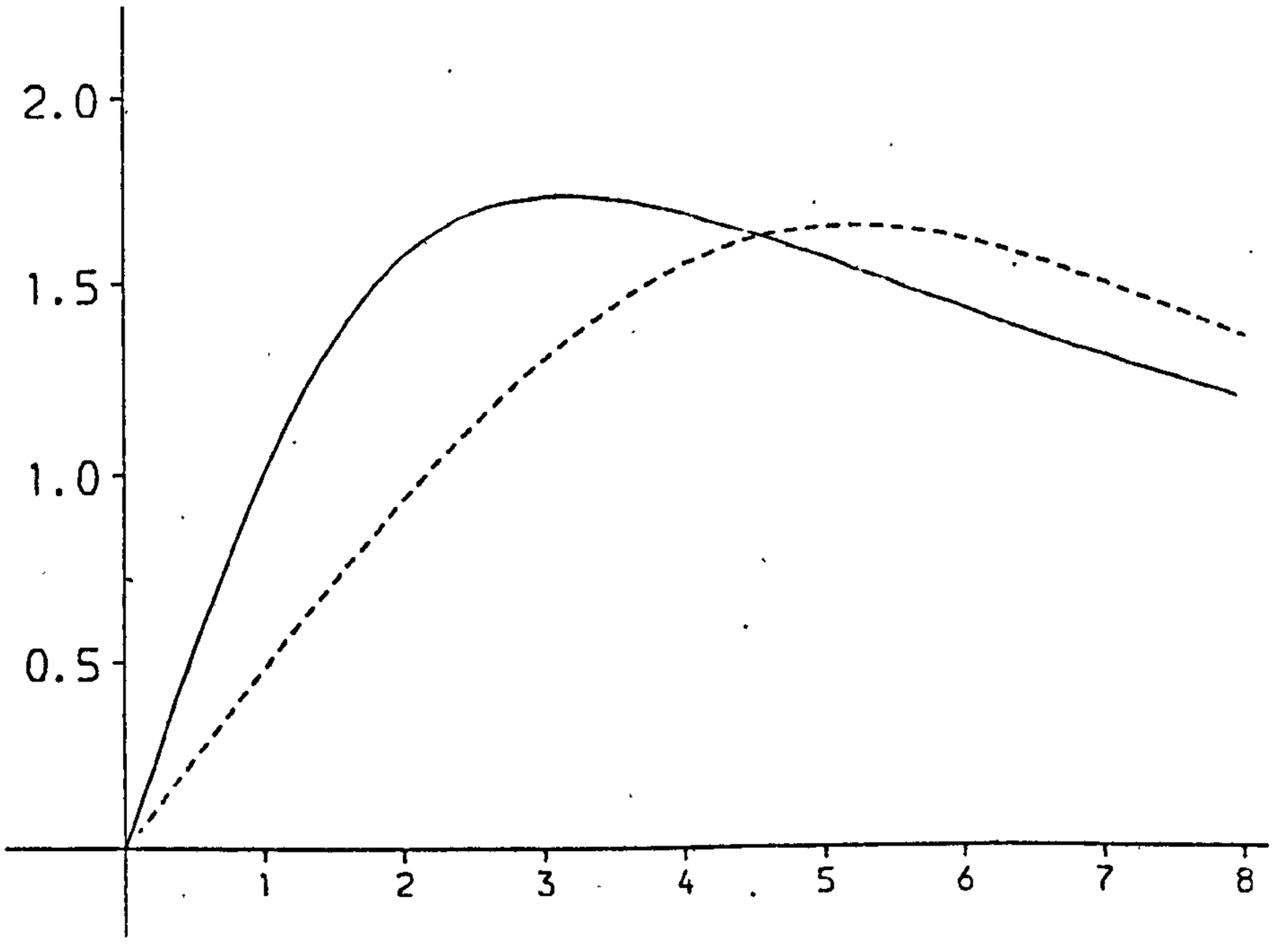
(a)


2.0
1.75
1.5
1.25
1.0
0.75
0.5
0.25
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0

(b)


1.5
1.25
1.0
0.75
0.5
0.25
1
2
3
4
5
6
7
8

(c)


2.0
1.5
1.0
0.5
1
2
3
4
5
6
7
8

2.3 Scale parameter case.

2.3.1 Introduction.

Consider the case of known location, $\theta=0$, and unknown scale parameter $\sigma>0$. The likelihood is now

$$p(y|\sigma) = \sigma^{-1}p(\sigma^{-1}y).$$

and the score functions of p are related by

$$g_\sigma(y|\sigma) = \sigma^{-1}\left[1-yg_y(y|\sigma)\right] , \qquad (2.3.1)$$

with $g_y(y|\sigma) = \sigma^{-1}g(\sigma^{-1}y)$, g(u) being the likelihood score function, $-\frac{\partial}{\partial u}\ln p(u)$, $u \in \mathbb{R}$.

It may seem reasonable to suppose that outlier-rejection could be achieved by adopting an outlier-prone density p. That this is not so can be seen simply by examining the posterior score function with respect to σ, given a prior $\pi(\sigma)$,

$$-\frac{\partial}{\partial\sigma}\ln p(\sigma|y) = -\frac{\partial}{\partial\sigma}\ln \pi(\sigma) + \sigma^{-1}\left[1-\sigma^{-1}yg(\sigma^{-1}y)\right] \qquad (2.3.2)$$

In order that the posterior score converge to the prior as $|y| \to \infty$, for all σ, we require that

$$ug(u) \to 1 \quad \text{as } |u| \to \infty.$$

None of our robust likelihoods satisfy this condition; it demands tails which are heavier than those of any density discussed so far. In general the densities of §2.2 are such that either ug(u) diverges, or, for robust densities, ug(u) converges to some constant not equal to unity. For example, with a Student t-k likelihood,

$$ug(u) \to 1+k \quad \text{as } |u| \to \infty,$$

and thus the observation has a limited effect as it becomes extreme.

Now we follow the same line of argument as that of §2.2.3 in choosing a prior representing strong information in some sense about the scale parameter.

Define $\lambda = \sigma^{-2}$. The usual normal theory analysis requires a Gamma prior distribution for λ for the conjugate analysis, and we adopt such a prior, noting that the discussion of a normal prior for location is valid here too, in that a gamma prior for λ is reasonable if, for example,

(i) we have a set of "good" observations to which we assign a normal likelihood leading to a gamma form in λ (and so a gamma prior before making these observations leads to a gamme posterior);

(ii) More formally, if we model the likelihood as

$$(1-\epsilon)^{-1}\sigma\phi(\sigma^{-1}y) + \epsilon\sigma^{-1}p(\sigma^{-1}y),$$

then the first term in the expansion of the likelihood of a set of good observations will be a gamma form in λ and will tend to dominate the other terms.

Within this framework, the following result is applicable,

Theorem 2.3.1

Let the prior be $\pi(\lambda) = G\left[\frac{\alpha}{2}, \frac{\beta}{2}\right]$, $\alpha>0, \beta>0$

given by

$$\pi(\lambda) \propto \lambda^{\frac{\alpha-1}{2}} e^{-\frac{\lambda\beta}{2}}, \quad \lambda>0.$$

Define the marginal density $p_y(y) = \int_0^\infty \lambda^{\frac{1}{2}} p(\lambda^{\frac{1}{2}}y).\pi(\lambda)d\lambda$ and the score and information functions

$$g_y(y) = -\frac{\partial}{\partial y} \ln p_y(y), \quad G_y(y) = \frac{\partial}{\partial y} g_y(y).$$

Then

$$\text{(i)}\quad E[\lambda|y] = \beta^{-1}.\{(\alpha+1) - yg_y(y)\},$$

$$\text{(ii)}\quad \text{var}[\lambda|y] = \beta^{-2}.\{2(\alpha+1) - 3yg_y(y) - y^2G_y(y)\}.$$

Proof:

By definition $E[\lambda|y] = \int_0^\infty \lambda^{3/2}p(\lambda^{\frac{1}{2}}y)\pi(\lambda)d\lambda.p_y(y)^{-1}$

or $\left[E[\lambda|y] - \alpha/\beta\right]p_y(y) = \int_0^\infty (\lambda-\alpha/\beta).\pi(\lambda).\lambda^{\frac{1}{2}}p(\lambda^{\frac{1}{2}}y)d\lambda.$

Now, by the special form of π,

$$(\lambda-\alpha/\beta)\pi(\lambda) = -2\beta^{-1}.\frac{\partial}{\partial\lambda}[\lambda\pi(\lambda)]. \qquad (2.3.3)$$

hence

$$\left[E[\lambda|y]-\alpha/\beta\right].p_y(y) = \int_0^\infty -2\beta^{-1}\frac{\partial}{\partial\lambda}[\lambda\pi(\lambda)]\lambda^{\frac{1}{2}} \qquad p(\lambda^{\frac{1}{2}}y)d\lambda$$

which, on integrating by parts gives

$$\{\left[-2\beta^{-1}\lambda\pi(\lambda)\lambda^{\frac{1}{2}}p(\lambda^{\frac{1}{2}}y)\right]_0^\infty + 2\beta^{-1}\int_0^\infty\lambda\pi(\lambda)\frac{\partial}{\partial\lambda}[\lambda^{\frac{1}{2}}p(\lambda^{\frac{1}{2}}y)]d\lambda\}$$

$$= \beta^{-1}\int_0^\infty \lambda\pi(\lambda)\{\lambda^{-\frac{1}{2}}p(\lambda^{\frac{1}{2}}y) + \lambda^{-\frac{1}{2}}y\frac{\partial}{\partial y}p(\lambda^{\frac{1}{2}}y)\}d\lambda$$

where we have used the identity

$$\frac{\partial}{\partial\lambda}p(\lambda^{\frac{1}{2}}y) = \frac{y}{2\lambda}p(y\lambda^{\frac{1}{2}}).$$

Therefore

$$p_y(y)\left[E[\lambda|y]-\alpha/\beta\right] = \beta^{-1}\{\int_0^\infty\lambda^{\frac{1}{2}}p(\lambda^{\frac{1}{2}}y)\pi(\lambda)d\lambda + y\int_0^\infty\pi(\lambda)\frac{\partial}{\partial y}[\lambda^{\frac{1}{2}}p(\lambda^{\frac{1}{2}}y)]d\lambda\}$$

$$= \beta^{-1}\{p_y(y) + y\frac{\partial p}{\partial y}y(y)\}$$

on interchanging the orders of integration and differentiation in the second term. The result (i) follows immediately.

For (ii) note that, using (2.3.3),

$$E[\lambda(\lambda-\alpha/\beta)|y]\ p_y(y) = -2\beta^{-1}\int_0^\infty \frac{\partial}{\partial\lambda}[\lambda\pi(\lambda)].\lambda^{3/2}\,p(\lambda^{\frac{1}{2}}y)d\lambda.$$

Integration by parts gives

$$-2\beta^{-1}\{[\lambda\pi(\lambda)\lambda^{3/2}\,p(\lambda^{\frac{1}{2}}y)]_0^\infty - \int_0^\infty \lambda\pi(\lambda)\,\frac{\partial}{\partial\lambda}\{\lambda^{3/2}\,p(\lambda^{\frac{1}{2}}y)\}d\lambda$$

$$= \beta^{-1}\int_0^\infty \lambda\pi(\lambda).[3\lambda^{\frac{1}{2}}p(\lambda^{\frac{1}{2}}y) + \lambda^{\frac{1}{2}}y\ \frac{\partial}{\partial y}\,p(\lambda^{\frac{1}{2}}y)]d\lambda$$

$$= 3\beta^{-1}p_y(y)E[\lambda|y] + y\beta^{-1}\frac{\partial}{\partial y}\left[p_y(y).E[\lambda|y]\right]$$

on interchanging the orders of integration and differentiation in the second term.

So, using (i)

$$E[\lambda(\lambda-\alpha/\beta)|y]p_y(y) = 3\beta^{-2}p_y(y)[a+1-yg_y(y)]$$

$$+ y\beta^{-1}\{p_y(y).\frac{\partial}{\partial y}E[\lambda|y] + \frac{\partial p_y}{\partial y}(y)E[\lambda|y]\}.$$

The second term here is

$$p_y(y)y\beta^{-2}.\{-yG_y(y) - (\alpha+1)g_y(y) + yg_y(y)^2\}\ ,$$

and therefore

$$E[\lambda(\lambda-\alpha/\beta)|y] = \beta^{-2}\{3(\alpha+1) - g_y(y).[3y+(\alpha+2)y] + y^2g_y^2(y) - y^2G_y(y)\}.$$

Using the identity

$$\mathrm{var}[\lambda|y] = E[\lambda(\lambda-\alpha/\beta)|y] + \alpha/\beta E[\lambda|y] - E[\lambda|y]^2$$

the result follows. □

Note that the prior mean is α/β and so $E[\lambda|y]$ can be written as

$$E[\lambda|y] = E[\lambda] + \beta^{-1}.[1-yg_y(y)].$$

Similarly $\mathrm{var}[\lambda|y] = \mathrm{var}[\lambda] + \beta^{-2}[2-3yg_y(y)-y^2G_y(y)]$.

As with the location problem, the marginal density defines the posterior mean and variance via the derivatives of the log density. (Again, higher moments can be derived similarly). We note the following immediate properties of p_y.

(i) $p_y(y)$ is unimodal symmetric at zero.

(ii) $p_y(y)$ is generally rather heavier-tailed than $p(y)$ in the sense that it lacks high order moments.

For $k>0$, $E[y^k] = E\left[E[y^k]\right]$

Now $E[y^k|\lambda] = \lambda^{-k/2} \int_{-\infty}^{\infty} u^k p(u)du$, thus, if p has moments of order k, i.e. $E[u^k] = \int_{-\infty}^{\infty} u^k p(u)du < \infty$, then

$$E[y^k] = E[u^k].E[\lambda^{-k/2}]$$

$$< \infty \text{ if and only if } E[\lambda^{-k/2}] < \infty.$$

For $\pi(\lambda) = G\left[\frac{\alpha}{2},\frac{\beta}{2}\right]$, $E[\lambda^{-k/2}] < \infty$ only for $k < \alpha$, and so $E[y^k]$ only exists for $k < \alpha$, when $E[u^k] < \infty$. This heavy-tailedness of p^y is reflected in the score function. Since $E[\lambda|y] \gtrsim 0$, we have, by the Theorem

$$(\alpha+1) - yg_y(y) \gtrsim 0 \quad \text{, for all } y,$$

hence, for $|y| \neq 0$, $|g_y(y)| \lesssim |y|^{-1}(\alpha+1)$.

Note that with a normal likelihood,

$$g_y(y) = (\alpha+1)y/(\beta+y^2)$$

and therefore

$$\lim_{|y|\to\infty} E[\lambda|y] = \lim_{|y|\to\infty} \mathrm{var}[\lambda|y] = 0$$

indicating the lack of robustness of a normal analysis. For robust likelihoods (and heavy-tailed non-robust ones), $yg_y(y)$, typically converges to a constant not equal to $\alpha+1$, i.e. $E[\lambda|y]$ tends to a non-zero constant as $|y|\to\infty$.

Similarly, using this bound on $yg_y(y)$ and the fact that $\text{var}[\lambda|y] > 0$, we can easily bound $G_y(y)$ by

$$|G_y(y)| < y^{-2}.3(\alpha+1) \quad , \; y \neq 0,$$

and hence show that $\text{var}[\lambda|y]$ is bounded above by $8(\alpha+1)\beta^{-2}$. Again, for non-normal likelihoods, $\text{var}[\lambda|y]$ has some finite non-zero limiting value as $|y|\to\infty$.

2.3.2. Location/Scale.

Now the likelihood is

$$p(y|\theta,\sigma) = \sigma^{-1}p(\sigma^{-1}[y-\theta]) \quad , \qquad \sigma>0, \; -\infty<\theta<\infty.$$

In order that we can apply the ideas of the separate location and scale problems we need to adopt a special form of joint prior distribution as follows:

Take a joint prior for (θ,σ) such that the conditional prior $\pi(\theta|\sigma)$ is scaled by σ, (centred at zero for simplicity),

i.e. $\pi(\theta|\sigma) = \sigma^{-1}\pi(\sigma^{-1}\theta)$,

and a marginal prior for σ, $\rho(\sigma)$.

Then we have

(i) The conditional posterior $\pi(\theta|y,\sigma) \propto p(\sigma^{-1}(y-\theta))\pi(\sigma^{-1}\theta)$.

(ii) The marginal posterior $\rho(\sigma|y)$ is obtained as follows

Set $z = \sigma^{-1}y$ and $\phi = \sigma^{-1}\theta$. Then

$$p(y|\sigma) = \int_{-\infty}^{\infty} \sigma^{-1}p[\sigma^{-1}(y-\theta)]\sigma^{-1}\pi(\sigma^{-1}\theta)d\theta$$

$$= \sigma^{-1} \int_{-\infty}^{\infty} p(z-\phi).\pi(\phi)d\phi$$

is unimodal symmetric at zero and scaled by σ putting us into the framework of §2.3.1.

This scheme is used in the normal theory model with the joint prior being of the normal/gamma form i.e. π a normal density and ρ that of the square root of an inverse gamma random variable. Following the ideas of this Chapter, we investigate the implications of this prior with a non-normal likelihood.

Define $\lambda = \sigma^{-2}$.

Now $p(\theta,\lambda)$ is such that $p(\theta,\lambda) = \pi(\theta\ \lambda)\rho(\lambda)$, where

$$\pi(\theta|\lambda) = N\left[m, c^2\lambda^{-1}\right] ,$$

and $$\rho(\lambda) = G\left[\alpha/2, \beta/2\right] .$$

Corollary 2.3.1.

Define $p(y|\sigma) = \sigma^{-1}p_1\left[\sigma^{-1}(y-m)\right] = \int_{-\infty}^{\infty} \sigma^{-2}p\left[\sigma^{-1}(y-\theta)\right].\pi(\theta|\lambda=\sigma^2)d\theta$, and the score and information functions

$$\sigma^{-1}.g_1\left(\sigma^{-1}(y-m)\right) = -\frac{\partial}{\partial y} \ln\ p(y|\sigma) \quad ,$$

$$\sigma^{-2}G_1\left(\sigma^{-1}(y-m)\right) = \sigma^{-1}\frac{\partial}{\partial y}\ g_1\left(\sigma^{-1}(y-m)\right).$$

Then

$$E\left[\theta|y,\sigma\right] = m + c^2\sigma.g_1\left[(y-m)\sigma^{-1}\right],$$

and

$$var\left[\theta|y,\sigma\right] = \sigma^2\left[.c^2-c^4G_1\left[(y-m)\sigma^{-1}\right]\right].$$

Proof. Apply Masreliez's Theorem, noting that from (ii) above, $p(y|\sigma)$ is unimodal and symmetric about m.

Corollary 2.3.2.

Define further $p(y) = p_2(y) = \int_0^\infty \sigma^{-1} p_1[\sigma^{-1}(y-m)] . \rho(\sigma) d\sigma$

and

$$g_2(y) = -\frac{\partial}{\partial y} \ln p_2(y) \ , \quad G_2(y) = \frac{\partial g_2}{\partial y}(y).$$

Then

$$E[\lambda|y] = b^{-1}\{(a+1) - y g_2(y)\},$$

and

$$var[\lambda|y] = b^{-2}\{2(a+1) - 3y g_2(y) - y^2 G_2(y)\}.$$

Proof. Apply Theorem 2.3.1 noting that p_2 is unimodal symmetric at m.

Now we have a framework on which to build a scheme for recursive estimation of location/scale parameters with non-normal likelihoods. In Chapter 3 we use these results in investigating models more general than the simple scalar parameter problems of this Chapter; in particular linear time-series models.

Appendix 2.

A2.1

Lemma 2.1.1.

Let $\theta \in \mathbb{R}$ have prior $\pi(\theta)>0$, $\theta \in \mathbb{R}$, and let $y \in \mathbb{R}$ be related to θ via a likelihood $p(y|\theta)>0$ for all y. Assume that p is twice differentiable with respect to y. If $g(y|\theta)$, $G(y|\theta)$ are the likelihood score and information functions and $g(y)$, $G(y)$ those of the marginal density of y, then

(i) $g(y) = E[g(y|\theta)|y]$

(ii) $g^2(y)-G(y) = E[g(y|\theta)^2-G(y|\theta)|y]$.

Proof:

Using the property $E\left[\frac{\partial}{\partial y} \ln p(\theta|y)|y\right] = 0$

(Cox & Hinkley, (1974), p110), and the relation

$$\frac{\partial}{\partial y} \ln p(\theta|y) = \frac{\partial}{\partial y} \ln p(y|\theta) - \frac{\partial}{\partial y} \ln p(y) \qquad (1)$$

we have (i) immediately.

For (ii) we use the identity

$$E\left[\left(\frac{\partial}{\partial y} \ln p(\theta|y)\right)^2|y\right] = E\left[-\frac{\partial^2}{\partial y^2} \ln p(\theta|y)|y\right]$$

and on expanding the square using (1) we have

$$E[g^2(y|\theta)+g^2(y)-2g(y)g(y|\theta)|y] = E[g^2(y|\theta)|y] - g^2(y)$$

from (i)

$$= E[G(y|\theta) - G(y)|y]$$

and (ii) follows. □

A2.2 Survey of heavy tailed, unimodal, symmetric densities.

(i) Contaminated normal.

Widely used in robustness studies, and in particular, in Bayesian approaches to robust estimation, (See, for example, Box and Tiao (1968) and Box (1980)), the contaminated normal density

$$p(y) = (1-\epsilon)\phi(y) + \epsilon\sigma^{-1}\phi(\sigma^{-1}y),$$

where $0<\epsilon<1$ and $\sigma>1$, is not outlier-prone. The score function illustrates the treatment that an observation receives from such a likelihood when used in the location problem. There are three distinct regions:

a) "Small" values of y are essentially assigned to the $\phi(y)$ component of p(y);

b) "Medium" values are problematic; is y an outlier or not?

c) "Large" values of y are assigned to the $\sigma^{-1}\phi(\sigma^{-1}y)$ component of p.

Use of this density is not problematic; we can essentially consider the two components separately in the usual way, as in Harrison and Stevens (1976) for example.

(ii) Exponential-power

Box and Tiao (1973) pioneered the use of their family in robustness studies; the density is written as

$$p(y) = c(\beta).\exp\{-\tfrac{1}{2}|y|^{\beta}\}, \qquad 0<\beta.$$

For $0<\beta\leq 2$, we have a unimodal symmetric and heavy-tailed density, with the Laplace, or double-exponential at $\beta=1$, and the normal at $\beta=2$. Box and Tiao restrict attention to $\beta\geq 1$, the densities for $\beta<1$ being extremely leptokurtic. However, the range $\beta<1$ is just that

section of this family which are outlier-prone, the others being outlier-resistant. To see this consider the conditions of 2.2.1.

Firstly, condition (iii) of (2.1.8); "uniformity" in the tails.

For $y>0$, $h>0$

$$-\ln\left[\frac{p(y+h)}{p(y)}\right] = -y^{\beta} + (y+h)^{\beta}$$

$$= y^{\beta}\left|\left(1+\frac{h}{y}\right)^{\beta}-1\right| .$$

So for $y^{-1}h<1$,

$$-\ln\left[\frac{p(y+h)}{p(y)}\right] = y^{\beta}\left\{-\frac{\beta h}{y} + \frac{\beta(\beta-1)}{2}\frac{h^2}{y^2}+\ldots\right\} \quad . \qquad (1)$$

Thus

a) When $0<\beta<1$, (1) is $\frac{-\beta h}{y^{1-\beta}} + 0\left(\frac{1}{y^{2-\beta}}\right)$

which tends to zero as $y\to\infty$, so (2.1.8) is satisfied.

b) When $1\leq\beta<2$ we have from (1)

$$\frac{p(y+h)}{p(y)} \sim \exp(\beta h) \text{ as } y\to\infty$$

So (2.1.8) is not satisfied. The outlier-resistance for $1<\beta\leq 2$ is noted by O'Hagan (1979) and follows from the strong unimodality of the density for such β.

Notice that a) and β) indicate that (2.1.8) will only be satisfied for densities which decay no faster than $\exp\{-k|y|\}$ as mentioned in §2.1.

The outlier-proneness for $\beta<1$ now follows since, clearly, the score function is decreasing function of $y>0$.

(iii) Student t.

The score of a Student t density with k degrees of freedom is

$$(k+1)y/(k+y^2), \qquad k>0,$$

and ultimately redescends to zero.

To prove (iii) of (2.2.8), note that

$$\left\{\frac{p(y+h)}{p(y)}\right\}^{\left(\frac{2}{k+1}\right)} = \frac{k+y^2}{k+(y+h)^2} \to 1 \text{ as } |y| \to \infty$$

and so the Student t distributions are outlier-prone.

This family provides a reasonably tractible density form with robust properties and allows for a choice of robustness parameter in the degrees of freedom k. We shall use the t distributions extensively in following Chapters.

(iv) Stable densities.

The symmetric stable distribution of index a, $1 \leq a \leq 2$, has standard characteristic function

$$\chi(t) = \exp - |t|^a , \; t \in \mathbb{R}$$

and a (regular) density on $\mathbb{R}$ which is unimodal at zero. (See, for example, Ibragimov and Linnik (1971)). The moments of p(y) exist only for order less than a, and so the distributions can be seen to be heavy-tailed, lying between the Cauchy (a=1), and the normal (a=2). p(y) is given by the inverse Fourier transform of χ which can be expressed as

$$p(y) = \sigma^{-1} \int_0^\infty \cos(ty) e^{-t^a} dt.$$

Thus p(y) is continuously differentiable in y. The behaviour of p(y), g(y) and G(y) is not well known and we examine asymptotic expansions of these functions. The following Lemma provides the means of doing this; it is a simple extension of Theorem 2.4.2 of Ibragimov and Linnik (1971) (page 55), and the proof follows their proof with relevant minor changes.

Lemma A2.2.1.

For any a, $1<a<2$, $k=0,1,2,\ldots$ and $x>0$, define

$$I_k(x) = \int_0^\infty \exp\{-it-(t/x)^a\}.t^k dt.$$

Then $$I_k(x) = (-i)^k \sum_{n=0}^{N} \frac{(-1)^{n+1}}{n!x^{na}} \Gamma(na+k+1).\left\{\sin\left(\frac{n\pi a}{z}\right) - i\cos\left(\frac{n\pi a}{z}\right)\right\}$$

$$+ O(x^{-(N+1)a-1}), \quad \text{for all } N\geq 1.$$

Proof.

Note that the above mentioned Theorem calculates $I_0(x)$. We follow that proof by substituting $te^{i\phi}$ for t in the integrand, where $\phi = -\pi/2a$.

Then

$$I_k(x) = e^{i\phi(k+1)} \int_0^\infty \exp\{-te^{i(\pi/2+\phi)} + i(t/x)^a\} t^k dt$$

which, on expanding

$$e^{i(t/x)^a} = \sum_{n=0}^{N} \frac{(t/x)^{na}}{n!} i^n + \gamma\frac{(t/x)^{(N+1)a}}{(N+1)!}, \quad |\gamma| \leq 1,$$

gives

$$I_k(x) = \sum_{n=0}^{N} \frac{e^{i\phi(k+1)}}{n!x^{na}} i^n \int_0^\infty t^{na+k} \exp\{-te^{i(\pi/2+\phi)}\}\, dt$$

$$+ O\left(\frac{x^{-(N+1)a}}{(N+1)!} \int_0^\infty e^{-t\cos(\pi/2+\phi)}\, t^{(N+1)a+k} dt\right)$$

Making a further rotation with $te^{i\theta}$ replacing t, where now $\theta = -\pi/2-\phi$, we have

$$\int_0^\infty t^{na+k} \exp\{-te^{i(\pi/2+\phi)}\}dt = e^{i\theta(1+na+k)} \cdot \Gamma(na+k+1)$$

and so

$$I_k(x) = e^{-\frac{ikn}{2}} \sum_{n=0}^{N} e^{\frac{i\phi_i n_\ell i\theta(na+1)}{n!x^{na}}} \Gamma(na+k+1) + O\left(\frac{\Gamma((N+1)a+k+1)}{(N+1)!\, x^{(N+1)a}}\right)$$

which, on substituting the values of ϕ and θ gives the required result.

As a check note that

$p(y) = \mathrm{Re}\left[\pi^{-1}y^{-1}I_0(x)\right]$ gives the asymptotic expansion of Ibragimov and Linnik.

Lemma A2.2.2.

As $y\to\infty$, $g(y) = cy^{-1} + O\left(y^{-(a+2)}\right)$,

and $G(y) = (d+c^2)y^{-2} + O\left(y^{-(a+3)}\right)$,

where

$$c = \Gamma(a+2)/\Gamma(a+1) \text{ and } d = \Gamma(a+3)/\Gamma(a+1).$$

Proof: Since $p(y) = \pi^{-1}\int_0^\infty \cos(ty)e^{-t^a}dt$,

then $-\frac{\partial p(y)}{\partial y} = \pi^{-1}\int_0^\infty t\sin(ty)e^{-t^a}dt$

$= \pi^{-1}y^{-2}\int_0^\infty u\sin(u)e^{-(u/y)a}du$

$= -\pi^{-1}y^{-2}\,\mathrm{IM}\left[I_1(y)\right]$.

which, by Lemma A2.2.1, is

$$y^{-1}\cdot\left\{\frac{\Gamma(a+2)\sin(\pi a/2)}{y^a} + O(y^{-2a-1})\right\}.$$

Thus

$$g(y) = y^{-1}\left\{\frac{\Gamma(a+2)\sin(\pi a/2)}{y^a} + O(y^{-2a-1})\right\}$$

$$\times\left\{\frac{\Gamma(a+1)\sin(\pi a/2)}{y^a} + O(y^{-2a-1})\right\}^{-1}$$

$= y^{-1}\left\{c + O(y^{-a-1})\right\}$, as required.

Similarly, since $\frac{\partial^2 p(y)}{\partial y^2} = \pi^{-1}y^{-3}\,\mathrm{Re}\left[I_2(y)\right]$, the expression for G follows.

NB. Dumouchel (1973) proves that the asymptotic expansion of Ibragimov and Linnik, $p(y) = \pi^{-1}y^{-1}\ \mathrm{Re}[I_0(y)]$ is differentiable term by term as a function of y and we could use this to prove Lemma 2.2.2 much more simply. However, the method of Lemma 2.2.1 was used in the example of the discussion of Masreliez's Theorem in §2.2.3 and so we retain it. It is trivial to check that this approach leads to the same result.

Now, we cannot deduce that the stable distributions are outlier-prone. Certainly the uniformity of the tails condition is satisfied immediately from the asymptotic expansion for p(y). However, it is not clear that g(y) is monotonically redescending to zero. Numerical studies suggest a form similar to Student t scores and we conjecture that the stable distributions are outlier-prone.

(v) Others. The following distributions provide shapes near to normality and all lie on the border of outlier-proneness/resistance, with the exception of (d).

(a) Logistic.

With $p(y) \alpha \mathrm{Sech}^2(y/2)$ we have a score function

$g(y) = \tan h(y/2)$, monotonically increasing but bounded.

Further $G(y) = \frac{1}{2}\ \mathrm{sech}^2(y/2)$.

(b) Huber k (See Example 2.2.5 for definition).

$$p(y)\ \alpha \begin{cases} \phi(y) & , \quad |y| \leq k; \\ \exp -k|y| & , \quad \text{otherwise}, \quad k \geq 0. \end{cases}$$

For k=0 this is just the Laplace.

(c) Extreme type.

A shape similar to the logistic is provided by the score

$$g(y) = (1-e^{-|y|}).\mathrm{sgn}(y),$$

corresponding to a density

$$p(y) \propto \exp\{-|y|+e^{-|y|}\}.$$

This is essentially a smoothed version of the Laplace density providing a means of removing the irregularity at the origin.

(d) Normal/uniform.

When $y = x/u$ where $x \sim N(0,1)$ and $u \sim U[0,1]$ are independent, then

$$p(y) \propto y^{-2}\left[1-e^{-y^2/2}\right]$$

and so

$$g(y) = 2y^{-1} - \left[1-e^{-y^2/2}\right]^{-1} y\, e^{-y^2/2}$$

This is similar in form to the Cauchy score although we note that all moments of $p(y)$ exist.

(vi) The ϵ-contaminated family.

The classical theory of robust estimation has often been concerned with providing sampling theory procedures that perform well when the errors are generated by a contaminated normal mixture of the forms

$$p_h(y) = (1-\epsilon)\phi(y) + \epsilon h(y), \quad 0 \leq \epsilon < 1, \text{ h symmetric.}$$

The family of such mixtures i.e.

$$= \{p_h | h \text{ symmetric}\}$$

is called the ϵ-contaminated family. Huber (1964) and (1977) adopted this family as the focus for his development of minimax robust estimation, and mixtures of this form have been discussed as providing reasonable approximations to "real-life" error distributions. The use of a normal contaminant h in both classical and Bayesian studies is essentially a parsimonious attempt to model the data when h is unknown. Although this works well as shown by Box

and Tiao (1968) and, in a time-series context, Harrison and Stevens (1976), when the contaminant has a large variance, we have seen that outlier-rejection cannot be obtained in terms of posterior convergence to prior in the location problem with a normal mixture. However, with robust h, we have the following result.

Lemma 2.2.3

Let h be unimodal symmetric at zero. If the distribution H with density h is outlier-prone, then so is $P_H = (1-\epsilon)\Phi + \epsilon H$.

Proof:

h decays no faster than $e^{-k|y|}$, some $k>0$ and therefore note that

$\lim_{y\to\infty} \{\phi(y)h^{-1}(y)\} = 0.$

Thus $\left\{\frac{p_h(y+\delta)-p_h(y)}{p_h(y)}\right\} = \frac{(1-\epsilon)\left\{\frac{\phi(y+\delta)}{\phi(y)} - 1\right\}\frac{\phi(y)}{h(y)} + \epsilon\left\{\frac{h(y+\delta)}{h(y)} - 1\right\}}{\left\{(1-\epsilon)\frac{\phi(y)}{h(y)} + \epsilon\right\}}$

$\to 0$ as $|y|\to\infty$ for $\delta>0$.

So (iii) of (2.1.8) holds.

Further, if g_h is the score of p_h, then the score of the mixture g_p is

$$g_p(y) = -\left\{(1-\epsilon)\frac{\phi'(y)}{h(y)} + \epsilon\frac{h'(y)}{h(y)}\right\}\left\{(1-\epsilon)\frac{\phi(y)}{h(y)} + \epsilon\right\}^{-1}.$$

Now both $\frac{\phi'(y)}{h(y)}$ and $\frac{\phi(y)}{h(y)}$ tend to zero as $|y|\to\infty$, so

$$g_p(y) - g_h(y) \to 0 \text{ as } y\to\infty.$$

Thus g_p behaves like g_h for large y. □

Finally note that $g_p(y)$ always lies between y and $g_h(y)$, since if in general

$$p(y) = \sum_{j=1}^{k} \pi_j p_j(y),$$

then Lemma 2.2.1 implies that

$$g(y) = -\frac{\partial}{\partial y} \ln p(y) = \sum_{j=1}^{k} \pi_j^* g_j(h),$$

where $g_j(y)$ is the score of $p_j(y)$, and

$$\pi_j^* \alpha \pi_j . p_j(y), \quad \text{with} \sum_{j=1}^{k} \pi_j^* = 1.$$

Hence, for all y,

$$\min_j \{g_j(y)\} \leq g(y) \leq \max_j \{g_j(y)\}.$$

A2.3. Scale mixtures of normals.

Continuous (and discrete) scale mixtures of normal densities provide useful methods of generating samples from symmetric distributions and in computing characteristics of sampling theory estimators as used by Andrews et al (1972) and Relles (1970). This family provides a natural framework in which to explore possible alternatives to normality and for completeness it is interesting to note that all of the robust likelihoods discussed above are continuous scale mixtures of normals.

Let $x \sim N[0,1]$ and $v>0$ be independent of x, with $y = xv$. Then

(i) It is well known that, if $v^{-2} \sim X_n^2$, $n>0$, then y is Student t-n.

(ii) Following Feller (1966, p.172), if v^2 is stable of index $a<1$, then y is stable of index 2a. In particular if $a \geq \frac{1}{2}$, then y is symmetric stable and heavy-tailed.

(iii) Andrews and Mallows (1974) show that if v/2 has the asymptotic distribution of the Kolmogorov distance statistic then y has a logistic distribution.

(iv) A further result of Andrews and Mallows is that if $v^2/2$ is exponential, then Y has a double-exponential distribution. In fact this is a special case of the following result which, as far as we are aware, has not appeared in the literature.

Lemma 2.3.1.

The exponential power distributions of index $0<a<2$ are scale mixtures of normals.

Proof Let $p(y) = ke^{-|y|^a}$, $0<a\leq 2$.

Now $p(y)$ is the characteristic function of a stable random variable of index a, thus, if f denotes the density of such a random variable, we have,

$$p(y) = k \int_{-\infty}^{\infty} e^{iyt} f(t)dt.$$

Moreover, from (ii) above,

$$f(t) = \int_0^{\infty} v^{-1}\phi(v^{-1}t).g(v)dv$$

where $g(v)$ is the density of v when v^2 is stable of index $a/2$. So

$$p(y) = k \int_0^{\infty} g(v) \int_{-\infty}^{\infty} e^{iyt} v^{-1}\phi(v^{-1}t)dt\ dv$$

$$= k \int_0^{\infty} g(v)e^{-y^2v^2/2}\ dv$$

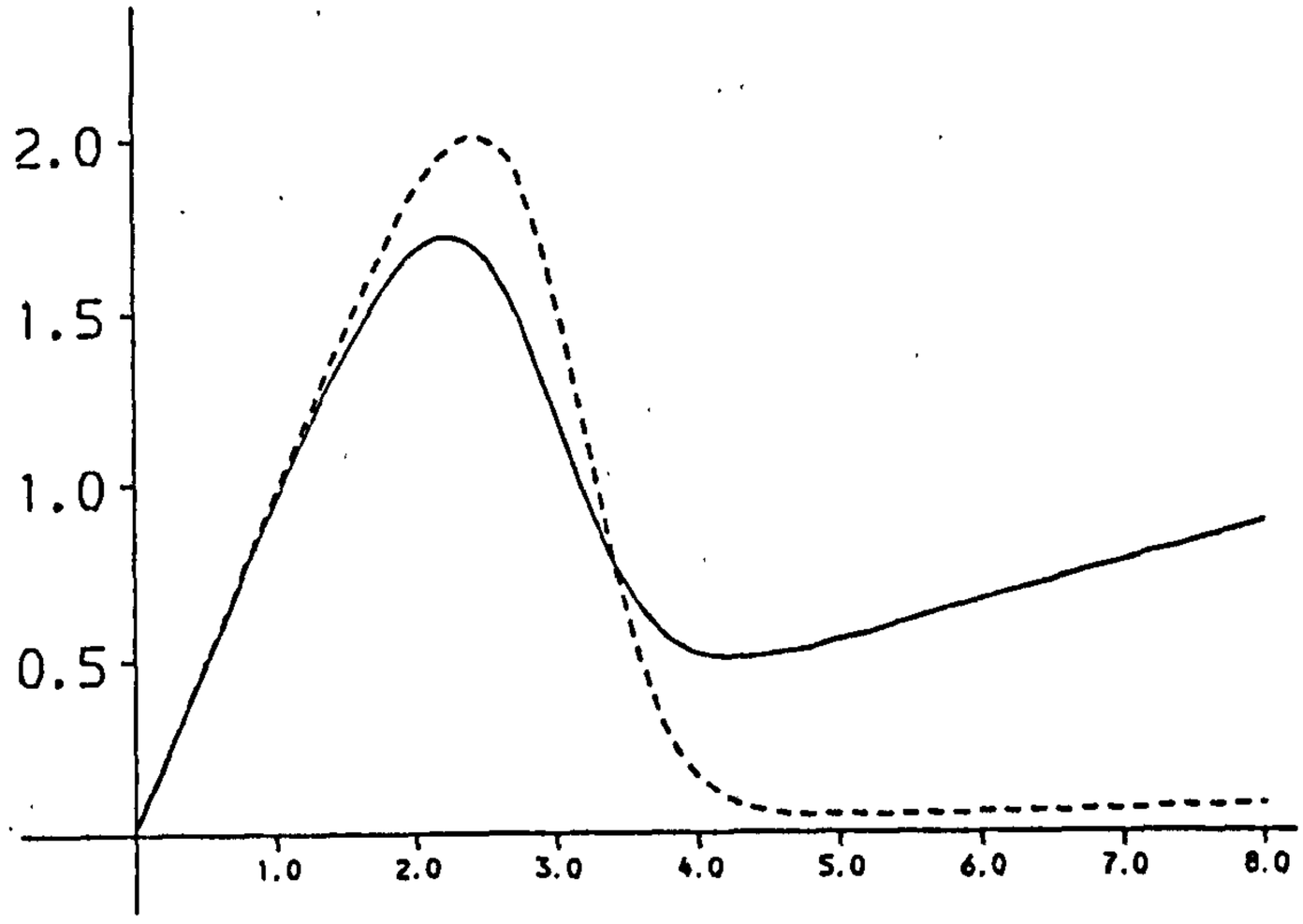
$$\alpha \int_0^{\infty} u^{-1}g(u^{-1}).u^{-1}\phi(u^{-1}y)du.$$

and the result follows. The mixing density is then proportional to $u^{-1}g(u^{-1})$.

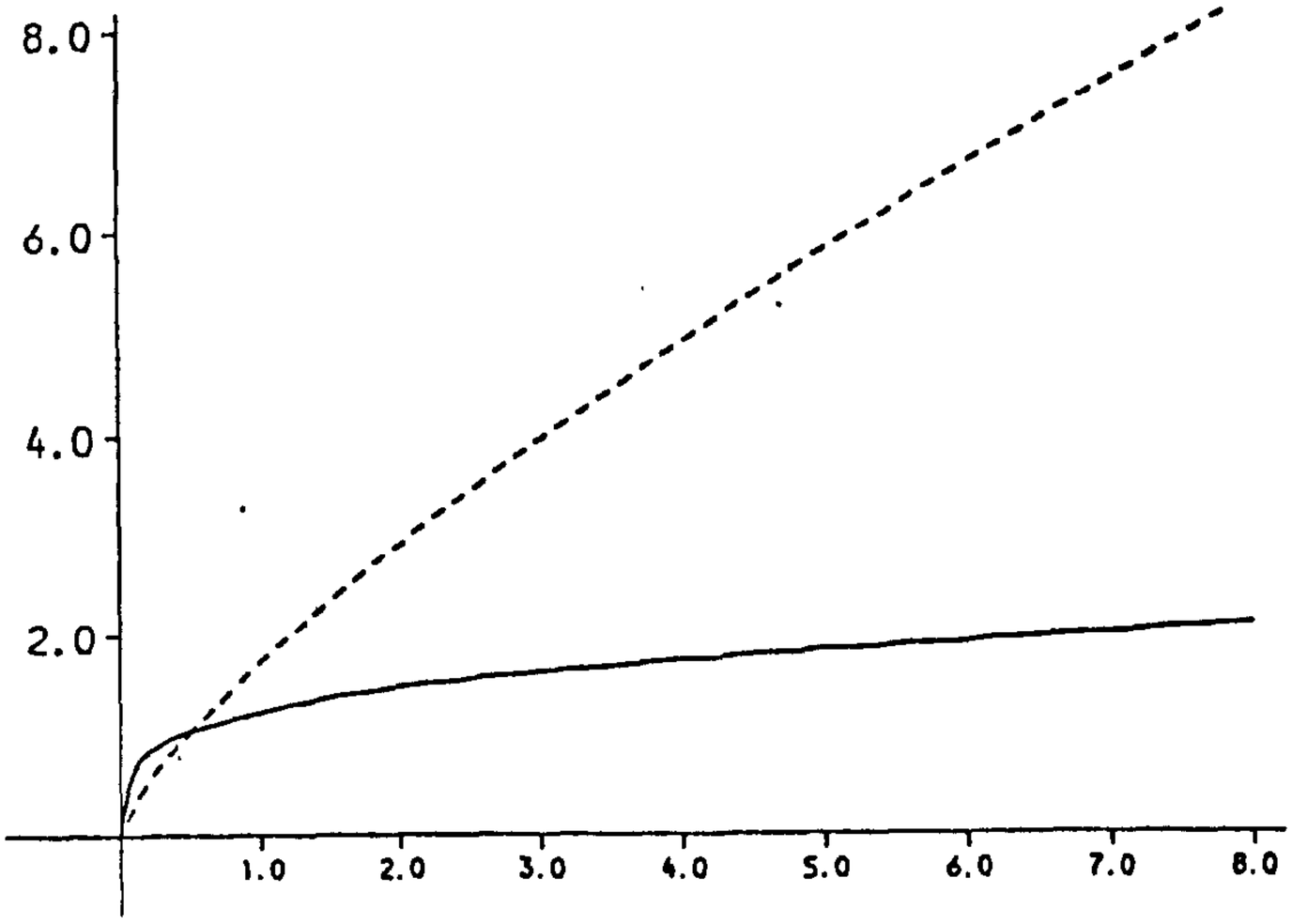
The following figures illustrate the above examples. In each figure the first named score function is given by the full line and the second by the dashed line.

A. CN(0.1,9) , CN (0.1,100)
B. EXPONENTIAL POWER, B= 1.25 , 1.75
C. EXPONENTIAL POWER, B= 1.00 , 0.50

A.



B.



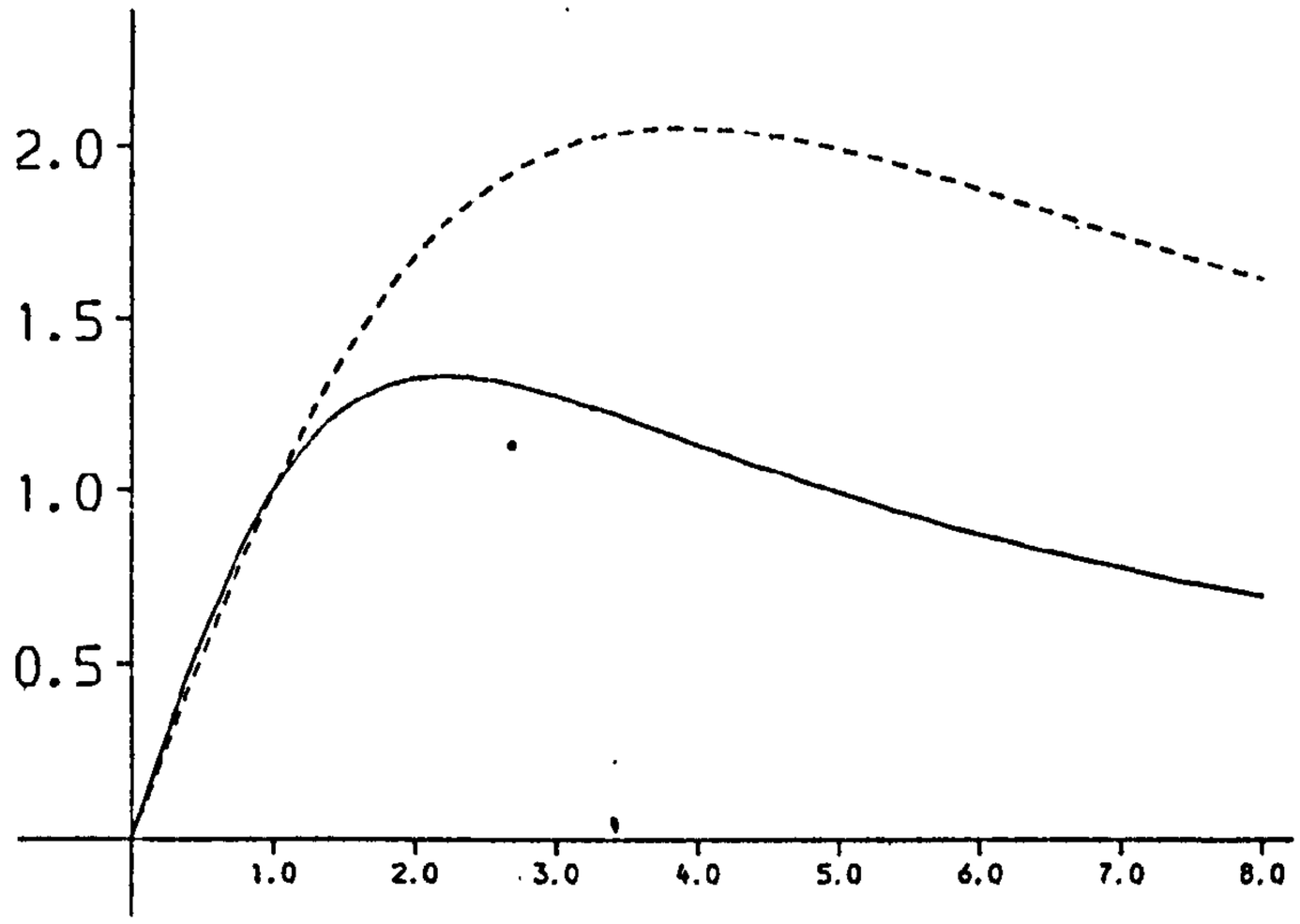
C.

A. STUDENT T , K= 5, 15

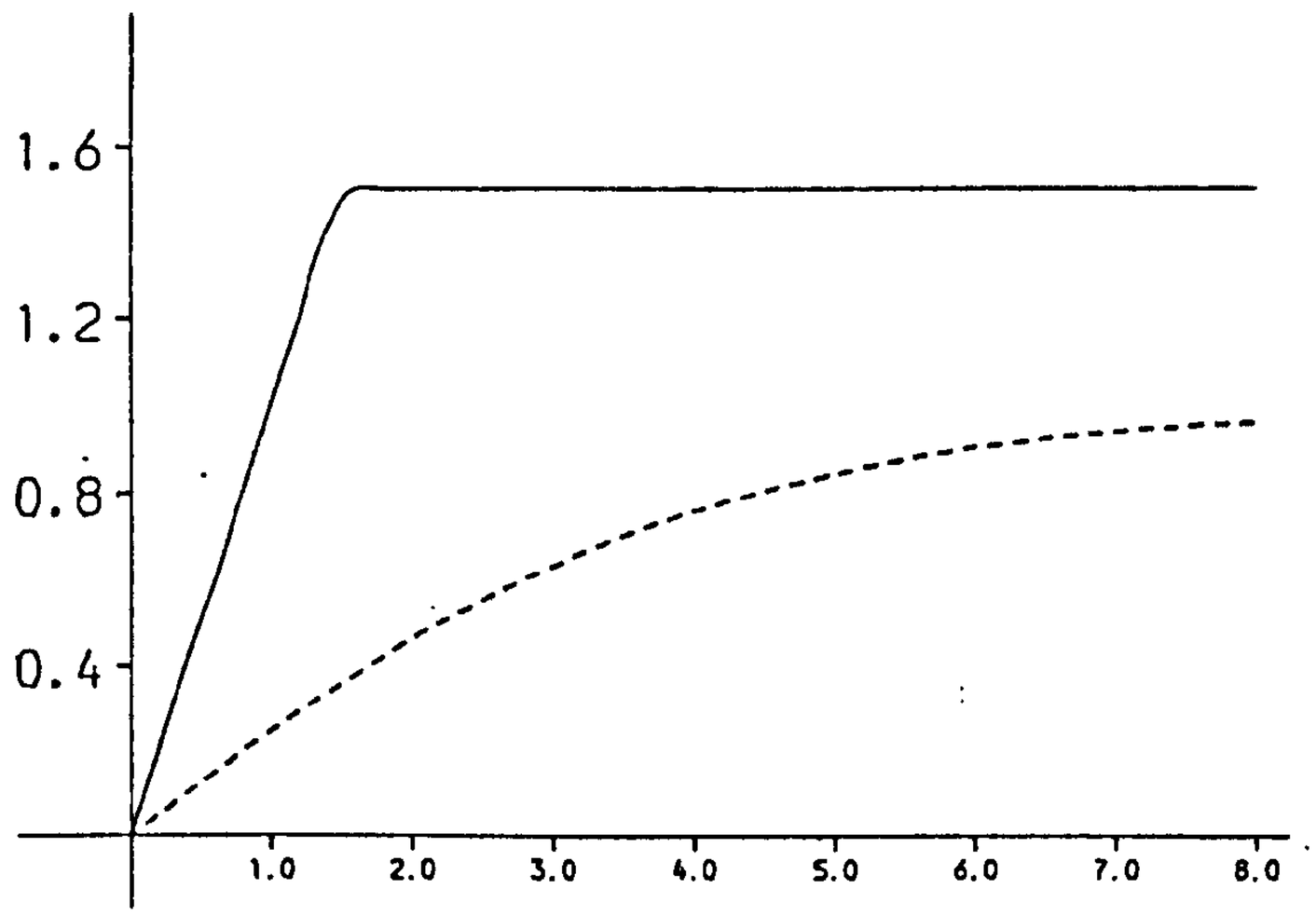
B. HUBER(1.5) , LOGISTIC

C. EXTREME TYPE , NORMAL/UNIFORM

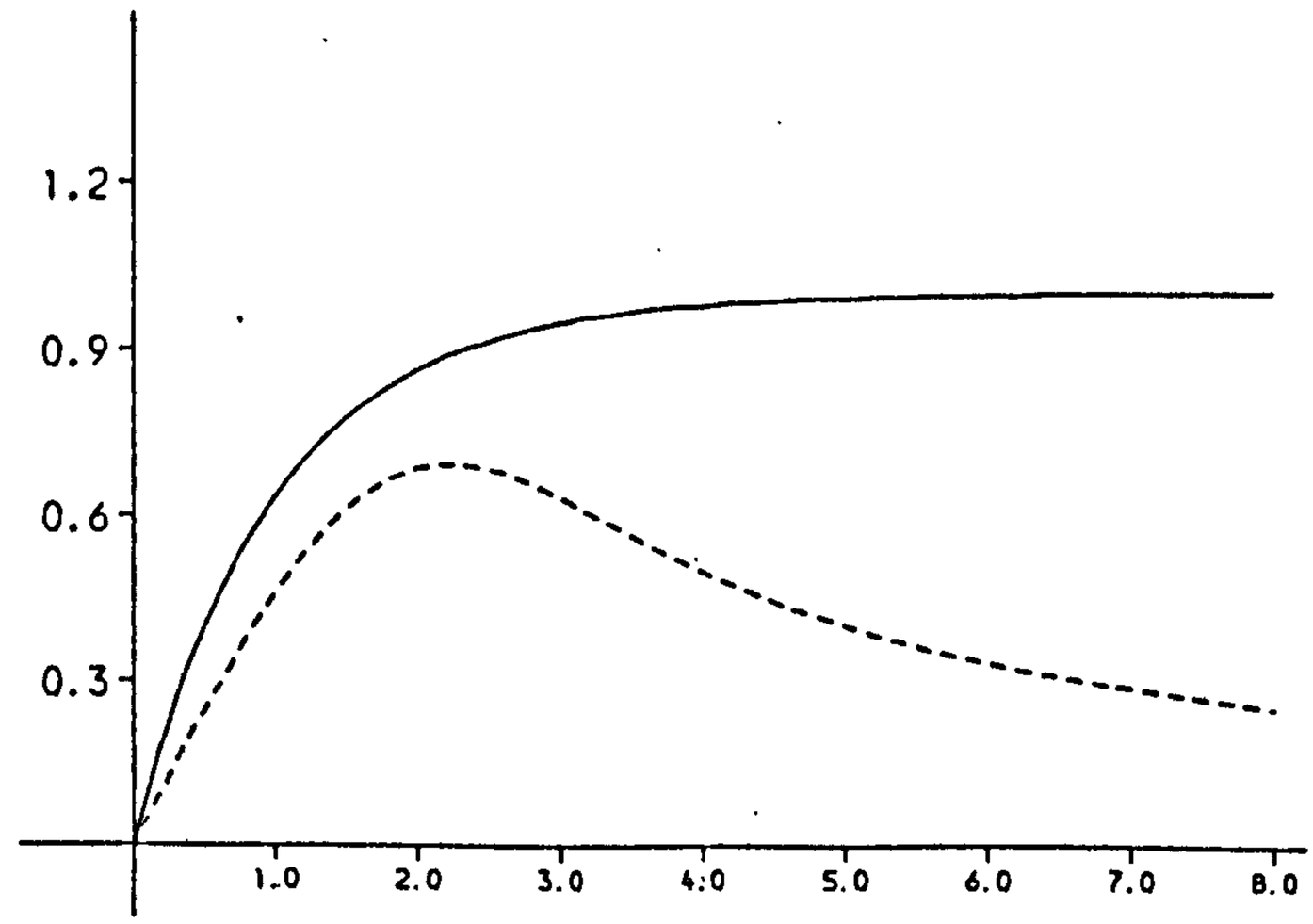
A.



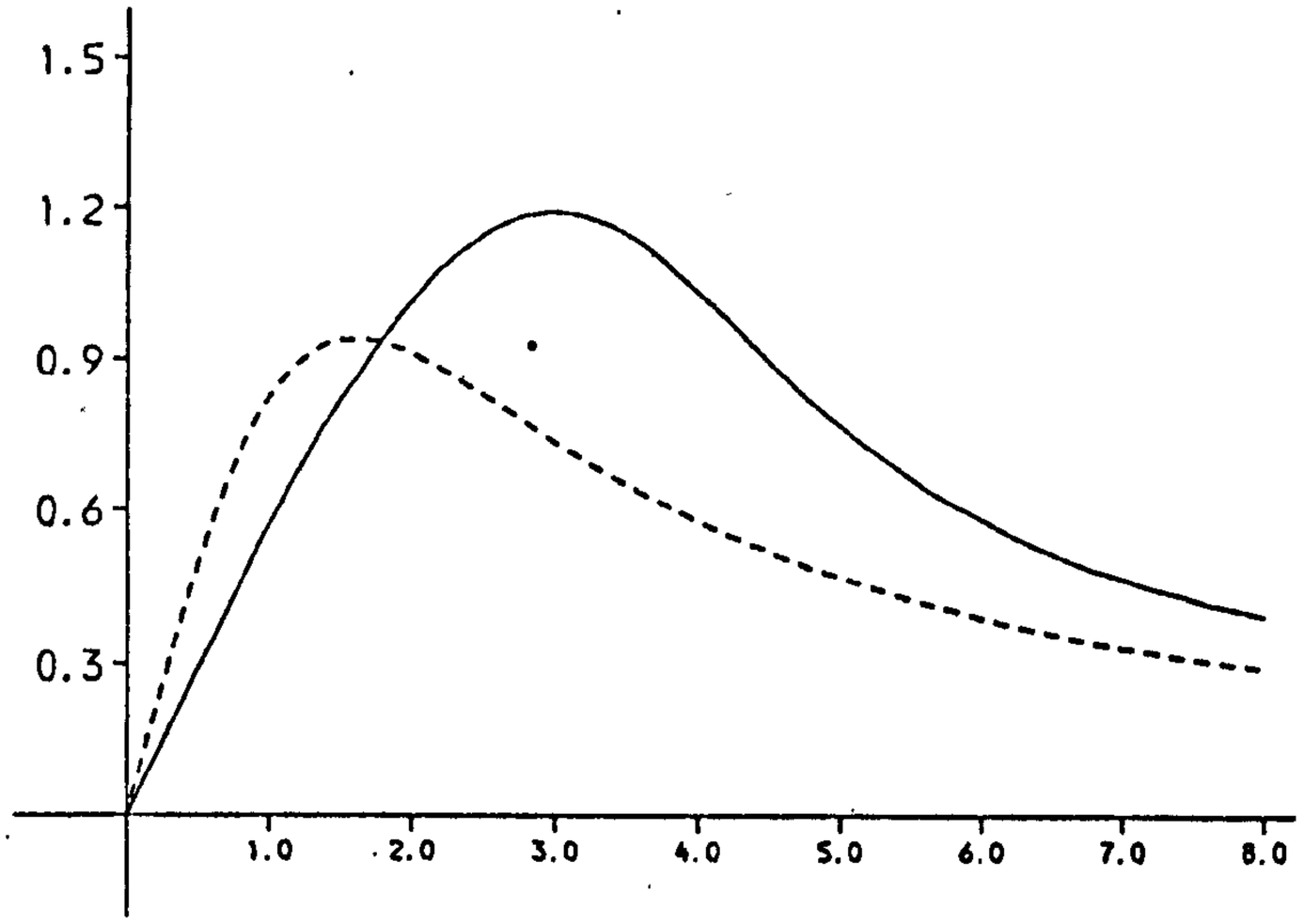
B.



C.



STABLE, INDEX A= 1.25 , 1.75



CHAPTER 3. The dynamic linear model.

3.1 Introduction

In this and the next Chapter we study outlier problems in the dynamic linear time series model described as follows:

Let $\underset{\sim}{Y}_1$, $\underset{\sim}{Y}_2$,... be a sequence of (m×1) vectors of observations. At time n, $\underset{\sim}{Y}_n$ is related to the (p×1) vector of parameters $\underset{\sim}{\theta}_n$ by the observation equation

$$\underset{\sim}{Y}_n = H_n \underset{\sim}{\theta}_n + \underset{\sim}{\nu}_n \quad , n=1,2,... \tag{3.1.1}$$

where the (m×p) regression matrix H_n is known at time n and $\{\underset{\sim}{\nu}_n;\ n=1,2,...\}$ is a sequence of zero mean, independent random vectors each with the same distribution, and $\underset{\sim}{\nu}_n$ is independent of $\underset{\sim}{\theta}_{n-1}$ given the past observations $D_{n-1} = \{\underset{\sim}{Y}_1,..., \underset{\sim}{Y}_{n-1}\}$.

The system parameter vector $\underset{\sim}{\theta}_n$ obeys the Markov evolution

$$\underset{\sim}{\theta}_n = G_n \underset{\sim}{\theta}_{n-1} + \underset{\sim}{\omega}_n \quad , n=1,2,... \tag{3.1.2}$$

where the (p×p) transition matrix G_n is known at time n, and $\{\underset{\sim}{\omega}_n;\ n=1,2,...\}$ is a sequence of zero mean, independent and identically distributed random vectors, with $\underset{\sim}{\omega}_n$ independent of $\underset{\sim}{\theta}_{n-1}$ given D_{n-1}. In addition we assume that the sequences $\{\underset{\sim}{\nu}_n\}$ and $\{\underset{\sim}{\omega}_n\}$ are independent.

This model formulation was used by Harrison and Stevens (1976) and as such does not coincide with the usual control theoretic state space formulation (See, for example, Anderson and Moore (1979)). $\underset{\sim}{\theta}_n$ is a stochastic parameter vector which can be interpreted as a time varying regression vector, although, in many cases, it is not unreasonable to treat $\underset{\sim}{\theta}_n$ as a state vector as in the control theoretic context. In Chapter 4 we do just that in a discussion of the state space representation of autoregressive and autoregressive-moving average models for which it is desirable, from the point of

view of modelling outliers, to consider such a representation. In particular we discuss the problem of outliers in the evolution equation of a state vector and clearly this is essentially a problem of modelling the errors in (3.1.2), in so far as the technical details are the same.

In this Chapter we restrict attention to possible outliers in the observation equation (3.1.1). Here we must make an important distinction between types of models as follows. If H_n is a matrix of regressors not involving the data D_{n-1}, then an outlier in (3.1.1) at time n will affect only $\underset{\sim}{Y}_n$ and not future observations (assuming of course that G_n does not involve D_{n-1} either). In the terminology of Fox (1972), such an observation will be called an additive outlier. However if H_n (and/or G_n) depends upon D_{n-1}, and in particular upon Y_{n-1}, then the effect of an outlier at time n is clearly carried through to time n+1, n+2,..., via the regression matrix (and/or transition matrix G_n). Again following Fox we call this an innovations outlier, using this terminology for more general models than the autoregressions discussed by Fox.

In distinguishing these two types of outliers in autoregressive/moving average models we must resort to the state space formulation and, as mentioned above, this is done in the next Chapter. For the moment we examine (3.1.1) as it stands, making no distinction between data dependent and independent matrices H_n and G_n.

Now the Markovian nature of the model leads naturally to a sequential approach to estimation of $\underset{\sim}{\theta}_n$, forecasting future Y_{n+j}, j=1,2,..., and smoothing i.e. "forecasting" into the past, values of $\underset{\sim}{\theta}_{n-j}$, j=1,2,..., and a retrospective analysis is generally extremely unwieldy. Typically the model will be much simplified with scalar observations and/or parameters, and time independent matrices

G_n and/or H_n.

In a sequential analysis, there are two major operations required for the estimation of $\underset{\sim}{\theta}_n$ within a coherent framework:

(i) the so-called time-update

$$p(\underset{\sim}{\theta}_n|D_{n-1}) = \int p(\underset{\sim}{\theta}_n|\underset{\sim}{\theta}_{n-1}).p(\underset{\sim}{\theta}_{n-1}|D_{n-1})d\underset{\sim}{\theta}_{n-1}$$

providing the parameter predictive density on the left hand side.

(ii) the prior to posterior update

$$p(\underset{\sim}{\theta}_n|D_n) \propto p(\underset{\sim}{\theta}_n|D_{n-1}).p(y_n|\underset{\sim}{\theta}_n,D_{n-1})$$

It is immediately clear that the linear/normal framework makes these two operations simple and tractible and leads to the Kalman filter recursive algorithm for the mean and covariance matrix of the normal posterior distribution derived in (ii). Changing the normality assumption for either $\underset{\sim}{\nu}_n$ or $\underset{\sim}{\omega}_n$ destroys this analytic tractibility and in order to obtain "exact" results within a Bayesian framework we must resort to numerical integration techniques. Specifically we must calculate the parameter predictive density of (i) pointwise for each value of $\underset{\sim}{\theta}_n$ in $\mathbb{R}^p$ and similarly for the prior to posterior analysis. Clearly this is not feasible. The computational effort required is prohibitive and some form of approximation to the Bayesian analysis is desirable. Our intention here is to examine some possible approximations in the case where we assume non-normality of the observational errors $\underset{\sim}{\nu}_n$.

Since we are assuming that outliers may occur only in (3.1.1), we retain the standard normality assumption for $\underset{\sim}{\omega}_n$,

$$\underset{\sim}{\omega}_n \sim N[0,W_n], \quad n=1,2,...$$

where the W_n are known positive definite matrices. Usually $W_n = W$ for all n. Further, we assume that, at time n, the density $p(\underset{\sim}{\theta}_{n-1} | D_{n-1})$ is approximately normal

$$(\underset{\sim}{\theta}_{n-1} | D_{n-1}) \sim N[\underset{\sim}{m}_{n-1}, C_{n-1}]$$

where $\underset{\sim}{m}_{n-1}$ and C_{n-1} are functions of D_{n-1}. The validity of this approximation is discussed in detail later, and it turns out to be justifiable in certain circumstances. We remark that it is in the first instance a convenient approximation from the point of view of tractability of analysis, for it implies that the time update (i) above is the same as the usual normal theory, giving

$$(\underset{\sim}{\theta}_n | D_{n-1}) \sim N[\underset{\sim}{a}_n, P_n]$$

where $$\underset{\sim}{a}_n = G_n \underset{\sim}{m}_{n-1},$$

and $$P_n = G_n C_{n-1} G_n^T + W_n.$$

Furthermore, this assumption enables us to use the more general form of Masreliez's result, our Theorem 2.4.1, to obtain the posterior mean and covariance matrix for $\underset{\sim}{\theta}_n | D_n$ in closed form. In general $p(\underset{\sim}{\theta}_{n-1} | D_{n-1})$ is not of course, exactly normal, but, whilst admitting that an examination of the exact density has no substitute as far as coherent inference and rational decision making are concerned, there seems much justification in the context of our model for behaving pragmatically and adopting useful approximations to the full Bayesian analysis. Given the data D_{n-1}, we are at liberty to examine $p(\underset{\sim}{\theta}_{n-1} | D_{n-1})$, (with all the associated problems of numerical integration for moments and marginal posterior distributions, and visualization of possibly high dimensional densities), but for the purpose of proceeding to the next observation stage, a sensible approximation which provides a tractible prior usefully summarizing our beliefs about $\underset{\sim}{\theta}_{n-1} | D_{n-1}$ has much to recommend itself. In fact we shall see

that in the problem of interest here such an approach will generally result in very little loss of precision and, indeed, this approach seems to be implicit in the mixture modelling of Harrison and Stevens (1976) where an exploding mixture of posterior distributions at each observation stage is "collapsed" to an approximation for the same reasons.

3.2 Scalar observations: filtering with heavy-tailed errors.

3.2.1 General comments and prior specification.

In modelling the error distribution of the ν_n with a view to robustifying the usual conjugate normal analysis we assume that p_ν is unimodal, symmetric at zero and heavy-tailed relative to the normal density ϕ, in the sense that the score function g_ν is bounded above by a linear function

$$|g_\nu(u)| = \left|\frac{\partial}{\partial u} \ln p_\nu(u)\right| < k|u|, \quad \text{for some } k>0.$$

This admits distributions which are not outlier-prone such as the exponential family of index β between one and two, whereas from a practical viewpoint we believe that a restriction to redescending score functions is desirable. In particular the Student t distributions provide useful alternatives to normality.

Following the discussion of §3.1, our prior for $\underset{\sim}{\theta}_n|D_{n-1}$ is $N[\underset{\sim}{a}_n, P_n]$ and within this framework the following result was proved by Masreliez as the general form of Theorem 2.4.1:-

$$E[\underset{\sim}{\theta}_n|D_n] = \underset{\sim}{a}_n + P_n \underset{\sim}{h}_n g(y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n), \tag{3.2.1}$$

and

$$\text{var}[\underset{\sim}{\theta}_n|D_n] = P_n - P_n \underset{\sim}{h}_n \underset{\sim}{h}_n^T P_n . G(y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n), \tag{3.2.2}$$

where, setting $u_n = y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n$, we have

$$g(u_n) = -\frac{\partial}{\partial y_n} \ln p(y_n) \text{ and } G(u_n) = \frac{\partial g}{\partial u_n}(u_n), \qquad (3.2.3)$$

with

$$p(y_n) = p(y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n) = \int_{\mathbb{R}^p} p_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n) \cdot p(\underset{\sim}{\theta}_n | D_n) d\underset{\sim}{\theta}_n .$$

So the correction made to the prior mean $\underset{\sim}{a}_n$ for $E[\underset{\sim}{\theta}_n | D_n]$ is a linear transformation of the smoothed score (or influence function)

$$\underset{\sim}{h}_n g(y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n) = E\left[-\frac{\partial}{\partial \underset{\sim}{\theta}_n} \ln p_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n) | D_n \right].$$

The linear transformation P_n essentially weights the correction term according to the uncertainty in the prior. We can relate this result to the empirical influence function discussed by Cox and Weisberg (1980) in the context of a static regression model corresponding to $\underset{\sim}{\theta}_n = \underset{\sim}{\theta}_m$ for all n,m. In this case $\underset{\sim}{a}_n = \underset{\sim}{m}_{n-1}$ and $P_n = C_{n-1}$. The empirical influence function IF_n for y_n given D_{n-1} is used by Cox and Weisberg to quantify the effect of y_n on the estimates of regression parameters and, from a Bayesian point of view, would correspond to

$$IF_n = \underset{\sim}{m}_{n-1} - \underset{\sim}{m}_n ,$$

a difference in posterior means. Of course Cox & Weisberg operate wholly within a normal model when IF_n is a linear function of y_n. Under the above conditions Mazreliez's result implies that, for non-normal p_ν,

$$IF_n = -P_n \underset{\sim}{h}_n g(y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n),$$

involving directly the classical influence function of p_ν.

Now in Masreliez's original paper (1975), he uses the above result to compute recursive estimates of $\underset{\sim}{\theta}_n$ in some examples. However, the error density he uses is a contaminated normal mixture for which a closed form expression can be obtained for the marginal score function. In fact this procedure coincides with the mixture modelling

of Harrison and Stevens (1976) when using such an error density. The philosophy behind its implementation in this framework is that the use of a heavy-tailed likelihood classifies observations according to how extreme they are and then collapsing the true mixture of posterior distributions to a single normal with the same mean and covariance matrix leads to an analysis close to the full explosive Bayesian analysis. Indeed Masreliez's numerical results compare the "collapsed" filters with the exact posterior mean and covariance matrix and his results indicate that the approximation is excellent in the cases he studied.

What about using alternative heavy-tailed error densities, and in particular Student t distributions? Well in general we cannot use Masreliez's result without calculating the marginal score by numerical integration. The exceptional case is for the Huber family of distributions as discussed in example 2.4.3 where the score function was computed. The expressions for g and G are somewhat tedious to compute, but this does give us a means of using the Huber distributions, (much used in the classical approach to robust estimation) in this time series model, without resorting to numerical techniques. Indeed, Masreliez and Martin (1977) use the Huber family in their minimax development of the filtering algorithms of Masreliez's original work, and their results are extremely encouraging. Broadly speaking the robust filters behave similarly to the Kalman filters for near normal data and yet have all the benefits of rejecting outliers which derive from the use of robust likelihoods. We discuss this further in the next section.

3.2.2. Review of Masreliez and Martins' work.

In their (1977) paper, Masreliez and Martin do not actually calculate the marginal score and information functions in order to provide the true mean and covariance matrix defined in (3.2.1) and

(3.2.2), but rather they use an approximation based upon a scaled version of the likelihood as follows:

Approximate $p(y_n)$ by the scaled likelihood

$$p(y_n) \simeq \sigma_n^{-1} p_\nu(\sigma_n^{-1} u_n) \ , \quad u_n = y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n, \tag{3.2.4}$$

where σ_n is a scale factor to be defined. Note that p_ν has been assumed to have scale parameter known and equal to unity and thus, if p_ν is normal, (3.2.4) holds exactly with

$$\sigma_n^2 = q_n^2 + 1, \quad \text{where } q_n^2 = \underset{\sim}{h}_n^T P_{n-1} \underset{\sim}{h}_n. \tag{3.2.5}$$

In view of this, Masreliez and Martin suggest that for general heavy-tailed p_ν, σ_n^2 be defined as in (3.2.5). Further justification for this is given by Martin (1979) with reference to the contaminated normal density. If $p_\nu(u_n) = CN[\epsilon; 1, \sigma^2] = (1-\epsilon)\phi(u_n) + \epsilon\sigma^{-1}\phi(\sigma^{-1}u_n)$; then the marginal density is of the same form

$$u_n \sim CN[\epsilon; 1+q_n^2, \sigma^2+q_n^2] \ .$$

Setting $\sigma_0^2 = 1+q_n^2$ and $\sigma_1^2 = \sigma^2+q_n^2$, Martin shows that the marginal score g can be written as

$$g(u_n) = \frac{u_n}{\sigma_0^2} \ . \left[1 - b(u_n).\{1-\sigma_0^2/\sigma_1^2\}\right]$$

where

$$b(u_n)^{-1} = 1 + (1-\epsilon).\epsilon^{-1}.(\sigma_0^2/\sigma_1^2).\exp\left[-\frac{u_n^2}{2}(\sigma_0^{-2}-\sigma_1^{-2})\right] .$$

For $\sigma^2 >> 1$, we have $g(u_n) \simeq \sigma_0^{-1} g_\nu(\sigma_0^{-1} u_n)$.

Thus the approximation used in this work is as follows:

$$E[\underset{\sim}{\theta}_n | D_n] \simeq \underset{\sim}{m}_n = \underset{\sim}{a}_n + P_{n-1}\underset{\sim}{h}_n . \sigma_n^{-1} . g_\nu(\sigma_n^{-1} u_n), \tag{3.2.6}$$

and

$$\text{var}[\underset{\sim}{\theta}_n | D_n] \simeq C_n = P_n - P_{n-1}\underset{\sim}{h}_n . \underset{\sim}{h}_n^T P_{n-1} . \sigma_n^{-2} . G_\nu(\sigma_n^{-1} u_n). \tag{3.2.7}$$

Using this scheme and assuming approximate normality of $p(\underset{\sim}{\theta}_n|D_n)$ at each stage, extremely encouraging results are reported by Masreliez and Martin for problems involving scalar $\underset{\sim}{\theta}_n$, as mentioned at the end of the last section. Further uses of this approach are discussed by Martin (1979) and (1980), and Kleiner et al (1979), as part of a larger study of robust estimation of power spectra. From our point of view an attractive feature of this scheme is that predictive distributions, (both for comparison of alternative model structures/error distributions and for forecasting), are available directly. A further positive connection between this approach and the Bayesian analysis is a coincidental result applicable when the filters are based on Huber densities. In this special case, we can show by reference to example 2.4.2 of §2.4, that the equation (3.2.6) actually defines the posterior mode, and thus provides an optimal point estimate at each stage. Furthermore in this case we can rewrite (3.2.7) as

$$C_n^{-1} = P_n^{-1} + \underset{\sim}{h}_n \underset{\sim}{h}_n^T . G_\nu(\sigma_n^{-1} u_n)$$

due to the special form of G_ν, and so the information matrix is being used as a proxy for the "precision" matrix from a normal likelihood.

Of course this is not true for general p_ν, and there are several apparent disadvantages of this scheme to be noted.

(i) The approximation to posterior moments obtained by this scaling of the likelihood to obtain an approximate marginal density is based on heuristic considerations alone. Are there any more formal approaches to the problem?

(ii) In the case of the Huber likelihood, the filtering algorithms are not smooth functions of the observations, as they should

be (the exact mean and covariance matrix are). Indeed, the information function G_ν is zero outside the central normal part of the likelihood, implying that $C_n = P_n$ there, whereas in the central part C_n is the normal theory value. On the boundary there is a discontinuity. The true posterior covariance is a continuous function of y_n and this discontinuity of C_n is not desirable. Can this be remedied?

(iii) Following (ii) what happens in the extreme case of a double exponential likelihood, where G_ν is zero almost everywhere?

(iv) In certain cases, C_n may not be positive definite! All we require for this to happen is that $\sigma_n^{-2} G_\nu(\sigma_n^{-1} u_n)$ be large enough that C_n be negative definite as follows: $C_n > 0$ if and only if

$$I - \underset{\sim}{h}_n \underset{\sim}{h}_n^T . P_n . \sigma_n^{-2} G_\nu(\sigma_n^{-1} u_n) > 0.$$

Now if $G_\nu(\sigma_n^{-1} u_n) = 0$ then $C_n = P_n > 0$. Otherwise, for positive definite C_n we require, from the above inequality, that

$$G_\nu(\sigma_n^{-1} u_n) < 1 + q_n^{-2}, \quad \text{where } q_n^2 = \underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n .$$

It is clear that this may not be satisfied, in particular if q_n is large. For example, when p_ν is Student t with k degrees of freedom the maximum value of G_ν is $1+k^{-1}$, thus restricting q_n to be no greater than k.

This is clearly undesirable.

(v) Further undesirable behaviour of G_ν occurs with the exponential power family where G_ν does not exist at the

origin and tends to infinity as u_n tends to zero.

(ii) How can we justify further the normal approximation to $p(\underset{\sim}{\theta}_n|D_n)$?

We now consider alternative approaches in order to try to avoid the above drawbacks. In the next section, we attempt to answer (i) above by an approximation to the Bayesian analysis.

3.2 The Gradient algorithm.

The discussion preceeding (3.2.8) of the special form of Masreliez and Martin's recursions in the case of a Huber likelihood are reminiscent of certain aspects of asymptotic distribution theory in that the mode and information matrix are used as mean and precision matrix of the normal approximation. We follow this further in this section.

Heuristically, if $p(\underset{\sim}{\theta}_n|D_{n-1})$ is concentrated about $\underset{\sim}{a}_n$, and if $p_\nu(y_n - \underset{\sim}{h}_n^T\underset{\sim}{\theta}_n)$ is approximately quadratic in $\underset{\sim}{\theta}_n$ in a neighbourhood of $\underset{\sim}{a}_n$, then, expanding the log likelihood as a function of $\underset{\sim}{\theta}_n$ about $\underset{\sim}{a}_n$ in a Taylor series we obtain

$$\ell n.p_\nu(y_n - \underset{\sim}{h}_n^T\underset{\sim}{\theta}_n) = \ell n.p_\nu(y_n - \underset{\sim}{h}_n^T\underset{\sim}{a}_n) + \underset{\sim}{\tilde{\theta}}_n^T.\underset{\sim}{h}_n.g_\nu(y_n - \underset{\sim}{h}_n^T\underset{\sim}{a}_n)$$

$$- \tfrac{1}{2}\,\underset{\sim}{\tilde{\theta}}_n^T\underset{\sim}{h}_n\underset{\sim}{h}_n^T\underset{\sim}{\tilde{\theta}}_n.G_\nu(y_n - \underset{\sim}{h}_n^T\underset{\sim}{a}_n) + r(y_n;\underset{\sim}{\theta}_n)$$

where $r(y_n;\underset{\sim}{\theta}_n) = O(||\underset{\sim}{\tilde{\theta}}_n||^2)$, and $\underset{\sim}{\tilde{\theta}}_n = \underset{\sim}{\theta}_n - \underset{\sim}{a}_n$.

If we ignore the remainder term $r(y_n;\theta_n)$, we have an approximate log posterior given by

$$\ell n.p(\underset{\sim}{\theta}_n|D_n) \simeq \text{constant} - \tfrac{1}{2}\,\underset{\sim}{\tilde{\theta}}_n^T P_n^{-1}\underset{\sim}{\tilde{\theta}}_n + \underset{\sim}{\tilde{\theta}}_n^T.\underset{\sim}{h}_n.g_\nu(u_n)$$

$$- \tfrac{1}{2}\,\underset{\sim}{\tilde{\theta}}_n^T\underset{\sim}{h}_n\underset{\sim}{h}_n^T\underset{\sim}{\theta}_n.G_\nu(u_n)$$

where $u_n = y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n$.

So, if we define $\overline{C}_n = P_n^{-1} + \underset{\sim}{h}_n \underset{\sim}{h}_n^T G_\nu(u_n)$, then

$$\ell n\ p(\underset{\sim}{\theta}_n | D_n) \simeq \text{constant} + \underset{\sim}{\theta}_n^T \underset{\sim}{h}_n . g_\nu(u_n) - \tfrac{1}{2}\ \underset{\sim}{\theta}_n^T\ \overline{C}_n\ \underset{\sim}{\theta}_n .$$

Now if $\overline{C}_n$ is non-singular, define $C_n = \overline{C}_n^{-1}$ i.e. $\overline{C}_n = C_n^{-1}$, and we have

$$(\underset{\sim}{\theta}_n | D_n) \sim N[\underset{\sim}{m}_n, C_n]$$

where
$$\underset{\sim}{m}_n = \underset{\sim}{a}_n + C_n \underset{\sim}{h}_n . g_\nu(u_n), \qquad (3.2.8),$$

and
$$C_n^{-1} = P_n^{-1} + \underset{\sim}{h}_n \underset{\sim}{h}_n^T . G_\nu(u_n). \qquad (3.2.9).$$

Note that when G_ν is zero, then $C_n = P_n$. Otherwise we can rewrite (3.2.8) as

$$\underset{\sim}{m}_n = \underset{\sim}{a}_n + P_n \underset{\sim}{h}_n . [\underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n . G_\nu(u_n) + 1]^{-1} g_\nu(u_n), \qquad (3.2.10)$$

and (3.2.9) as

$$C_n = P_n - P_n \underset{\sim}{h}_n . [\underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n . G_\nu(u_n) + 1]^{-1}\ \underset{\sim}{h}_n^T P_n G_\nu(u_n) \qquad (3.2.11)$$

and from these two equations we can calculate $\underset{\sim}{m}_n$ and C_n without inverting the matrix C_n^{-1}.

The recursions (3.2.10), (3.2.11) are similar to the updating algorithms for posterior mode and covariance matrix in asymptotic Bayesian theory and as such are essentially a one-step version of a stochastic gradient (or Newton-Raphson) type algorithm. To see this note that $\underset{\sim}{\theta}_n^*$ the posterior mode satisfies

$$\underset{\sim}{f}(\underset{\sim}{\theta}_n^*) = 0$$

where

$$\underset{\sim}{f}(\underset{\sim}{\theta}_n) = P_n^{-1}(\underset{\sim}{\theta}_n - \underset{\sim}{a}_n) - \underset{\sim}{h}_n . g_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n) \qquad (3.2.12)$$

$\underset{\sim}{m}_n$ is then a first step in the iteration of

$$\underset{\sim}{\theta}_n^i = \underset{\sim}{\theta}_n^{i-1} - \left[\frac{\partial \underset{\sim}{f}^T(\underset{\sim}{\theta}_n^{i-1})}{\partial \underset{\sim}{\theta}_n^{i-1}}\right].\underset{\sim}{f}(\underset{\sim}{\theta}_n^{i-1}), \quad i=1,2,\ldots$$

with $\underset{\sim}{\theta}^0 = \underset{\sim}{a}_n$ as starting point. Having noted this, it is clear that the recursions are really useful only when P_n is small. However, algorithms of this type have been used with some success. For some example, Martin and Masreliez (1975) propose a similar algorithm in the simple location problem, when $\underset{\sim}{\theta}_n$ = constant for all n. Polyak and Tsyphin (1980) discuss related algorithms with G_ν replaced by a constant and in particular by the expected information matrix at $\underset{\sim}{a}_n$ rather than the observed matrix. These refinements are geared towards producing asymptotically efficient estimation algorithms for fixed parameters $\underset{\sim}{\theta}_n$ and we discuss such aspects in detail in Chapter 6. At the moment we are interested in approximating finite sample posterior means and covariance matrices in time series, and such asymptotic considerations are spurious.

One further related work is that of Vere Jones (1975) in an entirely different context. He proposes algorithms of a stochastic process for forecasting various point processes. Again Vere Jones uses non-normal processes in general.

Notice that when p_ν is normal, the recursions above reduce to the Kalman filter. It might be hoped that for p_ν near to normality, for example with a Student t distribution with a large degree of freedom parameter, the algorithms above will behave like the Kalman filter for "good" data whilst retaining the attractive outlier resistant properties of a t distribution based analysis. Indeed this is the case and is illustrated at the end of this Chapter in numerical examples, with a modified algorithm introduced in the next section. A further positive remark about the gradient algorithm is that it does not conflict with Masreliezs' theorem,

as follows:

If we adopt $\underset{\sim}{m}_n$ defined by (3.2.10), then, using Masreliezs' result,

$$g(u_n) \simeq (q_n^2 G_\nu(u_n)+1)^{-1} g_\nu(u_n)$$

where $q_n^2 = \underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n$.

Thus

$$G(u_n) = \frac{\partial g(u_n)}{\partial u_n} \simeq (q_n^2 G_\nu(u_n)+1)^{-1} G_\nu(u_n) + t_n$$

where $t_n = (q_n^2 G_\nu(u_n)+1)^{-2} D_\nu(u_n) q_n^2$

and

$$D_\nu(u_n) = -\frac{\partial G}{\partial u_n}(u_n).$$

The second order approximation to the log likelihood assumes that D_ν is zero. Thus, from (3.2.2), using G above we have

$$C_n = P_n - P_n \underset{\sim}{h}_n \underset{\sim}{h}_n^T P_n . (q_n^2 G_\nu(u_n)+1)^{-1} G_\nu(u_n)$$

which is just the value given in (3.2.11), so the gradient approximation is consistent with Masrelieq's link between $\underset{\sim}{m}_n$ and C_n.

However, there are several serious drawbacks to the gradient algorithm, all of which are shared by the algorithm of §3.2.2. We note the following:

(i) Again the mean $\underset{\sim}{m}_n$ suffers the problem of having discontinuities for some p_ν. In particular, the discontinuity in G_ν for a Huber likelihood now appears in the mean $\underset{\sim}{m}_n$ as well as in C_n.

(ii) The algorithms cannot be based on the double-exponential distribution since G is zero almost everywhere and thus

C_n as defined will not reproduce the behaviour of the posterior covariance matrix.

(iii) In some cases, for example that of the Student t density, G_ν can take negative values. This means that $C_n > P_n$ which is as it should be for such observations as discussed in Chapter 2. However it may be that C_n as defined above will not be positive definite. Clearly this behaviour is unreasonable.

As mentioned above, we discuss theoretical aspects of this algorithm in Chapter 6. The practical problems are of interest here and we introduce a new approach which avoids all the above drawbacks and provides what we believe to be the most useful algorithm for this model. It has considerable intuitive appeal and also a very strong theoretical basis.

3.2.4. The modal recursions.

To motivate what follows consider the modal equation (3.2.12). The gradient algorithm was derived as a second order approximation to the solution of (3.2.12). The recursions of this section are derived as, essentially, a first order approximation as follows:

Since we assume that p_ν is symmetric we have $p_\nu(u)$ as a function of u^2, say $f(u^2/2)$ where f is a positive function on $[0,\infty)$. Unimodality of p_ν implies f decreases on $[0,\infty)$. Clearly then, the score g_ν is defined by

$$g_\nu(u) = -\frac{\partial}{\partial u} \ell n\ p_\nu(u) = -f'(u^2/2).u/f(u^2/2).$$

$$= \psi(u).u, \text{ say}$$

where $\psi_\nu(u) = -f'(u^2/2)/f(u^2/2)$.

The following properties of ψ_ν are immediate:

(i) $\phi_\nu(u)$ is positive for all real u, since f, and hence ℓn f, is a decreasing function of u^2.

(ii) $\psi_\nu(u)$, being a function of u^2, is symmetric about zero.

(iii) $\psi_\nu(u)$ is differentiable since $p_u(u)$ is assumed twice differentiable.

Now, we can write the modal equation (3.2.12) in the form

$$P_n^{-1}(\underset{\sim}{\theta}_n^* - \underset{\sim}{a}_n) - \underset{\sim}{h}_n \cdot \psi_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*) \cdot (y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*) = \underset{\sim}{0}$$

or $$\underset{\sim}{\theta}_n^* = \left[P_n^{-1} + \underset{\sim}{h}_n \underset{\sim}{h}_n^T \psi_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*)\right]^{-1} \cdot \left[P_n^{-1} \underset{\sim}{a}_n + \underset{\sim}{h}_n \cdot \psi_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*) y_n\right]$$

since $P_n^{-1} + \underset{\sim}{h}_n \underset{\sim}{h}_n^T \psi_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*)$ is nonsingular by virtue of the positivity of ψ_ν. Rearranging this expression we obtain

$$\underset{\sim}{\theta}_n^* = \underset{\sim}{a}_n + \left[P_n^{-1} + \underset{\sim}{h}_n \underset{\sim}{h}_n^T \psi_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*)\right]^{-1} \underset{\sim}{h}_n \psi_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*) \cdot (y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n)$$

$$= \underset{\sim}{a}_n + P \underset{n\sim}{h}_n \cdot (1 + \underset{\sim}{h}_n^T P \underset{n\sim}{h}_n \psi_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*))^{-1} \psi_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n^*) \cdot (y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n). \quad (3.2.13)$$

Now (3.2.13) can be used to calculate $\underset{\sim}{\theta}_n^*$ iteratively, substituting $\underset{\sim}{\theta}_n^{i-1}$ into the right hand side and calculating the left hand side as $\underset{\sim}{\theta}_n^i$. If we begin with $\underset{\sim}{\theta}_n^0 = \underset{\sim}{a}_n$ and approximate the solution by the one-step estimate $\underset{\sim}{\theta}_n^1 = \underset{\sim}{m}_n$, we obtain

$$\underset{\sim}{m}_n = \underset{\sim}{a}_n + P \underset{n\sim}{h}_n \cdot \left(1 + \underset{\sim}{h}_n^T P \underset{n\sim}{h}_n \psi_\nu(u_n)\right)^{-1} g_\nu(u_n) \qquad (3.2.14)$$

as our "modal recursion".

In order to calculate the corresponding covariance matrix we use Masreliezs' result as follows:- identifying the marginal score $g(u_n)$ with $\left(1+q_n^2\, \psi_\nu(u_n)\right)^{-1} g_\nu(u_n)$, $q_n^2 = \underset{\sim}{h}_n^T P \underset{n\sim}{h}_n$, we only need to differentiate with respect to u_n to find

Lemma 3.2.1.

If $g(u_n) = \left(1+q_n^2\ \psi_\nu(u_n)\right)^{-1} g_u(u_n)$

then defining $G(u_n) = \frac{\partial g(u_n)}{\partial u_n}$, we have

$$G(u_n) = \psi_\nu(u_n).\left(1+q_n{}^2\psi_\nu(u_n)\right)^{-1} + u_n\dot{\psi}_\nu(u_n).\left(1+q_n{}^2\psi_\nu(u_n)\right)^{-2}$$

and $P_n - P_{n\sim n\sim n}h\ h^T.P_n.G(u_n)$ is <u>always</u> positive definite when $\psi_\nu(u)$ is non-increasing for u>0.

Proof:

$$G(u_n) = \left(1+q_n^2\psi_\nu(u_n)\right)^{-1}G_\nu(u_n) - \left(1+q_n^2\psi_\nu(u_n)\right)^{-2}g_\nu(u_n).q_n^2\dot{\psi}_\nu(u_n).$$

Since $g_\nu(u_n) = \psi_\nu(u_n).u_n$, then $G_\nu(u_n) = \psi_\nu(u_n) + \dot{\psi}_\nu(u_n).u_n$

and so

$$G(u_n) = \left(1+q_n^2\psi_\nu(u_n)\right)^{-2}\{\psi_\nu(u_n) + \dot{\psi}_\nu(u_n)u_n + q_n^2\ \psi_\nu{}^2(u_n)$$

$$+ q_n^2\ \psi_\nu(u_n)\dot{\psi}_\nu(u_n).u_n - g_\nu(u_n)q_n^2\dot{\psi}_\nu(u_n)\}.$$

$$= \left(1+q_n^2\ \psi_\nu(u_n)\right)^{-1}\psi_\nu(u_n) + \left(1+q_n^2\psi_\nu(u_n)\right)^{-2}\dot{\psi}(u_n)u_n,$$

as required.

Further, since $\psi_\nu(u_n)$ is nonincreasing for u_n>0 and symmetric about zero, then

$$-\dot{\psi}(u_n).u_n \geq 0 \text{ always.}$$

Hence

$$C_n = P_n - P_{n\sim n\sim n}h\ h^TP_n.\left(1+q_n^2\psi_\nu(u_n)\right)^{-1}\psi_\nu(u_n) + P_{n\sim n\sim n}h\ h^TP_n.\phi_n(u_n)$$

$$= \left[P_n+h_{\sim n\sim n}h^T.\psi_\nu(u_n)\right]^{-1} + P_{n\sim n\sim n}h\ h^TP_n\phi_n(u_n)$$

where $\phi_n(u_n) = -\left(1+q_n^2\psi_\nu(u_n)\right)^{-2}\dot{\psi}_\nu(u_n)u_n \geq 0.$

So C_n is always positive definite, and further, is always greater than $\left[P_n+h_{\sim n\sim n}h^T\psi_\nu(u_n)\right]^{-1}$ for non-zero u_n. □

The modal recursions are now

$$\underset{\sim}{m}_n = \underset{\sim}{a}_n + P_n \underset{\sim}{h}_n \left(1+q_n^2 \psi_\nu(u_n)\right)^{-1} g_\nu(u_n) \qquad (3.2.15)$$

and

$$C_n = P_n - P_n \underset{\sim}{h}_n \underset{\sim}{h}_n^T P_n \; G(u_n) \qquad (3.2.16)$$

with $G(u_n)$ as in the above Lemma.

We note that

(i) The Kalman filter obtains when p_ν is normal.

(ii) Otherwise $\underset{\sim}{m}_n$ looks like the posterior mean for the normal model with a variance

$$\text{var}[u_n] = \psi_\nu^{-1}(u_n)$$

depending upon y_n.

(iii) $C_n = C_n^* + R_n$ where C_n^* is the posterior covariance matrix from the normal model with variance $\psi_\nu^{-1}(u_n)$ as above. The R_n term corrects the posterior variance and can be thought of as an adjustment reflecting uncertainty about the "estimate" $\psi_\nu^{-1}(u_n)$ of the error variance. This adjustment will be positive i.e. an increased covariance matrix, when by using Lemma 3.2.1, $\psi_\nu(u_n)$ is decreasing for positive u_n. This is true for <u>all</u> the densities of Chapter 2 with the exception of mixtures of the form

$$(1-\epsilon)p_1(u) + \epsilon p_2(u)$$

(and with more components); in particular the contaminated normal densities. However for such mixtures we apply the recursions separately to each component of the mixture and then C_n is always positive definite, as follows:-

For a k-component mixture,

$$p_u(u) = \sum_{j=1}^{k} p_j(u),$$

then

$$p(\underset{\sim}{\theta}_n|D_n) \propto \sum_{j=1}^{k} p(\underset{\sim}{\theta}_n|D_{n-1}).p_j(y_n - \underset{\sim}{h}_n^T\underset{\sim}{\theta}_n).\pi_j$$

$$\propto \sum_{j=1}^{k} p(\underset{\sim}{\theta}_n|D_{n-1},p_j).\pi_j^*$$

where $\pi_j^* \propto \int_{\mathbb{R}^p} p(\underset{\sim}{\theta}_n|D_{n-1}).p_j(y_n - \underset{\sim}{h}_n^T\underset{\sim}{\theta}_n)\, d\underset{\sim}{\theta}_n.\pi_j$, and $\sum_{j=1}^{k} \pi_j^* = 1$.

Thus using the modal recursions on each component $p(\underset{\sim}{\theta}_n|D_{n-1},p_j)$, we obtain values $\underset{\sim}{m}_{nj}$ and C_{nj}, and

$$\underset{\sim}{m}_n = E[\underset{\sim}{\theta}_n|D_n] \simeq \sum_{j=1}^{k} \pi_j^*.\underset{\sim}{m}_{nj},$$

$$C_n = \text{var}[\underset{\sim}{\theta}_n|D_n] \simeq \sum_{j=1}^{k} \pi_j^*\left[C_{nj} + (\underset{\sim}{m}_{nj} - \underset{\sim}{m}_n)(\underset{\sim}{m}_{nj} - \underset{\sim}{m}_n)^T\right]$$

Clearly using this approach to mixtures, the special case of normal components will result in the true mean and covariance matrix being calculated via the modal algorithms.

Returning to the criticisms of Masreliez and Martin's algorithms and the gradient algorithms we shall show in some examples that the problems of smoothness of $\underset{\sim}{m}_n$ and C_n have been solved by the modal algorithm. The final objection, that of assuming approximate posterior normality, is the subject of section 3.2.5. Following the examples, some figures are provided illustrating the normal approximation to the posterior for a scalar parameter θ_n. For several different likelihoods p, we plot the exact posterior density

$$p(\theta|y) \propto c^{-\frac{1}{2}}\phi(c^{-\frac{1}{2}}\theta).p(y-\theta)$$

for three values of y, 1,3 and 5. These appear as the full lines. The <u>modal</u> approximation as derived in this section appears as the

dashed line. Each page has three figures for the three values of the prior variance c = 1,3 and 5.

Examples 3.2.1.

(a) At normality we obtain the Kalman Filter with $\psi_\nu(u_n) = (1+q_n^2)^{-1}$.

(b) Exponential power family of index β, $0<\beta\leq 2$. Clearly $\psi_\nu(u) = \beta.|u|^{\beta-2}$, $u \neq 0$, and so the approximate marginal score is

$$g(u_n) = (q_n^2.\beta + |u_n|^{2-\beta})^{-1}\beta u_n$$

and information

$$G(u_n) = \beta\left[(\beta-1)|u_n|^{2-\beta}+q_n^2\beta\right]\left[q_n^2\beta + |u_n|^{2-\beta}\right]^{-2}.$$

In particular for $\beta=1$, the double exponential density, then

$$g(u_n) = (q_n^2 + |u_n|)^{-1}u_n$$

and $$G(u_n) = q_n^2.(q_n^2 + |u_n|)^{-2}.$$

(c) Student t with k degrees of freedom. Here we have

$$g(u_n) = (k+1)u_n/(k\sigma_n^2+u_n),$$

where

$$\sigma_n^2 = 1 + q_n^2(1+k^{-1}).$$

So the marginal score is approximately given by that of the likelihood with a different scale factor. In this case the modal recursions are like those of Masreliez and Martin although the definition of the scale factors differ.

(d) Huber (k).

The marginal score is $g(u_n) = \begin{cases} u_n.(1+q_n{}^2)^{-1}, & |u_n|\leq k, \\ ku_n.(|u_n|+q_n{}^2k)^{-1}. & \text{otherwise} \end{cases}$

and clearly g is continuous as a function of u_n.

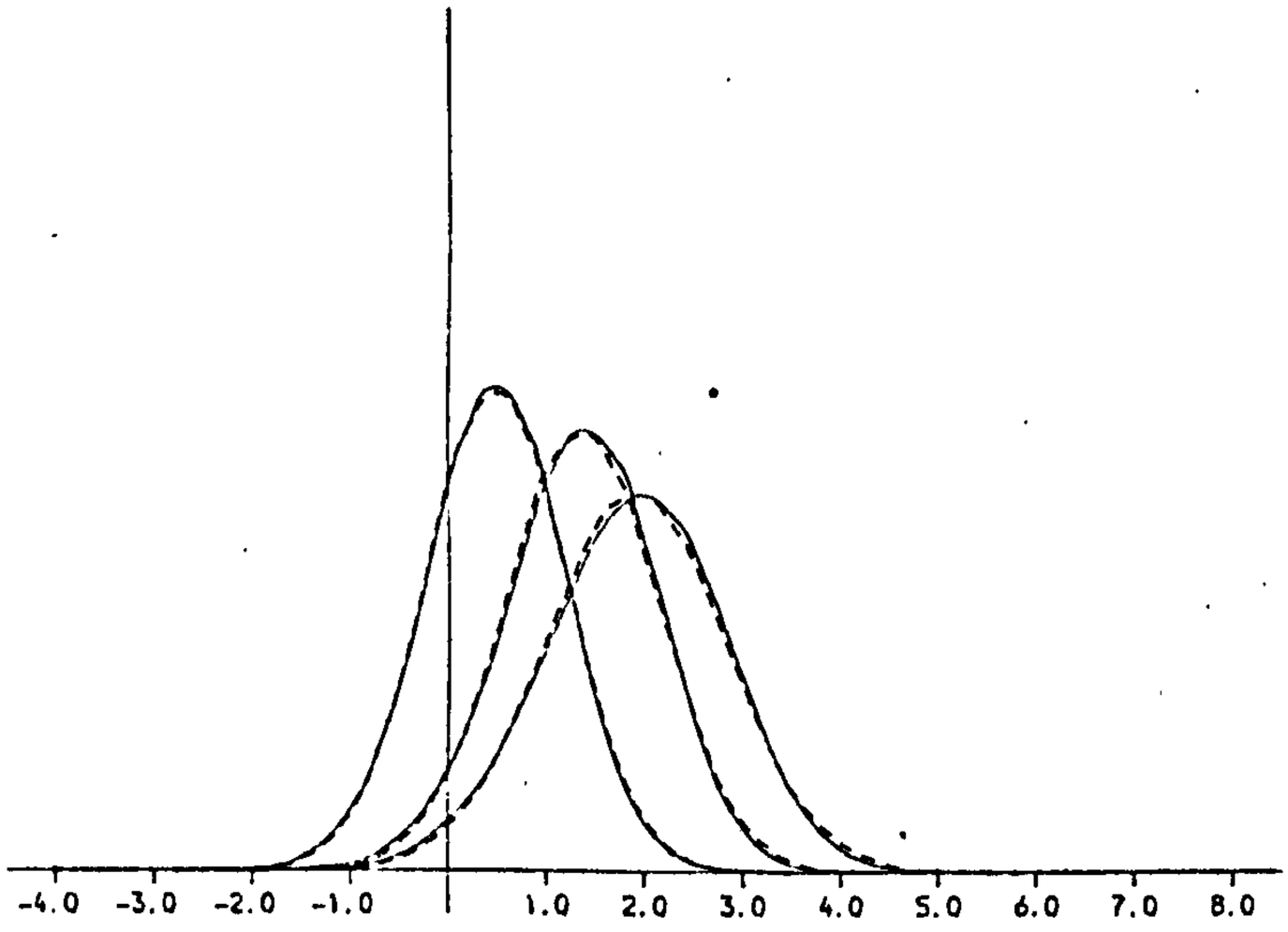
3.1(a)

STUDENT T-15

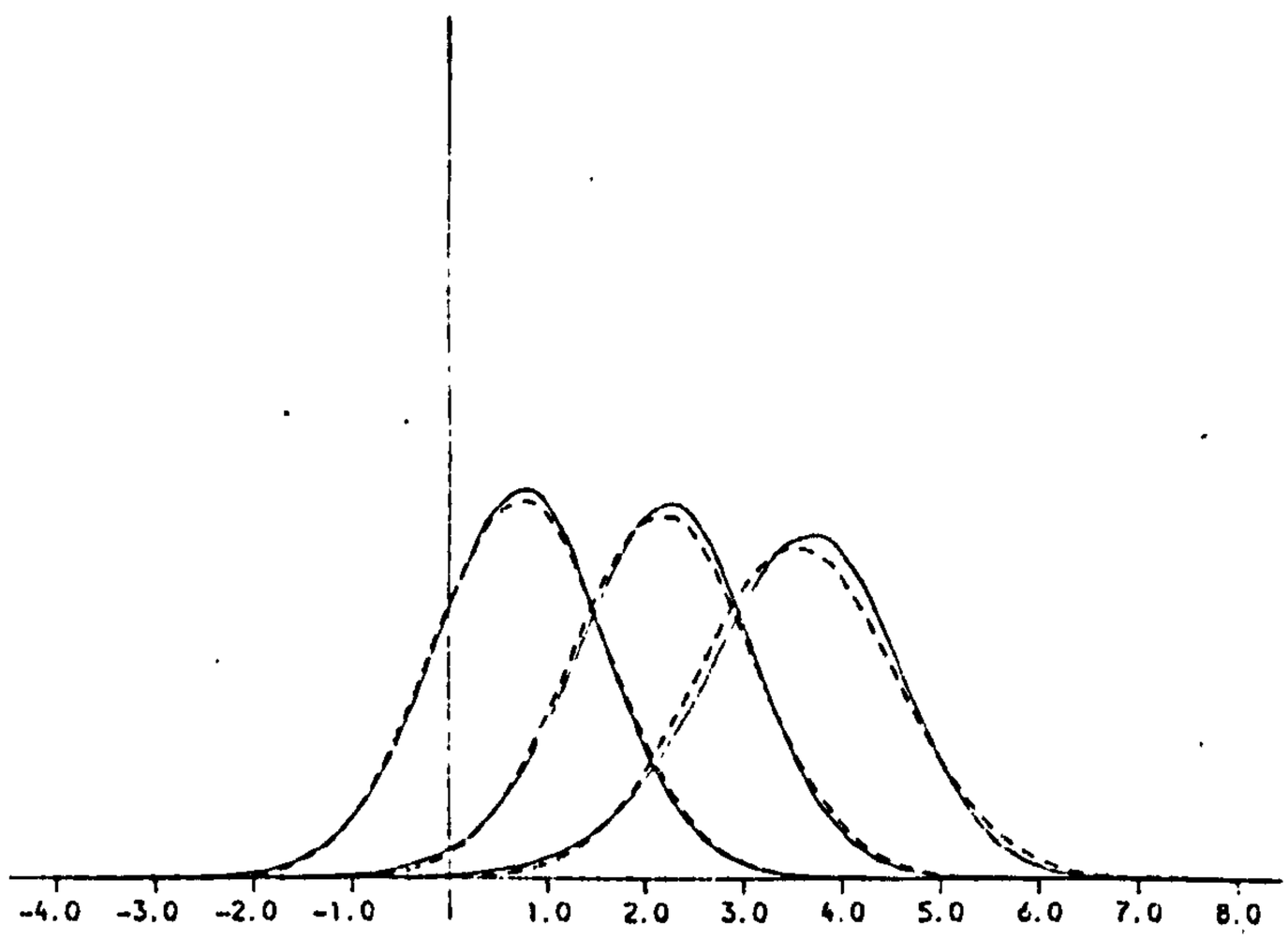
——— POSTERIOR

------ APPROXIMATION

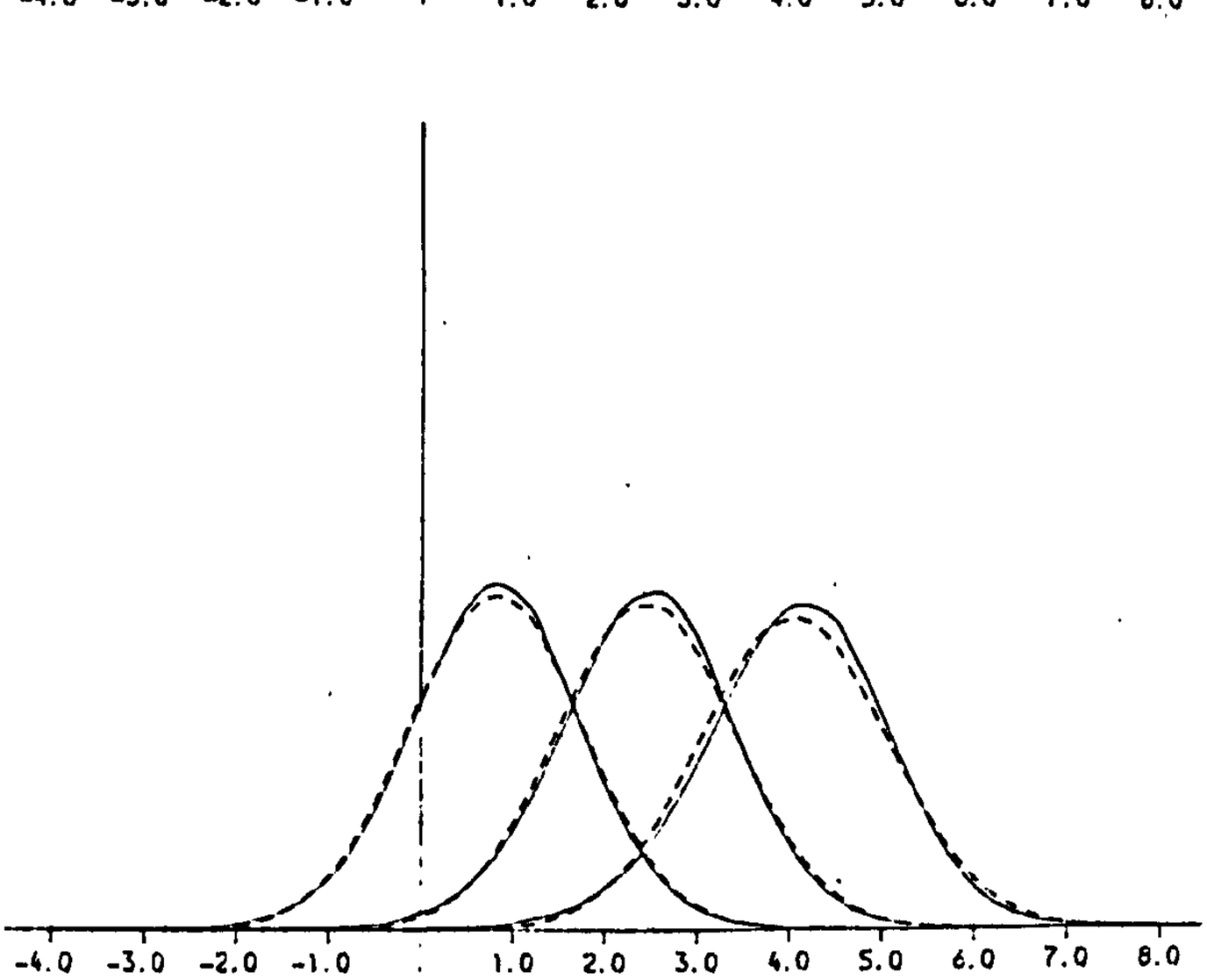
C=1



C=3



C=5

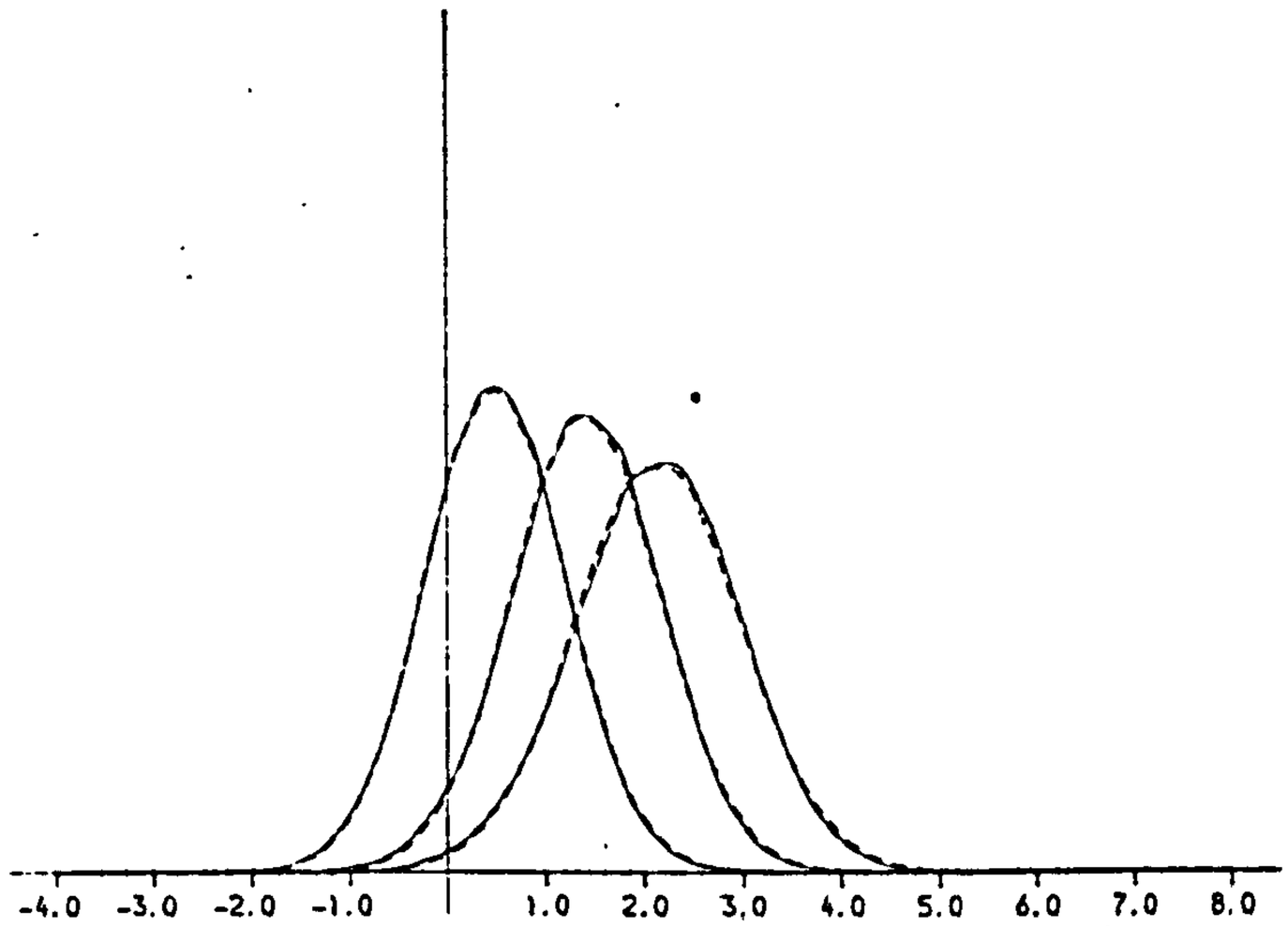


STUDENT T-25

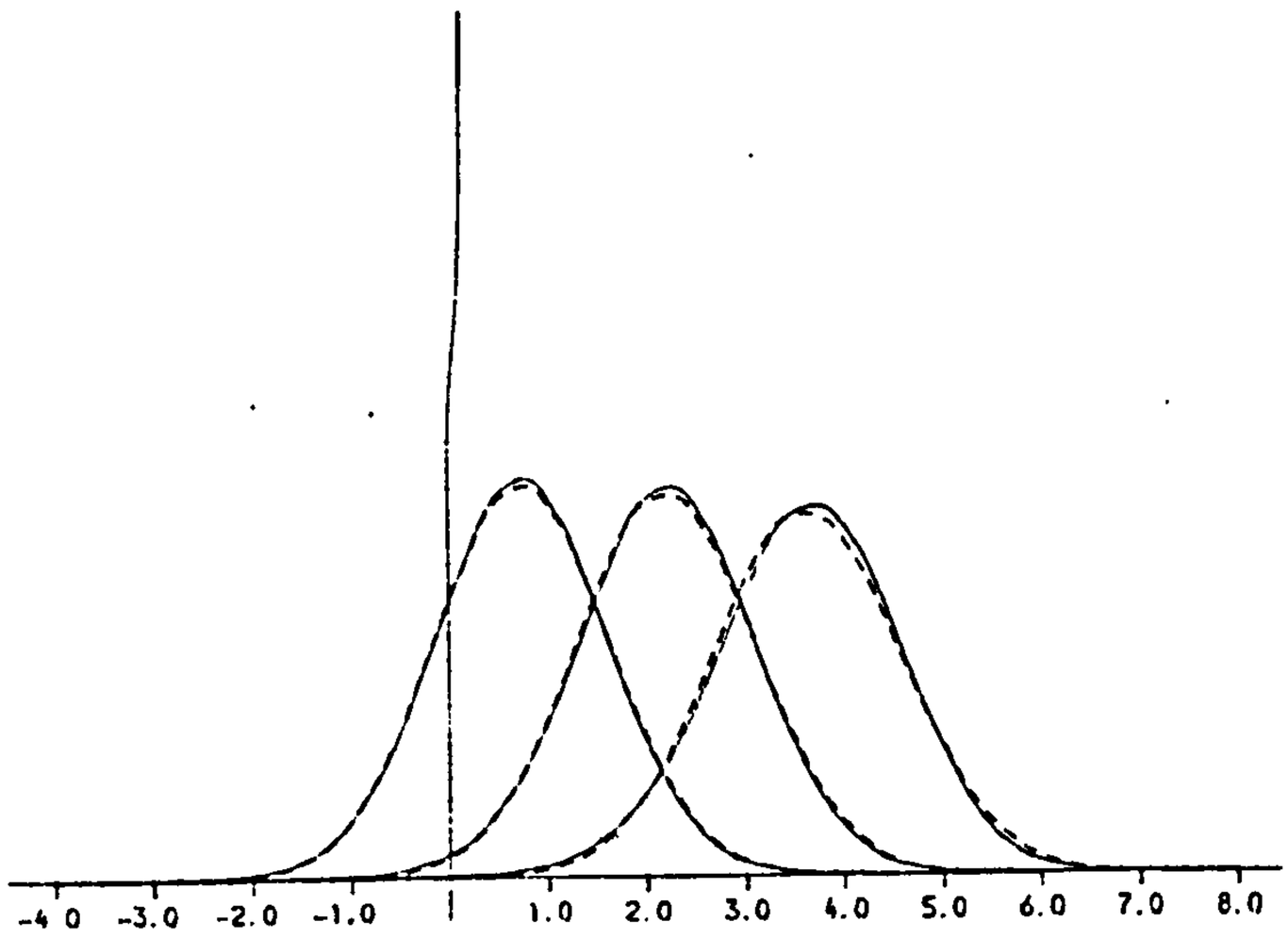
——— POSTERIOR

------ APPROXIMATION

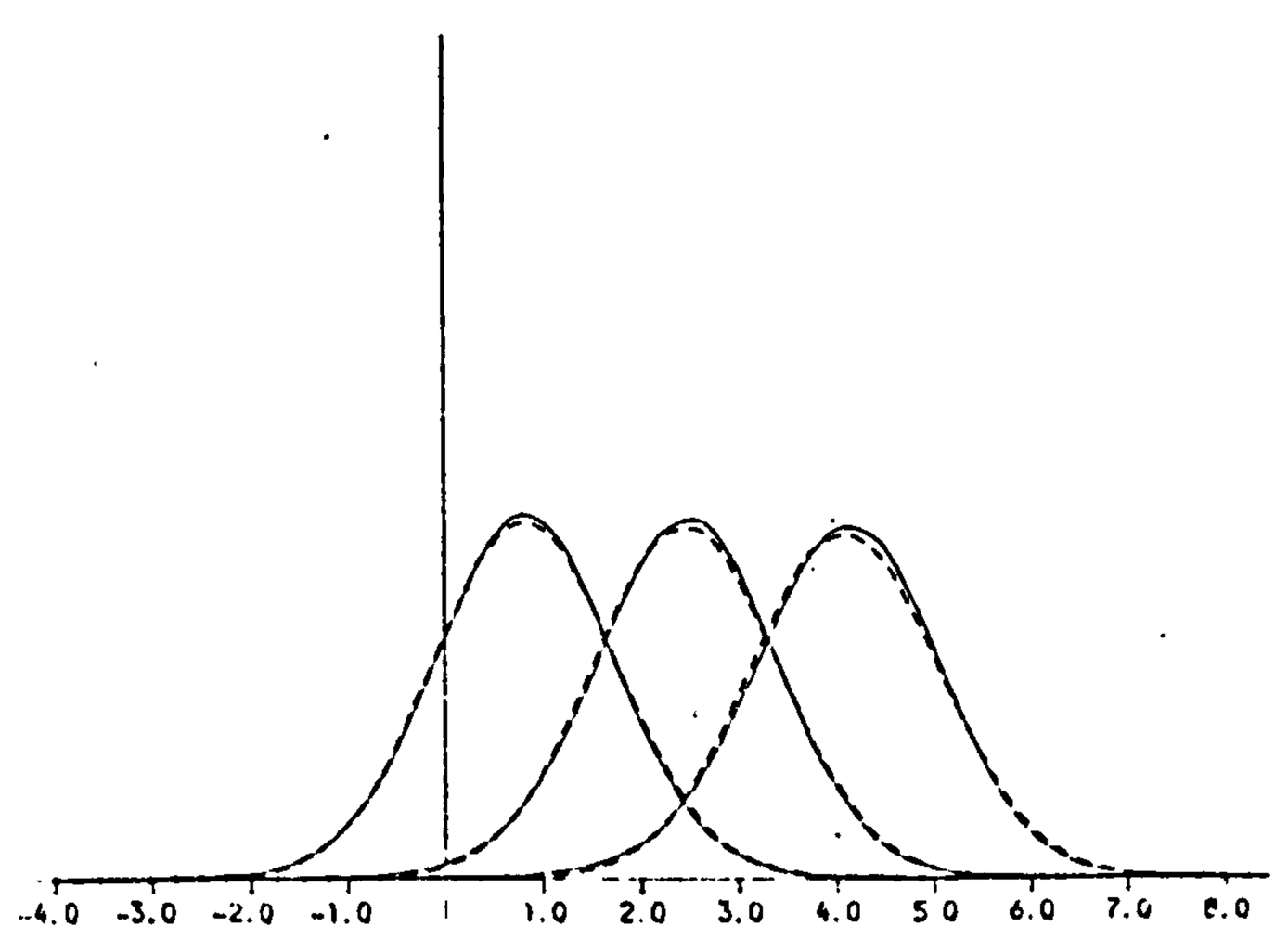
C=1



C=3



C=5



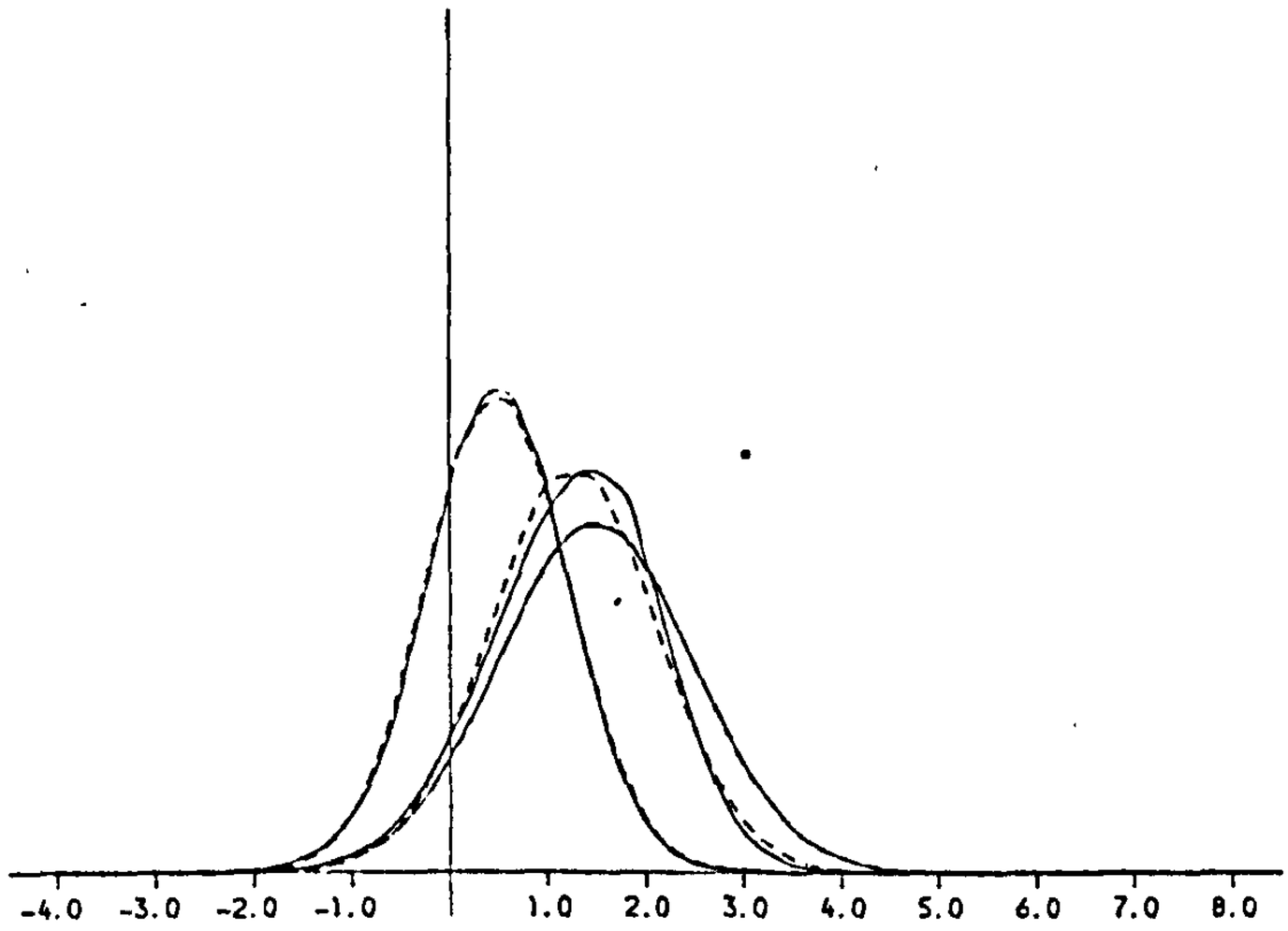
3.1(c)

HUBER - 1.5

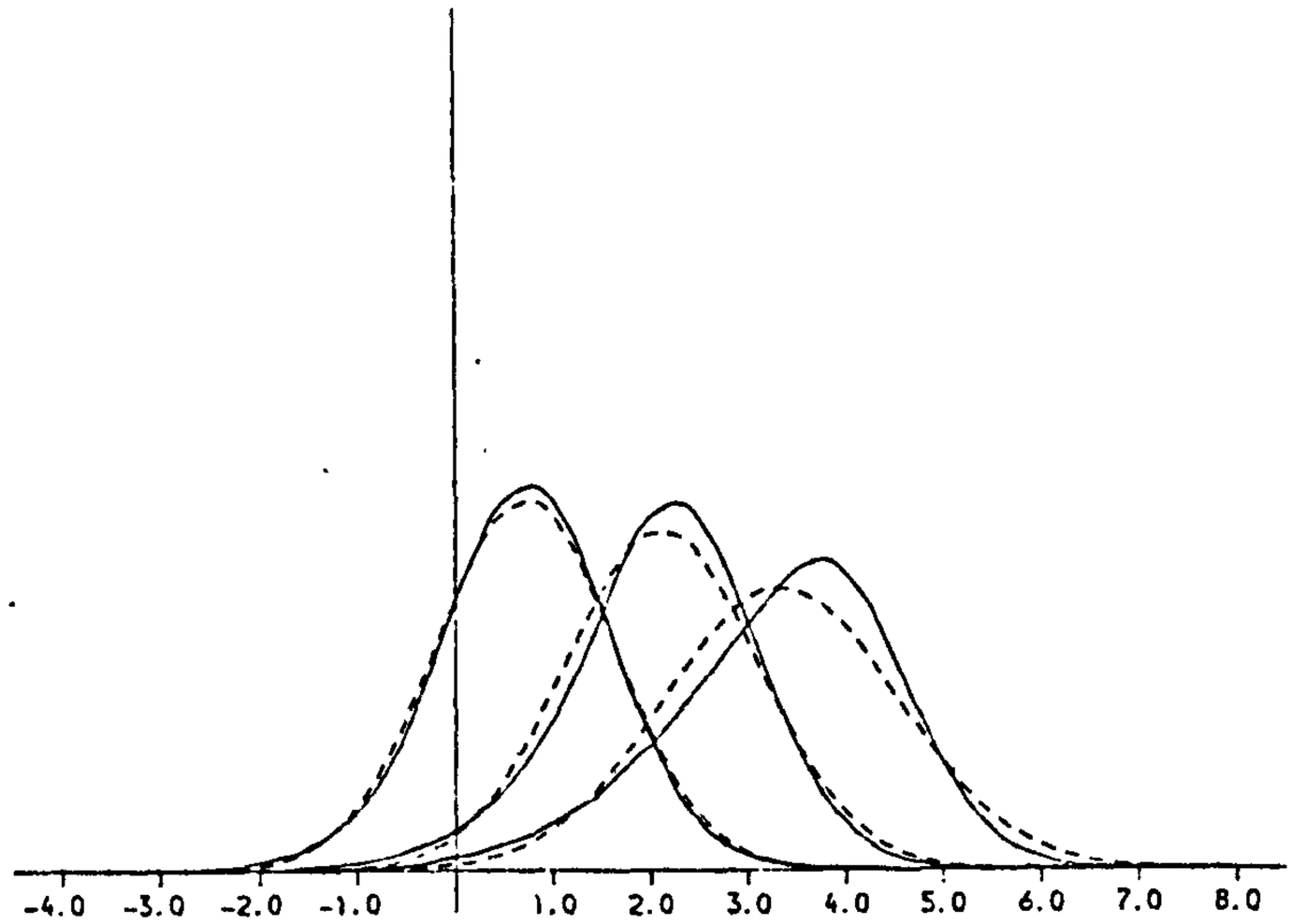
——— POSTERIOR

------ APPROXIMATION

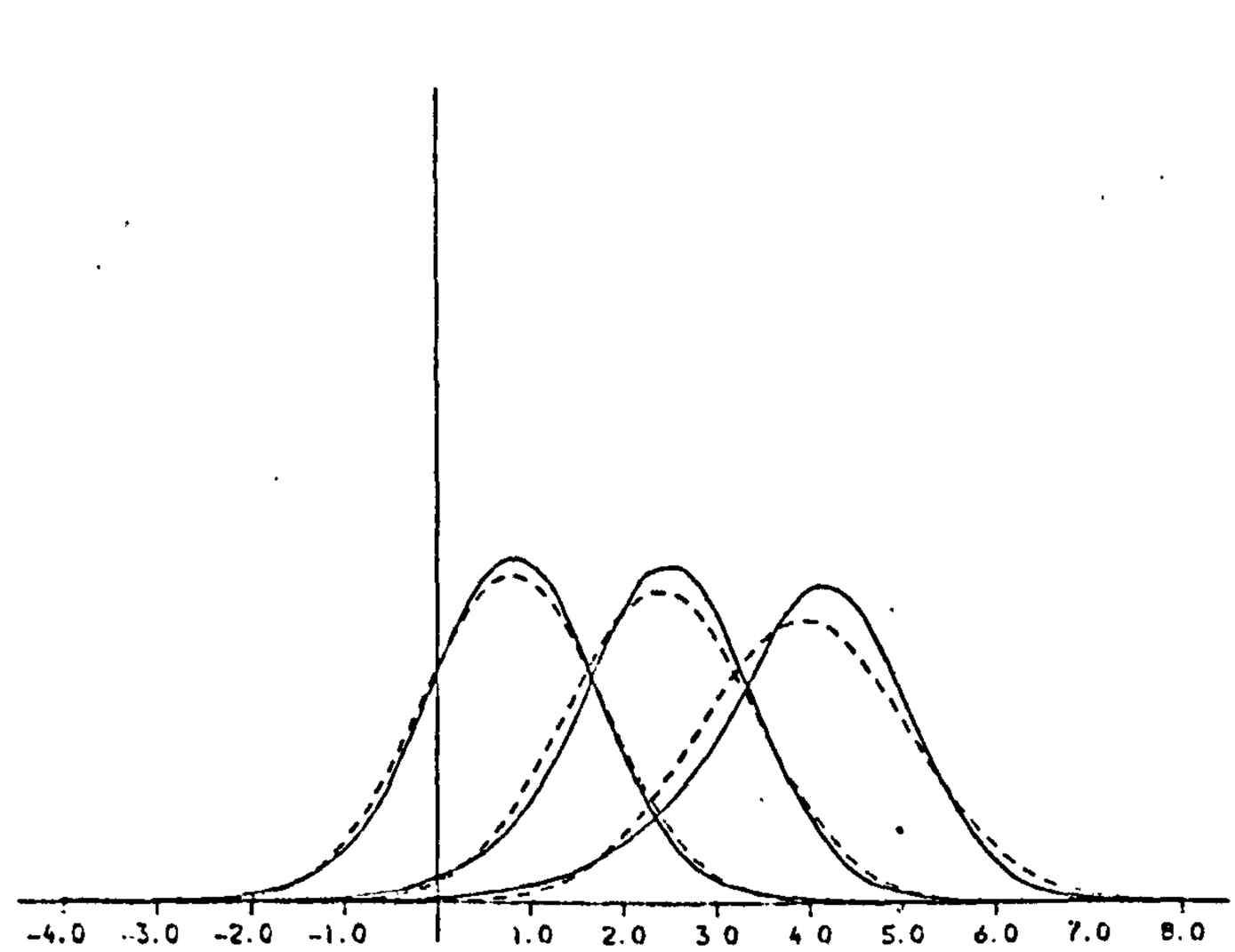
C=1



C=3



C=5



3.2.5. Scale mixtures of normal distributions.

The notes following the derivation of the modal recursions lead to the interpretation of the recursions being derived from the Kalman filter with a data-dependent "estimate" of the error variance used as a robustifying factor. We investigate this further in this section.

In Chapter 2 we discussed the family of scale mixtures of normal distributions and remarked that many of the symmetric distributions of interest are in this family. A general result about the representation of p_ν as

$$p_\nu(u) = \int_0^\infty N[0,t]\omega(t)dt \qquad (3.2.17)$$

was given by Chu (1968) as follows:

> If $p_\nu(u)$ is a function of $\frac{u^2}{2}$, then there exists a scalar function $\omega(t)$, $t\geq 0$, such that $p_\nu(u)$ can be expressed by (3.2.17) above.

Clearly the converse is trivial. Now the main condition for Chu's result is that, if $p_\nu(u) = f(u^2/2)$, then $f(s)$ vanishes faster than s^{-1} as $s\to\infty$. For details see Chu (1968). Clearly then, for all the symmetric distributions of Chapter 2 which decay at least as fast as the Cauchy, this result holds. However, there is no assurance that $\omega(t)$ as defined is a density function. Indeed there are heavy-tailed p_ν for which $\omega(t)$ is not even positive on $[0,\infty)$. As an example, take $p_\nu(u) \alpha [4+u^4]^{-1}$ i.e. $f(s) \alpha [1+s^2]^{-1}$. It is easily verified that the mixing function is $\omega(t) \alpha t^{-\frac{1}{2}}\sin(t/2)$.

Andrews & Mallows (1974) provide the conditions for $\omega(t)$ to be a density, as follows;

> if p_ν is symmetric about zero, then $\omega(t)$ is a density if and

only if

$$(-1)^k \frac{d^k}{dx^k} \cdot p_\nu(x^{\frac{1}{2}}) \geq 0, \quad \text{for } x>0 \text{ and } h = 1,2,3,\ldots .$$

Using this result we can verify the result of Appendix 2B, that the exponential power distributions are scale mixtures of normals when the index β lies in $[1,2]$. Our proof is, however, direct and provides the precise form of the mixing distribution.

Now for such distributions we have the decomposition of g_ν as $g_\nu(u) = \psi_\nu(u)u$ for all u. This representation of the score was actually an ingredient of the robust estimation of Ramsay and Novick (1980) in Bayesian regression problems, although this was not made explicit. They suggest robustifying standard analyses by multiplying the usual score functions, (in the normal case $g_\nu(u) \alpha u$), by a functions which limits the growth of g_ν and nullifies it asymptotically. In particular they replace u by a function proportional to $u.\exp\{-a|u|^{b/2}\}$, for some positive constants a and b; clearly this can be seen to be an embedding of the likelihood in the family of continuous scale mixtures with

$$\psi_\nu(s) = \exp\{-a|s|^{b/2}\}.$$

We shall now see how the modal recursions can be seen as an approximation to the analysis with a likelihood which is a scale mixture of normals.

Conditional analysis.

We work with the density written as

$$p_\nu(u) = \int_0^\infty \omega(\lambda).N[0,\lambda^{-1}]d\lambda.$$

So λ is the conditional precision parameter of the normal distribution of u given λ. In the model (3.2.1) and (3.2.2),

since the $\{v_n\}$ are independent, we have, for n = 1,2,..

$$(v_n|\lambda_n) \text{ are i.i.d. } N[0,\lambda_n^{-1}]$$

with $\{\lambda_1,\ldots,\lambda_n\}$ a sequence of independent, identically distributed random variables with density $\omega(\lambda)$.

Thus, given λ_n and $\underset{\sim}{\theta}_{n-1}|D_{n-1}$ as $N[\underset{\sim}{m}_{n-1},C_{n-1}]$, the usual normal theory leads to

$$(\underset{\sim}{\theta}_n|D_n,\lambda_n) \sim N[\underset{\sim}{m}_n(\lambda_n),C_n(\lambda_n)], \tag{3.2.18}$$

where

$$\underset{\sim}{m}_n(\lambda_n) = \underset{\sim}{a}_n + P_n\underset{\sim}{h}_n.(\lambda_n^{-1}+\underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n)^{-1}(y_n-\underset{\sim}{h}_n^T\underset{\sim}{a}_n), \tag{3.2.19}$$

and

$$C_n(\lambda_n) = P_n - P_n\underset{\sim}{h}_n(\lambda_n^{-1}+\underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n)^{-1}\underset{\sim}{h}_n^T P_n \tag{3.2.20}$$

with $\underset{\sim}{a}_n$, P_n as usual.

Therefore

$$(\theta_n|D_n) \sim \int_0^\infty N[\underset{\sim}{m}_n(\lambda_n),C_n(\lambda_n)]p(\lambda_n|D_n)d\lambda_n,$$

where

$$p(\lambda_n|D_n) \;\alpha\; \omega(\lambda_n)p(y_n|D_{n-1},\lambda_n)$$

and

$$(y_n|D_{n-1},\lambda_n) \sim N\left[\underset{\sim}{h}_n^T\underset{\sim}{a}_n,\lambda_n^{-1}+\underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n\right]. \tag{3.2.21}$$

In particular,

$$E[\underset{\sim}{\theta}_n|D_n] = \underset{\sim}{m}_n = \underset{\sim}{a}_n + P_n\underset{\sim}{h}_n.E[\lambda_n(1+\lambda_n q_n^2)^{-1}|D_n].(y_n-\underset{\sim}{h}_n^T\underset{\sim}{a}_n) \tag{3.2.22}$$

where $q_n^2 = \underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n$,

and $\text{var}[\underset{\sim}{\theta}_n|D_n] = C_n = P_n - P_n\underset{\sim}{h}_n.\underset{\sim}{h}_n^T.P_n.E[\lambda_n(1+\lambda_n q_n^2)^{-1}D_n]$

$$+ P_n\underset{\sim}{h}_n\underset{\sim}{h}_n^T P_n.(y_n-\underset{\sim}{h}_n^T\underset{\sim}{a}_n)^2.\text{var}[\lambda_n(1+\lambda_n q_n^2)^{-1}|D_n]$$

(3.2.23)

where the expectations on the right hand side are taken over $p(\lambda_n|D_n)$.

Note that $\lambda_n(1+\lambda_n q_n^2)^{-1}$ is the precision of y_n given λ_n, D_{n-1}. Defining the variable τ_n to be

$$\tau_n = q_n^2\lambda_n(1+q_n^2\lambda_n)^{-1},$$

then clearly $\tau_n \in [0,1]$, and we have the following, rather remarkable, observation;

In order to calculate both the posterior mean and covariance matrix of $\underset{\sim}{\theta}_n|D_n$, all we need is three numerical integrations over $[0,1]$. That is one for the normalizing constant

$$p(y_n|D_{n-1}) = \int_0^1 p(y_n|D_n,\tau_n)p(\tau_n)d\tau_n$$

and one each for the mean and variance of $\tau_n|D_n$. This is of the dimension p of $\underset{\sim}{\theta}_n$. If we compute these moments of $\underset{\sim}{\theta}_n$ directly, we require a total of $\frac{1}{2}(p+1)(p+2)$ numerical integrations over $\mathbb{R}^p$.

Those one dimensional integrals over $[0,1]$ can be done very efficiently and even fairly crude approximations via Simpson's rule provide excellent results.

Furthermore, the calculation of marginal posterior distributions for individual elements of $\underset{\sim}{\theta}_n$ and subsets of $\underset{\sim}{\theta}_n$ are rather difficult from the posterior distribution directly. However, using this approach we have, for example, if $\theta_{1n} = (\underset{\sim}{\theta}_n)_1$,

$$(\theta_{1n}|D_n) \sim \int_0^1 N\left[m_{1n}(\lambda_n), C_{11n}(\lambda_n)\right].p(\lambda_n|D_n)d\lambda_n$$

where $m_{1n}(\lambda_n) = \left(\underset{\sim}{m}_n(\lambda_n)\right)$, and $C_{11n}(\lambda_n) = \left(C_n(\lambda_n)\right)_{11}$, which provides an easier route to the marginal posteriors.

We feel that this approach has much to commend it; it provides exact values for the moments of $\underset{\sim}{\theta}_n$ and easier calculation of distributions of interest. However the modal recursions perform remarkably well as approximations, (as is illustrated later in §3.2.6 by means of numerical examples), and so we examine them further now.

Interpretation of modal recursions.

Defining $\tilde{\lambda}_n = \psi_u(u_n)$ we have the modal recursion for $\underset{\sim}{m}_n$ as

$$\underset{\sim}{m}_n = \underset{\sim}{a}_n + P_n.\underset{\sim}{h}_n.\tilde{\lambda}_n.(1+\tilde{\lambda}_n q_n^2)^{-1}(y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n)$$

$$= E[\underset{\sim}{\theta}_n | D_n, \lambda_n = \tilde{\lambda}_n].$$

The recursion for C_n is, similarly,

$$C_n = \text{var}[\underset{\sim}{\theta}_n | D_n, \lambda_n = \tilde{\lambda}_n] + R_n$$

where R_n is defined as $P_n \underset{\sim}{h}_n \underset{\sim}{h}_n^T P_n.\phi_n(u_n)$, $\phi_n(u_n) = \dfrac{-\psi_u(u_n).u_n.}{(1+q_n^2\psi_u(u_n))^2}$.

Relating these equations to (3.2.22) and (3.2.23) we see that we are approximating

$$E[\lambda_n(1+q_n^2\lambda_n)^{-1} | D_n] \simeq \tilde{\lambda}_n.(1+q_n^2\tilde{\lambda}_n)^{-1}, \qquad (3.2.24)$$

and

$$u_n.\ \text{var}[\lambda_n(1+q_n^2\lambda_n)^{-1} | D_n] \simeq -\tilde{\lambda}_n.(1+q_n^2\tilde{\lambda}_n)^{-2} \qquad (3.2.25)$$

As an example, for a Student t-k likelihood,

$$\tilde{\lambda}_n = (k+1).(k+u_n^2)^{-1}.$$

$\tilde{\lambda}_n$ is actually an approximation to the posterior mean of λ_n, given by

$$\tilde{\lambda}_n = E[\lambda_n | D_n, \underset{\sim}{\theta}_n = \underset{\sim}{a}_n] = E[\lambda_n | y_n, \underset{\sim}{\theta}_n = \underset{\sim}{a}_n] \qquad (3.2.26)$$

To see this, note that from Lemma 2.2.1 of Appendix 24 we have

$$-\frac{\partial}{\partial y_n}\ell n\ p(y_n|\underset{\sim}{\theta}_n) = E\left[-\frac{\partial}{\partial y_n}\ \ell n\ p(y_n|\theta_n,\lambda_n)|y_n,\underset{\sim}{\theta}_n\right]$$

with the expectation being over $p(\lambda_n|y_n,\underset{\sim}{\theta}_n)$. Thus

$$g_\nu(y_n-\underset{\sim}{h}_n^T\underset{\sim}{\theta}_n) = (y_n-\underset{\sim}{h}_n^T\underset{\sim}{\theta}_n).E[\lambda_n|y_n,\underset{\sim}{\theta}_n]$$

$$\psi_\nu(y_n-\underset{\sim}{h}_n^T\underset{\sim}{\theta}_n) = E[\lambda_n|y_n,\underset{\sim}{\theta}_n]$$

and the result (3.2.26) follows.

Now the problem of estimating $\{\lambda_n\}$ falls essentially into the category of many-parameter estimation discussed by Lindley (1971) and Leonard (1976). However, in our framework we have proper priors for both $\underset{\sim}{\theta}_n$ and λ_n at time n and generally, $\omega(\lambda_n)$ will be fairly concentrated (typically so will $p(\underset{\sim}{\theta}_n|D_n)$). For example, with a Student t-k likelihood $\omega(\lambda_n) = k^{-1}\chi^2_k$ with variance $2k^{-1}$. For k fairly large, say between 15 and 20, then ω is fairly concentrated and the posterior $p(\lambda_n|D_n)$ will tend to be concentrated too. Now for a related problem O'Hagan (1976) recommends marginal modes for precision parameters rather than joint modes. Following this we should consider replacing the approximate mean $\tilde{\lambda}_n$ by the mode λ_n^* of $p(\lambda_n|D_n)$ where λ_n^* is a solution

$$-\frac{2\partial}{\partial\lambda_n}\ \ell n\ \omega(\lambda_n) - \lambda_n^{-1} + q_n^2(\lambda_n q_n^2+1)^{-1}2^{-1} + \lambda_n u_n^2.(\lambda_n q_n^2+1)^{-2} = 0 \quad (3.2.27)$$

and which must be obtained iteratively in general.

Then

$$\underset{\sim}{m}_n^* = \underset{\sim}{m}_n(\lambda_n^*) = \underset{\sim}{a}_n + P_n\underset{\sim}{h}_n\lambda_n^*(1+\lambda_n^*q_n^2)^{-1}u_n,$$

and

$$C_n^* = P_n - P_n\underset{\sim}{h}_n\underset{\sim}{h}_n^TP_n.\ \frac{\partial}{\partial y_n}\ \{\lambda_n^*(1+\lambda_n^*q_n^2)^{-1}u_n\}$$

are the recursions corresponding to elimination of λ_n by maximising at marginal mode λ_n^*.

Clearly $\frac{\partial}{\partial y_n}\{\lambda_n^*(1+\lambda_n^* q_n^2)^{-1}u_n\}$ can be found by differentiating and then calculating $\frac{\partial}{\partial y_n}\lambda_n^*$ again by differentiation of (3.2.27). The details are routine, and, following our comments above there will typically be very little difference between λ_n^* and $\tilde{\lambda}_n$ and the numerical examples illustrate the effectiveness of the recursions using $\tilde{\lambda}_n$.

In summary then, the modal recursions have the nice interpretation of an approximate analysis with the nuisance parameters λ_n being replaced by estimates $\tilde{\lambda}_n$ rather than integrated out. In many cases, the fact that $p(\lambda_n|D_n)$ is rather concentrated makes this a reasonable approach. Further the conditional posterior distribution $p(\underset{\sim}{\theta}_n|D_n,\lambda_n=\tilde{\lambda}_n)$ is normal, justifying the normal approximation when using the modal recursions.

Further we can, if required, integrate λ_n out at each stage via a reasonably simple numerical integration, made exceptionally attractive by the fact that the calculations required are the same whatever the dimension p of the regression vector may be. The approximation of $p(\underset{\sim}{\theta}_n|D_n)$ by a normal density with the same mean and covariance can then be seen to be in the same spirit as the work of Harrison and Stevens (1976); here we have a general mixing density $p(\lambda_n|D_n)$ whereas the Harrison and Stevens model utilizes a discrete mixing distribution. Clearly future observations carry information about the current λ_n as long as $\underset{\sim}{\theta}_n$ is unknown. However the comments of §3.1 are pertinent in that it is expedient in this sequential analysis to adopt useful approximations to the full intractable analysis and eliminate λ_n before proceeding to the next observation stage.

3.2.6 Numerical Examples.

The following graphs provide an idea of the behaviour of the modal algorithms applied to a scalar markov model given by

$$y_n = \theta_n + \nu_n,$$
$$\theta_n = \theta_{n-1} + \omega_n.$$

The scale parameter of the density p_ν is unity as implicitly assumed throughout this Chapter and the variance of ω_n is taken as a constant, $W_n = R$ for all n.

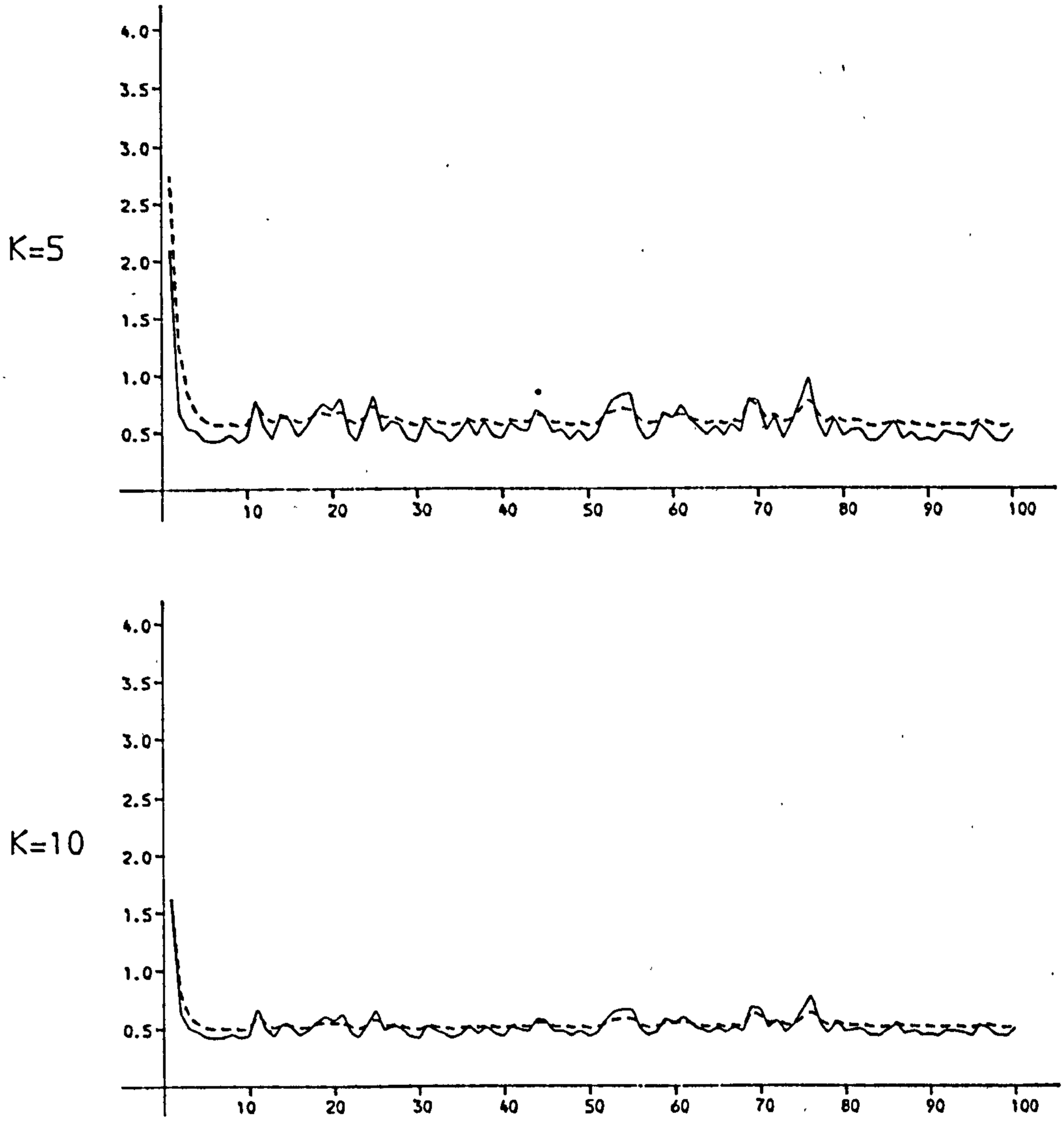
Figures 1. Comparison of Student t based filters, modal/exact.

Several realizations of the system above were generated and modal filters based on Student t distributions were used to track θ_n. The figures 3.2(a), 3.3(a) provide a display of the absolute tracking errors of the modal filters and the exact filters as calculated by the numerical integration discussed earlier. The distribution form which the $\{\nu_n\}$ were drawn is stated, as is the value of the parameter R. In each set of three figures the same data is used providing a comparison of the different properties of the filters for different degrees of freedom parameter k.

The figures 3.2(b), 3.3(b) display the corresponding values of the posterior variance of θ_n as generated by both the modal and exact algorithms.

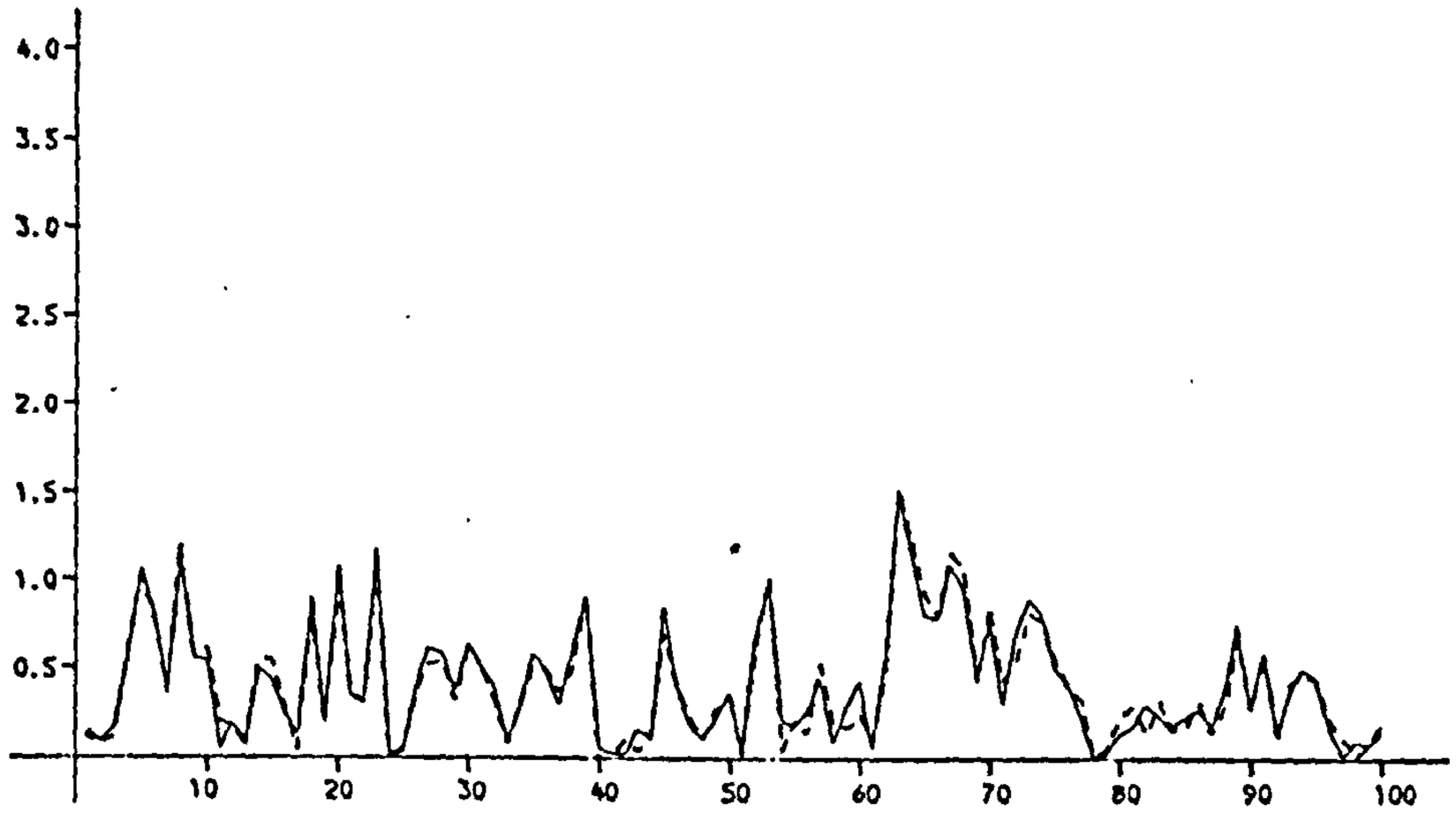
In all of these examples, we began with $m_0 = 0$ and $C_0 = 9$ for both filters, and $\theta_0 = 0$. Many more such simulations have been done and these graphs are typical of those simulations.

3.2(a)

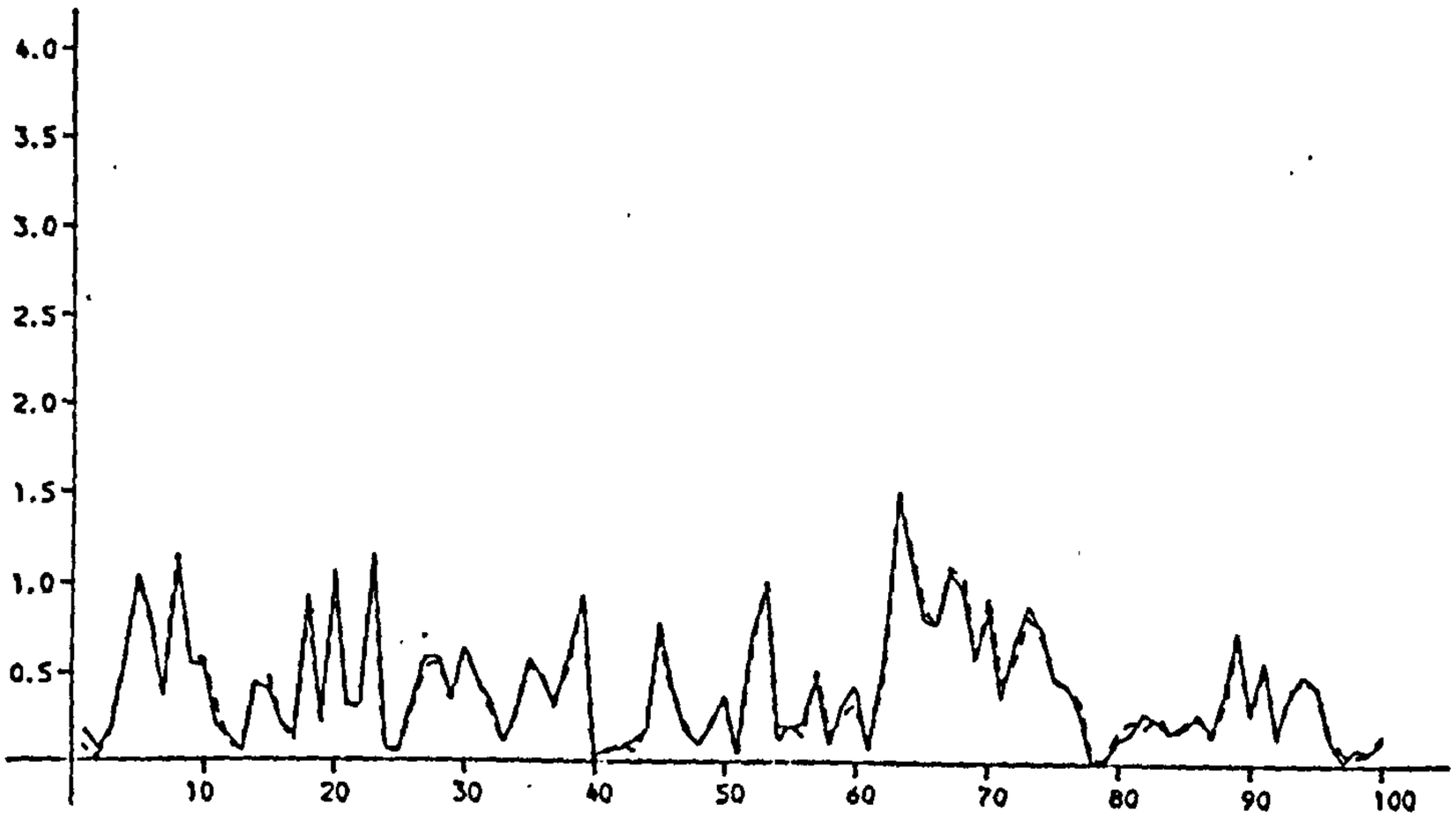


ABSOLUTE ERROR.
DATA FROM N(0, 1) - R=1/3
——— STUDENT T-K MODAL FILTER
-------- STUDENT T-K EXACT FILTER

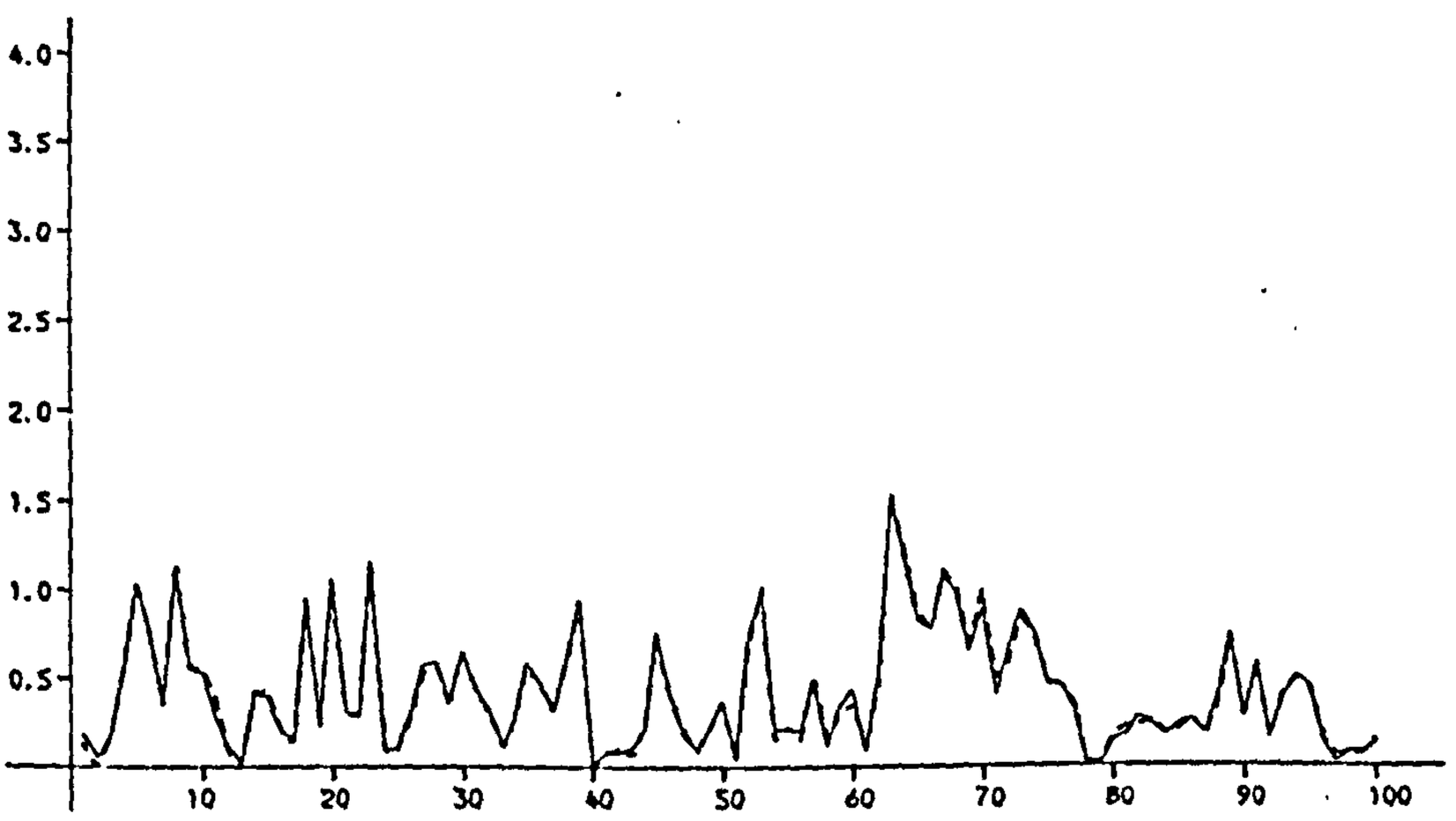
K=5



K=10

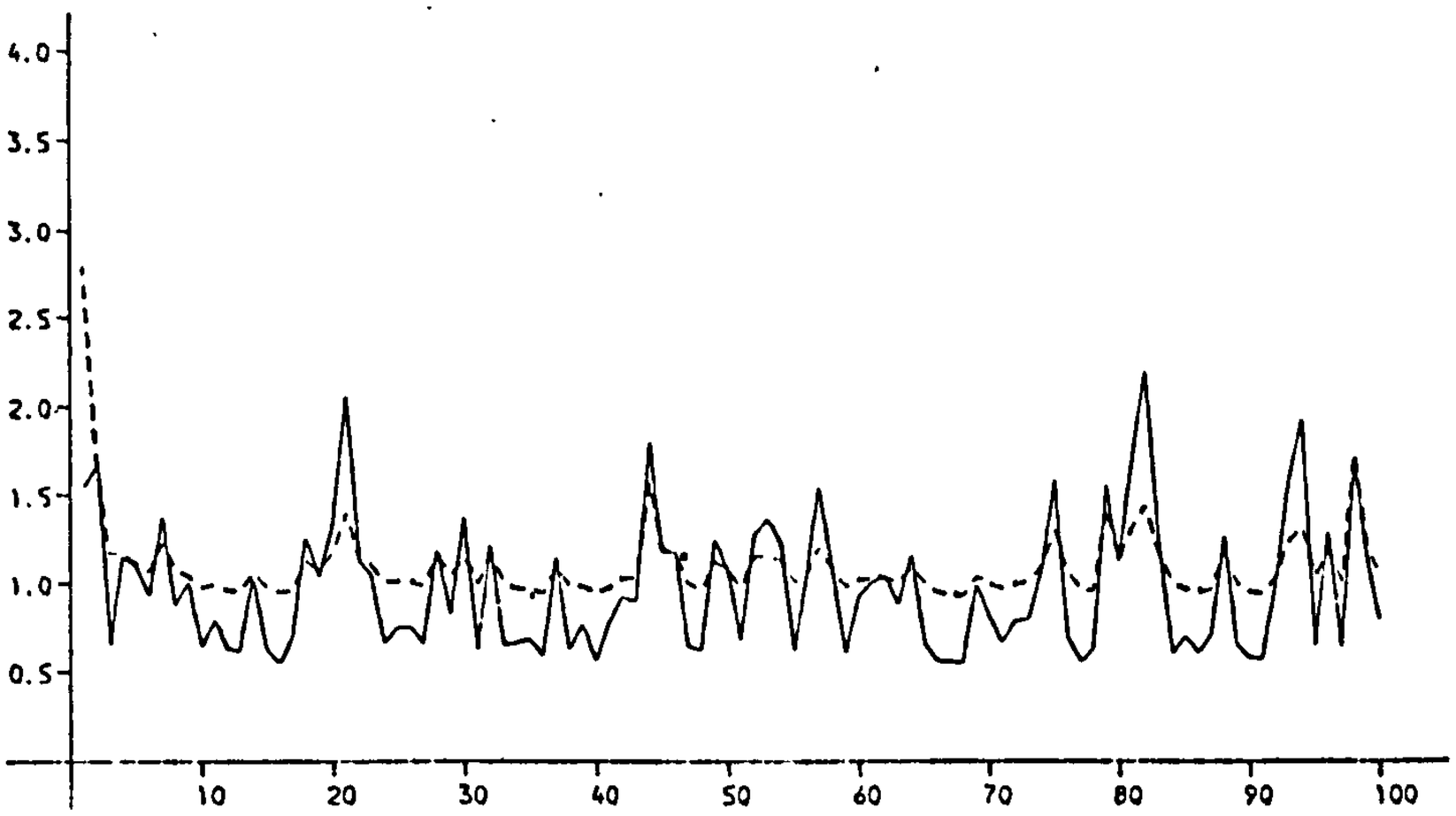


K=15

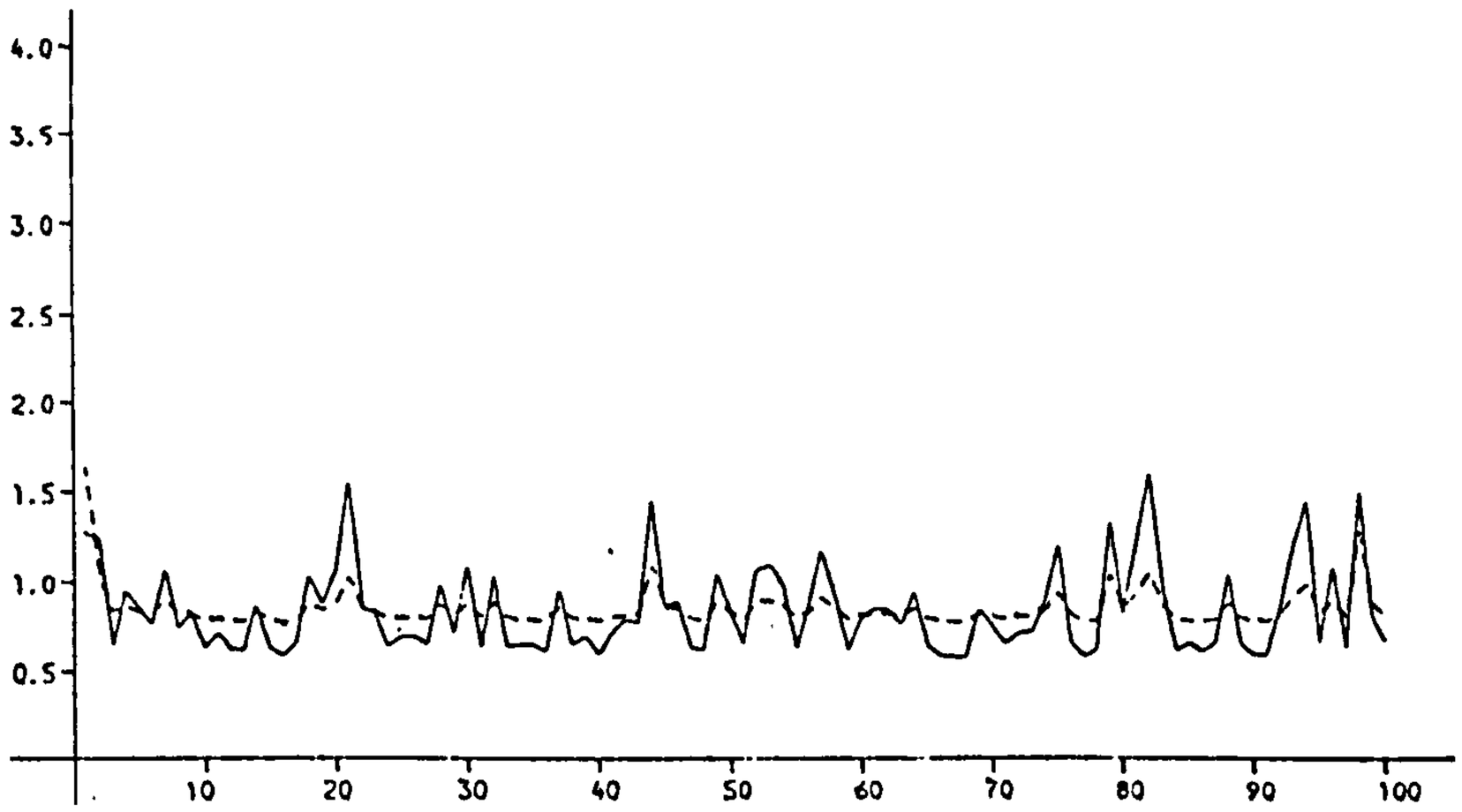


THEORETICAL POSTERIOR VARIANCE.
DATA FROM N(0, 1) - R=1/3
——— STUDENT T-K MODAL FILTER
-------- STUDENT T-K EXACT FILTER

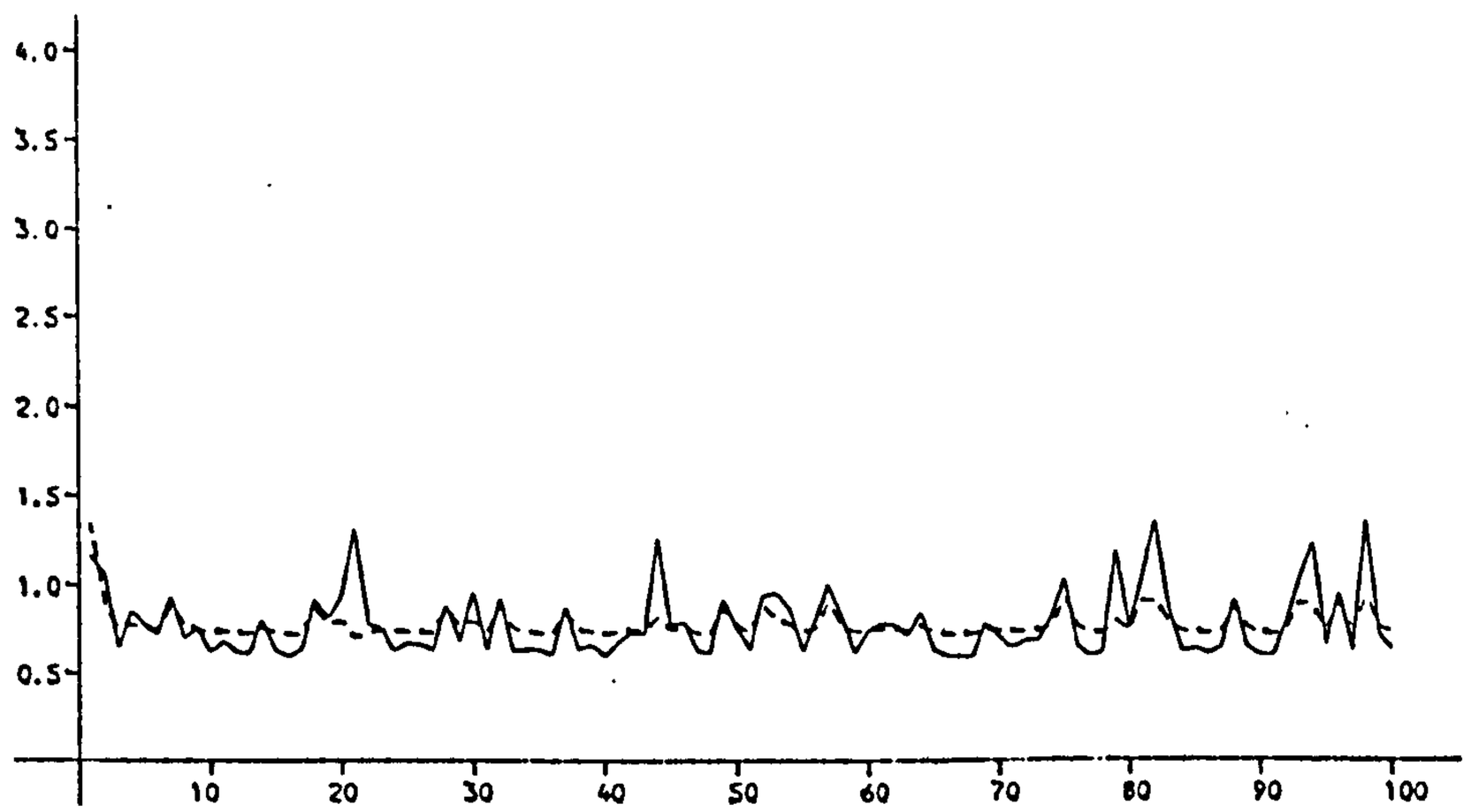
K=5



K=10

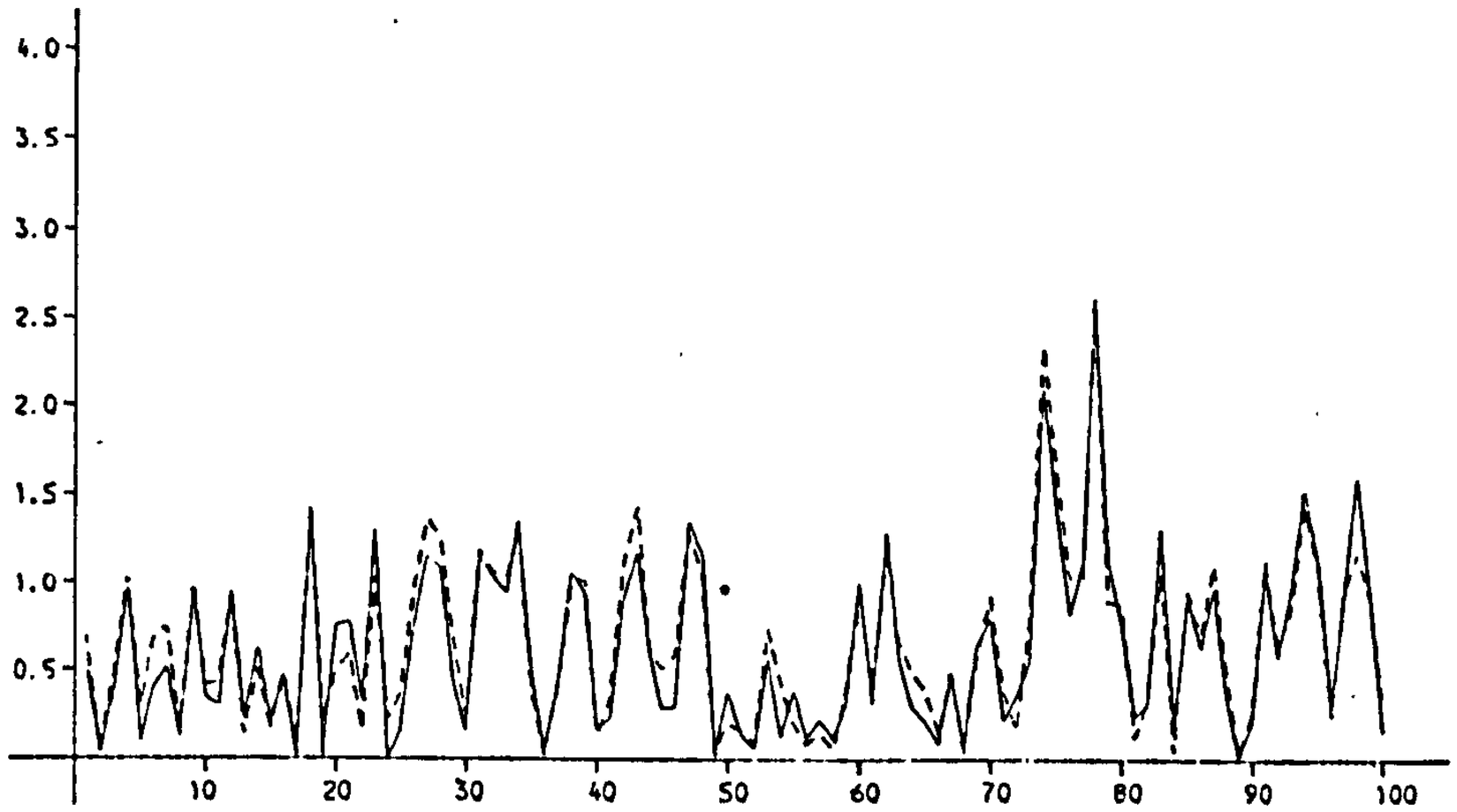


K=15

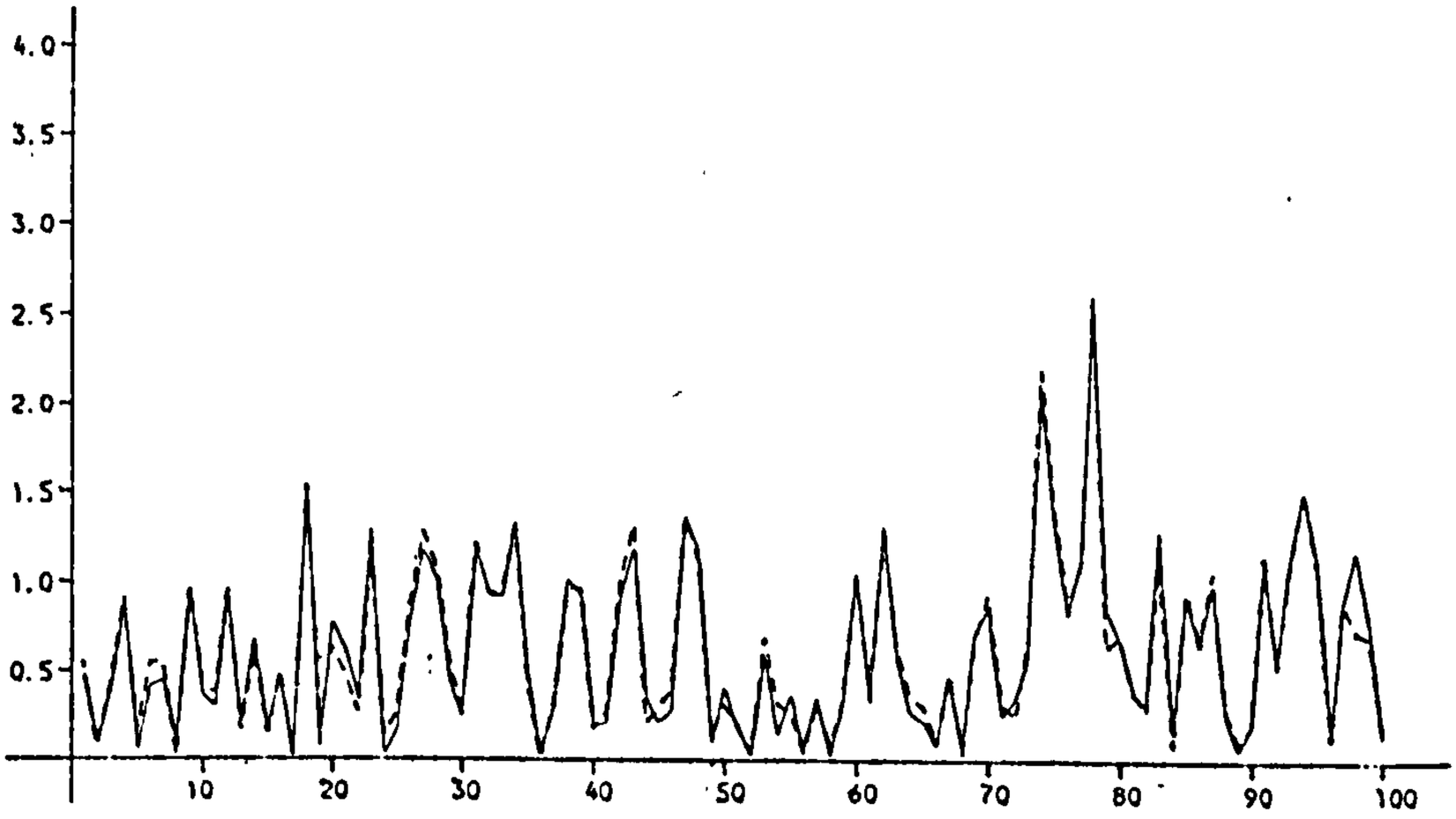


ABSOLUTE ERROR.
DATA FROM N(0, 1) - R=1
——— STUDENT T-K MODAL FILTER
-------- STUDENT T-K EXACT FILTER

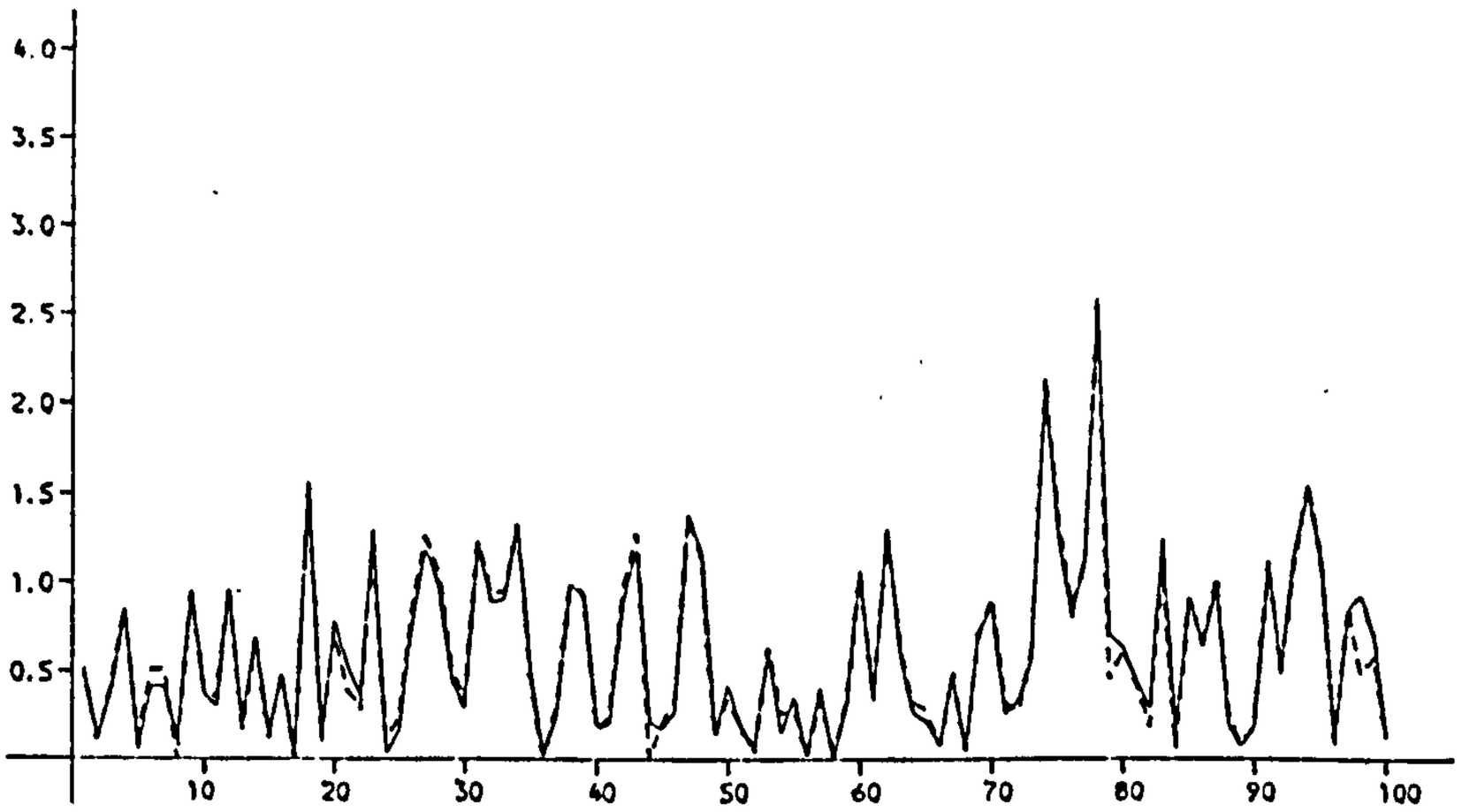
K=5



K=10



K=15



THEORETICAL POSTERIOR VARIANCE.
DATA FROM N(0,1) - R=1
——— STUDENT T-K MODAL FILTER
------- STUDENT T-K EXACT FILTER

Figures 2.

Several realizations of length 30 were generated with various error densities producing the $\{\nu_n\}$ in order to compare robust filters based on Student t densities with the Kalman filter. In each set of three figures we share the following:

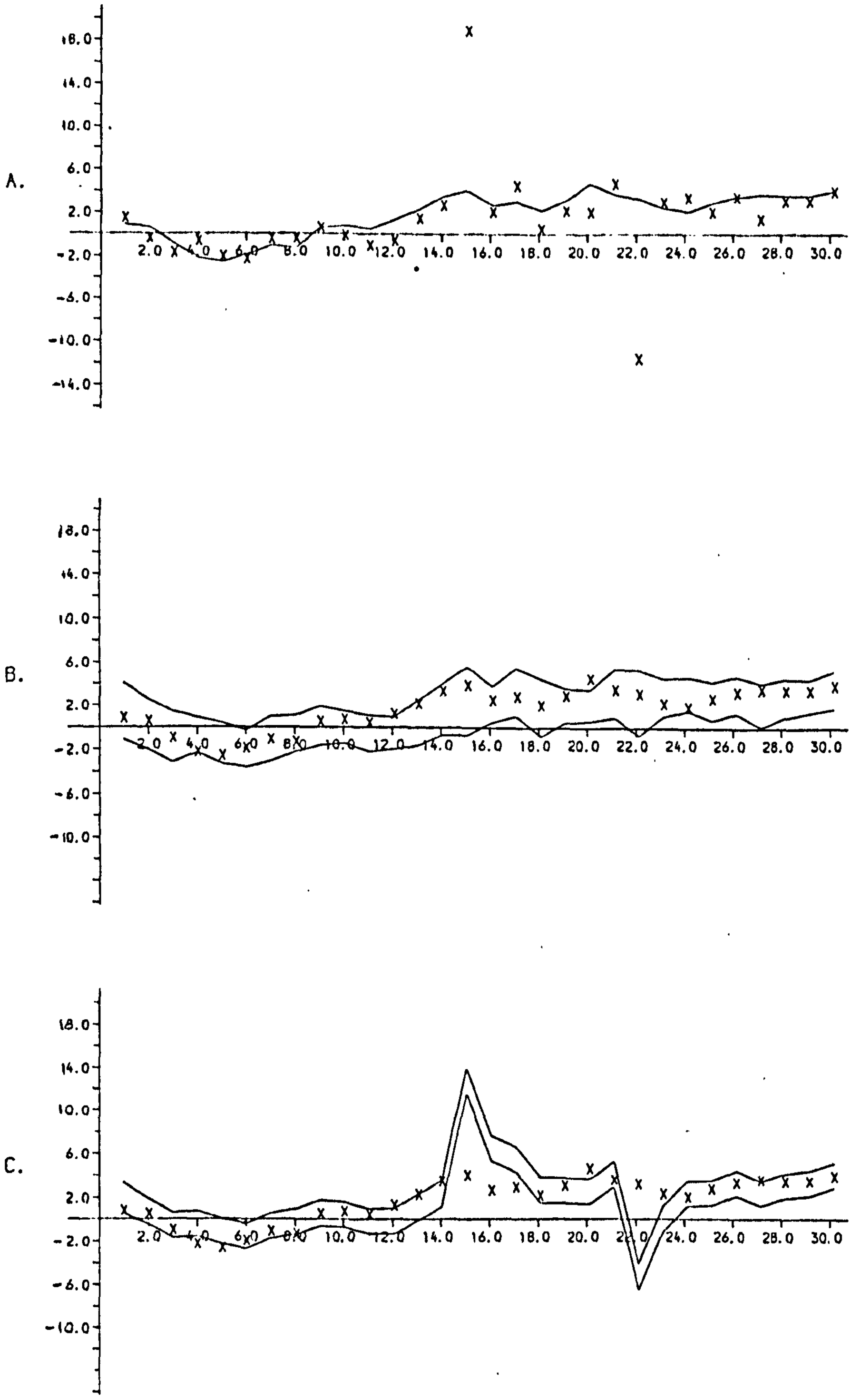
For the actual data generating distribution and value of R as stated on each set of figures, the upper figure A plots the process θ_n and the data y_n. The centre figure B plots the process θ_n and a (symmetric) 95% credible interval for θ_n

$$m_n \pm 1.96.\text{sqrt}(C_n),$$

where m_n, C_n are the mean and variance deriving from a robust modal algorithm based on the density as stated. The lower figure C provides the same plot based on the Kalman filter for figures 3.4a,b.

Both the chosen t density and the normal density for the Kalman filter have scale parameter unity. In each case, the prior specification was $m_0 = 0$, $C_0 = 9$ and we began with $\theta_0 = 0$.

For figures 3.5a,b the lower plot C displays the process θ_n and the 95% interval based on a filter derived from the true likelihood and employing the collapsing procedure of Harrison and Stevens. As shown by Masreliez and Martin (1977) and Masreliez (1975), the collapsing procedure produces results almost identical to those of the exact, but explosive, analysis. Again these figures are typical examples of more extensive numerical studies with different error densities.



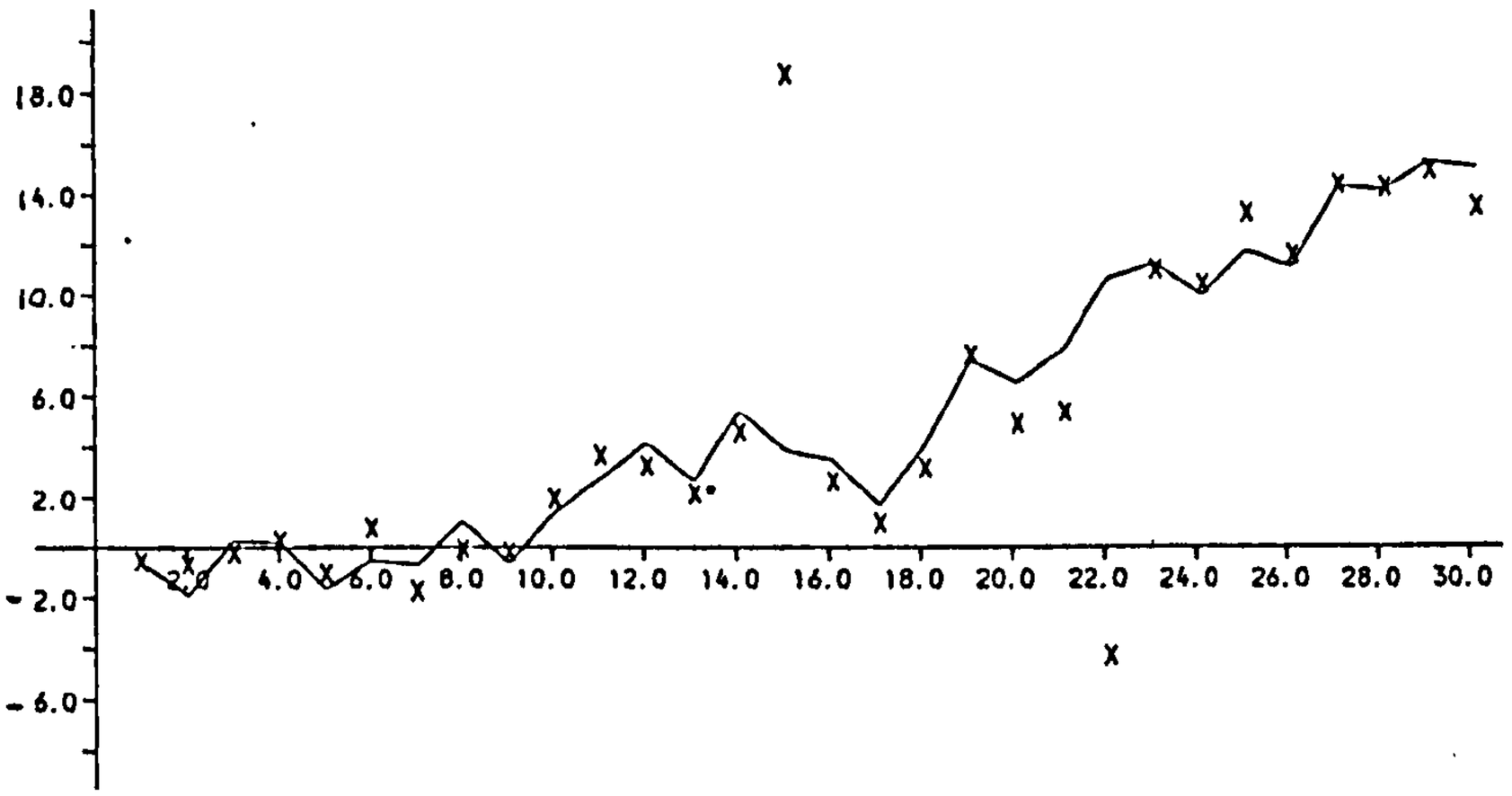
DATA FROM N(0, 1) + OUTLIERS - R=1

A. AR PROCESS AND PROCESS+NOISE

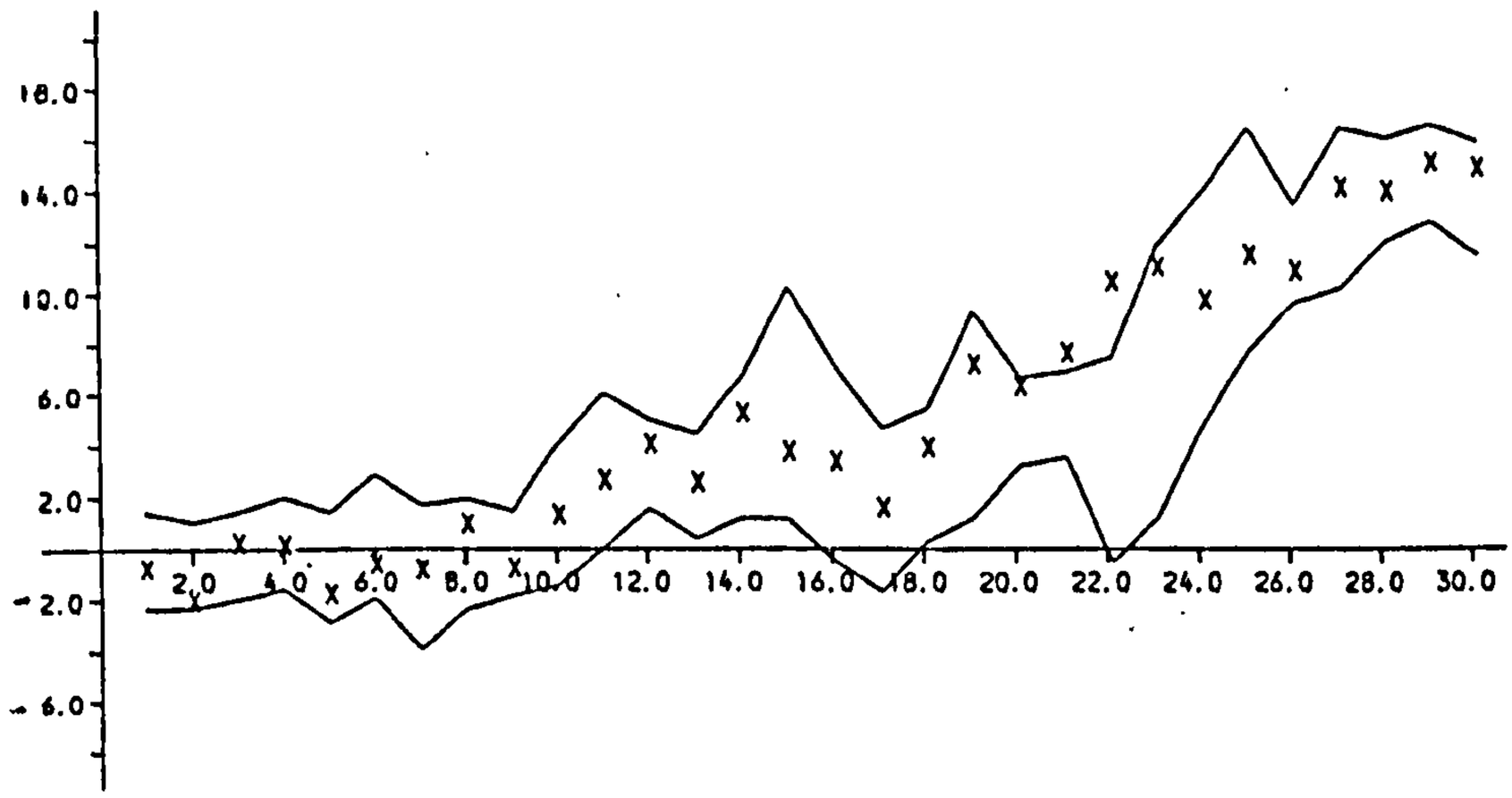
B. 95% INTERVAL - STUDENT T-5 FILTER

C. 95% INTERVAL - KALMAN FILTER

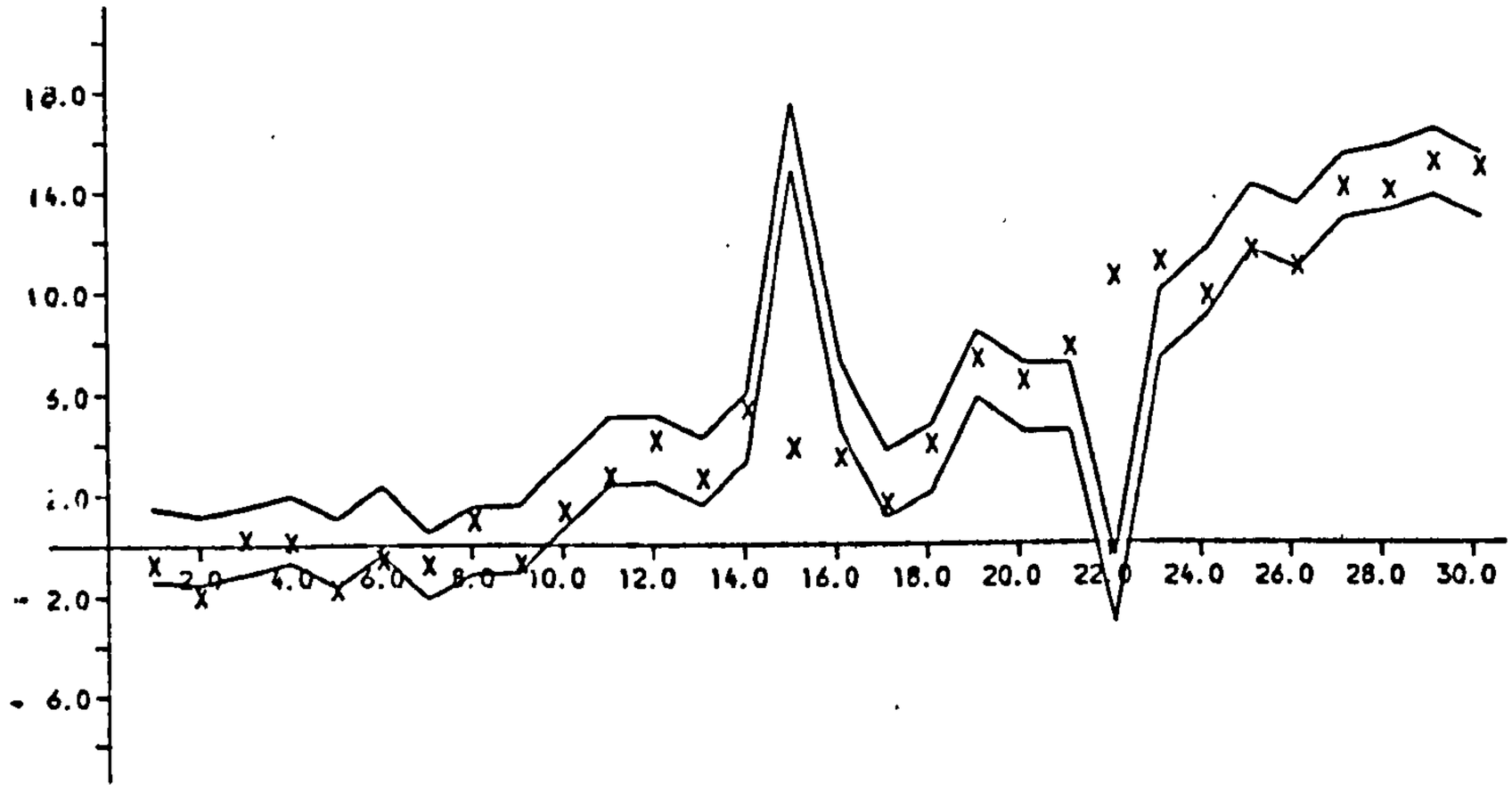
A.



B.



C.

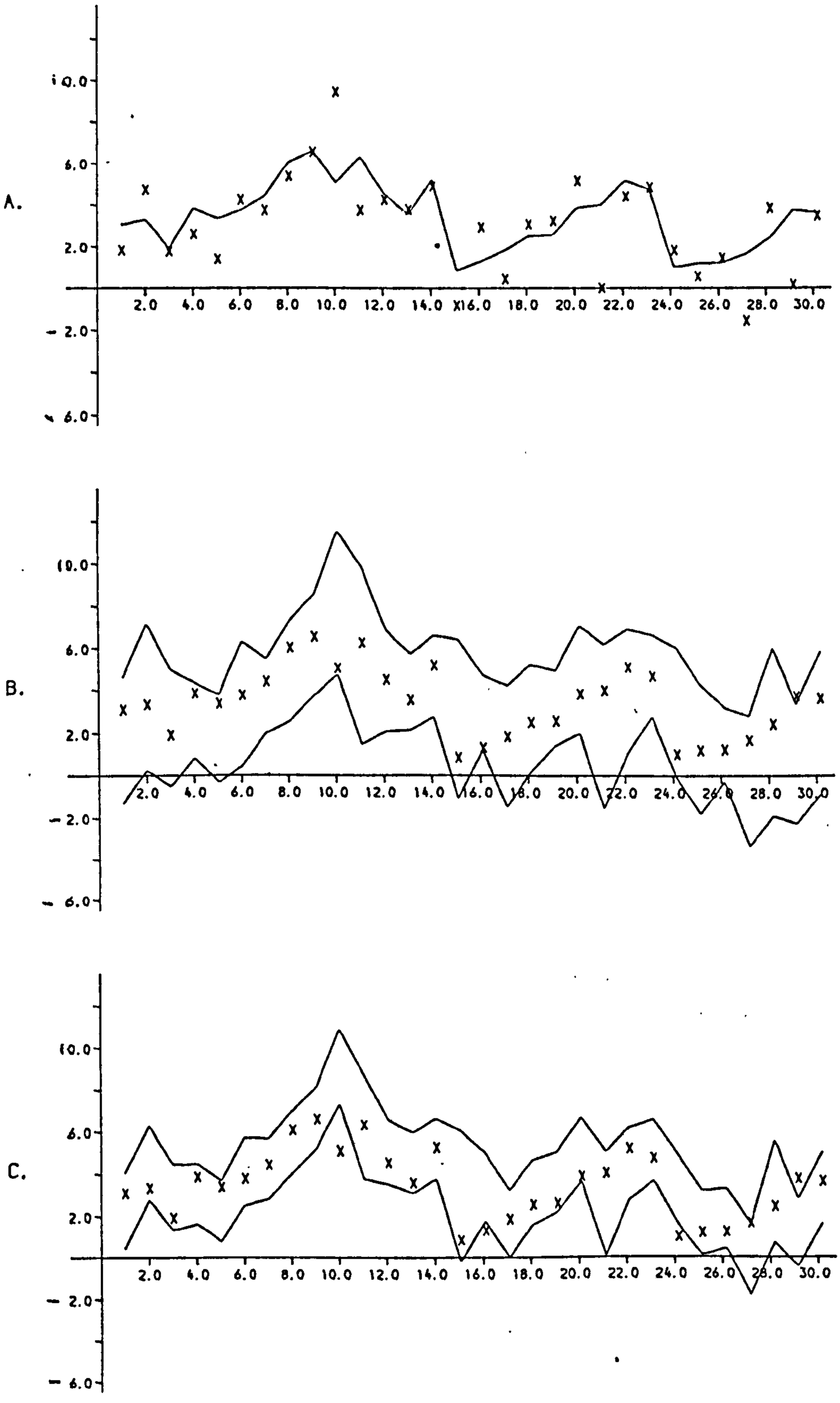


DATA FROM N(0, 1) + OUTLIERS - R=3

A. AR PROCESS AND PROCESS+NOISE

B. 95% INTERVAL - STUDENT T-S FILTER

C. 95% INTERVAL - KALMAN FILTER



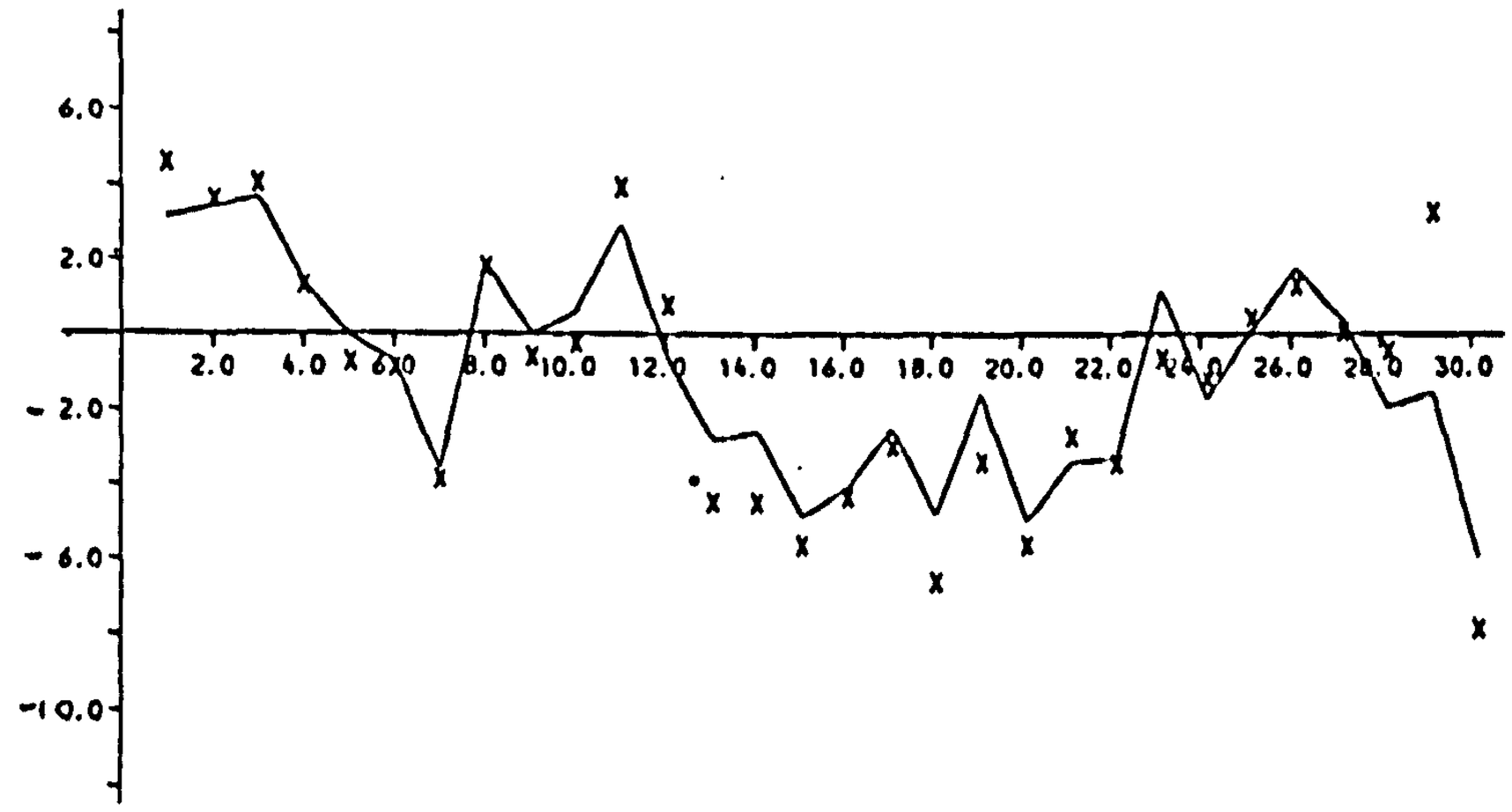
DATA FROM CN(0.1, 16) - R=1

A. AR PROCESS AND PROCESS+NOISE

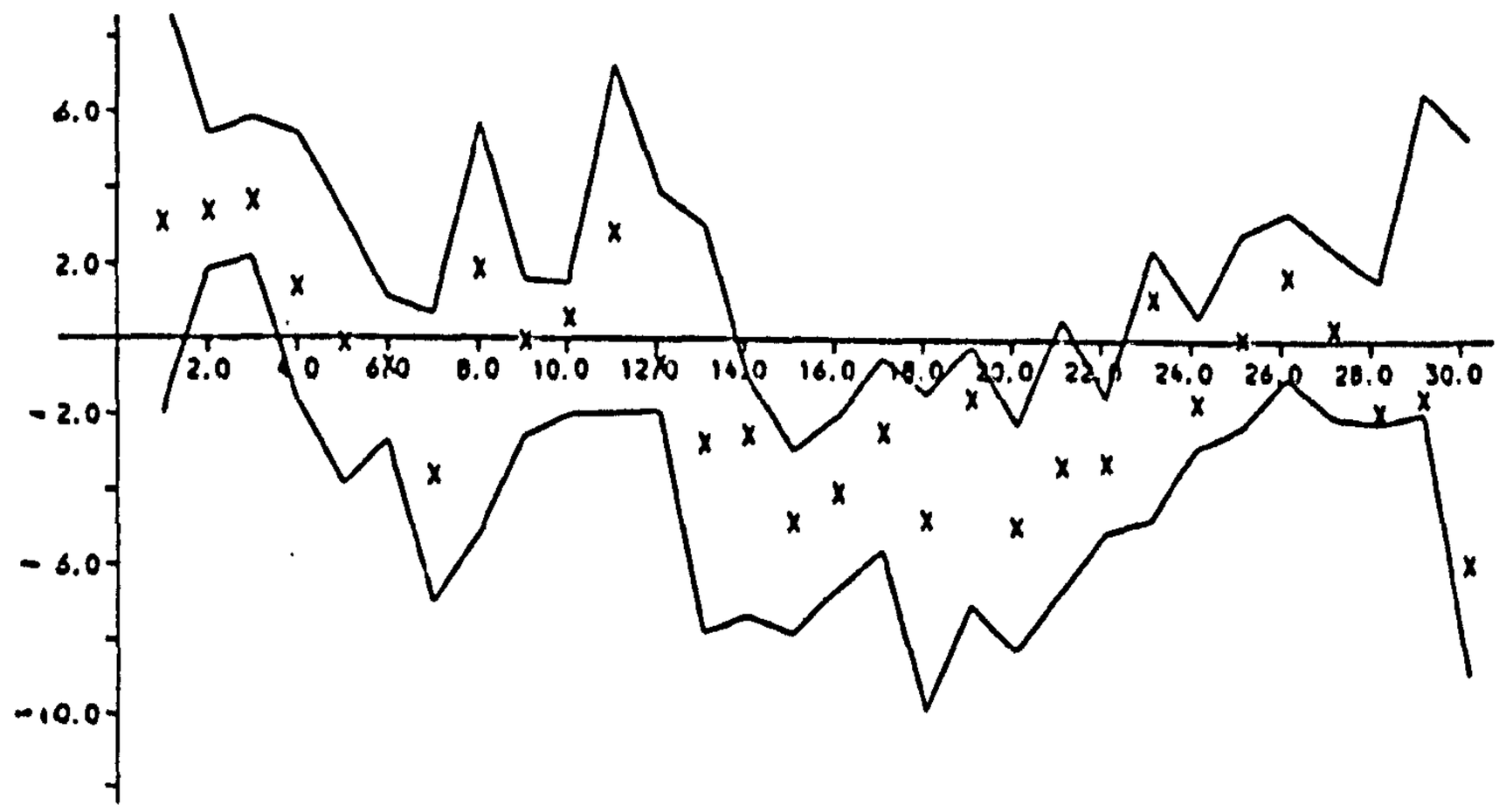
B. 95% INTERVAL - STUDENT T-5 FILTER

C. 95% INTERVAL - TRUE MIXTURE FILTER

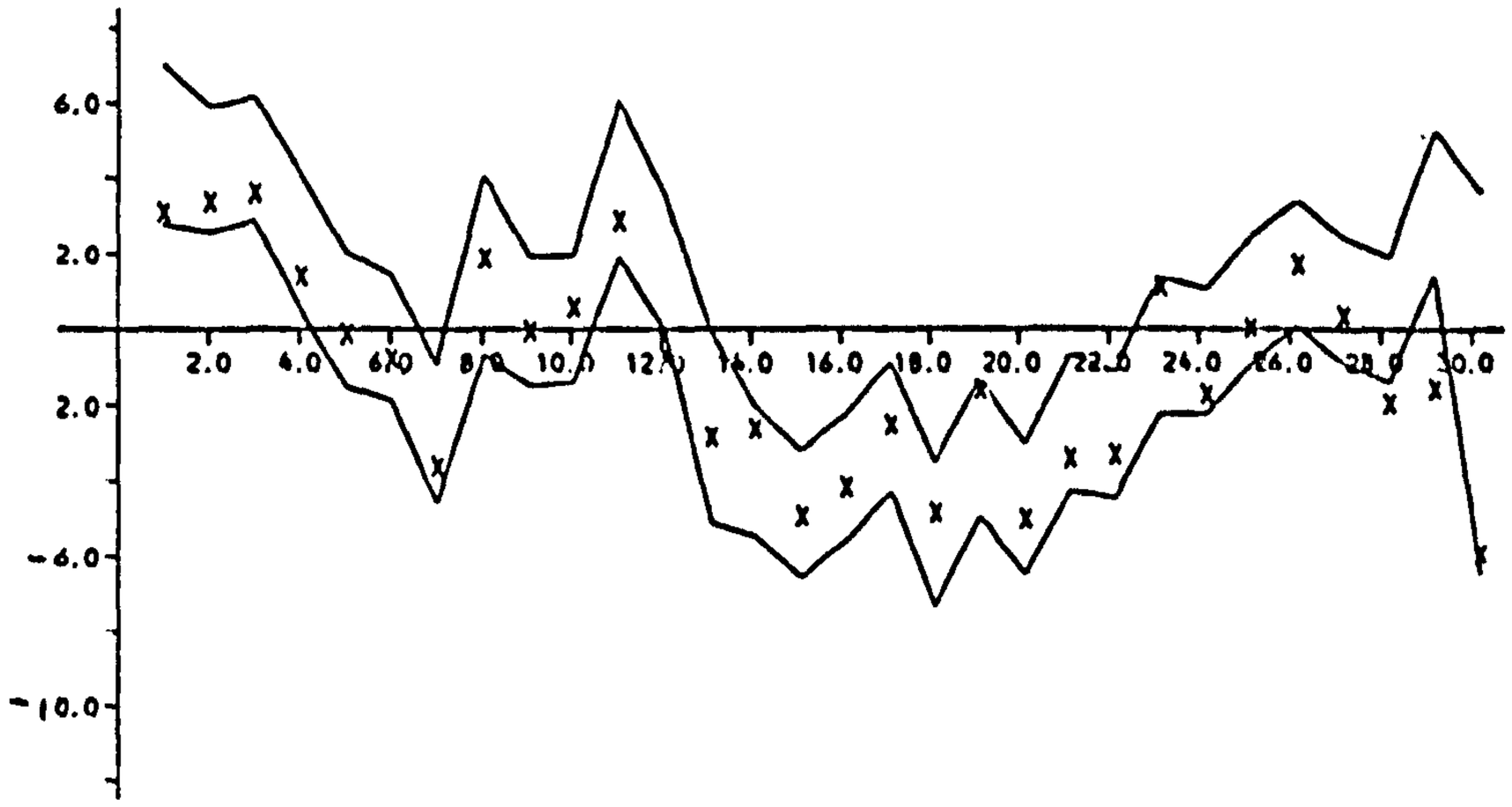
A.



B.



C.



DATA FROM CN(0.1,16) - R=3

A. AR PROCESS AND PROCESS+NOISE

B. 95% INTERVAL - STUDENT T-S FILTER

C. 95% INTERVAL - TRUE MIXTURE FILTER

Figures 3. Simulations.

Each of the following sets of six graphs have the following features.

One thousand sequences of fifteen observations on the markov system above were generated under various conditions. The observational errors for figures 3.6 to 3.8 were generated from the contaminated normal density $CN[0.1,16] = 0.9\phi(\epsilon) + \frac{0.1}{4}\phi\left(\frac{\epsilon}{4}\right)$. The value of the variance ratio R = W/V is as stated in each set of figures. Here V is the observational error variance given by $0.9 + (0.1)16 = 2.5$.

The full lines in the graphs (a) are the averages of the squared errors of the robust filters,

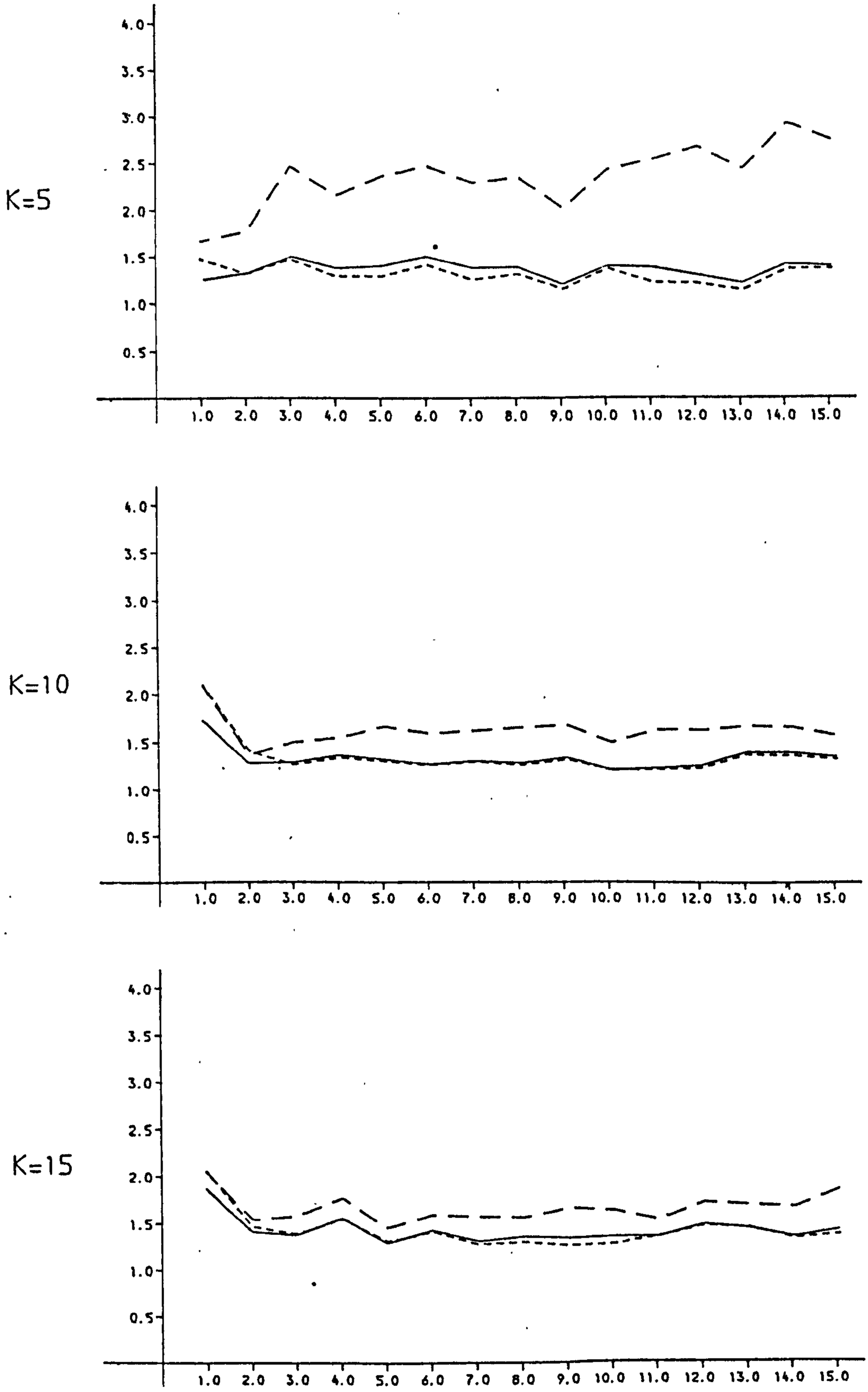
$$x_n = \sum_{j=1}^{1000} (\theta_n - m_{nj})^2/1000, \quad n=1,2,\ldots, 15.$$

where m_{nj} is the posterior mean from the robust filter for the jth sequence of the 1000 runs, at time n. The three graphs are for three values of the degrees of freedom k of the Student t density on which the filter is based.

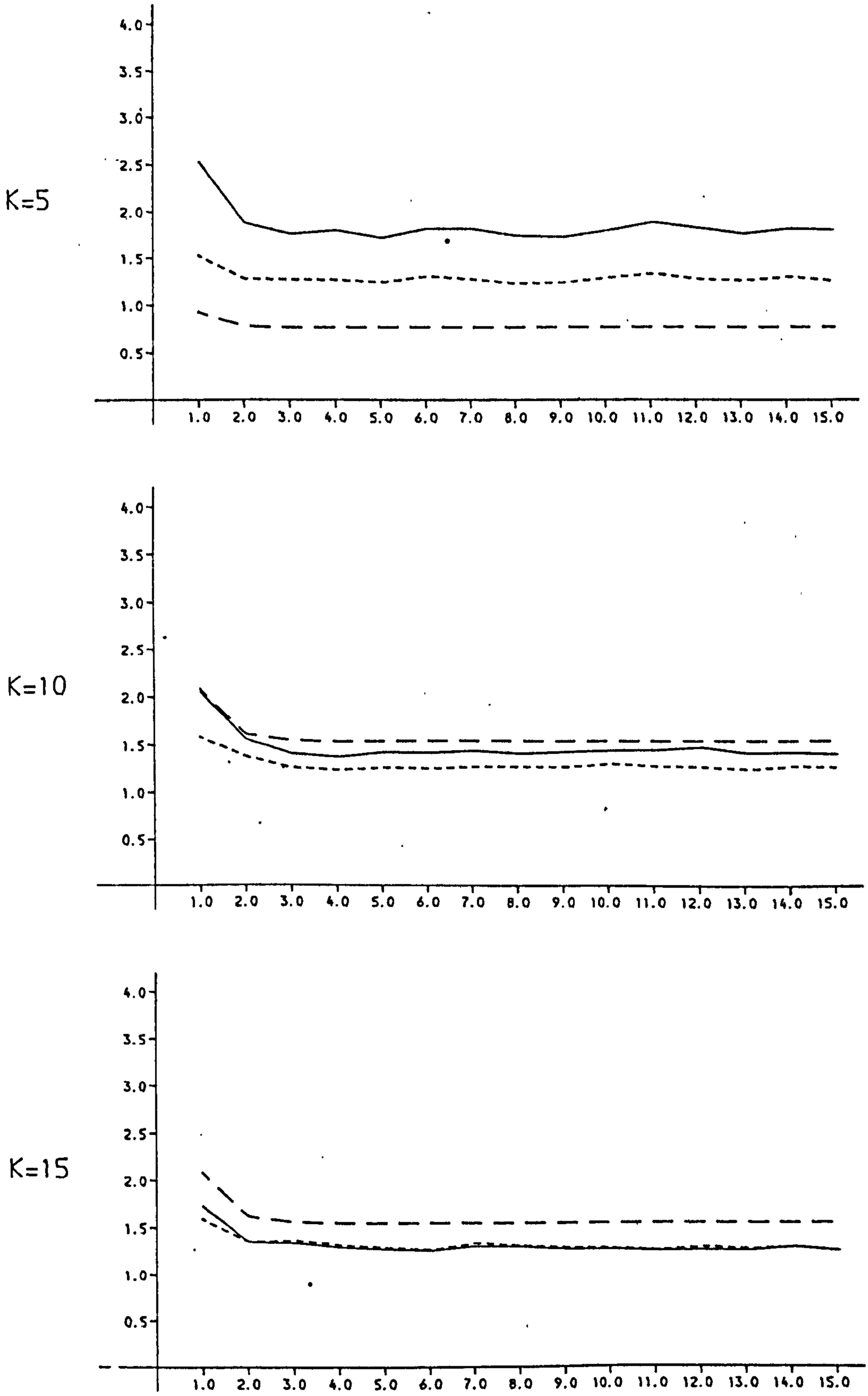
The short dash line is the mean squared error (as above) for the collapsed mixture filter as in Figures 2. The long dash line in the upper frame (k=5) is the average squared error for a Kalman filter based on the nominal $N[0,1]$ density, while that in the other two frames is that based on an $N[0,V]$ density i.e. using the true variance of the non-normal density.

Similarly the graphs b) display the theoretical posterior variances for each of the above mentioned filters,

$$y_n = \sum_{j=1}^{1000} c_n/1000, \quad n=1,2,\ldots,15.$$

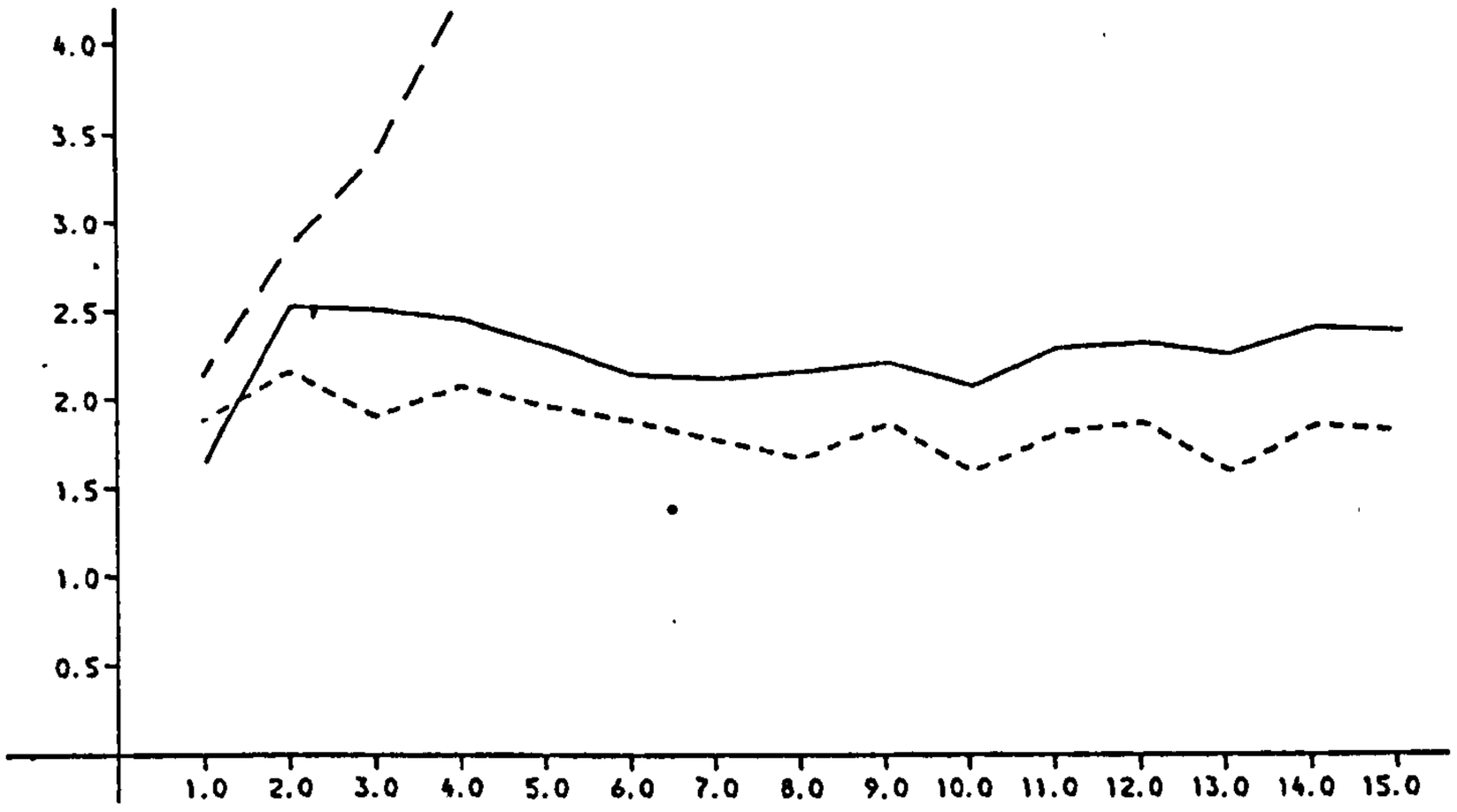


EXPERIMENTAL MEAN SQUARE ERROR.
DATA FROM CN(0.1,.16) R=1
——— STUDENT T-K FILTER
-------- MIXTURE FILTER
- — - — KALMAN FILTER

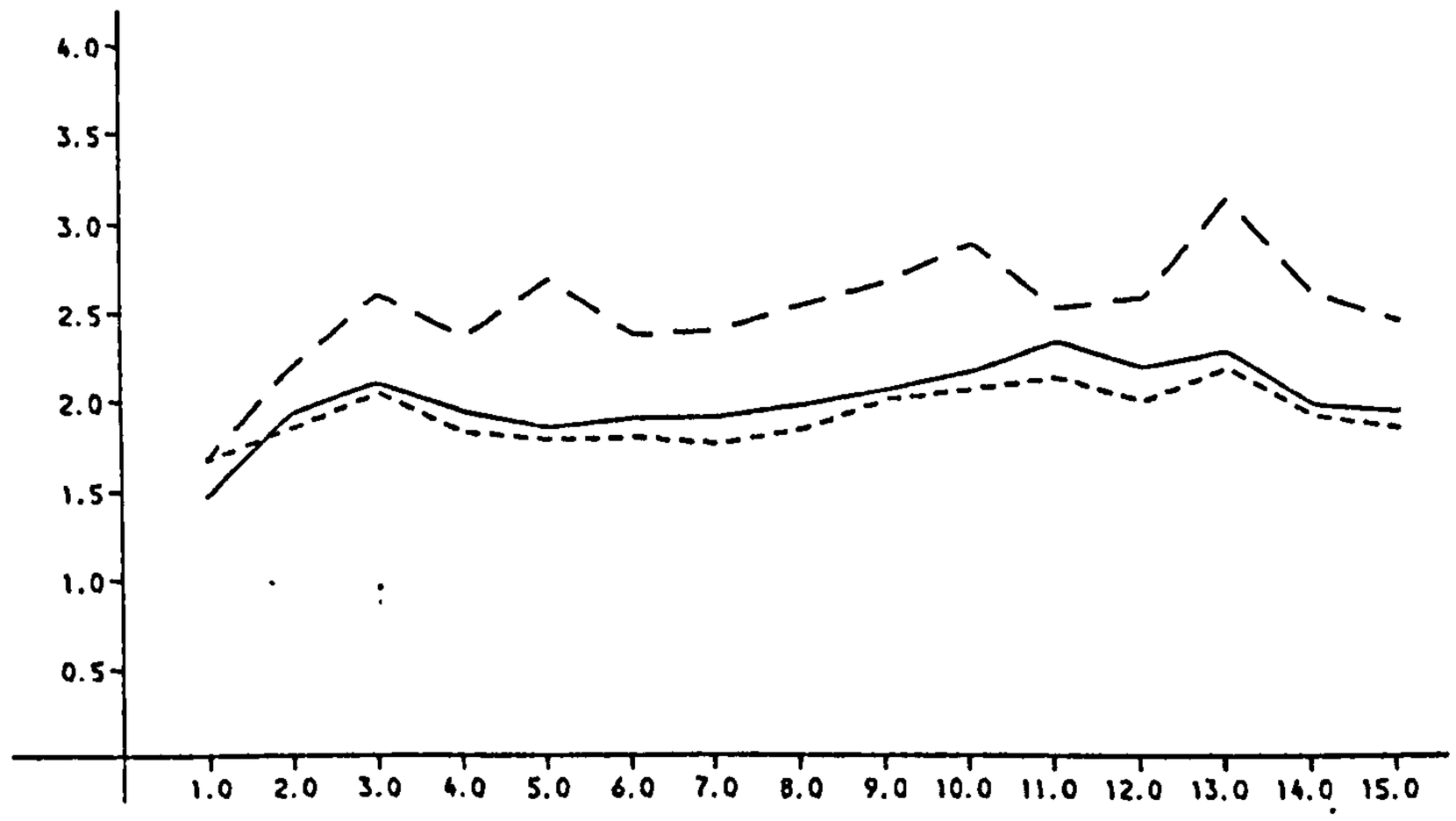


MEAN THEORETICAL POSTERIOR VARIANCE.
DATA FROM CN(0.1,16) R=1
——— STUDENT T-K FILTER
------- MIXTURE FILTER
- - - - KALMAN FILTER

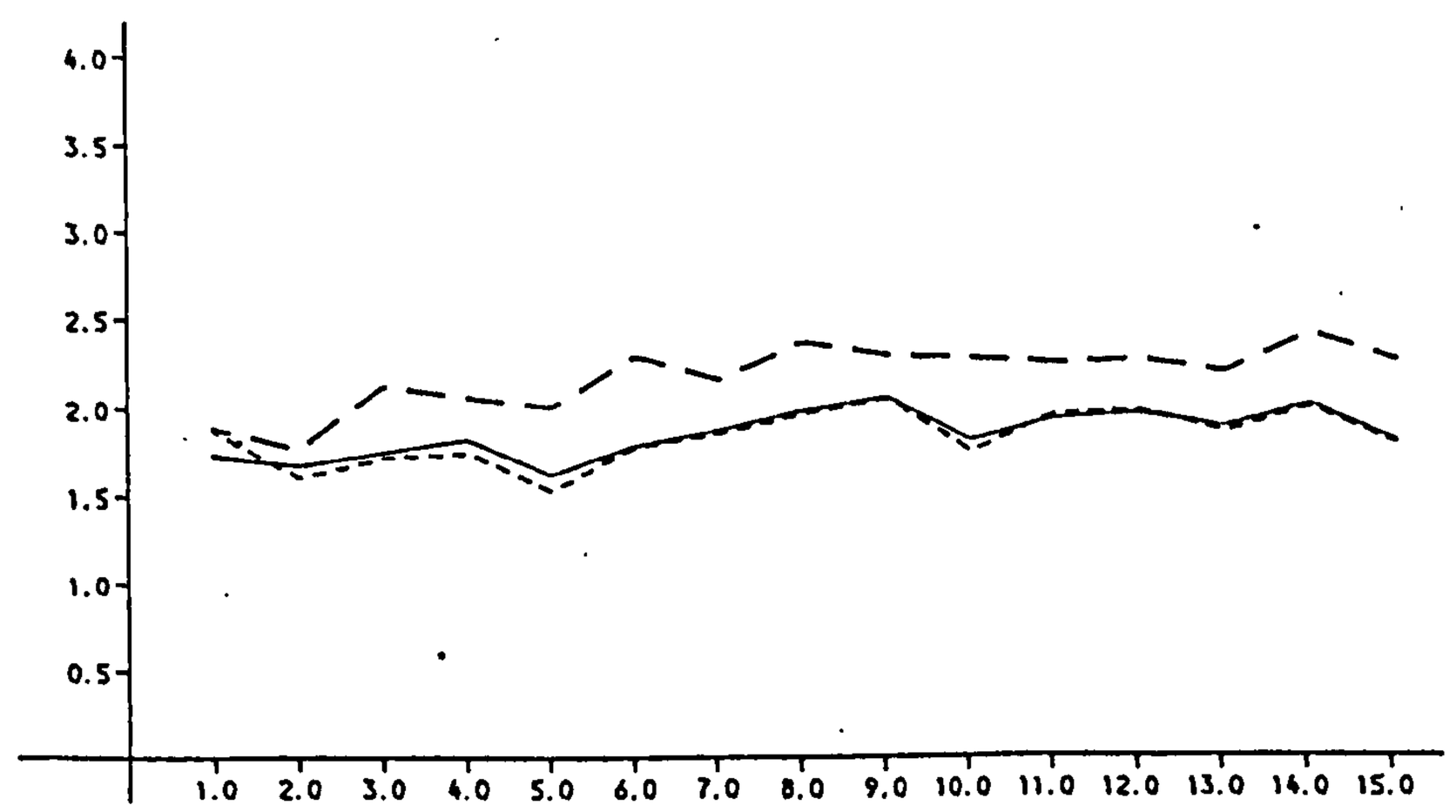
K=5



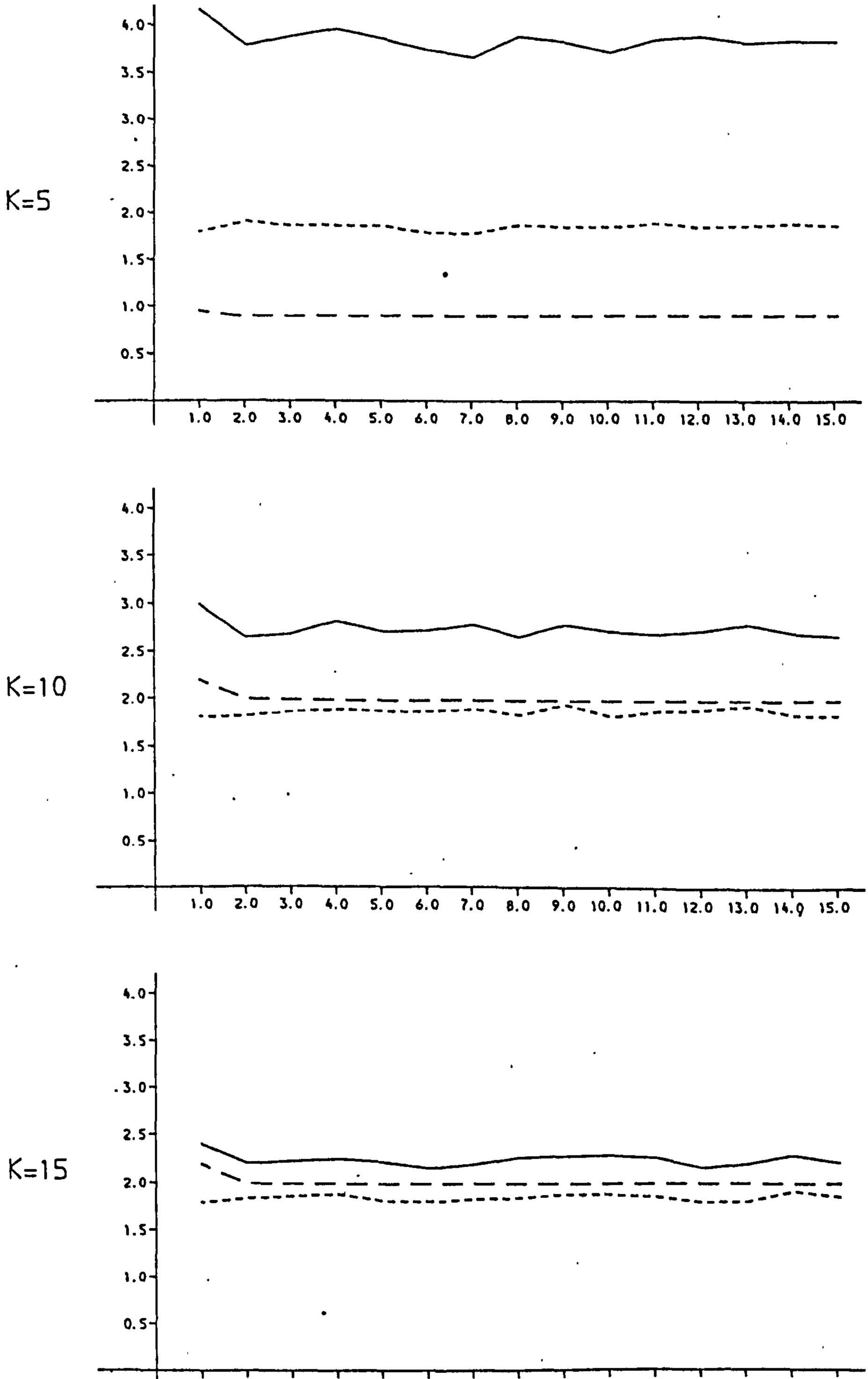
K=10



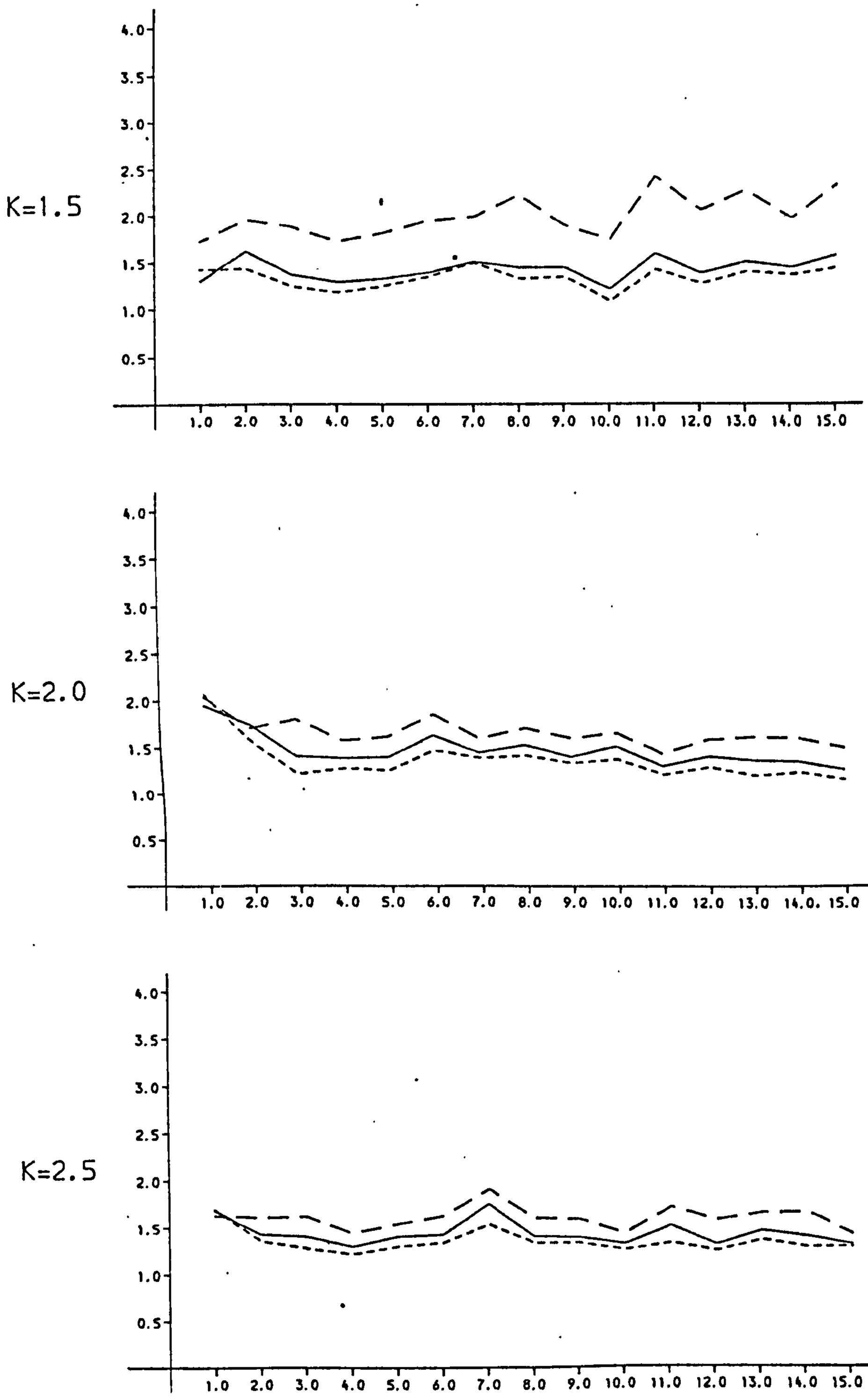
K=15



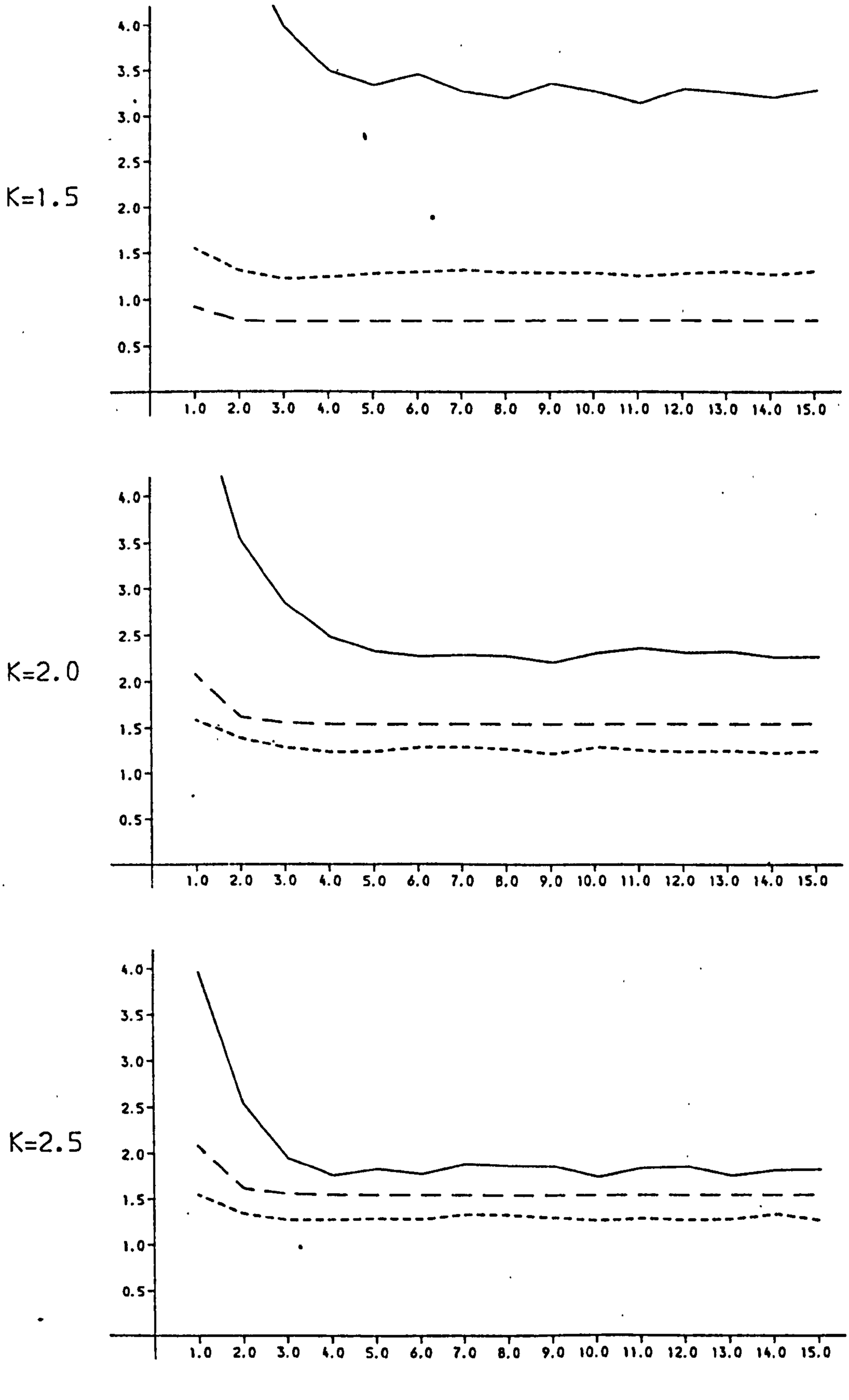
EXPERIMENTAL MEAN SQUARE ERROR.
DATA FROM CN(0.1,16) R=3
——— STUDENT T-K FILTER
-------- MIXTURE FILTER
- — - — KALMAN FILTER



MEAN THEORETICAL POSTERIOR VARIANCE.
DATA FROM CN(0.1,16) R=3
——— STUDENT T-K FILTER
------- MIXTURE FILTER
- — — - KALMAN FILTER



EXPERIMENTAL MEAN SQUARE ERROR.
DATA FROM CN(0.1,16) - R=1
——— HUBER-K FILTER
------- MIXTURE FILTER
- — - — KALMAN FILTER



MEAN THEORETICAL POSTERIOR VARIANCE.
DATA FROM CN(0.1, 16) - R=1
——— HUBER-K FILTER
- - - - - - MIXTURE FILTER

General conclusions.

From figures 1, we see that the modal filter performs, in general, similarly to the exact filter for Student t based algorithms although the latter behaves more smoothly for the range of values of both k and R considered i.e. k=5, 10, 15 and R= $\frac{1}{3}$, 1, 3. Also, the variance $C_n(\hat{\lambda}_n)$ generated by the modal recursion is in these examples almost exactly that given by the exact filter (requiring numerical integrations). Overall there seems very little to lose in using the modal filters.

From figures 2 we get an idea of the robustness provided by tudent t based filters. Use of a robust filter based on a t-k density with k between 5 and 10 requires no more computation effort than the Kalman filter and the performance on normal or near-normal data is similar. The pay-off comes when outliers occur with the robust filter maintaining a smooth and accurate track of the process while the Kalman filter behaves erratically, being totally misled by the aberrant observations.

From figures 3 we see that, for a range of priors and likelihoods, the values produced by the modal recursions for the theoretical posterior variances are typically very close to the simulation mean squared errors. Further we can see that the Kalman filter posterior variances are totally misleading when the data is heavy-tailed non-normal; the values are generally much smaller than the simulation mean squared errors. Finally, it is overwhelmingly clear that the robust filters outperform the Kalman filters in terms of mean squared errors for these simulations.

3.3. Prediction and Smoothing

The previous section discusses the calculation and approximation of the posterior distribution for the parameter vector $\underset{\sim}{\theta}_n$. We now

examine the implications of that analysis for the other important ingredients of these time-series problems, that is the calculation of the predictive densities

$$p(y_n|D_{n-1}),$$

and $$p(y_k|D_n), \quad k > n,$$

and also of the smoothed densities

$$p(\theta_k|D_n), \quad k < n.$$

The marginal (predictive) density $p(y_n|D_{n-1})$ is required for model criticism as in Box (1981), and in the related framework of adaptive estimation via mixture modelling.

The predictive densities for y_{n+h}, $k > 1$ form the basis from which forecasts of future behaviour are made. We discuss these first.

3.3.1 Prediction.

(i) $p(y_n|D_{n-1})$

This density is defined by $\int_{\mathbb{R}^p} p_\nu(y_n - \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n) \cdot p(\underset{\sim}{\theta}_n|D_{n-1}) d\underset{\sim}{\theta}_n$.

We have noticed that this density is essentially a "flattened" version of the likelihood in that convulution with the normal prior does not change the essential characteristics of the tail-behaviour. For the heavy-tailed likelihoods of interest, all the approximate filtering algorithms involve an approximation to $p(y_n|D_{n-1})$ as discussed below.

a) Masreliez and Martins' approach.

The scaled version of the likelihood is just

$$p(y_n|D_{n-1}) \simeq \sigma_n^{-1} p_\nu(\sigma_n^{-1} u_n), \quad u_n = y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n,$$

with $$\sigma_n^2 = 1 + q_n^2 \text{ and } q_n^2 = \underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n.$$

b) The gradient algorithm.

In deriving this algorithm in §3.2.3, the Taylor series approximation to the likelihood was

$$\ell n\ p_\nu(y_n - \underset{\sim}{h}_n^T\underset{\sim}{\theta}_n) \simeq \ell n\ p_\nu(u_n) + \underset{\sim}{\tilde{\theta}}_n^T.\underset{\sim}{h}_n.g_\nu(u_n) - \tfrac{1}{2}\underset{\sim}{\tilde{\theta}}_n^T\underset{\sim}{h}_n.\underset{\sim}{h}_n^T\underset{\sim}{\tilde{\theta}}_n.G_\nu(u_n)$$

where $\underset{\sim}{\tilde{\theta}}_n = \underset{\sim}{\theta}_n - \underset{\sim}{a}_n$.

Hence

$$p(y_n|D_{n-1}) \simeq p_\nu(u_n).\int_{\mathbb{R}^p} p(\underset{\sim}{\theta}_n|D_{n-1})\exp\left[\underset{\sim}{\tilde{\theta}}_n^T\underset{\sim}{h}_n g_\nu(u_n) - \tfrac{1}{2}\underset{\sim}{\tilde{\theta}}_n^T\underset{\sim}{h}_n h_n^T\underset{\sim}{\tilde{\theta}}_n G_\nu(u_n)\right]d\theta_n$$

$$= (2\pi)^{-p/2}|P_n|^{-\frac{1}{2}}p_\nu(u_n)\int_{\mathbb{R}^p}\exp\{\underset{\sim}{\tilde{\theta}}_n^T\underset{\sim}{h}_n.g_\nu(u_n) - \tfrac{1}{2}\underset{\sim}{\tilde{\theta}}_n^T.[\underset{\sim}{h}_n\underset{\sim}{h}_n^T G_\nu(u_n) + P_n^{-1}]\underset{\sim}{\tilde{\theta}}_n\}d\underset{\sim}{\theta}_n$$

$$= |C_n|^{\frac{1}{2}}|P_n|^{-\frac{1}{2}}p_\nu(u_n)\int_{\mathbb{R}^p} p(\underset{\sim}{\theta}_n|D_n)\ d\underset{\sim}{\theta}_n.\exp\{\tfrac{1}{2}g_\nu^2(u_n).\underset{\sim}{h}_n^T C_n\underset{\sim}{h}_n\} \qquad (3.3.1)$$

where $p(\underset{\sim}{\theta}_n|D_n)$ is the gradient approximation

$$(\underset{\sim}{\theta}_n|D_n) \sim N[\underset{\sim}{m}_n, C_n],$$

with $\underset{\sim}{m}_n$, C_n defined by (3.2.8) and (3.2.9).

Thus (3.3.1) is just

$$|C_n|^{\frac{1}{2}}|P_n|^{-\frac{1}{2}}p_\nu(u_n)\exp\{\tfrac{1}{2}g_\nu(u_n)\underset{\sim}{h}_n^T C_n.\underset{\sim}{h}_n\}$$

or, since $C_n^{-1} = P_n^{-1} + \underset{\sim}{h}_n\underset{\sim}{h}_n^T G_\nu(u_n)$, we have

$$p(y_n|D_{n-1}) \simeq [1+q_n^2 G_\nu(u_n)]^{-\frac{1}{2}}\exp\{\tfrac{1}{2}g_\nu(u_n)^2 q_n^2.(1+q_n^2 G_\nu(u_n))^{-1}\}\ .\ p_\nu(u_n) \qquad (3.3.2)$$

Notice that at normality, $(y_n|D_{n-1}) \sim N\left[\underset{\sim}{h}_n^T\underset{\sim}{a}_n, 1+q_n^2\right]$ as required.

Clearly both (i) and (ii) suffer from the problems outlined in 3.2 and in general cannot be recommended. The modal approximation does, however, lead to excellent approximations to predictive distributions with none of those drawbacks.

(iii) <u>The modal approximation</u>.

Given the modal recursion we have the approximate marginal score given by

$$g(u_n) \approx \left(1+q_n^2\psi_\nu(u_n)\right)^{-1} g_\nu(u_n)$$

or

$$p(y_n|D_{n-1}) \propto \exp\left\{-\int_{-\infty}^{\infty} \left(1+q_n^2\psi_\nu(u_n)\right)^{-1} g_\nu(u_n)du_n\right\}. \qquad (3.3.3)$$

Examples 3.3.1

(a) Student t likelihood with k degrees of freedom. $g(u_n)$ is now the score of a scaled Student t-k density, $\sigma_n^{-1}.T_k(\sigma_n^{-1}u_n)$, where $\sigma_n^2 = 1+q_n^2(1+k^{-1})$.

(b) Double exponential likelihood, $p_\nu(u) \propto \exp - |u|$.

Now

$$g(u_n) = (1+q_n^2|u_n|^{-1})^{-1}.\mathrm{sgn}(u_n)$$

$$= (|u_n|+q_n^2)^{-1}u_n.$$

Writing

$$g(u_n) = \mathrm{sgn}(u_n) - q_n^2\ \mathrm{sgn}(u_n).(|u_n|+q_n^2)^{-1},$$

we have

$$\int g(u_n)\ du_n = |u_n| - q_n^2\ \ell n(|u_n|+q_n^2) + \text{constant},$$

and so

$$p(y_n|D_{n-1}) \propto \exp\{-|u_n|\}.\ (|u_n|+q_n^2)^{q_n^2}\ .$$

[Notice that this is of the form $p_\nu(u_n).(|u_n|+q_n^2)^{q_n^2}$].

(c) Huber k likelihood.

Now $g(u_n) = \begin{cases} u_n(1+q_n^2)^{-1}, & |u_n| \leq k; \\ ku_n(|u_n|+kq_n^2)^{-1}, & \text{otherwise.} \end{cases}$

Then

$$p(y_n|D_{n-1}) \propto \begin{cases} \exp\{-\tfrac{1}{2}u_n^2.(1+q_n^{\ 2})^{-1}\}, & |u_n| \leq k; \\ (|u_n|+k^2q_n^2)^{k^2q_n^2}\exp\{-k|u_n|\}, & \text{otherwise.} \end{cases}$$

Of course we can calculate the exact marginal density very easily via the expression of p_ν as a scale mixture of normal densities.

Clearly

$$p(y_n|D_{n-1}) = \int_0^\infty p(y_n|D_{n-1},\lambda_n).\omega(\lambda_n)\, d\lambda_n$$

where

$$(y_n|D_{n-1},\lambda_n) \sim N\left[\underset{\sim}{h}_n^T \underset{\sim}{a}_n,\ \lambda_n^{-1}+q_n^2\right],$$

and $\omega(\lambda_n)$ is the mixing density. This one-dimensional integral can again be transformed easily to the unit interval and simple quadrature used to evaluate it. As earlier, this avoids integration over $\mathbb{R}^p$ and is independent of p, and is the recommended method of calculating the marginal density.

The following examples display the scaled approximation of Masreliez and Màrtin, our modal approximation and the exact predictive density for a few likelihoods with various prior specifications.

The likelihoods are Student t-k, with k=5, 10, 15 and double-exponential. The prior was centred at zero with $q^2 = c = \frac{1}{3}$, 1, 3.

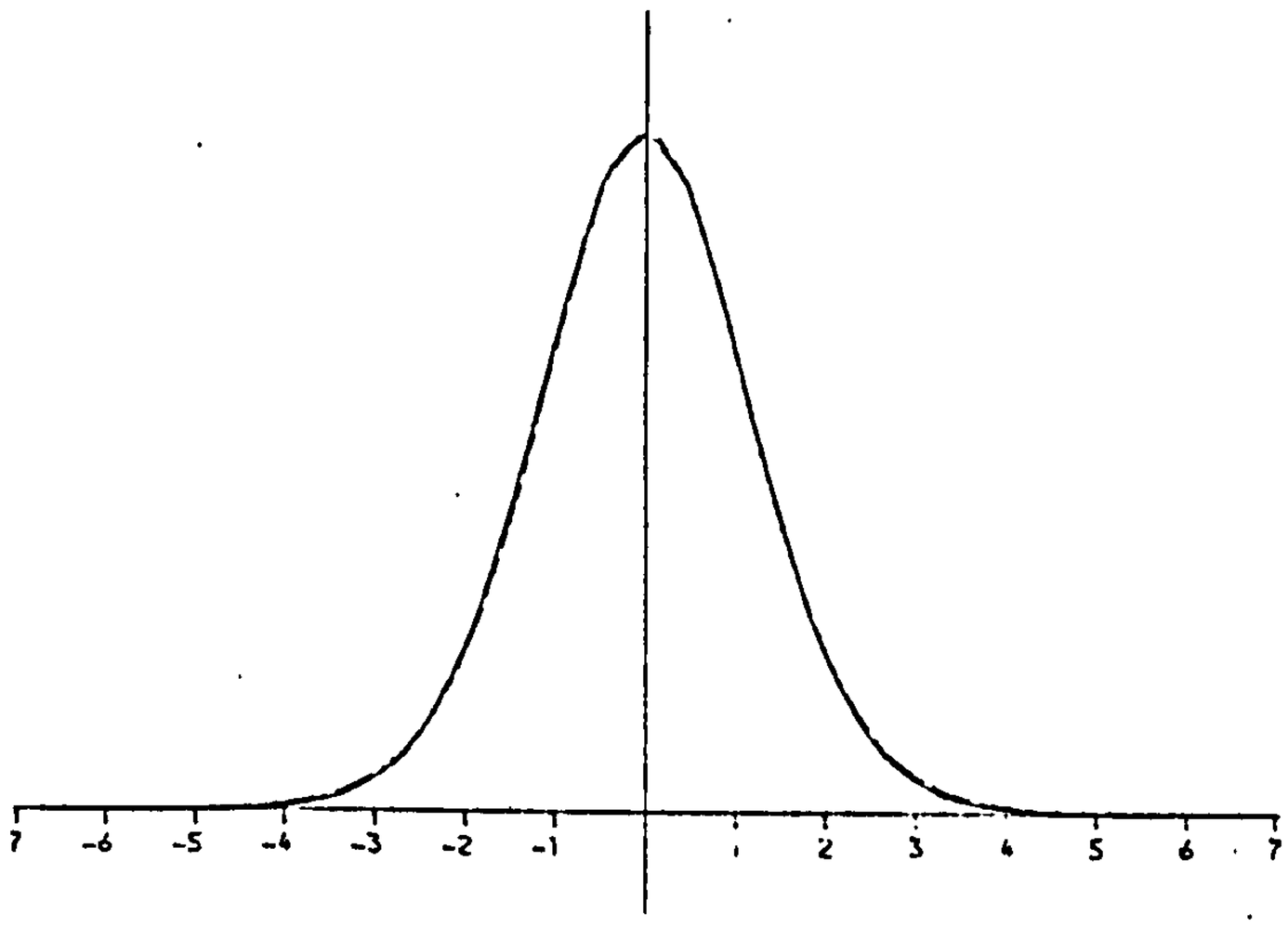
STUDENT T-15

——— MARGINAL

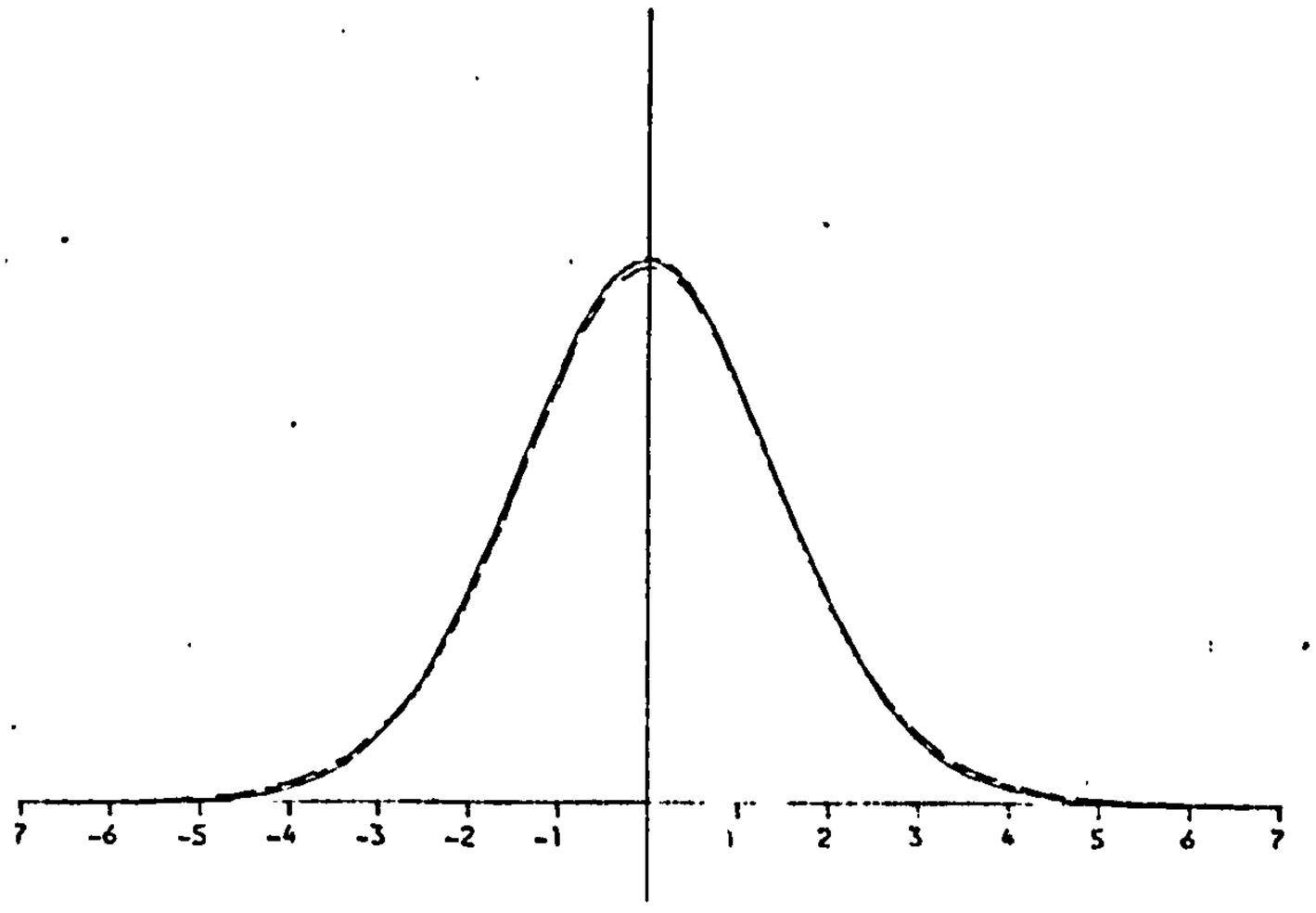
------ SCALED APPROXIMATION

– — – MODAL APPROXIMATION

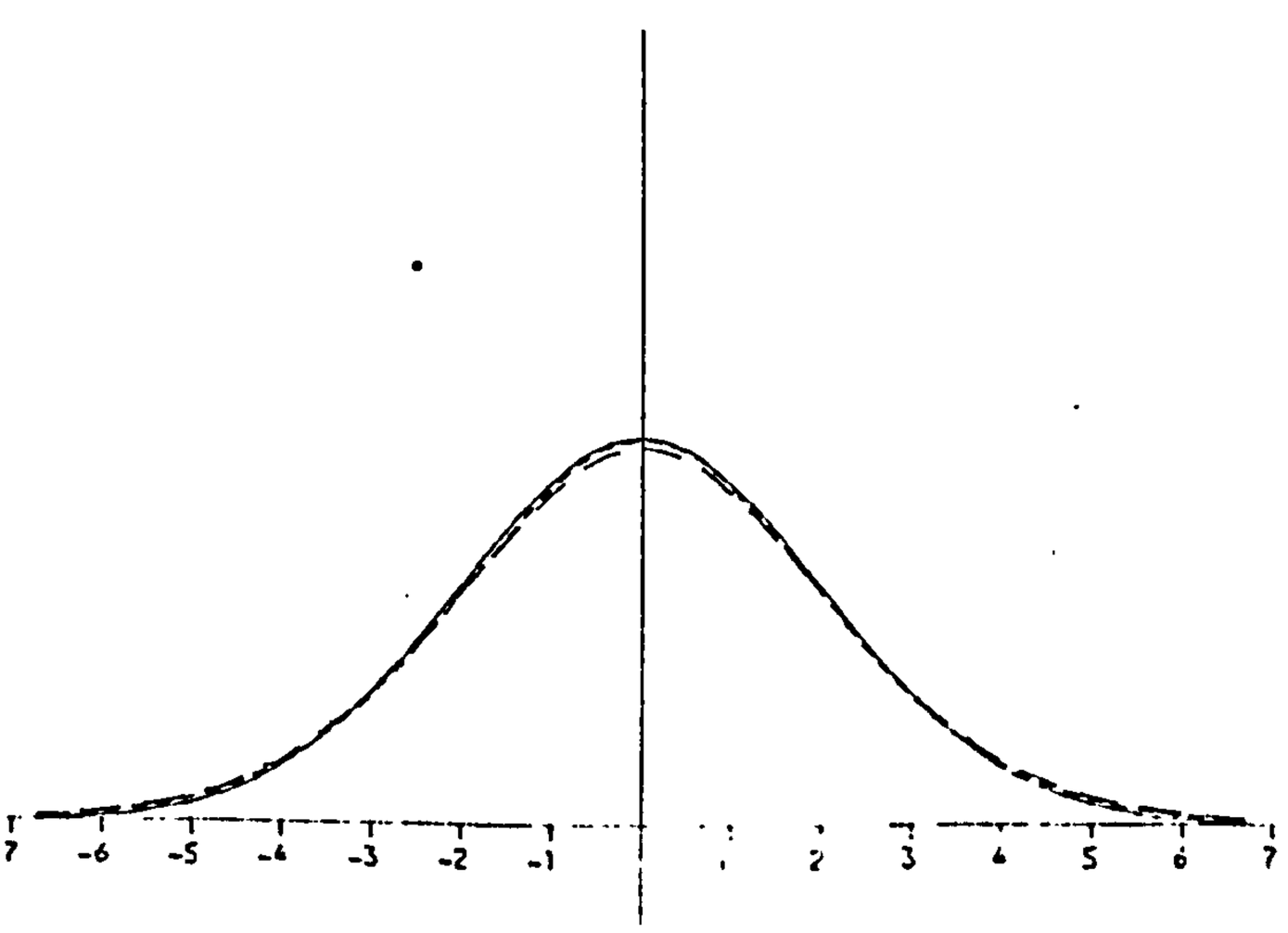
C=1/3



C=1



C=3



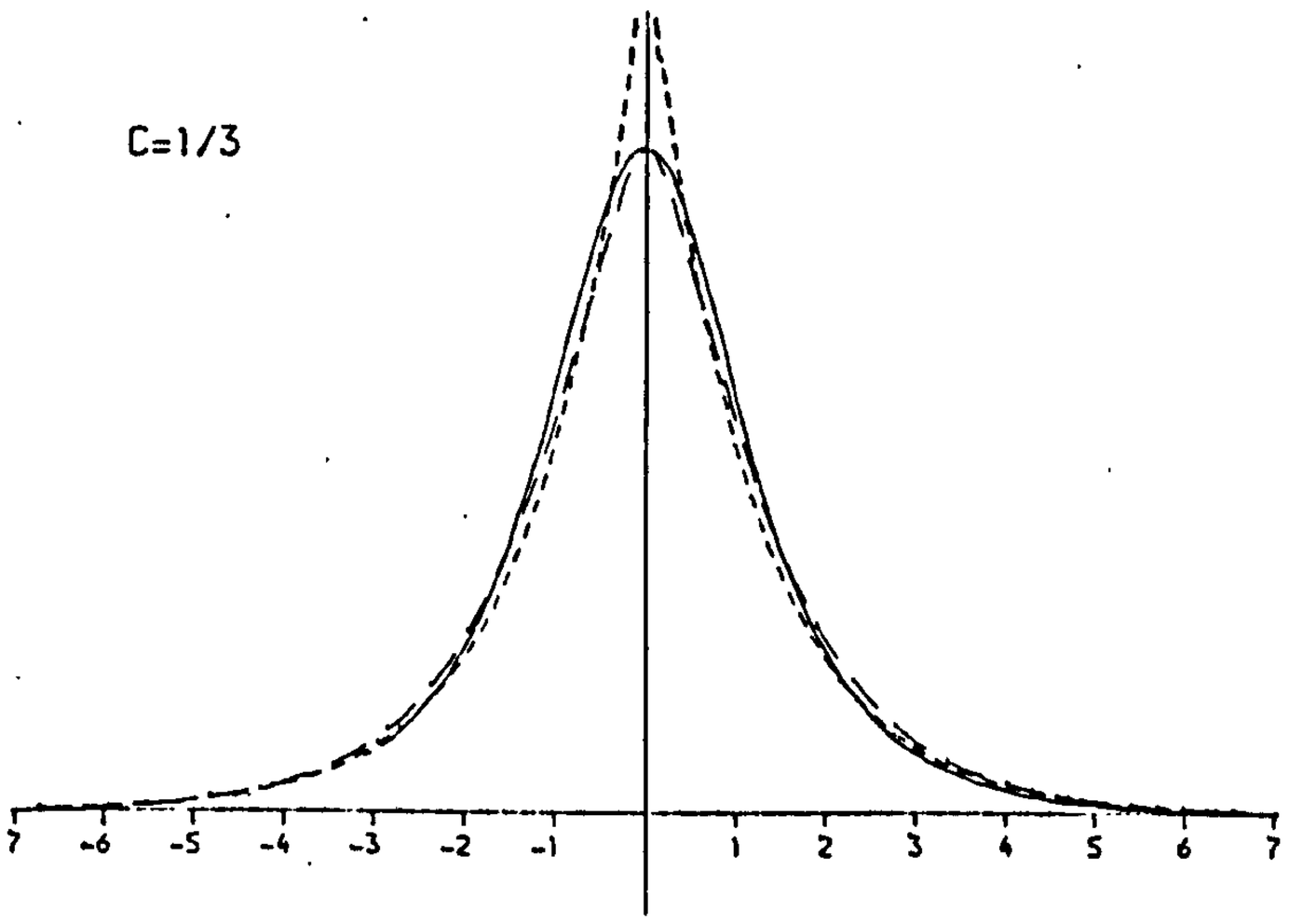
DOUBLE EXPONENTIAL

——— MARGINAL

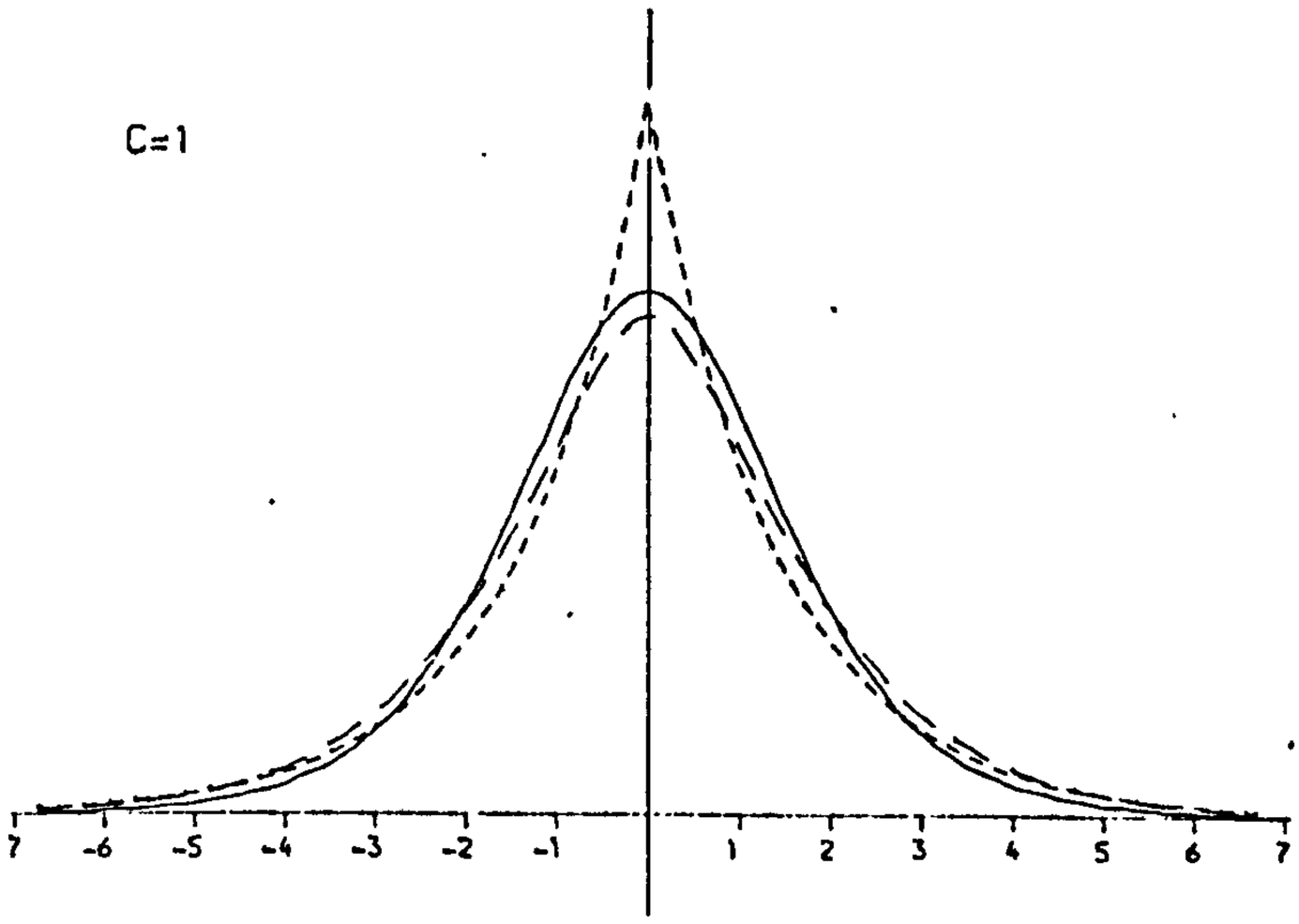
------ SCALED APPROXIMATION

– — – MODAL APPROXIMATION

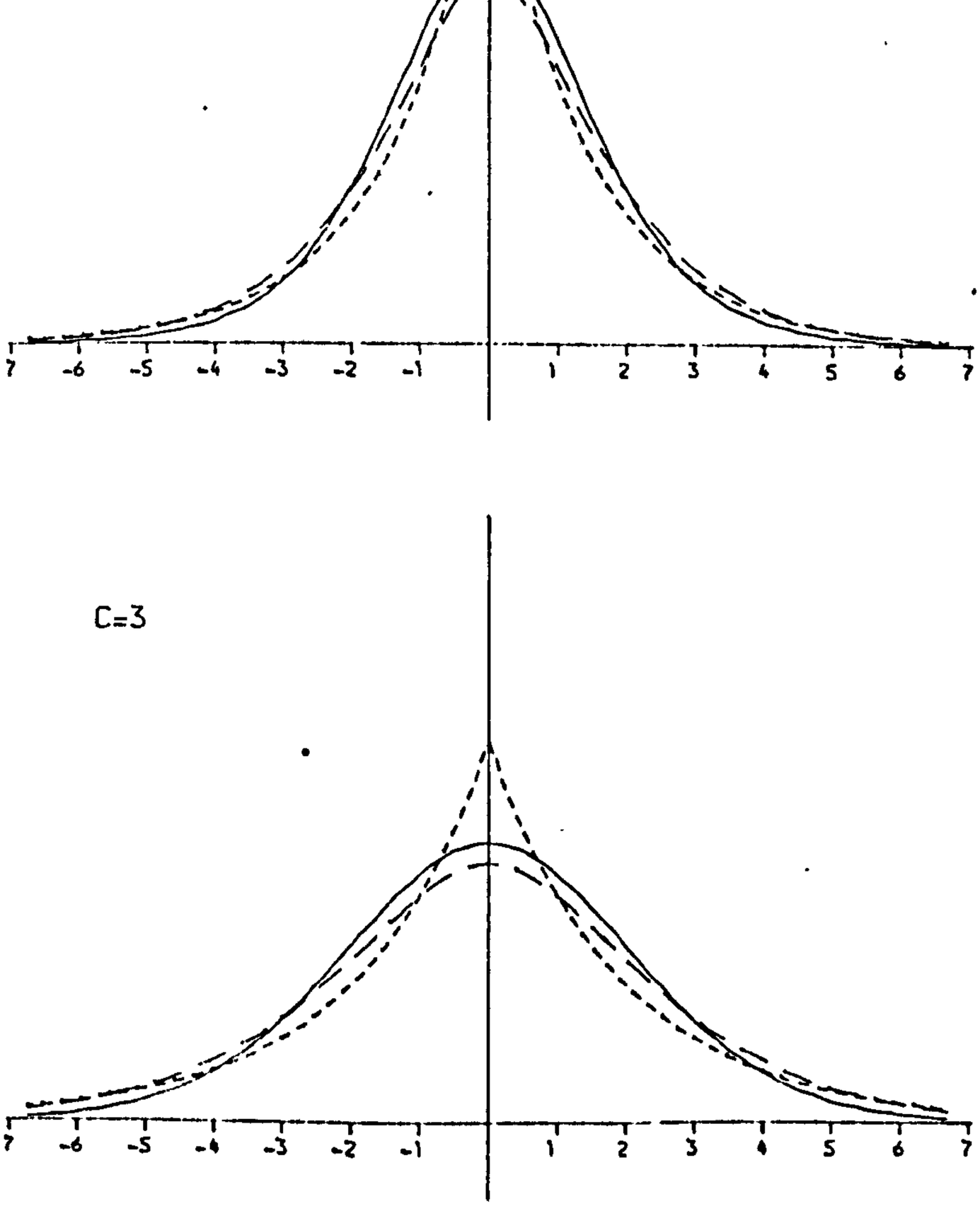
C=1/3



C=1



C=3



3.9(c)

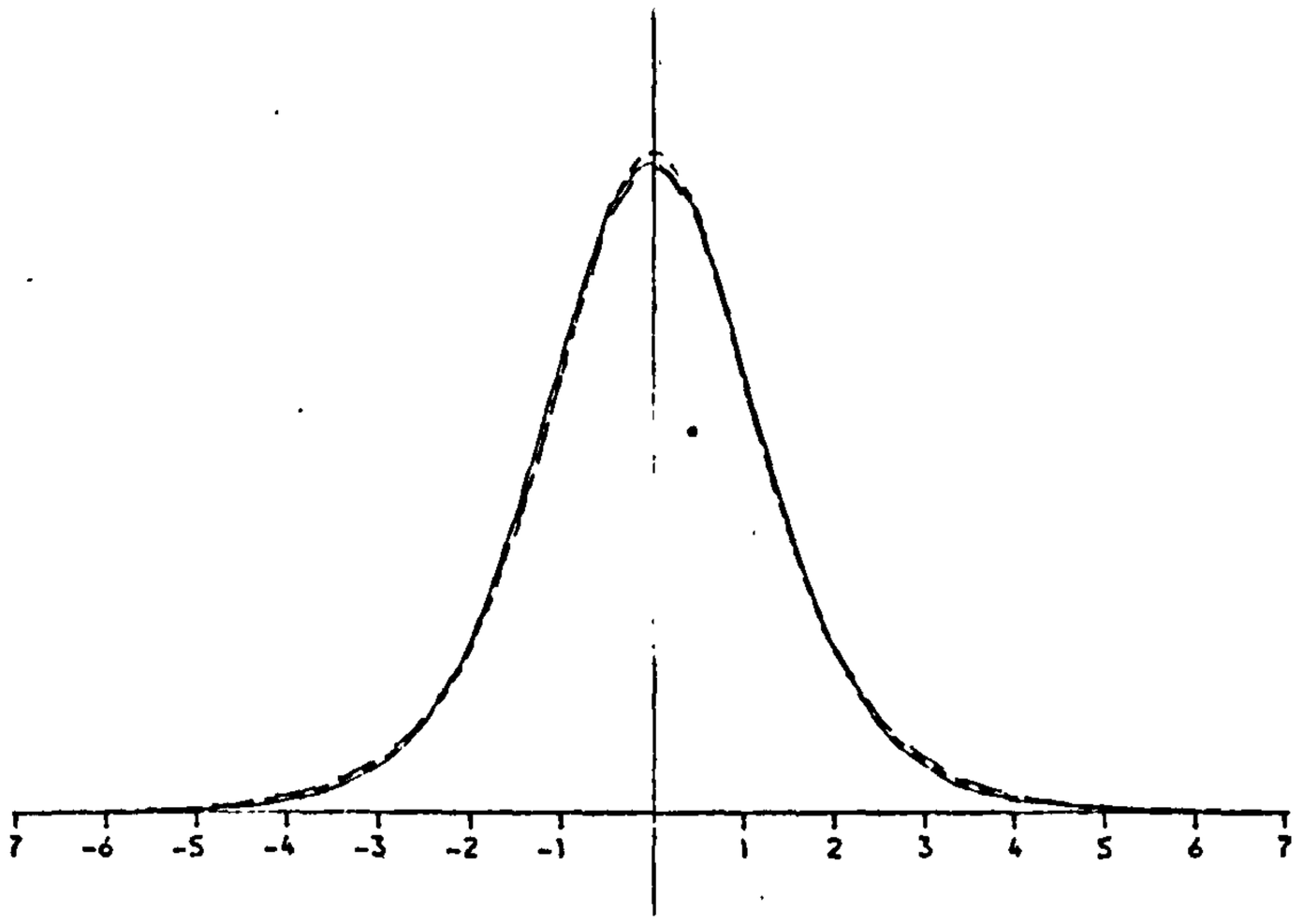
STUDENT T-S

——— MARGINAL

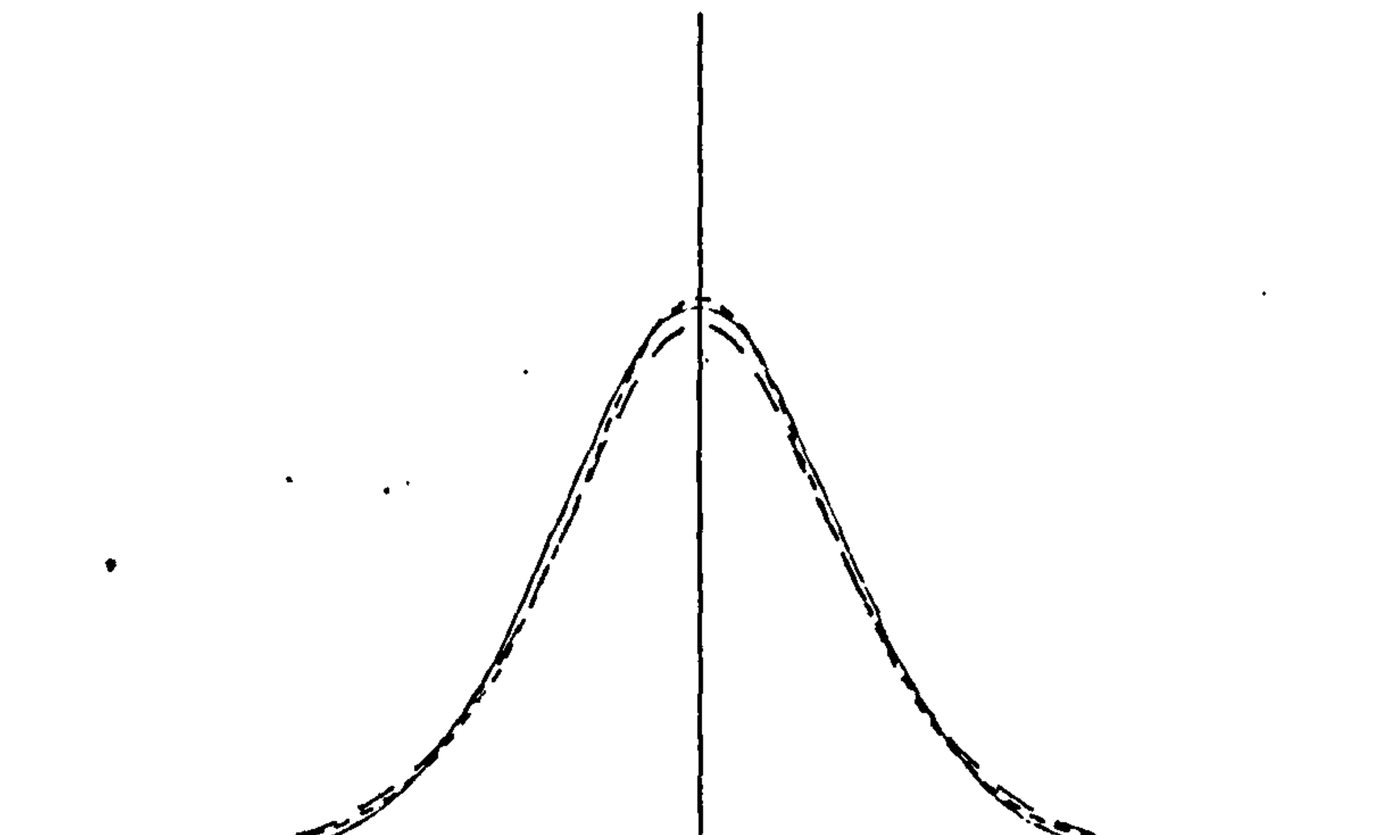
------ SCALED APPROXIMATION

– – – MODAL APPROXIMATION

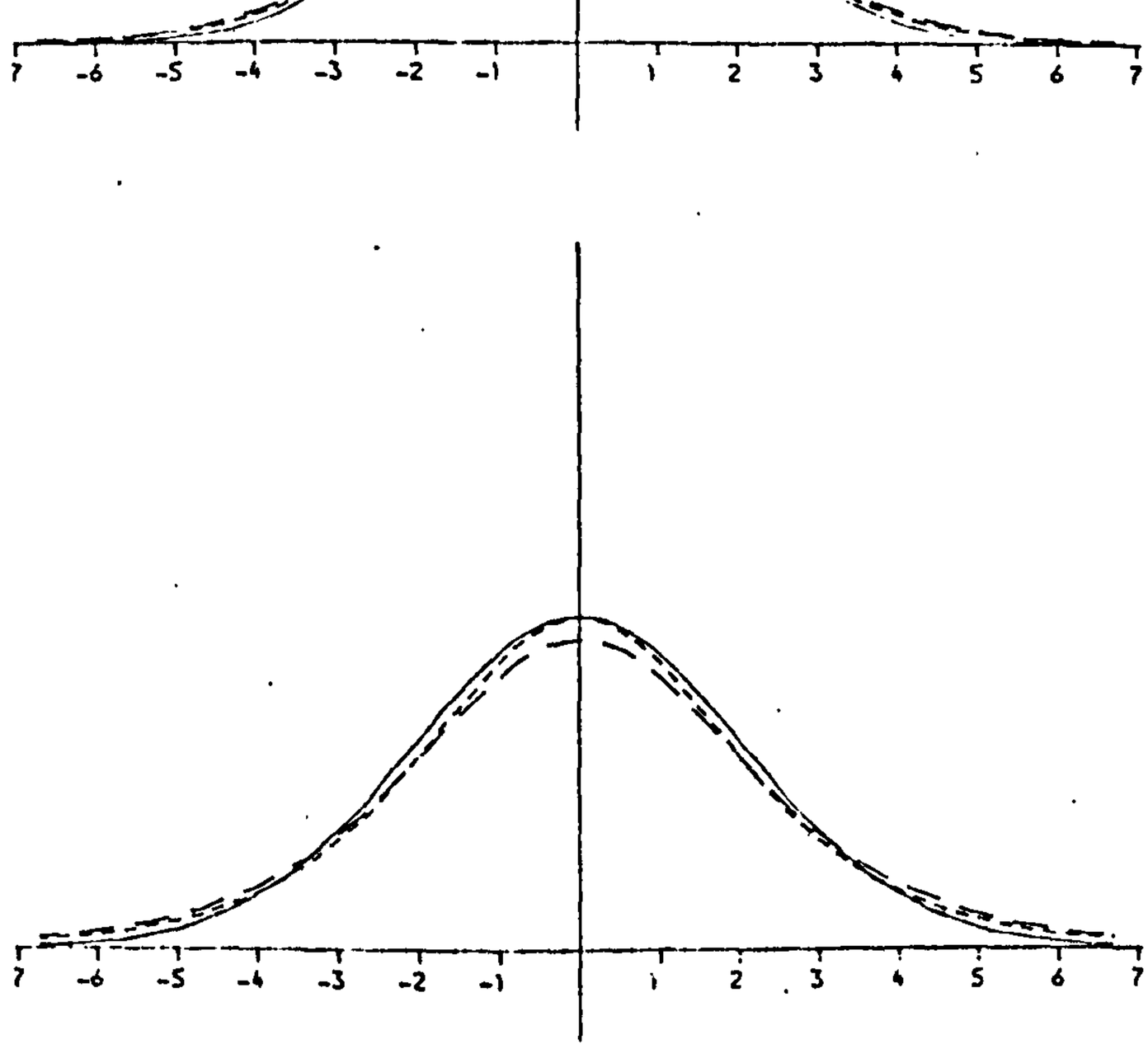
C=1/3



C=1



C=3



(ii) $\underline{p(y_{n+k}|D_n),\ k > 1.}$

Clearly, for $k \geq n$, given $\underset{\sim}{\theta}_n|D_n \sim N[m_n, C_n]$,

then

$$(\underset{\sim}{\theta}_k|D_n) \sim N\left[\underset{\sim}{t}_k^n,\ T_k^n\right]$$

where

$$\underset{\sim}{t}_k^n = G_k \underset{\sim}{t}_{k-1}^n ,$$

and

$$T_k^n = G_k T_{k-1}^n G_n^T + W_k, \qquad k=n+1,n+2,\ldots$$

Therefore the predictive density for $y_k|D_n$ is of the same form as the density of (i) above; the convolution of p_ν with a normal prior. As such the comments of (i) are appropriate.

The form of the predictor of $y_k|D_n$, $\underset{\sim}{h}_k^T \underset{\sim}{t}_k^n$, is of interest and we now consider a very special example

Special case: the scalar steady model.

For all n, $\quad y_n = \theta_n + \nu_n,$

$$\theta_n = \theta_{n-1} + \omega_n, \quad n=1,2,\ldots$$

This steady model is discussed by Harrison and Stevens (1976) as describing a slowly evolving trend θ_n. For this model, Smith (1979) examines a generalization of the normal theory to various other (exponential family) distributions for y_n, which involves discarding the linear evolution of θ_n in favour of a construction providing mathematical tractibility whilst retaining the notions of increased uncertainty and "steadiness" of the evolution $\theta_{n-1} \to \theta_n$. Smith considers the forms of the predictors of y_k, $k > n$ given D_n for such models and shows them to be exponentially weighted moving averages of past observations just as in the case of normality. In all the examples, the exponentially decaying weights are data-

independent, (essentially following from the use of the exponential family), and thus the predictors are linear in the observations. This linearity should immediately warn prospective users of the models of the sensitivity to outliers; some robustification is desirable.

Returning to the linear evolution and the use of heavy-tailed, non-normal errors, we see that in this case the predictor is

$$t_k^n = t_{k-1}^n = \ldots = m_n, \quad \text{for all } n.$$

Now

$$m_n = m_{n-1} + f_n.(y_n - m_{n-1})$$

where

$$f_n = E\left[\lambda_n p_n.(1+\lambda_n p_n)^{-1} \,|\, D_n\right]$$

directly from equation (3.2.22).

Clearly $0 \leq f_n \leq 1$. We can then rewrite m_n as

$$m_n = (1-f_n)m_{n-1} + f_n y_n$$

$$= \sum_{r=1}^{n} \alpha_r.y_r + \prod_{r=1}^{n} (1-f_r).m_0$$

where

$$\alpha_r = f_r. \prod_{j=r+1}^{n} .(1-f_j), \quad \text{for } 1 \leq r \leq n.$$

Clearly $\sum_{r=0}^{n} \alpha_r = 1$ for all n, and so, as $n \to \infty$, the steady state predictor is

$$t_k^n = m_n \simeq \sum_{r=1}^{n} \alpha_r\, y_r$$

which is a weighted moving average of past observations. However now not only do the weights α_r decay exponentially but they are data-dependent in such a way as to downweight the contribution of aberrant observations, providing robust predictors, when p_ν is

Note that if we use the modal recursions as an approximation then the same conclusions hold, with now f_n being replaced by $\tilde{f}_n$ where

$$\tilde{f}_n = \tilde{\lambda}_n p_n . (1+\tilde{\lambda}_n p_n)$$

as in (3.2.24).

3.3.2. Smoothing.

It is often of interest and importance to calculate the smoothed density $p(\underset{\sim}{\theta}_k | D_n)$, $k < n$, at time n. In particular if changes occur in the model structure we can only decide what has happened after receiving further data for confirmation. It turns out that a straightforward recursive system of equations can be derived for the smoothed densities given our assumption of approximate posterior normality at each stage (whether we use the modal algorithm or the exact method).

Martin (1979) proved this result for the scaled algorithm of Masreliez and Martin (1977). His proof, however, is unnecessarily long and misses the essential point that the approximate posterior normality is the key factor in developing the smoothed densities. We prove the result directly, and much more simply, in the following Theorem.

Theorem 3.3.1.

In the model of (3.2.1) and (3.2.2), if $p(\underset{\sim}{\theta}_n | D_n) \simeq N[\underset{\sim}{m}_n, C_n]$ for all n, then

$$p(\underset{\sim}{\theta}_k | D_n) \simeq N\left[\underset{\sim}{t}_k^n, T_k^n\right] \quad , \quad k \leq n,$$

where

$$\underset{\sim}{t}_k^n = \underset{\sim}{m}_k + C_k . G_{k+1}^T . P_{k+1}^{-1} . \left[\underset{\sim}{t}_{k+1}^n - \underset{\sim}{a}_{k+1}\right],$$

and

$$T_k^n = C_k - C_k.G_{k+1}^T.P_{k+1}^{-1}\ \left[P_{k+1}-T_{k+1}^n\right]\ P_{k+1}^{-1}G_{k+1}C_k,$$

where

$$\underset{\sim}{a}_{k+1} = G_{k+1}\ \underset{\sim}{m}_k,\ P_{k+1} = G_{k+1}\ C_k\ G_{k+1}^T + W_{k+1}.$$

Proof: For $k \leq n$, $p(\underset{\sim}{\theta}_k|D_n) = \int_{\mathbb{R}^p} p(\underset{\sim}{\theta}_k|\underset{\sim}{\theta}_{k+1},D_n).p(\underset{\sim}{\theta}_{k+1}|D_n)d\underset{\sim}{\theta}_{k+1}$, (1)

Thus (1) defines $p(\underset{\sim}{\theta}_k|D_n)$ recursively. Setting k=n-1, we have

$$p(\underset{\sim}{\theta}_{n-1}|\underset{\sim}{\theta}_n,D_n) = p(\underset{\sim}{\theta}_{n-1}|\underset{\sim}{\theta}_n,D_{n-1})p(y_n|\underset{\sim}{\theta}_{n-1},\underset{\sim}{\theta}_n,D_{n-1}).p(y_n|\underset{\sim}{\theta}_n,D_{n-1})^{-1}. \quad (2)$$

The first term of (2) is, by Bayes Theorem,

$$p(\underset{\sim}{\theta}_{n-1}|\underset{\sim}{\theta}_n,D_{n-1}) = p(\underset{\sim}{\theta}_{n-1}|D_{n-1}).p(\underset{\sim}{\theta}_n|\underset{\sim}{\theta}_{n-1},D_{n-1}).p(\underset{\sim}{\theta}_n|D_{n-1})^{-1}$$

so $(\underset{\sim}{\theta}_{n-1}|\underset{\sim}{\theta}_n,D_{n-1}) \sim N_{\underset{\sim}{\theta}_{n-1}}\left[\underset{\sim}{m}_{n-1},C_{n-1}\right]\ N_{\underset{\sim}{\theta}_n}\left[G_n\underset{\sim}{\theta}_{n-1},W_n\right]$

$$= N_{\underset{\sim}{\theta}_{n-1}}\left[\underset{\sim}{x}_{n-1},R_{n-1}\right],$$

where $\underset{\sim}{x}_{n-1} = \underset{\sim}{m}_{n-1} + C_{n-1}\ G_n^T.P_n^{-1}(\underset{\sim}{\theta}_n - \underset{\sim}{a}_n)$,

and

$$R_{n-1} = C_{n-1} - C_{n-1}\ G_n^T\ P_n^{-1}\ G_n\ C_{n-1}.$$

For the second term in (2) note that

$$p(y_n|\underset{\sim}{\theta}_{n-1},\underset{\sim}{\theta}_n,D_{n-1}) = p_\nu(y_n - h_{\underset{\sim}{n}\underset{\sim}{n}}^T\underset{\sim}{\theta}_n), \text{ not involving } \underset{\sim}{\theta}_{n-1},$$

and so (2) becomes

$$(\underset{\sim}{\theta}_{n-1}|\underset{\sim}{\theta}_n,D_n) \sim N_{\underset{\sim}{\theta}_{n-1}}\left[\underset{\sim}{x}_{n-1},\ R_{n-1}\right]. \quad (3)$$

Hence, from (1)

$$(\underset{\sim}{\theta}_{n-1}|D_n) \sim \int_{\mathbb{R}^p} N_{\underset{\sim}{\theta}_{n-1}}\left[\underset{\sim}{x}_{n-1},R_{n-1}\right].\ N_{\underset{\sim}{\theta}_n}\left[\underset{\sim}{m}_n,C_n\right]d\underset{\sim}{\theta}_n$$

$$= N_{\underset{\sim}{\theta}_{n-1}}\left[t_{n-1}^n,\ T_{n-1}^n\right],$$

where $\underset{\sim}{t}^{n}_{n-1} = E\left[\underset{\sim}{x}_{n-1} | D_n\right] = \underset{\sim}{m}_{n-1} + C_{n-1}\, G_n^T P_n^{-1} . \left[\underset{\sim}{m}_n - \underset{\sim}{a}_n\right]$,

and, similarly, $T^n_{n-1} = E\left[R_{n-1} + \left(\underset{\sim}{x}_{n-1} - \underset{\sim}{t}^n_k\right)\left(\underset{\sim}{x}_{n-1} - \underset{\sim}{t}^n_k\right)^T | D_n\right]$,

so

$$T^n_{n-1} = R_{n-1} + C_{n-1}\, G_n^T\, P_n^{-1}\, C_n\, P_n^{-1}\, G_n\, C_{n-1},$$

$$= C_{n-1} - C_{n-1}\, G_n^T\, P_n^{-1} \left[P_n - C_n\right] P_n^{-1} G_n\, C_{n-1}.$$

Noting that $\underset{\sim}{t}^n_n = \underset{\sim}{m}_n$, $T_n{}^n = C_n$, the result stated holds for $k = n-1$.

We prove the general result by induction.

Assume that for some $k < n$, the stated result holds. Then the first term in the integrand of (1) is calculated as for the case $k=n-1$.

Set $D_{n,k} = (y_{k+1}, \ldots, y_n)$. Then, by Bayes Theorem,

$$p(\underset{\sim}{\theta}_k | \underset{\sim}{\theta}_{k+1}, D_n) = p(\underset{\sim}{\theta}_k | \underset{\sim}{\theta}_{k+1}, D_k) . p(D_{n,k} | \underset{\sim}{\theta}_k, \underset{\sim}{\theta}_{k+1}, D_k) p(D_{n,k} | \underset{\sim}{\theta}_{k+1}, D_k)^{-1}$$

But, given $\underset{\sim}{\theta}_{k+1}$, $D_{n,k}$ is independent of $\underset{\sim}{\theta}_k$ and the final two terms cancel. The first is, by Bayes Theorem,

$$p(\underset{\sim}{\theta}_k | \underset{\sim}{\theta}_{k+1}, D_k) \;\alpha\; p(\underset{\sim}{\theta}_k | D_k) p(\underset{\sim}{\theta}_{k+1} | \underset{\sim}{\theta}_k, D_k)$$

which, as in (3) above, is just $N_{\underset{\sim}{\theta}_k}\left[\underset{\sim}{x}_k, R_k\right]$, (4)

where

$$\underset{\sim}{x}_k = \underset{\sim}{m}_k + C_k .\, G^T_{n+1} . P^{-1}_{n+1} . (\underset{\sim}{\theta}_{k+1} - \underset{\sim}{a}_{k+1}),$$

and

$$R_k = C_k - C_k . G^T_{k+1} P^{-1}_{k+1} . G_{k+1} . C_k .$$

Returning now to (1) we have

$$(\underset{\sim}{\theta}_k | D_n) \sim \int_{\mathbb{R}^p} N_{\underset{\sim}{\theta}_k}\left[\underset{\sim}{x}_k, R_k\right] . N_{\underset{\sim}{\theta}_{k+1}}\left[\underset{\sim}{t}^n_{k+1},\; T^n_{k+1}\right] d\underset{\sim}{\theta}_{k+1}$$

$$= N_{\underset{\sim}{\theta}_k}\left[\underset{\sim}{t}^n_k,\; T_k{}^n\right],$$

where

$$\underset{\sim}{t}_k^n = E[\underset{\sim}{x}_k | D_n] = \underset{\sim}{m}_k + C_k\, G_{k+1}^T\, P_{k+1}^{-1}.(\underset{\sim}{t}_{k+1}^n - \underset{\sim}{a}_{k+1}),$$

and

$$T_k^n = E[R_k + (\underset{\sim}{x}_k - \underset{\sim}{t}_k^n)(\underset{\sim}{x}_k - \underset{\sim}{t}_k^n)^T | D_n],$$

$$= R_k + C_k\, G_{k+1}^T\, P_{k+1}^{-1}\, T_{k+1}^n\, P_{k+1}^{-1}\, G_{k+1}\, C_k,$$

$$= C_k = C_k\, G_{k+1}^T\, P_{k+1}^{-1}\, [P_{k+1} - T_{k+1}^n] P_{k+1}^{-1}\, G_{k+1}\, C_k.$$

The proof now follows by induction. □

So Theorem 3.3.1 provides the means of recursively calculating the smoothed densities. Notice that the recursions are the same as those of the normal model, but now robust. In practice only one or two steps back are required i.e. calculation of $p(\underset{\sim}{\theta}_{n-1}|D_n)$ and $p(\underset{\sim}{\theta}_{n-2}|D_n)$, the Markov nature of the model means that information about $\underset{\sim}{\theta}_k$, $k < n$, carried in Y_n gets "weaker" as $|n-k|$ grows and so $p(\underset{\sim}{\theta}_k|D_n)$ will be little different from $p(\underset{\sim}{\theta}_k|D_{n-1})$ as $|n-k| \to \infty$.

To see this in the special case of the steady scalar model set $p = 1$, $G_n = 1 = \underset{\sim}{h}_n$ for all n, and $w_n = \omega$. Then by the Theorem

$$|T_k^n - T_k^{n-1}| = \left(\frac{C_n}{C_n+\omega}\right)^2 \left|T_{k+1}^n - T_{k+1}^{n-1}\right|$$

$$= \left(\frac{C_n}{C_n+\omega}\right)^{2(n-k)} . \omega \;, \to 0 \text{ as } |n-k| \to \infty.$$

3.4. Vector Observations.

Until now we have concentrated for sake of clarity of exposition, on scalar observations. We now return to the general vector observations model of §3.1, and extend the ideas of this Chapter to this general case. We note that the development is parallel to that for scalar observations and so we content ourselves with a simple presentation.

The density p_ν of $\underset{\sim}{u}_n$ is now a multivariate density for which we assume the following:-

(i) $p_\nu(\underset{\sim}{u})$ is continuous and positive for $\underset{\sim}{u} \in \mathbb{R}^n$;

(ii) $p_\nu(\underset{\sim}{u})$ twice piecewise differentiable in $\underset{\sim}{u}$;

(iii) $p_\nu(\underset{\sim}{u})$ unimodal at zero and spherically symmetric;

(iv) $p_\nu(\underset{\sim}{u})$ heavy tailed relative to the standard multivariate normal density in the sense that

$$\left[\frac{\partial}{\partial u_i} \ell n\ p_\nu(\underset{\sim}{u})\right] \lesssim k u_i, \quad k > 0,\ u_i \in \mathbb{R}.$$

as a function of u_i, $i=1,\ldots,m$.

The multivariate form of Masreliez's result, Theorem 3.2.1, is applicable when $(\underset{\sim}{\theta}_n | D_{n-1}) \sim N[\underset{\sim}{a}_n, P_n]$.

Define the marginal density p by

$$p(\underset{\sim}{y}_n) = p(\underset{\sim}{y}_n - H_n \underset{\sim}{a}_n) = \int_{\mathbb{R}^p} p_\nu(\underset{\sim}{y}_n - H_n \underset{\sim}{\theta}_n).N[\underset{\sim}{a}_n, P_n]\, d\underset{\sim}{\theta}_n$$

and its score function $\underset{\sim}{g}$ by

$$\underset{\sim}{g}(\underset{\sim}{u}_n) = -\frac{\partial}{\partial \underset{\sim}{Y}_n} \ell n\ p(\underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n), \quad \underset{\sim}{u}_n = \underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n.$$

Further, define the information matrix G by

$$G(\underset{\sim}{u}_n) = \frac{\partial}{\partial \underset{\sim}{Y}_n}^T \underset{\sim}{g}(\underset{\sim}{u}_n).$$

Then

$$E[\underset{\sim}{\theta}_n | D_n] = \underset{\sim}{a}_n + P_n H_n^T . g(\underset{\sim}{u}_n)$$

and

$$\mathrm{Var}[\underset{\sim}{\theta}_n | D_n] = P_n - P_n H_n^T G(\underset{\sim}{u}_n) H_n P_n.$$

The score $\underset{\sim}{g}$ has ith component

$$g_i(\underset{\sim}{u}_n) = -\frac{\partial}{\partial Y_i} \ell n\ p(\underset{\sim}{u}_n), \quad i=1,\ldots,m$$

each of which mirrors the behaviour of the likelihood score in the same way as in the scalar model of §3.2.3. Now the spherical symmetry of p_ν, that is

$$p_\nu(\underset{\sim}{u}) = f(\underset{\sim}{u}^T\underset{\sim}{u}/2), \quad \underset{\sim}{u} \in \mathbb{R}^m,$$

implies that

$$\underset{\sim}{g}_\nu(\underset{\sim}{u}) = \psi_\nu(\underset{\sim}{u})\ \underset{\sim}{u}, \text{ say}$$

where $\psi_\nu(\underset{\sim}{u})$ is a function of $\underset{\sim}{u}^T\underset{\sim}{u}$ given by

$$\psi_\nu(\underset{\sim}{u}) = -f'(\underset{\sim}{u}^T\underset{\sim}{u}/2)/f(\underset{\sim}{u}^T\underset{\sim}{u}/2).$$

This decomposition of g_ν leads immediately to the multivariate analogues of the modal recursions, as follows;

$\underset{\sim}{m}_n = E[\underset{\sim}{\theta}_n | D_n]$ is given approximately by

$$\underset{\sim}{m}_n \simeq \underset{\sim}{a}_n + P_n H_n^T\left[I+\psi_\nu(\underset{\sim}{u}_n).H_n P_n H_n^T\right]^{-1} \underset{\sim}{g}_\nu(\underset{\sim}{u}_n). \qquad (3.4.1)$$

Identifying this with Masreliez's result we have

$$g(\underset{\sim}{u}_n) \simeq \left[I+\psi_\nu(\underset{\sim}{u}_n).H_n P_n H_n^T\right]^{-1} g_\nu(\underset{\sim}{u}_n). \qquad (3.4.2)$$

Hence $G(\underset{\sim}{u}_n) = \dfrac{\partial}{\partial \underset{\sim}{u}_n^T} g(\underset{\sim}{u}_n)$ can be calculated as follows:

Set $\phi(\underset{\sim}{u}) = \psi_\nu^{-1}(\underset{\sim}{u})$ and $Q_n = H_n P_n H_n^T$. We have

$$g(\underset{\sim}{u}_n) = \left[I\phi(\underset{\sim}{u}_n)+Q_n\right]^{-1}\underset{\sim}{u}_n$$

or

$$\phi(\underset{\sim}{u})\ g(\underset{\sim}{u}_n) + Q_n\ g(\underset{\sim}{u}_n) = \underset{\sim}{u}_n$$

and, differentiating with respect to $\underset{\sim}{u}_n$, we obtain,

$$G(\underset{\sim}{u}_n).\left[I\phi(\underset{\sim}{u}_n)+Q_n\right] = I - g(\underset{\sim}{u}_n)\ \frac{\partial\phi(\underset{\sim}{u}_n)}{\partial \underset{\sim}{u}_n^T}$$

Let $R_n = \left[I\phi(\underset{\sim}{u}_n)+Q_n\right]^{-1}$.

Then

$$G(\underset{\sim}{u}_n) = \left(I - R_n \underset{\sim}{u}_n \frac{\partial\phi(\underset{\sim}{u}_n)}{\partial \underset{\sim}{u}^T_n}\right) R_n$$

$$= R_n\left[R_n^{-1} - \underset{\sim}{u}_n \frac{\partial\phi(\underset{\sim}{u}_n)}{\partial \underset{\sim}{u}^T_n}\right] R_n \qquad (3.4.3)$$

The modal recursion for C_n is then

$$C_n \simeq P_n - P_n H_n^T G(\underset{\sim}{u}_n) H_n P_n \qquad (3.4.4)$$

with $G(\underset{\sim}{u}_n)$ given by (3.4.3).

Example 3.4.1.

Let p_ν be multivariate Student t-k,

$$p_\nu(\underset{\sim}{u}) \propto \left[k+\underset{\sim}{u}^T\underset{\sim}{u}\right]^{-(k+m)/2}$$

Then $\phi(\underset{\sim}{u}) = \psi_\nu(\underset{\sim}{u})^{-1} = (k+m)^{-1}(k+\underset{\sim}{u}^T\underset{\sim}{u})$, leading to

$$\frac{\partial\phi(\underset{\sim}{u})}{\partial \underset{\sim}{u}^T} = (k+m)^{-1}\underset{\sim}{u}^T.$$

(3.4.3) then becomes

$$G(\underset{\sim}{u}_n) = \left[I+\psi_\nu(\underset{\sim}{u}_n)Q_n\right]^{-1}\left\{I - \frac{\underset{\sim}{u}_n\underset{\sim}{u}_n^T}{\left[k+\underset{\sim}{u}_n^T\underset{\sim}{u}_n\right]}\left[I+\psi_\nu(\underset{\sim}{u}_n)Q_n\right]^{-1}\right\}\psi_\nu(\underset{\sim}{u}_n) \qquad (3.4.5)$$

Again it is clear that, when p_ν is actually a scale mixture of normals,

$$p_\nu(\underset{\sim}{u}) = \int_0^\infty N_\lambda\left[\underset{\sim}{0}, I\lambda^{-1}\right]\omega(\lambda)d\lambda,$$

then $\psi_\nu(\underset{\sim}{u}_n)$ plays the role of an estimate of λ_n; as in §3.2.5,

$$\psi_\nu(\underset{\sim}{u}_n) = E\left[\lambda_n \middle| D_n, \underset{\sim}{\theta}_n = \underset{\sim}{a}_n\right]$$

Further comment on this development, and on the exact analysis v scale mixtures, imitates §3.2 exactly and so we persue it no further.

CHAPTER 4. The D.L.M.; scale problems.

4.1 Introduction

We now turn to the estimation of unknown scale parameters and covariance matrices of the error distributions in the models of Chapter 3. We assumed throughout that the scale parameters of observational error densities were known and unity in the case of scalar observations and that multivariate observations had spherically symmetric error distributions. In general this will not be the case and successful implementation of filtering algorithms will depend upon effective estimation of scale parameters and matrices of elliptically symmetric error densities.

In §4.3 we concentrate on the scalar model of §3.2 and examine features of the posterior distributions for scale parameters of heavy-tailed densities. In the simpler context of location estimation rather than time-series, a variety of classical procedures have been proposed for scale estimation. In particular scale analogues of robust M-estimators have been found useful, as in Huber (1973) and (1978), and these approximate the coherent solution in that maximum likelihood can be viewed as an approximation to a posterior modal solution in some cases. In the more complex time-series problems, Martin (1979) and (1980) utilizes such ideas to develop scale estimates for use in his recursive filtering algorithms discussed in §3.2.1. In particular the latter reference provides a discussion of two possible methods.

The first is to calculate a time invariant estimate of the constant scale parameter via an approximate maximum likelihood approach similar to that outlined in Huber (1973). This then provides a global scale estimate and of course demands a retrospective analysis i.e. must be computed off-line. The second suggestion is to estimate the scale parameter sequentially by an auxilliary data-dependent

recursion. This scheme, preferred by Martin, is intuitively appealing and the success of Martins' algorithms bears out the usefulness of such an approach. Unfortunately the mentioned auxilliary recursion is not given in that reference. Also, as always, such a scheme suffers from a lack of formal justification and, more practically, distributional results for such estimators are not provided and so little feel for uncertainty involved with a point estimate of scale can be obtained.

We approach the problem within a coherent framework and the main difficulty lies in obtaining tractible forms for posterior distributions of scale parameters. In §4.3 we use ideas of Chapter 3 to develop approximations to the formal Bayesian analysis of unknown scale parameters in the scalar observations model. In §4.4 we develop this in the multivariate case. Here we encounter further problems of tractibility, for even in the case of the usual normal linear model, (without the added complications of the time series model), there is no tractible conjugate analysis when both the regression vector and the covariance matrix are unknown when, in addition, we have proper priors for these two a priori independent parameters. In order to surmount these problems in the non-normal case, we consider first the analysis in the normal model to obtain some idea of the sort of approach that might be appropriate. This is the subject matter of the next section.

4.2 The normal model: unknown covariance structure.

4.2.1 Unknown covariance scale parameter.

We take the model of §3.1,

$$\underset{\sim}{Y}_n = H_n \underset{\sim}{\theta}_n + \underset{\sim}{\nu}_n , \qquad (4.2.1)$$

$$\underset{\sim}{\theta}_n = G_n \underset{\sim}{\theta}_{n-1} + \underset{\sim}{\omega}_n , \quad n=1,2,... \qquad (4.2.2)$$

with the assumptions made in Chapter 3. The difference now is that the $\{\underset{\sim}{\nu}_n\}$ are normally distributed.

The most general assumption that we can make about the covariance structure of $\underset{\sim}{\nu}_n$ whilst retaining a tractible (conjugate) sequential analysis is as follows;

$$\text{Let} \quad (\underset{\sim}{\nu}_n|\lambda) \sim N[\underset{\sim}{0}, \lambda^{-1}V_n], \quad n=1,2,\ldots \tag{4.2.3}$$

where V_n is a known (m×m) covariance matrix and λ is an unknown scalar parameter. The prior to posterior analysis for $\underset{\sim}{\theta}_n$ and λ will now follow the usual conjugate theory of, for example, De Groot (1970, §9.10), if the prior covariance matrix of $\underset{\sim}{\theta}_n|D_{n-1}$ is also scaled by λ

$$\text{i.e.} \qquad (\underset{\sim}{\theta}_n|D_{n-1},\lambda) \sim N[\underset{\sim}{a}_n, \lambda^{-1}P_n] \tag{4.2.4}$$

and if, in addition, the prior for $\lambda|D_{n-1}$ is a Gamma distribution

$$(\lambda|D_{n-1}) \sim G[\alpha_{n-1}/2, \beta_{n-1}/2], \tag{4.2.5}$$

with α_{n-1}, β_{n-1} both positive.

With these assumptions, the posterior distribution is of the same normal/gamma form, given by

$$(\underset{\sim}{\theta}_n|D_n,\lambda) \sim N[\underset{\sim}{m}_n, \lambda^{-1}C_n] \tag{4.2.6}$$

and, defining $R_n = \text{Var}[\underset{\sim}{Y}_n|D_{n-1},\lambda].\lambda = H_nP_nH_n^T + V_n$,

$$(\lambda|D_n) \sim G[\alpha_n/2, \beta_n/2]$$

where $\underset{\sim}{m}_n$, C_n are given by the usual Kalman filter recursions and

$$\alpha_n = \alpha_{n-1} + 1,$$

$$\beta_n = \beta_{n-1} + (\underset{\sim}{Y}_n - H_n\underset{\sim}{a}_n)^T R_n^{-1}(\underset{\sim}{Y}_n - H_n\underset{\sim}{a}_n). \tag{4.2.7}$$

Clearly then, if (4.2.4) is to hold for all n, the covariance matrix of $\underset{\sim}{W}_n$ must also be scaled by λ.

i.e. $$P_n = G_n C_{n-1} G_n^T + W_n$$

and $$\text{var}\left[\underset{\sim}{\theta}_n | D_{n-1}\right] = \lambda^{-1} P_n \text{ if } \text{var}\left[\underset{\sim}{W}_n\right] = \lambda^{-1} W_n.$$

The system described above is essentially a generalization of the static linear model as in De Groot. The marginal posterior distribution for $\underset{\sim}{\theta}_n$ and the marginal for $\underset{\sim}{y}_n$ are available as multivariate t distributions

$$p(\underset{\sim}{\theta}_n | D_n) \propto \left[\beta_n + (\underset{\sim}{\theta}_n - \underset{\sim}{m}_n)^T C_n^{-1} (\underset{\sim}{\theta}_n - \underset{\sim}{m}_n)\right]^{-(\alpha_n+1)/2}$$

and

$$p(\underset{\sim}{Y}_n | D_{n-1}) \propto \beta_n^{-\alpha_n/2}$$

with β_n given by (4.2.7).

The predictive distributions for $y_{n+k}|D_n$, k=1,2,..., are also available as t distributions in the usual way.

This analysis provides an extremely useful method of learning the scaling of errors in the dynamic linear model and, indeed, is used extensively in Chapter 7 in a practical problem. There is little more to be said about this case and we turn to the (intractible) problem of an unknown covariance matrix in the normal model.

4.2.2. Unknown covariance matrix.

Retaining the model (4.2.1) and (4.2.2), we now assume that the covariance matrix of $\underset{\sim}{\nu}_n$ is unknown. Generally

$$\underset{\sim}{\nu}_n \sim N\left[\underset{\sim}{0}, \lambda_n^{-1} \Lambda^{-1}\right]$$

where both λ_n and Λ are unknown. This model is an important part of the robust estimation of $\underset{\sim}{\theta}_n$ and Λ in §4.4. For this section we

assume the λ_n known and equal to unity,

$$\underset{\sim}{\nu}_n \sim N[\underset{\sim}{0}, \Lambda^{-1}], \quad n=1,2,\ldots \quad .$$

When $\underset{\sim}{\theta}_n$ is known, the conjugate analysis is obtained by adopting a Wishart distribution for Λ (De Groot (1970), as follows; $(\Lambda|D_{n-1}) \sim W[\alpha_{n-1}, V_{n-1}]$ given by

$$p(\Lambda|D_{n-1}) \propto |\Lambda|^{(\alpha_{n-1}-m-1)/2} \exp\{-\tfrac{1}{2}\mathrm{tr}(V_{n-1}\Lambda)\}$$

where tr is the trace function. Then $E[\Lambda|D_{n-1}] = \alpha_{n-1} V_{n-1}^{-1}$. Defining $\underset{\sim}{u}_n = \underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n$, we have the likelihood, for known $\underset{\sim}{\theta}_n$, given by

$$(\underset{\sim}{Y}_n|\underset{\sim}{\theta}_n,\Lambda) \sim N[H_n\underset{\sim}{\theta}_n,\Lambda^{-1}]$$

and so

$$(\Lambda|\underset{\sim}{\theta}_n,D_n) \sim W[\alpha_n,V_n]$$

where

$$\alpha_n = \alpha_{n-1} + 1, \tag{4.2.8}$$

and

$$V_n = V_{n-1} + \underset{\sim}{u}_n.\underset{\sim}{u}_n^T. \tag{4.2.9}$$

The problems arise since $\underset{\sim}{\theta}_n$ is not known. For the full model we have a complex posterior distribution $p(\underset{\sim}{\theta}_n,\Lambda|D_n)$ when the prior for $\underset{\sim}{\theta}_n$ is the usual $N[\underset{\sim}{a}_n, P_n]$ and the prior for Λ is the Wishart distribution above, with $\underset{\sim}{\theta}_n$ and Λ independent. O'Hagan (1976) discusses the calculation of various joint and marginal modes in a similar framework and investigates the relationships between such point estimates. He also discusses the relative merits of such estimates, broadly concluding that for the covariance matrices, (and precision matrices such as Λ), the marginal modes provide "better" estimates than joint modes. We discuss the calculation of marginal modes and of approximations to $p(\underset{\sim}{\theta}_n|D_n)$ and $p(\Lambda|D_n)$ which derive from ideas similar to those used in the modal recursions of Chapter 3.

This is done in (b) and (c) below. In (a) we consider the joint distribution, joint modes and approximations to the Bayesian analysis derived from the joint density.

(a) Joint distribution

$$p(\underset{\sim}{\theta}_n, \Lambda | D_n) \propto |\Lambda|^{(\alpha_{n-1}-m-1)/2} \exp\{-\tfrac{1}{2}\mathrm{tr}(V_{n-1}\Lambda)\}.$$

$$\times \exp\left\{-\tfrac{1}{2}(\underset{\sim}{\theta}_n - \underset{\sim}{a}_n)^T P_n^{-1} (\underset{\sim}{\theta}_n - \underset{\sim}{a}_n)\right\}.$$

$$\times |\Lambda|^{\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n)^T \Lambda (\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n)\right\}. \qquad (4.2.10)$$

The joint modes $(\underset{\sim}{\theta}_n^*, \Lambda^*)$ are defined by the modal equations

$$\underset{\sim}{\theta}_n^* = \underset{\sim}{a}_n + P_n H_n^T . \left[H_n P_n H_n^T + \Lambda_n^{*-1}\right]^{-1} \left[\underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n\right], \qquad (4.2.11)$$

and

$$\Lambda_n^{*-1} = (\alpha_{n-1} - m)^{-1} \left[V_{n-1} + (\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n^*)\ (\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n^*)^T\right], \qquad (4.2.12)$$

which can be solved iteratively to provide the values of the modes for use as point estimates.

As an approximation, note that if we use as a starting point for such an iteration the prior means $\underset{\sim}{a}_n$ and Λ_{n-1}, we obtain one-step estimates

$$\underset{\sim}{\theta}_n^1 = \underset{\sim}{a}_n + P_n H_n^T . \left[H_n P_n H_n^T + \Lambda_{n-1}^{-1}\right]^{-1} \left[\underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n\right] \qquad (4.2.13)$$

and $$\Lambda_n^{1-1} = (\alpha_{n-1} - m)^{-1} \left[V_{n-1} + (\underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n) . (\underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n)^T\right]. \qquad (4.2.14)$$

Clearly the recursions (4.2.13) and (4.2.14) follow from a Taylor series approximation to $p(\underset{\sim}{\theta}_n, \Lambda | D_n)$, expanding as a function of $\underset{\sim}{\theta}_n$ and Λ about the prior means $\underset{\sim}{a}_n$ and Λ_{n-1} and retaining second order terms in $\underset{\sim}{\theta}_n$ but only first order terms in Λ. This approximation would imply that $(\underset{\sim}{\theta}_n, \Lambda | D_n)$ are approximately independent, with

$$(\underset{\sim}{\theta}_n | D_n) \sim N[\underset{\sim}{\theta}_n^1, C_n^1]$$

and

$$(\Lambda | D_n) \sim W[\alpha_n, V_n],$$

where

$$C_n^1 = P_n - P_n H_n^T \left[H_n P_n H_n^T + \Lambda_{n-1}^{-1}\right]^{-1} P H_n,$$

and

$$\alpha_n = \alpha_{n-1} + 1,$$

$$V_n = V_{n-1} + (\underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n)(\underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n)^T.$$

(b) Marginal for $\underset{\sim}{\theta}_n$.

$$p(\underset{\sim}{Y}_n | \underset{\sim}{\theta}_n, D_{n-1}) = \int_{\mathbb{R}^n} p(\underset{\sim}{Y}_n | \underset{\sim}{\theta}_n, \Lambda)\ p(\Lambda | D_{n-1}) d\Lambda$$

$$\alpha\ |V_{n-1} + (\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n)(\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n)^T|^{-(\alpha_{n-1}+1)/2}$$

$$\alpha\ \left(1 + (\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n)^T V_{n-1}^{-1} (\underset{\sim}{Y}_n - H \underset{\sim}{\theta}_n)\right)^{-(\alpha_{n-1}+1)/2}$$

$$\alpha\ \left(\alpha_{n-1} + (\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n)^T \Lambda_{n-1} (\underset{\sim}{Y}_n - H_n \underset{\sim}{\theta}_n)\right)^{-(\alpha_{n-1}+1)/2} \qquad (4.2.16)$$

where $\Lambda_{n-1} = E[\Lambda | D_{n-1}] = \alpha_{n-1} \cdot V_{n-1}^{-1}$.

Therefore $p(\underset{\sim}{\theta}_n | D_n)$ is proportional to the product of a normal prior and a multivariate Student t likelihood. We dealt with just such a problem in Chapter 3 and can directly apply those results as follows:

The likelihood can be expressed as a scale mixture of normals,

$$(\underset{\sim}{Y}_n | \underset{\sim}{\theta}_n, D_{n-1}) \sim \int_{\infty}^{\infty} N[H_n \underset{\sim}{\theta}_n, \lambda_n^{-1} \Lambda_{n-1}^{-1}] W(\lambda_n) d\ \lambda n$$

where $W(\lambda) = G[\alpha_{n-1}/2, \alpha_{n-1}/2]$ and λ_n is independent of Λ, $\underset{\sim}{\theta}_n$. Then

$$(\underset{\sim}{\theta}_n | D_n, \lambda_n) \sim N[\underset{\sim}{m}_n(\lambda_n), C_n(\lambda_n)]$$

where

$$\underset{\sim}{m}_n(\lambda_n) = \underset{\sim}{a}_n + P_n H_n^T \left[H_n P_n H_n\ \lambda_n + \Lambda_{n-1}^{-1}\right]^{-1} \lambda_n (\underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n)$$

and

$$C_n(\lambda_n) = P_n - P_n H_n^T \left[H_n P_n H_n \lambda_n + \Lambda_{n-1}^{-1}\right]^{-1} H_n P_n \lambda_n.$$

The modal recursions of Chapter 3 can now be used to provide an approximation given by

$$(\underset{\sim}{\theta}_n | D_n) \sim N\left[\underset{\sim}{m}_n(\tilde{\lambda}_n),\ C_n\right]$$

where $\tilde{\lambda}_n = \psi(\underset{\sim}{u}_n) = (\alpha_{n-1}+1).(\alpha_{n-1}+\underset{\sim}{u}_n^T \Lambda_{n-1} \underset{\sim}{u}_n)^{-1}$

and $\underset{\sim}{u}_n = \underset{\sim}{Y}_n - H_n \underset{\sim}{a}_n$.

Furthermore

$$\underset{\sim}{m}_n = \underset{\sim}{m}_n(\underset{\sim}{\lambda}_n) = \underset{\sim}{a}_n + P_n H_n^T R_n^{-1}.\psi(\underset{\sim}{u}_n).\underset{\sim}{u}_n \qquad (4.2.17)$$

where $R_n = H_n P_n H_n^T \psi(\underset{\sim}{u}_n) + \Lambda_{n-1}^{-1}$. The covariance corresponding to this is

$$C_n = C_n(\tilde{\lambda}_n) + S_n.$$

where $C_n(\tilde{\lambda}_n) = P_n - P_n H_n^T R_n^{-1} H_n P_n\ \psi(\underset{\sim}{u}_n)$

and $S_n = P_n H_n^T R_n^{-1}.\underset{\sim}{u}_n \underset{\sim}{u}_n^T.R_n^{-1} H_n P_n.(\alpha_{n-1}+\underset{\sim}{u}_n^T \Lambda_{n-1} \underset{\sim}{u}_n)^{-1}$.

Those recursions follow directly from §3.2.4 in the special case of a Student t likelihood.

Due to the excellent performance of the Student t based modal recursions in Chapter 3 we expect the above algorithm to perform well and illustrate it later with numerical examples. For large n, α_n behaves like n and (4.2.17), (4.2.18) behave like the Kalman Filter. A closer examination of the form of $\underset{\sim}{m}_n$ reveals a similarity between itself and the joint recursion of (4.2.13) above. The latter is essentially just $E\left[\underset{\sim}{\theta}_n | D_n,\ \Lambda=\Lambda_{n-1}\right]$ whilst the modal recursion (S.2.17) is equal to $E\left[\underset{\sim}{\theta}_n | D_n,\ \Lambda=\hat{\Lambda}_n\right]$, where $\hat{\Lambda}_n = \psi(\underset{\sim}{u}_n)\Lambda_{n-1}$ rather than just Λ_{n-1}. Clearly for large n there is no difference and numerical studies later indicate similar small sample performance.

(c) Marginal for Λ

$$P(\underset{\sim}{Y}_n|\Lambda,\ D_{n-1}) = N\left[H_n \underset{\sim}{a}_n, \Lambda^{-1}+Q_n\right]$$

where $Q_n = H_n P_n H_n^T$. So, immediately,

$$P(\Lambda|D_n) \alpha |\Lambda|^{(\alpha_{n-1}-m-1)/2} |\Lambda^{-1}+Q_n|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\ \mathrm{tr}\left[\Lambda V_{n-1}+(\Lambda^{-1}+Q_n)^{-1} \underset{\sim}{u}_n \underset{\sim}{u}_n^T\right]\right\}. \qquad (4.2.19)$$

Following the philosophy of Chapter 3, we approximate (4.2.19) with a density of the same functional form as the prior i.e. Wishart. If we increment the power of $|\Lambda|$ by one half - taking essentially one degree of freedom for $\underset{\sim}{Y}_n$, then "linearizing" the resultant exponent as a function of Λ will supply a Wishart form. To do this we require the following matrix derivatives:

(i) $\frac{\partial}{\partial \Lambda}\ \mathrm{tr}\left\{(\Lambda^{-1}+Q_n)^{-1} \underset{\sim}{u}_n \underset{\sim}{u}_n^T\right\} = (I+\Lambda Q_n)^{-1} \underset{\sim}{u}_n \underset{\sim}{u}_n^T\ (I+Q_n\Lambda)^{-1}$;

(ii) $\frac{\partial}{\partial \Lambda}\ \ell n|\Lambda^{-1}+Q_n| = -\Lambda^{-1}(\Lambda^{-1}+Q_n)^{-1}\Lambda^{-1}$.

These are special cases of results proved in Lemma 5.1 in the Appendix 5A.

The exponent to be linearized is then, from (4.2.19) on ignoring a term $|\Lambda|^{(\alpha_{n-1}-m)/2}$,

$$-\tfrac{1}{2}|\Lambda^{-1}+Q_n| - \tfrac{1}{2}\ \ell n|\Lambda| - \tfrac{1}{2}\mathrm{tr}\left[\Lambda V_{n-1}+(\Lambda^{-1}+Q_n)^{-1} \underset{\sim}{u}_n \underset{\sim}{u}_n^T\right].$$

So using a Taylor series expansion to first order about $\Lambda = \Lambda_{n-1}$ the prior mean we obtain

$$\text{constant} -\tfrac{1}{2}\mathrm{tr}\left\{\left[\Lambda-\Lambda_{n-1}\right].\left[-\Lambda_{n-1}^{-1}(\Lambda_{n-1}^{-1}+Q_n)^{-1}\Lambda_{n-1}^{-1} + \Lambda_{n-1}^{-1} + V_{n-1} + (I+\Lambda_{n-1}Q_n)^{-1} \underset{\sim}{u}_n \underset{\sim}{u}_n^T (I+Q_n\Lambda_{n-1})^{-1}\right]\right\}$$

by using (i) and (ii) above.

Rewriting we have

$$\text{constant } -\tfrac{1}{2}\text{tr}\{\Lambda V_n\}$$

where

$$V_n = V_{n-1} + D_n(\Lambda_{n-1})$$

and

$$\begin{aligned} D_n(\Lambda) &= -\Lambda^{-1}(\Lambda^{-1}+Q_n)^{-1}\Lambda^{-1} + \Lambda^{-1} + (I+\Lambda Q_n)^{-1}\underset{\sim}{u}_n\underset{\sim}{u}_n^T(I+Q_n\Lambda)^{-1} \\ &= (I+\Lambda Q_n)^{-1}\left[\underset{\sim}{u}_n\underset{\sim}{u}_n^T + Q_n(I+\Lambda Q_n)\right] \quad (I+Q_n\Lambda)^{-1}. \end{aligned} \tag{4.2.20}$$

Note that $p(\Lambda|D_n)$ is now approximately $W[\alpha_n, V_n]$ with $\alpha_n = \alpha_{n-1}+1$. Further the approximate posterior mode is

$$\Lambda_n^* = (\alpha_n - m).V_n^{-1}, \quad \text{if } \alpha_n > m$$

which can be seen to be an approximation to the true mode $\tilde{\Lambda}_n$ which is a solution of

$$\tilde{\Lambda}_n = (\alpha_n - m)\ [V_{n-1}+D_n(\tilde{\Lambda}_n)]^{-1}.$$

Clearly

$$\Lambda_n^* = (\alpha_n - m).[V_{n-1}+D_n(\Lambda_{n-1})]^{-1}$$

is a one-step approximation to $\tilde{\Lambda}_n$ with starting point Λ_{n-1}. Further $D_n(\Lambda)$ involves P_n i.e. takes into account the uncertainty about $\underset{\sim}{\theta}_n$. The joint recursions for $\underset{\sim}{\theta}_n$ and Λ of equations (4.2.13) and (4.2.14) have been used extensively in the engineering literature as in, for example Ljung (1978), and the book of Goodwin and Payne (1977). The usual approach is via approximate joint maximum likelihood estimation and examples are given in the above references with also convergence analyses. We much prefer the marginal modal recursions; firstly it is generally the marginal distributions that are of interest, and secondly the expression of the marginal likelihod for $\underset{\sim}{Y}_n$ given $\underset{\sim}{\theta}_n$, D_{n-1} as a Student t likelihood puts us into the framework of Chapter 3 with a special density form and we have seen that the modal recursions for t likelihood perform well.

§4.2.3 Numerical Examples.

The following sets of figures provide a comparison of the performances of the algorithms discussed above. For several data generating distributions we use the well tried and tested approximate joint maximum likelihood type algorithm together with our marginal modal recursions to track a 2-dimensional state vector $\underset{\sim}{\theta}_n$ with covariance matrix Λ unknown. In each case we take

$$W_n = \omega I_2, \ \omega = 0.1, \quad \text{for all } n.$$

Each set of figures, 4.1 to 4.4 has five plots. Plots A and B display the absolute errors in the two components of the recursions for $\underset{\sim}{\theta}_n$; thus plot A is of

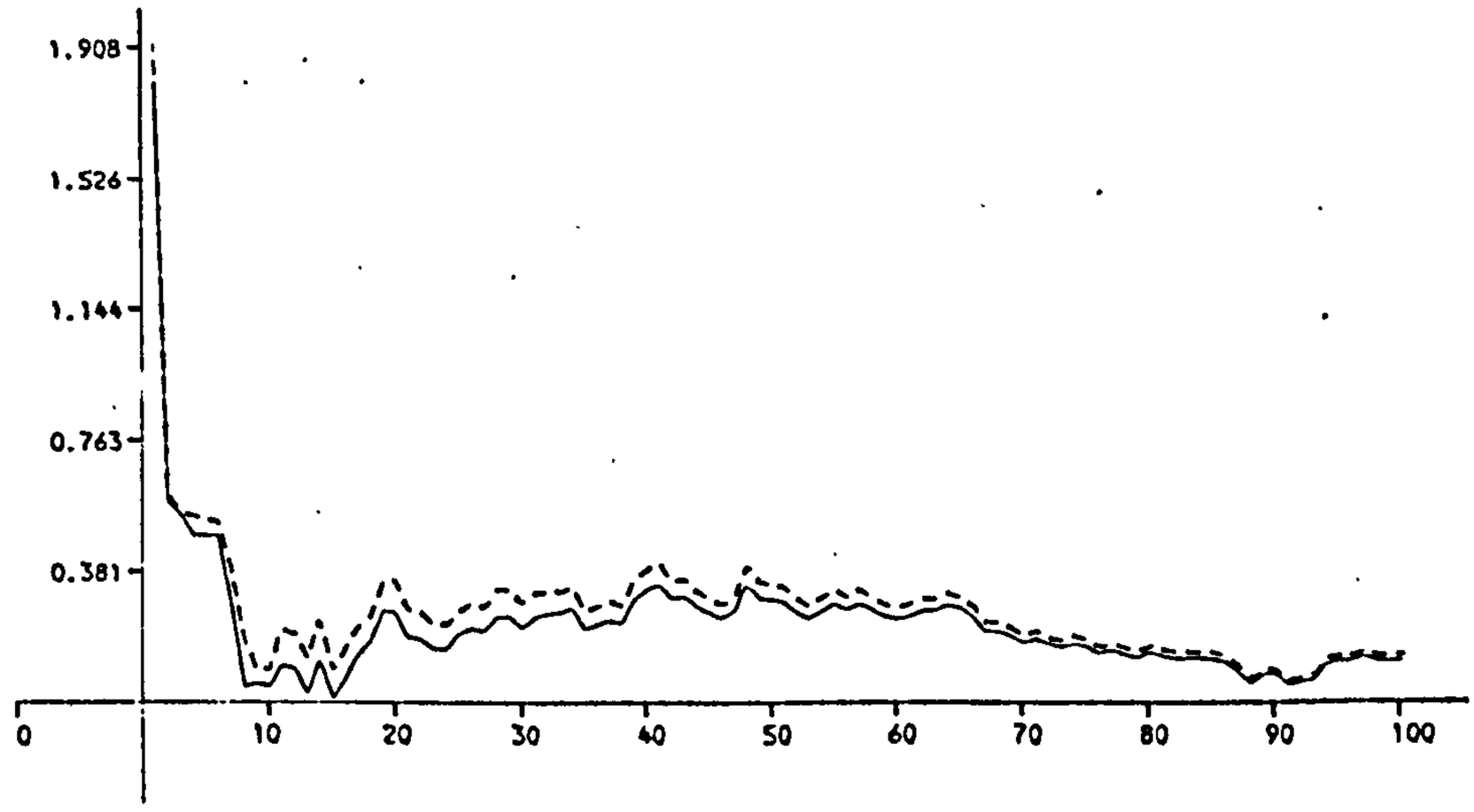
$$|\theta_{n1} - m_{n1}|, \quad \text{where} \quad \underset{\sim}{\theta}_n = (\theta_n, \theta_{n2})^T \quad \text{and}$$
$$\underset{\sim}{m}_n = (m_{n1}, m_{n1})^T,$$

for both filters.

Plots C, D and E in each set of figures display the absolute errors in the covariance matrix recursion in the same way. For the modal recursion for Λ we use the mean of the approximate posterior Wishart density as defined in (c) of §4.2.2. Plots C and D are of the diagonal elements, and E of the off-diagonal element. The priors taken were $\Lambda \sim W[1, I]$, $\underset{\sim}{\theta}_0 \sim N[\underset{\sim}{0}, 100\ I]$ with $\underset{\sim}{\theta}_0$ actually $\underset{\sim}{0}$. Many more numerical studies were undertaken with a variety of priors and starting values but these figures are typical. The two algorithms perform similarly in general but, as shown by the Cauchy example, 4.3, the marginal modal solution is much more effective and robust in non normal situations.

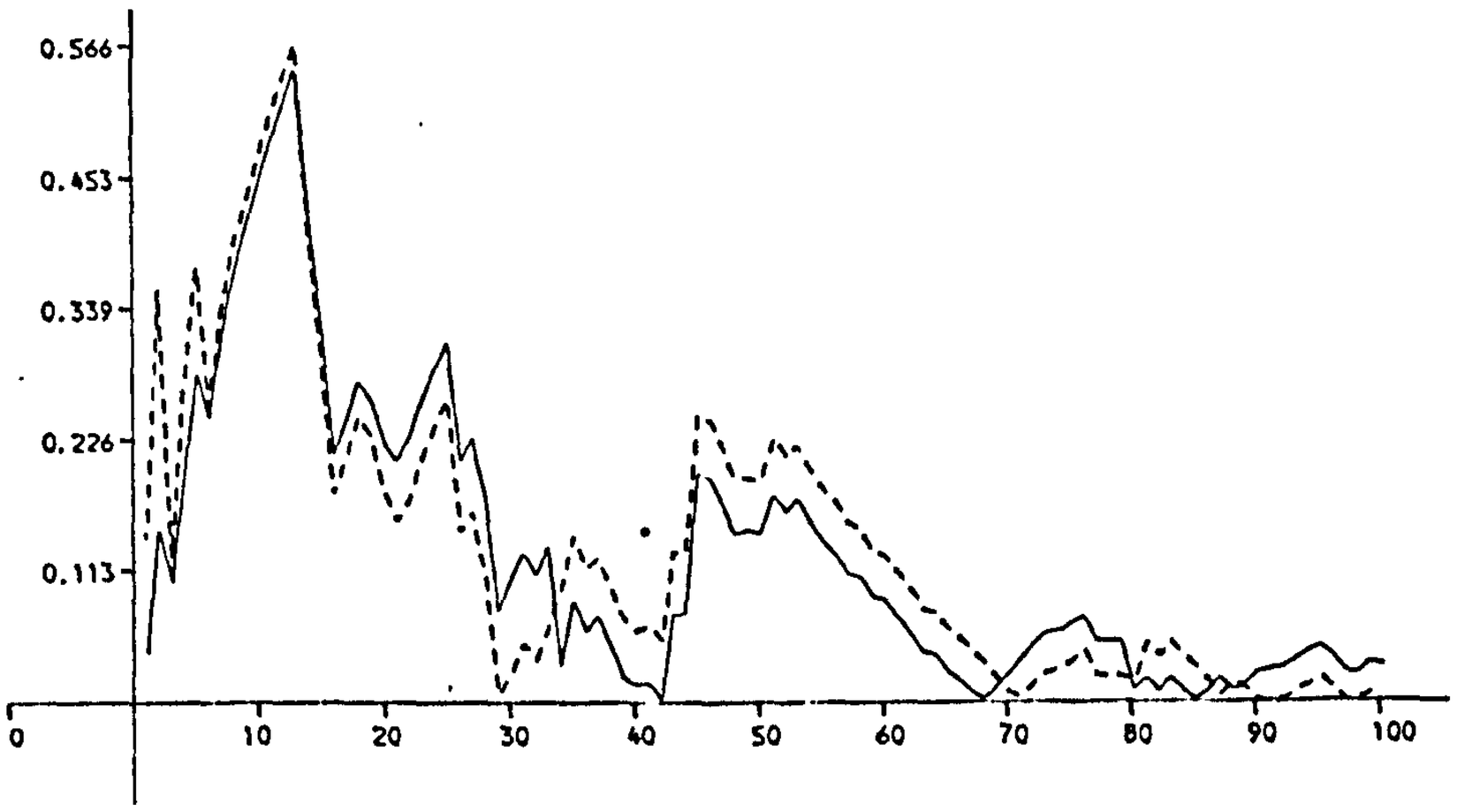
A.

B.

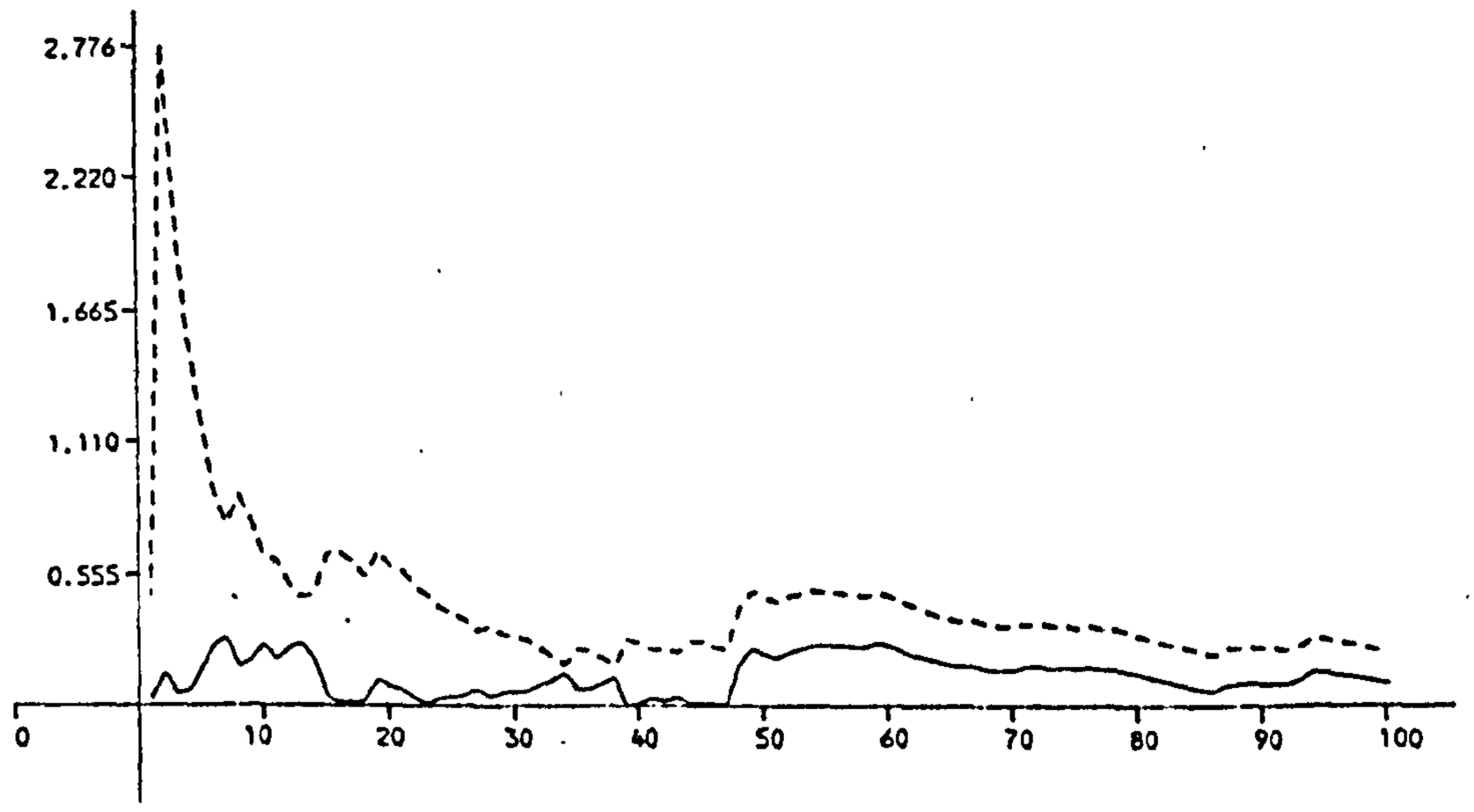


ABSOLUTE ERRORS - MEAN
DATA FROM N(0, I)
——— MODAL FILTER
-------- JOINT FILTER

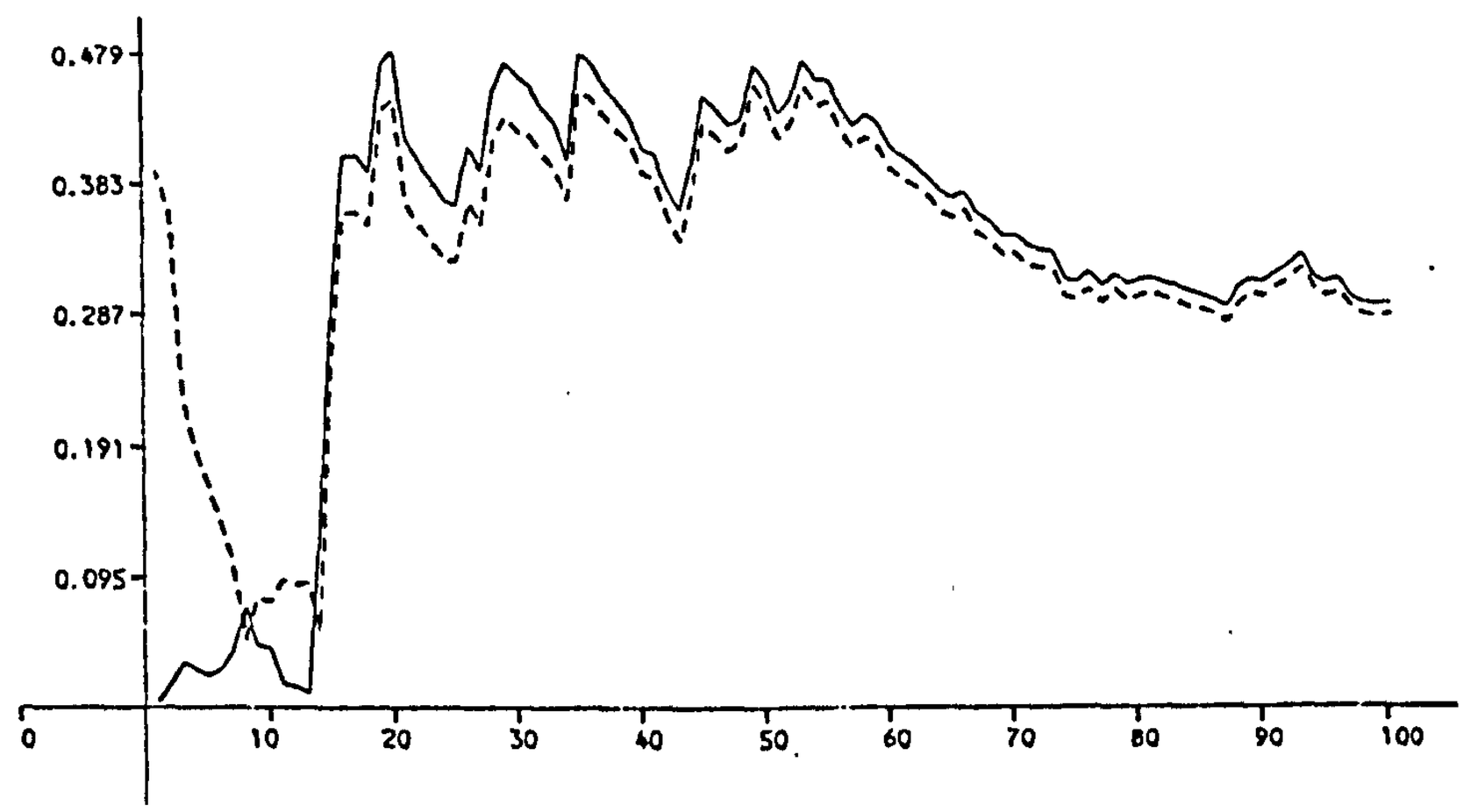
C.



D.

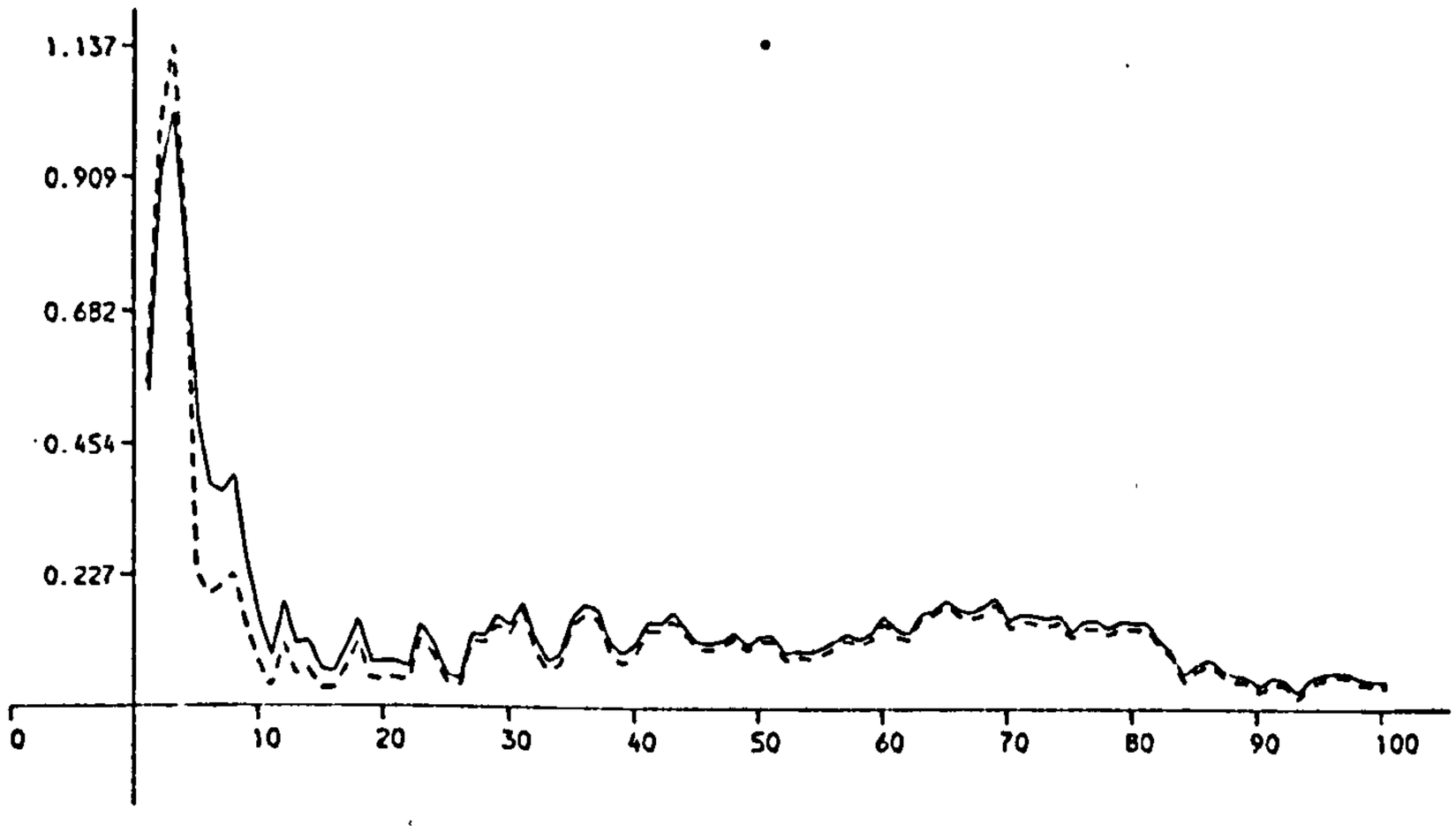


E.

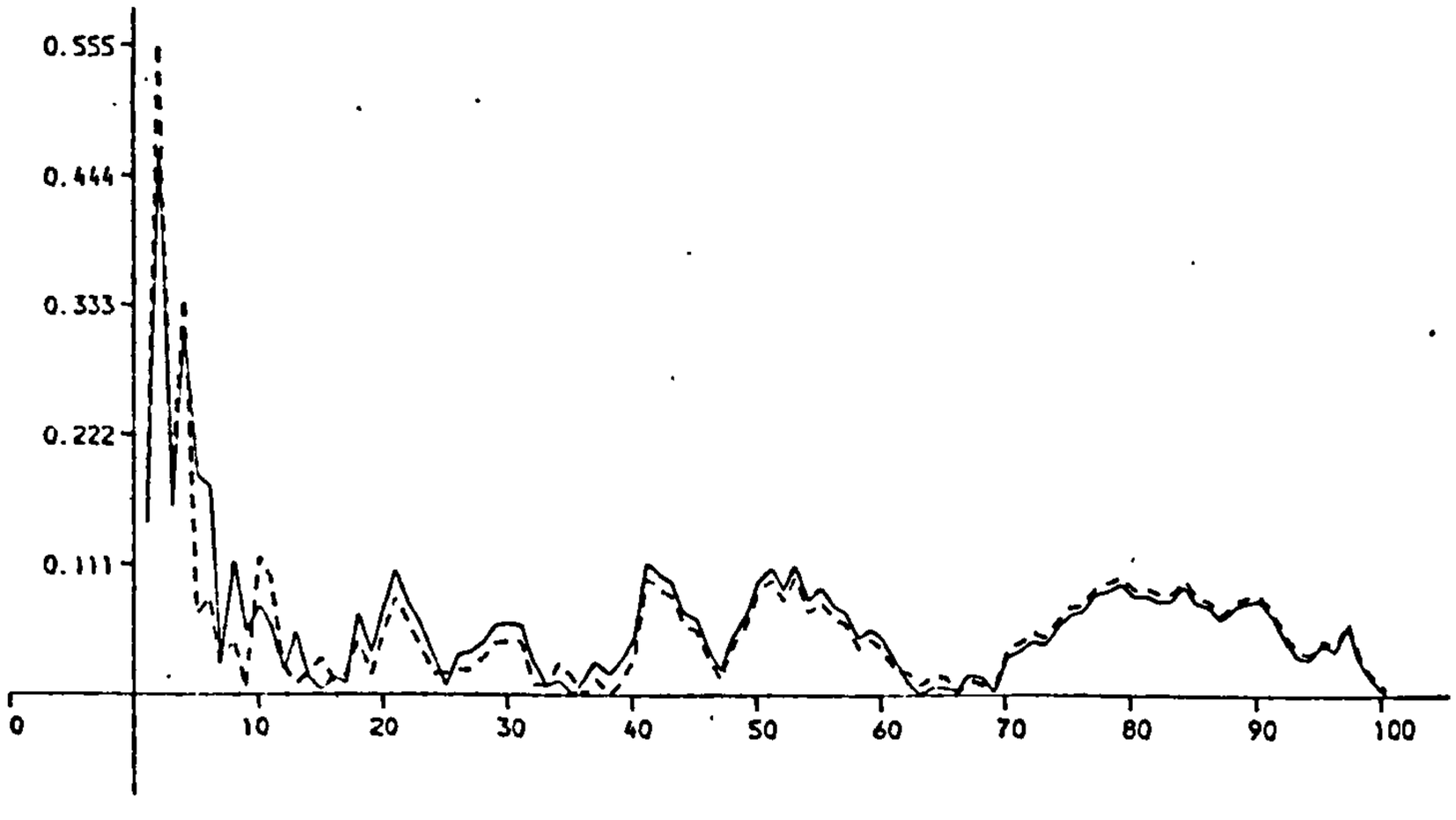


ABSOLUTE ERRORS - COVARIANCE
DATA FROM N(0, I)
——— MODAL FILTER
------- JOINT FILTER

A.

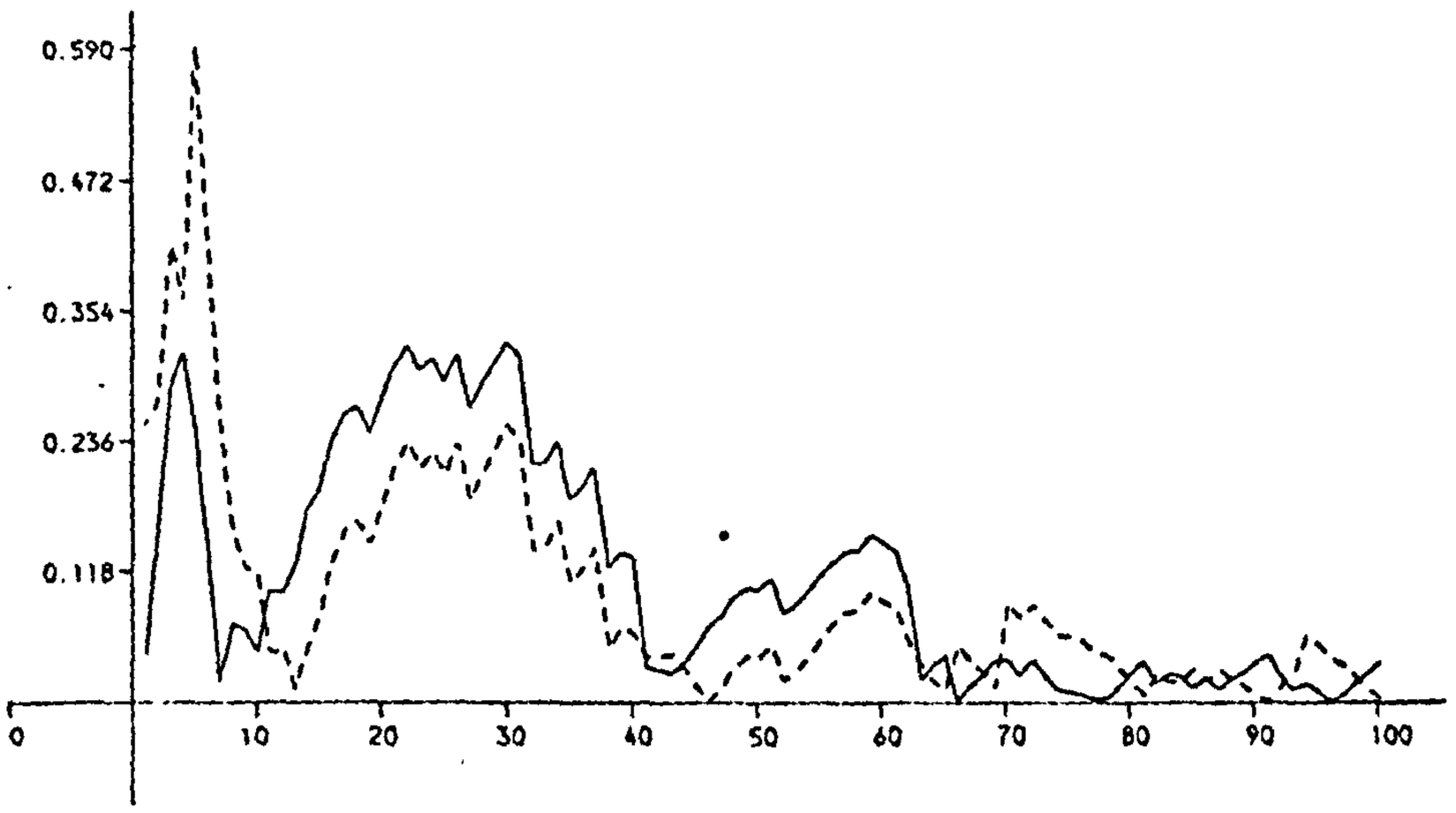


B.

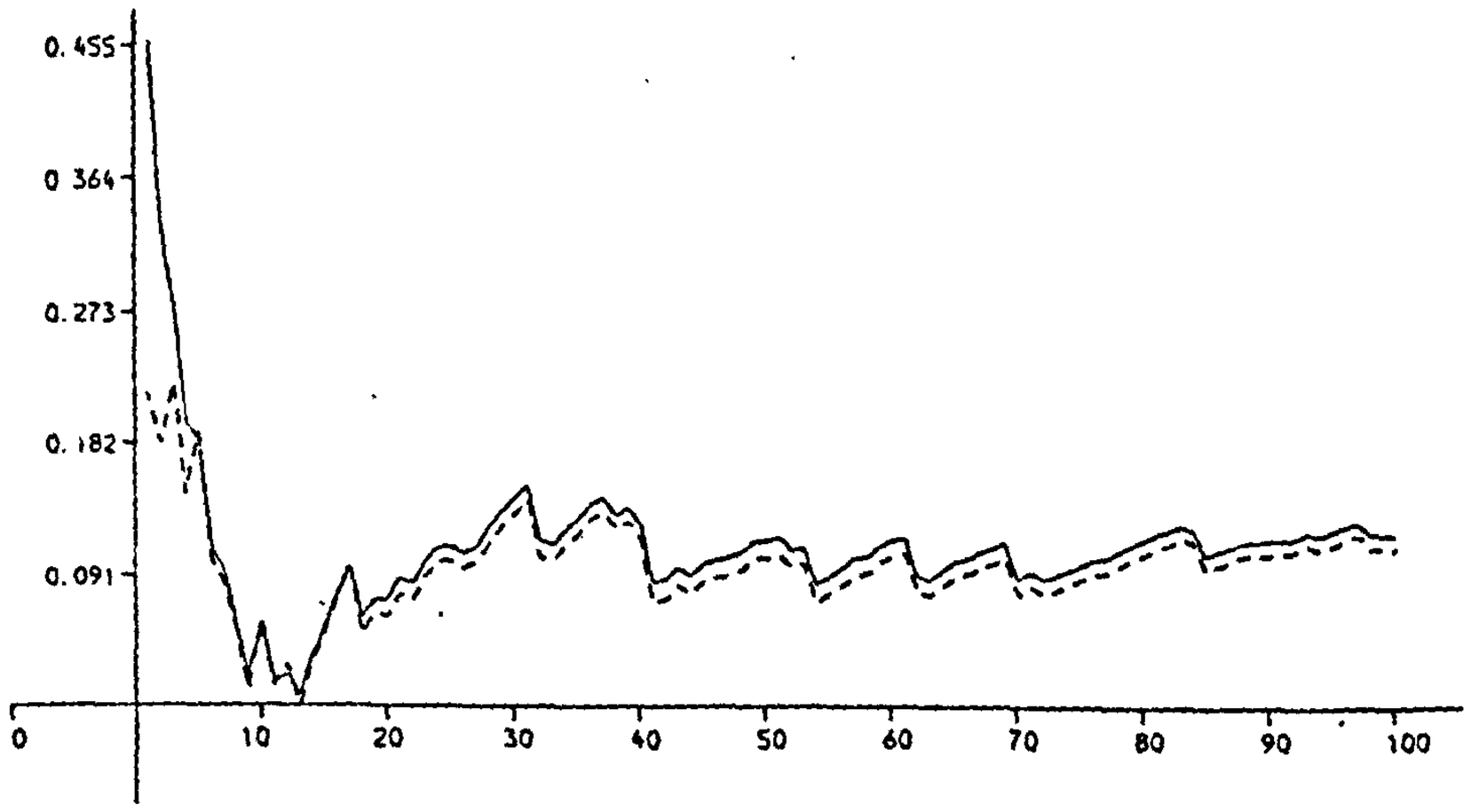


ABSOLUTE ERRORS - MEAN
DATA FROM N(0,V) - V = ((1,0.5),(0.5,0.5))
——— MODAL FILTER
-------- JOINT FILTER

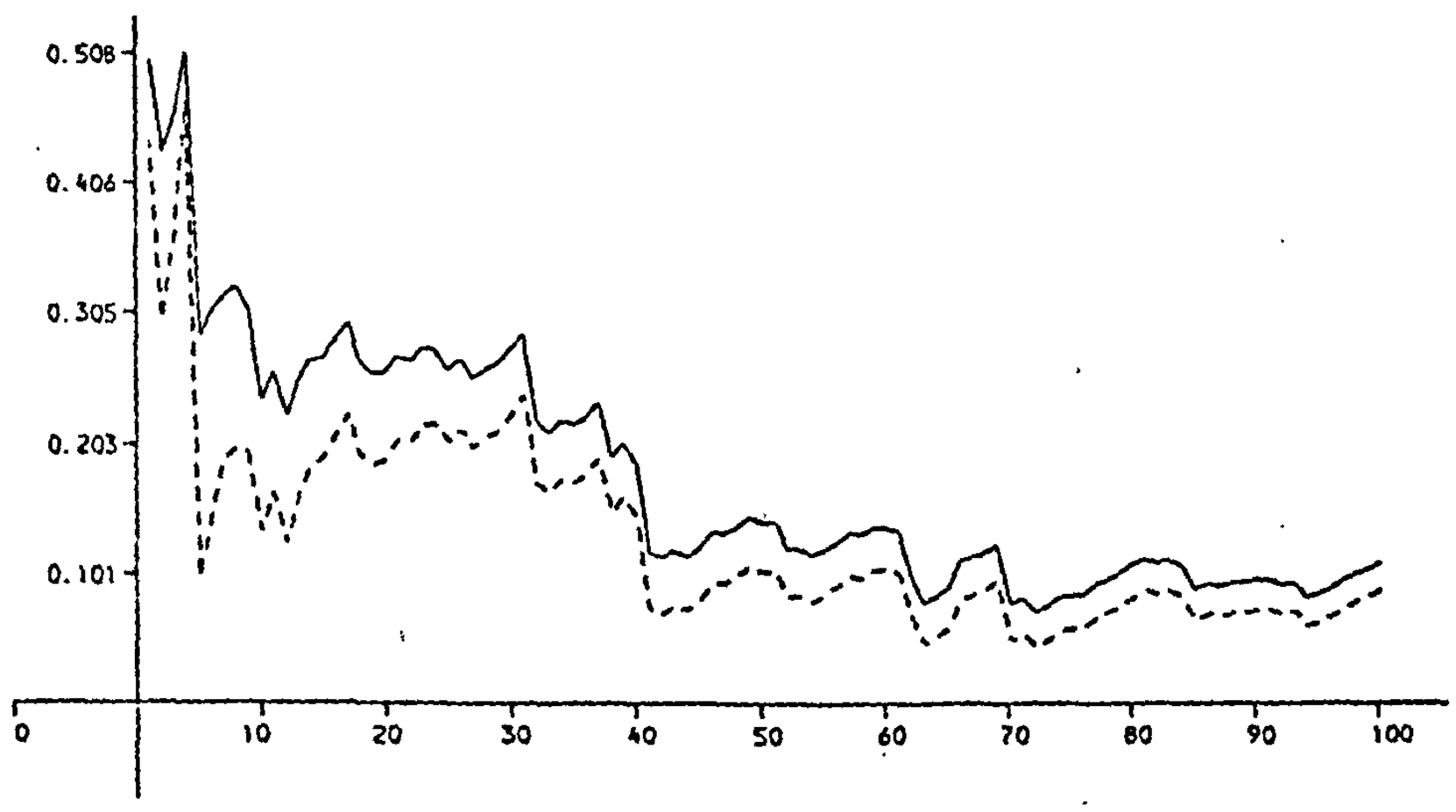
C.



D.

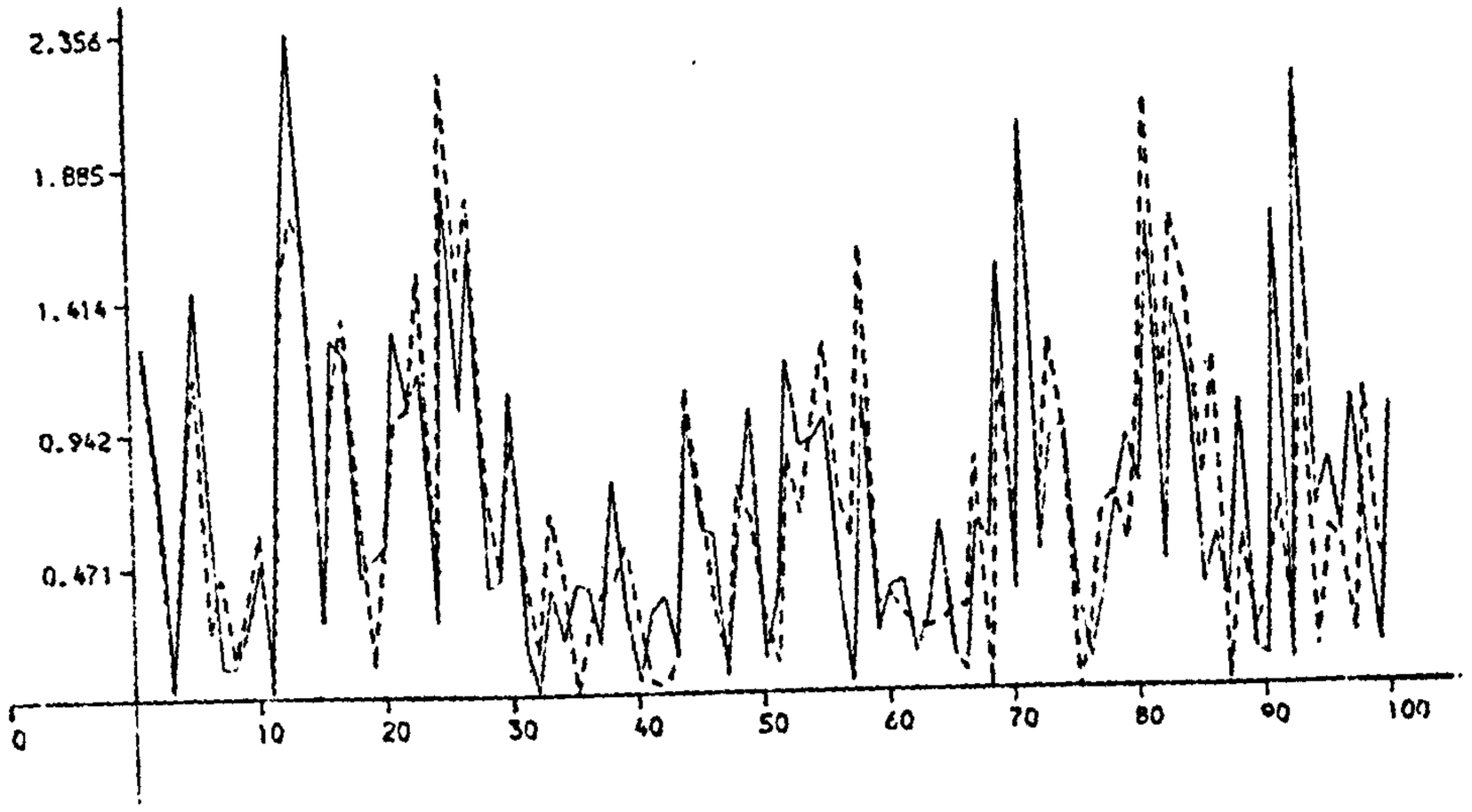


E.

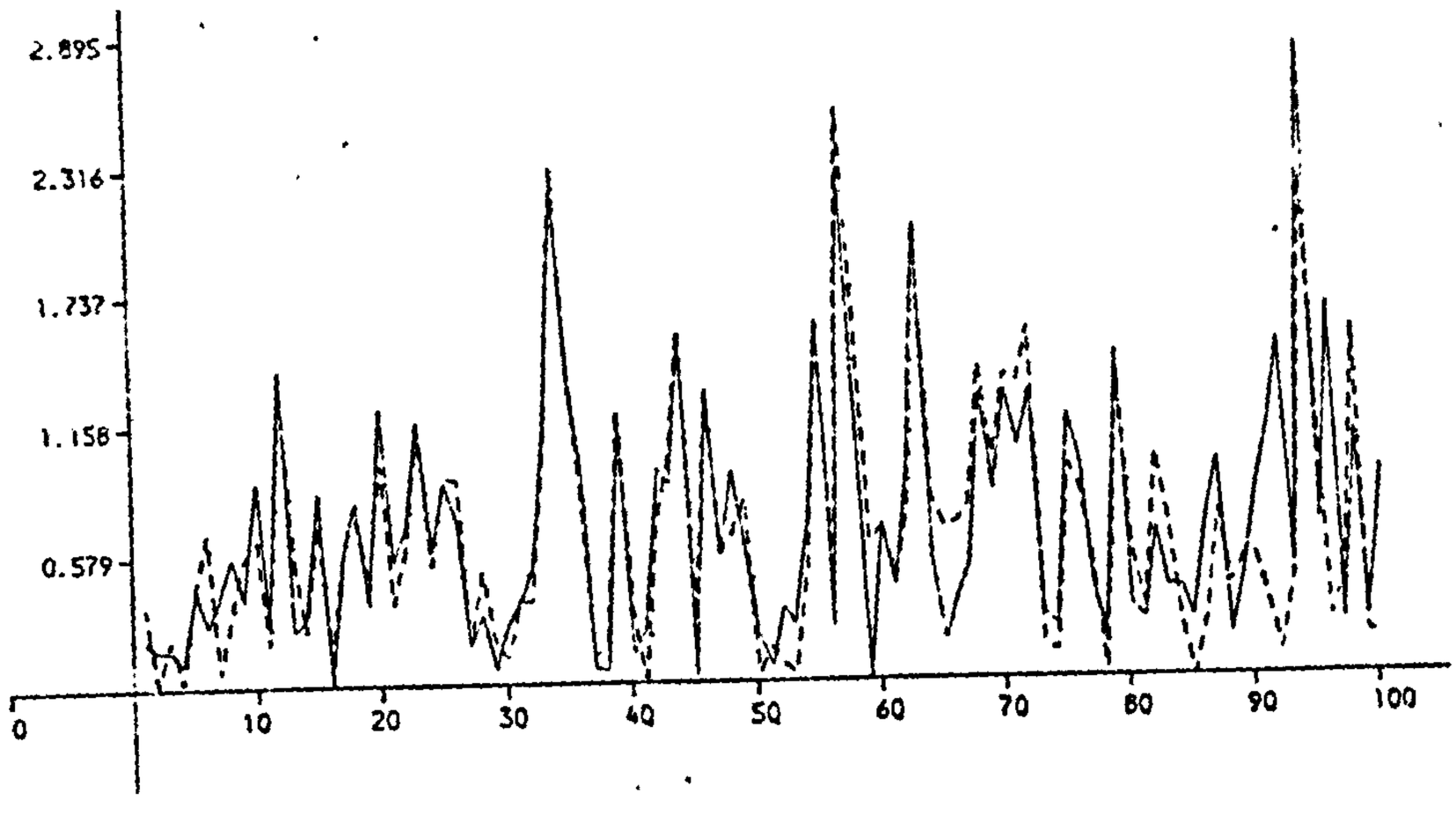


ABSOLUTE ERRORS - COVARIANCE
DATA FROM N(0,V) - V = ((1,0.5),(0.5,0.5))
——— MODAL FILTER
-------- JOINT FILTER

A.

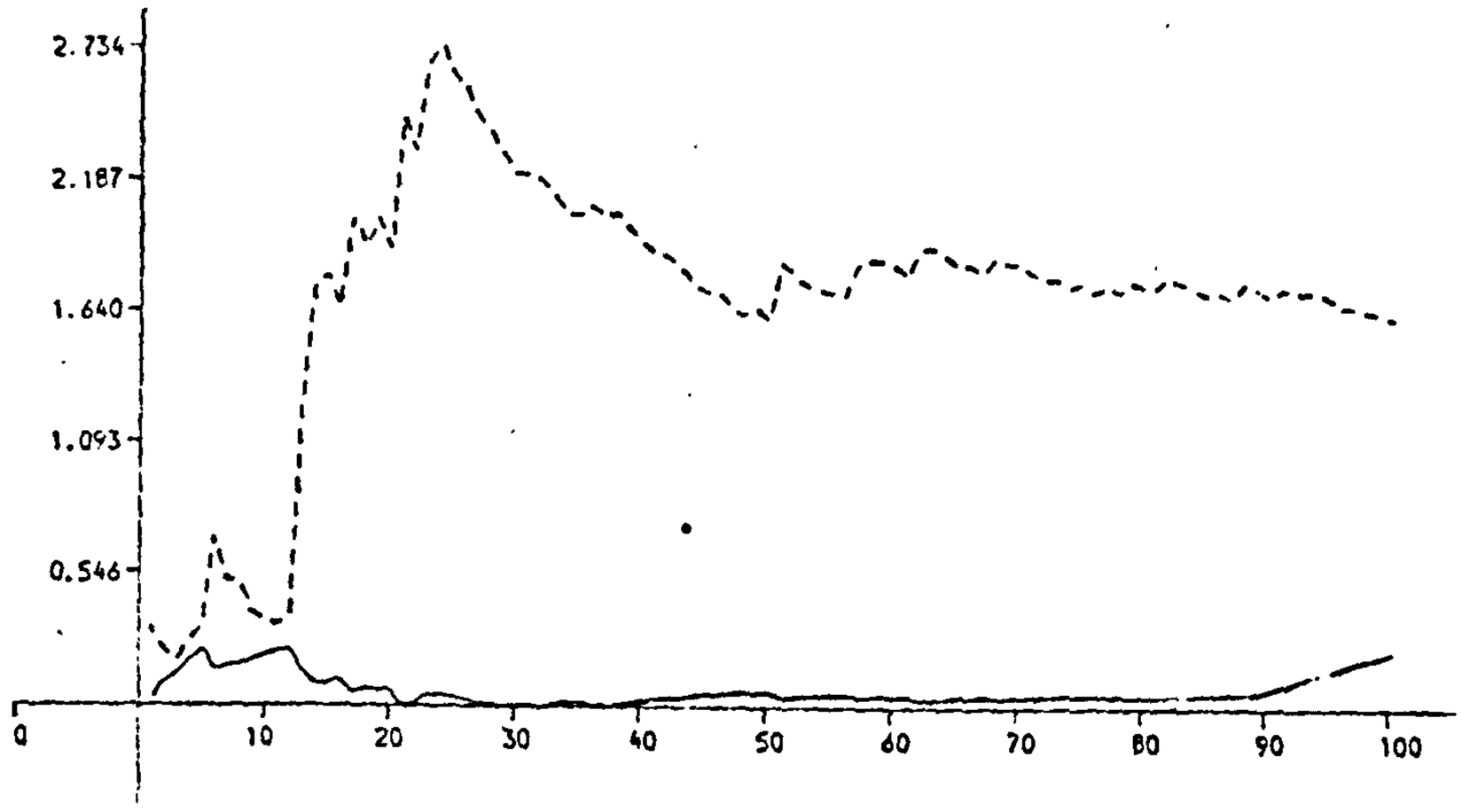


B.

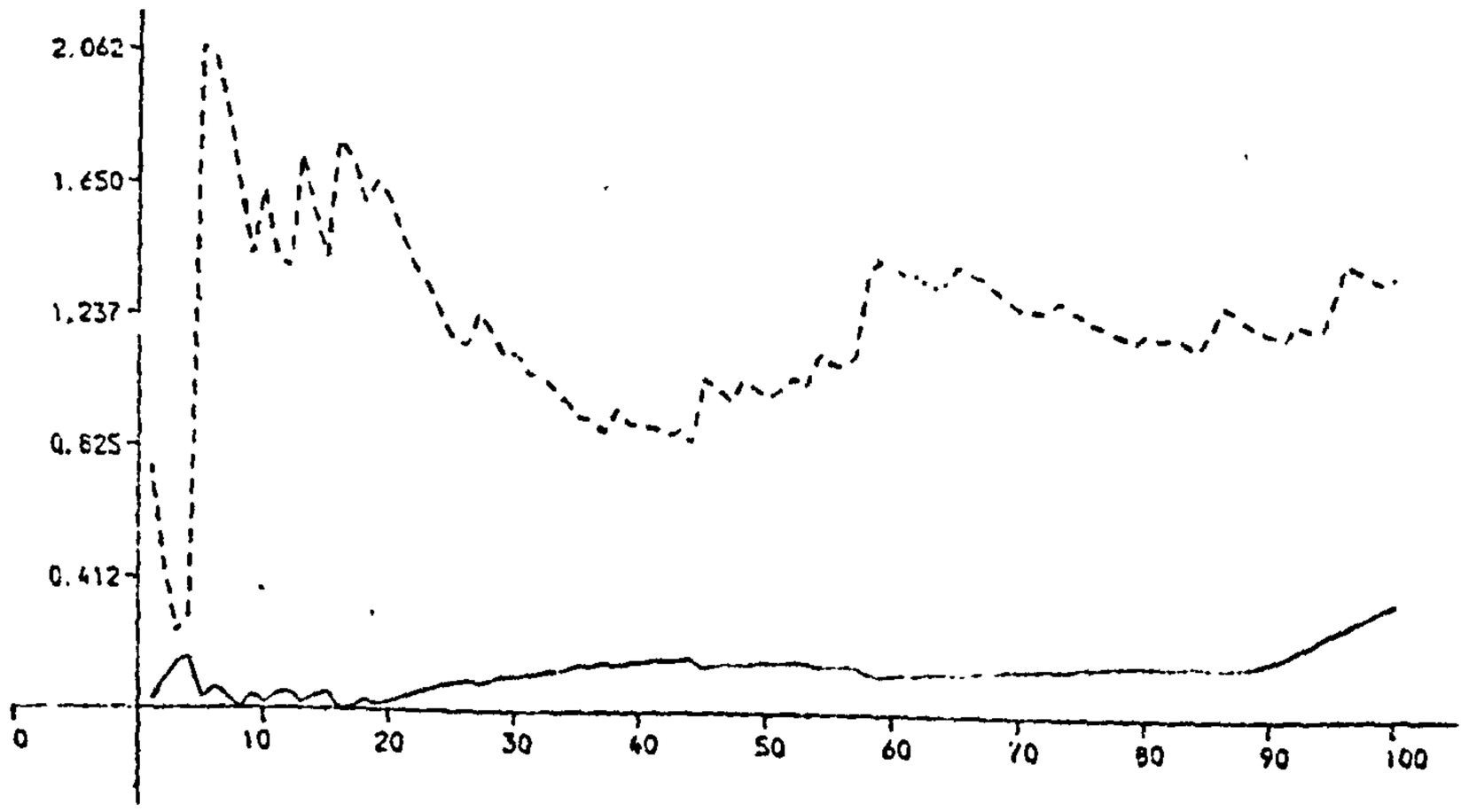


ABSOLUTE ERRORS - MEAN
DATA FROM CAUCHY
——— MODAL FILTER
- - - - - - - - JOINT FILTER

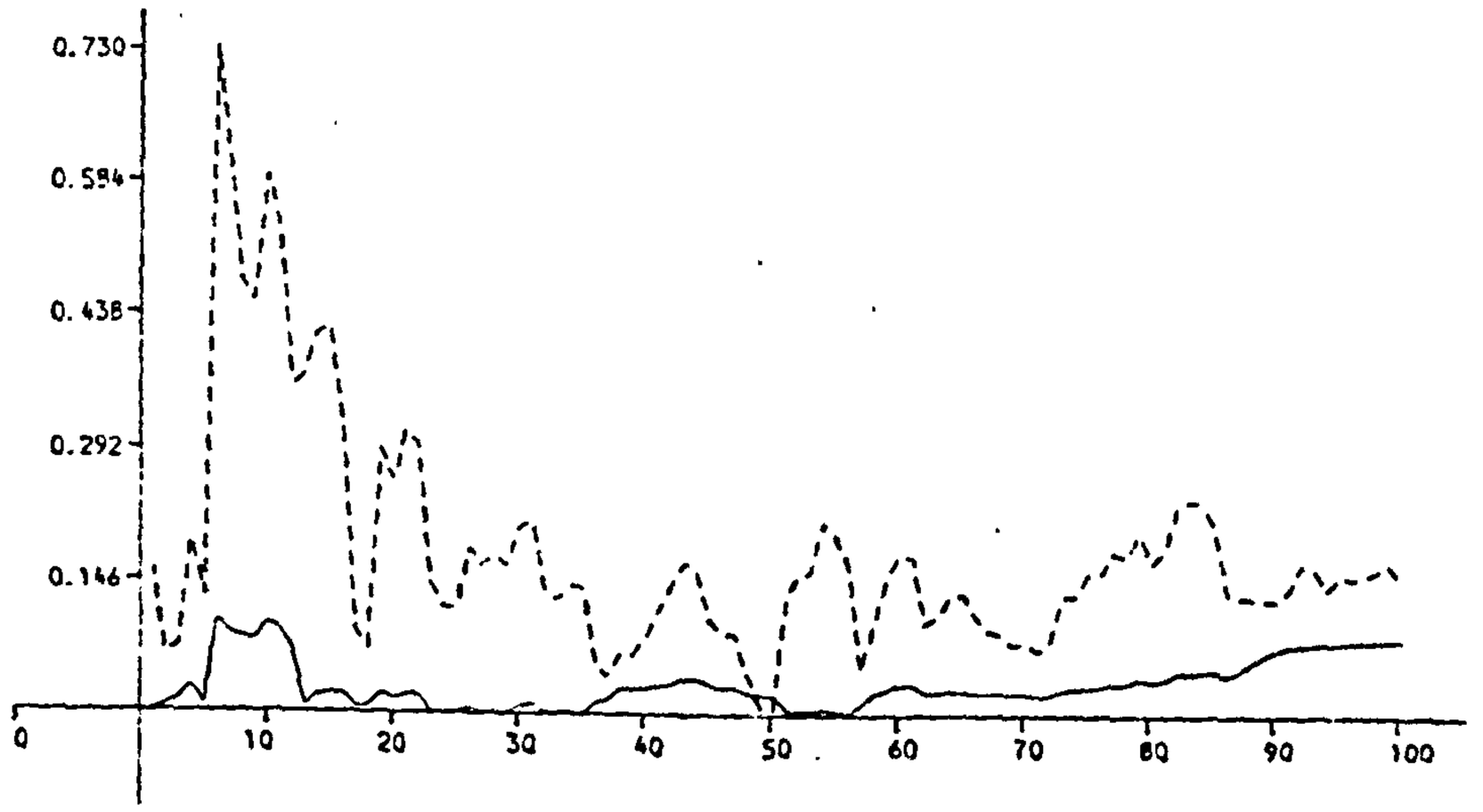
C.



D.



E.



ABSOLUTE ERRORS - COVARIANCE
DATA FROM **CAUCHY**
——— MODAL FILTER
- - - - - - - JOINT FILTER

4.3. Non-normality: heavy-tailed error distributions.

4.3.1. Scalar observations: unknown scale parameters.

Consider the scalar observations model of (3.2.1) and (3.2.2) with observation equation

$$y_n = \underset{\sim}{h}_n^T \underset{\sim}{\theta}_n + \nu_n \qquad (4.3.1)$$

We now assume that the heavy-tailed unimodal symmetric error density p_ν of ν_n is known up to a scale parameter σ,

$$p_\nu(\nu_n|\sigma) = \sigma^{-1} p_\nu(\sigma^{-1}\nu_n), \quad n = 1,2,... \qquad (4.3.2)$$

In Chapter 2 we discussed at some length the joint prior specification for $\underset{\sim}{\theta}_n$ and σ when $\underset{\sim}{\theta}_n$ was both scalar and constant i.e. the location/scale problem of §2.5.2 and §2.5.3. The comments of those sections are applicable here; if the prior for $\underset{\sim}{\theta}_n$ is scaled by σ with prior mean $\underset{\sim}{a}_n$ not involving σ^2, then the marginal likelihood of y_n given σ is also scaled by σ,

$$p(y_n|D_{n-1},\sigma) = \sigma^{-1} p(\sigma^{-1}[y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n]).$$

So we may as well begin our analysis by considering the case of $\underset{\sim}{\theta}_n$ known (and equal, say, to $\underset{\sim}{a}_n$). We adopt a Gamma prior for $\lambda = \sigma^{-2}$ as discussed in §2.5.3,

$$(\lambda|D_{n-1}, \underset{\sim}{\theta}_n = \underset{\sim}{a}_n) \sim G[\alpha_{n-1}/2, \beta_{n-1}/2].$$

(a) Known $\underset{\sim}{\theta}_n = \underset{\sim}{a}_n$.

Set $z_n = y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n$. Then $p(\lambda|D_n)$ is given by

$$p(\lambda|D_n) \propto G[\alpha_{n-1}/2, \beta_{n-1}/2] \cdot \lambda^{\frac{1}{2}} p_\nu(\lambda^{\frac{1}{2}} z_n) \qquad (4.3.3)$$

Again the route to a recursive updating algorithm for λ is the same as that adopted in §4.2; approximate the posterior by a density of the same functional form as the prior via a Taylor series expansion of the log likelihood. In this case we already have the factor $\lambda^{\frac{1}{2}}$ in

the likelihood so the power of λ in the posterior can be directly incremented by one half representing a degree of freedom for y_n. The remaining exponent in the likelihood is then expanded to first order as

$$\ell n\ p_\nu(\lambda^{\frac{1}{2}} z_n) \approx \text{constant} + (\lambda - \ell_{n-1}) \frac{\partial}{\partial \ell_{n-1}}\ \ell n\ p_\nu(z_n \ell_{n-1}^{\frac{1}{2}}) + \ldots$$

where ℓ_{n-1} is the prior mean $\ell_{n-1} = \alpha_{n-1}/\beta_{n-1}$.

Ignoring higher order terms leads to

$$(\lambda | D_n) \sim G[\alpha_n/2, \beta_n/2], \tag{4.3.4}$$

where $$\alpha_n = \alpha_{n-1} + 1,$$

and

$$\beta_n = \beta_{n-1} - 2 \frac{\partial}{\partial \ell_{n-1}} \ell n\ p_\nu(z_n \ell_{n-1}^{\frac{1}{2}}). \tag{4.3.5}$$

Now, if g_ν is the score of p_ν, then

$$-\frac{\partial}{\partial \lambda} \ell n\ p_\nu(z_n \lambda^{\frac{1}{2}}) = -\frac{1}{2\lambda^{\frac{1}{2}}} g_\nu(z_n \lambda^{\frac{1}{2}}) z_n,$$

therefore

$$\beta_n = \beta_{n-1} + \ell_{n-1}^{-\frac{1}{2}} . z_n\ g_\nu(z_n \ell_{n-1}^{\frac{1}{2}}).$$

Further the symmetry of p_ν implies that the score factors as $g_\nu(u) = \psi_\nu(u).u$, and hence

$$\beta_n = \beta_{n-1} + z_n^2\ \psi_\nu(z_n\ \ell_{n-1}^{\frac{1}{2}}) \tag{4.3.6}$$

Clearly ψ_ν acts as a "robustifier"; at normality $\psi_\nu = 1$ and (4.3.6) is exact and non-robust. Otherwise, for heavy-tailed p_ν, ψ_ν limits the inference of the squared residual z_n^2 on the factor β_n and hence on the posterior of λ. Note that the posterior mode ℓ_n^* is given by

$$\ell_n^{*-1} = \beta_n(\alpha_n-2)^{-1} \text{ for } \alpha_n > 2$$

$$= (\alpha_n-2)^{-1}\left[\beta_{n-1}+z_n\psi_\nu(\ell_{n-1}^{\frac{1}{2}}z_n)\right].$$

Clearly the exact mode is

$$\hat{\ell}_n^{-1} = (\alpha_n-2)^{-1}\left[\beta_{n-1}+z_n^2\,\psi_\nu(\hat{\ell}_n^{\frac{1}{2}}z_n)\right]$$

and so ℓ_n^* is a one-step approximation to $\hat{\ell}n$ with starting point ℓ_{n-1}.

Finally the posterior mean is given by $E[\lambda|D_n] \simeq \ell_n = \alpha_n/\beta_n$, whose inverse then satisfies

$$\alpha_n^2 = \ell_n^{-1} = \sigma_{n-1}^2 + \alpha_n^{-1}\left[z_n^2\,\psi_\nu(\ell_{n-1}^{\frac{1}{2}}z_n) - \sigma_{n-1}^2\right].$$

Alternatively, noting that

$$\frac{\partial}{\partial\sigma^2}\ell n\; p(z_n|\sigma^2) = -\frac{1}{2\sigma^4}\left[z_n^2\psi_\nu(\sigma^{-1}z_n) - \sigma^2\right],$$

we have the recursive algorithm for σ_n^2 defined by

$$\sigma_n^2 = \sigma_{n-1}^2 - 2\sigma_{n-1}^4\;\alpha_n^{-1}\left[-\frac{\partial}{\partial\sigma_{n-1}^2}\ell n\; p(z_n|\sigma_{n-1}^2)\right]. \qquad (4.3.7)$$

NB: Similar recursions can be derived for $E[\lambda^{-1}|D_n]$ and various modes of λ, λ^{-1} etc. They differ, in general, only by constant multipliers of the score function in (4.3.7).

So we obtain recursive algorithms for moments of λ which depend on the observations via the score function as in the case of $\underset{\sim}{\theta}_n$ when λ is known. As noted in Chapter 2, likelihoods which are "robust" for $\underset{\sim}{\theta}_n$ i.e. have bounded and redescending score functions, are not necessarily robust for λ in that the posterior $p(\lambda|D_n)$ will not converge to the prior as $|z_n|$ increases. Our approximate gamma posterior for λ behaves in this way: in general β_n does not converge to β_{n-1}, and clearly $\alpha_n = \alpha_{n-1} + 1$ implies no possibility of convergence to the prior.

In the case of normality $\ell_n \to 0$ as $|z_n| \to \infty$. For a robust analysis we should require some constant limit for ℓ_n hopefully not too small. From (4.3.6) this requires that ψ_ν decays at least as fast as z_n^{-2}, which is the case for the Student t family in particular. For the exponential power family the rate is always less than z_n^{-2} meaning $\ell_n \to \infty$ as $|z_n|$ does; in particular for index $0 < \beta < 2$ the rate is like $|z_n|^{\beta-2}$. In fact all the other heavy-tailed distributions of Appendix 2B but the stable family and the normal/uniform lead to this non-robust behaviour of ℓ_n. The stable ψ_ν function is asymptotically $O(z_n^{-2})$ so leading to behaviour similar to the student family. The normal/uniform behaves like the Cauchy in the tails.

Interpretation via scale mixtures of normals.

When $p_\nu(\nu_n) = \int_0^\infty N\left[0, \lambda_n^{-1}\right] \omega(\lambda_n) d\lambda_n$, $n = 1,2,\ldots$ with $\{\lambda_n\}$ independent positive random variables we can again proceed by a conditonal analysis.

$$p(\lambda|D_n,\lambda_n) \propto G\left[\alpha_{n-1}/2,\ \beta_{n-1}/2\right] .\ \lambda^{\frac{1}{2}} \exp\{-\lambda_n \lambda z_n^2/2\}$$

$$= G\left[(\alpha_{n-1}+1)/2,\ (\beta_{n-1}+\lambda_n z_n^2)/2\right]. \qquad (4.3.8)$$

and

$$p(\lambda|D_n) = \int_0^\infty p(\lambda|D_n,\lambda_n) p(\lambda_n|D_n) d\lambda_n, \qquad (4.3.9)$$

where

$$p(\lambda_n|D_n) \propto \omega(\lambda_n)\ p(y_n|\lambda_n,D_n),$$

and

$$p(y_n|\lambda_n,D_n) \propto \{\beta_{n-1}+\lambda_n z_n^2\}^{-(\alpha_{n-1}+1)/2} \qquad (4.3.10)$$

Thus $E\left[\lambda|D_n\right] = (\alpha_{n-1}+1)\ E\left[(\beta_{n-1}+\lambda_n z_n{}^2)^{-1}|D_n\right]$.

Clearly (4.3.6) can be viewed as an approximation, given by

$$E\left[\lambda|D_n,\lambda_n=\tilde{\lambda}_n\right], \text{ where } \tilde{\lambda}_n = \psi_\nu(z_n\ \ell_{n-1}^{\frac{1}{2}}). \qquad (4.3.11)$$

Note further that, using Lemma 2.1.1, we have

$$\lambda^{\frac{1}{2}} g_\nu(\lambda^{\frac{1}{2}} z_n) = \lambda z_n \psi_\nu(\lambda^{\frac{1}{2}} z_n) = E\left[-\frac{\partial}{\partial z_n} \ell n\ p(z_n|\lambda,\lambda_n)\,|\,z_n,\lambda\right]$$

$$= \lambda z_n\ E[\lambda_n|z_n,\lambda],$$

thus $$\overset{\sim}{\lambda}_n = E\left[\lambda_n|z_n,\lambda=\ell_{n-1}\right] = E\left[\lambda_n|D_n,\lambda=\ell_{n-1}\right] \qquad (4.3.12)$$

This development is just as in the location problem. The function ψ_ν is used in an approximation as a robustifier in an attempt to eliminate the nuisance parameter λ_n by substitution of an estimate rather than by integrating over $p(\lambda_n|D_n)$.

If we prefer to evaluate $E[\lambda|D_n]$ numerically, we approximate $p(\lambda|D_n)$ by $G[\alpha_n/2,\beta_n/2]$

where

$$\alpha_n = \alpha_{n-1} + 1,$$

and

$$\beta_n^{-1} = E\left[(\beta_{n-1}+\lambda_n a_n^2)^{-1}|D_n\right] = E\left[\beta_n(\lambda_n)^{-1}|D_n\right].$$

This approximation then has the same mean as $p(\lambda|D_n)$ and can be seen to provide the closest Gamma approximation in terms of the Kullback Leibler directed divergence (See Appendix 5B). As in the location case, we can evaluate β_n by a single integration over $[0,1]$ since

$$\beta_n^{-1} = \beta_{n-1}^{-1}\ E\left[(1+\lambda_n z_n^2 \beta_{n-1}^{-1})^{-1}|D_n\right]$$

and the subject of this expectation is contained in the unit interval.

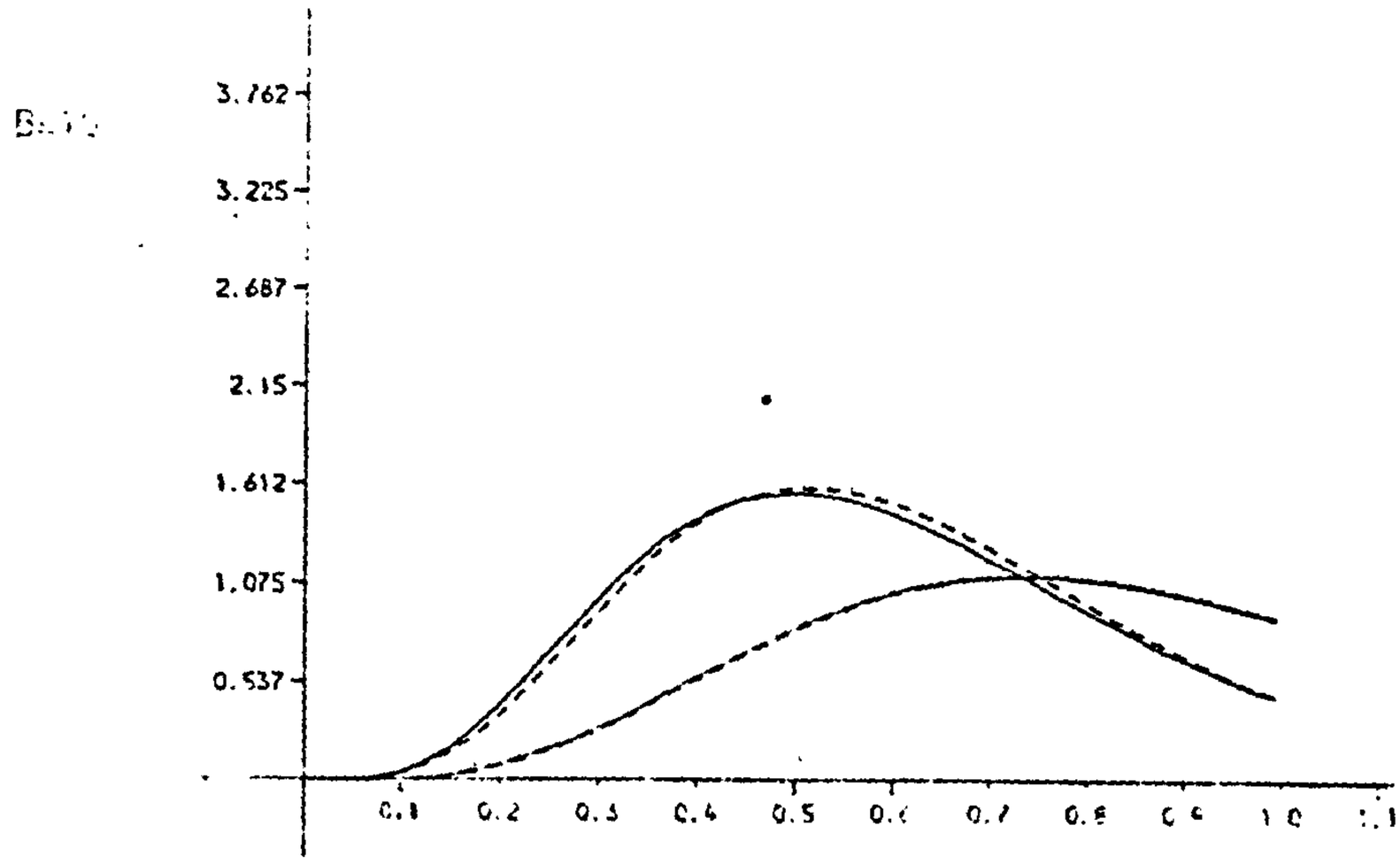
We provide some numerical examples of this approximation when the likelihood is in the Student t family. Figures 4.4 and 4.5 each contain three plots corresponding to different prior specification

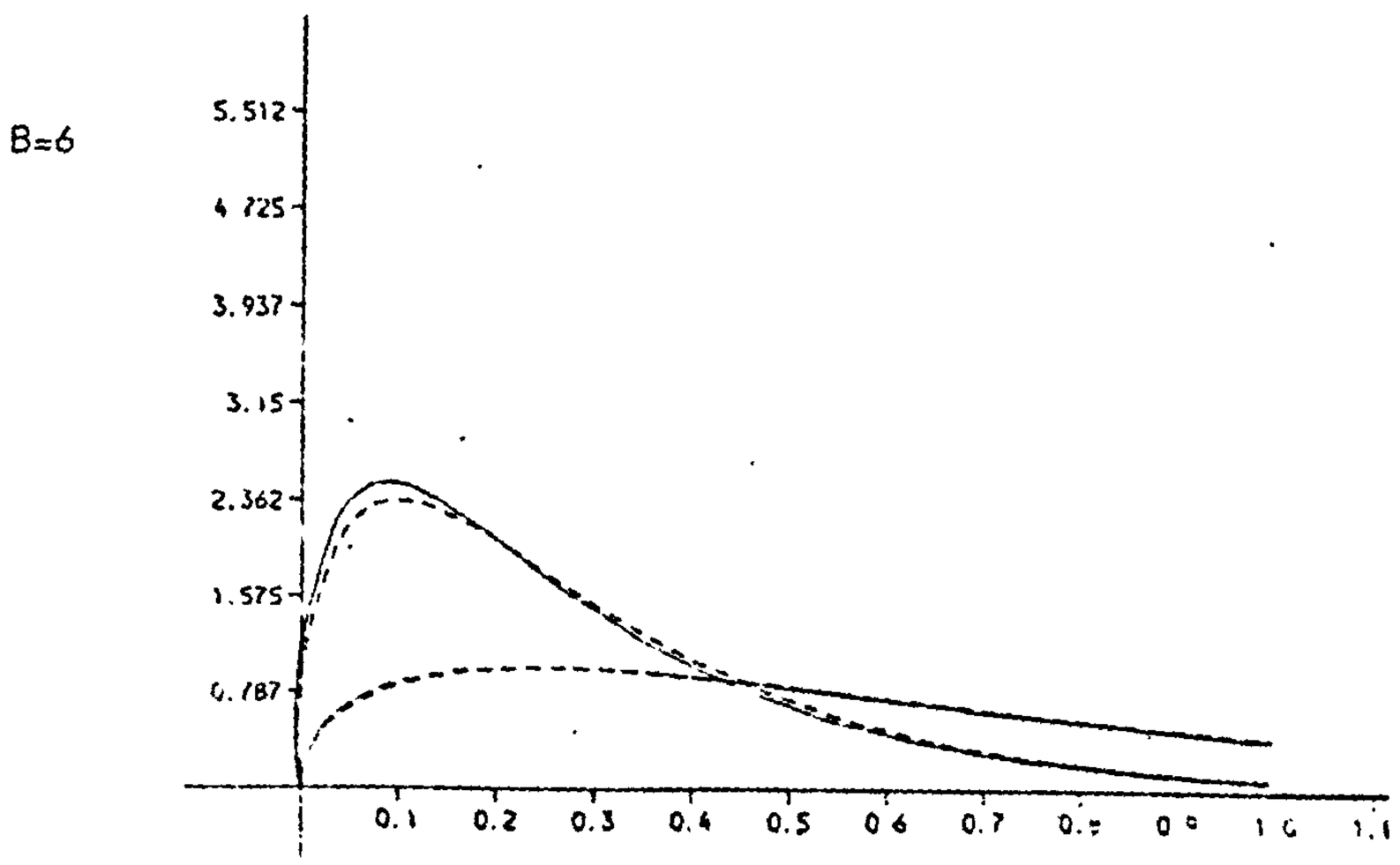
$$\lambda \sim G[b/2,b/2];$$

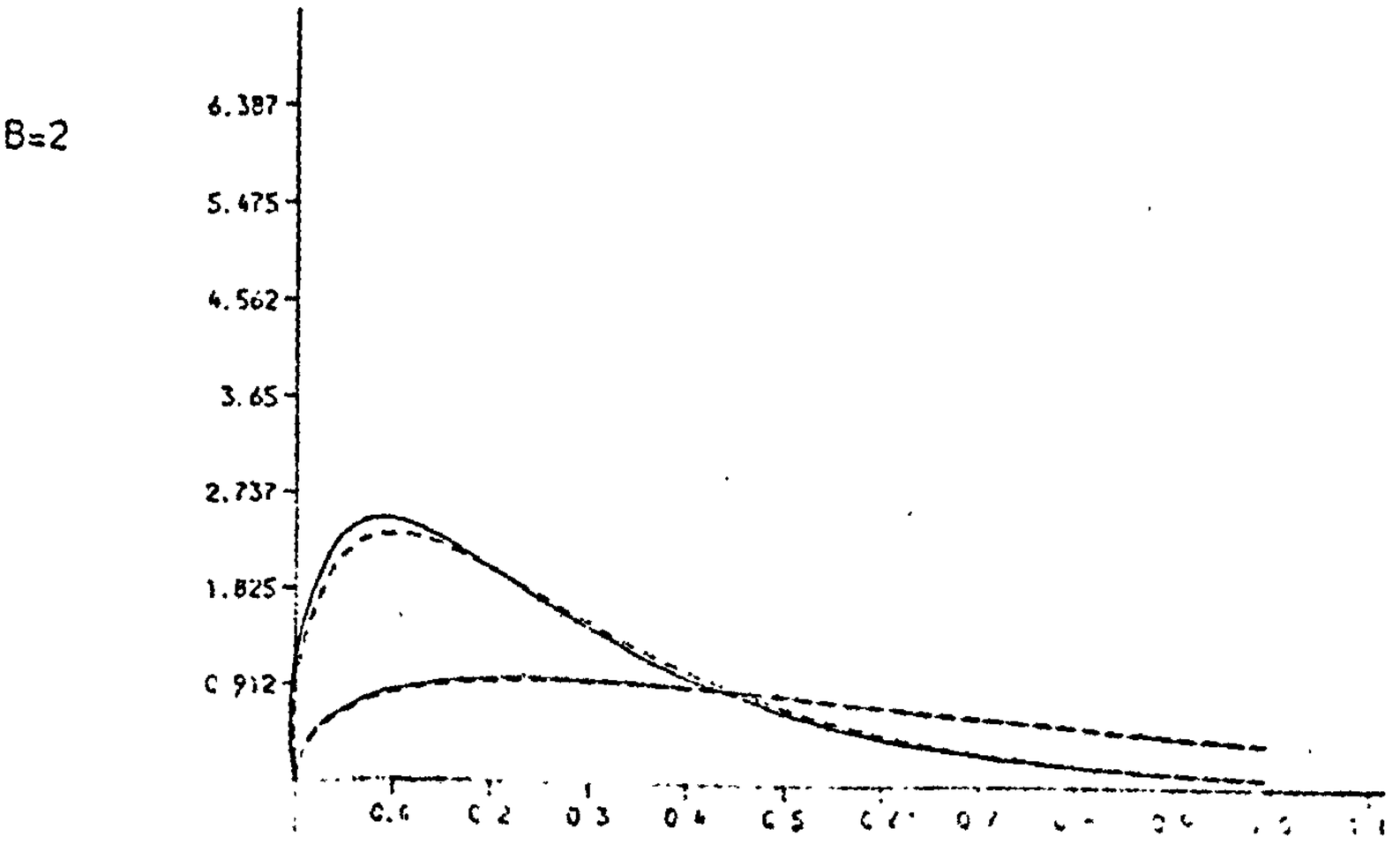
with b = 2, 6, 10. In each plot we have drawn the posterior density

for $\lambda|y$ and the approximate gamma density just discussed for the two values of y, 1.5 and 3.

STUDENT T-25
——— POSTERIOR
------ APPROXIMATION





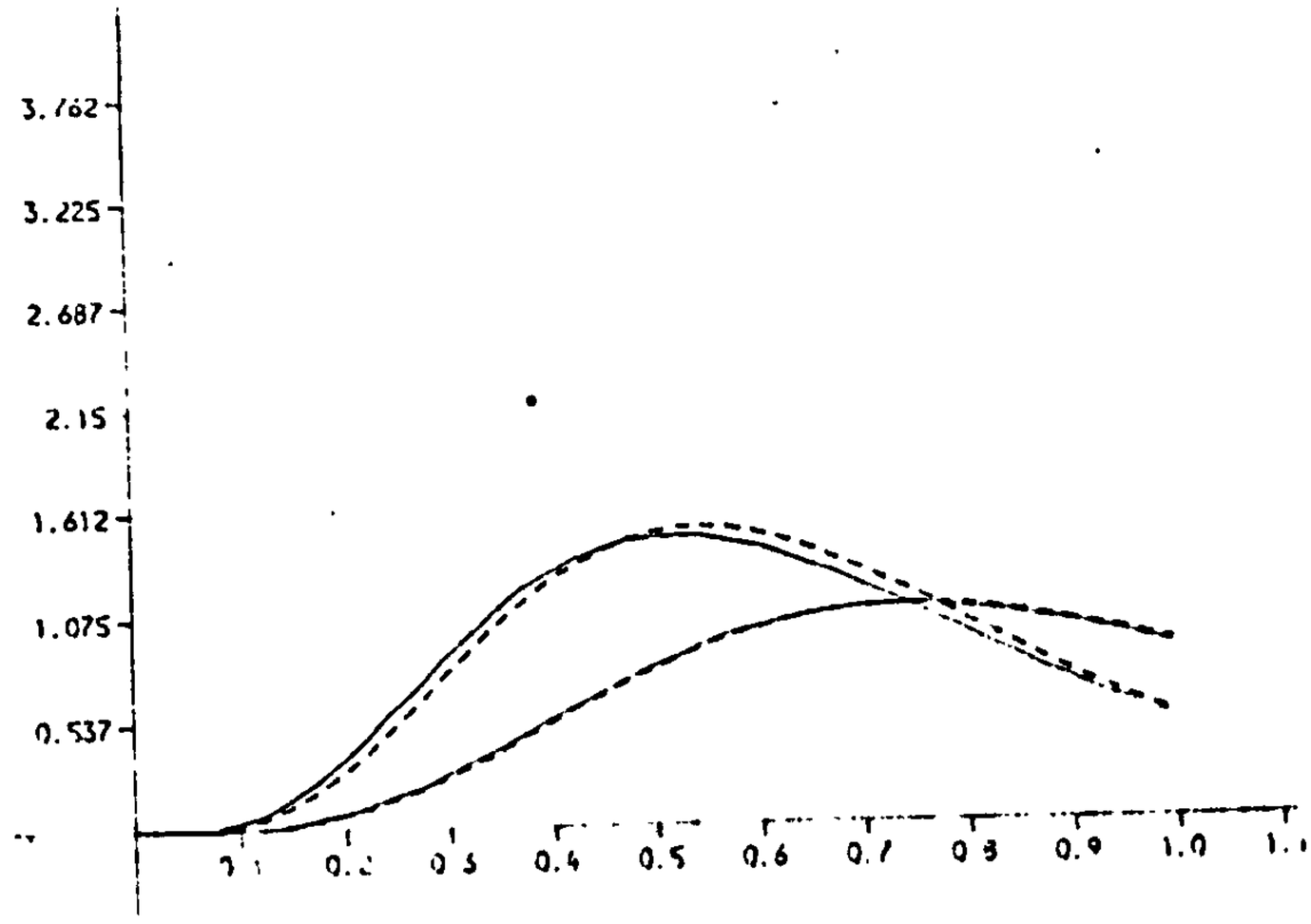


STUDENT T-IS

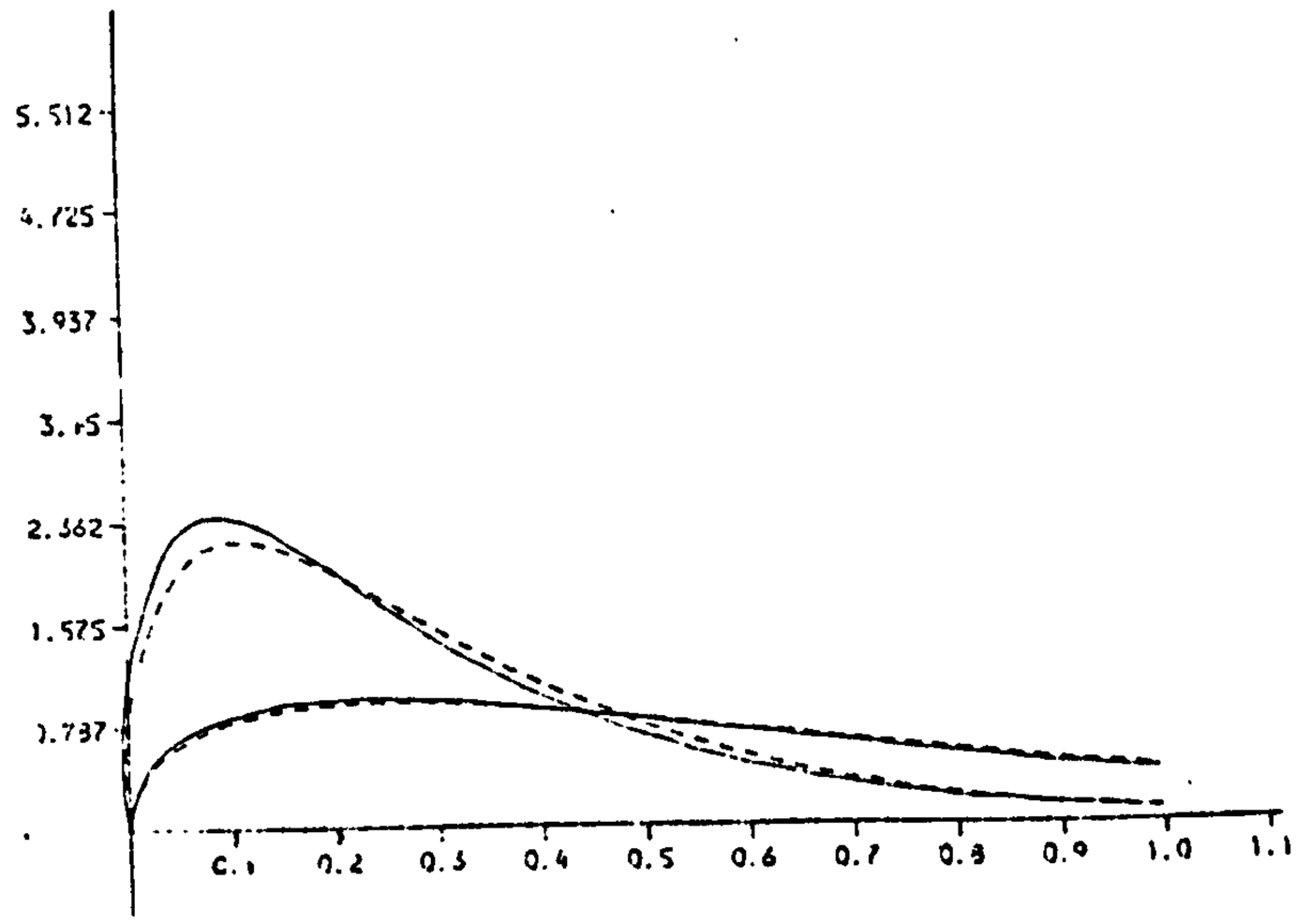
——— POSTERIOR

----- APPROXIMATION

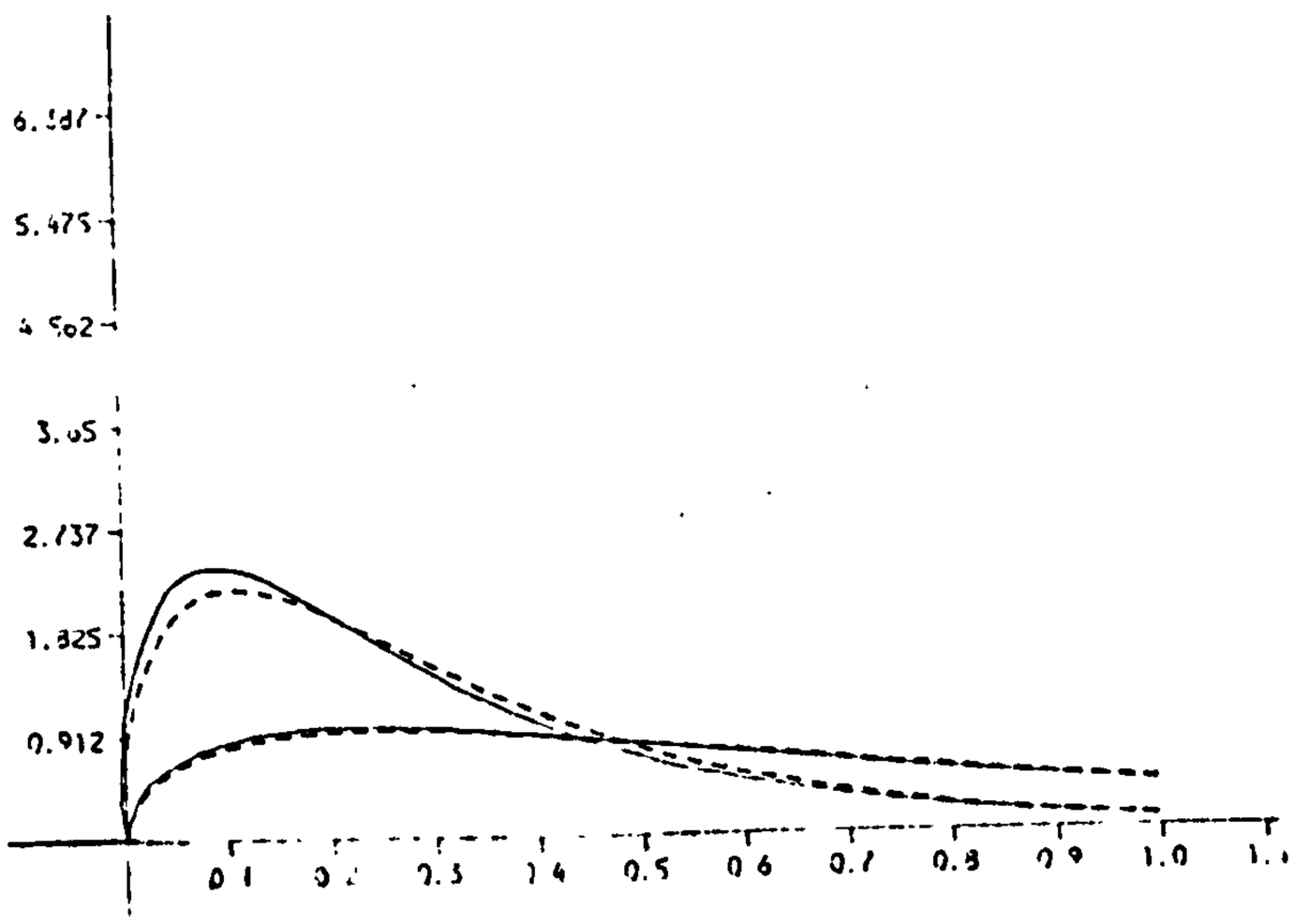
B=10



B=6



B=2



(b) The full model: unknown λ and $\underset{\sim}{\theta}_n$ via scale mixtures of normals.

We now use this approach to learning scale parameters in the model (4.3.1) with the parameter evolution (3.2.1). As in §4.2.1, we again scale the covariance matrix of the evolution errors $\underset{\sim}{\omega}_n$ by the unknown σ.

Thus the prior is

$$(\underset{\sim}{\theta}_{n-1} | D_{n-1}, \lambda) \sim N\left[\underset{\sim}{m}_{n-1}, \lambda^{-1} C_{n-1}\right],$$

leading to

$$(\underset{\sim}{\theta}_n | D_{n-1}, \lambda) \sim N\left[\underset{\sim}{a}_n, \lambda^{-1} P_n\right],$$

where

$$\underset{\sim}{a}_n = G_n \underset{\sim}{m}_{n-1} \text{ and } P_n = G_n C_{n-1} G_n^T + W_n.$$

In addition

$$\lambda | D_n \sim G\left[\alpha_{n-1}/2,\ \beta_{n-1}/2\right],$$

with λ independent of $\underset{\sim}{\omega}_n$ and ν_n.

Now, conditional on the mixing parameter λ_n we have

$$(y_n | \underset{\sim}{\theta}_n, \lambda, \lambda_n) \sim N\left[\underset{\sim}{h}_n^T \underset{\sim}{\theta}_n,\ \lambda^{-1}\lambda_n^{-1}\right].$$

Thus

(i) $$(\underset{\sim}{\theta}_n | D_n, \lambda,\ \lambda_n) \sim N\left[\underset{\sim}{m}_n(\lambda_n),\ \lambda^{-1}\ C_n(\lambda_n)\right]$$

where $$\underset{\sim}{m}_n(\lambda_n) = \underset{\sim}{a}_n + P_n \underset{\sim}{h}_n (\lambda_n q_n^2 + 1)^{-1} \lambda_n (y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n),$$

$$C_n(\lambda_n) = P_n - P_n \underset{\sim}{h}_n \underset{\sim}{h}_n^T P_n (\lambda_n q_n^2 + 1)^{-1} \lambda_n,$$

and

$$q_n^2 = \underset{\sim}{h}_n^T P_n \underset{\sim}{h}_n;$$

(ii) $$(\lambda | D_n, \lambda_n) \sim G\left[\alpha_n/2, \beta_n(\lambda_n)\right]$$

where $$\alpha_n = \alpha_{n-1} + 1,$$

and

$$\beta_n(\lambda_n) = \beta_{n-1} + (y_n - \underset{\sim}{h}_n^T \underset{\sim}{a}_n)^2 (\lambda_n q_n^2 + 1)^{-1} \lambda_n.$$

(iii) From (i) and (ii),

$$p(\underset{\sim}{\theta}_n|D_n,\lambda_n) \propto \{\beta_n(\lambda_n) + (\underset{\sim}{\theta}_n - \underset{\sim}{m}_n(\lambda_n))^T C_n^{-1}(\lambda_n)(\underset{\sim}{\theta}_n - \underset{\sim}{m}_n(\lambda_n))\}^{-(p+\alpha_n)/2}$$

Clearly one-dimensional integrations over [0,1] will suffice to calculate moments of λ and $\underset{\sim}{\theta}_n$ given D_n exactly. To obtain simple analogues of the modal recursions we follow the ideas of earlier sections by replacing λ_n by an estimate $\tilde{\lambda}_n$, given by

$$\tilde{\lambda}_n = E\left[\lambda_n|y_n,\underset{\sim}{\theta}_n=\underset{\sim}{a}_n,\lambda=\ell_{n-1}\right]$$

$$= \psi_\nu\left[\ell_{n-1}^{\frac{1}{2}}(y_n - \underset{\sim}{h}_n^T\underset{\sim}{a}_n)\right].$$

From (i) and (ii) above, we then have the recursion

$$\underset{\sim}{m}_n = \underset{\sim}{m}_n(\tilde{\lambda}_n) = \underset{\sim}{a}_n + P_n\underset{\sim}{h}_n(\tilde{\lambda}_n q_n^2+1)^{-1}\ell_{n-1}^{-\frac{1}{2}} g_\nu\left(\ell_{n-1}^{\frac{1}{2}}(y_n - \underset{\sim}{h}_n^T\underset{\sim}{a}_n)\right) \qquad (4.3.13)$$

The equation for C_n, the approximate posterior variance, is derived as in §3.2.4 and §3.2.5, as

$$C_n = C_n(\tilde{\lambda}_n) + R_n \qquad (4.3.14)$$

where R_n is an extra term (which is such that $C_n > C_n(\tilde{\lambda}_n)$) given by

$$R_n = P_n\underset{\sim}{h}_n\underset{\sim}{h}_n^T P_n\, \phi_n\left(\ell_{n-1}^{\frac{1}{2}}(y_n - \underset{\sim}{h}_n^T\underset{\sim}{a}_n)\right),$$

and $$\phi_n(u_n) = -\psi_\nu'(u_n)\, u_n \left(1 + q_n^2\, \psi_\nu(u_n)\right)^{-2}.$$

Then we approximate $p(\underset{\sim}{\theta}_n|D_n,\lambda) \approx p(\underset{\sim}{\theta}_n|D_n,\lambda,\lambda_n=\tilde{\lambda}_n)$

$$\approx N\left[\underset{\sim}{m}_n, \lambda^{-1}C_n\right]. \qquad (4.3.15)$$

Further, for λ, $\quad p(\lambda|D_n) \approx p(\lambda|D_n, \lambda_n = \tilde{\lambda}_n)$

$$\approx G[\alpha_n/2,\ \beta_n/2], \tag{4.3.16}$$

where $\beta_n = \beta_n(\tilde{\lambda}_n) = \beta_{n-1} + (y_n - h_{\sim n}^T a_{\sim n}).(\tilde{\lambda}_n q_n^2 + 1)^{-1}\tilde{\lambda}_n$.

$$= \beta_{n-1} + (y_n - h_{\sim n}^T a_{\sim n})\ \ell_{n-1}^{-\frac{1}{2}}\ (\tilde{\lambda}_n q_n^2 + 1)^{-1} g_\nu\left[\ell_{n-1}^{\frac{1}{2}}(y_n - h_{\sim n}^T a_{\sim n})\right]$$

The following figures provide some numerical examples concerning the model

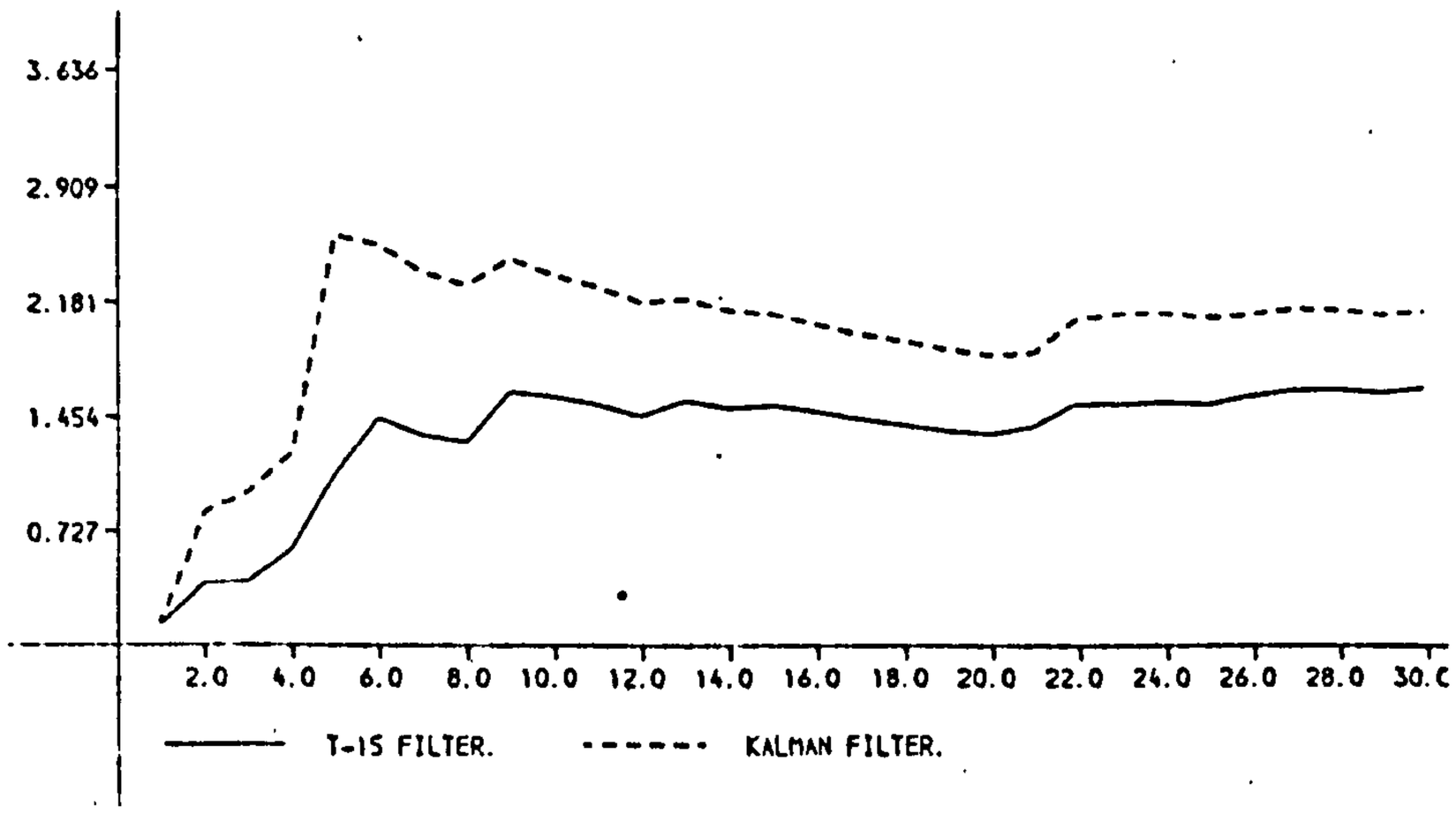
$$y_n = \theta_n + \nu_n,$$

$$\theta_n = \theta_{n-1} + \omega_n, \qquad n = 1,2,\ldots$$

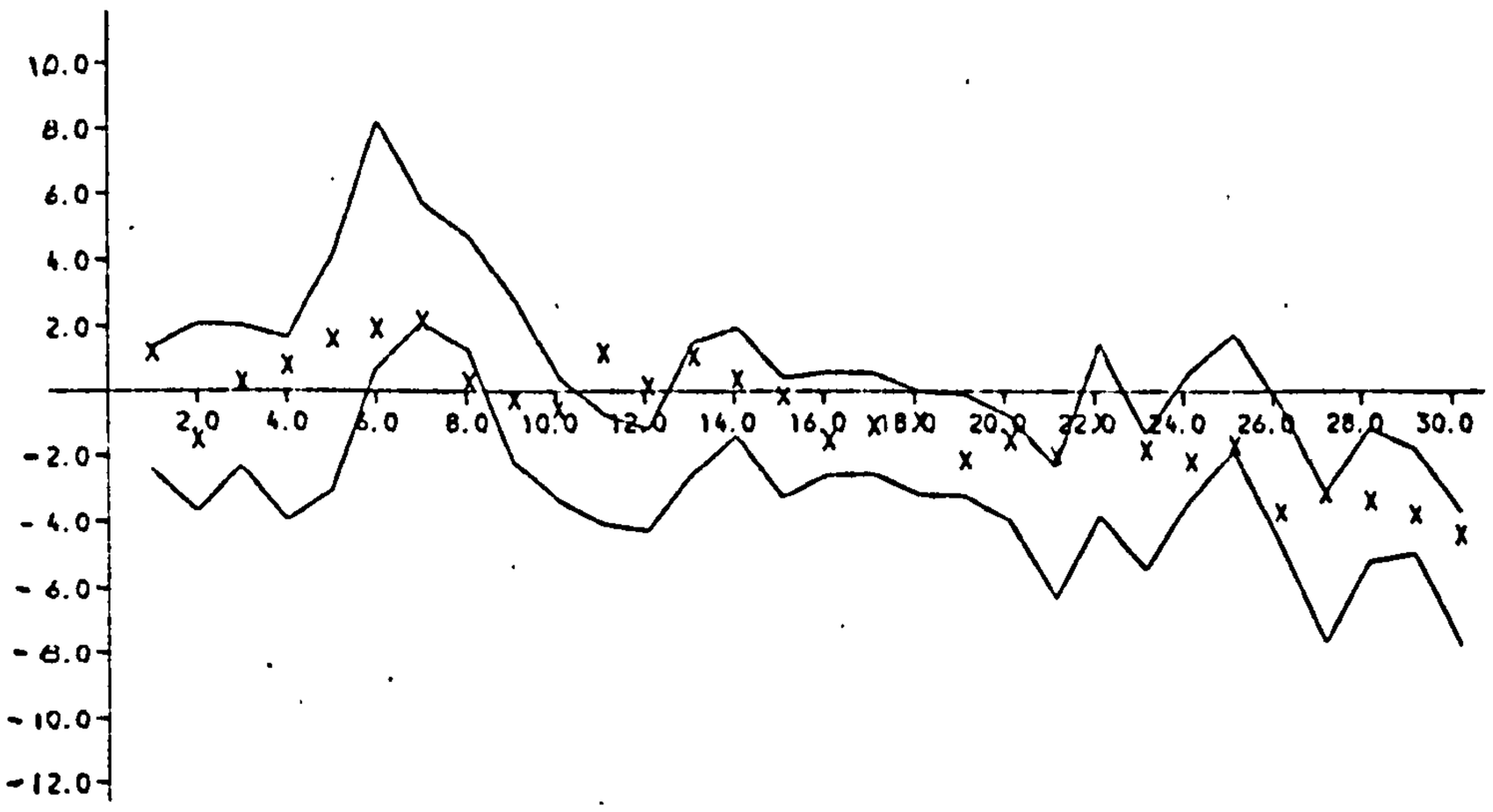
with $\omega_n \sim N[0,\omega]$ and for variour error densities p_ν as stated below each figure. The robust filters are based on Student t distributions for 5 and 15 degrees of freedom. The parameter R is the square of the ratio of the scale parameters, $R = \omega\lambda$. The upper frame in each set of three, A, provides a plot of the function β_n/α_n against n as an estimate of λ^{-1}. The frames B and C display 95% credible intervals for θ_n, which is plotted too.

We can again see the excellent performance of the robust filters on all sets of data and the "smoothing" effect of the choice of robustness parameter k (as the degrees of freedom of the model). The confidence intervals for k=5 are obviously wider than those for k=15. Again these figures are representative of more extensive numerical studies.

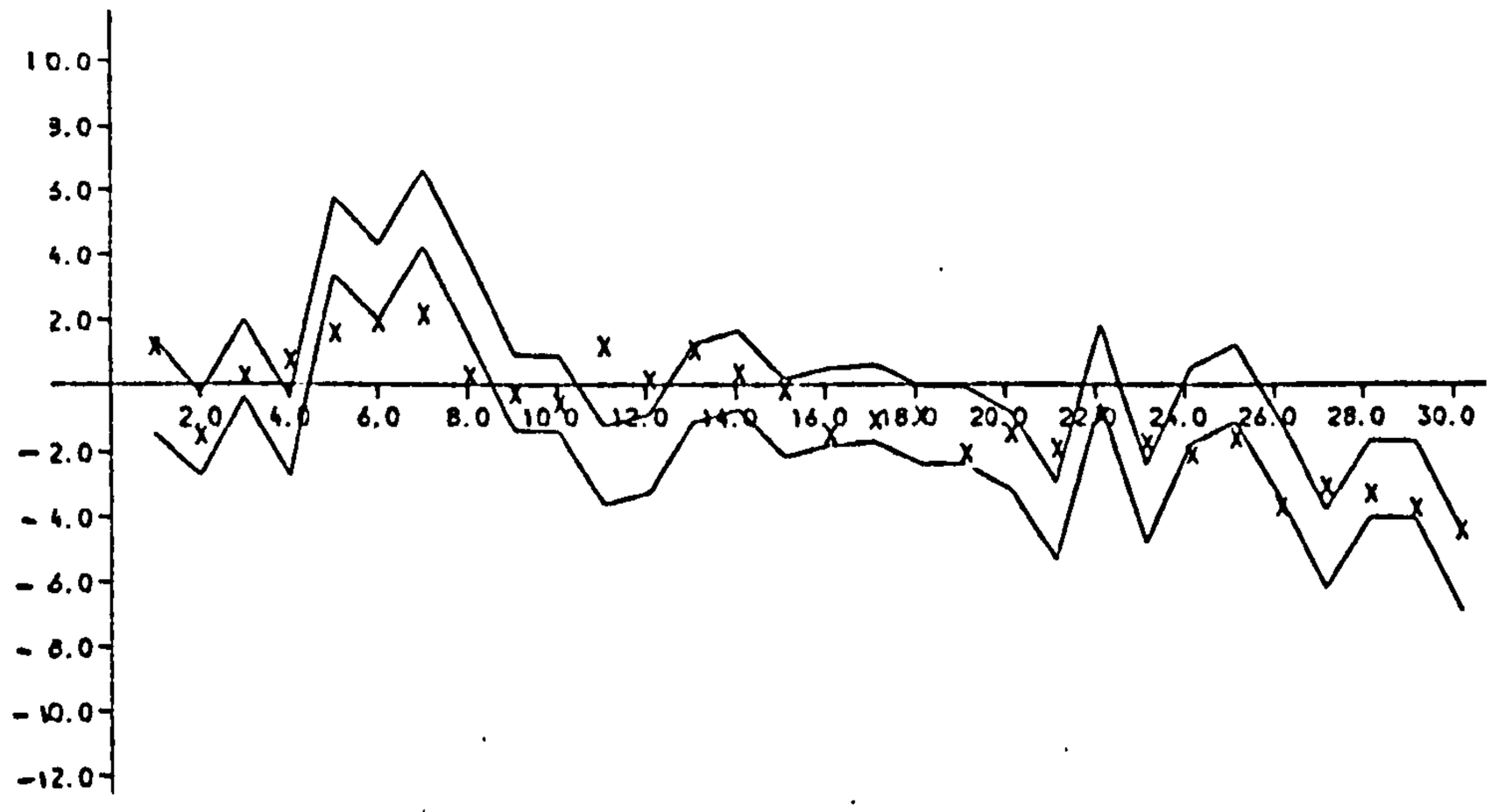
A.



B.



C.



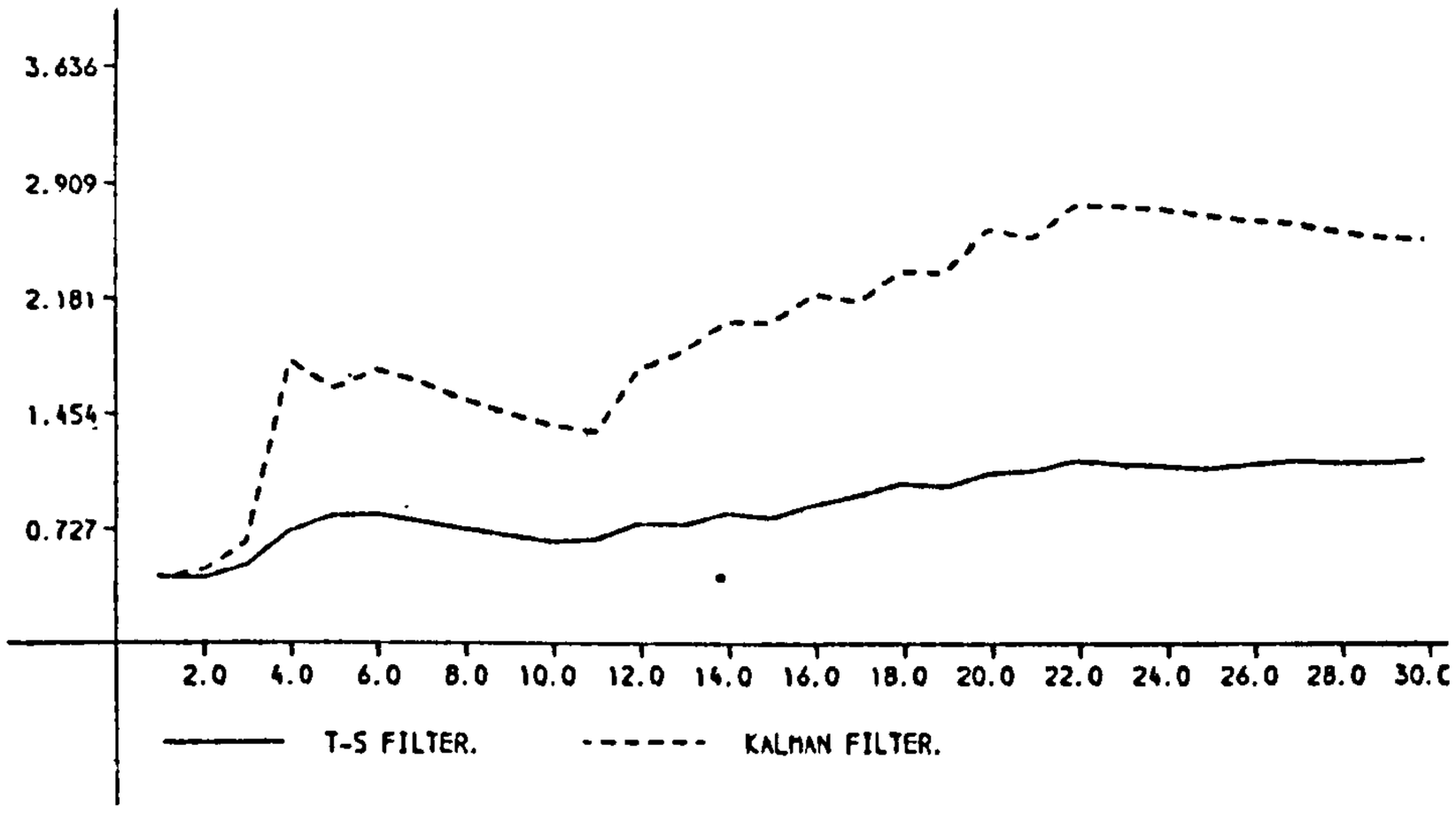
AR PROCESS (X) + DATA FROM N(0,9) - R=1

A. SCALE ESTIMATES

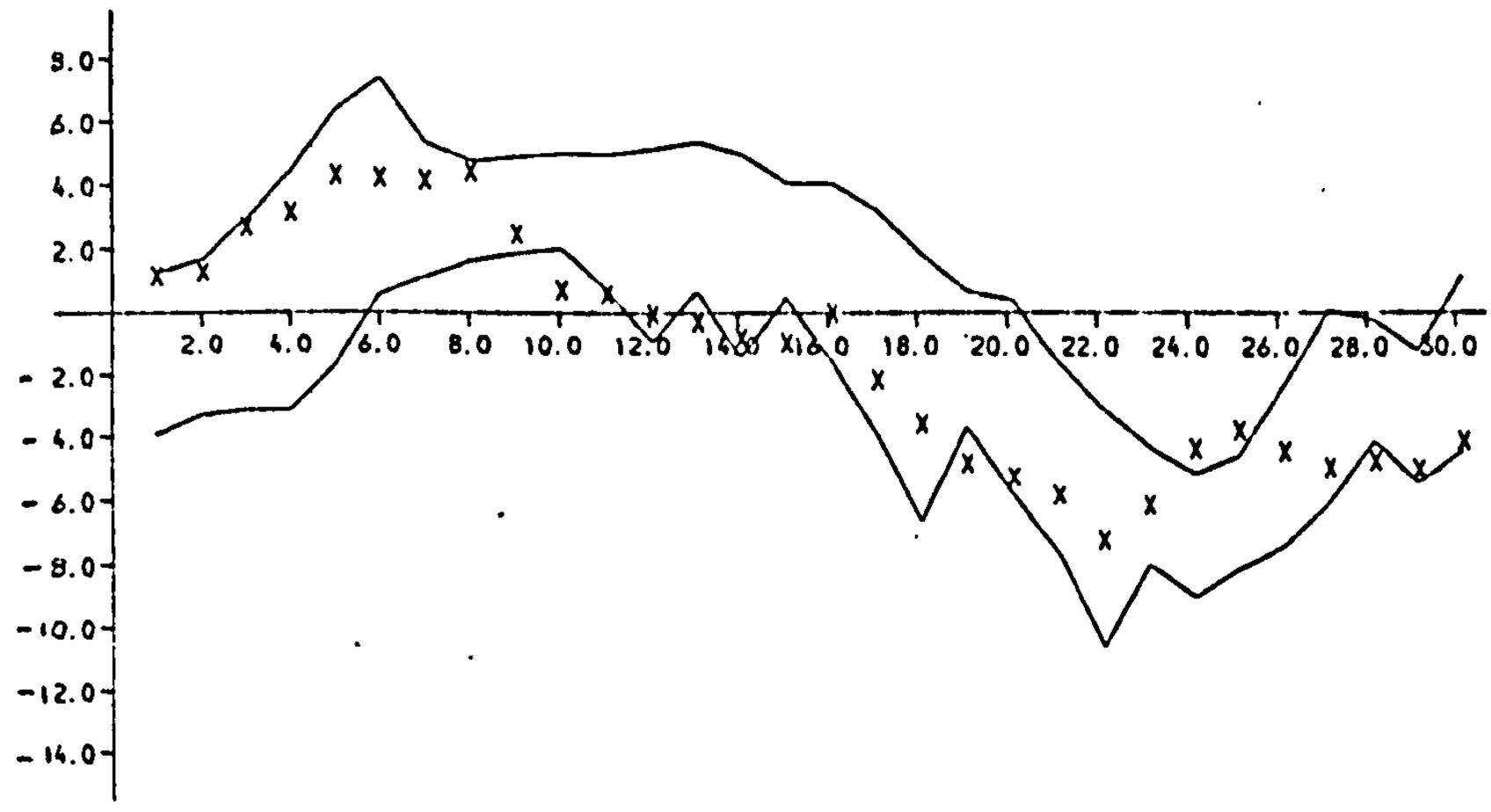
B. 95% INTERVAL - STUDENT T-15 FILTER + SCALE

C. 95% INTERVAL - KALMAN FILTER + SCALE

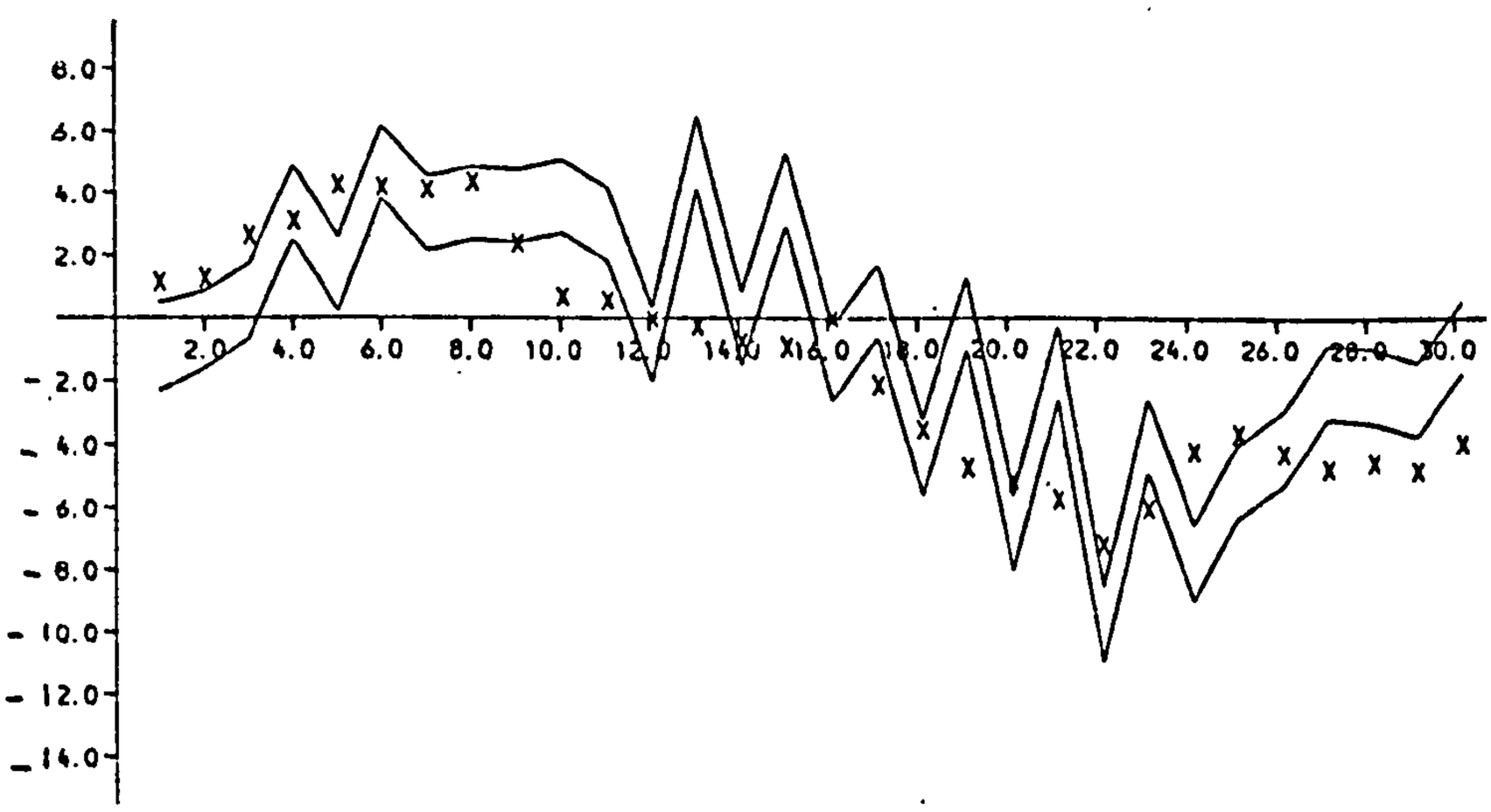
A.



B.



C.



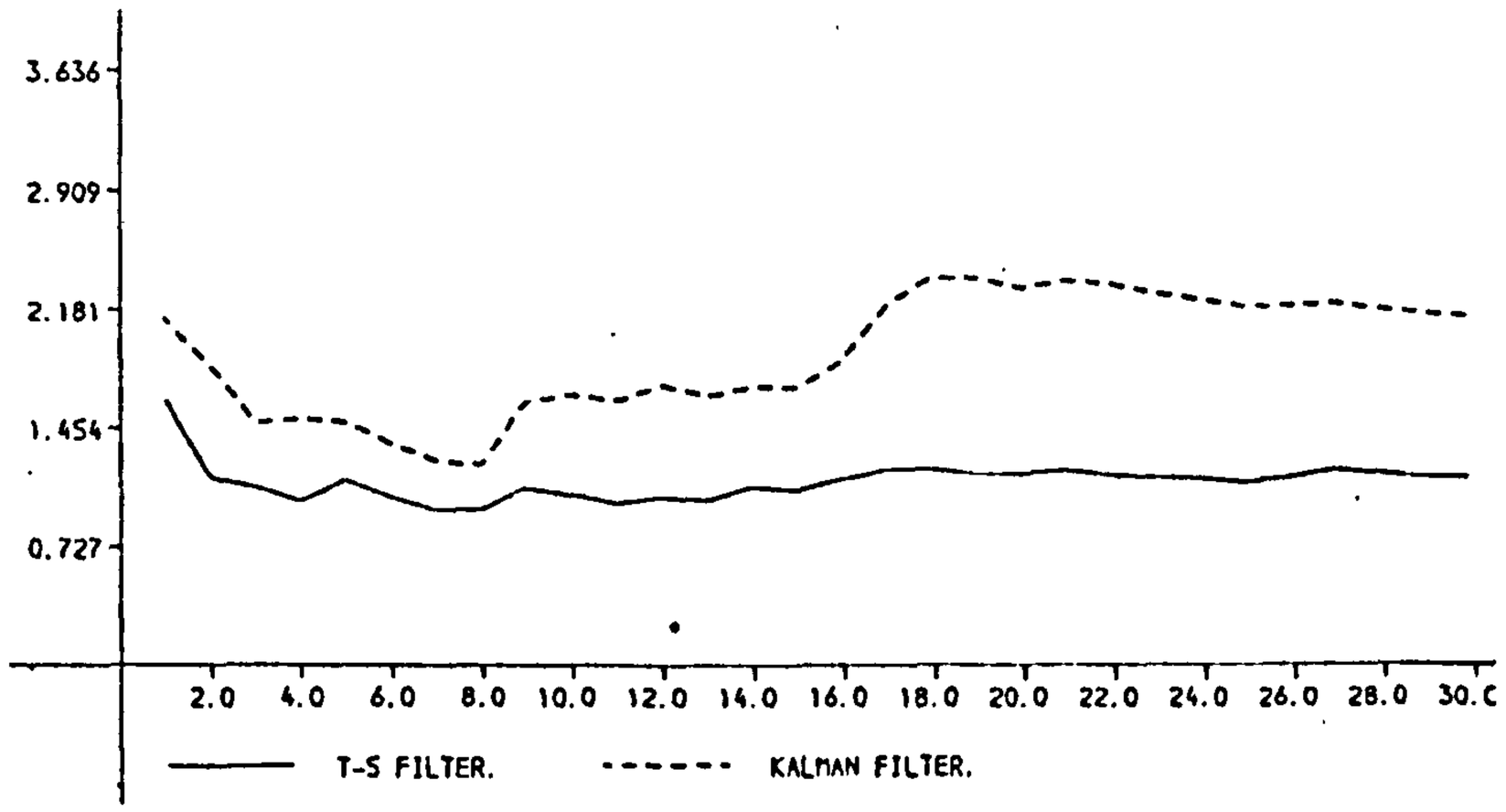
AR PROCESS (X) + DATA FROM N(0,9) - R=1

A. SCALE ESTIMATES

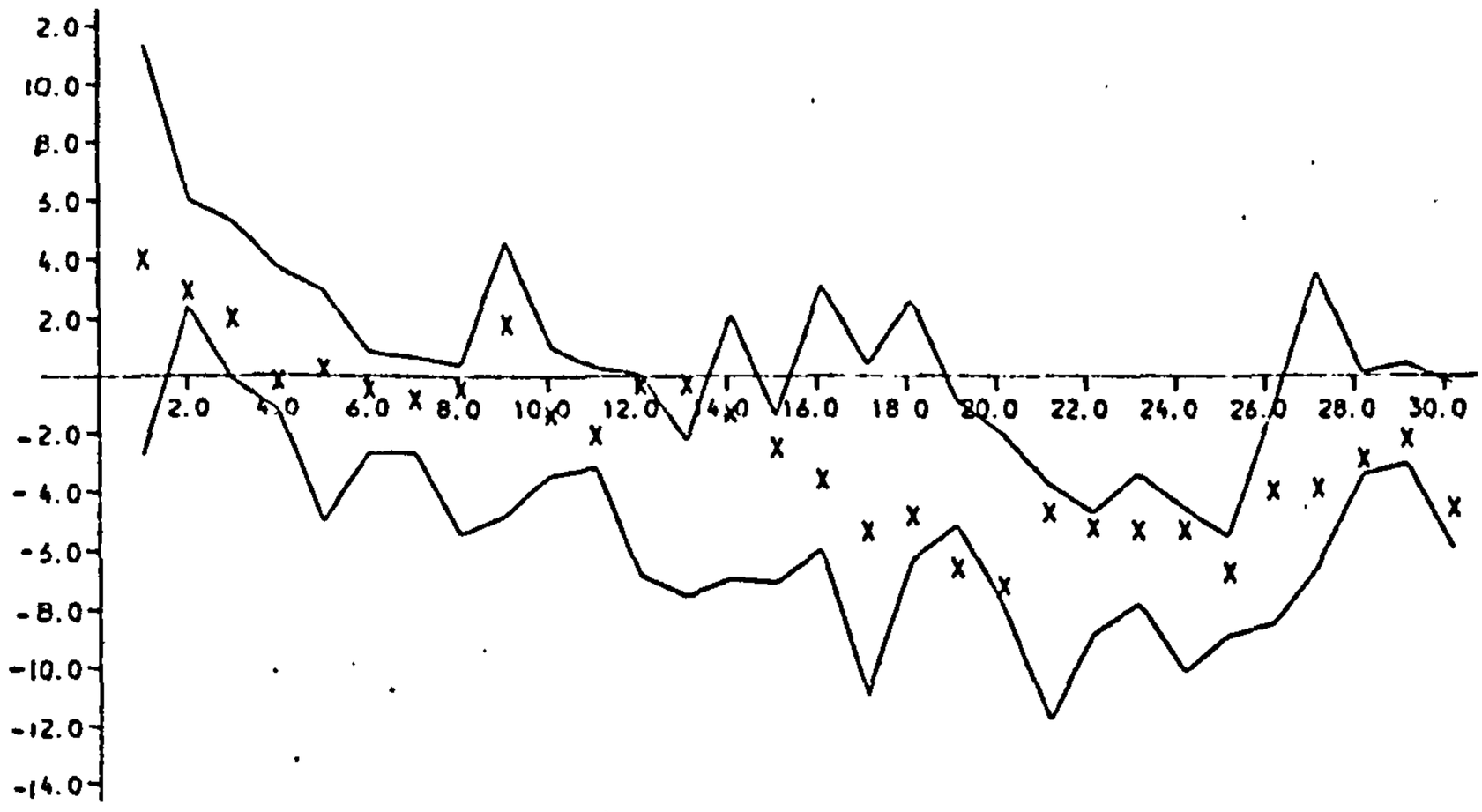
B. 95 INTERVAL - STUDENT T-S FILTER + SCALE

C. 95 INTERVAL - KALMAN FILTER + SCALE

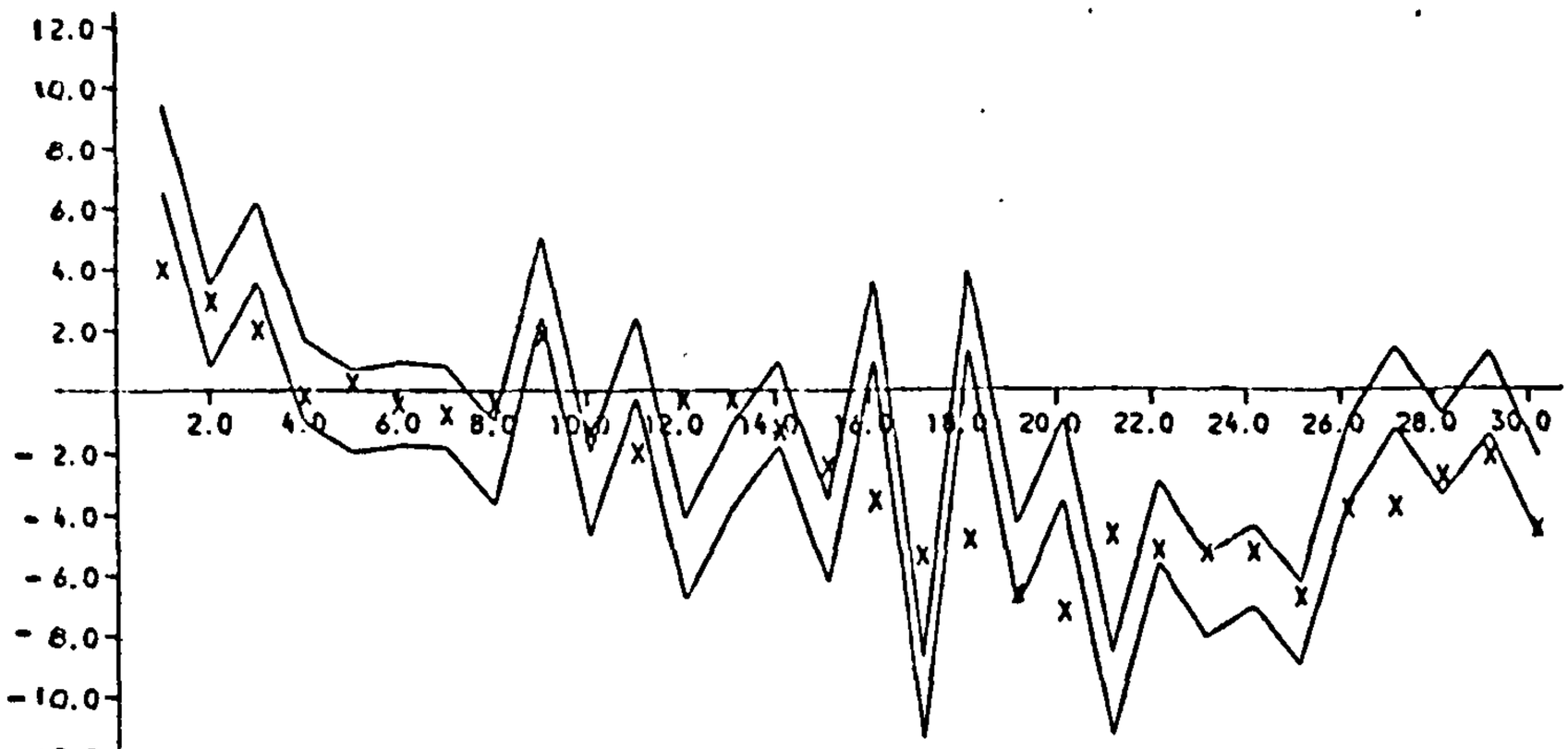
A.



B.



C.



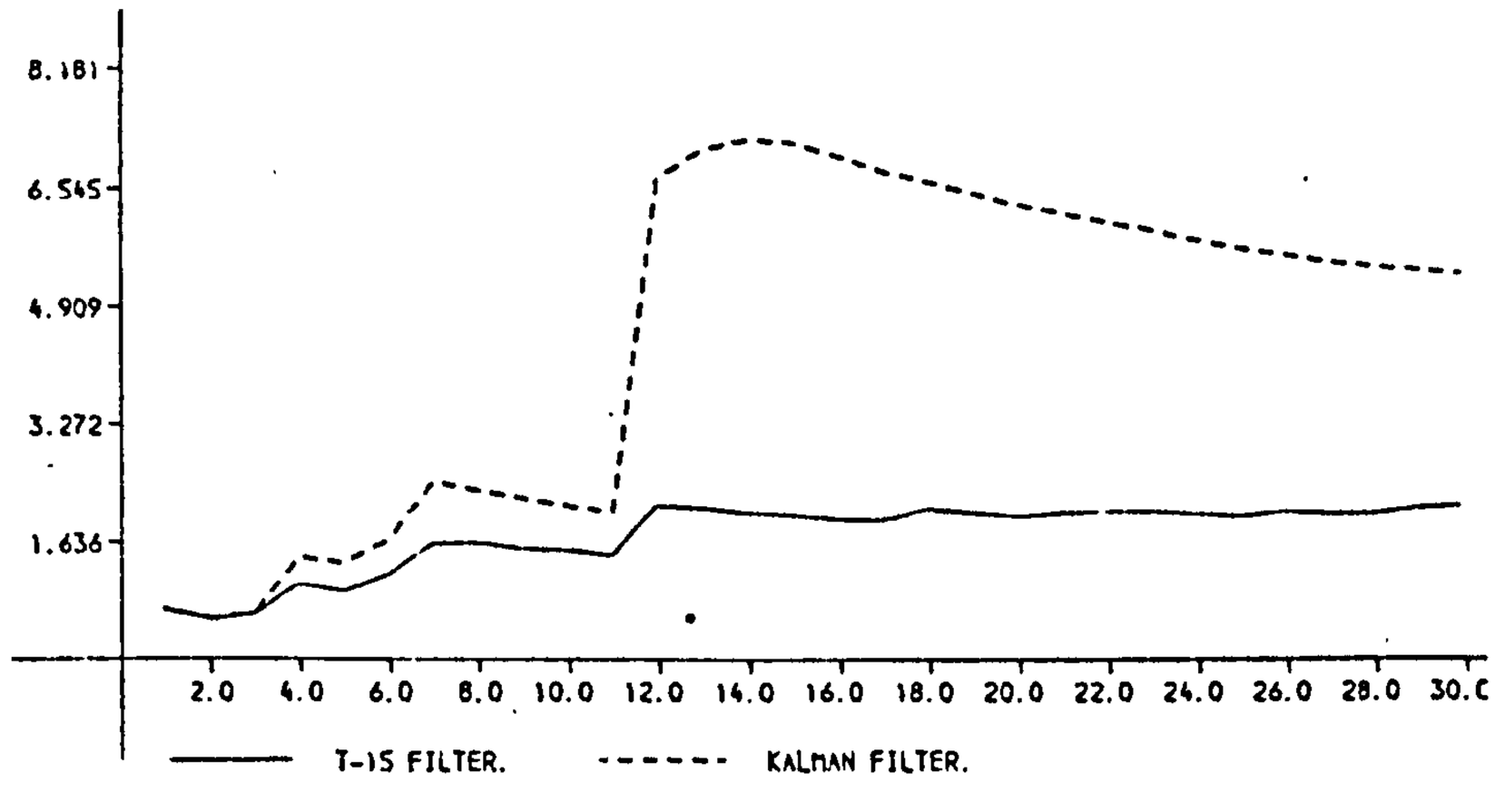
AR PROCESS (X) + DATA FROM N(0, 9) - R=3

A. SCALE ESTIMATES

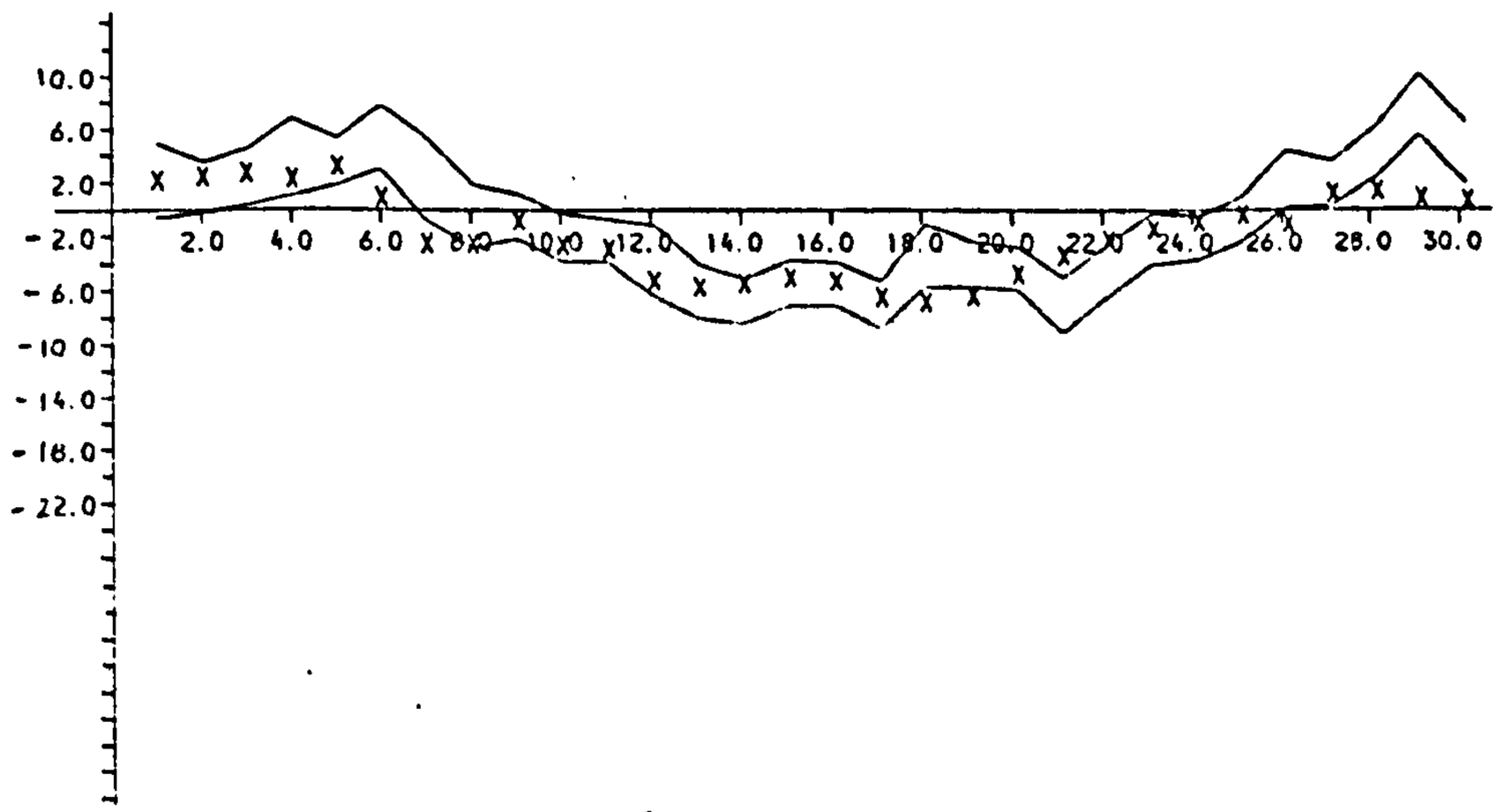
B. 95% INTERVAL - STUDENT T-S FILTER + SCALE

C. 95% INTERVAL - KALMAN FILTER + SCALE

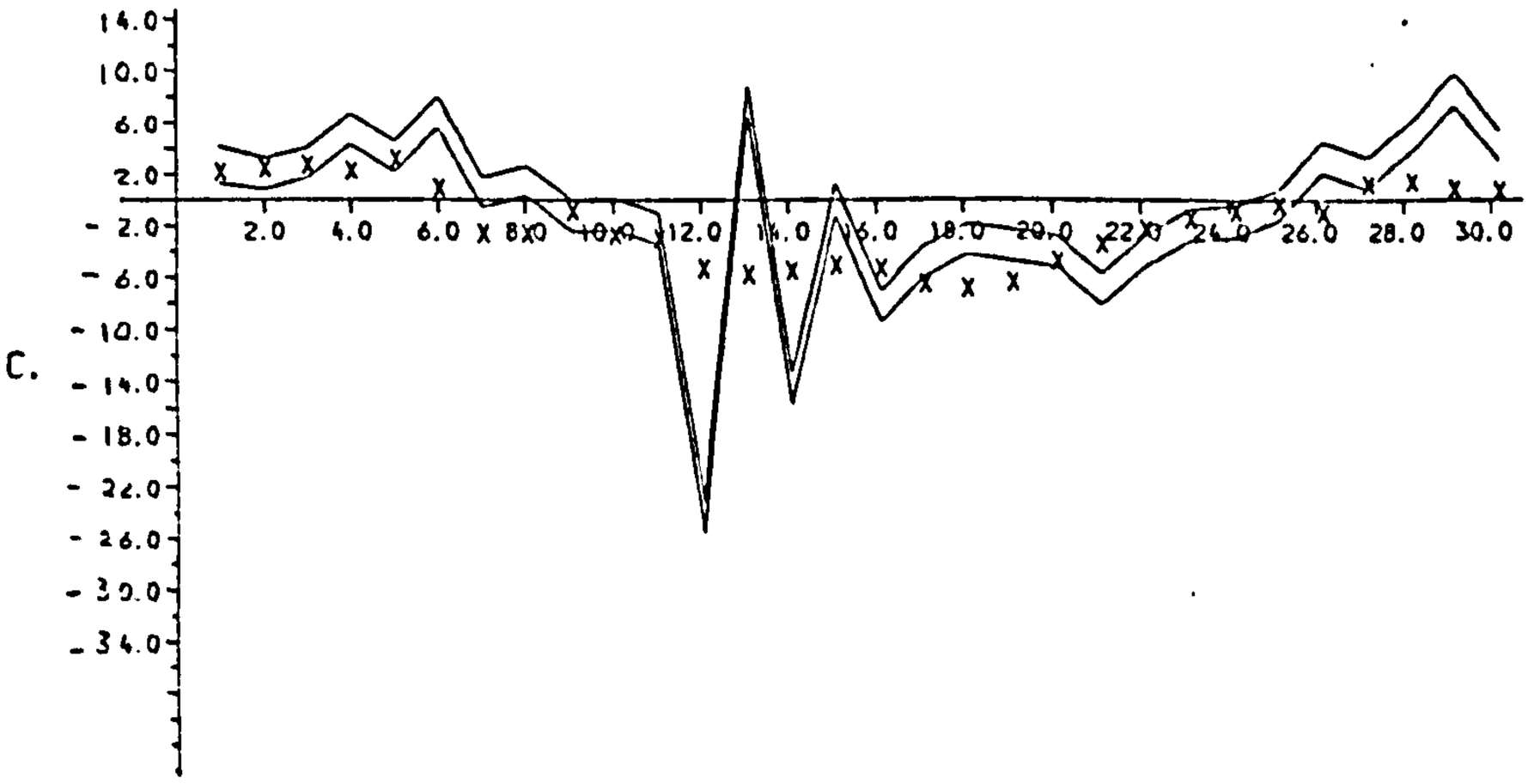
A.



B.



C.



AR PROCESS (X) + DATA FROM CAUCHY - R=1

A. SCALE ESTIMATES

B. 95% INTERVAL - STUDENT T-15 FILTER + SCALE

C. 95% INTERVAL - KALMAN FILTER + SCALE

4.3.2. Prediction and Smoothing.

(a) Prediction.

(i) The marginal density $p(y_n|D_{n-1})$.

We calculate $p(y_n|D_{n-1})$ just as in §3.3.1. From the modal recursion above we have

$$\beta_n = \beta_{n-1} + (y-h_{\sim n}^T a_{\sim n})^2 (\tilde{\lambda}_n q_n^2+1)^{-1} \tilde{\lambda}_n$$

with $$\tilde{\lambda}_n = \psi_\nu\left[\ell_{n-1}^{\frac{1}{2}}(y_n - h_{\sim n}^T a_{\sim n})\right].$$

Now, from (5.3.16) we have an approximation to the score of $p(y_n|D_n)$ by using our scale analogue (Theorem 2.3.1) of Masreliez's Theorem, just as we used the latter in §3.3.1.

From Theorem 2.3.1, we have, defining $u_n = y_n - h_{\sim n}^T a_{\sim n}$, that

$$E[\lambda|D_n] = \beta_{n-1}^{-1} \{\alpha_n - u_n\, g_1(u_n)\}$$

where

$$g_1(u_n) = -\frac{\partial}{\partial y_n} \ell_n\, p(y_n|D_n)$$

is the marginal score function.

Further from (4.3.16),

$$E[\lambda|D_n] \simeq \ell_n = \alpha_n\, \beta_n^{-1},$$

so, equating the two and rearranging we have

$$g_1(u_n) = \alpha_n\, \tilde{\lambda}_n\, u_n \{\beta_{n-1}(\tilde{\lambda}_n q_n^2+1) + \tilde{\lambda}_n\, u_n^2\}^{-1} \qquad (4.3.17)$$

This then defines the approximate marginal density via

$$p(y_n|D_n) \propto \exp\{-\int_{-\infty}^{\infty} g_1(u_n)du_n\}.$$

Example 4.3.2.

(i) Normal likelihood.

If $p_\nu(u) = \phi(u)$, then $\tilde{\lambda}_n = 1$ since $\omega(\lambda_n)$ is degenerate at $\lambda_n = 1$.

So $g_1(u_n) = \alpha_n\, u_n \{\beta_{n-1}(q_n^2+1) + u_n^2\}^{-1}$ which is the score of

$$p(y_n|D_n) \propto \{\beta_{n-1} + (q_n^2+1)^{-1}u_n^2\}^{-\alpha_n/2}$$

a scaled Student t distribution and the exact result.

(ii) Student t-k likelihood.

Now $\tilde{\lambda}_n = (k+1)\ (k+\ell_{n-1}u_n^2)^{-1}$, and so

$$g_1(u_n) = \alpha_n'\, u_n \{\gamma_n + u_n^2\}^{-1}$$

where

$$\alpha_n' = \alpha_n(k+1)\ (k+\alpha_n)^{-1} \quad ,$$

and

$$\gamma_n = \beta_{n-1}[q_n^2(k+1)+k]\ (k+\alpha_n)^{-1}.$$

Thus, again, $p(y_n|D_n)$ is a scaled Student t distribution although now with a different degrees of freedom parameter, α_n' ,

$$p(y_n|D_n) \propto \{\gamma_n+u_n^2\}^{-\alpha_n'/2}.$$

In the special case $q_n^2 = 0$, corresponding to $\underset{\sim}{\theta}_n$ known, and $\alpha_{n-1} = \beta_{n-1} = b$, we have

$$p(y_n|D_n) \propto \{1 + u_n^2/k'\}^{-(k'+1)/2}$$

where $k' = b\ k/(b+k+1)$.

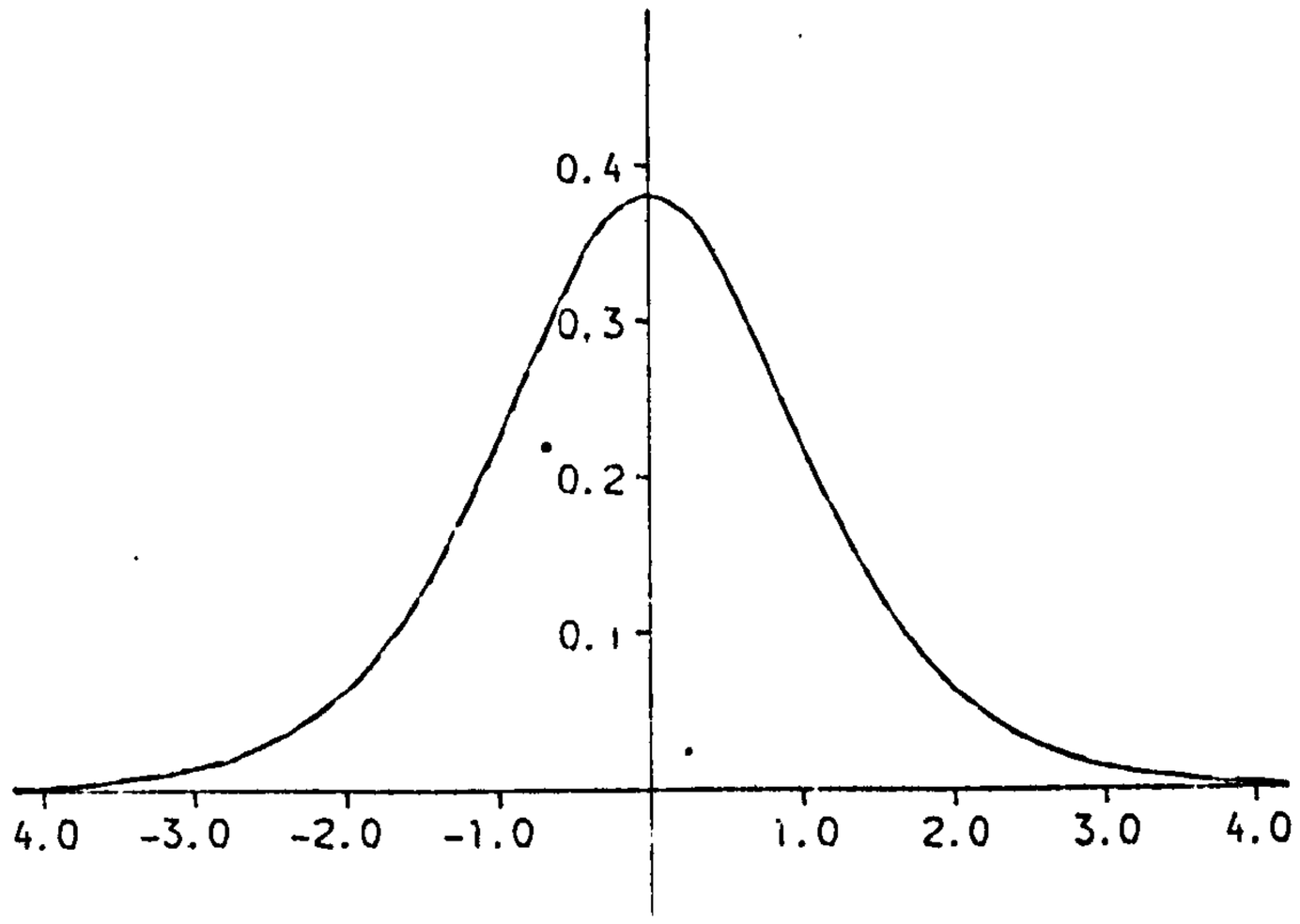
For various values of k we computed the true marginal density and this approximation for a range of b and these are displayed in the following figures.

STUDENT T-15

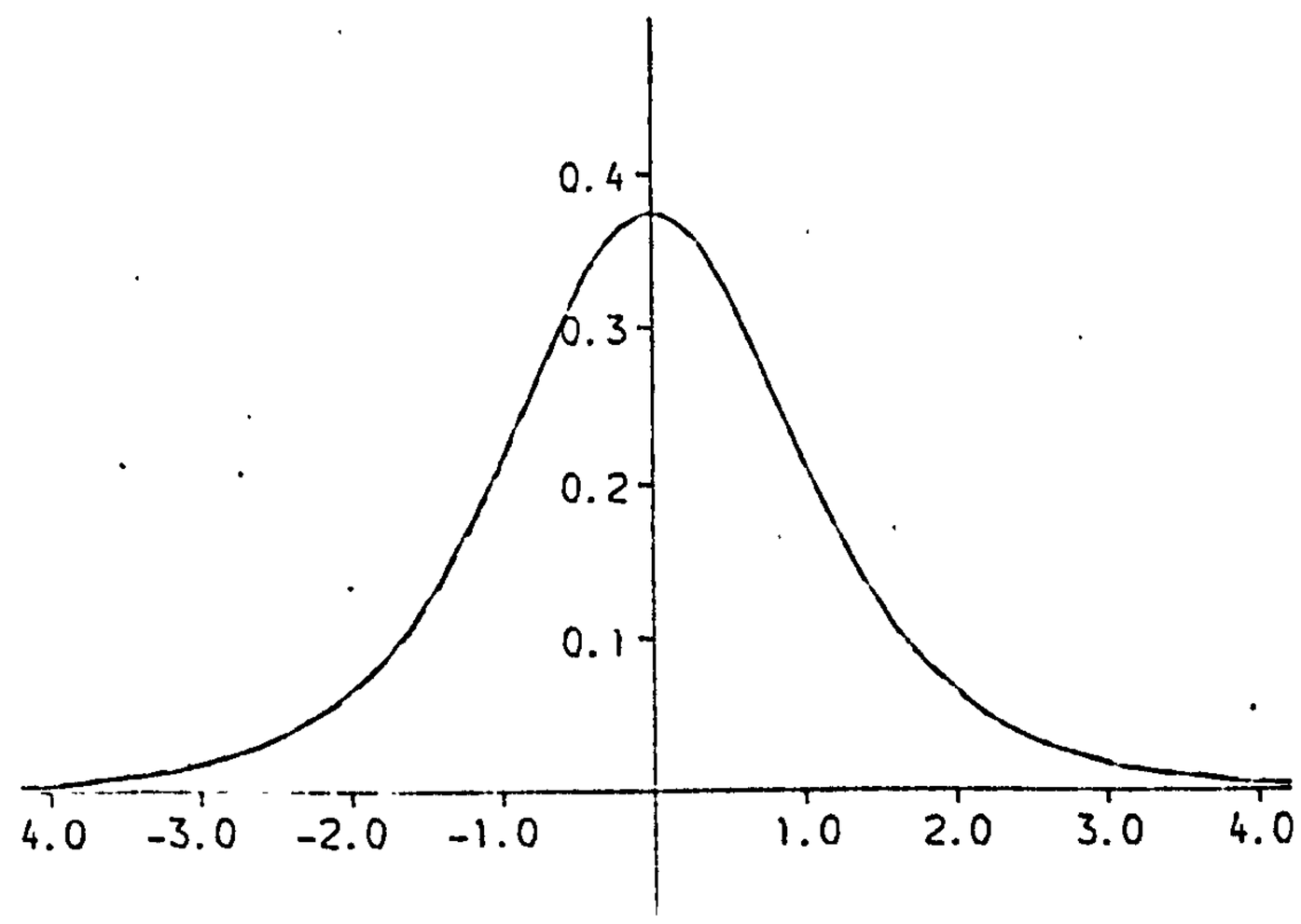
——— MARGINAL

- — - MODAL APPROXIMATION

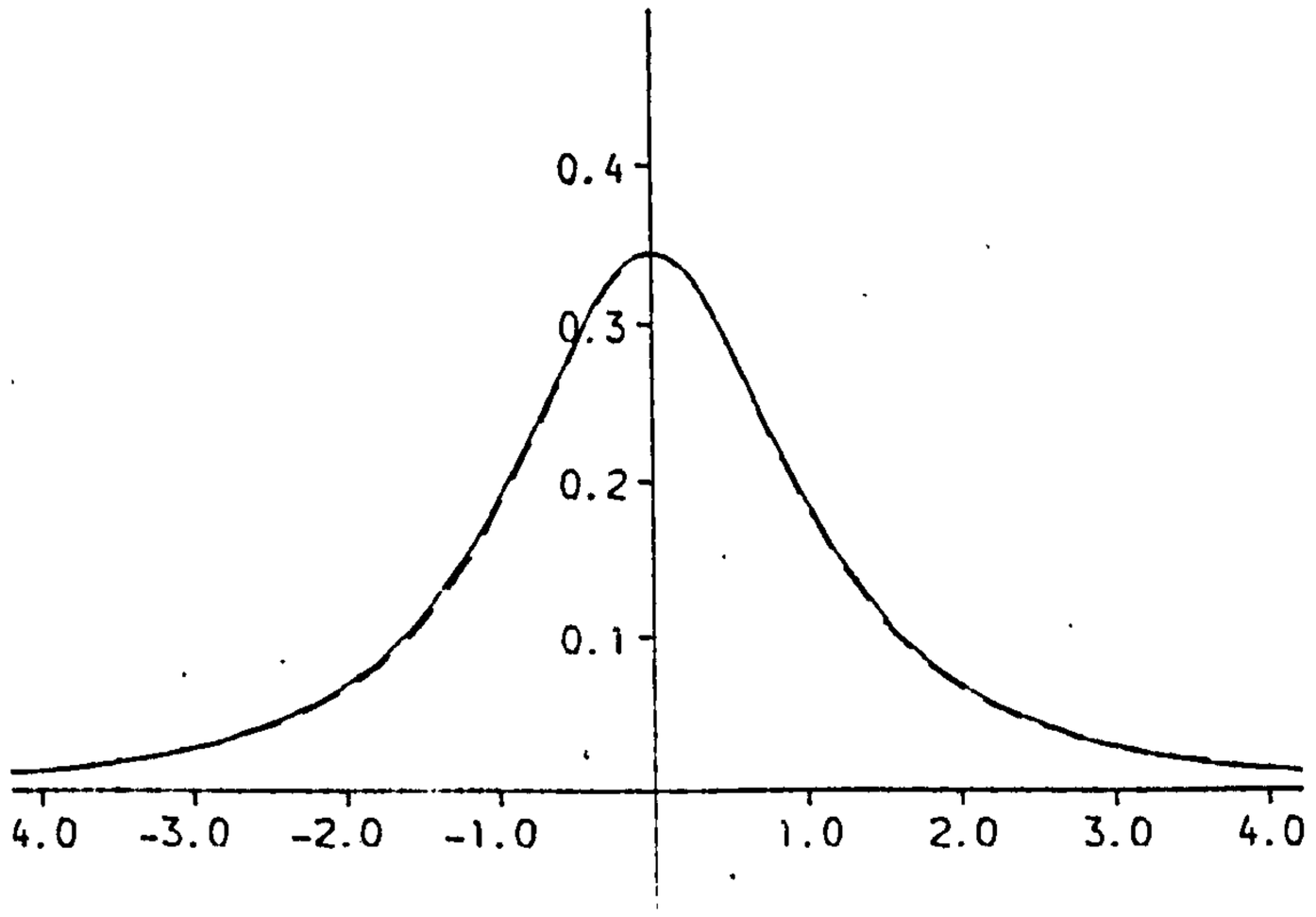
B=10



B=6



B=2

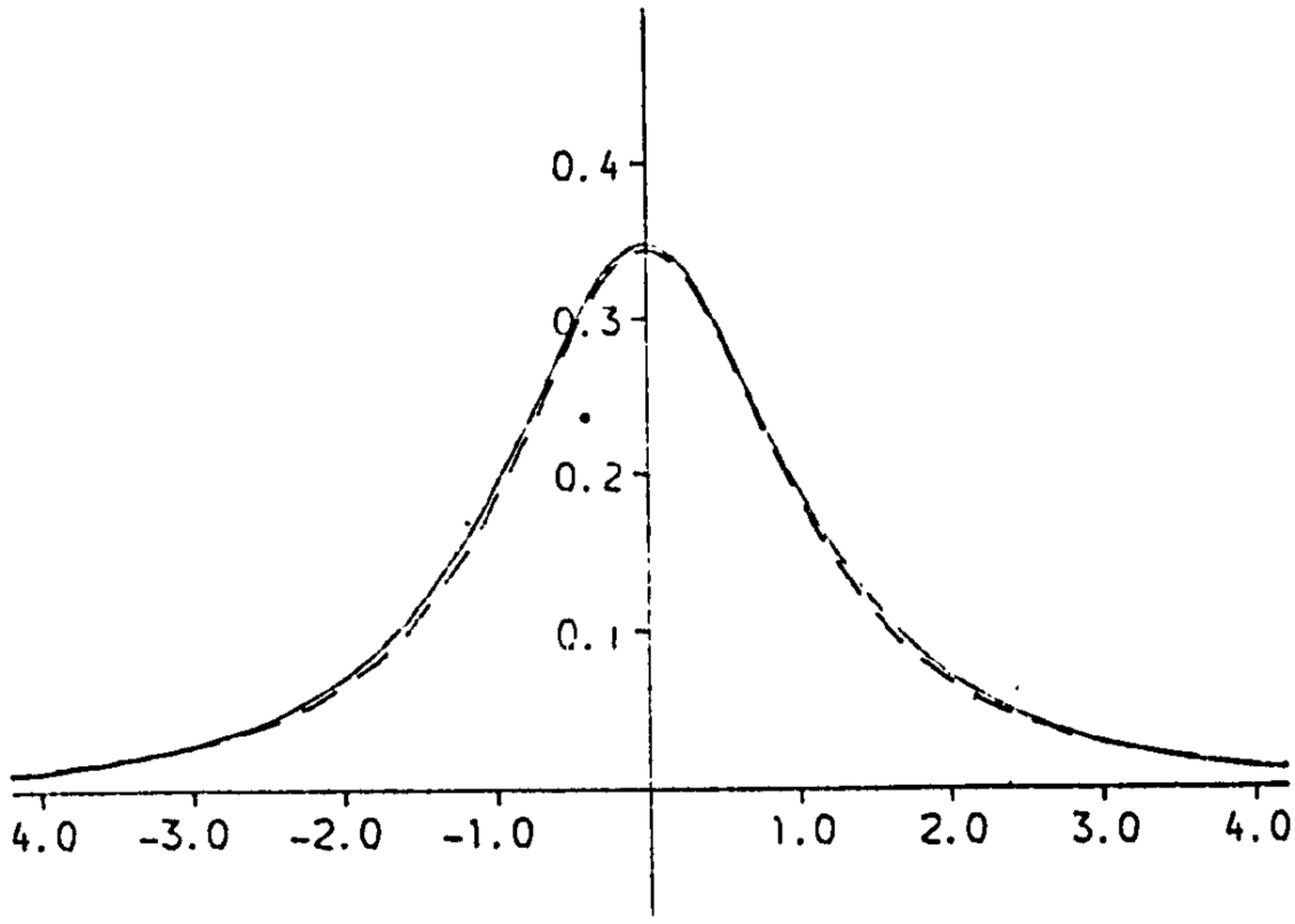


STUDENT T-5

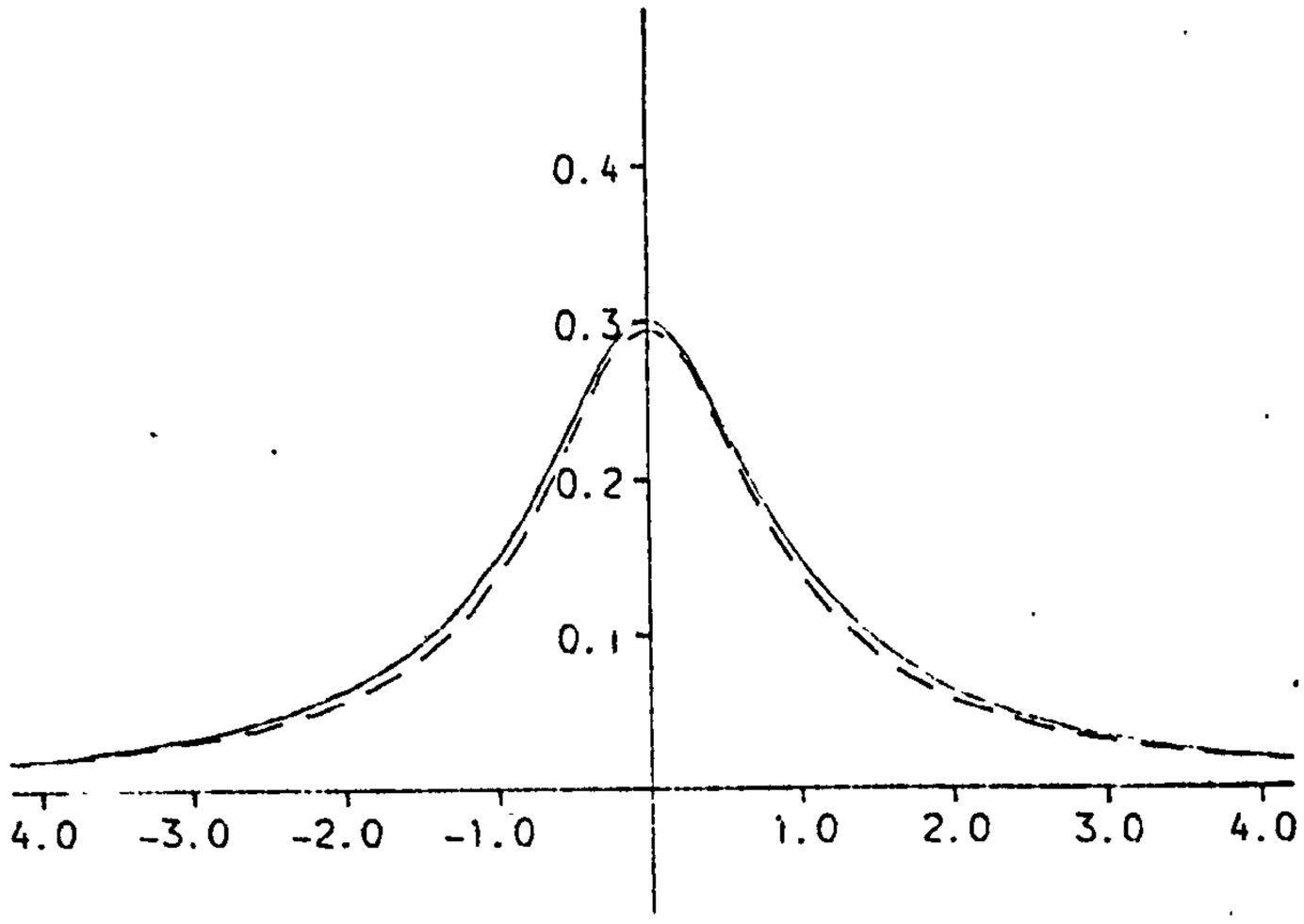
——— MARGINAL

– – – MODAL APPROXIMATION

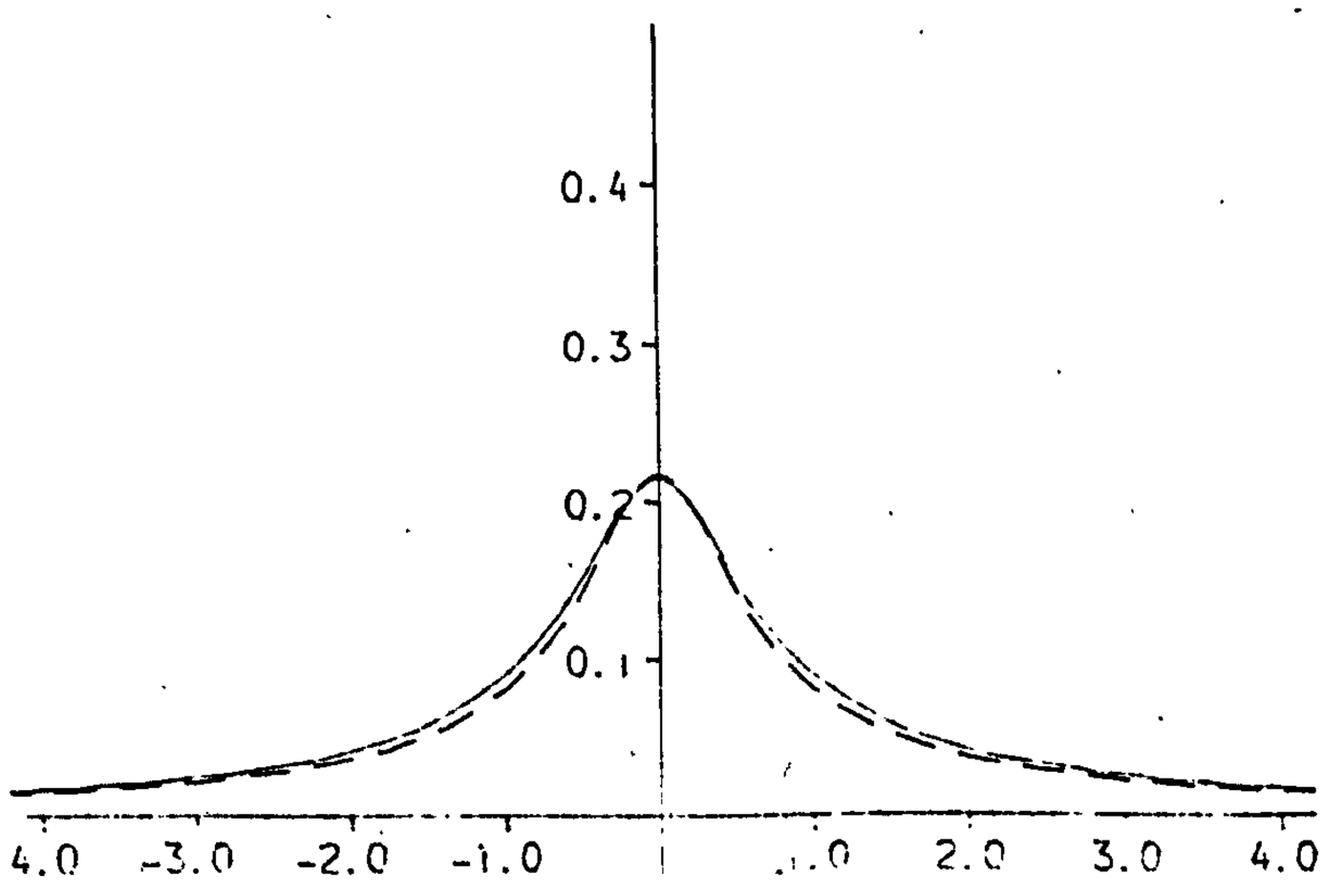
B=3



B=1



B=1/3



(ii) Predictive densities.

The calculation of predictive densities for future observations proceeds along the same lines.

For k>n,

$$p(y_k|D_n) = \int_0^\infty \int_{\mathbb{R}^p} \lambda^{\frac{1}{2}} p_\nu\left[\lambda^{\frac{1}{2}}(y_k - h_{\underset{\sim}{k}}^T \theta_{\underset{\sim}{k}})\right]\ p(\theta_{\underset{\sim}{k}}|\lambda, D_n) p(\lambda|D_n) d\theta_{\underset{\sim}{n}} d\lambda$$

Clearly $$(\theta_{\underset{\sim}{k}}|\lambda, D_n) \sim N\left[t_{\underset{\sim}{k}}^n, \lambda^{-1} T_k^n\right]$$

where, as in §3.3.1, for k = n+1, n+2,...

$$t_{\underset{\sim}{k}}^n = G_k\ t_{\underset{\sim}{k-1}}^n \text{ and } T_n^{\ n} = G_n\ T_{k-1}^n\ G_k^T + W_k.$$

Thus, defining

$$u_k^n = y_k - h_{\underset{\sim}{k}}^T t_{\underset{\sim}{k}}^n,$$

$$q_k^{n2} = h_{\underset{\sim}{k}}^T\ T_k^n\ h_{\underset{\sim}{k}},$$

and

$$\hat{\lambda}_k^n = \psi_\nu\left[\ell_n^{\frac{1}{2}} u_k^n\right],$$

we have, by analogy with (4.3.17),

$$-\frac{\partial}{\partial y_k} \ell n\ p(y_k|D_n) \simeq (1+\alpha_n)\hat{\lambda}_k^n \left\{\beta_n(\hat{\lambda}_k^n q_k^{n2}+1) + \hat{\lambda}_k^n\ u_k^{n2}\right\}^{-1}.u_k^n$$

Note that, both for the marginal density of $y_n|D_{n-1}$ and for these predictive densities, when $\theta_{\underset{\sim}{n}}$ is known as in §4.3.1(a), we have $q_n^2 = q_k^{n2} = 0$ for all k, and then the score above becomes

$$(1+\alpha_n)\ u_k^n \left\{\beta_n + u_k^{n2}\right\}^{-1}.$$

(b) Smoothing.

There is little to be said about the smoothed densities

$$p(\theta_{\underset{\sim}{k}}|D_n), \quad \text{for } k < n.$$

Invoking Theorem 3.3.1, we see that the appropriate approximation is

$$p(\underset{\sim}{\theta}_k|D_n,\lambda) = N\left[\underset{\sim}{t}^n_k,\ T^n_k\lambda^{-1}\right]$$

where $$\underset{\sim}{t}^n_k = \underset{\sim}{m}_k + C_k\ G^T_{k+1}\ P^{-1}_{k+1}\left[\underset{\sim}{t}^n_{k+1} - \underset{\sim}{a}_{k+1}\right],$$

and $$T^n_n = C_k - C_k\ G^T_{n+1}\ P^{-1}_{k+1}\left[P_{k+1} - T^n_{k+1}\right]\ P^{-1}_{n+1}\ G_{k+1}\ C_k.$$

Thus $$p(\underset{\sim}{\theta}_k|D_n) = \int_0^\infty p(\underset{\sim}{\theta}_k|D_n,\lambda)p(\lambda|D_n)d\lambda$$

$$\alpha\ \{\beta_n + (\underset{\sim}{\theta}_k - \underset{\sim}{t}^n_k)^T\ T^{n^{-1}}_k\ (\underset{\sim}{\theta}_k - \underset{\sim}{t}^n_k)\}^{-\alpha_n/2},$$

a scaled multivariate t density, exactly as in the case of a normal likelihood.

4.3.3. Vector observations: unknown covariance matrix.

Consider the model of §4.2.1 with now a heavy-tailed unimodal and elliptically symmetric error density

$$p_\nu(\underset{\sim}{u}|\Lambda) = |\Lambda|^{\frac{1}{2}}p_\nu(\Lambda^{\frac{1}{2}}\underset{\sim}{u}) = |\Lambda|^{\frac{1}{2}}f(\underset{\sim}{u}^T\Lambda\underset{\sim}{u}/2)$$

for some decreasing function f on $\mathbb{R}$, with $f > 0$.

We approach the problem of estimation of Λ via the construction of p_ν as a scale mixture of normal densities

$$p_\nu(\underset{\sim}{u}|\Lambda) = \int_0^\infty N\left[\underset{\sim}{0},\Lambda^{-1}\lambda^{-1}\right]\ \omega(\lambda)d\lambda$$

where ω is the mixing density. So for the model of §4.2.1, we assume that

$$\underset{\sim}{\nu}_n|\lambda_n \sim N\left[\underset{\sim}{0},\Lambda^{-1}\lambda_n^{-1}\right], \quad n = 1,2,\ldots$$

with

$$\{\lambda_n\} \text{ i.i.d. with density } \omega(\lambda).$$

Now we are in a position to use the development of §4.2.2. as follows:-

It at time n

(i) $(\Lambda|D_{n-1}) \sim W\left[\gamma_{n-1}, V_{n-1}\right]$

with

$$p(\Lambda|D_{n-1}) \propto |\Lambda|^{(\gamma_{n-1}-m-1)/2} \exp\{-\tfrac{1}{2}\,\mathrm{tr}(V_{n-1}\Lambda)\};$$

(ii) $(\underset{\sim}{\theta}_{n-1}|D_{n-1}) \sim N\left[\underset{\sim}{m}_{n-1},\ C_{n-1}\right]$

then the analogues of equations (4.2.17) to (4.2.20) lead to approximate posterior distributions conditional on λ_n are

(iii) $(\underset{\sim}{\theta}_n|D_n,\lambda_n) \sim N\left[\underset{\sim}{m}_n(\lambda_n),\ C_n(\lambda_n)\right]$

where

$$\underset{\sim}{m}_n(\lambda_n) = \underset{\sim}{a}_n + P_n H_n^T R_n^{-1}\,\psi(\underset{\sim}{u}_n)\underset{\sim}{u}_n \qquad (4.3.18)$$

with $\underset{\sim}{a}_n = G_n \underset{\sim}{m}_{n-1}$, $P_n = G_n\, C_{n-1}\, G_n^T + W_n$,

and $\underset{\sim}{u}_n = \underset{\sim}{y}_n - H_n \underset{\sim}{a}_n$, $R_n = H_n P_n H_n^T\,\psi(\underset{\sim}{u}_n) + \Lambda_{n-1}^{-1}$.

The function ψ is given by

$$\psi(\underset{\sim}{u}_n) = \lambda_n\,(\gamma_{n-1}+1).(\gamma_{n-1}+\lambda_n\,\underset{\sim}{u}_n^T\,\Lambda_{n-1}\underset{\sim}{u}_n)^{-1} \qquad (4.3.19)$$

where $\Lambda_{n-1} = E\left[\Lambda|D_{n-1}\right] = \gamma_{n-1}\,V_{n-1}^{-1}$.

Further

$$C_n(\lambda_n) = P_n - P_n H_n^T R_n^{-1} H_n P_n\,\psi(\underset{\sim}{u}_n) + S_n$$

where $S_n = P_n H_n R_n^{-1}\underset{\sim}{u}_n\underset{\sim}{u}_n^T R_n^{-1} H_n P_n\,(\gamma_{n-1}+\lambda_n\,\underset{\sim}{u}_n^T\,\Lambda_{n-1}\underset{\sim}{u}_n)^{-1}\lambda_n$.

(iv) $(\Lambda|D_n,\lambda_n) \sim W\left[\gamma_n,\ V_n\right]$

where

$$\gamma_n = \gamma_{n-1} + 1,$$

and

$$V_n = V_{n-1} + D_n(\Lambda_{n-1}|\lambda_n),$$

where the function D is defined by,

$$D_n(\Lambda|\lambda_n) = (I+\Lambda Q_n\lambda_n)^{-1}\left[\underset{\sim}{u}_n\underset{\sim}{u}_n^T + \lambda_n Q_n(I+\Lambda Q_n\lambda_n)\right](I+Q_n\Lambda\lambda_n)^{-1} \tag{4.3.20}$$

Now, as in 4.3.3(a) we produce modal recursions by using an estimate of λ_n, given by

$$\tilde{\lambda}_n = \psi_\nu\left[\Lambda_{n-1}^{\frac{1}{2}}\underset{\sim}{u}_n\right] = E\left[\lambda_n|y_n,\underset{\sim}{\theta}_n=\underset{\sim}{a}_n,\Lambda=\Lambda_{n-1}\right].$$

Thus the equation for $\underset{\sim}{m}_n$ is just $\underset{\sim}{m}_n = \underset{\sim}{m}_n(\tilde{\lambda}_n)$, given by

$$\underset{\sim}{m}_n(\tilde{\lambda}_n) = \underset{\sim}{a}_n + P_n H_n^T R_n^{-1}\psi(\underset{\sim}{u}_n)\underset{\sim}{u}_n$$

with now $\psi(\underset{\sim}{u}_n) = \tilde{\lambda}_n(\gamma_{n-1}+1)\ (\gamma_{n-1}+\tilde{\lambda}_n\underset{\sim}{u}_n^T\Lambda_{n-1}\underset{\sim}{u}_n)^{-1}$ and

$$R_n = H_n P_n H_n^T\psi(\underset{\sim}{u}_n) + \Lambda_{n-1}^{-1}.$$

from (4.3.19). The corresponding approximation to the posterior covariance matrix is given by

$$C_n = C_n - P_n H_n^T G(\underset{\sim}{u}_n)H_n P_n$$

where

$$G(\underset{\sim}{u}_n) = \frac{\partial}{\partial\underset{\sim}{u}_n}\ \{R_n^{-1}\psi(u_n)u_n\}\ .$$

Further details are routinely derived from these basic observations.

Appendix 4.

4A Lemma 4.1.

Let Λ, A, B be (m×m) symmetric matrices with Λ positive definite. Define the scalar functions f and g by

$$f(\Lambda) = \mathrm{tr}\left[(\Lambda^{-1}+A)^{-1}B\right]$$

and

$$g(\Lambda) = \ell n\left|\Lambda^{-1}+A\right|.$$

Then (i) $\frac{\partial f}{\partial \Lambda}(\Lambda) = (I+\Lambda A)^{-1}B\,(I+A\Lambda)^{-1}$

and (ii) $\frac{\partial g}{\partial \Lambda}(\Lambda) = -\Lambda^{-1}(\Lambda^{-1}+A)^{-1}\Lambda^{-1}$.

Proof. Define the (m×m) matrix X by $X = (\Lambda^{-1}+A)^{-1}$. Then, for all i,j,

$$\frac{\partial f}{\partial \Lambda_{ij}}(\Lambda) = \mathrm{tr}\left[\frac{\partial X}{\partial \Lambda_{ij}}\,B\right].$$

But $\Lambda^{-1}X + AX = I$, and so,

$$\frac{\partial \Lambda^{-1}}{\partial \Lambda_{ij}}\,X + (\Lambda^{-1}+A)\,\frac{\partial X}{\partial \Lambda_{ij}} = 0.$$

Now, to calculate $\frac{\partial \Lambda^{-1}}{\partial \Lambda_{ij}}$ note that

$$\Lambda^{-1}\Lambda = I \text{ so } \frac{\partial \Lambda^{-1}}{\partial \Lambda_{ij}}\,\Lambda + \Lambda^{-1}\,\frac{\partial \Lambda}{\partial \Lambda_{ij}} = 0.$$

So, since $\frac{\partial \Lambda}{\partial \Lambda_{ij}} = \underset{\sim}{\ell}_i\,\underset{\sim}{\ell}_j^T$ where $\underset{\sim}{\ell}_i$ is a zero column vector with unity in the ith position, we have

$$\frac{\partial \Lambda^{-1}}{\partial \Lambda_{ij}} = -\Lambda^{-1}\underset{\sim}{\ell}_i\,\underset{\sim}{\ell}_j^T\,\Lambda^{-1}.$$

Hence

$$\frac{\partial X}{\partial \Lambda_{ij}} = X\,\Lambda^{-1}\,\underset{\sim}{\ell}_i\,\underset{\sim}{\ell}_j^T\,\Lambda^{-1}\,X$$

so

$$\frac{\partial f}{\partial \Lambda_{ij}}(\Lambda) = \mathrm{tr}\left[X\Lambda^{-1}\underset{\sim}{\ell}_i\underset{\sim}{\ell}_j^T\Lambda^{-1}XB\right]$$

$$= \mathrm{tr}\left[\underset{\sim}{\ell}_j^T\,\Lambda^{-1}XBX\Lambda^{-1}\underset{\sim}{\ell}_i\right]$$

$$= \underset{\sim}{\ell}_i \Lambda^{-1} X B X \Lambda^{-1} \underset{\sim}{\ell}_j ,$$

and therefore

$$\frac{\partial f}{\partial \Lambda} (\Lambda) = \Lambda^{-1} X B X \Lambda^{-1},$$

and the result (i) follows.

For (ii) note that

$$\frac{\partial g(\Lambda)}{\partial \Lambda_{ij}} = \underset{\ell,m}{\Sigma} \left(\frac{\partial g(\Lambda)}{(\partial \Lambda^{-1})_{\ell m}} \right) \frac{\partial (\Lambda^{-1})}{\partial \Lambda_{ij}} \ell m$$

$$= \text{tr} \left[\frac{\partial g(\Lambda)}{\partial \Lambda^{-1}} \left(\frac{\partial \Lambda^{-1}}{\partial \Lambda} \right)_{ij}^{T} \right]$$

$$= \text{tr} \left[\frac{\partial g(\Lambda)}{\partial \Lambda^{-1}} \frac{\partial \Lambda^{-1}}{\partial \Lambda_{ij}} \right].$$

Now we know $\frac{\partial \Lambda^{-1}}{\partial \Lambda_{ij}} = - \Lambda^{-1} \underset{\sim}{\ell}_i \underset{\sim}{\ell}_j^T \Lambda^{-1}$, and, further,

$$\frac{\partial g(\Lambda)}{\partial \Lambda^{-1}} = \frac{\partial}{\partial \Lambda^{-1}} \ell n |\Lambda^{-1} + A| = (\Lambda^{-1} + A)^{-1} = X$$

and so

$$\frac{\partial g(\Lambda)}{\partial \Lambda_{ij}} = \text{tr} \left[-X \Lambda^{-1} \underset{\sim}{\ell}_i \underset{\sim}{\ell}_j^T \Lambda^{-1} \right]$$

$$= -\underset{\sim}{\ell}_i^T \Lambda^{-1} X \Lambda^{-1} \underset{\sim}{\ell}_j ,$$

$$\frac{\partial g(\Lambda)}{\partial \Lambda} = - \Lambda^{-1} X \Lambda^{-1}, \text{ as required.}$$

Appendix 4A: Lemma 4.2.

Let λ,ϕ be random variables with joint density

$$p(\lambda,\phi) = G_{\lambda}\left[a/2,b/2\right]\pi(\phi)$$

where b is a function of ϕ. Let $p(\lambda)$ be the marginal density of λ.

Define $f(\lambda) = G\left[a/2,\beta/2\right]$ and choose $\beta > 0$ such that

$$I(f,p) = \int_0^{\infty} \left\{p(\lambda)f(\lambda)^{-1}\right\}p(\lambda)d\lambda$$

is minimized as a function of β. I is the Kullback-Leibler directed divergence from f to p (Kullback, 1959) and

$$I(f,p) \geq 0 \text{ with equality if and only if } f = p$$

almost everywhere. Thus β satisfies

$$\frac{\partial I}{\partial \beta} = 0 \ ,$$

or

$$\frac{\partial}{\partial \beta} E\left[\frac{(a-2)}{2}\ell n(\lambda) + \frac{\beta\lambda}{2} - \frac{a}{2}\ell n\left(\frac{\beta}{2}\right) + \ell n\,\Gamma\left(\frac{a}{2}\right)\right] = 0.$$

Therefore

$$E\left[\frac{\lambda}{2} - \frac{a}{2\beta}\right] = 0$$

or

$$\frac{a}{\beta} = E[\lambda].$$

But

$$E[\lambda] = E\left[E[\lambda|\phi]\right] = E\left[a\ b(\phi)^{-1}\right]$$

and so β satisfies

$$\beta^{-1} = E\left[b(\phi)^{-1}\right].$$

CHAPTER 5. Classical time series models.

5.1. Autoregressive models.

As mentioned in Chapter 3 we must make a distinction between regressions in which the data enters into the matrix of regressors and those in which they do not. This distinction was noted in the context of autoregressions by Fox (1972) in a discussion of outliers in time series. The two basic models are described as follows, and following Kleiner et al (1979), we shall call them the innovations outliers (IO) model and the additive outliers (AO) model.

5.1.1. Innovations outliers.

Let
$$y_n = \sum_{j=1}^{p} \theta_j y_{n-j} + \nu_n \qquad (5.1.1)$$
be the observation equation. The so called innovation at time n is the observational error ν_n and a large innovation will have an effect on future observations since the aberrant observation will be used as a regressor. Suppose that, for example, y_n is uncontaminated with $\{\nu_n\}$ as i.i.d. $N[0,\sigma^2]$. Define the contaminated process $\{z_n\}$ by

$$\left.\begin{aligned} z_k &= y_k \quad , \ k < n, \\ z_n &= y_n + \delta, \quad \delta > 0 \\ z_k &= \sum_{j=1}^{p} \theta_j z_{k-j} + \nu_k, \ n < k. \end{aligned}\right\} \qquad (5.1.2)$$

Then we effectively have an outlying innovation, $\nu_n + \delta$, at time n.

Thus
$$z_{n+1} = y_{n+1} + \theta_1 \delta,$$
$$z_{n+2} = y_{n+2} + (\theta_1^2 + \theta_2)\delta,$$

and so on. In particular, for an AR(1) process,

$$z_{n+r} = y_{n+r} + \theta_1^r \delta, \quad \text{for } r > 0$$

and so the effect of the shift δ on the process decays as r increases with the actual observations $\{z_n\}$ given by the true process $\{y_n\}$ plus an exponentially decaying shift.

Abraham and Box (1979) discuss a retrospective Bayesian analysis of IO models within the framework of the "conditional model" described by Box (1979 , 1980). This approach assumes that outliers occur with some probability, α, say, and proceeds to calculate the posterior/predictive distributions of interest conditional upon knowing that a given subset $S = (y_{r_1},\ldots,y_{r_k})$ of the data are aberrant. The outliers are modelled by a non-zero shift, as with $\{z_n\}$ above, with the shift being the same for all outlying observations. Inferences are made by averaging the posterior/predictive distributions with respect to posterior probabilities of the given subsets S being aberrant. This procedure becomes computationally expensive with calculations required for each k = 1,2,..., and all possible subsets of size k, and is usually only performed for a small number of outliers, up to, say, 5% of the sample size, corresponding to a small α.

From the point of view of sequential estimation of course we cannot do this without performing a new analysis at each time point using all the data to that time. However, this innovations outlier model falls into the framework of Chapter 3, (although now, of course, we are taking $\underset{\sim}{\theta}_n = \underset{\sim}{\theta}$ to be fixed for all n; the general variable $\underset{\sim}{\theta}_n$ can be handled in the same way using the usual linear evolution equation). Given that we believe in the model (4.1.1) as the data generating mechanism, we need only adopt a heavy-tailed, near normal error density p_ν for the ν_n in order that outliers are automatically downweighted at the time of occurrence. However it is not clear that this limits the effect of the outlier at time n on $p(\underset{\sim}{\theta}|D_k)$, for k > n, for which the observed z_n is used as a fixed regressor. We examine the consequences in the special case of an

AR(1) process.

Special case p=1.

Now $\theta=\theta_1$ and the observations $\{z_n\}$ are related to the "clean" process $\{y_k\}$ by the equations

$$y_k = y_{k-1} + \nu_k \quad , \qquad k = 1,2,\ldots$$

$$z_n = y_n \quad , \qquad k < n$$

$$z_k = y_k + \theta^{n-k}\,\delta \quad , \qquad k \geq n,$$

where δ is the shift (assume $\delta > 0$ for clarity) at time n. Consider first the normal theory analysis, $p_\nu(u) = \phi(u)$.

(i) Kalman filter.

$$(\theta|D_n) \sim N[m_n, C_n]$$

where

$$m_n = m_{n-1} + C_{n-1}\, z_{n-1}(1+C_{n-1}z_{n-1}^2)^{-1}(z_n - z_{n-1}m_{n-1}) \qquad (5.1.3)$$

and

$$C_n^{-1} = C_{n-1}^{-1} + z_{n-1}^2. \qquad (5.1.4)$$

Clearly m_n is linear in δ, $m_n \to \infty$ with δ. C_n is constant however. So $p(\theta|D_n)$ moves along with the outlier, in the usual non-robust way associated with a normal likelihood.

Now at time n+1, we can rewrite m_{n+1} in the form

$$m_{n+1} = \left[C_{n-1}^{-1} + z_{n-1}^2 + z_n^2\right]^{-1} \left[C_{n-1}^{-1}m_{n-1} + z_n\, z_{n-1} + z_{n+1}\, z_n\right]$$

and, since $z_n = y_n + \delta$ and $z_{n+1} = y_{n+1} + \theta\delta$, we have

$$m_{n+1} = \left[\delta^{-2}\, C_{n-1}^{-1} + z_{n-1}^2\, \delta^{-2} + (y_n\, \delta^{-1}+1)^2\right]^{-1} \left[C_{n-1}^{-1}m_{n-1}\delta^{-2} + (y_n\delta^{-1}+1)\delta^{-1}z_{n-1} + (y_n\delta^{-1}+1)(y_{n+1}\delta^{-1}+\theta)\right]$$

$\to \infty$ as $\delta \to \infty$.

Further

$$C_{n+1}^{-1} = C_n^{-1} + z_n^2 \qquad \to \infty \text{ with } \delta.$$

Thus $p(\theta|D_{n+1})$ becomes degenerate about the <u>true value θ</u>, a rather remarkable observation. Of course in practice δ is finite.

<u>(ii) Robust filter.</u>

If p_ν is outlier-prone with score $g(u) = \psi(u)u$, (so $\psi(u)$ is bounded), then at time n we have, with a normal prior $(\theta|D_{n-1}) \sim N\left[m_{n-1}, C_{n-1}\right]$, that the posterior θ-score is

$$-\frac{\partial}{\partial\theta} \ell n\ p(\theta|D_n) = C_{n-1}^{-1}(\theta - m_{n-1}) - z_{n-1}(z_n - z_{n-1}\theta)\psi(z_n - z_{n-1}\theta)$$

So the posterior score converges to the prior score for all θ as δ (hence z_n) tends to infinity if $\psi(u)$ decays faster than u^{-1}. The posterior modal equation is

$$\theta = \left[C_{n-1}^{-1} + z_{n-1}^2\psi(z_n - z_{n-1}\theta)\right]^{-1} \left[C_{n-1}^{-1}\ m_{n-1} + z_{n-1}\ z_n\psi(z_n - z_{n-1}\theta)\right]$$

and the posterior mode (s) θ_n^* tend to m_{n-1} as δ tends to infinity when $\psi(u)$ decays faster than u^{-1}. Use of the modal recursions of Chapter 3 with, for example, Student t likelihoods, provides a robust analysis:

$$m_n = m_{n-1} + C_{n-1}\left(1 + C_{n-1}z_{n-1}^2\psi(u_n)\right)^{-1} g(u_n)$$

where $u_n = z_n - m_{n-1}$, implies $m_n \to m_{n-1}$ as $\delta \to \infty$. Also

$$C_n = C_{n-1} - C_{n-1}^2\ G_n(u_n),$$

with

$$G_n(u_n) = \frac{\partial}{\partial u_n}\left\{\left(1 + C_{n-1}^2 z_{n-1}^2\psi(u_n)\right)^{-1} g(u_n)\right\},$$

and $G_n(u_n) \to 0$ as $u_n \to \infty$ implies $C_n \to C_{n-1}$ as $\delta \to \infty$.

Going now to time n+1, the score of $p(\theta|D_{n+1})$ is given as

$$C_{n-1}^{-1}(\theta-m_{n-1}) - z_{n-1}(z_n-z_{n-1}\theta)\psi(z_n-z_{n-1}\theta) - z_n(z_{n+1}-z_n\theta)\psi(z_{n+1}-z_n\theta),$$

and so posterior mode(s) θ^*_{n+1} satisfy

$$\theta^*_{n+1}\left[C_{n-1}^{-1}+z_{n-1}^2\psi(z_n-z_{n-1}\theta^*_{n+1}) + z_n^2\psi(z_{n+1}-z_n\theta^*_{n+1})\right]^{-1}\left[C_{n-1}^{-1}m_{n-1}+z_n\, z_{n-1}\right.$$
$$\left.\psi(z_n-z_{n-1}\theta^*_{n+1}) + z_{n+1}z_n\psi(z_{n+1}-z_n\theta^*_{n+1})\right].$$

Since ψ is bounded, $\theta^* \to \theta$ with $\delta \to \infty$.

Use of the modal recursions similarly lead to $m_{n+1} \to \theta$ as $\delta \to \infty$, just as for the Kalman filter. However, unlike the Kalman filter, C_{n+1} does not necessarily tend to zero. Consider the Student t-k likelihood. Then

$$z_n^2 G_{n+1}(u_{n+1}) = (k+1)\left[k+C_n\, z_n^2(k+1) + u_{n+1}^2\right]^{-2}\left[k+C_n\, z_n^2(k+1) - u_{n+1}^2\right]z_n^2$$
$$= (k+1)\left[k\delta^{-2} + C_n(k+1)(y_n\delta^{-1}+1)^2 + \left[\epsilon_{n+1}\delta^{-1}+\theta-m_n\right]^2\right]^{-2}\cdot$$
$$\left[y_n\delta^{-1}+1\right]^2\left[k\delta^{-2} + C_n(k+1)(y_n\delta^{-1}+1)^2 - \left[\epsilon_{n+1}\delta^{-1}+\theta-m_n\right]^2\right]$$

where $u_{n+1} = z_{n+1} - z_n\, m_n = (y_{n+1}-m_n y_n) + \delta(0-m_n)$

and $\epsilon_{n+1} = y_{n+1} - m_n\, y_n$.

Therefore $\lim_{\delta\to\infty} C_{n+1} = \lim_{\delta\to\infty}\{C_n - C_n^2\, z_n^2 G_{n+1}(u_{n+1})\}$

$$= C_{n-1} - C_{n-1}^2(k+1)\,\frac{\left[C_{n-1}(k+1) - (\theta-m_{n-1})^2\right]}{\left[C_{n-1}(k+1) + (\theta-m_{n-1})^2\right]^2}$$

Note that this limit is non-zero unless $m_{n-1} = \theta$.

Example 5.1.1.

To illustrate consider the following numerical example of a sample of size 11 from an AR(1) process with $\theta = 0.5$. We took $m_0 = 0$ and $c_0 = 10$ and ran a Kalman filter together with a robust Student t-4 modal filter on the data, with an added shift of δ at time 10. Denote the Kalman filter mean and variance of $\theta|D_n$ by (m_n, c_n) and those of the robust filter by (x_n, s_n). After 9 "good" observations we had,

N = Normal theory posterior $N[0.353, 0.362]$,

R = Robust posterior $N[0.365, 0.495]$.

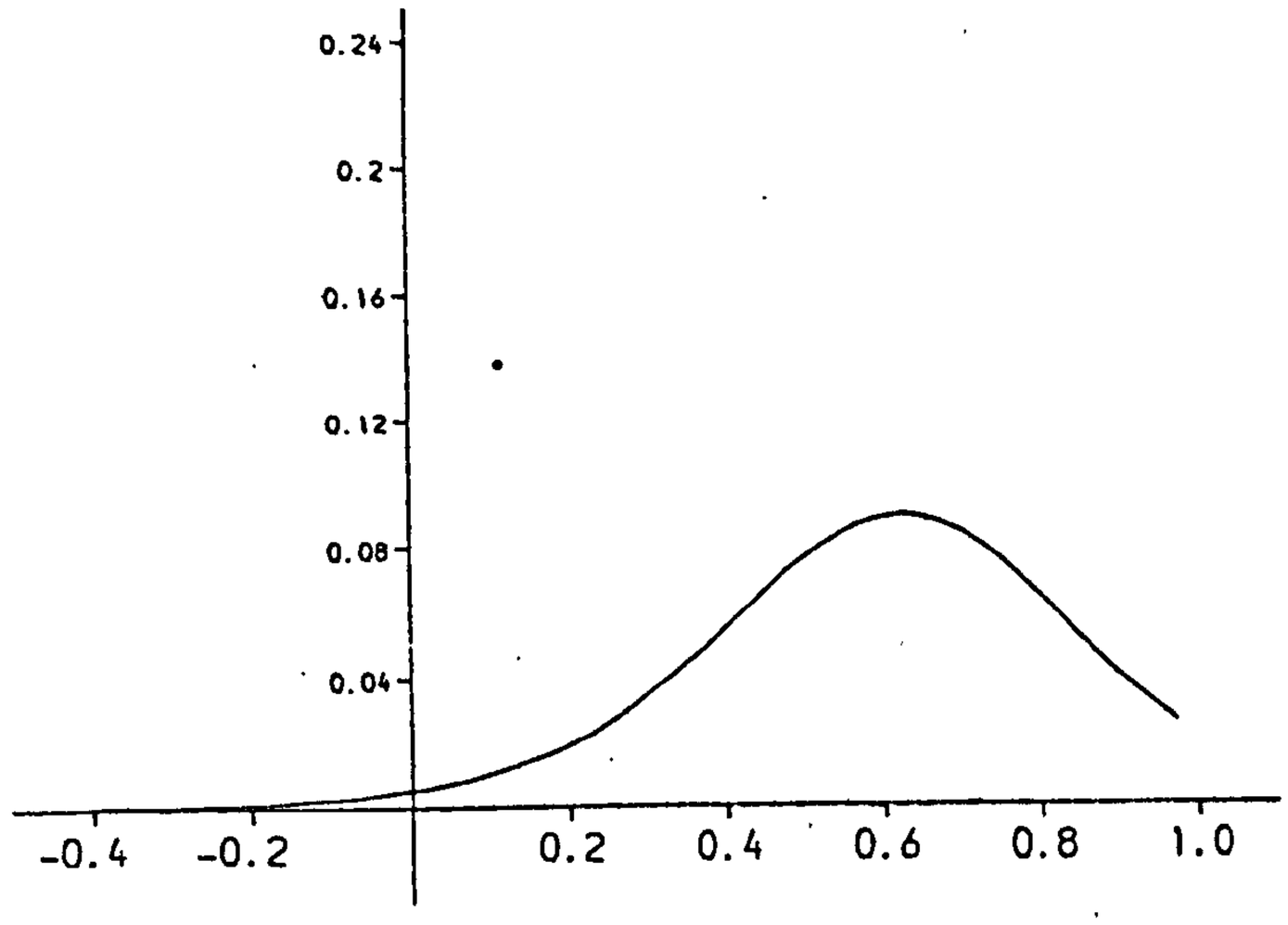
Shift		Time n: n = 10	Time n: n = 11
$\delta = 0$	N	(0.539, 0.339)	(0.535, 0.284)
	R	(0.621, 0.479)	(0.572, 0.335)
$\delta = 5$	N	(1.137, 0.339)	(0.604, 0.133)
	R	(0.538, 0.500)	(0.509, 0.127)
$\delta = 10$	N	(1.734, 0.339)	(0.575, 0.081)
	R	(0.471, 0.497)	(0.503, 0.078)
$\delta = 100$	N	(12.500, 0.339)	(0.510, 0.010)
	R	(0.380, 0.495)	(0.500, 0.094)

Example 5.1.2.

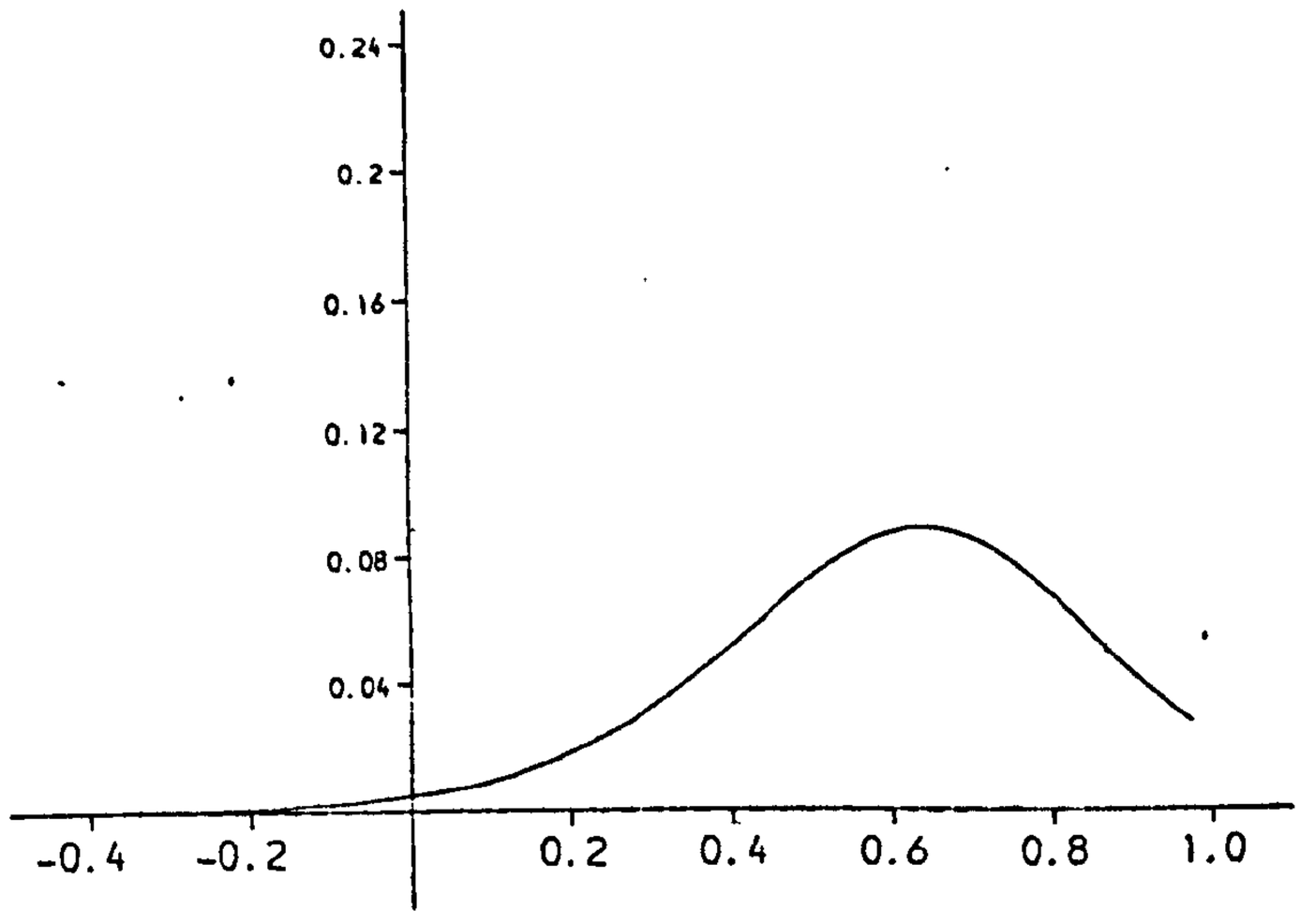
A second example is illustrated in Fig. 5.1. Using the same AR(1) process the posterior density is plotted at times n = 18, 19 and 20 and an innovations outlier is introduced at time n = 19. The likelihood was a Student t-4 density.

IO AT N=19
POSTERIOR AT TIME N

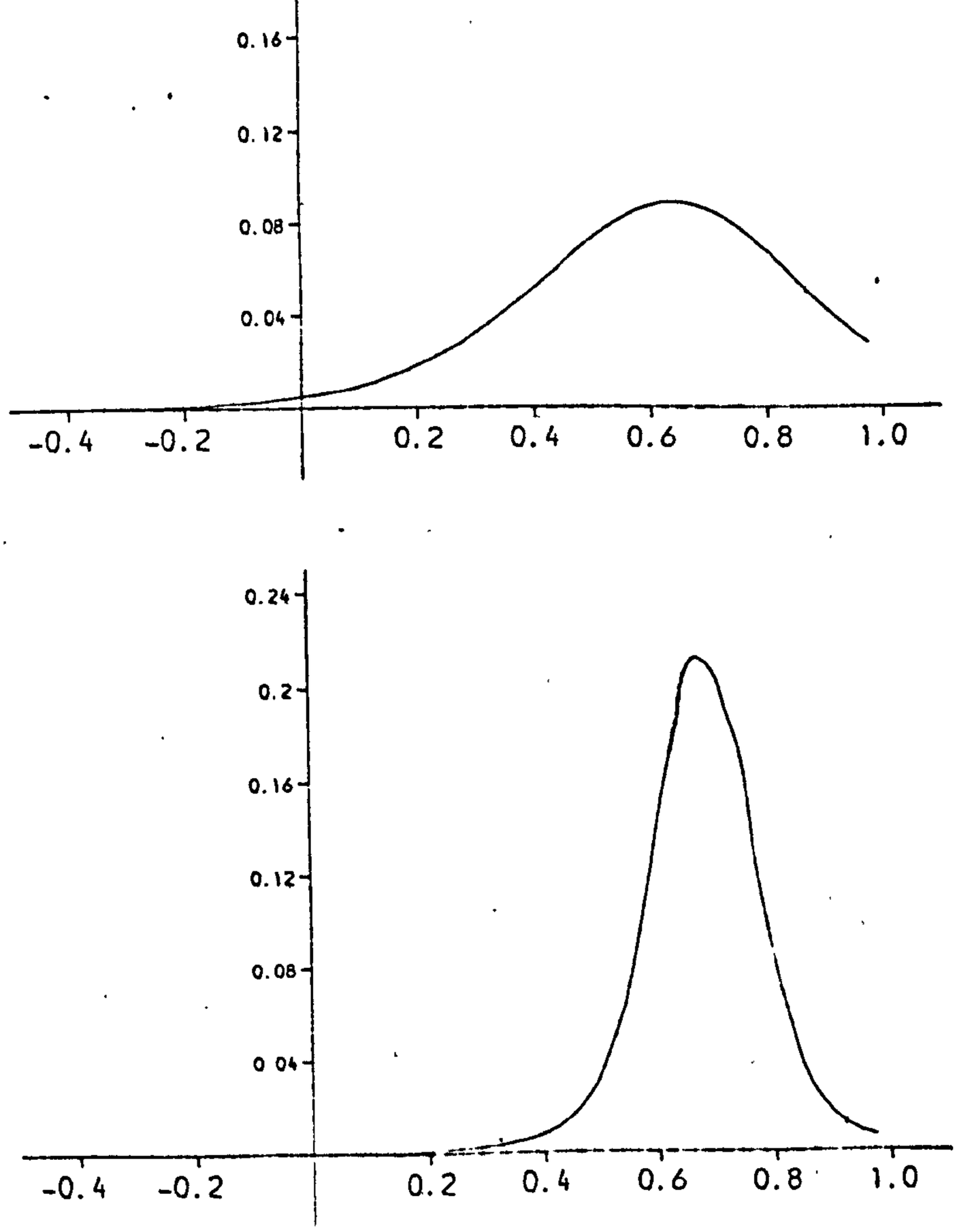
N=18



N=19



N=20



Unknown error variance.

The above assumes that the scale parameter of p_ν is known and unity. Now let $p(\nu_n) = \sigma^{-1} p_\nu(\sigma^{-1}\nu_n)$, $n = 1,2,\ldots$ with σ unknown. Define $\lambda = \sigma^{-2}$. Staying for definiteness with the simple AR(1) process the usual conjugate analysis for normal p_ν leads to

$$(\theta|\lambda, D_n) \sim N[m_n, \lambda^{-1} C_n]$$

and

$$(\lambda|D_n) \sim G[\alpha_n/2, \beta_n/2],$$

where m_n, C_n are defined in (4.1.3) and (4.1.4), $\alpha_n = \alpha_{n-1} + 1$ and

$$\beta_n = \beta_{n-1} + (z_n - z_{n-1} m_{n-1})^2 (1 + z_{n-1}^2 C_{n-1})^{-1}.$$

Clearly the usual non-robustness is apparent. On receipt of the next observation, z_{n+1}, we have $(\lambda|D_{n+1}) \sim G[\alpha_{n+1}/2, \beta_{n+1}/2]$ with, now, $\alpha_{n+1} = \alpha_n + 1$ and

$$\beta_{n+1} = \beta_n + (z_{n+1} - z_n m_n)^2 (1 + z_n^2 C_n)^{-1}.$$

Again $\beta \to \infty$ with δ, so $p(\lambda|D_n)$ and $p(\lambda|D_{n+1})$ become degenerate at the origin as $\delta \to \infty$.

We can easily see that use of an outlier prone likelihood p_ν with either the exact recursions on the modal recursions of §4.3.1. will lead to a robust analysis with the same features as the models of Chapter 4: $(\lambda|D_n) \sim G[\alpha_n/2, \alpha_n/2]$ where both β_n and β_{n+1} are bounded above as functions of δ and reach the upper bounds as $\delta \to \infty$.

5.1.2. Additive outliers.

Let $x_n = \sum_{j=1}^{p} \theta_j x_{n-j} + \nu_n$ be a pth order autoregression which is unobservable in general, the data $\{y_n\}$ deriving from

$$y_n = x_n + \epsilon_n, \quad n = 1,2,\ldots \quad . \tag{5.1.2}$$

The general idea is that x_n is a "clean" process e.g. ν_n are normally distributed and that outliers occur with "large" values of ϵ_n. Such an outlier produces a single aberrant observation; the effect does not carry over to future data. This model comes closer to the concept of outlying observations and, as mentioned in Kleiner et al (1979) and Martin (1979), additive outliers provide more realistic models as they tend to occur much more frequently in practice than innovations outliers. Of course we now have an inherently non-linear model (non-linear in the parameters $\underset{\sim}{\theta}$). In their discussion of outliers in autoregressions, Abraham and Box (1979) remark that numerical integrations are required to obtain marginal posterior and predictive distributions in their retrospective analysis. The problem is far worse in a sequential framework where these numerical integrations must be done at each observation stage, even with normality of the errors $\{\epsilon_n\}$. However a conditional analysis together with a fairly crude summation replacing the numerical integration provides excellent results and we discuss the analysis now, beginning with normally distributed $\{\epsilon_n\}$. We provide the general treatment with the variance of the process errors ν_n unknown.

(a) Normal observational errors.

Martin (1979) discusses the use of the state space form of the model of this section when the variance of the errors and the parameter vector $\underset{\sim}{\theta}$ are known. This form is

$$y_n = \underset{\sim n}{h}^T \underset{\sim n}{x} + \epsilon_n, \qquad (5.1.3)$$

$$\underset{\sim n}{x} = G \underset{\sim n-1}{x} + \underset{\sim n}{\nu}, \qquad n = 1,2,\ldots \qquad (5.1.4)$$

where $\underset{\sim n}{x}^T = (x_n, x_{n-1}, \ldots, x_{n-p+1})$ is the state vector;

$$G = \begin{pmatrix} \theta_1 & \theta_2 & \cdots & \theta_p \\ & & & \\ & I_{p-1} & & 0 \\ & & & \\ & & & 0 \end{pmatrix}$$ is the state transition matrix;

$$\underset{\sim}{\nu}_n^T = (\nu_n, 0, \ldots, 0),$$

and $\underset{\sim}{h}^T = (1, 0, \ldots, 0).$

$\underset{\sim}{\theta}$ is now playing the role of a parameter vector in the control theoretic terminology. Martin's work is concerned with tracking the state $\underset{\sim}{x}_n$ using robust filtering algorithms derived from the scaled recursions of Chapter 3, §3. We are interested in such a track but of main interest is the unknown parameter vector $\underset{\sim}{\theta}$ which Martin assumes to be known (or uses an off-line estimate for $\underset{\sim}{\theta}$). Furthermore, $\mathrm{var}[\nu_n]$ will be unknown in general.

Our analysis assumes the following;

(i) $\{\nu_n\}$ are i.i.d. $N[0,\lambda^{-1}]$ with λ unknown.

(ii) $\{\epsilon_n\}$ are i.i.d. $N[0,\sigma^2\lambda^{-1}]$, with σ^2 known, and $\{\epsilon_n\}$ independent of $\{\nu_n\}$.

(iii) λ a priori has a gamma distribution. $G[\alpha_0/2, \beta_0/2]$.

(iv) $$(\underset{\sim}{x}_0 | \lambda, \underset{\sim}{\theta}) \sim N[\underset{\sim}{m}_0, \lambda^{-1} C_0]$$

with $\underset{\sim}{m}_0$, C_0 not involving λ. (See Appendix 5A for the calculation of $\underset{\sim}{m}_0$, C_0.)

Then for given $\underset{\sim}{\theta}$, the usual Kalman recursions obtain as follows:

$$(\underset{\sim}{x}_n | \lambda, \underset{\sim}{\theta}, D_n) \sim N[\underset{\sim}{m}_n(\underset{\sim}{\theta}), C_n(\underset{\sim}{\theta})\lambda^{-1}],$$

where $$\underset{\sim}{m}_n(\underset{\sim}{\theta}) = \underset{\sim}{a}_n(\underset{\sim}{\theta}) + P_n(\underset{\sim}{\theta})\underset{\sim}{h}_n(\sigma^2 + \underset{\sim}{h}^T P_n(\underset{\sim}{\theta})h)^{-1}\{y_n - \underset{\sim}{h}^T \underset{\sim}{a}_n(\underset{\sim}{\theta})\},$$

and $$C_n(\underset{\sim}{\theta}) = P_n(\underset{\sim}{\theta}) - P_n(\underset{\sim}{\theta})\underset{\sim}{h}\,\underset{\sim}{h}^T P_n(\underset{\sim}{\theta})\,(\sigma^2 + \underset{\sim}{h}^T P_n(\underset{\sim}{\theta})\underset{\sim}{h})^{-1},$$

with, as usual, $\underset{\sim}{a}_n(\underset{\sim}{\theta}) = G\,\underset{\sim}{m}_{n-1}(\underset{\sim}{\theta})$, $P_n(\underset{\sim}{\theta}) = G\,C_{n-1}(\underset{\sim}{\theta})G^T + V$

and

$$V_{11} = 1, \; V_{ij} = 0 \text{ otherwise.}$$

Further $$(\lambda|\underset{\sim}{\theta},D_n) \sim G[\alpha_n/2,\beta_n(\underset{\sim}{\theta})/2],$$

where

$$\alpha_n = \alpha_{n-1} + 1,$$

and

$$\beta_n(\underset{\sim}{\theta}) = \beta_{n-1}(\underset{\sim}{\theta}) + (y_n - \underset{\sim}{h}^T \underset{\sim}{a}_n(\underset{\sim}{\theta}))^2 (\sigma^2 + \underset{\sim}{h}^T P_n(\underset{\sim}{\theta})\underset{\sim}{h})^{-1}.$$

Note that $\underset{\sim}{h}^T G = \underset{\sim}{\theta}^T$ so $y_n - \underset{\sim}{h}^T \underset{\sim}{a}_n(\underset{\sim}{\theta}) = y_n - \underset{\sim}{\theta}^T \underset{\sim}{m}_{n-1}(\underset{\sim}{0})$ and, similarly,

$$\underset{\sim}{h}^T P_n(\underset{\sim}{\theta})h = [P_n(\underset{\sim}{\theta})]_{11} = \underset{\sim}{\theta}^T C_{n-1}(\underset{\sim}{\theta})\ \underset{\sim}{\theta}$$

with $$\underset{\sim}{h}^T V \underset{\sim}{h} = 1.$$

The calculation of $p(\underset{\sim}{\theta}|D_n)$ is then straightforward, defined pointwise by the recursive equation

$$p(\underset{\sim}{\theta}|D_n) \propto p(\underset{\sim}{\theta}|D_{n-1})\ p(y_n|\underset{\sim}{\theta},D_{n-1})$$

where

$$p(y_n|\theta,D_{n-1}) \propto (\sigma^2 + \underset{\sim}{h}^T P_n(\underset{\sim}{\theta})h)^{-\frac{1}{2}}\ \beta_n(\underset{\sim}{\theta})^{-\alpha_n/2}\ \beta_{n-1}(\underset{\sim}{\theta})^{\alpha_{n-1}/2}$$

For small p, in particular for p = 1 or 2, the use of a fairly coarse grid of values for $\underset{\sim}{\theta}$ leads to a useful procedure.

The model (5.1.3) and (5.1.4) was developed in order to provide an additive outlier generating structure by the introduction of the errors $\{\epsilon_n\}$. As such, the distribution of those errors should suffice to produce outliers and nothing more since the usual variation, i.e. the innovations are already modelled as the $\{\nu_n\}$. Following Martin (1980), the parsimonious model for the enor density p_ϵ of the i.i.d. $\{\epsilon_n\}$ is taken as

$$p_\epsilon(\epsilon) = (1-\tau)\delta_0 + \pi p(\epsilon), \quad 0 < \pi < 1,$$

where π is "small", δ_0 represents a point mass at the origin, and p is a unimodal, symmetric heavy-tailed density. Thus the error density for the observations $\{y_n\}$ is that of $u_n = \nu_n + \epsilon_n$ which, when $\{\nu_n\}$ are i.i.d. $N[0,1]$, is given by

$$(1-\tau)\phi(u_n) + \pi f(u_n),$$

where f is the convolution of p with ϕ. [As an aside note that to be consistent with this error model in the innovations outlier case of §4.1.1, we ought to take our p_ν as a mixture of a normal with a heavy tailed density. This is just the prescription of Appendix 2].

Now the analysis (a) above applies to normal p. This will produce a partially robust analysis but full outlier-rejection can only be obtained by using an outlier-prone density. In the following we use a general p with the normal analysis (a) being a special case.

(b) Non normality.

Assume that

(i) $\{\nu_n\}$ given λ are i.i.d. $N[0,\lambda^{-1}]$;

(ii) $\{\epsilon_n\}$ given λ are i.i.d. $\lambda^{\frac{1}{2}} p_\epsilon(\lambda^{\frac{1}{2}}\epsilon)$ where p_ϵ is the mixture

$$(1-\pi)\delta_0 + \pi\, p(\epsilon) \quad , \qquad 0 < \pi < 1,$$

with p unimodal, symmetric heavy tailed;

(iv) $$(\underset{\sim}{x}_0 | \lambda, \underset{\sim}{\theta}) \sim N[\underset{\sim}{m}_0, \lambda^{-1} C_0]$$

where $\underset{\sim}{m}_0$, C_0 do not involve λ.

(v) $\lambda \sim G[\alpha_0/2\ \beta_0/2]$.

The analysis adopted involves collapsing mixtures of normal distributions in the way of Harrison and Stevens (1976) as we did in Chapters 3 and 4 and requires at time n that, approximately,

$$(\underset{\sim}{x}_{n-1} | \underset{\sim}{\theta}, \lambda, D_{n-1}) \sim N\left[\underset{\sim}{m}_{n-1}(\underset{\sim}{\theta}), \lambda^{-1}\, C_{n-1}(\underset{\sim}{\theta})\right] \qquad (5.1.5)$$

with $\underset{\sim}{m}_{n-1}(\underset{\sim}{\theta})$, $C_{n-1}(\underset{\sim}{\theta})$ independent of λ,

and

$$(\lambda|\underset{\sim}{\theta},D_{n-1}) \sim G\left[\alpha_{n-1}/2,\ \beta_{n-1}(\underset{\sim}{\theta})/2\right]. \tag{5.1.6}$$

Given these assumptions,

$$(\underset{\sim}{x}_n|\underset{\sim}{\theta},\lambda,D_{n-1}) \sim N\left[\underset{\sim}{a}_n(\underset{\sim}{\theta}),\lambda^{-1}P_n(\underset{\sim}{\theta})\right] \tag{5.1.7}$$

with $\underset{\sim}{a}_n(\underset{\sim}{\theta})$, $P_n(\underset{\sim}{\theta})$ as given in (a) above. Now we examine the components of the analysis separately.

(i) $p(\underset{\sim}{x}_n,\lambda|\underset{\sim}{\theta},D_n)$

We have the joint normal/gamma prior $p(\underset{\sim}{x}_n,\lambda|\underset{\sim}{\theta},D_{n-1})$ given above, so the posterior is just proportional to

$$p(\underset{\sim}{x}_n,\lambda|\underset{\sim}{\theta},D_{n-1})\ \{(1-\pi)\delta_{r_{n=0}} + \pi\lambda^{\frac{1}{2}}p_\epsilon(\lambda^{\frac{1}{2}}r_n)\}$$

where $r_n = y_n - \underset{\sim}{h}^T\underset{\sim}{x}_n = y_n - x_n$.

So

$$p(\underset{\sim}{x}_n,\lambda|\underset{\sim}{\theta},D_n) = (1-\pi^*)p_1(\underset{\sim}{x}_n,\lambda|\underset{\sim}{\theta},D_n) + \pi^*p_2(\underset{\sim}{x}_n,\lambda|\underset{\sim}{\theta},D_n)$$

The functions p_1, p_2, π^* are as follows.

p_1 is the posterior when $\epsilon_n = 0$ so

$$y_n = \underset{\sim}{h}^T\underset{\sim}{x}_n = x_n = \underset{\sim}{h}^TG\ \underset{\sim}{x}_{n-1} + \underset{\sim}{h}^T\underset{\sim}{\nu}_n$$

$$= \underset{\sim}{\theta}^T\underset{\sim}{x}_{n-1} + \nu_n.$$

Therefore $(\underset{\sim}{x}_{n-1}|y_n=x_n,\underset{\sim}{\theta},D_n,\lambda) \sim N\left[\underset{\sim}{t}^n_{n-1}(\underset{\sim}{\theta}),\ T^n_{n-1}(\underset{\sim}{0})\lambda^{-1}\right]$,

and $(\lambda|y_n = x_n,\underset{\sim}{\theta},\ D_n) \sim G\left[\alpha_n/2,\beta_{n1}(\underset{\sim}{\theta})/2\right]$.

where $\underset{\sim}{t}^n_{n-1}(\underset{\sim}{\theta}) = \underset{\sim}{m}_{n-1}(\underset{\sim}{\theta}) + C_{n-1}(\underset{\sim}{\theta})\ \underset{\sim}{\theta}\left(1+\underset{\sim}{\theta}^TC_{n-1}(\underset{\sim}{\theta})\underset{\sim}{\theta}\right)^{-1}\left(y_n-\underset{\sim}{\theta}^T\underset{\sim}{m}_{n-1}(\underset{\sim}{\theta})\right)$,

$$T^n_{n-1}(\underset{\sim}{\theta}) = C_{n-1}(\underset{\sim}{\theta}) - C_{n-1}(\underset{\sim}{\theta})\underset{\sim}{\theta}\underset{\sim}{\theta}^TC_{n-1}(\underset{\sim}{\theta})\ \left(1+\underset{\sim}{\theta}^TC_{n-1}(\underset{\sim}{\theta})\underset{\sim}{\theta}\right)^{-1},$$

$$\alpha_n = \alpha_{n-1} + 1,$$

and

$$\beta_{n1}(\underset{\sim}{\theta}) = \beta_{n-1} + \left(1+\underset{\sim}{\theta}^T C_{n-1}(\underset{\sim}{\theta})\underset{\sim}{\theta}\right)^{-1} \left(y_n - \underset{\sim}{\theta}^T \underset{\sim}{m}_{n-1}(\underset{\sim}{\theta})\right)^2. \qquad (5.1.8)$$

Further $p(\underset{\sim}{x}_n | \lambda, \underset{\sim}{\theta}, D_n, y_n = x_n)$ is the (singular) normal density

$$N\left[\left(y_n : \underset{\sim}{t}^{n\,T}_{n-1}(\underset{\sim}{\theta})\right)^T,\ \lambda^{-1} \begin{pmatrix} 0 & \vdots & \underset{\sim}{0}^T \\ \cdots & \vdots & \cdots \\ \underset{\sim}{0} & \vdots & T^n_{n-1}(\underset{\sim}{0}) \end{pmatrix} \right],$$

$$N\left[\underset{\sim}{m}_{n1}(\underset{\sim}{\theta}),\ \lambda^{-1} C_{n_1}(\theta)\right], \quad \text{say} \qquad (5.1.9)$$

The second component p_2 is proportional to the product of a normal/gamma prior with a heavy-tailed likelihood and thus the methods of Chapter 3 are directly applicable with the extras of Chapter 4 to deal with the scaling λ. We use the modal recursions of Chapter 4, §4.3.1.

Let $g(\epsilon) = -\frac{\partial}{\partial \epsilon} \ell n\ p(\epsilon) = \psi(\epsilon).\epsilon$ and express $p(\epsilon_n)$ as a scale mixture of normals.

$$p(\epsilon_n) = \int_0^\infty N\left[0, \lambda_n^{-1}\right] \omega(\lambda_n) d\lambda_n.$$

Define the prior mean for the scale of the $\{\nu_n\}$ to be

$$E\left[\lambda \middle| \underset{\sim}{\theta}\ D_{n-1}\right] = \ell_{n-1}(\underset{\sim}{\theta}) = \alpha_{n-1}/\beta_{n-1}(\underset{\sim}{\theta}).$$

Then if $\tilde{\lambda}_n(\underset{\sim}{\theta}) = E\left[\lambda_n \middle| \underset{\sim}{\theta}, \underset{\sim}{x}_n = \underset{\sim}{a}_{n-1}(\underset{\sim}{\theta}),\ \lambda = \ell_{n-1}(\underset{\sim}{\theta}), D_{n-1}\right]$

$$= \psi\left[\ell^{\frac{1}{2}}_{n-1}(\underset{\sim}{\theta})\ \left(y_n - \underset{\sim}{\theta}^T m_{n-1}(0)\right)\right],$$

the modal recursions are given by

$$\underset{\sim}{m}_{n2}(\underset{\sim}{\theta}) \simeq \underset{\sim}{a}_n(\underset{\sim}{\theta}) + \underset{\sim}{P}_n(\underset{\sim}{\theta})\underset{\sim}{h}\ (q^2_n(\underset{\sim}{\theta})\tilde{\lambda}_n(\underset{\sim}{\theta})+1)^{-1}\tilde{\lambda}_n(\underset{\sim}{\theta})u_n(\underset{\sim}{\theta}) \qquad (5.1.10)$$

with $q^2_n(\underset{\sim}{\theta}) = \underset{\sim}{h}^T P_n(\underset{\sim}{\theta})\underset{\sim}{h}$ and $u_n(\underset{\sim}{\theta}) = y_n - \underset{\sim}{\theta}^T m_{n-1}(\underset{\sim}{\theta})$.

Further

$$C_{n2}(\underset{\sim}{\theta}) \simeq P_n(\underset{\sim}{\theta}) - P_n(\underset{\sim}{\theta})\underset{\sim}{h}\,\underset{\sim}{h}^T P_n(\underset{\sim}{\theta})\ \phi_n\left(\ell_{n-1}^{\frac{1}{2}}(\underset{\sim}{\theta})u_n(\underset{\sim}{\theta})\right). \tag{5.1.11}$$

where $\phi_n(u)$ is calculated as in equation (4.3.15) of §4.3.1.

If, also, $\alpha_n = \alpha_{n-1} + 1$

and

$$\beta_{n2}(\underset{\sim}{\theta}) = \beta_{n-1}(\underset{\sim}{\theta}) + u_n^2(\underset{\sim}{\theta}).(\tilde{\lambda}_n(\underset{\sim}{\theta})q_n^2+1)^{-1} \tag{5.1.12}$$

then the modal algorithm gives the joint normal/gamma posterior

$$p(\underset{\sim}{x}_n, \lambda | D_n, \underset{\sim}{\theta}, \epsilon_n \neq 0)$$

as

$$(\underset{\sim}{x}_n | \lambda, D_n, \underset{\sim}{\theta}, \epsilon_n \neq 0) \sim N\left[\underset{\sim}{m}_{n2}(\underset{\sim}{\theta}), \lambda^{-1}C_{n2}(\underset{\sim}{\theta})\right],$$

and

$$(\lambda | D_n, \underset{\sim}{\theta}, \epsilon_n \neq 0) \sim G\left[\alpha_n/2,\ \beta_{n2}(\underset{\sim}{\theta})/2\right].$$

The function π^* is defined via the predictive densities for y_n in (ii) below. The approximation to the joint posterior $p(\underset{\sim}{x}_n, \lambda | D_n, \underset{\sim}{\theta})$ will be in the spirit of Chapter 3 and is made by collapsing to a single normal gamma

$$(\underset{\sim}{x}_n | \lambda, D_n, \underset{\sim}{\theta}) \sim N\left[\underset{\sim}{m}_n(\underset{\sim}{\theta}),\ \lambda^{-1}C_n(\underset{\sim}{\theta})\right]$$

and

$$(\lambda | D_n, \underset{\sim}{\theta}) \sim G\left[\alpha_n/2,\ \beta_n(\underset{\sim}{\theta})/2\right]$$

where $\underset{\sim}{m}_n(\underset{\sim}{\theta}) = (1-\pi^*)\ \underset{\sim}{m}_{n1}(\underset{\sim}{\theta}) + \pi^* \underset{\sim}{m}_{n2}(\underset{\sim}{\theta})$,

$$C_n(\underset{\sim}{\theta}) = (1-\pi^*)\ \left[C_{n1}(\underset{\sim}{\theta}) + \left(\underset{\sim}{m}_n(\underset{\sim}{\theta}) - \underset{\sim}{m}_{n1}(\underset{\sim}{\theta})\right)\ \left(\underset{\sim}{m}_n(\underset{\sim}{\theta}) - \underset{\sim}{m}_{n1}(\underset{\sim}{\theta})\right)^T\right]$$

$$+ \pi^*\ \left[C_{n2}(\underset{\sim}{\theta}) + \left(\underset{\sim}{m}_n(\underset{\sim}{\theta}) - \underset{\sim}{m}_{n2}(\underset{\sim}{\theta})\right)\ \left(\underset{\sim}{m}_n(\underset{\sim}{\theta}) - \underset{\sim}{m}_{n2}(\underset{\sim}{\theta})\right)^T\right],$$

and $\beta_n(\underset{\sim}{\theta})^{-1} = (1-\pi^*)\ \beta_{n1}(\underset{\sim}{\theta})^{-1} + \pi^* \beta_{n2}(\underset{\sim}{\theta})^{-1}$

(ii) Predictive densities for $y_n|\theta$.

We calculate first $p_1(y_n|\underset{\sim}{\theta}) = p(y_1|\underset{\sim}{\theta}, D_{n-1}, \epsilon_n=0)$. Since $y_n = x_n$, we have

$$(y_n|\lambda, D_{n-1}, \underset{\sim}{\theta}) \sim N\left[\underset{\sim}{\theta}^T m_{n-1}(\theta), \lambda^{-1}\left(1+\underset{\sim}{\theta}^T C_{n-1}(\underset{\sim}{\theta})\underset{\sim}{\theta}\right)\right]$$

and so $p_1(y_n|\underset{\sim}{\theta}) \propto \left(1+\underset{\sim}{\theta}^T C_{n-1}(\underset{\sim}{\theta})\underset{\sim}{\theta}\right)^{-\frac{1}{2}} \left\{\beta_{n-1}(\underset{\sim}{\theta}) + \frac{(y_n - \underset{\sim}{\theta}^T m_{n-1}(\underset{\sim}{\theta}))^2}{\left(1+\underset{\sim}{\theta}^T C_{n-1}(\underset{\sim}{\theta})\underset{\sim}{\theta}\right)}\right\}^{-\frac{\alpha_n}{2}} \beta_{n-1}(0)^{\alpha_{n-1}/2}$

(5.1.13)

Also $(1-\pi^*) \propto (1-\pi)p_1(y_n|\underset{\sim}{\theta})$. (5.1.14)

For $p_2(y_n) = p(y_n|\underset{\sim}{\theta}, D_{n-1}, \epsilon_n \neq 0)$ we refer to §4.3.2 of Ch. 4, where this marginal density is derived when the modal recursions are used as an approximation. Equation (4.3.17) gives the marginal score which can be used to find $p_2(y_n|\underset{\sim}{\theta})$. In particular, if we use a Student t density for p then we obtain the marginal given in examples 4.3.2 (ii).

Further $\pi^* \propto \pi p_2(y_n)$, and so π^* can be obtained using, in addition, (5.1.14).

(iii) Posterior for $\underset{\sim}{\theta}$.

$$p(\underset{\sim}{\theta}|D_n) \propto p(\underset{\sim}{\theta}|D_{n-1})p(y_n|\underset{\sim}{\theta}, D_{n-1}),$$

is calculated again pointwise, using

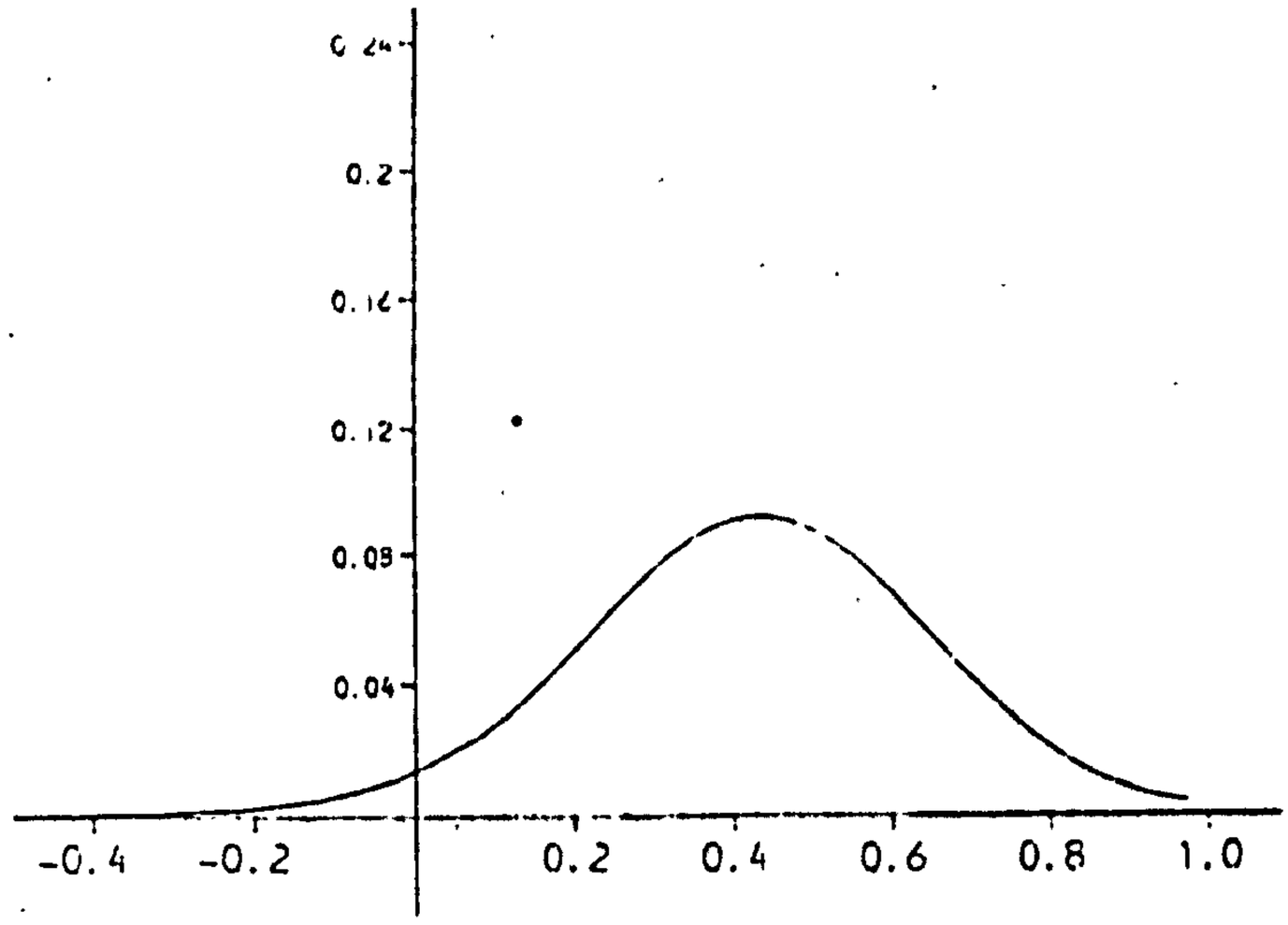
$$p(y_n|\underset{\sim}{\theta}, D_{n-1}) = (1-\pi)p_1(y_n|\underset{\sim}{\theta}) + \pi p_2(y_n|\underset{\sim}{\theta}).$$

Example 5.1.3.

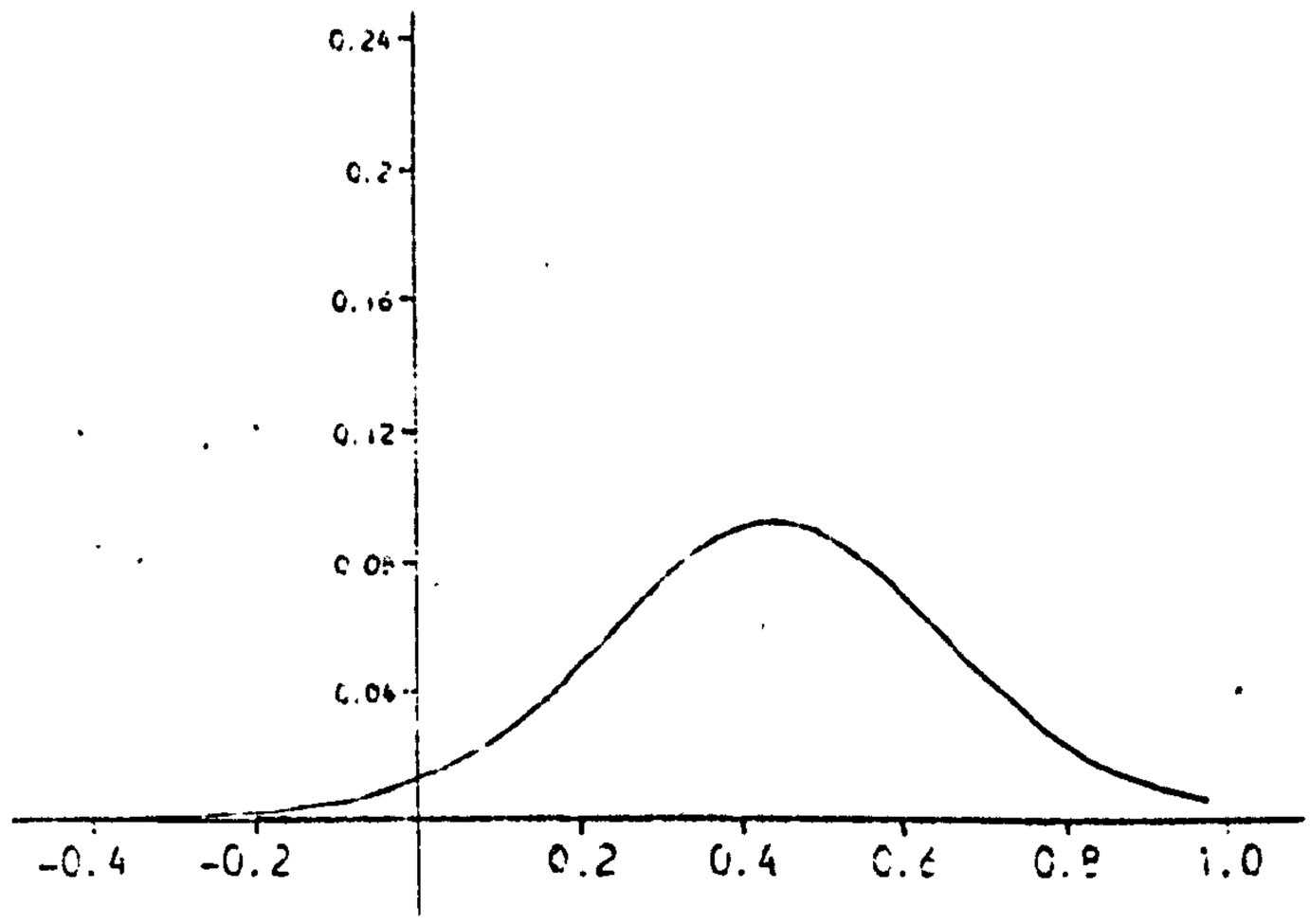
For illustration the model of Example 5.1.2 was used in the same way but now with an additive outlier at time n = 19. The posterior densities at times n = 18, 19 and 20 are shown in Figure 5.2.

AO AT N=19
POSTERIOR AT TIME N

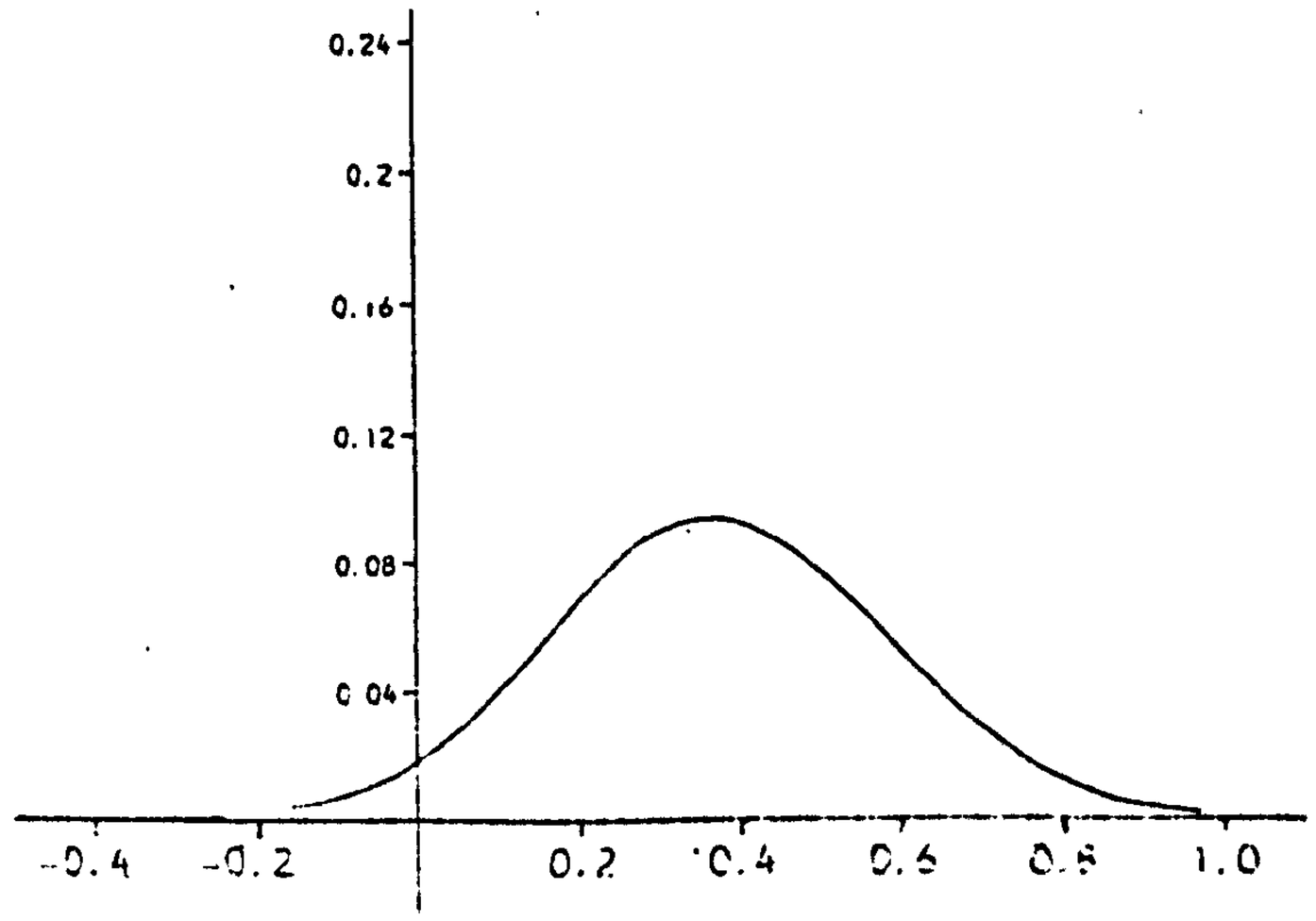
N=18



N=19



N=20



5.1.3. Modelling outliers of unknown type.

How do we approach the problem of modelling outliers in autoregressions when both types are admitted? Martin (1979) mentions that it seems difficult to obtain efficient classical M-estimators of autoregressive parameters when both types occur. In particular, the performance of robust M-estimators (and the analysis of §5.1.1) is seriously degraded if additive outliers (which are not specifically modelled in §5.1.1) occur. See Miller (1980) for an example of this and Martin (1980) for discussion. Clearly modelling additive outliers as in §5.1.2 solves the problem and what we need to do is simply model both types of outliers, as follows.

Let $y_n = \underset{\sim}{h}^T \underset{\sim}{x}_n + \epsilon_n$,

and

$$\underset{\sim}{x}_n = G\underset{\sim}{x}_{n-1} + \underset{\sim}{\nu}_n, \qquad n = 1,2,\ldots$$

where $\underset{\sim}{h}$, $\underset{\sim}{x}_n$, G, $\underset{\sim}{\nu}_n$ are as before. However now take $p_\nu(\nu_n)$ to be heavy-tailed non-normal and

$$p_\epsilon(\epsilon_n) = (1-\pi)\delta_0 + \pi p(\epsilon_n) \text{ with p also heavy-tailed.}$$

In order to base the analysis of this model on previous ideas we express both p_ϵ and p_ν as scale mixtures of normals.

Then $p_\epsilon(\epsilon_n|\lambda) = \int_0^\infty N\left[0, \lambda^{-1}\lambda_n^{-1}\right]\omega(\lambda_n)d\lambda_n$

where $\omega(\lambda_n) = (1-\pi)\delta_0 + \pi\omega'(\lambda_n)$ and

$$p(\epsilon_n|\lambda) = \int_0^\infty N\left[0, \lambda^{-1}\lambda_n^{-1}\right]\omega'(\lambda_n)\, d\lambda_n$$

and, further,

$$p_\nu(\nu_n|\lambda) = \int_0^\infty N\left[0, \lambda^{-1}\mu_n^{-1}\right]\, u(\mu_n)d\mu_n$$

where u and ω' are specified densities on $\mathbb{R}^+$. Finally λ_n and μ_n are independent and independent of λ_r, μ_r for $r \neq n$. Now the conditional analysis given $f_n = (\lambda_n, \mu_n)$ proceeds in the usual way and we use the modal approximations to eliminate f_n. However we adopt a specific density u for μ_n as follows. If ν_n is normal, then we have the AO model of §5.1.3 i.e. normal prior, non-normal likelihood. To use this in the case of AO and IO structure, we take p_ν as a contaminated normal mixture, i.e. the special case

$$u(\mu) = (1-\gamma)\delta_1 + \gamma\delta_v$$

where $v >> 1$ and $0 < \gamma < 1$, with γ small.

The analysis now follows §5.1.2 for each component of the resulting mixture posterior of normal/gamma densities for $(\underset{\sim}{x}_n, \lambda)$,

$$p(\underset{\sim}{x}_n, \lambda \,|\, \underset{\sim}{\theta}, \ D_n).$$

This is a 4-component mixture just as in the Multi-state model of Harrison and Stevens (1976) with the addition of a more general error density for ϵ_n providing a means of using an outlier prone distribution via the modal approximations and also with a scale parameter in $\lambda^{-\frac{1}{2}}$.

Now in the case of a single AO or IO generating model, a "surprisingly large" observation indicates unequivocably the occurrence of an outlier of that type and the analysis reflects this, ignoring the outlier. However in the model of this section complications arise just as in the Harrison-Stevens system, and these problems underly the comments of Martin (1979) on distinguishing outlier type.

If y_n is "large", we cannot know at time n whether we have had an AO or an IO and, since the latter corresponds to a change in the

level of the state vector $\underset{\sim}{x}_n$ but the former does not, then the corresponding components of $p(\underset{\sim}{x}_n|\underset{\sim}{\theta}, D_n)$ will be centred some distance apart leading to the possibility of a bimodal posterior. A further observation will help to distinguish the outlier types at time n via calculation of $p(\underset{\sim}{x}_n|\underset{\sim}{\theta}, D_{n+1})$, although the occurrence of an outlier at time n+1 would complicate matters. Mallows (1980), in a discussion of the related problem of smoothing time series, suggests just such a behaviour as being required for a fully robust analysis.

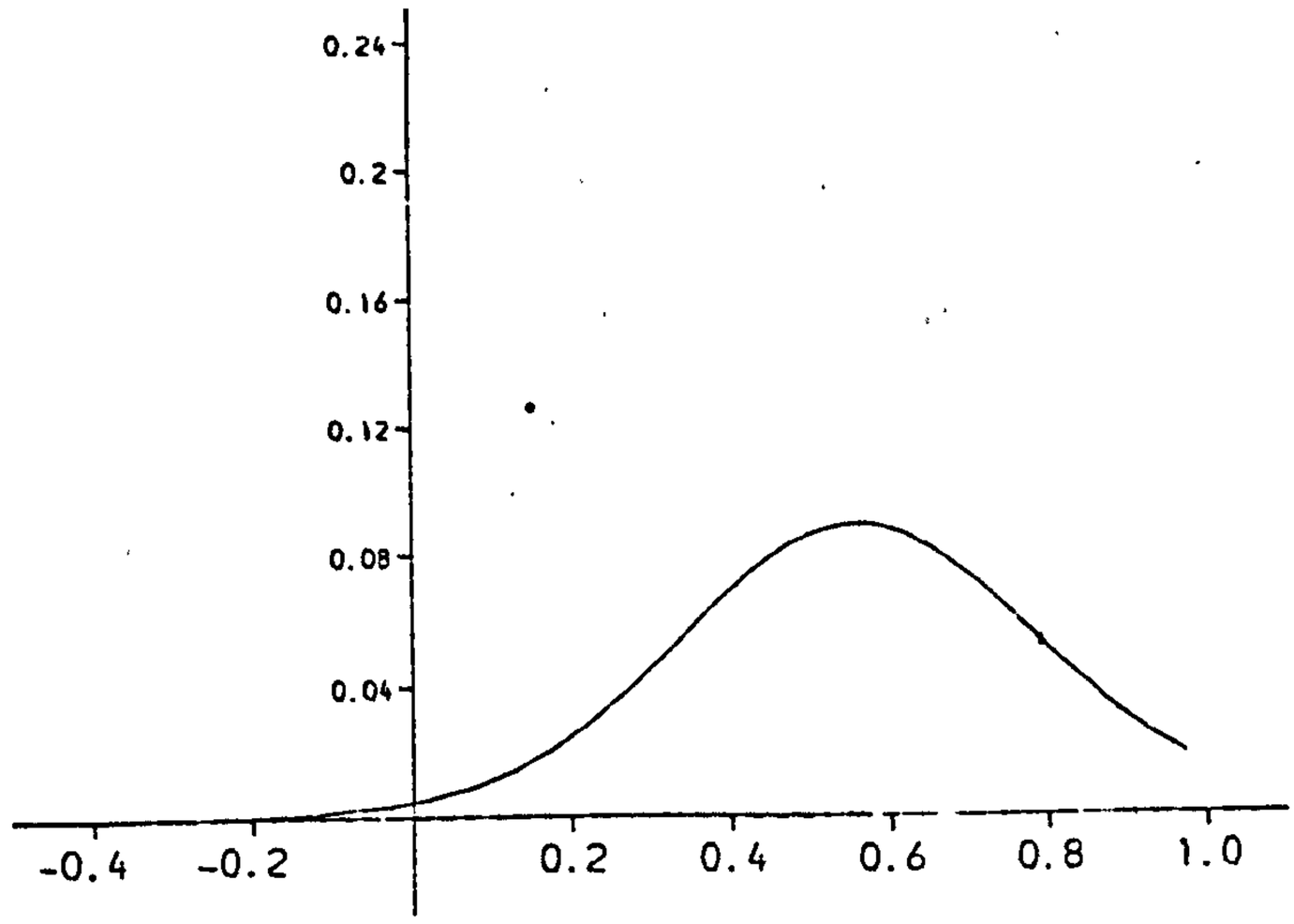
In order to cope with this behaviour we follow Harrison and Stevens proposal of not collapsing $p(\underset{\sim}{x}_n,\lambda|\underset{\sim}{\theta}, D_n)$ to a single joint normal/gamma density but instead retaining the full 4-state mixture as our prior for time n+1. Thus we require 4 parallel analyses at each observation stage.

Example 5.1.4.

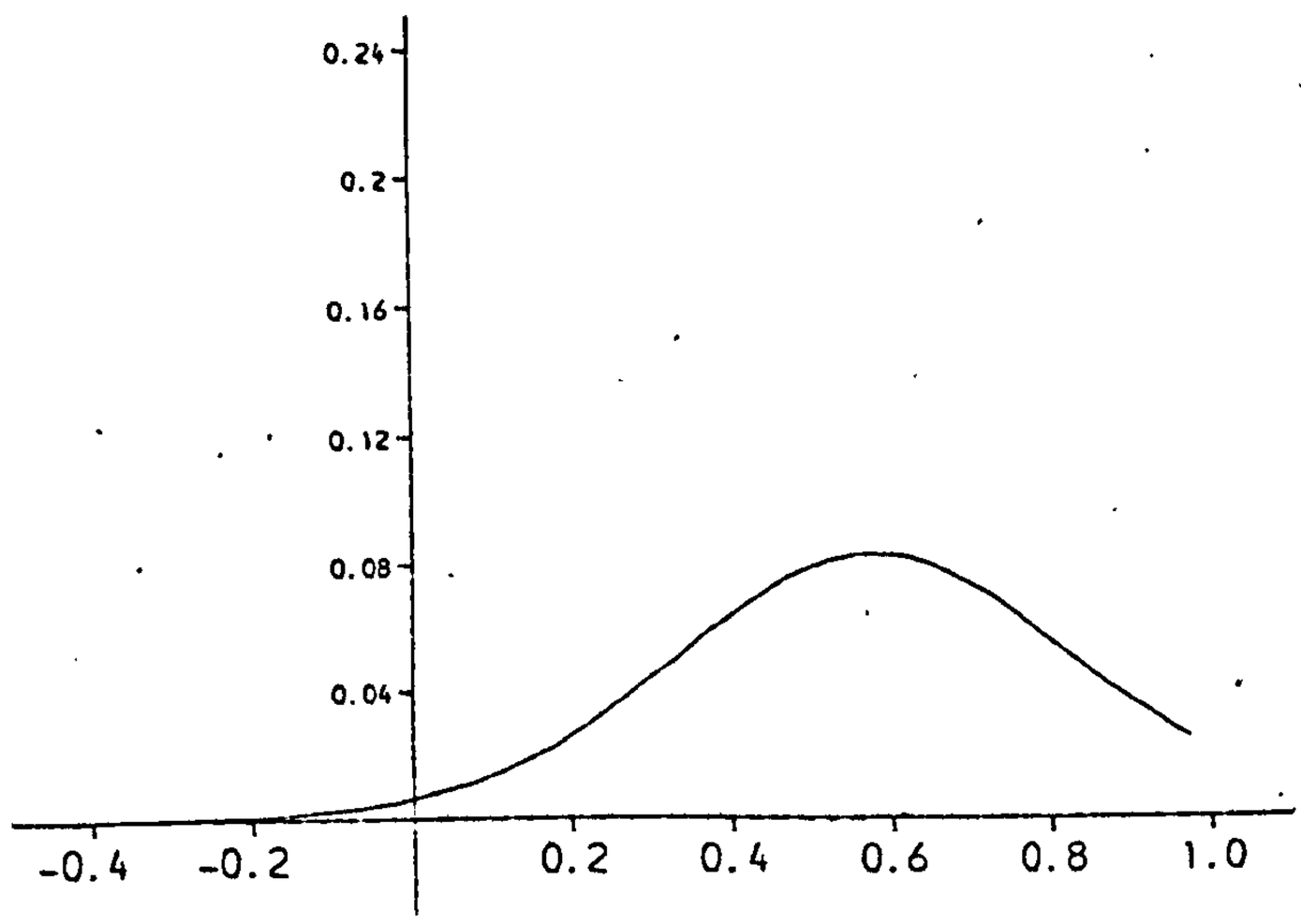
Figure 5.3 provides a plot of the posterior densities for the AR(1) process of the earlier examples. This time both innovation and additive outliers occur at n = 19.

IO + AO AT N=19
POSTERIOR AT TIME N.

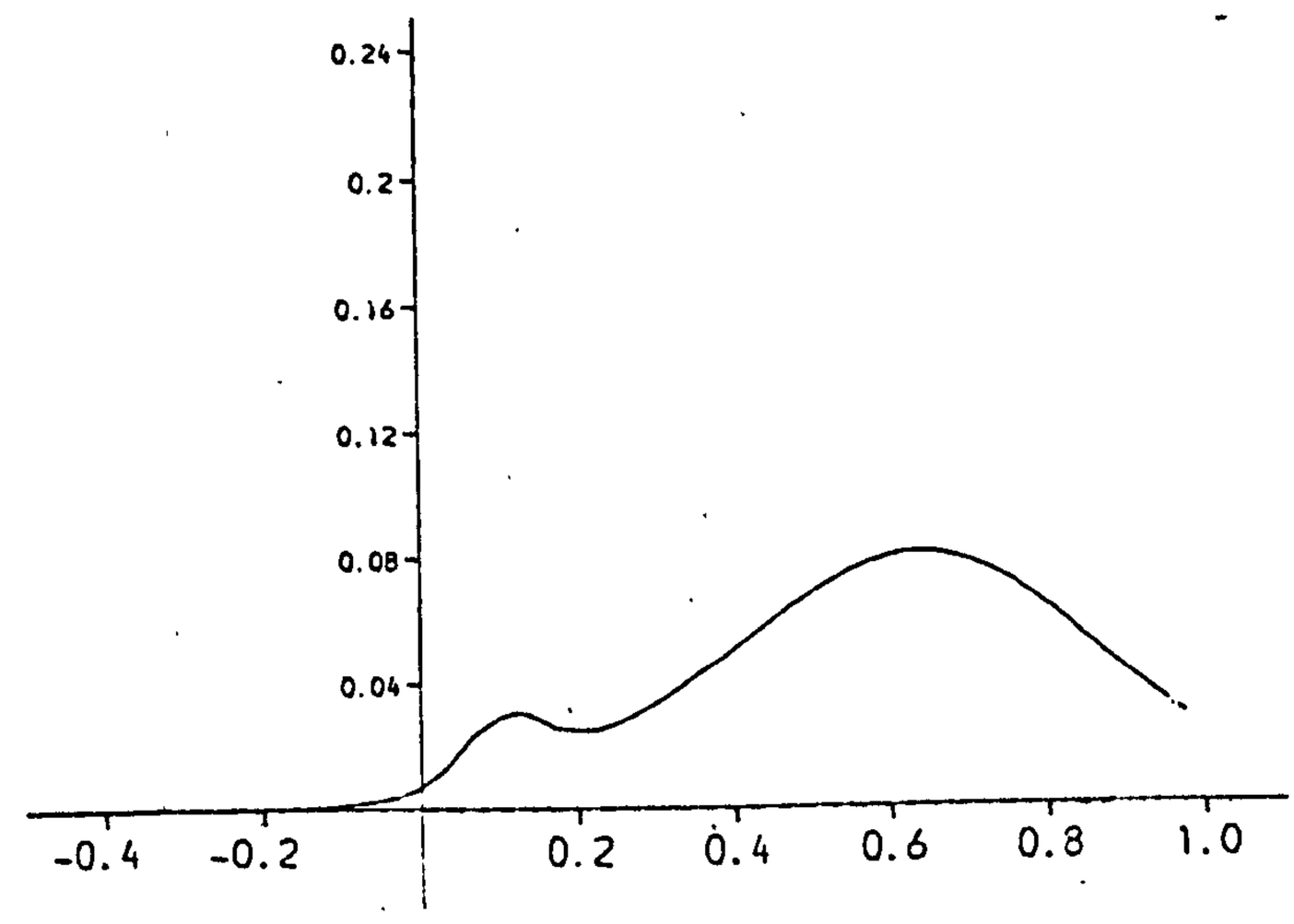
N=18



N=19



N=20



5.2. Autoregressive - moving average models.

5.2.1. Normal errors: state space representation.

The general ARMA model is inherently non-linear in the moving average parameters and so cannot be directly analysed using Kalman filtering techniques in a sequential processing of the observations. However, using a state-space representation of the model we are able to utilize Kalman filtering techniques as we did with the AO-AR model in §5.1.1 to compute the posterior distribution for the parameters pointwise. Priestley (1980) describes in detail the state space representation of ARMA systems and this method is used by Gardner et al (1980) in calculating likelihood functions recursively for ARMA models. Clearly the approach extends easily to move general models and Harvey and Phillips (1979) use the state space form for estimation in regression models with ARMA errors, again from a maximum likelihood point of view. The Bayesian analysis of ARMA models discussed in this section can easily be generalised in the same way.

We begin by defining the state space model.

We have the general ARMA(p,q) model for observations $\{y_n\}$ given by

$$y_n - \sum_{j=1}^{p} y_{n-j}\,\theta_j = \nu_n - \sum_{j=1}^{q} \nu_{n-j}\,\phi_j \qquad , n = 1,2,\ldots \qquad (5.2.1)$$

where the $\{\nu_n\}$ are i.i.d. $N[0,\lambda^{-1}]$ and $\underset{\sim}{\theta}^T = (\theta_1,\ldots,\theta_p)$, $\underset{\sim}{\phi}^T = (\phi_1,\ldots,\phi_q)$ are the unknown parameters. The state space model is defined as follows.

Let $r = \max(p,q)$ and define $\theta_j = \phi_i = 0$ for $j = p+1,\ldots r$ and $i = q+1,\ldots,r$. Further define the new parameters π_j by

$$\pi_j = \theta_j - \phi_j \qquad ,j = 1,\ldots,r.$$

Then, if $x_n = y_n - \nu_n$, $n = 1,2,\ldots$, the state vector $\underset{\sim}{x}_n$ is given by

$$\underset{\sim}{x}_n^T = (x_n,\ldots,x_{n-r+1}).$$

The regression matrix is $\underset{\sim}{h}^T = (1,0,\ldots,0)$ for all n and the state transition matrix is

$$G = \begin{pmatrix} \phi_1 & \phi_2 \cdots \phi_r \\ 1 & 0 \ldots . 0 \\ & \ddots \\ 0 \ldots . 1 & 0 \end{pmatrix} = \left(\begin{array}{c|c} \underset{\sim}{\phi}_r^T & \\ \hline I_{r-1} & \underset{\sim}{0}_{r-1} \end{array}\right),$$

with the model

$$y_n = \underset{\sim}{h}^T \underset{\sim}{x}_n + \nu_n \qquad n = 1,2,\ldots \tag{5.2.2}$$

$$\underset{\sim}{x}_n = G\, \underset{\sim}{x}_{n-1} + \underset{\sim}{\omega}_n \tag{5.2.3}$$

where now $\underset{\sim}{\omega}_n$ is not a vector of errors but a constant at time n,

$$\underset{\sim}{\omega}_n^T = (\omega_n, 0,\ldots,0)$$

with

$$\omega_n = \sum_{j=1}^{r} y_{n-j}\, \pi_j. \tag{5.2.4}$$

Now the calculation of $p(\underset{\sim}{\theta},\underset{\sim}{\phi}|D_n)$ proceeds in essentially the same way as for the additive outliers autoregression of §5.1.2. We apply the Kalman filter to $(\underset{\sim}{x}_n|\underset{\sim}{\theta},\underset{\sim}{\phi})$ and calculate the predictive density for $(y_n|\underset{\sim}{\theta},\underset{\sim}{\phi})$ from that analysis.

Clearly if, at time n,

$$(\underset{\sim}{x}_{n-1}|\underset{\sim}{\theta},\underset{\sim}{\phi},\lambda,D_{n-1}) \sim N\left[\underset{\sim}{m}_{n-1},\lambda^{-1}C_{n-1}\right], \tag{5.2.5}$$

where $\underset{\sim}{m}_{n-1}$, C_{n-1} are functions of $\underset{\sim}{\theta},\underset{\sim}{\phi}$, then

$$(\underset{\sim}{x}_n|\underset{\sim}{\theta},\underset{\sim}{\phi},\lambda,D_n) \sim N\left[\underset{\sim}{m}_n,\lambda^{-1}C_n\right]$$

where $\underset{\sim}{m}_n$, C_n are found from the usual normal analysis as

$$\underset{\sim}{m}_n = \underset{\sim}{a}_n + P_n \underset{\sim}{h}\, (\underset{\sim}{h}^T P_n \underset{\sim}{h}+1)^{-1}\, (y_n - \underset{\sim}{h}^T \underset{\sim}{a}_n),$$

$$C_n = P_n - P_n \underset{\sim}{h}\, \underset{\sim}{h}^T P_n (\underset{\sim}{h}^T P_n \underset{\sim}{h}+1)^{-1},$$

$$\underset{\sim}{a}_n = G\underset{\sim}{m}_{n-1} + \underset{\sim}{\omega}_n,$$

$$P_n = G\, C_{n-1} G^T.$$

Notice that $\underset{\sim}{h}^T G = \underset{\sim}{\phi}^T$ so $u_n = y_n - \underset{\sim}{h}^T \underset{\sim}{a}_n = y_n - \underset{\sim}{\phi}^T \underset{\sim}{m}_{n-1} - \omega_n$,

is just the residual given $\underset{\sim}{\phi}, \underset{\sim}{\theta}$ and $\underset{\sim}{x}_{n-1} = \underset{\sim}{m}_{n-1}$.

Further $p(y_n | \underset{\sim}{\theta}, \underset{\sim}{\phi}, \lambda, D_{n-1}) = N\left[\underset{\sim}{h}^T \underset{\sim}{a}_n, (\underset{\sim}{h}^T P_n \underset{\sim}{h}+1)\lambda^{-1}\right]$.

So if $(\lambda | \underset{\sim}{\theta}, \underset{\sim}{\phi}, D_{n-1}) \sim G\left[\alpha_{n-1}/2, \beta_{n-1}/2\right]$,

(with β_{n-1} depending upon $\underset{\sim}{\theta}, \phi$),

we have $\alpha_n = \alpha_{n-1} + 1$,

$$\beta_n = \beta_{n-1} + u_n^2 \, (\underset{\sim}{h}^T P_n \underset{\sim}{h}+1)^{-1},$$

and $(y_n | \underset{\sim}{\theta}, \underset{\sim}{\phi}, D_{n-1})$ has a t density, proportional to

$$\left\{\beta_{n-1} + u_n^2 (\underset{\sim}{h}^T P_n \underset{\sim}{h}+1)^{-1}\right\}^{-(\alpha_{n-1}+1)/2}.$$

Therefore

$$p(\underset{\sim}{\theta}, \underset{\sim}{\phi} | D_n) \propto p(\underset{\sim}{\theta}, \underset{\sim}{\phi} | D_{n-1}) p(y_n | \underset{\sim}{\theta}, \underset{\sim}{\phi}, D_{n-1})$$

is updated pointwise just as in the AO AR models of §5.1.2. This is the form of computation used by Gardner et al (1980) in computing the likelihood function for maximum likelihood estimation of ARMA models with normal errors. Details of numerical performance as far as accuracy and speed are concerned are discussed in depth in that reference. However of course, such an algorithm is extremely sensitive to outliers due to the normality assumptions and we intend to protect against this. It is quite clear that the form of model (5.2.2) and (5.2.3) is just the same as that of the AO AR model, then the method of analysis is the same. We briefly discuss the results.

a) Innovations outliers.

Innovations outliers are modelled by giving the $\{\nu_n\}$ a heavy-tailed density p_ν. From the model (5.2.2) and (5.2.3) we see that we have the model of Chapter 4 for updating $p(\underset{\sim}{x}_n, \lambda | \underset{\sim}{\theta}, \phi, D_{n-1})$ and the analysis is the same as that of §5.1.2 for the AO AR model. We obtain the marginal density $p(y_n | \underset{\sim}{\theta}, \underset{\sim}{\phi}, D_{n-1})$ as a Student t form and $p(\underset{\sim}{\theta}, \underset{\sim}{\phi} | D_n)$ is then calculated pointwise in a recursive fashion.

b) Additive outliers.

This case is somewhat different. To model additive outliers we have to introduce a third level in the hierarchical model as follows.

Replace y_n of (5.2.2) by a further process variable z_n. Let ϵ_n be the "outlier" at time n, with the density of §5.1.2,

$$p_\epsilon(\epsilon_n) = (1-\pi)\delta_0 + \pi p(\epsilon_n) \qquad ,0 < \pi < 1.$$

where p is outlier-prone. Assume $\{\epsilon_n\}$ are i.i.d. given λ, with density $\lambda^{\frac{1}{2}} p_\epsilon(\lambda^{\frac{1}{2}}\epsilon)$.

Then the observations $\{y_n\}$ satisfy

$$y_n = z_n + \epsilon_n,$$

with

$$z_n = \underset{\sim}{h}^T \underset{\sim n}{x} + \nu_n,$$

$$\underset{\sim n}{x} = G \underset{\sim n-1}{x} + \underset{\sim n}{\omega}, \qquad n = 1,2,\ldots$$

where $\{\nu_n\}$ are i.i.d. $N[0,\lambda^{-1}]$.

Now, as defined, this model cannot be directly analysed by the method of §5.1.2. due to the appearance of the only partially observable $\{z_n\}$ in $\underset{\sim n}{\omega}$

i.e. $\underset{\sim n}{\omega} = (\omega_n, 0, \ldots, 0)$ where $\omega_n = \sum_{j=1}^{r} \cdot \pi_j \, z_{n-j}$.

To overcome this we augment the state vector as follows. Define a new state vector $\underset{\sim}{g}_n$ by

$$\underset{\sim}{g}_n = (g_n, g_{n-1}, \ldots, g_{n-2r}) = (\underset{\sim}{z}_n^T, \underset{\sim}{x}_n^T),$$

where $$\underset{\sim}{z}^T = (z_n, z_{n-1}, \ldots, z_{n-r}).$$

The new state equation is

$$\underset{\sim}{g}_n = H\underset{\sim}{g}_{n-1} + \underset{\sim}{\nu}_n \qquad (5.2.6)$$

where

$$H = \begin{bmatrix} \pi_1, \ldots, \pi_r & \phi_1, \ldots \phi_r \\ I_{r-1} & 0 \\ \cdots\cdots\cdots & \cdots \\ \pi_1, \ldots, \pi_r & \\ 0 & G \end{bmatrix}$$

and $\underset{\sim}{\nu}_n^T = (\nu_n, 0, \ldots, 0)$ is $(2r \times 1)$.

For example, with an MA(1) model, $H = \begin{bmatrix} \phi_1 & -\phi_1 \\ \phi_1 & -\phi_1 \end{bmatrix}$.

The corresponding observation equation is

$$y_n = \underset{\sim}{h}^T \underset{\sim}{g}_n + \epsilon_n \qquad (5.2.7)$$

where now $\underset{\sim}{h}$ is $(2r \times 1)$, $\underset{\sim}{h}^T = (1, 0, \ldots, 0)$.

Thus given the normal/gamma prior for $(y_{n-1}, \lambda | \underset{\sim}{\theta}, \underset{\sim}{\phi}, D_{n-1})$ we have, say,

$(y_{n-1} | \lambda, \underset{\sim}{\theta}, \underset{\sim}{\phi}, D_{n-1}) \sim N[\overline{m}_{n-1}, \lambda^{-1} C_{n-1}]$ where as usual $\underset{\sim}{m}_{n-1}$, C_{n-1} are functions of $\underset{\sim}{\theta}$ and $\underset{\sim}{\phi}$ but not λ, and

$$(\lambda | \underset{\sim}{\theta}, \underset{\sim}{\phi}, D_{n-1}) \sim G\left[\alpha_{n-1}/2, \beta_{n-1}/2\right], \text{ where}$$

again β_{n-1} depends on $\underset{\sim}{\theta}$ and $\underset{\sim}{\phi}$.

Thus $(\underset{\sim}{y}_n | \lambda, \underset{\sim}{\theta}, \underset{\sim}{\phi}, D_{n-1}) \sim N\left[\underset{\sim}{a}_n, \lambda^{-1} P_n\right]$

with $\underset{\sim}{a}_n = H \underset{\sim}{m}_{n-1}$, $P_n = H C_{n-1} H^T + V$,

where $V_{11} = 1$ and $V_{ij} = 0$ otherwise.

Then the posterior density is a mixture of two components, as is the predictive density for y_n, according as ϵ_n is zero or not just as in §5.1.2.

i) $\epsilon_n = 0$, hence $y_n = z_n$. So $p(y_n|\lambda,\underset{\sim}{\theta},\underset{\sim}{\phi},D_n)$ is a singular normal distribution

$$N\left[\underset{\sim}{m}_n, \lambda^{-1}C_n\right]$$

with $\underset{\sim}{m}_n$, C_n derived via the Kalman filter. Note that we will have the first element of $\underset{\sim}{m}_n$ as y_n and the first row and column of C_n set to zeros. See equations (5.1.8) and (5.1.9).

The predictive density $p(y_n|\lambda,\underset{\sim}{\theta},\underset{\sim}{\phi},D_{n-1})$ is just $p(z_n|\lambda,\underset{\sim}{\theta},\underset{\sim}{\phi},D_{n-1})$ and so we simply take the prior Student t density $p(z_n|\underset{\sim}{\theta},\underset{\sim}{\phi},D_{n-1})$ as the likelihood component in the posterior update

$$p(\underset{\sim}{\theta},\underset{\sim}{\phi}|D_n) \propto p(\underset{\sim}{\theta},\underset{\sim}{\phi}|D_{n-1})p(z_n|\underset{\sim}{\theta},\underset{\sim}{\phi},D_{n-1}).$$

ii) $\epsilon_n \neq 0$. In this case we apply directly the modal equations just as in §5.1.2. See that section for details, equations (5.1.10) to (5.1.14) inclusive.

(c) Both types. It suffices to note that taking a contaminated normal mixture density for ν_n in (5.2.6) puts us into the framework of the AR IO and AO model and the analysis parellels §5.1.3.

Appendix 5A. Calculation of initial mean and variance matrix for state vectors in models of this Chapter.

For all outlier models of this Chapter we require the values of $\underset{\sim}{m}_0 = \underset{\sim}{m}_0(\underset{\sim}{\theta},\underset{\sim}{\phi})$ and $c_0 = c_0(\underset{\sim}{\theta},\underset{\sim}{\phi})$, the prior mean vector and covariance matrix of the state vector $\underset{\sim}{y}_n$ of (5.2.6). [The AR models are a special case with $\underset{\sim}{\phi} = \underset{\sim}{0}$]. As in Gardner et al (1980), taking the state equation at time n = 0 with $\underset{\sim}{g}_{-1}$ defined as $\underset{\sim}{g}_0$, we have

$$\underset{\sim}{g}_0 = H\,\underset{\sim}{g}_0 + \underset{\sim}{\nu}_0 .$$

Thus $\underset{\sim}{m}_0(\underset{\sim}{\theta},\underset{\sim}{\phi}) = \underset{\sim}{m}_0 = H\,\underset{\sim}{m}_0$ and so $\underset{\sim}{m}_0 = \underset{\sim}{0}$.

Further $C_0 = H\,C_0\,H^T + V$, where $V_{11} = 1$ and $V_{ij} = 0$ otherwise. Gardner et al (1980) provide an algorithm for the solution of this equation as a function of C_0.

Example 1. AR(1) model, $c_0 = \theta^2\,c_0 + 1$ or $c_0 = (1-\theta^2)^{-1}$.

Example 2. MA(1) model,

$$C_0 = \begin{pmatrix} -\phi & \phi \\ -\phi & \phi \end{pmatrix} C_0 \begin{pmatrix} -\phi & -\phi \\ \phi & \phi \end{pmatrix} + V$$

and so, if $C_0 = \begin{pmatrix} c_1 & c_3 \\ c_3 & c_2 \end{pmatrix}$, then

for $\phi = 0$, $C_0 = V$. Otherwise $c_2 = c_3 = \phi^{-2}$, $c_1 = 1 + \phi^{-2}$.

CHAPTER 6

Asymptotic theory of recursive algorithms.

6.1. Introduction

In this Chapter we examine some of the recursive algorithms of earlier Chapters in greater mathematical detail in order to obtain results about asymptotic behaviour. The recursions were constructed as approximations to the formal Bayesian analyses and examples have shown that their behaviour in finite samples is excellent. This Chapter is concerned with special cases of the models of the previous Chapters and, in particular, with constant parameters, where the notion of convergence is relevant.

Suppose we have a random sample of real valued observations $\{Y_1,\ldots,Y_n\}$ whose distribution depends upon an unknown real parameter θ. The problem of sequentially estimating θ by the sampling theory method of Stochastic Approximation (S.A.), originally developed by Robbins and Monro (1951), has been considered by several authors. Recent contributions by Kashyap, Blaydon and Fu (1970) and Fabian (1978), have provided widely applicable results about the asymptotic properties of S.A. algorithms, and Martin and Masreliez (1975), and Poljak and Tsyphin (1979) have applied such schemes to estimation in models such as those of Chapter 3. The basic form of S.A. depends upon the existence of an observable sequence $\{Z_1,\ldots,Z_n\}$ such that

$$E\left[Z_n \mid \theta, D_{n-1}\right] = 0 \text{ for all } n$$

where, as usual, $D_r = \{Y_1,\ldots,Y_r\}$. A sequence of estimates of θ is then defined recursively by

$$\theta_n = \theta_{n-1} - a_n Z_n$$

where

$$\sum_{n=1}^{\infty} a_n = \infty \text{ and } \sum_{n=1}^{\infty} a_n^2 < \infty$$

i.e: $\{a_n\}$ is harmonic.

Given further conditions on the sequence $\{Z_n\}$ (regularity conditions for the distribution of the Z_n), convergence of θ_n to θ in some sense can often be proved and, generally, the (sampling theoretic) asymptotic distribution of θ_n shown to be normal. Sacks (1958) proves results of this kind and Fabian (1978) provides a thorough analysis of both convergence and asymptotic normality of a general S.A. scheme. Fabian also discusses the so-called "asymptotically efficient" algorithms, providing a sequence of estimates θ_n whose sampling theoretic variance approaches the Cramer-Rao lower bound as n increases. The first investigation of such efficient schemes appears to have been the work of Sakrison (1965) in an engineering problem. Later Anbar (1973) and Abdelhamid (1973) considered transformations of the original observation $\{Y_n\}$ which lead to asymptotically efficient S.A. estimates.

We shall be concerned with these efficient schemes, the basic form of which is as follows. Suppose that the common density f of the Y_n is twice differentiable in θ with score function

$$g(y|\theta) = -\frac{\partial}{\partial\theta} \ell n\, f(y|\theta)$$

and Fisher Information

$$I(\theta) = E\left[\frac{\partial^2}{\partial\theta^2} g(y|\theta)\,|\,\theta\right].$$

[The extension to vector $\underset{\sim}{\theta}$ and $\underset{\sim}{Y}$ is obvious].

Then θ_n is defined by the recursion

$$\theta_n = \theta_{n-1} - n^{-1} A(\theta_{n-1})^{-1}\, g(y_n|\theta_{n-1}) \qquad (6.1.1)$$

where $A(x)$ is bounded above and below away from zero and $A(\theta) = I(\theta)$.

The recursion is intuitively attractive; for a "regular" problem

$$E[g(y|\theta)|\theta] = 0 \tag{6.1.2}$$

and so (6.1.1) is of the form of a stochastic gradient algorithm for finding the zero(s) of the regression function

$$M(X) = E[g(y|X)|\theta] \tag{6.1.3}$$

with the gain function $n^{-1} A(X)^{-1}$ being chosen to provide the correct asymptotic variance i.e. the Cramer-Rao lower bound.

From a Bayesian viewpoint (6.1.1) is attractive for the following reason.

Note that $g(y|X) = -\frac{\partial}{\partial X} \ell n \left\{\frac{f(y|X)}{f(y|\theta)}\right\}$, $f(y|\theta) \neq 0$,

so M(X) is the derivative with respect to X of the Kullback-Leibler directed divergence

$$K(\theta,X) = E\left[\ell n \left\{\frac{f(y|X)}{f(y|\theta)}\right\} \mid \theta\right]$$

(subject to regularity conditions). Thus the S.A. scheme (6.1.1) is constructed to locate the zeros of

$$M(X) = \frac{\partial}{\partial X} K(\theta,X) \tag{6.1.4}$$

and, since $K(\theta,X)$ is positive for $X \neq \theta$ with an absolute minimum at θ, (Kullback, 1959), then θ is one of the possible limits of θ_n. Now Berk (1966) discusses the asymptotic form of the posterior distribution $p(\theta|D_n)$ and shows that, subject to regularity conditions, $p(\theta|D_n)$ asymptotically concentrates on the set of values X in the range of θ such that $K(\theta,X)$ is minimized. Hence the efficient S.A. scheme asymptotically favours the same values as $p(\theta|D_n)$.

However, there are two major problems associated with these algorithms. For a Bayesian, an important question is that of the lack of a coherent basis for such sampling theory schemes. Secondly, and more practically, they are designed specifically for asymptotic optimality and small sample performance may be poor. Ljung (1978) discusses this and references illustrations of just how bad small sample behaviour can be for certain models.

Our recursions, (both the modal and exact forms), are constructed as approximate Bayes' "estimators" for any sample size and thus provide at least a partial solution to the above mentioned criticisms. A coherent basis exists and small sample performance has been illustrated by way of example and is generally excellent. We proceed now to examine the asymptotic properties of our recursions and discuss the meaning of the corresponding approximate posterior distributions.

We require a S.A. convergence result for our analysis and this appears in the Appendix 6. It is a generalization and extension of the result of Kashyap,Blaydon and Fu (1970) mentioned above, and provides convergence results for static regression and simple location problems, and for joint regression/scale estimation.

§6.2 Location and regression.

§6.2.1 Scalar location estimation.

Consider a random sample $\{y_1,\ldots,y_n\}$ from a unimodal, symmetric distribution with density p having unknown location θ. Adopting a normal likelihood as a model, the usual conjugate analysis leads to a posterior

$$(\theta|D_n) \sim N[m_n,C_n], \tag{6.2.1}$$

where

$$m_n = m_{n-1} + C_n\,(y_n - m_{n-1}), \tag{6.2.2}$$

and $$C_n^{-1} = C_{n-1}^{-1} + 1, \tag{6.2.3}$$

or $$C_n = C_{n-1} - C_{n-1}^2 \, (C_{n-1}+1)^{-1}. \tag{6.2.4}$$

Now (6.2.2) fits into the S.A. framework with the current "estimate" of θ, m_n, given by the previous estimate plus a correction term proportional to the current residual. The constant of proportionality, C_n, decays harmonically with n as evidenced by (6.2.3).

If we choose a heavy-tailed, non-normal likelihood, then our recursions of Chapter 3 lead to an approximate posterior distribution (6.2.1), with

$$m_n = m_{n-1} + C_{n-1} \, g_n(y_n - m_{n-1}), \tag{6.2.5}$$

and $$C_n = C_{n-1} - C_{n-1}^2 \, G_n(y_n - m_{n-1}), \tag{6.2.6}$$

where $g_n(u)$ is skew-symmetric about the origin and has a zero there, and $G_n(u) = \partial g_n(u)/\partial u$.

Now (6.2.5) resembles the efficient S.A. recursions with the difference that g_n is not the simple likelihood score function. For the exact recursion g_n is the score of the convolution of the likelihood with the $N[m_{n-1}, C_{n-1}]$ density; for the modal recursion

$$g_n(u) = \left(1+C_{n-1}\psi(u)\right)^{-1} g(u)$$

where $g(u) = \psi(u)\, u$ is the likelihood score. In both cases we have, essentially, a "smoothed" form of g(u) providing the response to the observation y_n. Furthermore, the gain function C_{n-1} is data dependent and, by analogy with the S.A. form, is required to behave harmonically with n. The fact that C_n may be greater than C_{n-1} is relevant in improving the small sample behaviour of the algorithm. Note that, for both our recursions, as $C_{n-1} \to 0$, $g_n(u) \to g(u)$ for all u leading

to the simple S.A. recursive form (6.1.1).

Now we can rewrite (6.2.6) in the form of (6.2.3), as

$$C_n^{-1} = C_{n-1}^{-1} + \gamma_n(y_n - m_{n-1}), \tag{6.2.7}$$

where $$\gamma_n(u) = G_n(u)\ [1 - C_{n-1} G_n(u)]^{-1}. \tag{6.2.8}$$

The denominator in (6.2.8) is positive by virtue of the positivity of C_n, C_{n-1} and (6.2.6). This form proves useful in examining the convergence of m_n, which we now do.

Note that, from (6.2.7), defining $A_n = n\, C_n$ then

$$A_n^{-1} = n^{-1} \sum_{r=1}^{n} \gamma_r(u_r) + n^{-1}\, C_0^{-1} \tag{6.2.9}$$

where $u_r = y_r - m_{r-1}$. In view of the above discussion, we expect the sequence $\{A_n\}$ to have a positive limit as $n \to \infty$, and this is just the sort of condition used by Martin and Masreliez (1975). In fact the following condition suffices in this simple scalar problem.

Condition 6.2.1.

There exist $m > \epsilon > 0$ such that, for all n,

$$\epsilon < A_n < m. \tag{6.2.10}$$

Notes (i) A stronger assumption will be required for the general regression model of §6.2.2.

(ii) For some likelihoods $\gamma_n(u)$ is always positive and, in that case, truncating such that $\gamma_n(u_n) > \gamma > 0$ for all n and u leads to an upper bound as in (6.2.10). For Student t and normal/uniform likelihoods, $\gamma_n(u)$ may take negative values and in such cases we must truncate A_n^{-1} below away from zero as necessary. Similarly, for most likelihoods $\gamma_n(u)$ is bounded above for all

n and u and so A_n^{-1} is bounded above. In exceptional cases a truncation of A_n^{-1} directly is again necessary.

Of course in practice, since we process only a finite sample, we simply use the recursions without modification. Truncations of this kind, as used by Martin and Masreliez (1975) for example, are purely technical devices for use in the convergence proof.

Now the other conditions for convergence of the recursions are given in the following result.

Lemma 6.2.1. Define $u_n = y_n - x$, and

$$r_n(x|\theta) = E\left[A_{n-1}\, g_n(u_n)\,|\,\theta, \bar{x}\right].$$

Then, for all n, if our score function satisfies $|g_n(u)| < k|u|$,

(i) $r_n(x|\theta) = 0$ if and only if $x = \theta$;

(ii) $\inf_{\epsilon<|x-\theta|<\epsilon^{-1}} (\theta-x)\, r_n(x|\theta) > 0$, for $\epsilon > 0$;

(iii) $E\left[A_{n-1}^2\, g_n^2(u_n)\,|\,\theta, \bar{x}\right] \leq h\left[1+(\theta-x)^2\right]$, $h > 0$.

Proof:

$g_n(u)$ is skew-symmetric about zero and $p(u)$ is symmetric, therefore

$$\int_{-\infty}^{\infty} g_n(y-x)p(y-\theta)dy$$

$$= \int_{-\infty}^{\infty} g_n(y)\,\{p(y-a) - p(y+a)\}dy$$

where $a = x-\theta$.

So $r_n(x|\theta) = r_n(a)$ and, for both a and y positive, $g_n(y) > 0$ and $p(y-a) > p(y+a)$. Thus $r_n(a) > 0$. Also, by symmetry, $r_n(a) < 0$ for $a < 0$. Finally $r_n(a) = 0$ and so (i) and (ii) hold.

Further $E\left[A_{n-1}^2\, g_n^2(u_n)\,|\,\theta,x\right] \leq M^2 E\left[g_n^2(u)\,|\,\theta,x\right]$ by condition 6.1.1. Now for our heavy-tailed likelihoods, $|g_n(u)| < k|u|$, $k > 0$ for all n. Therefore

$$E\left[g_n^2(u_n)\,|\,\theta,x\right] < k^2\{\mathrm{var}\left[y_n|\theta\right] + (\theta-x)^2\}$$

$$< h\left[1+(\theta-x)^2\right], \quad h > 0$$

whenever the second moment of p is finite. If p has no variance, then we must use a bounded score function. In fact if $|g_n(u)| < k$ for all n,u as, for example, with a Student t density, then (iii) holds directly without the requirement of the existence of any moments of p.

Using Lemma 6.2.1 and Theorem A6.1 we immediately deduce the following:

Theorem 6.2.1.

If condition 6.2.1 holds then

$$m_n \to \theta \text{ with probability one.}$$

Proof:

The result follows as a special case of Theorem A6.1 with

(i) $A_n = 1$, $\Delta_n = 0$ for all n,

(ii) $g_{\underset{\sim}{n+1}}(y_{\underset{\sim}{n+1}}|\underset{\sim}{\theta},\underset{\sim}{X}) = n\, C_n\, g_{n+1}(y_{n+1}-x)$,

(iii) $r_{\underset{\sim}{n+1}}(\underset{\sim}{X}|\underset{\sim}{\theta}) = n\, C_n\, E\left[g_{n+1}(y_{n+1}-x)\,|\,\theta,x\right]$,

and the properties established in Lemma 6.2.1.

Asymptotic distribution.

Under regularity conditions on the score and information functions of the assumed likelihood, Heyde and Johnstone (1978) examine the asymptotic form of $p(\theta|D_n)$. Given that the posterior

mode θ_n is consistent for θ, that the information function $G(y-\theta)$ is continuous around θ and that

$$P_n^{-1} = \sum_{k=1}^{n} G(y_i-\theta_n)$$

tends to infinity with n, their result is that $p(\theta|D_n)$ is asymptotically normal with mean θ_n and variance P_n.

With the stronger assumption that $n^{-1}P_n^{-1}$ converges to a finite non-zero limit, we show that our approximate normal posterior distribution agrees with this result and so our approximation is "asymptotically efficient".

Corollary 6.2.1.

If $n^{-1}P_n^{-1} \to P > 0$ as $n \to \infty$,

Then $n^{-1}\left[C_n^{-1}-P_n^{-1}\right] \to 0$, with probability one.

Proof: By virtue of (6.2.1), $C_n \to 0$ as $n \to \infty$. For the modal recursion with likelihood score $g(u) = u\psi(u)$ and information $G(u)$, then

$$G_n(u) = \frac{G(u)}{\left[1+C_{n-1}\psi(u)\right]} - \frac{g(u)\dot{\psi}(u)\,C_{n-1}}{\left[1+C_{n-1}\psi(u)\right]^2}$$

So $G_n(u) - G(u) \to 0$ as $n \to \infty$ for all u. Similarly this is true for the exact recursion. Further, from (6.2.8), $\gamma_n(u) - G_n(u) \to 0$ as $n \to \infty$ and thus, since

$$A_n^{-1} = n^{-1}\,C_n^{-1} = n^{-1}C_0^{-1} + n^{-1}\sum_{r=1}^{n}\gamma_r(u_r)$$

we have

$$n^{-1}\left[C_n^{-1} - \sum_{r=1}^{n} G(u_r)\right] \to 0 \quad \text{as } n \to \infty.$$

By continuity of G and consistency of m_n and θ_n we have

$$n^{-1} \sum_{r=1}^{n} G(y_r - x_r) - \sum_{r=1}^{n} G(y_r - \theta) \to 0$$

with probability one for $x_r = m_r$ and $x_r = \theta_r$. So $n^{-1}(C_n^{-1} - P_n^{-1}) \to 0$ almost surely. □

Note that since P_n and C_n are both $O(n^{-1})$, this implies that there exists $M_n = O(n)$ such that

$$M_n(C_n - P_n) \to 0$$

with probability one.

§6.2.2. The Regression Case.

Given vector observations $\{\underset{\sim}{Y}_n; n=1,2,\ldots\}$ in $\mathbb{R}^m$ such that

$$\underset{\sim}{Y}_n = H_n \underset{\sim}{\theta} + \underset{\sim}{\nu}_n, \quad n = 1,2,\ldots \quad \underset{\sim}{\theta} \in \mathbb{R}^p,$$

where the $\underset{\sim}{\nu}_n$ are independent identically distributed with a unimodal symmetric density $p(\underset{\sim}{\nu})$, the generalizations of (6.2.5) are

$$\underset{\sim}{M}_n = \underset{\sim}{M}_{n-1} + C_{n-1} H_n^T \underset{\sim}{g}_n(\underset{\sim}{u}_n) \tag{6.2.11}$$

$$C_n = C_{n-1} - C_{n-1} H_n^T G_n(\underset{\sim}{u}_n) H_n C_{n-1}, \tag{6.2.12}$$

$\underset{\sim}{u}_n = \underset{\sim}{Y}_n - H_n \underset{\sim}{M}_{n-1}$.

Now both Fabian (1978) and Kashyap, Blaydon and Fu (1970) use conditions on the recursion (6.2.11) that we cannot generally satisfy. Poljak and Tsypkin (1979) replace the gain matrix C_{n-1} with $n^{-1}C$ for a positive definite matrix C and subsequently note that their convergence results hold for gains of the form C_{n-1} such that $n\,C_{n-1}$ converges to C in some sense as $n \to \infty$. We adopt a similar assumption.

Condition 6.2.2.

There exist $M>\delta>0$, $K>0$, $\epsilon_n>\epsilon>0$ and positive definite symmetric matrices Δ_n such that, if $A_n = n\, C_n$,

(i) $\delta < ||A_n|| < M$

(ii) $A_n = C + \Delta_n$, $||\Delta_n|| < Kn^{-\epsilon n}$.

for some positive definite symmetric matrix C.

Note that $C_n^{-1} = C_0^{-1} + \sum_{r=1}^{n} \Gamma_r(\underset{\sim}{u}_r)$ (6.2.13)

where $\Gamma_r(\underset{\sim}{u}) = \left[I - C_{r-1} H_r^T G_r(\underset{\sim}{u}) H_r\right]^{-1} \left[H_r^T G_r(\underset{\sim}{u}) H_r\right]$ (6.2.14)

So Condition 6.2.2 requires that $n^{-1} \sum_{r=1}^{n} \Gamma_r(\underset{\sim}{u}_r)$ is bounded above and below (away from zero) in norm, and has a positive definite limit C with the difference between it and its limit decaying as an inverse power of n. A similar condition is used by Sakrison (1965) although he uses further restrictions to define a particular value of C.

Theorem 6.2.2.

Let Condition 6.2.2 hold. Suppose we assume a heavy-tailed likelihood and that $E\left[\,||H_n||^{\,4}\right] < \ell < \infty$. Then $\underset{\sim}{M}_n$ converges almost surely to $\underset{\sim}{\theta}$.

[If we adopt a robust likelihood then the condition on the H_n can be weakened to $E\left[\,||H_n||^{\,2}\right] < \ell < \infty$.]

Proof:

$$\text{Let } \underset{\sim}{r}_n(\underset{\sim}{X}|\underset{\sim}{\theta}) = E\left[H_n \underset{\sim}{g}_n(\underset{\sim}{Y}_n - H_n \underset{\sim}{X}) \,\middle|\, \underset{\sim}{\theta}, \underset{\sim}{X}\right]$$

$$= E\left[H_n \underset{\sim}{g}_n(\underset{\sim}{a}_n + \underset{\sim}{\nu}_n) \,\middle|\, \underset{\sim}{\theta}, \underset{\sim}{X}\right] = \underset{\sim}{r}_n(\underset{\sim}{a}_n),$$

where $\underset{\sim}{a}_n = H_n(\underset{\sim}{\theta} - \underset{\sim}{X})$. Clearly, as in the scalar case of Lemma 6.1.1, the

symmetry and unimodality of $p(\underset{\sim}{\nu})$ and to skew-symmetry of $g_n(\underset{\sim}{u})$ imply that $\underset{\sim n}{r}(\underset{\sim n}{a}) = \underset{\sim}{0}$. Furthermore, $(\underset{\sim}{\theta}-\underset{\sim}{X})^T \underset{\sim n}{r}(\underset{\sim}{X}|\underset{\sim}{\theta}) = E\left[\underset{\sim n}{a}^T \underset{\sim n}{g}(\underset{\sim n}{a}+\underset{\sim n}{\nu}) \,|\, \underset{\sim}{\theta},\underset{\sim}{X}\right]$ is positive for $\underset{\sim n}{a} \neq \underset{\sim}{0}$ by the same conditions. Now for our heavy-tailed likelihoods we have either

$$\text{(a)} \quad \|g_n(\underset{\sim}{u})\| < K\|\underset{\sim}{u}\| \qquad \text{or (b)} \quad \|g_n(\underset{\sim}{u})\| < K,$$

for all n, u.

In case (a), $E\left[\|H_n \underset{\sim n}{g}(\underset{\sim n}{a}+\underset{\sim n}{\nu})\| \;|\; \underset{\sim}{\theta},\underset{\sim}{X}\right]$

$$< K\, E\left[\|H_n\|\left(\|\underset{\sim n}{a}\| + \|\underset{\sim n}{\nu}\|\right) \,|\, \underset{\sim}{\theta},\underset{\sim}{X}\right]$$

$$< K\, E\left[\|H_n\|^2 \|\underset{\sim}{\theta}-\underset{\sim}{X}\|\right] + b\, E\left[\|\underset{\sim n}{\nu}\|\right]$$

$$< c\|\underset{\sim}{\theta}-\underset{\sim}{X}\| + d, \quad \text{say, } c, d > 0,$$

when $E\left[\|\underset{\sim n}{u}\|\right] < \infty$.

Similarly, using $E\left[\|H_n\|^4\right] < \ell < \infty$,

$$E\left[\|H_n \underset{\sim n}{g}(\underset{\sim}{a}+\underset{\sim n}{\nu})\|^2 \,|\, \underset{\sim}{\theta},\underset{\sim}{X}\right] < f\|\underset{\sim}{\theta}-\underset{\sim}{X}\|^2 + g, \; f, g > 0;$$

when $E\left[\|\underset{\sim n}{u}\|^2\right] < \infty$.

In case (b),

$$E\left[\|H_n \underset{\sim n}{g}(\underset{\sim n}{a}+\underset{\sim n}{\nu})\| \;|\; \underset{\sim}{\theta},\underset{\sim}{X}\right] < K \;\; E\left[\|H_n\|\right] < \ell', \text{ say}$$

and $E\left[\|H_n \underset{\sim n}{g}(\underset{\sim n}{a}+\underset{\sim n}{\nu})\|^2 \,|\, \underset{\sim}{\theta},\underset{\sim}{X}\right] < m$, say.

[Note that using a robust likelihood as in (b) requires only that $E\left[\|H_n\|^2\right] < \infty$.] .

The above conditions are just those of Theorem A6.1 so we deduce that M_n converges with probability one and

$$\lim_{n\to\infty} E\left[\underset{\sim n}{a}^T \underset{\sim n}{g}(\underset{\sim n}{a}+\underset{\sim n}{\nu}) \,|\, \underset{\sim}{\theta}, \underset{\sim}{X}=\underset{\sim n-1}{M}\right] = 0$$

However this implies that $\lim_{n\to\infty} H_n(\underset{\sim}{\theta}-\underset{\sim}{M}_{n-1}) = 0$ with probability one by the positivity condition and hence $\underset{\sim}{M}_n$ converges to $\underset{\sim}{\theta}$. □

Corollary 6.2.2.

As in the scalar case, the correct asymptotic distribution is provided i.e: our approximation is "asymptotically efficient".

Proof:

As in Corollary 6.2.1, $C_n \to 0$ and $n^{-1} C_n^{-1} - n^{-1} P_n^{-1} \to 0$ almost surely,

where $P_n^{-1} = \sum_{r=1}^{n} H_r^T G(\underset{\sim}{Y}_r - H_r \underset{\sim}{\theta}_n) H_r$ is assumed $O(n)$,

G being the information matrix of p and $\underset{\sim}{\theta}_n$ the mode of $p(\underset{\sim}{\theta}|D_n)$. Note that, since P_n and C_n are $O(n^{-1})$, there exist a matrix M_n of order n such that $M_n(C_n - P_n) \to 0$ almost surely.

Example.

As an example consider the innovations outlier model for autoregressions discussed in §5.1.1.

$$y_n = \sum_{r=1}^{p} \theta_r y_{n-r} + \nu_n \qquad n = 1,2,\ldots$$

So here $H_n = \underset{\sim}{h}_n^T = (y_{n-1},\ldots,y_{n-p})$.

Defining $\Gamma_r = \Gamma_r(u_r)$ of (6.2.14) we have

$$H_r^T \Gamma_r(u_r) H_r = \Gamma_r A_r$$

where A_r is a $p\times p$ matrix with ij element

$$(A_r)ij = y_{r-i}\, y_{r-j}.$$

Thus $n^{-1} C_n^{-1} = n^{-1} \sum_{r=1}^{n} \Gamma_r + n^{-1} C_0^{-1}$ has a finite limit when, for example, $\mathrm{var}[\Gamma_r y_{r-i} y_{r-j}]$ is uniformly bounded for all r, i, j. In particular, if Γ_r is bounded above and is positive for all r

(as, for example, with the logistic likelihood), this condition coincides with the requirement of Theorem 6.2.2, that the fourth moment of $p(\nu)$ exists. Further, stationarity of the process is required by Theorem 6.2.2 implying a restricted range of values for $\underset{\sim}{\theta}$. In such cases the restriction of recursions to lie within such a range does not affect convergence and, in practice, is not always necesssary.

§6.3. Scale problems.

§6.3.1. Simple scale problem for a random sample.

Given $\{Y_n;\ n=1,2,\}$ consisting of independent observations with common unimodal, symmetric and heavy-tailed density with unknown scale σ, the approximation derived in §4.3.1 leads to a posterior gamma distribution for $\lambda=\bar{\sigma}^2$

$$(\lambda|D_n) \sim G[\alpha_n/2, \beta_n/2] \tag{6.3.1}$$

where

$$\alpha_n = \alpha_{n-1} + 1, \tag{6.3.2}$$

$$\beta_n = \beta_{n-1} + y_n^2\psi(y_n/\sigma_{n-1}), \tag{6.3.3}$$

and the likelihood score is $g(n) = u\psi(u)$. Further the mean of (6.3.1) is $\alpha_n/\beta_n = \sigma_n^{-2}$ where

$$\sigma_n^2 = \sigma_{n-1}^2 + \alpha_n^{-1}\left[y_n^2\psi(y_n/\sigma_{n-1}) - \sigma_{n-1}^2\right]. \tag{6.3.4}$$

To prove convergence we assume that σ_n^2 satisfies

Condition 6.3.1. There exists $M > \epsilon > 0$ with $\epsilon < \sigma_n^2 < M$ for all n.

[Note that this truncation is irrelevant in practice: simply choose $\epsilon(M)$ to be the smallest (largest) numbers available on whatever machine is used for calculation].

Theorem 6.3.1. If Condition 6.3.1 holds and our likelihood is heavy-tailed then σ_n^2 converges with probability one to the solution σ_0^2 of $\sigma_0^2 = E\left[y^2\psi(y/\sigma_0)\right]$.

Proof: $\alpha_n = n + \alpha_0$ so noting the truncation of condition 6.3.1, $A_n = \sigma_{n-1}^2 \alpha_n^{-1}$ is harmonic. For our heavy-tailed likelihoods, $u^2\psi(u)$ is increasing in u; (this is easily checked for the likelihoods of Appendix 2.) Thus

$$M(\sigma^2) = E\left[(y^2/\sigma^2)\psi(y/\sigma) - 1\right]$$

is decreasing in σ^2. Furthermore $M_n(\sigma^2)$ is continuous in σ^2 and

i) $\lim_{\sigma^2 \to 0} M(\sigma^2) > 0$ (in some cases ∞);

ii) $\lim_{\sigma^2 \to \infty} M(\sigma^2) < 0$.

For example, with a Student t-k likelihood, the limit in i) is k, and that in ii) is -1.

So there exists a unique root σ_0^2 of $M(\sigma^2) = 0$, and, in addition, $(\sigma_0^2 - \sigma^2)\, M(\sigma^2) > 0$ for $\sigma^2 \neq \sigma_0^2$. Finally, since either $|g(u)| < k$ for all u or $|g(u)| < k|u|$ for all u we can find a constant $C > 0$ satisfying

$$E\left[\left\{(y^2/\sigma^2)\psi(y/\sigma) - 1\right\}^2\right] < C\left[1 + (\sigma^2 - \sigma_0^2)^2\right] .$$

as a function of σ^2. In particular if p is one of our robust likelihoods then $u^2\psi(u)$ is bounded above by a constant. Otherwise the existence of the fourth moment of the true likelihood is a sufficient condition for this bound.

Now these conditions lead to the satisfaction of those of Theorem A6.1. Alternatively, (since A6.1 is a very general result), the original convergence result of Kashyap, Blaydon and Fu (1970) may be

used directly to give probability one convergence of σ_n^2 to σ_0^2 as stated. □

Note that, as in §4.3.1, we have the identity

$$y^2\psi(y|\sigma) - \sigma^2 = -2\sigma^4 \frac{\partial}{\partial\sigma^2}\ell n\ p(y|\sigma^2) \qquad (6.3.5)$$

where $p(y|\sigma^2) = \frac{1}{\sigma}\ p\left(\frac{y}{\sigma}\right)$ is the likelihood. So (6.3.4) is a gradient algorithm and the convergence point σ_0^2 satisfies

$$E\left[-\frac{\partial}{\partial\sigma_0^2}\ \ell n\ p(y|\sigma_0^2)\right] = 0,$$

which, as in Berk (1966), is the set of concentration of the posterior distribution $p(\sigma^2|D_n)$. For the asymptotic distribution note that $\alpha_n = n + \alpha_0$ and $\text{var}[\lambda|D_n] \simeq \alpha_n/\beta_n^2 = \alpha_n^{-1}\ \sigma_n^{-4}$, therefore (6.3.1) is asymptotically normal

$$(\lambda|D_n) \sim N\left[\sigma_n^{-2},\ \alpha_n^{-1}\ \sigma_n^{-4}\right] \quad \text{as } n \to \infty$$

This agrees with the asymptotic form of $p(\lambda|D_n)$ when the likelihood is normal. Otherwise we could obtain asymptotic agreement by eventually using a second order approximation to $p(\lambda|D_n)$ rather than the first order scheme of §4.3.1. Clearly it would be preferable to approximate the posterior of, for example, $\ell n(\lambda)$ which will be closer to normality.

Having said this, the heuristic justification and experimental verification of the accuracy of the original approximation of (6.3.1) leads us to believe that there is little practical gain in adopting the asymptotically efficient form.

§6.3.2. Joint regression/scale problem.

Consider now the model

$$y_n = \underset{\sim}{h}_n^T\ \underset{\sim}{\theta} + \nu_n, \qquad n=1,2,\ldots$$

with the assumptions of §6.2.2 and §6.3.1. Directly from §4.3.1 we have the approximations

$$\text{(i)}\quad (\underset{\sim}{\theta}|D_n,\lambda) \sim N\left[\underset{\sim}{M}_n, \lambda^{-1} C_n\right],$$

where $\underset{\sim}{M}_n$, C_n are defined by (6.2.11, 12) with the difference that $u_n = (y_n - \underset{\sim}{h}_n^T \underset{\sim}{M}_{n-1})/\sigma_{n-1}$;

$$\text{(ii)}\qquad (\lambda|D_n) \sim G\left[\alpha_n/2, \beta_n/2\right],$$

where α_n, β_n are defined by (6.3.2, 3) with y_n replaced by $\epsilon_n = \sigma_{n-1}\, u_n$, $\psi(u)$ replaced by

$$\psi_n(u) = \left(1+q_n^2\psi(u)\right)^{-1}\psi(u)$$

and $q_n^2 = \underset{\sim}{h}_n^T C_{n-1} \underset{\sim}{h}_n$.

Now we can write

$$\underset{\sim}{X}_{n+1} = \underset{\sim}{X}_n + n^{-1} A_n \underset{\sim}{r}_n(\epsilon_{n+1})$$

where $\underset{\sim}{X}_n^T = (\underset{\sim}{M}_n, \sigma_n^2)$,

$$A_n = \begin{pmatrix} n\,C_n & \vdots & \underset{\sim}{0} \\ \cdots & \vdots & \cdots \\ \underset{\sim}{0}^T & \vdots & n\alpha_{n+1}^{-1} \end{pmatrix},$$

and $\underset{\sim}{r}_n(\epsilon_{n+1})^T = \left(\underset{\sim}{h}_{n+1}^T g_n(u_{n+1}),\ \epsilon_{n+1}^2\psi_{n+1}(u_{n+1}) - \sigma_n^2\right)$,

and use Theorem A6.1 to show convergence of $\underset{\sim}{X}_n$ as follows.

Theorem 6.3.2. Given conditions 6.2.1 and 6.3.1, the use of a heavy-tailed likelihood means that $\underset{\sim}{X}_n$ converges with probability one to $(\underset{\sim}{\theta}, \sigma_0^2)^T$.

Proof: Under the assumptions above, the appearance of σ_n^2 in the recursions for $\underset{\sim}{M}_n$ and C_n is of no importance; it essentially means that we are adopting a different scale factor for each observation. In particular Theorem 6.2.2 holds for the modified $\underset{\sim}{M}_n$ recursion and

$\underset{\sim n}{M} \to \underset{\sim}{\theta}$ with probability one irrespective of the behaviour of σ_n^2 (subject to Condition 6.3.1 of course).

For σ_n^2 note that the zero of the regression function

$$E\left[\epsilon_{n+1}^2 Y_{n+1}(u_{n+1}) - \sigma_n^2\right] = 0 \qquad (6.3.7)$$

depends on n and we have a dynamic stochastic approximation scheme such as discussed by Uosaki (1974) and Ruppert (1979). We could use the results of these works in modified form to show convergence for our scheme but it is much simpler to proceed as follows.

We have

$$\psi_{n+1}(u) = \psi(u) - q_{n+1}^2 \psi(u)^2 / \left(1+q_{n+1}^2 \psi(u)\right)$$

and so

$$\sigma_{n+1}^2 = \sigma_n^2 + \alpha_n^{-1}\left[\epsilon_{n+1}^2 \psi(u_{n+1}) - \sigma_n^2\right] - \alpha_n^{-1} q_{n+1}^2 M_n(\epsilon_{n+1})$$

where $M_n(\epsilon) = \epsilon^2 \psi^2(\epsilon/\sigma_n) / \left(1+q_{n+1}^2 \psi(\epsilon/\sigma_n)\right)$.

Now if $E\left[\underset{\sim n}{M}(\epsilon_{n+1}) \middle| \underset{\sim}{\theta}, \underset{\sim n}{m}\right]$ is bounded above, condition (i) of Theorem A6.1 holds since both α_n^{-1} and q_{n+1}^2 are $O(n^{-1})$. Indeed, for robust likelihoods $M_n(\epsilon)$ is bounded above for all n; for our other heavy-tailed likelihoods, $E\left[M_n(\epsilon_{n+1}) \middle| \underset{\sim}{\theta}, \underset{\sim n}{m}\right]$ is bounded if the second moment of p exists. The proof of Theorem 6.3.1 can now be used to show that the σ_n^2 part of the recursion for $\underset{\sim n}{X}$ also satisfies the conditions of Theorem A6.1; the details are the same. From this we deduce that $\underset{\sim n}{X}$ converges with probability one to the stated point $(\underset{\sim}{\theta}, \sigma_0^2)$.

As far as the asymptotic distribution is concerned, the comments of the previous two sections are pertinent. Full asymptotic efficiency is not achieved for $(\sigma^2 | D_n)$ but a minor modification would permit this. In practice, the original scheme is more than adequate as an approximation to the true posterior distribution.

Appendix 6.

Theorem A6.1. [Based on Kashyap, Blaydon and Fu (1970)].

Let $\{\underset{\sim}{Y}_n;\ n=1,2,\ldots\}$ be a sequence of $(m\times 1)$ random vectors with distribution depending upon the $(p\times 1)$ parameter vector $\underset{\sim}{\theta}$; $\{g_n;\ n=1,2,\ldots\}$ a sequence of functions of $\mathbb{R}^m$ to $\mathbb{R}^p$ with $\underset{\sim}{g}_n(\underset{\sim}{Y}_n) = \underset{\sim}{g}_n(\underset{\sim}{Y}_n|\underset{\sim}{\theta})$ possible depending upon D_{n-1}; $\{\underset{\sim}{A}_n;\ n=1,2,\ldots\}$ a sequence of $(p\times p)$ positive definite matrices satisfying Condition 6.2.2.

Define $\underset{\sim}{M}_{n+1} = \underset{\sim}{M}_n + n^{-1}\, A_n\, \underset{\sim}{g}_{n+1}(\underset{\sim}{y}_{n+1}|\underset{\sim}{M}_n)$, $\|\underset{\sim}{M}_0\| < \infty$,

and

$$\underset{\sim}{r}_n(\underset{\sim}{X}|\underset{\sim}{\theta}) = E\left[\underset{\sim}{g}_n(\underset{\sim}{y}_n|\underset{\sim}{X})\,|\,\underset{\sim}{\theta},\underset{\sim}{X}\right].$$

If

(i) $\underset{\sim}{r}_n(\underset{\sim}{\theta}|\underset{\sim}{\theta}) = \underset{\sim}{0}$;

(ii) $\inf_{\sigma<\|\underset{\sim}{X}-\underset{\sim}{\theta}\|<\alpha^{-1}} (\underset{\sim}{\theta}-\underset{\sim}{X})^T \underset{\sim}{r}_n(\underset{\sim}{X}|\underset{\sim}{\theta}) \geq 0$, for all $\alpha > 0$;

(iii) $E\left[\|\underset{\sim}{g}_n(\underset{\sim}{y}_n|\underset{\sim}{X})\| \mid \underset{\sim}{\theta},\underset{\sim}{X}\right] < k\,\|\underset{\sim}{\theta}-\underset{\sim}{X}\| + b$, $b,k > 0$;

(iv) $E\left[\|\underset{\sim}{g}_n(\underset{\sim}{y}_n|\underset{\sim}{X})\|^2 \mid \underset{\sim}{\theta},\underset{\sim}{X}\right] < h\left[1+\|\underset{\sim}{\theta}-\underset{\sim}{X}\|^2\right]$, $h > 0$;

then $\underset{\sim}{M}_n$ converges with probability one to $\underset{\sim}{\theta}_0$ where

$$(\underset{\sim}{\theta}_0-\underset{\sim}{\theta})^T \lim_{n\to\infty} \underset{\sim}{r}_n(\underset{\sim}{\theta}_0|\underset{\sim}{\theta}) = 0.$$

In particular $\underset{\sim}{\theta}_0 = \underset{\sim}{\theta}$ if either (i) is a unique solution or $\lim_{n\to\infty} \underset{\sim}{r}_n(\underset{\sim}{X}|\underset{\sim}{\theta})$ has only $\underset{\sim}{\theta}$ as a possible zero.

Proof:

Let $\underset{\sim}{X}_n = \underset{\sim}{M}_n - \underset{\sim}{\theta}$, $q_n = \underset{\sim}{X}_n^T\, C^{-1}\, \underset{\sim}{X}_n$, and $\underset{\sim}{M}^n = \{\underset{\sim}{M}_0,\ldots,\underset{\sim}{M}_n\}$.

Then

$$q_{n+1} = q_n + \frac{2}{n}\,\underset{\sim}{g}_{n+1}^T(\underset{\sim}{Y}_{n+1}|\underset{\sim}{M}_n)\,\underset{\sim}{X}_n + \frac{2}{n}\,\underset{\sim}{g}_{n+1}^T(\underset{\sim}{Y}_{n+1}|\underset{\sim}{M}_n)\,\Delta_n\, C\,\underset{\sim}{X}_n$$

$$+ \frac{1}{n^2}\,\underset{\sim}{g}_{n+1}^T(\underset{\sim}{Y}_{n+1}|\underset{\sim}{M}_n)\, P_n\, \underset{\sim}{g}_{n+1}(\underset{\sim}{Y}_{n+1}|\underset{\sim}{M}_n),$$

where $P_n = (C+\Delta_n)\ C^{-1}(C+\Delta_n)$.

Now there exist constants $a_1,\ldots,a_5$ such that, by Condition 6.2.2, (iii) and (iv),

$$E\left[g_{\underset{\sim}{n+1}}^T(Y_{\underset{\sim}{n+1}}|M_{\underset{\sim}{n}})\ \Delta_n\ C\ X_{\underset{\sim}{n}}|\theta_{\underset{\sim}{}},M^n\right] < \frac{a_1}{n^{\epsilon_n}}\ q_n + \frac{a_2}{n^{\epsilon_n}}$$

and

$$E\left[g_{\underset{\sim}{n+1}}^T(Y_{\underset{\sim}{n+1}}|M_{\underset{\sim}{n}})\ P_n\ g_{\underset{\sim}{n+1}}(Y_{\underset{\sim}{n+1}}|M_{\underset{\sim}{n}})\Big|\theta_{\underset{\sim}{}},M^n\right]$$

$$< \left[1+q_n\right]\left[a_3 + \frac{a_4}{n^{\epsilon_n}} + \frac{a_5}{n^{2\epsilon_n}}\right].$$

Thus

$$E\left[q_{n+1}|\theta_{\underset{\sim}{}},M^n_{\underset{\sim}{}}\right] \leq \frac{2}{n} E\left[g_{\underset{\sim}{n+1}}^T(Y_{\underset{\sim}{n+1}}|M_{\underset{\sim}{n}})X_{\underset{\sim}{n}}|\theta_{\underset{\sim}{}},M^n\right]$$

$$+ q_n\left[1+\mu_n\right] + f_n, \qquad (1)$$

where $\sum_{n=1}^{\infty} \mu_n$ and $\sum_{n=1}^{\infty} f_n$ both converge,

and from (ii) we deduce that $E\left[q_{n+1}|\theta_{\underset{\sim}{}},M^n_{\underset{\sim}{}}\right] \leq {}_n\left[1+\mu_n\right] + f_n.$ (2)

From this point on, the proof of Kashyap, Blaydon and Fu (1970) can be followed (using (1) and (2)), to show that $X_{\underset{\sim}{n}}$ converges almost surely and, also,

$$\lim_{n\to\infty} (M_{\underset{\sim}{n}}-\theta_{\underset{\sim}{}})^T\ r_{\underset{\sim}{n+1}}\ (M_{\underset{\sim}{n}}|\theta_{\underset{\sim}{}}) = 0_{\underset{\sim}{}} \text{ with}$$

probability one. The result follows.

Chapter 7. Multi-State Modelling.

7.1 Introduction.

This Chapter comprises a report on some work done in conjunction with the Renal Unit at the City Hospital in Nottingham. The problem is one of monitoring medical time series made up of observations on various chemical indicators from patients in post operative care. In particular, in monitoring kidney transplant recipients, sequences of chemicals in the blood and urine are routinely collected in order to assess renal function and, in this context, the identification and examination of trends in the data provides a useful guide to the state of the transplant at any time. Clearly some form of statistical processing of the data would appear to be necessary to clarify the patterns of behaviour evidenced in this, and other, medical time series.

Previous analyses of medical time series have focused on monitoring steady behaviour of the variables of interest. In particular, Chik et al (1979), in a study of foetal heart rate variability, apply simple exponentially weighted moving average (EWMA) techniques to smooth the data for presentation to clinicians. Hill and Endresen (1978) are concerned with monitoring heart rates and blood pressures of patients in intensive care. They adopt essentially the same approach for smoothing data, although they base their analysis on Kalman filtering techniques.

Our coherent approach uses the dynamic Bayesian linear model as a flexible basis for sensible analyses of time series which exhibit structural changes, some of paramount importance, others incidental. In particular, abrupt changes of certain kinds may indicate some form of relapse and, in renal care, the possibility of rejection of a transplant. The multi-state model of Harrison and Stevens (1976) models such phenomena automatically, whereas

alternative monitoring techniques, such as those used by Chik et al and Hill and Endreson, break down on the occurrence of such events. Stoodley and Mirnia (1979) adopt a linear growth model as used by Harrison and Stevens, and devise an automatically resetting CUSUM to detect changes. They compare their system favourably with the multi-state model, commenting that the latter requires a great deal of expertise for satisfactory operation and that the purpose built computer program they used was very demanding of time and storage. We return to those comments in §7.3.

In §7.2 we discuss aspects of the data from the Renal Unit in Nottingham and describe our construction of a model. §7.3 outlines the use of the linear growth model in a multi-state framework and discusses some problems encountered in our early attempts at analysis. Theoretical extensions of the model to overcome such problems are then presented. Finally §7.4 contains examples of outputs of the analyses and a discussion of the problems of making inferences about changes in renal function from the apparent changes in the monitored indicators.

7.2 Kidney Transplant Study.

7.2.1 Transplant Data.

Historical perspective.

In caring for kidney transplant recipients, doctors are aware that dramatic changes in the function of the transplanted organ can occur suddenly and lead to a serious relapse and possible rejection of the kidney. Observations are made on various chemical indicators which hopefully provide useful guidelines to renal function at any time. The problem we are concerned with is to analyse such data sequentially in order to detect any changes as soon as possible and assess their significance.

The state of renal function is generally gauged by clinicians via an unobservable factor termed the Glomerular Filtration Rate (GFR), which measures the rate of clearance of various substances through the kidneys. In order to estimate GFR, the blood and urine concentrations of several chemicals are measured and related to kidney function. One of the most important of such indicators is the chemical serum creatinine which is easily measured in the blood. Our study centres on creatinine although others, notably plasma concentrates of urea and a chemical called β2 microglobulin, are also of interest and are analysed similarly.

Under normal renal function GFR is constant and creatinine is excreted at a constant rate. A fall off in GFR is indicated by an increased blood concentration of creatinine and so it is this event that we are concerned with detecting in connection with possible rejection episodes.

In recent years, investigations of the behaviour of plasma creatinine with changing GFR have argued that reciprocal plasma creatinine is approximately linear with time. Knapp et al (1977) discuss this and Trimble (1980) examines this in depth, also considering an alternativee log transformation. The reciprocal transform is used by Smith and Cook (1980) in an attempt to identify change points in creatinine series using Bayesian regression techniques. This successful study treats reciprocal creatinine readings as independent observations from a straight line regression model against time. The technique of fitting piecewise linear functions of time to the data provides inferences about the change point corresponding to the time of onset of a rejection episode. Clearly the analysis is retrospective whereas our study requires a sequential approach and a recognition of the time series nature of the data.

In addition to the reciprocal transform, a time dependent correction is made to the creatinine data in order to take into account fluctuations in the level of body-water of the patient. An increase in body-water dilutes the plasma concentration of creatinine and this occurs in particular at the onset of rejection, just the time when we want to detect an increase. To correct this distortion, a routine adjustment involving body weight is made to the raw creatinine readings. Further discussion of this appears in Knapp et al (1977) and Trimble (1980) where the dramatic improvement in the linearity of reciprocal creatinine with time when this correction is made is illustrated.

Further features of the data.

(i) The observation on plasma creatinine are taken notionally at 8 hourly intervals over what is usually a period of several weeks in the case of (ultimately) successful transplants. The general pattern of behaviour is that an early period of poor renal function is followed by a gradual improvement and then a more erratic period which (hopefully) settles down as the organ is accepted and reaches its steady functioning level. The data on the graphs in §7.4.1 illustrate this.

(ii) Dialysis treatment is often provided in the early stages of post operative care to support the transplanted kidney. This has the effect of a short term improvement in renal function evidenced by a sudden decrease in the level of plasma creatinine followed by a slower decrease to the original level.

(iii) As the level of creatinine increases the data becomes more noisy. This behaviour will tend to obscure subtle changes in function if not properly modelled.

(iv) The observations are not always equally spaced; there are missing values and, sometimes, extra readings within an eight hourly period.

(v) The data contains outliers due to laboratory measurement and recording techniques and transfer to computer storage.

(vi) A different set of data, on urine measurements which we analyse similarly, exhibits "seasonal" variation due to diurnal body rythms.

Clearly any successful analysis must observe these features and provide a means of explaining them in the model adopted. §7.3 discusses the model.

7.2.2 The System Model.

Steady Renal function.

We now investigate the use of the approximately linear evolution with time of the reciprocal, body-water corrected serum creatinine.

Let ϕ_{0t} be the (unobservable) creatinine level at time t and ϕ_t the body-water corrected value,

$$\phi_t = W_t \cdot \phi_{0t} \qquad (7.2.1)$$

where W_t is a known factor depending upon body weight. We suppose that, for a time period $[S,T]$ of steady renal function,

$$\phi_t^{-1} = \mu + \beta t, \quad t=S, S+1,\ldots,T. \qquad (7.2.2)$$

Now we measure a value X_{0t}, say, of ϕ_{0t} subject to error. The following two main sources of error are apparent.

(i) Analytical Error.

As noted in §7.2.1, the variation in the data increases with the level. This would usually suggest a log transformation to achieve constant variance but to do this would mean losing the linear structure (7.2.2). Instead we introduce an additive error a_t whose distribution depends upon ϕ_{0t}.

Specifically, we take, approximately,

$$(a_t|\phi_{0t}) \sim N[0,\ c^2\phi_{0t}^2] \tag{7.2.3}$$

for some constant c. The quality control laboratory at the City Hospital in Nottingham were able to supply an approximate value for the coefficient of variation c for our data, of about 10%.

(ii) Reporting Error.

The data is measured in units of μmol/l (micro moles per litre) and is quoted to the nearest 10 of such units. Thus each reading is automatically subject to an additive, zero mean, symmetric error u_t say, with range $[-5,5]$.

Incorporating those two errors into a measurement model we have

$$X_{0t} = \phi_{0t} + a_t + u_t$$

or, if X_t denotes the body-water corrected observation

$$X_t = W_t\, X_{0t} = \phi_t + W_t(a_t+u_t) \tag{7.2.4}$$

Define $\theta_t = \phi_t^{-1}$ and the datum $y_t = X_t^{-1}$. Then, from (7.2.4) our series is given by

$$y_t = \theta_t.(1+S_t)^{-1} \tag{7.2.5}$$

where

$$S_t = \theta_t W_t (a_t+u_t) = \phi_{0t}^{-1}.(a_t+u_t) \qquad (7.2.6)$$

We now approximate (7.2.5) in the following way. If $|S_t|<<1$, then $(1+S_t) \simeq 1-S_t$, and so, from (7.2.5)

$$y_t \simeq \theta_t(1-S_t). \qquad (7.2.7)$$

This assumption can indeed be seen to be reasonable for our data and is the subject of Appendix 7.A(a).

At this stage we consider one final source of error.

(iii) Timing Error.

Observations are nominally timed at units of eight hours apart. Allowing for a timing error of at most 30 minutes of the nominal time gives us a third source of variation in the form of a zero-mean symmetric error r_t say, and in terms of the timing units of 8 hours, $|r_t| \leq 1/16$.

In the light of this, we correct (7.2.7) by adjusting $\theta_t = \mu+\beta t$ by a factor βr_t, leading to

$$y_t \simeq \theta_t + \nu_t \qquad (7.2.8)$$

where

$$\nu_t = -S_t\theta_t + (1-S_t)\beta r_t \qquad (7.2.9)$$

In Appendix 7.A(b) we discuss the error ν_t, concluding with the approximation

$$(\nu_t|\theta_t) \sim N[0, \theta_t^2.c^2] \qquad (7.2.10)$$

which is adequate for nearly all t.

7.3 Multi-State Modelling.

7.3.1 The Linear Growth Model.

The linear evolution of θ_t (7.2.2), and the structure (7.2.8)

fit neatly into the linear growth model of HS(1976), given by

$$y_t = \theta_t + \nu_t, \tag{7.2.11}$$

$$\theta_t = \theta_{t-1} + \beta_t + \gamma_t, \tag{7.2.12}$$

$$\beta_t = \beta_{t-1} + \delta_t, \tag{7.2.13}$$

where γ_t, δ_t are additional zero-mean, independent normal errors.

[A seasonal extension of this model was also developed for the urine data mentioned in §7.2.1(vi); details follow HS (1971) and (1976).]

To model changes define the states M_{tj}, j=1,...,4 at time t to be (1) steady state, (2) change in level, θ_t, (3) change in the trend β_t, (4) outlier, respectively. These are all features noted in §7.2.1 and are constructed by setting, in state M_{tj},

$$\mathrm{Var}[\nu_t | M_{tj}, \theta_t] = c^2\theta_t^2 . R_{\nu j}$$

$$\mathrm{Var}[\gamma_t | M_{tj}] = c^2 R_{\gamma j},$$

$$\mathrm{Var}[\delta_t | M_{tj}] = c^2 R_{\delta j},$$

where the multipliers $R_{\cdot j}$ are given in the table below.

	j	$R_{\nu j}$	$R_{\gamma j}$	$R_{\delta j}$
M_{t1} = Steady State	1	1	0	0
M_{t2} = Level Change	2	1	90	0
M_{t3} = Trend Change	3	1	0	60
M_{t4} = Outlier	4	100	0	0

The following prior information is also input:

(i) Prior probabilities of states M_{tj},

$$\underset{\sim}{p0}^T = (p0_1,...,p0_4) \text{ where, for } j=1,...,4,$$

$$p0_j = \mathrm{Prob}[M_{tj}], \text{ for all } t.$$

The values chosen for our data are

$$(0.85,\ 0.06,\ 0.07,\ 0.02),$$

admitting a priori a relatively high incidence of changes in level and trend with comparatively fewer outliers.

(ii) Prior distribution for the state vector $\underset{\sim}{X}_t^T = (\theta_t, \beta_t)$ at $t = 0$, taken to be normal,

$$\underset{\sim}{X}_0 \sim N[\underset{\sim}{m}_0,\ c^2 C_0]$$

where we originally took

$$\underset{\sim}{m}_0^T = (0,\ 0), \quad C_0 = \text{diag}(100,100).$$

This prior distribution essentially represents vague prior knowledge of (θ_0, β_0). In fact the clinicians are generally fairly confident that the level of creatinine at time zero is about 1000, corresponding to $\theta_0 \simeq 0.001$, and that initial growth is negligible, $\beta_0 \simeq 0$. Thus a more realistic prior is adopted, with these values defining $\underset{\sim}{X}_0$, and C_0 given by diag (0.01, 0.01). Clearly, considering the range of values for θ_t, this still represents a fairly diffuse prior distribution.

Now if we know $\text{Var}[\nu_t | M_{tj}]$ for all t, j, then the prior to posterior analysis outlined by Harrison and Stevens obtains and this has the following basic features.

At time t,

(a) the conditional posterior for $\underset{\sim}{X}_t$ is given by a "collapsed" density

$$(\underset{\sim}{X}_t | D_t,\ M_{tj}) \sim N[\underset{\sim}{m}_{tj},\ C_{tj}.\ c^2]; \qquad (7.2.14)$$

(b) the posterior probabilities for M_{tj} are calculated,

$$p_{tj} = \text{Prob}[M_{tj} | D_t]; \qquad (7.2.15)$$

(c) the unconditional posterior for $\underset{\sim}{X}_t$ is

$$p(\underset{\sim}{X}_t|D_t) = \sum_{j=1}^{4} p_{tj}.p(\underset{\sim}{X}_t|D_t, M_{tj}). \qquad (7.2.16)$$

However, $\mathrm{Var}[\nu_t|M_{tj},\theta_t] \simeq c^2\theta_t^2 R_{\nu j}$, for all t,j. To use the normal analysis, we replace this with the assumption

$$\mathrm{Var}[\nu_t|M_{tj}] \simeq \sigma_{tj}^2.R_{\nu j}, \qquad (7.2.17)$$

where

$$\sigma_{tj}^2 = c^2.E[\theta_t|D_{t-1}, M_{tj}]^2 \qquad (7.2.18)$$

for all t,j

using the expected level of the series rather than the true level.

Probabilities of States.

The probabilities given in (a) above are the interesting elements from the point of view of assessing renal function. If, for example, these weights favour M_{t2} then we suspect a change in level at time t as occurs after dialysis treatment. More importantly an abrupt change in trend from a positive to negative direction implies an increase in plasma creatinine consistent with a deterioration in renal function and possible onset of rejection. So in the first instance we look for high values of p_{t3}, together with a shift of $p(\beta_t|D_t)$ to favouring negative growth. Of course we should examine a plot of $p(\beta_t|D_t)$ in order to determine the directions of trend but a useful guideline will be a negative value of $E[\beta_t|D_t]$ or a negative value of $E[\beta_t|D_{t+1}]$.

On the receipt of an observation y_t that is not consistent with M_{t1} it is impossible to distinguish between possible changes immediately. A further observation y_{t+1} will in general clarify the issue and our procedure is to treat the weights p_{tj} as providing only a general indication of the state of the system and taking a low value of p_{t1} to mean some form of instability has arisen at time t. We then

calculate

$$q_{tj} = \text{Prob}\left[M_{tj} \mid D_{t+1}\right], \quad j=1,\ldots,4,$$

after observing y_{t+1}. These "one-step back" or "smoothed" probabilities give us a more concrete basis for assessing changes at time t and are calculated already during the collapsing procedure at time t+1. Although we generally base our inferences on the q_{tj}, it is sometimes not clear even one step on what has happened and a further confirmatory observation must be taken. For example a level change at time t may be obscured by an outlier at t+1. Thus we calculate

$$r_{tj} = \text{Prob}\left[M_{tj} \mid D_{t+2}\right], \quad j=1,\ldots,4$$

which, although generally of little use, provide a confirmation of changes in some, albeit infrequent, events.

7.3.2 Unknown Variation.

Early implementation of HS.

The analysis above used a given value of the coefficient of variation c appearing in the variance structure of the model. Operating like this was moderately successful in the early stages of the study although sensitivity to chosen values gave problems requiring a period of experience with the system to tune to optimal values.

In the light of this we adopted HS's approach involving a discrete "grid" of values for c and, viewing c as a random variable, the assignment of a prior distribution over this grid. This leads to a tractible analysis within the multi-state framework although now, of course, we have a 4N state model where N is the size of the grid. So at each stage a 4N component model "explodes" to a $(4N)^2$ state model which must then be collapsed back to 4N states.

It is just this large dimensionality that leads Stoodley and Mirnia (1979) to complain about the demand that this approach makes on computer time and space. We agree with this entirely, a less cumbersome and more elegant approach is required. Furthermore, the choice of a grid is somewhat arbitrary and the resulting mixture "likelihood" suffers from a lack of identifiability in that some variance levels may be reproduced within the mixture.

An alternative approach.

In the light of the above comments we adopt the following approach which allows for a more realistic modelling of the distribution of unknown variance parameters and also fits nicely into the multi-state framework. (The details are essentially a special case of the investigations of §4.3).

Define $\lambda = c^{-2}$. Then at time t, we have a conditional likelihood given by

$$(y_t | \underset{\sim}{X}_t, \lambda, M_{tj}, M_{t-1,i}, D_{t-1}) \sim N\left[\underset{\sim}{h}^T \underset{\sim}{X}_t, \lambda^{-1} \sigma^2_{tj}\right],$$

where $\underset{\sim}{h}^T = (1, 0)$ and σ^2_{tj} is given in (7.2.18).

Suppose that the prior for $\underset{\sim}{X}_t$, λ given $M_{t-1,i}$, D_{t-1} is of the conjugate normal/gamma form

$$(\underset{\sim}{X}_t | \lambda, M_{t-1i}, D_{t-1}) \sim N\left[\underset{\sim}{a}_{ti}, \lambda^{-1} P_{ti}\right],$$

$$(\lambda | M_{t-1i}, D_{t-1}) \sim G\left[\frac{n_{t-1}}{2}, \frac{b_{t-1i}}{2}\right].$$

Then, given λ, the usual analysis leads to a normal posterior for X_t given λ, M_{tj}, M_{t-1i}, D_t. For λ we have

$$(\lambda | M_{tj}, M_{t-1i}, D_t) \sim G\left[n_t/2, B_{tij}/2\right]$$

where

$$n_t = n_{t-1} + 1,$$

$$B_{tij} = b_{t-1i} + (y_t - \underset{\sim}{h}^T \underset{\sim}{a}_{ti})^2/(\underset{\sim}{h}^T P_{ti} \underset{\sim}{h} + \sigma^2_{tj}), \quad \text{for all } i,j.$$

So

$$p(\underset{\sim}{X}_t, \lambda | M_{tj}, D_t) = \sum_{i=1}^{4} p(\underset{\sim}{X}_t, \lambda | M_{tj}, M_{t-1i}, D_t) . p(M_{t-1i} | M_{tj}, D_t)$$

Now the conditional posterior for $\underset{\sim}{X}_t | \lambda, M_{tj}, D_t$ can be collapsed as usual to a single normal with the same moments. Similarly $p(\lambda | M_{tj}, D_t)$ is a mixture of gammas,

$$\sum_{i=1}^{4} G[n_t/2, B_{tij}/2] . p(M_{t-1i} | M_{tj}, D_t)$$

which we collapse to the nearest gamma density as defined by the minimum Kullback-Leibler divergence approximation of Appendix 4B, $G[n_t/2, b_{tj}/2]$ where

$$b_{tj}^{-1} = \sum_{i=1}^{k} B_{tij}^{-1} . p(M_{t-1i} | M_{tj}, D_t).$$

Finally then we have the (collapsed) joint normal/gamma mixture for $p(\underset{\sim}{X}_t, \lambda | D_t)$ given by

$$\sum_{j=1}^{k} p_{tj} . NG[\underset{\sim}{X}_t, \lambda | \underset{\sim}{M}_{tj}, C_{tj}, n_t, b_{tj}]$$

where $NG[\underset{\sim}{X}, \lambda | \underset{\sim}{M}, C, n, b]$ denotes the density of

$$(\underset{\sim}{X} | \lambda) \sim N[\underset{\sim}{M}, \lambda^{-1} C],$$

and

$$\lambda \sim G[n/2, b/2].$$

All that remains is to calculate $p(M_{t-1i} | M_{tj}, D_t)$ and $p_{tj} = p(M_{tj} | D_t)$, which are obtained simply as follows.

Clearly $p(y_t | M_{tj}, M_{t-1i}, D_{t-1})$ is a Student t density, proportional to

$$(\sigma^2_{tj} + \underset{\sim}{h}^T P_{ti} \underset{\sim}{h})^{-\frac{1}{2}} \quad B_{tij}^{-n_t/2}$$

Now $p(M_{t-1i}|M_{tj}, D_t) \propto p(y_t|M_{tj}, M_{t-1i}, D_{t-1}) \cdot p_{t-1i}$

and $p(M_{tj}|D_t) \propto p(y_t|M_{tj}, D_{t-1})\ p_{0j}$,

where $p(y_t|M_{tj}, D_{t-1}) = \sum_{i=1}^{4} p(y_t|M_{tj}, M_{t-1i}, D_{t-1})\ p(M_{t-1i}|M_{tj}, D_t)$.

With this method of learning the unknown coefficient of variation in the creatinine, and other, series, we found that we obtained a much more robust analysis together with increased sensitivity in the detection and estimation of changes. In the next section we present some examples of graphical output from same analyses.

7.4. Implementation.

7.4.1. Graphical Output.

Figures 7.4 a) - d) illustrate the output of a computer program written for the analysis of creatinine, and other time series. Figs a) and b) have the following common features.

(i) The upper graph displays the corrected creatinine plotted on an inverted reciprocal scale. Thus a change of trend from negative to positive on this graph indicates a deterioration in renal function.

(ii) The second graph is of posterior probabilities of <u>some form</u> of instability at the current time, as measured by $1-p_{t1}$.

(iii) The third graph is of the "one-step back" probability of a <u>positive</u> trend change. Thus we display q_{t3} when $E[\beta_t|D_{t+1}] < 0$.

(iv) The lower graph is of the "two-step back" probabilities, r_{tj} subject, again, to a negative value of $E[\beta_t|D_{t+2}]$.

(v) In addition the circles plotted on the lower two graphs are of the (appropriately lagged) probabilities of a change in

level, providing a little extra information about the state of the system.

Figure 7.4 a) is a typical series, whilst 7.4 b) is somewhat special being extremely quiet and serving only to illustrate the detection of level changes due to dialysis treatment (indicated by HD at the top of the upper graph), and the detection of the onset of Δ rejection at day 45.

Figs. c) and d) have the same basic features. Fig. c) is concerned with urea rather than creatinine, urea being the other important indicator we studied. Fig. d) displays the output of the analysis of both creatinine and urea on the same graphs; viewing the two together is often helpful for the clinician.

7.4.2. Inferences and Decisions.

In order to make inferences about the state of the kidney and to make recommendations that the clinicians take the relevant action, we have to be guided by past experience with the data in interpreting the probabilities of changes. At this stage in the study no formal decision theoretic analysis and explication of clinicians utilities has been made and we resort to the following simple procedure.

Each of the creatinine series (about 30 to date, with average length 65 days) were analysed. For a set of "levels", $0 = \alpha_1 < \alpha_2 < \ldots < \alpha_m = 1$, we recorded the number and timing of the occurrences of posterior probabilities p_{t3} (concerning possible rejections) that exceeded the chosen levels. The set of such at a level α will be called the set of "positives at level α", $P(\alpha)$. [A similar procedure was followed using the "one step back" probabilities q_{t3} leading to a set of positives $Q(\alpha)$, say.]

Following this, the data was given to a group of consultant clinicians who, armed with patients records, results of various physiological tests, and the benefit of professional hindsight, classified all observations as either non-rejection or "probable or definite rejection". Further medical details can be found in the thesis of Trimble (1980). The latter group of events will be called the "rejections", and number R, the former are called the "non-rejections", and number N.

Thus at a level α, the proportion of "true positives" is given by $T(\alpha)$, where

$$R.T(\alpha) = \text{Number of "rejections" in } P(\alpha).$$

Further the proportion of "false positives" is given by $F(\alpha)$, where

$$N.F(\alpha) = \text{Number of "non-rejections" in } P(\alpha).$$

Now suppose that we adopt the procedure of interpreting a probability p_{t3} (or q_{t3}) as indicating a possible rejection if it exceeds a cutoff value α_0, say. Our problem is to choose α_0 in some rational way. Suppose the clinician deems it K times more important that he detects a rejection than that he receives a false alarm. Then we model his utility function as

$$\begin{cases} K, & \text{if he receives a vindicated warning,} \\ -1, & \text{if the warning is false,} \end{cases}$$

and the expected gain at a cutoff level α is

$$G(\alpha) = K \text{ Prob}\left[\text{rejection} \,\middle|\, \text{positive at level } \alpha\right] - \text{Prob}\left[\text{non-rejection} \,\middle|\, \text{positive at level } \alpha\right].$$

On the basis of our data, we estimate this by

$$G(\alpha) = K.T(\alpha) - F(\alpha).$$

So our problem is simply to maximize $G(\alpha)$ as a function of α_0.

Assuming $G(\alpha)$ is differentiable, then the optimal level α_0 satisfies

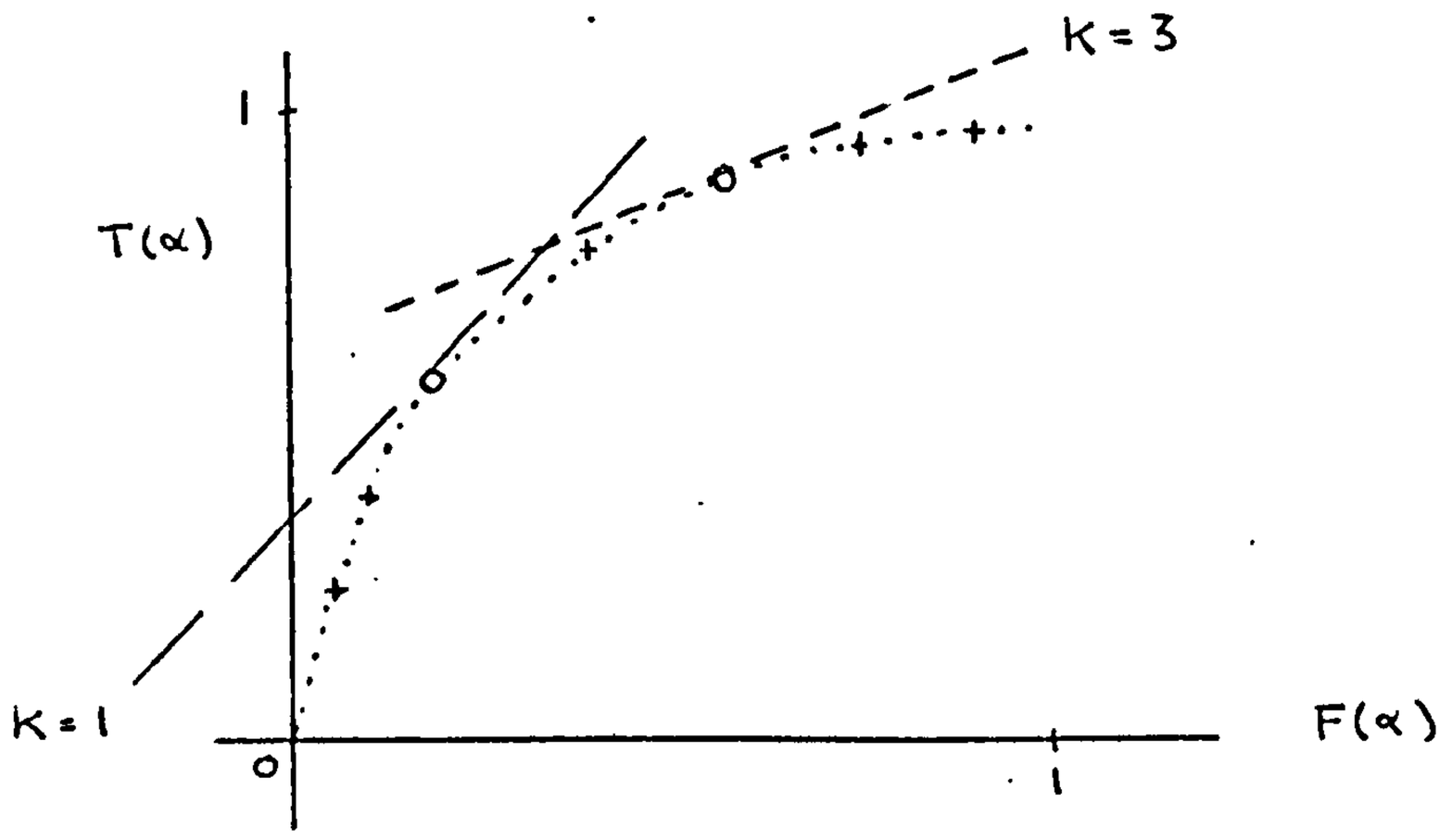
$$\frac{K.\partial T(\alpha)}{\partial \alpha} - \frac{\partial F(\alpha)}{\partial \alpha} = 0$$

or, since $\frac{\partial T(\alpha)}{\partial \alpha} = \frac{\partial T(\alpha)}{\partial F(\alpha)} \cdot \frac{\partial F(\alpha)}{\partial \alpha}$ and assuming $\frac{\partial F(\alpha)}{\partial \alpha} \neq 0$, we have

$$\frac{\partial T}{\partial F} = K^{-1}, \quad \text{at } \alpha = \alpha_0$$

So a plot of the pairs $\left(F(\alpha_i), T(\alpha_i)\right)$ for our set of levels α_i, i=1,2,...,M will give us an approximation to T as a function of F and we can then interpolate to find α_0.

Graphically



The procedure is intuitively reasonable: we move α down the curve from the upper right and hope to decrease F faster than T.

In particular for creatinine, the cutoff value obtained for "on-line" probabilities p_{t_3} was about 0.15 when K=3. This corresponds to a posterior/prior odds ratio of about 2.4. With this cutoff level, 31 out of 32 "rejections" were detected, and, in the one missed episode, clinical diagnosis was made on the basis of several extra observations made during one day, which information was not used in the analysis. For the "one-step back" probabilities, the posterior/

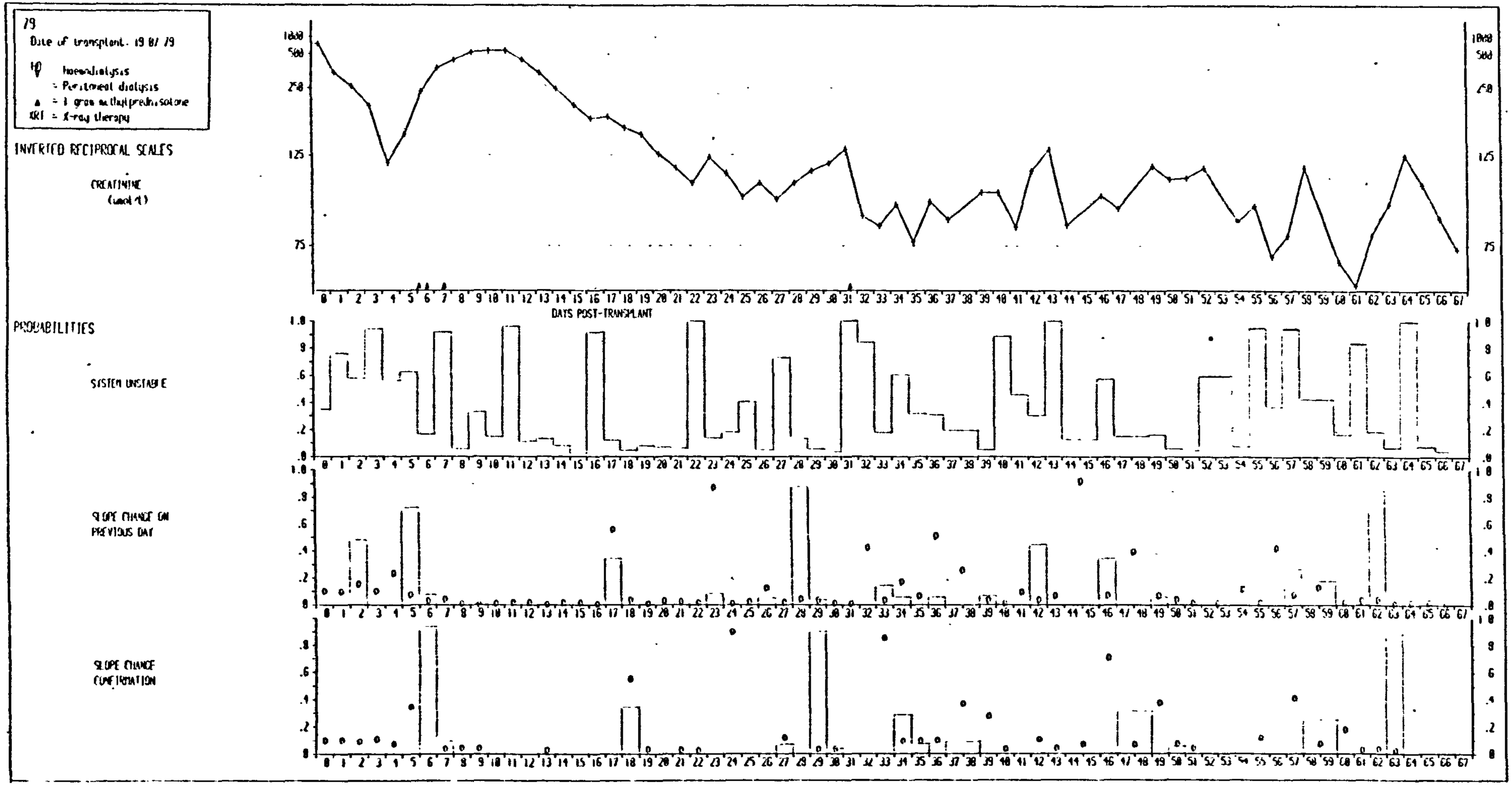
prior odds ratio was about 3.5, indicating the improvement in the expected performance due to using a further observation. So the system appears capable of detecting rejections and, with this procedure, the number of false alarms remains acceptable to the clinician. Furthermore, for those rejection episodes that were treated by the doctors at the time, the computerised monitoring system detected a possible rejection (i.e. p_{t3} reached cutoff level) at the time, and, on average, treatment followed 1.5 observations later. In the light of this we feel that such a system offers much in the way of an aid to clinicians in detecting and diagnosing transplant rejections.

7.5. Conclusions.

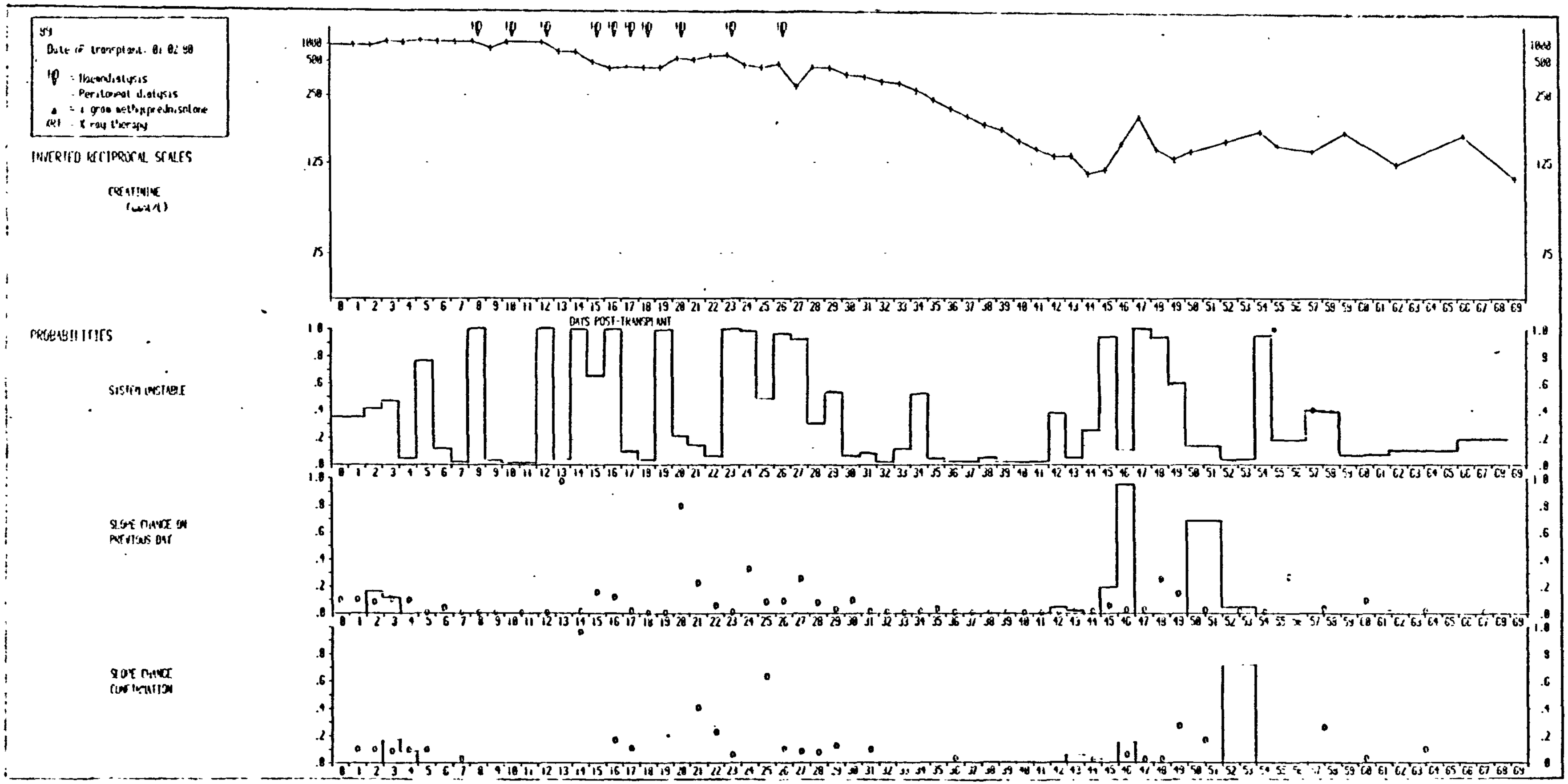
The early results of our study of kidney transplant data reported above illustrate the flexibility and practical utility of multi-state modelling in a non-forecasting problem. Clearly there remains much to be done in this area, notably analyses of other indicators and then the construction and implementation of multi-variate models in order to tie together indicators.

It is quite clear from this study that successful implementation of such an approach will not be achieved without a thorough investigaton of the physical system being analysed in order to move towards a model which both captures some of the structure of the data, and can be adapted to a reasonably simple mathematical form. Furthermore, a detailed study of error characteristics seems desirable since, if neglected, the sensitivity of analysis will be considerably impaired. Finally, a coherent basis for learning about unknown variation as detailed in §7.4 is also valuable in improving performance and in adding to our knowledge about the time series being studied.

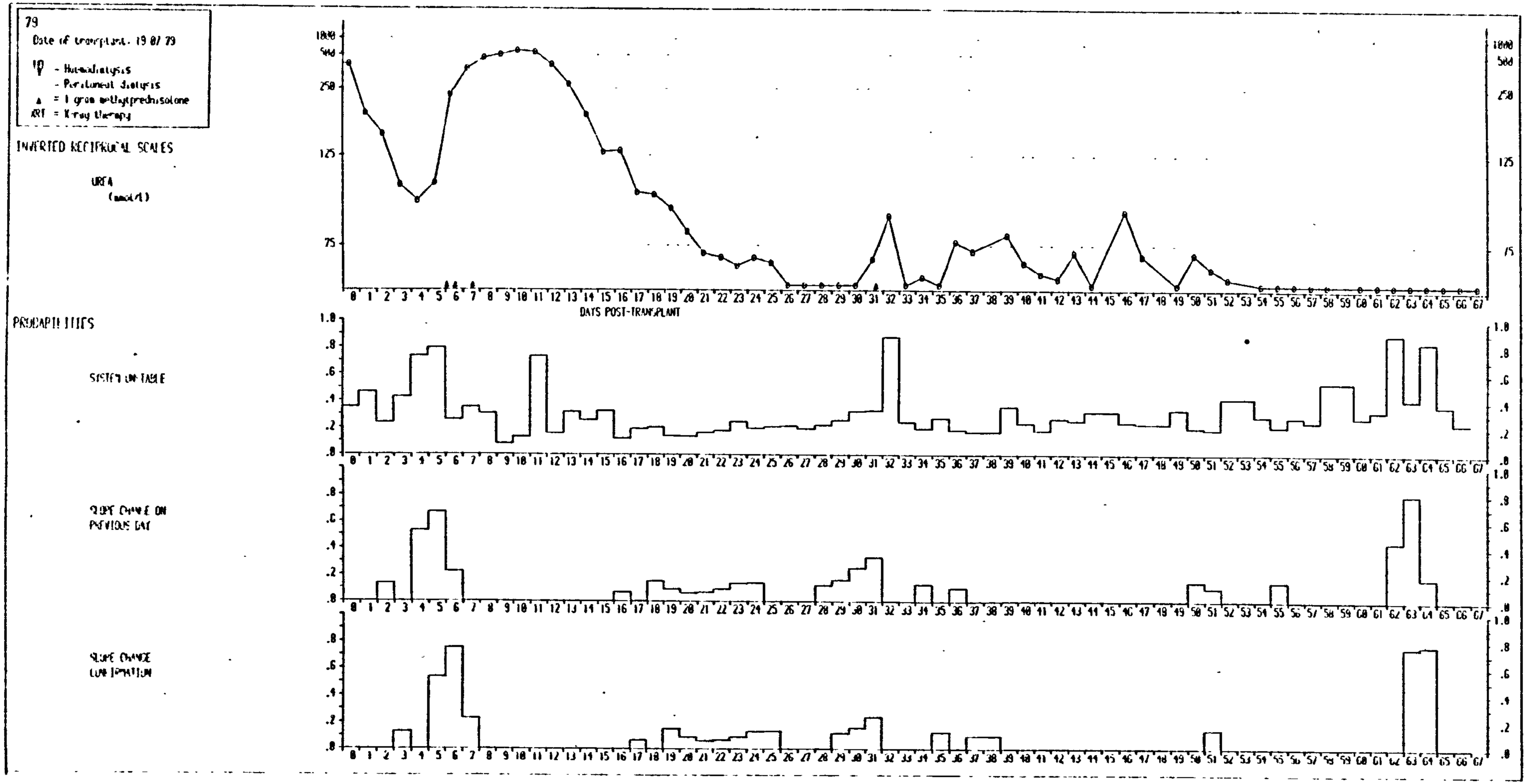
7.4(a)


79
Date of transplant. 19 07 79
HD Haemodialysis
= Peritoneal dialysis
▲ = 1 gram methylprednisolone
XRT = X-ray therapy
INVERTED RECIPROCAL SCALES
CREATININE
(umol/l)
1000
500
250
125
75
DAYS POST-TRANSPLANT
PROBABILITIES
SYSTEM UNSTABLE
SLOPE CHANGE ON PREVIOUS DAY
SLOPE CHANGE CONFIRMATION

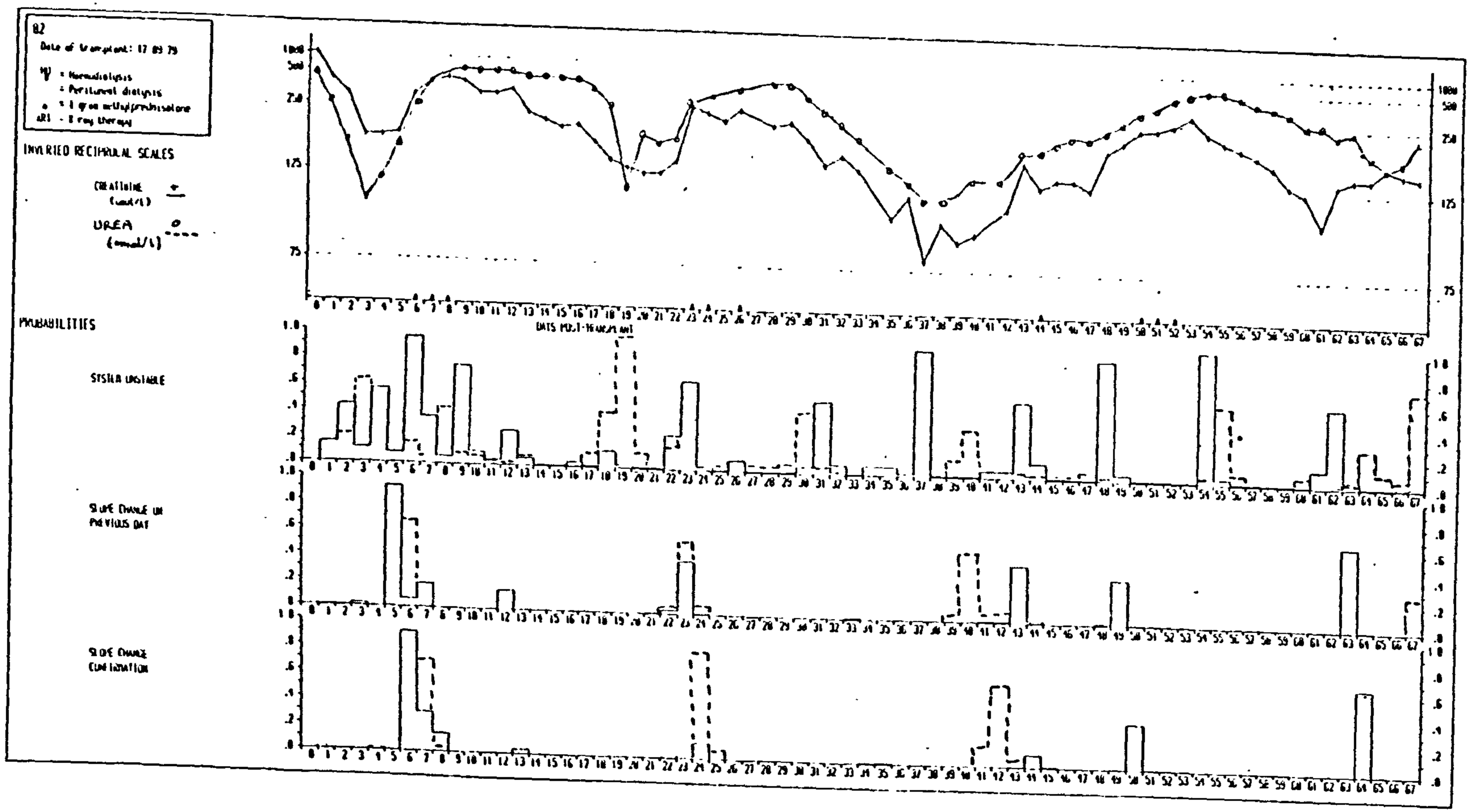
7.4(b)


SM
Date of transplant: 01 02 90
HD = Haemodialysis
PD = Peritoneal dialysis
a = 1 gram methylprednisolone
XRT = X ray therapy
INVERTED RECIPROCAL SCALES
CREATININE
(umol/l)
1000
500
250
125
75
DAYS POST-TRANSPLANT
PROBABILITIES
STATE UNSTABLE
SLOPE CHANGE ON PREVIOUS DAY
SLOPE CHANGE CONFIRMATION
1.0
.8
.6
.4
.2
0

7.4(c)


79
Date of transplant. 19 07 79
- Haemodialysis
- Peritoneal dialysis
▲ = 1 gram methylprednisolone
XRT = X-ray therapy
INVERTED RECIPROCAL SCALES
UREA
(mmol/l)
1000
500
250
125
75
DAYS POST-TRANSPLANT
PROBABILITIES
SIGN ON TABLE
SLOPE CHANGE ON PREVIOUS DAY
SLOPE CHANGE CONFIRMATION
1.0
.8
.6
.4
.2
.0

7.4(d)



82
Date of transplant: 17 09 79
= Haemodialysis
= Peritoneal dialysis
= 1 gram methylprednisolone
= X ray therapy
INVERTED RECIPROCAL SCALES
CREATININE
UREA
PROBABILITIES
SYSTEM UNSTABLE
SLOPE CHANGE ON PREVIOUS DAY
SLOPE CHANGE CONFIRMATION
DAYS POST-TRANSPLANT

Acknowledgements.

It is convenient at this point to acknowledge the contributions of the personnel of the Renal unit at the City Hospital in Nottingham. In particular, thanks are due to Dr. M.S. Knapp, who initiated the study, and to Mr. I. Trimble of the University Medical School who provided the biological and medical background to the problem and whose help in formulating the models of this study was invaluable.

Appendix 7.A.

(a) $$s_t = \phi_{0t}^{-1}(a_t+u_t)$$

where $a_t \sim N[0, c^2\phi_{0t}^2]$, $c \simeq 0.1$,

and $|u_t| < 5$.

For serum creatinine we have that, in general,

$$10 < \phi_{0t} < 1000.$$

Indeed almost all the values exceed 100.

Thus $\phi_{0t}^{-1} < k$ where k is at most 0.1 and typically closer to 0.01, hence

$$|s_t| < |\phi_{0t}^{-1} a_t| + 5k.$$

But $\phi_{0t}^{-1} a_t \sim N[0,c^2]$ and so, with high probability,

$$|s_t| < 5(c+k).$$

Thus $|s_t| < 1$ with $|s_t| << 1$ for most t, and thus it is not unreasonable to suppose that

$$(1+s_t)^{-1} \simeq 1-s_t.$$

(b) $$\nu_t = -s_t\ \theta_t + (1-s_t)\ \beta r_t$$

where $|r_t| < 1/16$.

Now, given θ_t, ν_t is clearly symmetrically distributed about zero. Further, since $\phi_{0t}^{-1} = W_t\ \theta_t$ we have

$$\mathrm{Var}[\nu_t|\theta_t] = E[s_t^2\theta_t^2 + (1+s_t^2-2s_t)\beta^2 r_t^2 - 2s_t\theta_t(1-s_t)\beta r_t|\theta_t]$$

$$= \theta_t^4 W_t^2 E[u_t^2] + \theta_t^2 c^2 + \beta^2 E[r_t^2]\ \{1 + W_t^2\theta_t^2 E[u_t^2] + c^2\}.$$

Now θ_t is at most 0.1, typically $O(10^{-2})$ and β is of the same order. From §7.2.2 (iii) and assuming that the timing error r_t is approximately normally distributed we have $E[r_t^2] \simeq 4 \times 10^{-4}$. $E[u_t^2]$ is O(1) and $c^2 \simeq 10^{-2}$.

Taking these points into consideration, we see that

$$\mathrm{Var}[\nu_t|\theta_t] \simeq \theta_t^4\ W_t^2\ E[u_t^2] + \theta_t^2\ c^2$$

on ignoring terms smaller by $O(10^{-2})$. Now for θ_t at the upper end of its range, that is near 0.1, both terms remaining are important. However, for θ_t of order 10^{-2} and smaller the variance of ν_t is approximately given by $\theta_t^2\ c^2$. This approximation essentially assumes that the contributions of timing and reporting errors are negligible compared with the analytic error a_t, and an approximately normal distribution for ν_t follows.

REFERENCES

ABRAHAM, B. and BOX, G.E.P. (1979). Bayesian analysis of some outlier problems in time series. Biometrika, 66, 229-236.

ALSPACH, D.L. and SORENSON, H.W. (1971). Recursive Bayesian estimation using Gaussian Sums. Automatica, 7, 465-479.

ANDERSON, B.D.O. and MOORE, J.B. (1979). Optimal Filtering. (Information and System Sciences Series), Prentice-Hall: New Jersey.

ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H., TUKEY, J.W. (1972). Robust Estimates of Location. Princeton: University Press.

ANDREWS, D.F. and MALLOWS, C.L. (1974). Scale mixtures of normal distributions. J.R. Statist. Soc. B. 36, 99-102.

BARNDORFF-NEILSEN, O. (1978). Information and Exponential Families in Statistical Theory. Wiley.

BOX, G.E.P. (1979). Robustness in the strategy of scientific model building. in Robustness in Statistics. (eds: R.L. Launer and G.N. Wilkinson). Academic Press: New York.

BOX, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. J.R. Statistic. Soc. A, 143, 383-430.

BOX, G.E.P. and JENKINS, G.M. (1971). Time Series Analysis: Forecasting and Control. Holden-Day: San Francisco.

BOX, G.E.P. and TIAO, G.C. (1962). A further look at robustness via Bayes' Theorem. Biometrika 49, 419-432.

BOX, G.E.P. and TIAO, G.C. (1968). A Bayesian approach to some outlier problems. Biometrika, 55, 119-128.

BOX, G.E.P. and TIAO, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley: Reading, Mass.

CHIK L. SOKOL, R.J., ROSEN, M.G., PILLAY, S.K. and S.E. JARRELL.(1979). Trend analysis of interpartrum data: a basis for a computerized fetal monitor. Clinical Obstretrics and Gynecology, 22, 665-679.

CHU, K.C. (1973). Estimation and detection for linear systems with elliptical random variables. I.E.E.E. Trans.Aut. Con., 18, 499-505.

COOK, R.D. and WEISBERG S. (1980). Characterization of an empirical influence function for detecting influential cases in regression. Technometrics, 22, 495-508.

COX, D.R. and HINKLEY, D.V. (1974). Theoretical Statistics. London: Chapman & Hall. New York: Wiley.

DAWID, A.P. (1973). Posterior expectations for large observations. Biometrika, 60, 664-667.

DEGROOT, M.H. (1970). Optimal Statistical Decisions. McGraw-Hill: New York.

Du MOUCHEL, W.H. (1973). On the asymptotic normality of the maximum likelihood estimator when sampling from a stable distribution. Ann. Statist., 1, 948-957.

FABIAN, V. (1978). On asymptotically efficient recursive estimation. Ann. Statist., 6, 854-866.

FELLER, W. (1966). An Introduction to Probability Theory and its Applications. Wiley: New York.

FOX, A.J. (1972). Outliers in time series. J.R. Statist. Soc., 34, 350-363.

GARDNER, G., HARVEY, A.C. and PHILLIPS, G.D.A. (1980). An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. Applied Statistics, 29, 311-317.

GOODWIN, G.C. and PAYNE, R.L. (1977). Dynamic System Identification: Experiment Design and Data Analysis. Academic Press: New York.

HAMPEL, F. (1974). The influence function and its role in robust estimation. J. Amer. Statist. Ass., 69, 383-393.

HARRISON, P.J. and STEVENS, C.F. (1971). A Bayesian approach to short term forecasting. Operational Research Quarterly. 22, 341-362.

HARRISON, P.J. and STEVENS, C.F. (1976). Bayesian Forecasting (with discussion). J.R. Statist. Soc. B, 38, 205-247.

HARVEY, A.C. and PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. Biometrika, 66, 49-58.

HEYDE, C.C. and JOHNSTONE, I.M. (1979). On asymptotic posterior normality for stochastic processes. J.R. Statist. Soc., B, 184-189.

HILL, D.W. and ENDRESON, J. (1978). Trend recording and forecasting in intensive therapy. British Journal of Clinical Equipment January, 1978, 4-14.

HUBER, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Statist., 35, 73-101.

HUBER, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist., 1, 799-821.

HUBER, P.J. (1977). Robust Statistical Procedures. Philadelphia: Society for Industrial and Applied Mathematics.

IBRAGIMOV, I.A. and LINNIK, Yu.V. (1971). Independent and Stationary Sequences of Random Variables. (ed: J.F.C. Kingman). Wolters-Noordhoff, Groningen, The Netherlands.

KALMAN, R.E., (1963). New methods in Wiener filtering theory. in Proceedings of the 1st Symposium on Engineering Applications of Random Function Theory and Probability.

KASHYAP, R.L., BLAYDON, C.C. and FU, K.S. (1970). Stochastic approximation. in Adaptive, Learning and Pattern Recognition Systems: Theory and Applications. (eds: J.M. Mendel and K.S. Fu). Academic Press. New York and London.

KLEINER, B., MARTIN, R.D. and THOMPSON D.J. (1979). Robust estimation of power spectra (with discussion). J.R. Statist. Soc., B, 313-351.

KNAPP, M.S., BLAMEY, R., COVE-SMITH, R. and HEATH, N. (1977). Monitoring the function of renal transplants. The Lancet, 1183.

KULLBACK, S. (1959). Information Theory and Statistics. Wiley: New York.

LINDLEY, D.V. (1971). The estimation of many parameters. in Foundations of Statistical Inference. (eds: V.P. Godambe and D.A. Sprott). Holt, Rinehart and Winston of Canada Ltd. Toronto, Montreal.

LJUNG, L. (1978). On recursive prediction error identification algorithms. Research report LiTH-ISY-I-0226. Dept. of Electrical Engineering, Linkopking University, Sweden.

LUKACS, E. and LAHA, R.G. (1964). Applications of Characteristics Functions. Griffin's Statistical Monographs and courses.

MARTIN, R.D. (1979). Robust estimation of time-series autoregressions. in Robustness in Statistics (eds: R.L. Launer and G. Wilkinson). Academic Press: New York.

MARTIN, R.D. (1978). Robust estimation of autoregression models. in Directions in Time Series. (eds: D.R. Brillinger and G.C. Tiao).

MARTIN, R.D. and MASRELIEZ, C.J. (1975). Robust estimation via stochastic approximation. I.E.E.E. Trans. Inf. Theory, 21, 263-271.

MASRELIEZ, C.J. (1975). Approximate non-Gaussian filtering with linear state and observation relations. I.E.E.E. Trans. Aut. Con., 20, 107-110.

MASRELIEZ, C.J. and MARTIN, R.D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman Filter. I.E.E.E. Trans. Aut. Con., 22, 361-371.

MILLER, R.B. (1978). Comment on paper of R.D. Martin in Directions in Time Series. (eds: D.R. Brillinger and G.C. Tiao).

O'HAGAN, A. (1976). On posterior joint and marginal modes. Biometrika, 63, 329-333.

O'HAGAN, A. (1979). On outlier rejection phenomena in Bayes inference. J.R. Statist. Soc., B, 41, 358-367.

O'HAGAN, A. (1981). A moment of indecision. Biometrika, 68, 329-330.

OSAKI, K. (1974). Some generalizations of dynamic stochastic approximation processes. Ann. Statist., 2, 1042-1048.

POLJAK, B.T. and TSYPKIN, Ja. Z. (1979). Adaptive estimation algorithms: convergence, optimality, robustness. Automation and remote control, 3, 71-84.

POLJAK, B.T. and TSYPKIN, Ja. Z. (1980). Robust identification. Automatica, 16, 53-63.

PRIESTLEY, M.B. (1978). System identification, Kalman filtering, and stochastic control. in Directions in Time Series. (eds: D.R. Brillinger and G.C. Tiao).

RAMSAY, J.O. and NOVICK, M.R. (1980). PLU robust Bayesian decision theory: point estimation. Journal of the American Statistical Association, 75, 901-907.

RELLES, D.A. and ROGERS, W.H. (1977). Statisticians are fairly robust estimators of location. Journal of the American Statistical Association, 72, 107-111.

ROBBINS, H. and MUNRO, S. (1951). On a stochastic approximation method. Ann. Math. Statist. 24, 400-407.

RUPPERT, D. (1979). A new dynamic stochastic approximation procedure. Ann. Statist., 7, 1179-1195.

SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. Ann. Math. Statist., 29, 373-405.

SAKRISON, D.J. (1966). Efficient recursive estimation; application to estimating the parameters of a covariance function. Internat. J. Engrg. Sci., 3, 461-483.

SMITH, A.F.M. (1979. Change-point problems: approaches and applications. in Bayesian Statistics, Proceedings of the First International Meeting in Valencia, Spain. (eds: J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Snith) University Press-Valencia.

SMITH, A.F.M. and COOK, D.G. (1980). Straight lines with a change point: a Bayesian analysis of some renal transplant data. Applied Statistics, 29, 180-189.

SMITH, J.Q. (1979). A generalization of the Bayesian steady forecasting model. J.R. Statist. Soc. B, 41, 378-387.

STOODLEY, K.D.C. and MIRNIA, M. (1978). The automatic detection of transients, step changes and slope changes in the monitoring of medical time series. The Statistician, 28, 163-170.

TRIMBLE, I. (1980). Unpublished M. Phil. thesis. University of Nottingham Medical School.

VERE-JONES, D. (1975). On updating algorithms and inference for stochastic point processes. in Perspectives in Probability and Statistics (ed: J. Gani), 239-259. London: Academic Press.

WEST, M. (1981). Robust sequential approximate Bayesian estimation. J.R. Statist. Soc. B, 43, 157-166.