Bayesian Edge-Detection in Image Processing

by David A. Stephens

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy.

February 1990

PAGE NUMBERING AS IN THE ORIGINAL THESIS

To my family

Table of Contents

Abstract

Problems associated with the processing and statistical analysis of image data are the subject of much current interest, and many sophisticated techniques for extracting semantic content from degraded or corrupted images have been developed. However, such techniques often require considerable computational resources, and thus are, in certain applications, inappropriate. The detection localised discontinuities, or edges, in the image can be regarded as a pre-processing operation in relation to these sophisticated techniques which, if implemented efficiently and successfully, can provide a means for an exploratory analysis that is useful in two ways. First, such an analysis can be used to obtain quantitative information relating to the underlying structures from which the various regions in the image are derived about which we would generally be *a priori* ignorant. Secondly, in cases where the inference problem relates to discovery of the unknown location or dimensions of a particular region or object, or where we merely wish to infer the presence or absence of structures having a particular configuration, an accurate edge-detection analysis can circumvent the need for the subsequent sophisticated analysis. Relatively little interest has been focussed on the edge-detection problem within a statistical setting.

In this thesis, we formulate the edge-detection problem in a formal statistical framework, and develop a simple and easily implemented technique for the analysis of images derived from two-region single edge scenes. We extend this technique in three ways; first, to allow the analysis of more more complicated scenes, secondly, by incorporating spatial considerations, and thirdly, by considering images of various qualitative nature. We also study edge reconstruction and representation given the results obtained from the exploratory analysis, and a cognitive problem relating to the detection of objects modelled by members of a class of simple convex objects. Finally, we study in detail aspects of one of the sophisticated image analysis techniques, and the important general statistical applications of the theory on which it is founded.

Acknowledgements.

I would like to thank my supervisor Professor Adrian Smith for his support and encouragement throughout the period of my research, and the other former and current members of the Statistics Group at the University of Nottingham for their many collective professional and other contributions. I thank in particular Dr. Cliff Litton, Jon Wakefield, and Nick Polson for their advice and ideas.

This work was carried out whilst I was a full-time research student financed by a Science and Engineering Research Council grant, and the thesis itself was completed whilst I was employed as a Research Assistant at the University of Nottingham on a project funded by the SERC Complex Stochastic Systems Initiative, for both of which I am profoundly grateful.

Chapter 1 : Statistical Image Processing.

(1.1) Introduction.

The statistical approach to the solution of inference problems in science and engineering proceeds as follows. First, we construct a modelling framework in which the problem and any subsequent analysis may be formulated and interpreted. We then design and perform an informative experiment in order to gather data. If necessary, we then might make transformations of the data, or carry out an exploratory analysis to discover broad trends and investigate general structure. We would then finally proceed with a detailed analysis to complete the inferential process, and attempt to report a coherent and relevant solution to the problem, conditional on the data observed.

Now suppose that, given a suitable framework, the collated data takes the form of a set of observations spatially configured in at least two dimensions, relating to the physical or measurable attributes of a collection of subsets or regions in again at least two dimensions (either identically or in projection), with the relationship being regarded as stochastic rather than deterministic. Suppose that we have interest in making inferences about these (unobservable) attributes and their spatial inter-relation. Then the corresponding exploratory and detailed inference problems are referred to as statistical **image processing** and **image analysis.**

In this introductory chapter, we give a brief indication of the history of the development of image processing techniques and note several important and influential references, and list a selection of some of the most important fields of application. We also attempt to provide a motivation for the use of specifically statistical techniques discussed and developed in this thesis. Later in this chapter, we present a glossary of important terms, and discuss what we regard to be the fundamental problems in image processing. Finally, we set out the aims and intentions of the work presented in this thesis.

(1.1.1) History and Applications.

The problems associated with the collection and processing of image or signal data are familiar in scientific and engineering circles, and research extending over the last 30 years, in conjunction with tremendous advances in computer and other related technology, has given rise to an extensive literature. It is only relatively recently, however, that these problems have been embraced and addressed by the statistical community, whereas previously the majority of frontier work had been carried out in the research departments of electrical and electronic engineering, and computer science, in both industrial companies and academic institutions.

Some idea of the how the subject has developed over this period can be gained with reference to several periodical publications. Journals associated with the Institute of Electrical and Electronic Engineers (I.E.E.E.) have been and are popular media for the presentation of both statistical and non-statistical work; see, in particular, *I.E.E.E. Proceedings*, and *Transactions* on *Information Theory*, on *Acoustics, Sound, and Signal Processing*, and latterly on *Pattern Analysis and Machine Intelligence*. Other useful specialist journals include *Pattern Recognition* and *Computer Vision, Graphics, and Image Processing*. It is from such sources that the majority of our background references will be drawn, and jointly serve as an orientation for the work that we shall present. An important and comprehensive introductory text with an emphasis on aspects of "classical" image processing and analysis techniques is that of Rosenfeld and Kak (1982). See also Andrews and Hunt (1977), Pratt (1978), and Schowengerdt (1986).

Prior to 1970, little concerted effort had been applied to the problem in a specifically statistical framework. Problems such as classification and discrimination that, as we shall see later, are closely related to the image analysis problem, had been studied extensively, but with the relevance being consequential rather that motivative. Early works in which statistical analysis was fully considered are Fukunaga (1972) and Duda and Hart (1973). These again are excellent introductory texts. An important influence on the development of statistical image analysis was Besag (see, for example, Besag (1974,1975,1977,1978)) who pioneered work on spatial probability structures and statistics, although at that time the link was still largely incidental. In last ten years, the growth of interest in the subject has been rapid, and many important and interesting papers have appeared. We note two in particular - Geman and Geman (1984) and Besag (1986) - which have motivated much subsequent research. We shall see and discuss further the particular relevance of each of these papers in the remainder of this chapter, and in several subsequent chapters. Note also Ripley (1988) as an important reference regarding spatial data analysis.

The works listed above, and the extensive references that they contain, represent a comprehensive bibliography of published work relating to image processing and analysis. A recent addition to the literature is a Special Issue of the *Journal of Applied Statistics*, which contains a useful introductory paper by Dubes and Jain (1989), and an overview of current research.

Many practical applications of image processing and analysis exist. Without quoting specific examples or providing details, the most important of these applications relate to agronomy (inference about land-use from satellite images), astronomy (studying the motion of galaxies), industrial processing (automated manufacturing and quality control), medicine (internal body imaging), and the military (intelligence, reconnaissance, defence/offence systems), relating variously to the imaging techniques of, for example, photography, tomography,

radiography etc. Although aspects of the problem differ in each of these examples, a single general framework and terminology will suffice for all of them.

(1.2) Terminology and notation.

As indicated above, before we can perform any statistical image processing or analysis, we must first construct a framework in which the analysis and its results may be interpreted. Prior to this, however, it is also necessary to develop an unambiguous and definitive terminology with which we may communicate. From a statistical perspective, the terms that we use and their meanings are widely recognised and understood. In the specific context of image processing, however, no formal syntax or semantics have ever been established in the general literature, with terms and definitions being either duplicated or otherwise insufficiently precise. A series of papers of Rosenfeld in the journal *Computer Graphics and Image Processing*, and the important book by Rosenfeld and Kak have gone some way to developing the fundaments of a language for the subject, to which a statistical aspect has been added by Fukunaga, Duda and Hart, and, perhaps most significantly, Besag (1986), who has had the most profound contemporary effect on the attitudes of statisticians, stimulating a great deal of the current interest and research. In this thesis, we follow generally the definitions and terminology introduced in these references, but also hopefully we will exclude any ambiguities, inconsistencies, and redundancies. Later, we shall present a glossary of important terms and the interpretation they will have in our subsequent work. We begin by introducing some notation necessary for our statistical formulation.

Recall that in the image processing data analysis problem, we are to observe in some space, subsequently denoted $\mathfrak{Y}$, data that arise indirectly from the physical or measurable aspects of a collection of entities in some other space, denoted $\boldsymbol{\Theta}$. We shall denote the observed data by $\boldsymbol{Y}$, and the unobservable quantities from which $\boldsymbol{Y}$ is derived by $\boldsymbol{\theta}$. In addition, and perhaps more usefully, we define $S_{\boldsymbol{\theta}}$ and $S_{\boldsymbol{Y}}$ as the physical regions in (at most) Euclidean 3-space in which $\boldsymbol{\theta}$ and $\boldsymbol{Y}$ - qualitatively interpreted at this stage as the characteristics of the collection of entities of interest and the observed data, respectively - are located and spatially configured. This definition is somewhat abstract, but its interpretation will become evident in the light of the examples we give below. Now, due to the practical considerations of the data collection procedure, it is necessary to impose some discretisation on the region $S_{\boldsymbol{Y}}$. In the same way, but now for reasons of statistical convenience and ease of implementation, we also consider it important to impose discretisation on $S_{\boldsymbol{\theta}}$. Thus, it is clear that, in general, we shall assume that $\boldsymbol{Y}$ is a vector contained in Euclidean m-space, and that $\boldsymbol{\theta}$ is a vector contained in Euclidean M-space. The elements of these vectors are commonly referred to as **pixels** (picture elements). Consequently, we shall refer to the vector of unobservable pixel values $\boldsymbol{\theta}$ in $S_{\boldsymbol{\theta}}$ as the **true scene** pixel values, and similarly we shall refer to the vector of observed data pixel

values $\boldsymbol{Y}$ in $S_{\boldsymbol{Y}}$ and derived from θ as the **image** pixel values.

The interpretation of the terms defined above is most readily demonstrated by means of a simple example. Consider the case where S_θ and $S_{\boldsymbol{Y}}$ are both planar regions (that is, S_θ and $S_{\boldsymbol{Y}}$ have the same dimension, $S_{\boldsymbol{Y}}$ is not a projection of S_θ), and furthermore suppose S_θ and $S_{\boldsymbol{Y}}$ coincide exactly in a rectangular region of dimensions (l_1, l_2), denoted by S. Using the natural coordinate system with axes parallel to the boundaries of S, we may perform the necessary discretisation of $S_{\boldsymbol{Y}}$ by imposing a $n_1 \times n_2$ grid of rectangular pixels each of size $(l_1/n_1) \times (l_2/n_2)$ on the region, and subsequently recording data elements per pixel. Thus, $\boldsymbol{Y}$ is a real vector having $n_1 \times n_2$ elements, each of which may be univariate or multivariate quantities. Similarly, we may produce a discretisation of S_θ by imposing on it a scaled version of this grid, consisting of $N_1 \times N_2$ rectangular pixels of size $(l_1/N_1) \times (l_2/N_2)$, and consequently θ is a real vector having $N_1 \times N_2$ elements. Generally, we shall regard the elements of θ as taking values on some set of integers rather than the whole real line, for reasons that we discuss below. Figure 1 depicts the results of such a discretisation, with $N_1 = N_2 = 16$, and $n_1 = n_2 = 8$, producing square pixels when the region S is chosen to be square. The discretised versions of S_θ and $S_{\boldsymbol{Y}}$ are depicted in figures 1(a) and (b) respectively.

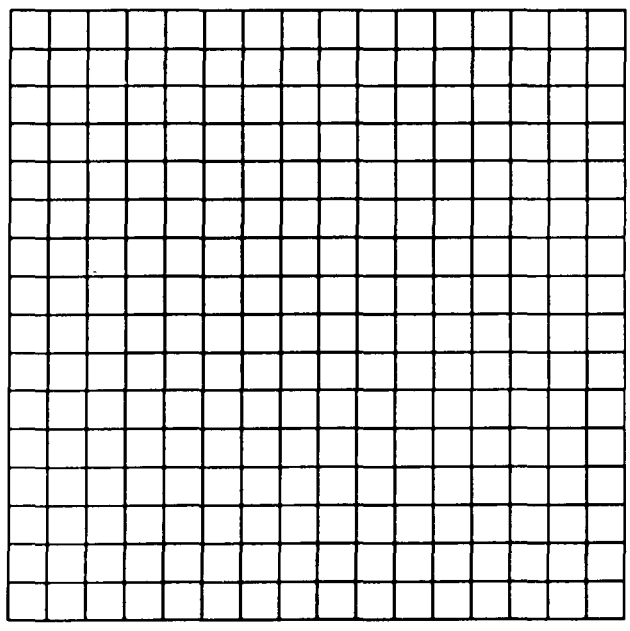


Fig 1(a) : S_θ

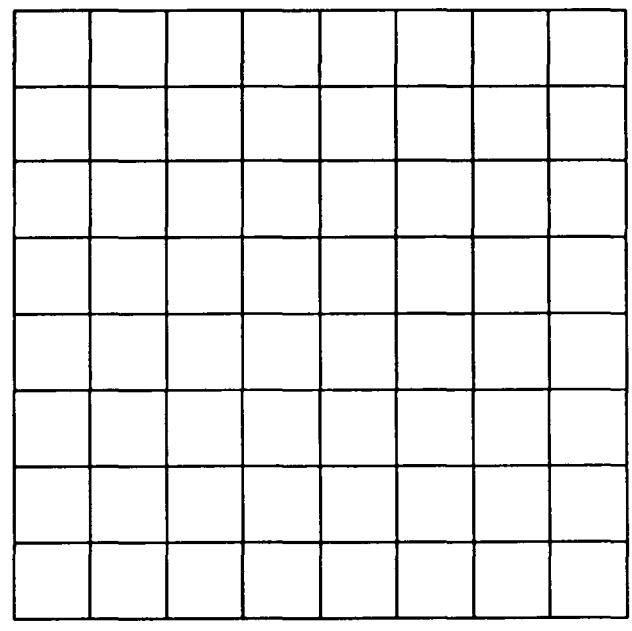


Fig 1(b) : $S_{\boldsymbol{Y}}$

Clearly, this merely represents one possible version of the many different types of discretisation that may be used. It has several important features. First, the grid used to discretise S_θ is finer (or of a higher **resolution**) than that used to discretise $S_{\boldsymbol{Y}}$. Generally, we might regard the grid used to discretise $S_{\boldsymbol{Y}}$ as being in some sense of fixed resolution due to the practical considerations of data collection as mentioned above. Therefore, we regard as coherent the use of a grid with higher resolution as a representation of elements in the true scene. Thus, loosely, we might regard $\boldsymbol{Y}$ as a corrupted version of a projection of θ, and, in our original notation, $M \geq m$. Secondly, it is clear that, in this particular instance, the pixels in $S_{\boldsymbol{Y}}$ correspond

independently to one group of four pixels in S_θ with no overlapping. Such considerations will be of importance in the subsequent statistical analysis. Thirdly, the actual nature of the processing problem may be such that S_θ and S_Y have different dimensions - for instance, we may have a two-dimensional image derived from a three-dimensional true scene, as is the case in many medical imaging examples - in which case S_θ and S_Y coincide only in projection. It is evident that each of these points reduce to the following: that, subsequent to the discretisation procedure, we must define the precise form of the function used when mapping pixels in S_θ to pixels in S_Y.

Having established the correspondence of pixels in S_Y to (blocks of) pixels in S_θ, we now proceed to discuss the introduction of randomness into imaging process. It is clear that if this process is regarded as purely deterministic, then statistical techniques are not required, and we return to well-established and moderately successful (if intuitively unsatisfactory) non-statistical techniques. Our interpretation of the imaging process, or the **image model**, is a conventional statistical one, namely that

$$Data \;=\; Structure * Noise \tag{1.1}$$

(see, for example, Smith(1986)), where the terms "*Data*" and "*Structure*" in (1.1) correspond respectively to "image" and "true scene" as defined above, "*Noise*" corresponds to the inherent but undesirable stochastic element, and * is an operator defining precisely how the *Structure* and *Noise* interact. This interpretation of the term "*Noise*" is very close indeed to its common interpretation in the image processing context, where a **noise-process** is regarded as acting to **corrupt** the underlying **signal**. Hence, we denote the noise-process by ε, and thus we may formally re-write (1.1) in the image processing context as

$$f : (\theta, \varepsilon) \rightarrow Y \quad , \tag{1.2}$$

where f is merely some function involving the operation * and the pixel correspondence described above. We could qualify the precise form of f (see Geman and Geman (1984) for a mathematical exposition, and, for example, Rosenfeld and Kak (1982) for the image processing aspects), or merely regard it as some "black-box" operation. We favour the latter of these options, except where specific knowledge of the operation is relevant to our subsequent modelling assumptions. We hence refer to f as the **image-formation process**, and generally consider its form to be a consequence of the mathematical rather than the physical aspect of the image processing problem (although, of course, the former will typically be motivated by the latter). We discuss the particular aspects of the choice of f in more detail in a later chapter, and we shall see that this choice can be regarded as part of the Bayesian *a priori* model elaboration

procedure, and that generally f can be assumed to take some familiar form.

Having made the largely notational and mathematical definitions above, we now seek to define terms with a interpretation specific to the image processing context. For instance, we have so far regarded the unobservable θ as measurable aspects of entities in Θ. We shall from this point on refer to these entities as **texture regions**, or merely **textures**. This terminology is largely adopted from traditional 2-D signal processing, and we shall regard it as generically rather than specifically defined; that is, its actual interpretation will be contextual. Note that, here, the texture regions are regions in (at most) 3-space, rather than being merely planar, and that the usage of the term may or may not coincide with the more common usage. In certain practical instances, it may be useful to refer to individual texture regions as **objects**, and to the residual part of any true scene as **background** in the usual way. We shall also refer to particular configurations of texture regions, or indeed regions of a particular configuration as **patterns**, and to the boundaries between adjacent textures as **edges**, in the usual way. None of these definitions conflict to any great degree with those in the literature. Later, however, we shall be making further definitions that will supercede to some extent those made previously. First, we introduce the probabilistic formulation and notation used subsequently in this thesis.

As indicated previously, we shall attempt to solve the particular problems in image processing that we study within a Bayesian framework, and inference about the unobservable θ will be made conditionally on the observed data via some form of posterior distribution or density derived from a set of qualitative and quantitative prior assumptions. For example, we shall make decisions via posterior probabilities and carry out estimation procedures on the basis of posterior distributions, each in conjunction with the appropriate loss-functions for incorrect decisions. It falls beyond the scope of this thesis to review the formal (decision-theoretic) justification for the use of Bayesian methodology, but we feel that, first, in general, it provides the most intuitively satisfying method of solution to statistical problems, and second, specific to the image processing context, we shall see that this form of inferential procedure is of considerable use in the modelling of such complex stochastic systems.

It is clear that, since we shall adopt a statistical approach to image processing problems, it will be necessary to refer notationally to certain forms of probability distributions and densities. We thus introduce the following notation. We shall write the marginal form for one variable and the joint and conditional forms for two variables as

$$[\,.\,] \;,\; [\,.\,,.\,] \;,\; \text{and} \;\; [\,.\,|\,.\,]$$

respectively, with the obvious extension for higher numbers of variables. We shall also represent the marginalisation process, of variable θ_1 with respect to variable θ_2, say, as

$$[\theta_1] = \int [\theta_1, \theta_2] ,$$

or, equivalently,

$$[\theta_1] = \int [\theta_1 \mid \theta_2][\theta_2] ,$$

with no other reference being made to the integrator variable. Despite the minimalist nature of this notation, its interpretation in every context will be entirely obvious. For discrete probabilities, we shall occasionally adopt the usual Pr(.) notation.

Having introduced the basic terminology that we shall use when referring to problems in image processing, and the necessary statistical notation, we now proceed to describe and discuss several important areas within the subject that we study in later chapters. The terminology that we use will, on occasion, differ slightly in interpretation relative to more traditional interpretations. However, we believe our terminology to be sensible and consistent.

(1.3) Problems in image processing.

The problems that we discuss below can be regarded as fundamental problems in image processing, and themselves include virtually all other problems of interest in the subject. In our description, for definiteness, we shall make specific reference to the situation depicted in figure 1, where S_θ and S_Y coincide exactly in a planar rectangular region S, with each discretised into grids of rectangular pixels, and where the resolutions of the two grids may or may not be equal. We begin with what we regard as the most important problem for solution.

(1.3.1) Image segmentation.

The image segmentation problem can be presented simply as follows. Given the observed (image) data Y, corresponding to pixel values in S_Y, our objective is to allocate each of the elements in the unobservable (true scene) vector θ, corresponding to pixel values in S_θ, to one (or occasionally more) of the textures or texture regions in Θ. We shall refer to the allocation of pixels to textures as **classification**, both in a transitive and intransitive sense - for example, we might validly refer to the "true scene pixel classification" for pixel i, say, meaning the actual underlying value of θ_i, or equally as validly to a "pixel classification procedure" as the mechanism by which the pixels are allocated. Our interpretation of the term **image segmentation** here is identical to that of terms such as image restoration or reconstruction that are frequently used elsewhere in the literature. We feel that "segmentation" describes the nature of the problem more satisfactorily than either of these options.

We shall see later that the segmentation problem can be approached, broadly, in three ways, which we shall refer to as estimation, probabilistic classification, and non-probabilistic classification. Two of these three approaches, estimation and probabilistic classification, are derived using slightly different statistical assumptions, and have different ultimate objectives, but are generally closely related. The third, non-probabilistic classification, uses low-level or quasi-statistical arguments, and can be thought of as principally exploratory data analytic.

We regard segmentation as the fundamental problem in image processing, since it is the prime objective in the majority of fields of application. We thus regard the remaining problems described below as either preliminary or ancilliary to this primary objective. Later, we discuss two cognitive problems where further inferences concerning, for example, presence or absence of objects or patterns in the true scene are made either subsequent or in parallel to segmentation. First, we describe a problem that can regarded as an important preliminary step in the processing of image data.

(1.3.2) Edge-detection.

Generally, in the context of image analysis, our interpretation of the nature of the unobservable true scene is that it is comprised of broadly homogeneous texture regions configured in some way in relation to each other. Inherent in this interpretation of the true scene is the concept of boundaries between textures, or edges as defined above. Clearly, it is of interest to be able to identify the positions of these edges. We discuss in more detail in a later section the justification of our interest in the discernment of edge regions, or **edge-detection**, and at greater length in a later chapter, where we shall note its importance as a preliminary stage in the image processing procedure.

We shall see that, despite the considerable literature concerned with edge-detection methodology and applications, little has been done to formulate the problem in a formal probabilistic framework. This latter task is the primary concern of this thesis, and we shall see later that, in fact, the edge-detection problem can be approached in a decision-theoretic setting by appealing to other well-known statistical techniques. We now describe a third important problem in the image processing context.

(1.3.3) Object detection.

Consider a texture region configuration in which one texture region is completely spatially contained within another texture that itself extends to cover the remaining region of S_Y. In such a situation, we refer to these two texture regions as object and background respectively, as indicated briefly above. In this situation, we frequently wish to make inferences concerning the location, dimensions, and orientation of the object relative to the background, or

relative to the chosen coordinate axes, rather than merely seeking a solution to the segmentation problem. Such problems are common in many applications; for example in medical imaging using data collected using tomographic methods, or in military reconnaissance using remote sensing.

The nature of the inference problem here in this **object detection** problem is fundamentally different to that of the segmentation problem described above. It is cognitive rather than merely observational, and thus requires a different approach to its solution. Again, few attempts have been made in the literature to formulate this problem in a statistical (or at least estimative) setting. We shall attempt such an approach in a later chapter of this thesis.

Finally, we describe one further problem of a cognitive nature that is related to another aspect of image analysis.

(1.3.4) Pattern recognition.

We defined the term "pattern" above to mean a particular configuration of texture regions in the true scene. Implicit in this definition is the fact that such a configuration must have some characteristic quality that allows discrimination between it and other configurations. Thus, for any given image, we might wish to make inference relating to the presence or absence of patterns of a particular type. We shall refer to this inferential problem as **pattern recognition**, and note that it is practically relevant in many fields of application; for example, in the regulation of industrial and engineering processes, and in the machine processing of printed characters.

Clearly, the pattern recognition problem has links with areas of mathematics outside of statistics. The connection with sophisticated techniques concerned with shape analysis and morphology is obvious, but we might also note links with artificial intelligence, and also with the mathematical formulation of psychological concepts. The major part of this broad spectrum of ideas obviously falls beyond the scope of this thesis. However, we shall see in a later chapter the relevance of pattern recognition to simple problems of object detection.

We have described what we believe to be the four problems of primary interest in the area of image processing. We have noted that generally we regard image segmentation as our ultimate goal, but of necessity this must be proceeded by an edge-detection analysis, possibly complemented by other inferential procedures, such as object detection and pattern recognition. The structure of this thesis is broadly along these lines. For the remainder of this introductory chapter, we describe in detail the approaches developed previously in an attempt to solve two of these problems, namely image segmentation and edge-detection. Later, we give an account of various statistical and non-statistical edge-detection procedures that have appeared in the literature. First, we present a summary of the techniques that have been

proposed as a solution to the image segmentation problem. Due to the extensive literature on this subject, and bearing in mind that our objective is to formulate the problem in a decision-theoretic, probabilistic framework, we shall restrict our survey to purely statistical approaches to the problem.

(1.4) Statistical approaches to image segmentation.

As mentioned above, the problem of image segmentation has been approached in a statistical framework using techniques that fall into three broad categories. First, it has been viewed as an estimation problem, where the elements of θ, the true scene pixel classification values, are regarded as unknown parameters that may be estimated using classical (maximum-likelihood) or Bayesian (maximum probability) techniques. Secondly, it has been viewed as what we shall call a probabilistic classification problem, approached via such procedures as cluster analysis, discriminant analysis, and predictive classification, where the elements of θ are allocated to textures according to their fidelity to texture characteristics. (We note in passing that there are mathematical links between the maximum probability approach to estimation described above and the minimum distance approaches inherent in probabilistic classification, although the two approaches can be regarded as conceptually distinct.) Thirdly, the image segmentation problem has been viewed as what we shall call a non-probabilistic classification problem. We use this somewhat catch-all category to describe intuitively reasonable and effective segmentation techniques that do not fall readily into either of the other categories, but nevertheless still use some form of statistical methods. It will become apparent later precisely which sorts of techniques we include in this category. We now proceed to discuss each of these approaches in greater detail.

(1.4.1) Estimation.

Before presenting a summary of estimation oriented techniques and procedures, we introduce the specific forms of notation that we are to use. Recall equations (1.1) and (1.2), and suppose that the stochastic relationship between θ and Y due to ε and quantified through f is such that f is known. Then we write the probabilistic dependence of Y on θ as $[Y \mid \theta]$, with all other aspects of the dependence being suppressed at this stage. In a classical statistical framework, subsequent inference about θ is frequently made via maximum-likelihood methods, so that, in particular, an estimate of θ, denoted by $\hat{\theta}$, is given by

$$\hat{\theta} = \arg\max_{\theta} [Y \mid \theta]$$

In a Bayesian decision-theoretic framework, inferences are made via the posterior distribution for θ given Y, denoted by $[\theta \mid Y]$. In this framework, the required estimate for θ is derived

from $[\theta \mid Y]$ and an appropriate loss function for incorrect actions, denoted by $l(\theta, \hat{\theta})$. The usual Bayesian risk analysis can then be used to show that if this loss function takes the form

$$l(\theta, \hat{\theta}) = \begin{cases} 0 & \hat{\theta} = \theta \\ 1 & \text{otherwise} \end{cases} \tag{1.3}$$

then the optimal choice of $\hat{\theta}$, that is, the choice of $\hat{\theta}$ that minimises the expectation of the loss function taken with respect to $[\theta \mid Y]$, can be shown to satisfy

$$\hat{\theta} = \arg\max_{\theta} [\theta \mid Y] \quad , \tag{1.4}$$

that is, the estimate for θ is the joint mode of the joint posterior distribution for θ, $[\theta \mid Y]$. Such an estimate is termed the **maximum *a posteriori***, or M.A.P., estimate. Alternately, if the loss function takes the form

$$l(\theta, \hat{\theta}) = \sum_{i=1}^{M} l(\theta_i, \hat{\theta}_i) \tag{1.5}$$

and $l(\theta_i, \hat{\theta}_i)$ takes a similar form to (1.3), then the optimal choice for $\hat{\theta}$ is comprised of the M elements $\hat{\theta}_i$ that satisfy

$$\hat{\theta}_i = \arg\max_{\theta_i} [\theta_i \mid Y] \tag{1.6}$$

for $i = 1, \ldots, M$, where $[\theta_i \mid Y]$ is the marginal posterior distribution of the single parameter θ_i. Such an estimate is termed the **marginal posterior modal**, or M.P.M., estimate.

In an image segmentation context, much attention has been paid to evaluating the estimate of θ given by (1.4), the M.A.P. estimate, and markedly less so to the estimate given by (1.6), the M.P.M. estimate. However, we feel that is somewhat misguided, due to the contextual interpretation of the two respective loss functions. Informally, the loss function in (1.3) can be interpreted as "all incorrect segmentations derived from the image data are equally as bad", whereas that in (1.5) can be interpreted as "how bad a segmentation is depends directly on the number of incorrectly classified pixels". In the vast majority of practical applications, the latter loss function is clearly more appropriate than the former. For example, we would regard the segmentation in figure 2(a) when the true scene comprised a centrally positioned square region on a background as inferior to the segmentation depicted in figure 2(b). However, under the loss function in (1.3) the two segmentations are regarded as equally incorrect.

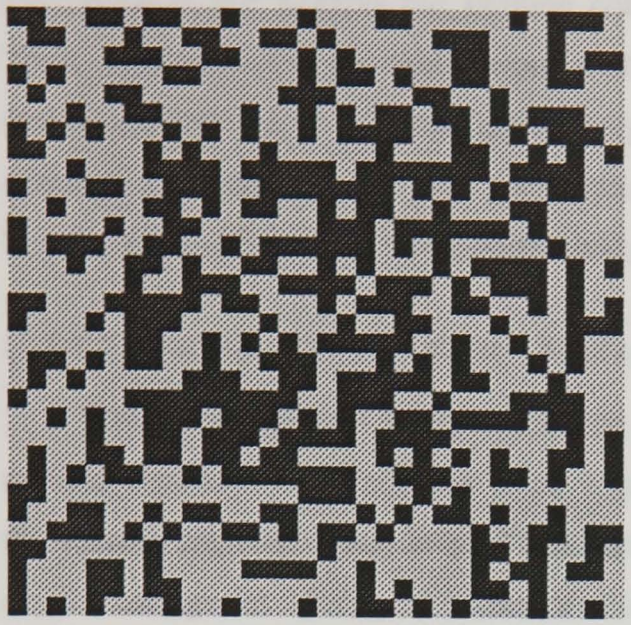


Fig 2(a) : segmentation (a)

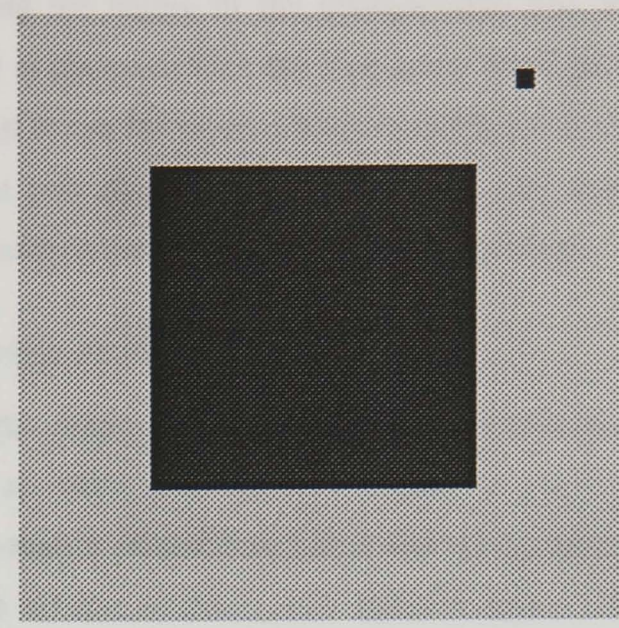


Fig 2(b) : segmentation (b)

Clearly, before we are able to report the required estimates we must first evaluate the joint posterior distribution, $[\theta \mid Y]$, or the set of marginal posterior distributions, $[\theta_i \mid Y]$ for $i = 1, \ldots, M$. In the Bayesian paradigm, evaluation of these posterior forms involves the specification of a prior distribution for the unknown parameters of interest. The nature of this prior distribution is that it should both qualitatively (through the choice of functional form) and quantitatively (through the choice of prior parameters) reflect our subjective opinions and beliefs relating to these parameters.

In the image segmentation context, for definiteness, we shall consider the specification of a joint prior structure for the true scene parameters, denoted by $[\theta]$, and the evaluation of the joint posterior distribution $[\theta \mid Y]$. It is clear that the joint and marginal prior and posterior distributions are, in fact, deterministically related, and that specification of a joint structure induces a marginal structure. Via Bayes theorem we have that

$$[\theta \mid Y] \propto [Y \mid \theta][\theta], \tag{1.7}$$

where $[Y \mid \theta]$ is the likelihood function. The problem thus reduces to specifying interesting forms for likelihood and prior, and identifying and evaluating the posterior distributional form that appears in (1.7).

Despite the fact that we might regard regard the form of $[Y \mid \theta]$ as fixed (by f), it is still strictly a consequence of our (prior) modelling assumptions, and thus we might view (1.7) as equivalent to a simple prior-posterior probability map. However, due to the practical considerations of the data collection process, the form of $[Y \mid \theta]$ is often restricted to be one of a small number of familiar functions. We are typically more at liberty to choose the form of

$[\theta]$ since it will express our opinion concerning the nature of the unobservable true scene. Thus, with respect to the other terms in (1.7), it is the model for the true scene pixel parameters as represented by $[\theta]$ that offers the greatest scope for more refined modelling and subsequently improved analysis. We now discuss one general approach to specifying $[\theta]$ that has broadly been accepted as being particularly relevant to the image segmentation problem.

(1.4.1.1) Modelling the true scene - Markov Random Fields.

Recall our general interpretation of the true scene as comprising homogeneous texture regions separated by edges, the latter regarded as small-scale features relative to the size of the texture regions. Thus, in any localised sub-region of the true scene we would generally expect contiguous blocks of pixels of the same texture to exist, with isolated pixels of any texture rarely occurring. In the light of this interpretation, one possible specification for $[\theta]$ may be constructed as follows. First, consider the conditional distribution of parameter θ_i given the parameter values at all other pixels, $\theta_{(i)}$, denoted by $[\theta_i \mid \theta_{(i)}]$. Then, because of our interpretation of the true scene, it would seem that a realistic modelling assumption is given by

$$[\theta_i \mid \theta_{(i)}] \equiv [\theta_i \mid \theta_{\partial i}] \tag{1.8}$$

where the vector $\theta_{\partial i}$ has elements which are the true scene parameter values for pixels in a locality or **neighbourhood** of pixel i. This assumption is appealing in the image segmentation context because it reflects our opinion concerning the local nature of the true scene. Thus, via (1.8), we have an appealing conditional prior structure for the parameters θ. Probability distributions exhibiting the property in (1.8) (and a further important property relating to positivity) are referred to as **Markov Random Fields**, or M.R.F.s, or also as **Gibbs** distributions - see Besag(1974) for further information. One important feature of such distributions (demonstrated by Besag and other authors) is that specification of the forms for the i conditional distributions $[\theta_i \mid \theta_{\partial i}]$, for $i = 1, \ldots, M$, completely specifies a unique joint structure for $[\theta]$ under weak regularity conditions. Such prior distributions have frequently been used to model the true scene in statistical image segmentation procedures. However, the implied joint distribution is complex, taking the form

$$[\theta] = \frac{1}{Z} \exp\left\{ - \sum_{c \in C} V_c(\theta) \right\} \tag{1.9}$$

where C is the set of **cliques** (subsets of pixels in which each element is a neighbour of all other elements, and in which all neighbours of a member pixel are contained), V_c is the clique **potential** for clique c specified on a scale relative to all other types of clique, and Z is the

normalising constant for the distribution. The interpretation of the joint structure is somewhat less appealing than that of the local conditional structure, but it is nevertheless important, as we must at all times consider the global implications of local assumptions. Again, these definitions are most easily explained by means of a simple example. Suppose that, in two dimensions, for any pixel i internal to the true scene pixel grid (that is, not on a boundary or on the corner of the grid), $\theta_{\partial i}$ comprises parameter values in all pixels horizontally, vertically and diagonally adjacent to i. Then each pixel of this type is contained within ten distinct types of clique - one of order one, four of order two, four of order three, and one of order four. These cliques, and suggestions for choices of clique potentials, are presented, for example, in Derin and Cole (1986). Interpreted using non-Bayesian terminology, $[\theta]$ is often viewed as acting as a smoothness or regularisation constraint that penalises "rough" segmentations.

As mentioned above, the appeal of the property (1.8) is somewhat lessened by the complex nature of $[\theta]$ in (1.9) and of the resulting posterior distribution from which the estimates are to be derived. Also, the complexity of the induced prior marginal structure prevents straightforward evaluation of the marginal posterior estimates; in fact, the marginal posterior distributions are virtually unobtainable using standard techniques when a prior of this form is used. Fortunately, a sophisticated technique for the optimisation of the joint posterior distribution has been developed, and was presented originally in the segmentation context by Geman and Geman (1984). We now discuss the algorithm developed there, and subsequent important developments.

(1.4.1.2) Stochastic Relaxation and Simulated Annealing - the Gibbs Sampler.

We present here a version of the original algorithm that differs somewhat in emphasis from that given originally by Geman and Geman, taking account of recent important developments in this area. Consider the following procedure. For each pixel i, we may write down the **full conditional posterior distribution** for parameter θ_i, denoted by $[\theta_i \mid Y, \theta_{(i)}]$, as

$$[\theta_i \mid Y, \theta_{(i)}] \propto [Y \mid \theta]\,[\theta_i \mid \theta_{(i)}] \tag{1.10}$$

where $[\theta_i \mid \theta_{(i)}] \equiv [\theta_i \mid \theta_{\partial i}]$ for a suitable neighbourhood system defining $\theta_{\partial i}$. Under suitable assumptions concerning $[Y \mid \theta]$, the conditional distributions given by (1.10) are straightforward to evaluate due to the simple nature of the conditional prior for θ_i given $\theta_{\partial i}$. The **Gibbs Sampler**, proposed by Geman and Geman, proceeds as follows. After assigning initial values to each of the θ_i in some arbitrary fashion, sample iteratively from each of the M full conditional distributions $[\theta_i \mid Y, \theta_{(i)}]$, with the values of the conditioning variables chosen on each iteration to be equal to the variates most recently obtained for those variables

by the iterative procedure. The iterative updating of the θ_i by the sampled values is referred to as **stochastic relaxation** (labelling, substitution). Under certain conditions (that the set of full conditional distributions uniquely define the joint distribution, and that each conditional distribution is sampled from "infinitely often"), Geman and Geman prove that as the number of iterations tends to infinity the joint sample of θ_i tends in distribution to $[\theta \mid Y]$ (see Geman and Geman (1984) for full details and proofs). Clearly, this technique is also of potential use elsewhere in Bayesian statistics, and we shall see such examples of its use later.

In the segmentation problem we therefore have a technique that allows us "eventually" to sample from the joint posterior distribution of the true scene pixel parameters $[\theta \mid Y]$. To derive, say, M.A.P. estimates, however, we must find the mode of this joint posterior distribution. Again due to Geman and Geman, a maximisation technique is available via the Gibbs Sampler. Instead of sampling iteratively from $[\theta_i \mid Y, \theta_{(i)}]$ on every iteration, we sample from $\{[\theta_i \mid Y, \theta_{(i)}]\}^{1/T}$ for some $T > 0$ (which is still relatively straightforward due to its discrete nature), and change T between iterations, starting with T large but decreasing it to zero as the iteration number increases. As T tends to zero, the set of sampled values are concentrated on the mode of the joint posterior distribution. The parameter T is referred to as a **temperature**, and the optimisation technique as **simulated annealing**, reflecting an analogy with techniques and processes in thermodynamics. In practice, the temperature must be altered according to a schedule that "cools" the system very slowly, and over a long time scale. Thus, although the technique is attractive, it can be computationally very demanding.

The Gibbs Sampler algorithm and associated annealing techniques as presented by Geman and Geman have subsequently been developed in a number of ways. The original authors suggest the use of an edge-process to lie along the edges between pixels, to act so as to restrict the influence of adjacent pixels lying within different texture regions. Ripley (1988) describes how the implementation of the algorithm may be improved and devises an appealing adaptive scheme to define the priority with which the pixel sites are visited. He also advocates the use of an exponential cooling schedule, as opposed to the logarithmic schedule proposed by Geman and Geman, and discusses the important issue of the assessment of convergence of the iterative process. We return to the problems associated with convergence at a later stage. Another amendment to the Gibbs Sampler as it was originally presented may be obtained by noting that we can use the identical technique to sample eventually from the marginal posterior distributions for the pixel parameters, $[\theta_i \mid Y]$ - this is merely a consequence of the theorems proved by Geman and Geman, and those stated later by Tanner and Wong (1987) in a related context. We therefore now have a technique enabling us to evaluate the M.P.M. estimates for the unknown pixel parameters that were previously unobtainable under this prior structure. More profoundly, the work of Tanner and Wong, and that of Gelfand and Smith (1990) and Gelfand *et al.* (1989), demonstrates that the Gibbs Sampler methodology may be

applied to a much wider range of statistical inference problems than those associated with image processing. For example, Gelfand *et al.* (1988) show that the algorithm may be used to compute estimates of marginal posterior densities for unknown parameters in a wide variety of normal data models, and Carlin *et al.* (1989) apply the methodology to inference problems in changepoint models: see also Hills (1989) for further theoretical and practical considerations, and Roberts and Polson (1990) for a general investigation of the nature and convergence of the procedure. Returning to the image segmentation problem, however, in practice, various difficulties may arise. Besides the problems concerned with computational load and the assessment of convergence of the algorithm, Greig *et al.* (1989) show that, in a special case when an exact form for the M.A.P. estimate is available, the exact estimate and the estimate derived using the Gibbs Sampler and annealing often differ considerably. It is also widely recognised that the choice of hyperparameters in the conditional prior distribution $[\theta_i \mid \theta_{(i)}]$ influences the nature of the posterior distributions greatly. In particular, the M.A.P. estimate derived from an inappropriate specification of prior hyperparameters often allocates all pixels to one texture. These negative features are difficult to understand and foresee. Although in simple cases we may study analytically and hence gain some understanding of the nature of the prior, the nature of the posterior is considerably more complex. We must be aware therefore that, despite the attractions of Gibbs distribution models and the stochastic relaxation and simulated annealing techniques, a number of practical and theoretical difficulties remain. Indeed, several authors (for example, Blake and Zisserman (1987)) doubt the usefulness of Gibbs models in problems of modelling true scenes in the image analysis problem. We believe, however, that the M.R.F./Gibbs structure captures the qualitative aspect of our prior opinion quite adequately.

Several models have been proposed to represent the spatial structure relevant in image segmentation problems. In the important papers of Besag (1974,1975), the distinction is drawn between so-called Conditional Markov (Autoregression) (CM or CAR) models and Simultaneous Autoregression (SAR) models, each of which can used to specify spatial structure, although the two approaches are not equivalent because of the different covariance structures involved in the specifications - indeed, Molina and Ripley (1989, section 3) suggest that we can

> "think of the CAR prior corresponding to first differences being white noise but the SAR prior to second differences being white noise. "

See, for example, Kanal (1980), Kashyap *et al.* (1981), Ripley (1981,1988), Kashyap and Chellappa (1983), Kunsch (1987) and Kent and Mardia (1988) for further discussion concerning the representation of spatial structure. In particular, the formulation above, and those of Hassner and Sklansky (1980), Cross and Jain (1981), Derin *et al.* (1984), Geman and Geman (1984), Derin and Elliot (1987), and Cohen and Cooper (1987) fall into the former category,

whereas, for example, those of Woods (1981), Khotanzad and Chen (1987), and Woods *et al.* (1987) fall into the latter. The choice of one approach over the other usually a matter of personal taste, or due to algorithmic considerations. In this thesis, we shall adopt the former approach, wherein, in the continuous case, the conditional expectations of the pixel values are simple linear sums of their neighbouring pixel values. Also, Besag (1975) considered the analysis of non-lattice data involving spatial structure. These spatial models have applications in each of the image segmentation techniques that we shall discuss.

We have discussed in some detail the estimative approach to the image segmentation problem, as it is with this approach that this thesis will be primarily concerned. We now discuss briefly the two other statistical approaches to this problem that have been adopted in the literature.

(1.4.2) Probabilistic classification.

Probabilistic classification techniques such as cluster analysis or discriminant analysis derived from classical likelihood or Bayesian posterior or predictive probability formulations are familiar in statistics. They are largely concerned with the optimal allocation of each element of the data set to one of a number of classes, with optimality defined in terms of minimum distance (classical) or maximum probability (Bayesian) criteria. Hence links with maximum probability approaches in the estimation context as described above are apparent, although there is a clear difference in emphasis between the two approaches.

In the image segmentation context, probabilistic classification techniques require that we allocate each pixel to one of a number of textures about which we have some prior opinion or information relating to physical nature. In the case where no spatial structure is assumed, for a fixed number of textures K, denoted $T_0, \ldots, T_{k-1}$, the simplest probabilistic classification rules are defined as follows. In a classical/maximum-likelihood framework, we allocate pixel i to texture T_j, here denoted by $\theta_i \in T_j$, if

$$[Y_i \mid \theta_i \in T_j] > [Y_i \mid \theta_i \in T_k] \quad , \quad k \neq j \tag{1.11}$$

which in the Normal case reduces to a minimum-distance criterion, and, in a Bayesian framework, $\theta_i \in T_j$ if

$$\Pr(\theta_i \in T_j \mid Y_i) > \Pr(\theta_i \in T_k \mid Y_i) \quad , \quad k \neq j \quad , \tag{1.12}$$

the maximum posterior probability criterion. The relationship between the estimative and allocative maximum-likelihood and Bayesian approaches is evident through the forms of (1.11) and (1.12) and the forms that appeared in section (1.4.1), particularly equation (1.6). Again,

we concentrate here on the Bayesian formulation. In the usual way, we may rewrite the terms in (1.12) as

$$\Pr(\theta_i \in T_j \mid Y_i) \propto [Y_i \mid \theta_i \in T_j] \Pr(\theta_i \in T_j) \tag{1.13}$$

where the first term on the right-hand side is now generally an integrated likelihood derived from the modelling and prior assumptions, and the second is the prior probability of pixel i being a member of texture T_j. Thus, comparisons of the sort necessary to achieve a maximum posterior probability classification of each pixel in the image defined by (1.12) are straightforward through (1.13). A complicating factor is that the parameters appearing in the likelihood, the **texture parameters**, are generally unknown and so must be estimated using **training data** from regions in the image known *a priori* to be derived from individual textures, or have some prior distribution specified for them, or be estimated as part of an iterative classification procedure such as the K-means algorithm described by Hartigan (1975). See also Duda and Hart (1973) as another general reference on the use of discriminatory probabilistic techniques in the image analysis problem.

General features of this simple approach are that the conditional distributions of data values given the texture parameters take on well-known and tractable forms (usually Normal), that there is no structure in the true classification, and that each datum point is allocated to only one of the K possible classes. In the image processing context, the first of these is necessary and acceptable, but the second and third perhaps seem inappropriate. We have described above the advantages of introducing some from of spatial structure into our prior specification for the true scene pixel classification, and such a spatial element has been introduced into the classification problem by, for example, Fu and Yu (1980), Switzer (1980,1983), Mardia (1984), Haslett (1985), and Klein and Press (1987). We also recognised in our initial formulation of the image processing problem that each observed image pixel value may be derived from a number of true scene pixels, and in a similar way we might entertain the possibility that each pixel value in the image was an observation from the mixture of the K individual texture probability distributions $[Y_i \mid \theta_i \in T_k]$ weighted by some unknown factor w_{ik}, $k = 0, \ldots, K-1$. Interest would then lie in making inferences concerning the w_{ik} for each i, interpreted as representing the proportion of pixel i belonging to texture T_k. Such an approach is termed **fuzzy classification**, and the procedure is described in more detail by Kent and Mardia (1988); see also the recent work reported by Cannon *et al.* (1986) and Gath and Geva (1989). We accept the appeal of such an approach in the classification context, but note that the M.P.M. technique described above, with the Gibbs Sampler used to compute approximate marginal posterior distributions for the true scene classification parameters, seems itself to provide a suitably fuzzy rule, which we "harden" by recording only the posterior modal texture

for each pixel.

Finally, we turn to the third category of classification techniques that we loosely describe as being non-probabilistic, although in general these techniques will have some basis in probability. The techniques are largely informal in nature, but are nevertheless worthy of mention.

(1.4.3) Non-probabilistic classification.

The first example that we study in this section relates to much of the work discussed in (1.4.1) and (1.4.2) above. It was proposed initially by Besag (1983), and discussed later in more detail by the same author (Besag (1986)). We regard it as an informal technique, as the theory underlying the algorithm through which it is implemented in somewhat incomplete, yet one which neatly captures several of the important aspects described above.

(1.4.3.1) Iterated Conditional Modes.

Consider the modelling of image-formation and noise process leading to (1.2), and the Bayesian formulation of the estimative approach to image segmentation described in (1.4.1). Consider in particular the modelling assumptions relating to the M.R.F. prior for the true scene pixel classifications defined through (1.8) and (1.9), leading to the full conditional posterior distribution for parameter θ_i, $[\theta_i \mid Y, \theta_{(i)}]$, given by (1.10). Besag proposed that, rather than sampling randomly and iteratively from the each of these full conditional distributions and using annealing eventually to locate the joint maximum or marginal maxima, we should at each stage merely locate the mode of each univariate distribution $[\theta_i \mid Y, \theta_{(i)}]$ deterministically, and then use the modal ordinate as the current value of the parameter θ_i to be used as the value of the conditioning variable in the subsequent iterative procedure. This technique captures the important spatial element, and the maximisation step reflects our interest in the modal estimates. Besag termed this technique **Iterated Conditional Modes**, or I.C.M.. In practice the algorithm often produces adequate segmentations in a remarkably few number of iterations, and thus is less computationally demanding than the Gibbs Sampler. However, the problems of choice of hyperparameters in the M.R.F. prior and the assessment of convergence of the algorithm still remain. Indeed, as the algorithm proceeds, the quality of the segmentation often deteriorates. Despite this, the I.C.M. technique can be regarded as more robust (in the short term) to the nature of the prior field specification; that is, segmentations consisting entirely of one texture will only occur after a large number of iterations.

Besag also discusses an estimation scheme for the texture parameters appearing in the likelihood that will in general be *a priori* unknown - this point was not discussed by Geman and Geman when the Gibbs Sampler was introduced initially. The estimation technique proposed by Besag (1974,1986) is that of maximum pseudo-likelihood, where the estimates are

those values maximising a function derived as the product of the likelihood functions of non-independent sets of variables. Although not a true maximum-likelihood estimate for the vector of unknown parameters, this estimate can be proved to have certain similar attractive properties (see, for example, Besag (1977) on efficiency, Geman and Graffigne (1987) on consistency, and Lakshmanan and Derin (1989) for further discussion), and can be regarded as an adequate approximation to the true maximum-likelihood estimate. The problem of parameter estimation in M.R.F.s specifically in the image processing context has never been adequately solved in a Bayesian framework.

Thus the I.C.M. technique, although appealing in many ways, actually involves some rather *ad hoc* procedures, and, unlike the Gibbs Sampler, has little theoretical justification - there are no convergence results equivalent to those proven by Geman and Geman. Note also that the resulting segmentation has no associated measure of uncertainty, one principal justification for the use of statistical methods in image segmentation. Its chief role currently is to act as a pre-processing procedure for other, more formal, techniques.

(1.4.3.2) Thresholding.

The second informal segmentation technique that we discuss is known as **thresholding**. Consider the problem of allocating each pixel in the image to precisely one of K textures. Suppose that the textures are homogeneous, and numbered so that the texture mean levels $\mu_0, \ldots, \mu_{K-1}$ form a monotone increasing sequence. Then by choosing $K-1$ constants $t_1, \ldots, t_{K-1}$ that also form a monotone increasing sequence, we might classify pixel i to texture T_j if the realisation of variable Y_i, denoted by y_i, lies in the interval between t_j and t_{j+1}, with t_0 and t_K suitably defined as negative and positive infinity, respectively; that is,

$$t_j \leqslant y_i < t_{j+1} \quad \Rightarrow \quad \theta_i \in T_j \quad , \quad j = 0, \ldots, K-1 \, . \tag{1.14}$$

Clearly, such a procedure is related to the simple probabilistic classification methods described above, under assumptions of normality, common noise variance across the image and a maximum-likelihood/minimum-distance criterion, or a maximum posterior probability criterion under a vague prior specification. More generally, it can be regarded as a simple non-parametric segmentation technique - this is its familiar interpretation in classical image processing. However, the segmentations obtained are sensitive to the particular threshold values chosen, whether the choice be made using exploratory methods (histograms), through information from training data, or prior knowledge of the true scene. Ridler and Calvard (1978) proposed a simple adaptive thresholding procedure. More recently, Mardia and Hainsworth (1988) developed a spatial thresholding method by incorporating the prior knowledge of spatial structure discussed above, and presented a comparison of techniques for a number of

images. See also Perez and Gonzalez (1987) for another adaptive thresholding algorithm.

The iterative and adaptive thresholding techniques described above perform remarkably well despite their simple nature. Again, however, as for the I.C.M. technique, the segmentation carries with it no associated measure of uncertainty, and thus might similarly be regarded as a pre-processing operation to be carried out prior to a more sophisticated analysis.

Finally, we mention briefly two further approaches to image segmentation that are implemented as optimisation procedures. First, variants of the EM-algorithm (Dempster *et al.* (1977)) have commonly been used to obtain estimative maximum-likelihood segmentations from noise-corrupted images. Most recently, Silverman *et al.* (1990) developed a version of the algorithm to reproduce images from data in the context of a positron-emission tomography experiment by including a smoothing step. This iterative algorithm produced more than adequate results in practice, but again convergence issues proved difficult. Also, again, point estimates only are obtained by such procedures. Secondly, Gull and Skilling (1985) proposed maximum entropy as a methodology and criterion for solution of the segmentation problem. Such a technique commonly involves considerable computational expense, and also Molina and Ripley (1989) question the validity of the approach to image segmentation, due to the nature of the entropy "prior" function.

(1.5) Edge-detection.

As mentioned above, edge-detection must be regarded as an important preliminary operation in any form of image analysis. It is evident, for example, that although the specification of M.R.F. prior forms for the true scene classification via simple local conditional distributions is adequate for pixels internal to a large homogeneous texture region, such simple assumptions will not be appropriate at or near texture boundaries. Naturally, therefore, the edge-detection problem has received considerable attention in the classical image processing literature. A review of edge-detection techniques can be found in Rosenfeld and Kak (1982, chapter 10). We note in particular the work of Nevatia and Babu (1980), who used simple thinning and thresholding techniques as the basis of a line-finding algorithm, and the work of Marr and Hildreth (1980), Haralick (1984), Nalwa and Binford (1986), Chen and Medioni (1989), De Micheli *et al.* (1989), and Zhou *et al.* (1989), who used a variety of techniques based on localised differential operators under simple statistical (Gaussian) assumptions for the image model. Such techniques commonly involve numerical differentiation or approximation to differentiation, and subsequent optimisation of the first derivative (**extrema** methods), or location of positions where the second derivative is zero (**zero-crossings** methods): see also Torre and Poggio (1986) for a discussion of the two related (but non-equivalent) techniques and a description of a regularisation approach to the edge-detection

problem, and the work of Canny (1986), who also uses an essentially regularisation-based approach. As such techniques commonly revolve around local operations performed in series for different sub-images or **windows** within the complete image, subjective choices must be made concerning the size of window used and the precise way in which the results from different windows of possibly different sizes are to be combined: see Lu and Jain (1989) for a discussion of such problems. This class of techniques often produces impressive results. However, we are of the opinion that the localised nature of such techniques is, in fact, in direct conflict with our interpretation of many edge-detection problems. We justify this opinion in more detail in chapter 2, where we shall see that a re-formulation of the edge-detection problem is indicated.

We briefly mention other edge-detection techniques which display a rather more formal statistical nature. Mascarenhas and Prado (1980) devised a complex Bayesian multiple hypothesis testing procedure from decision-theoretic principles. Cooper and Sung (1983) also adopted a Bayesian approach using a multiple-window optimal boundary finding algorithm. Recently, Bouthemy (1989) proposed a likelihood ratio hypothesis test for the detection of moving edges. Finally, Kashyap and Eom (1989) also devised a likelihood ratio test for edge-detection in images with more than one texture. This last technique is interesting as it attempts to locate edges by inspection of the data in relatively large segments in adjacent rows/columns of the image. We shall see the relevance of such an approach to our own work in chapter 2.

In general, therefore, we regard the complete edge-detection problem to be composed of three sub-problems; the detection stage itself, and subsequent **localisation** (removal of false edge-points etc.) and **reconstruction** or **representation** of the edge. In this thesis (chapters 2 to 6) we discuss various aspects of each of these problems. We be concerned in particular with a new approach to the detection stage of the problem derived from a Bayesian decision-theoretic viewpoint.

(1.6) Plan of thesis.

The structure of this thesis will be as follows. In chapter 2, we consider the edge-detection problem in more detail, and attempt to formulate it in a formal decision-theoretic framework. In this framework, we shall see that in certain circumstances the edge-detection problem in image processing can be interpreted as a familiar problem in a more general statistical context. On the basis of this analogy, we develop an edge-detection scheme with reference to an image derived from a simple true scene, with emphasis being placed on the need for computational efficiency. We discuss the advantages of our scheme over the local and *ad hoc* techniques described in section (1.5). In chapter 3, we adapt our formulation for the analysis of images derived from more complex true scenes, such as those containing convex objects and multiple texture regions. We shall often see that exact analysis is possible but

computationally demanding, and thus we shall seek to develop various approximation strategies. In chapter 4, we incorporate notions of localised pixel dependence and edge continuity into our original formulation. In chapter 5, we study the performance of our proposed scheme in the analysis of images derived using a range of image-formation processes. In chapter 6, we consider the reconstruction of edges from the sets of edge-points returned by our edge-detection scheme, and develop a procedure for the estimation of location, dimension, and orientation parameters for a particular class of simple convex objects. We also consider the detection of single or multiple objects in images using a variation of our original edge-detection scheme. Finally, in chapter 7, we show how the edge-detection routines developed in previous chapters can be incorporated into segmentation schemes at the early stages of the sophisticated procedures described in section (1.4). Also, in the context of the related segmentation problem, we develop an amended version of the Gibbs Sampler algorithm to overcome the difficulties associated with the estimation of texture parameters mentioned briefly in section (1.4.3.1).

Chapter 2 : Edge–Detection in Image Processing.

Edge-detection in its broadest sense is a segmentation technique based on the detection of localised discontinuities in an image true scene that arise at texture boundaries. It is widely regarded as an important first step in image processing for a number of reasons. First, and very loosely, most of the "information" contained in an image is to be found at the texture boundaries (Rosenfeld and Kak (1982)). Secondly, any presumed global structure concerning the spatial nature of the image true scene may be held to be invalid in the vicinity of texture boundaries - for instance, our qualitative belief about local dependencies motivates the choice of the Markov Random Field as a prior for the image true scene, but it is impractical to consider the precise dependence structure at each pixel. Thus generally we assume some symmetric form for the dependencies, e.g. neighbourhood systems holding over all pixels. Whereas this is an acceptable assumption for the majority of pixels, it is not necessarily so for those pixels near texture boundaries , and so it is of interest to investigate these pixels further. (The problem of "breakdown" in dependence structure is tackled by Geman and Geman (1984)) by means of a "line process" in conjunction with the more common pixel "intensity process".) Thirdly, and somewhat related to both of the above points, if initially we restrict attention to the efficient and accurate detection of local discontinuities, then any subsequent image analysis will be (a) presumably itself more efficient and accurate (note the importance of a good initial realisation of the line process in the work of Geman and Geman), and (b) perhaps rendered unnecessary, depending on the underlying decision problem. (In cognitive problems - shape analysis, pattern recognition - pixel-by-pixel classification of an image is not the real problem. For instance, the nature of the edges of a circle and a square are sufficiently different to enable us to distinguish between them, despite their topological similarity.) Consequently, we might expect considerable reductions in processing time if the edge-detection problem can be dealt with effectively. Finally, and more esoterically, psychological and physiological evidence indicates the actual use of edge-detection in biological visual perception systems.

Note that throughout the above discussion, we refer to the detection of "localised" discontinuities, and this is entirely accurate. However, our interest is in localised discontinuities between larger homogeneous regions , and thus localised detection methods (gradient operations, differencing) that operate over a small sub-grid of pixels may be seen to be inappropriate - our interpretation of an "edge" or "edge-point" at the true scene (unobserved) level is independent of the field-of-vision (entire image or image segment) but this is not the case at the observation level. This point is evinced by the poor performance of localised detection methods when used to analyse images with relatively high levels of noise-corruption, an example of which we shall see later.

(2.1) Edge-detection - simple example.

Consider the simplest possible and yet still interesting true scene for the edge-detection problem. Region S_θ consists of two textures T_1, T_2, the nature of each being governed by the (vector) parameters θ_1, θ_2, respectively, to be thought of as representing mean levels, scale or covariance parameters etc.. The two textures are separated by a simple edge (defined by a single curve in the plane), and thus there is an abrupt change in the parameters controlling pixels on either side of the edge. Figure 3 depicts such a true scene.

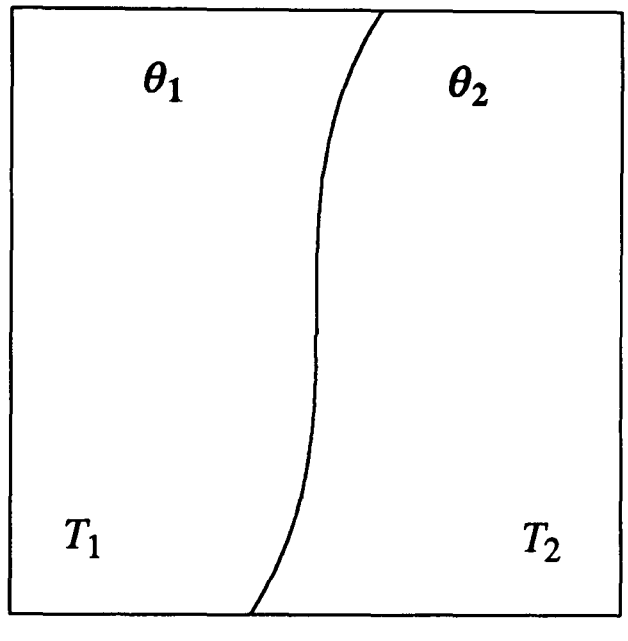


Fig 3 : Simple edge in region S_θ

The task that remains is straightforwardly stated: in the light of data Y, the image, observed on S_Y and presumed to be some noise-corrupted version of the true scene, make inferences concerning the location of the simple edge. The inference will take the form of, say, reporting of edge-points in Cartesian coordinates, or of some parametric or non-parametric curve to represent the edge, or merely of some highlighted version of the observed image. This simple problem is the basis of all edge-detection problems, since, clearly, the region S_Y may be considered as the entire picture or some segment of the entire picture containing one simple edge only. The subsequent classification of pixels in S_θ will follow on the basis of the inferences made about the position of the edge.

We proceed to consider a simple (but common) version of this simple problem : Assume that the image-formation process $f(\theta, \varepsilon)$ corrupts each cell in region S_θ independently with additive Gaussian white noise, that there is a 1-1 correspondence between pixels in S_θ and S_Y, and that the observed pixel image consists of univariate observations, so that

$$f : \theta \rightarrow Y$$

and

$$Y_{ij} = \theta_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2.1}$$

a standard linear "Signal + Noise" model. This is a commonly assumed image-formation model in the analysis of satellite data, and thus the "edge" in question in this particular simple edge-detection problem can be thought of as, for instance, a land-usage boundary, with the image Y being the collection of reflectances/radiances in a particular "band" recorded over all pixels. Figure 4 depicts a typical image realisation based on the simple edge of figure 3 and the image-formation process above.

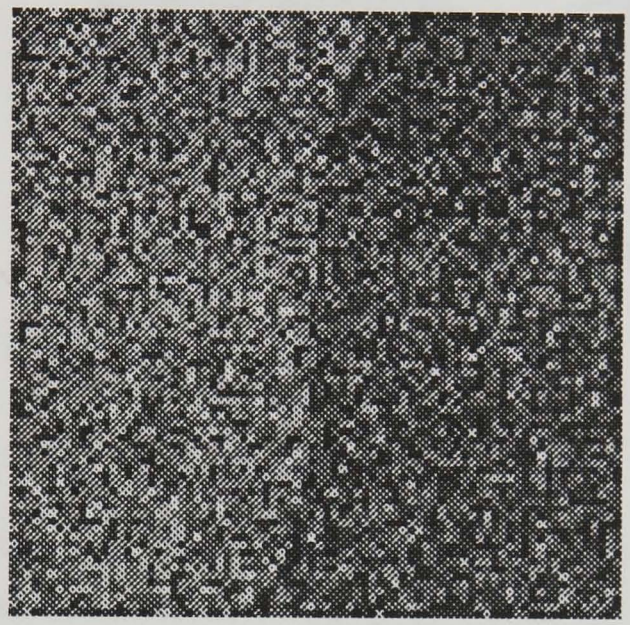


Fig 4 : Image containing simple edge

In this 80×80 pixel image, the mean level at every pixel is equal across each texture (i.e. $\theta_{ij} = \theta_k$ if pixel (i, j) lies in T_k, $k = 1, 2$) with the textures having different mean levels ($\theta_1 = 0.0$, $\theta_2 = 1.0$) and common variance ($\sigma^2 = 1.0$). Figure 4 is a six-level representation of the image.

It is clear, from this simple example, that localised edge-detection techniques that operate over a small sub-grid of pixels do not adequately reflect the nature of the edge-detection problem. An edge can only be discerned as such if it marks an abrupt change in some feature of the image between one large region and another. Techniques that do not take this into account cannot hope to capture the edge structure correctly. To make a visual analogy, in figure 4, we perceive the left half of the whole image to be "lighter" than the right, thus making our task of segmentation relatively easy, whereas were we to inspect 3 x 3 or 5 x 5 sub-images then much of the edge structure would be destroyed. Figure 4 gives some indication of this problem. The 5 x 5 sub-images of figure 4 are taken from different parts of the entire image. Figure 5(a) is a portion of texture 1, centred at pixel (12,12), figure 5(b) depicts an edge region centred at pixel (40,50), and figure 5(c) is a portion of texture 2, centred at pixel (62,50). It is not straightforward to distinguish which sub-image contains the edge.

It should be noted that this example is, in terms of Signal-Noise ratio, (defined here simply as the absolute value of the ratio $(\theta_1 - \theta_2)/\sigma$), relatively extreme (i.e. the ratio here is low, 1.0) and we might expect localised methods to perform adequately in less extreme cases. However it is important to note such fundamental flaws in the localised methods.

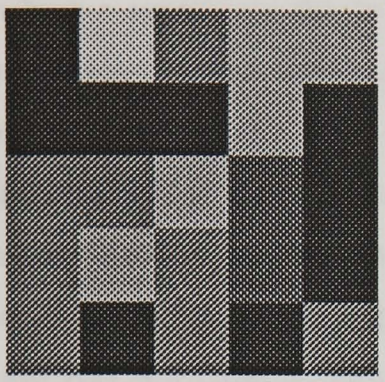


Fig 5(a) : Non-edge

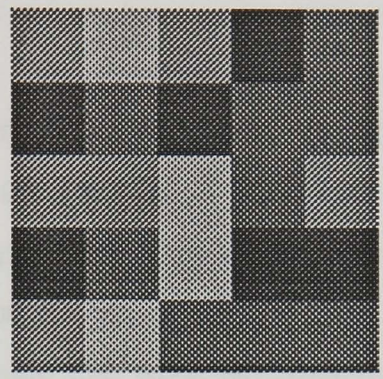


Fig 5(b): Edge

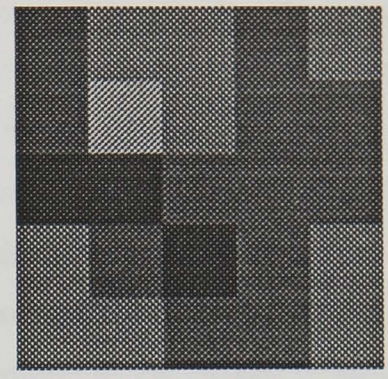


Fig 5(c) : Non-edge

Hence we seek an alternative approach to the edge-detection problem.

(2.2) Changepoint approach to edge-detection

We wish to formulate the edge-detection problem in such a way as to incorporate the notion that an edge should be interpreted as an abrupt (i.e. localised) change in some more large-scale feature. Consider the simple edge of figure 3, and a single row (j say) in the image matrix, as depicted in figure 6.

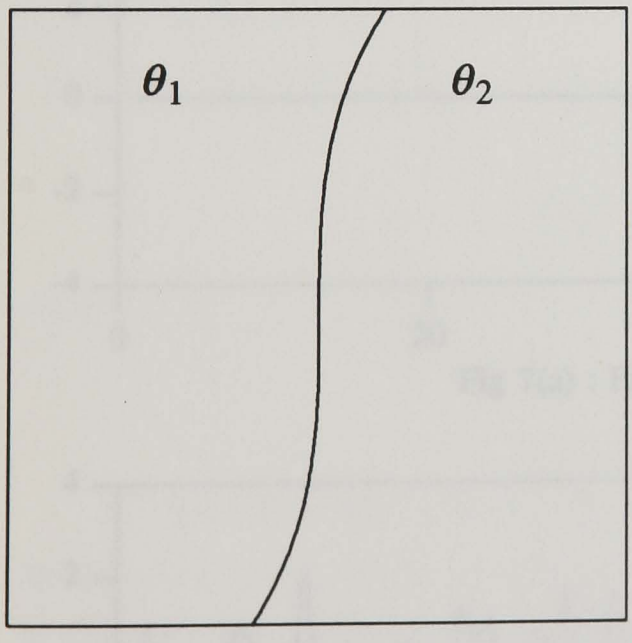


Fig 6(a) : Simple edge in region S_θ

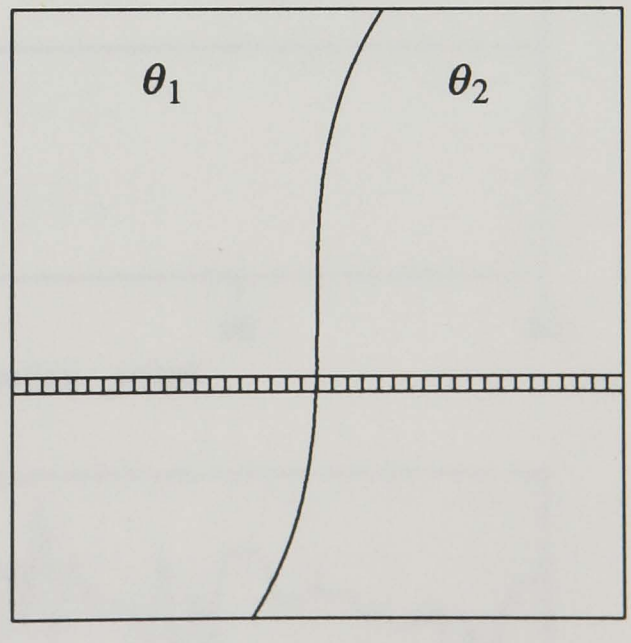


Fig 6(a) : Row j from data on region S_Y

Under the image-formation process (2.1) and assuming homogeneity of textures ($\theta_{ij} = \theta_k$ if pixel (i, j) lies in T_k, $k = 1, 2$) it is clear that the distribution of each of the data elements Y_{ij} in row j is as follows.

For some r ($1 \leqslant r \leqslant 80$),

$$Y_{1j}, \ldots, Y_{rj} \sim N(\theta_1, \sigma^2)$$

$$Y_{r+1j}, \ldots, Y_{80j} \sim N(\theta_2, \sigma^2) \qquad (2.2)$$

where r represents the unknown (and unobservable) edge-point position in row j. The edge-detection problem now reduces to that of making inference about r in single or over a number of adjacent rows. This is the familiar statistical problem of **changepoint analysis** and **identification**: see, for example, the reviews of Shaban (1980) and Zacks (1982). Hence the edge-detection problem in image-processing can be formulated so as to be essentially a practical application of changepoint analytic techniques. Figure 7 further illustrates this point. Figure 7(a) is a cross-section of a single row (row 50) from the true scene of the image in figure 2. It is of the same form as the representations of "ideal edges" in the image-processing literature, with the edge clearly visible between points 40 and 41 on the horizontal scale. Figure 7(b) is the same row taken from the noise-corrupted image. It is reminiscent of, for instance, time-series plots from system-monitoring operations, an area in which prospective/retrospective identification of changepoints is of some importance. Thus the use of changepoint analysis in edge-detection problem is intuitively reasonable. Note that the position of the underlying shift in mean-level (i.e. the edge) is barely discernible in figure 7(b), due to the noise-corruption, so that localised tests for shift in mean-level would be of little use.

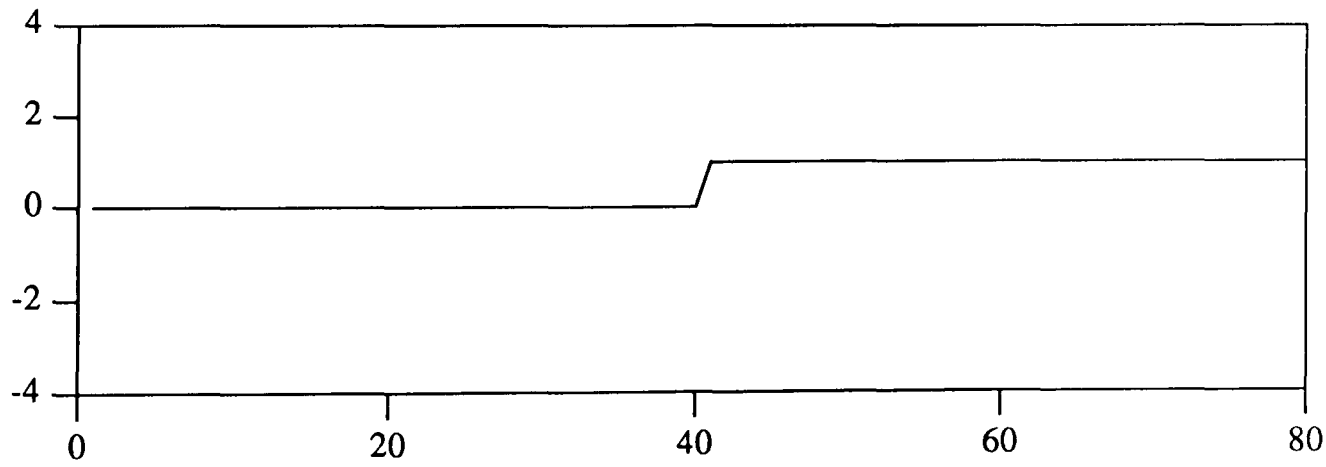


Fig 7(a) : Edge cross-section - actual

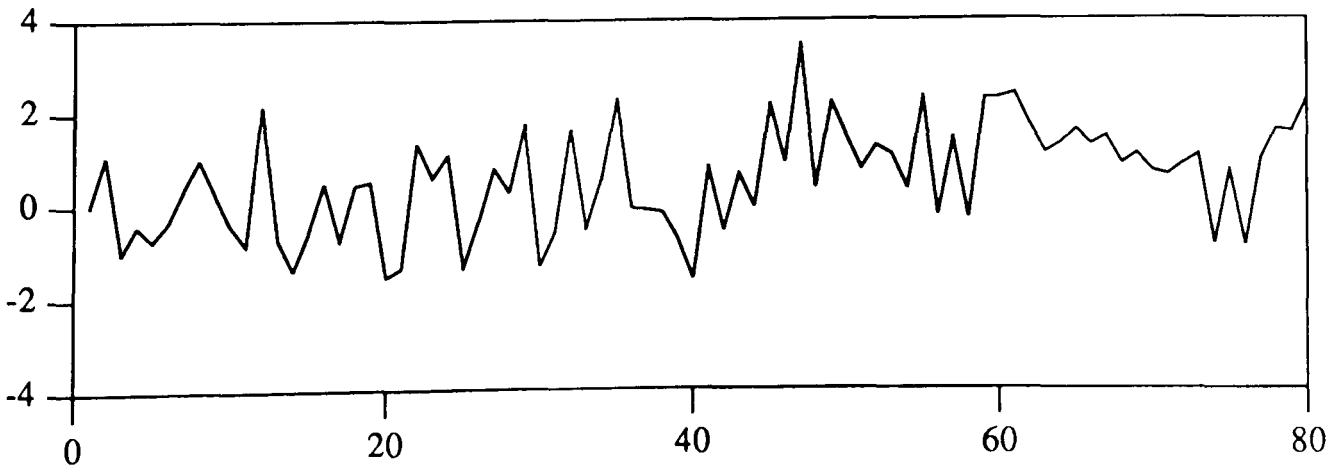


Fig 7(b) : Edge cross-section - noise-corrupted

A changepoint approach to edge-detection was in fact loosely suggested by Rosenfeld and Kak (1982, vol. 2, pp. 108-110), and proposed more fully by Basseville (1981). However, Basseville used a prospective scheme derived from Hinkley's cumulative sum procedure (Hinkley (1971)). We feel that a retrospective scheme is more attractive as it reflects the global rather than local aspects of the edge-detection problem.

A number of approaches to the changepoint problem appear in the literature, including those based on non-parametric (Pettitt (1980), Hinkley (1971)) or likelihood (Hinkley (1970)) formalisms. The approach we adopt here is Bayesian - see, for example, Chernoff and Zacks (1964), Broemeling (1972,1974), Smith (1975), Booth and Smith (1982). In the Bayesian formulation, inference is made via a posterior distribution for the unknown changepoint position, derived from prior assumptions concerning the functional relation between data and population parameters, and prior beliefs about those parameters. Before attempting to formulate the edge-detection problem in this way, we first formally introduce Bayesian approach to changepoint identification, and the necessary notation and terminology.

(2.3) Bayesian retrospective changepoint identification.

We adopt the following notation. Let $Y = (Y_1, \ldots, Y_n)$ be a sequence of random variables, and $y = (y_1, \ldots, y_n)$ be a realisation of these variables. Let θ be the vector of parameters of the sampling distribution, and ψ be a vector of hyperparameters appearing in the specification $[\theta \mid \psi]$ of the prior distribution for θ. Following e.g. Smith (1975), we make the following definition. The sequence of random variables $Y_1, \ldots, Y_n$ has a **changepoint** at r $(1 \leq r \leq n)$ if

$$Y_1, \ldots, Y_r \sim [Y_i \mid \theta_1]_1$$

$$Y_{r+1}, \ldots, Y_n \sim [Y_i \mid \theta_2]_2$$

where

$$[\,. \mid \theta_1]_1 \neq [\,. \mid \theta_2]_2$$

In this particular context, our emphasis will be on **retrospective** changepoint identification, that is, given a realisation $y_1, \ldots, y_n$ of the process, our objective is (primarily) to make inferences about the unknown changepoint position, r. Inference will be made via the posterior distribution of r, denoted by $[r \mid Y, \psi]$. From Bayes theorem, we have that

$$[r \mid Y, \psi] \propto [Y \mid r, \psi]\,[r]\,. \tag{2.3}$$

The first term on the right-hand side of (2.3) is the marginal distribution of Y given r and ψ,

and can be re-expressed as the likelihood function for Y integrated over the prior for θ, namely

$$[Y \mid r, \psi] = \int [Y \mid r, \theta, \psi]\,[\theta \mid r, \psi] \tag{2.4}$$

if θ is wholly or partially unknown, and simply as the likelihood itself if θ is completely known (in which case we identify ψ as θ). If indeed θ is unknown, then any subsequent inference about these parameters will be made via the posterior distribution of θ

$$[\theta \mid Y, \psi] = \sum [\theta \mid r, Y, \psi]\,[r \mid Y, \psi] \tag{2.5}$$

where

$$[\theta \mid r, Y, \psi] \propto [Y \mid r, \theta, \psi]\,[\theta \mid r, \psi]\,. \tag{2.6}$$

We make certain assumptions in order to simplify (2.3). First, in our formulation we specifically refer to ψ as a vector of hyperparameters, that is, parameters governing the nature of our prior belief. Hence, the conditional distribution of Y given r, θ, ψ is independent of ψ, i.e.

$$[Y \mid r, \theta, \psi] \equiv [Y \mid r, \theta]$$

Secondly, we shall, in general, regard θ as independent of r *a priori*, so that

$$[\theta \mid r, \psi] \equiv [\theta \mid \psi]\,.$$

Thirdly, we assume that $Y_1, \ldots, Y_n$ are conditionally independent given θ, and thus

$$[Y \mid r, \theta] = \prod_{i=1}^{n} [Y_i \mid r, \theta]\,.$$

Finally, we assume that all functional forms $[\,. \mid .\,]$ are known. Each of these assumptions are acceptable in the edge-detection context for specific choices of the image-formation process, as we shall see later. However, none of the assumptions is absolutely necessary and may be relaxed at a later stage. Therefore, from (2.3) and the assumptions above, the posterior distribution for r is given by

$$[r \mid Y, \psi] \propto \int \prod_{i=1}^{n} [Y_i \mid r, \theta] [\theta \mid \psi] [r] . \tag{2.7}$$

The final step in the Bayesian procedure is to report some estimate of r, $\hat{r}$, say, obtained via $[r \mid Y, \psi]$ and an appropriate loss function, rather than $[r \mid Y, \psi]$ itself. We shall, in general, assume a 0-1 loss function, i.e.

$$l(r, r^*) = \begin{cases} 0 & r^* = r \\ 1 & r^* \neq r \end{cases}.$$

The resulting estimate under this loss function satisfies

$$\hat{r} = \arg \max_{r} [r \mid Y, \psi] ,$$

i.e. $\hat{r}$ is the posterior mode. The modal ordinate is easily obtained from the discrete univariate posterior distribution.

Before returning specifically to the edge-detection problem, we discuss other aspects of Bayesian changepoint identification. The following general points arise from the formulation. First, there is an obvious and natural extension of the definition above from a single to a multiple-changepoint process. The Bayesian approach to the equivalent problems associated with multiple-changepoint sequences is identical to that above; i.e. we would make inference via $[r_1, r_2, \ldots, r_k \mid Y, \psi]$ where

$$\begin{aligned} [r_1, r_2, \ldots, r_k \mid Y, \psi] &\propto [Y \mid r_1, r_2, \ldots, r_k, \psi] [r_1, r_2, \ldots, r_k] \\ &= \int \prod_{i=1}^{n} [Y_i \mid r_1, r_2, \ldots, r_k, \theta] [\theta \mid \psi] [r_1, r_2, \ldots, r_k] . \end{aligned}$$

Secondly, $[r \mid Y, \psi]$ is simply a univariate, n-valued discrete distribution, and thus will be easily calculable, with straightforward optimisation, moment calculation, etc.. However, analytic results (concerned with, say, the properties of $[r \mid Y, \psi]$ when the distribution of Y is altered) will not be generally available. Finally, the precise nature of the "change" implied in the definition above is unspecified. We will consider here problems restricted to those in which the change is parametric, rather than distributional, so that

$$[\,.\mid.\,]_1 = [\,.\mid.\,]_2 \text{ but } \theta_1 \neq \theta_2 .$$

In the light of the above formulation, our primary interest will be in proposing various forms for $[Y \mid r, \theta], [\theta \mid \psi], [r]$ (likelihood - prior combinations) and examining the

resulting posterior forms $[r \mid Y, \psi]$. First, we discuss choices of $[Y \mid r, \theta]$ that will be particularly relevant in the image processing context.

(2.3.1) Forms for $[Y \mid r, \theta]$.

We shall, primarily, consider two forms for $[Y \mid r, \theta]$ (the likelihood), namely those arising from choosing $[Y_i \mid r, \theta]$ to have

(1) Normal

(2) Poisson

distributions, as these represent the two most relevant forms for the context in which the resulting posterior densities are to be used. (It is also convenient to choose the individual $[Y_i \mid r, \theta]$ so that $[Y \mid r, \theta]$ is easily formed from their product - e.g. choose from the exponential family - but that aside, we could assign $[Y_i \mid r, \theta]$ to reflect any of a wide range of image-formation processes.)

(2.3.1.1) $[Y_i \mid r, \theta]$ Normal.

Assuming Y_i to be conditionally normally distributed given $\theta = (\theta_1, \theta_2, \tau_1, \tau_2)$ the changepoint process of the definition above becomes

$$Y_1, \ldots, Y_r \sim N(\theta_1, \tau_1^{-1})$$

$$Y_{r+1}, \ldots, Y_n \sim N(\theta_2, \tau_2^{-1})$$

for some unknown r $(1 \leq r \leq n)$, where

$$\theta_1 \neq \theta_2$$

$$\theta_1 = (\theta_1, \tau_1)$$

$$\theta_2 = (\theta_2, \tau_2).$$

Under this scheme, three natural conditions lead to three forms for the likelihood $[Y \mid r, \theta]$.

(A) $\tau_1 = \tau_2 = \tau$ (Common precision)

$$[Y \mid r, \theta] = \prod_{i=1}^{n} [Y_i \mid r, \theta_1, \theta_2, \tau]$$

$$\propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{\tau}{2} \left[\sum_{i=1}^{r} (Y_i - \theta_1)^2 + \sum_{i=r+1}^{n} (Y_i - \theta_2)^2 \right] \right\}$$

(B) $\theta_1 = \theta_2 = \theta$ (Common mean)

$$[Y \mid r, \theta] = \prod_{i=1}^{n} [Y_i \mid r, \theta, \tau_1, \tau_2]$$

$$\propto \tau_1^{\frac{r}{2}} \tau_2^{\frac{(n-r)}{2}} \exp\left\{ -\frac{\tau_1}{2} \sum_{i=1}^{r} (Y_i - \theta)^2 - \frac{\tau_2}{2} \sum_{i=r+1}^{n} (Y_i - \theta)^2 \right\}$$

(C) $\theta_1 \neq \theta_2, \tau_1 \neq \tau_2$ (Different mean and precision)

$$[Y \mid r, \theta] = \prod_{i=1}^{n} [Y_i \mid r, \theta_1, \theta_2, \tau_1, \tau_2]$$

$$\propto \tau_1^{\frac{r}{2}} \tau_2^{\frac{(n-r)}{2}} \exp\left\{ -\frac{\tau_1}{2} \sum_{i=1}^{r} (Y_i - \theta_1)^2 - \frac{\tau_2}{2} \sum_{i=r+1}^{n} (Y_i - \theta_2)^2 \right\}$$

(2.3.1.2) $[Y_i \mid r, \theta]$ Poisson

Assuming Y_i to be conditionally Poisson distributed given $\theta = (\lambda_1, \lambda_2)$ the changepoint process of the definition above becomes

$$Y_1, \ldots, Y_r \sim Poisson(\lambda_1)$$

$$Y_{r+1}, \ldots, Y_n \sim Poisson(\lambda_2)$$

for some unknown r ($1 \leq r \leq n$). Here

$$\theta_1 = \lambda_1$$

$$\theta_2 = \lambda_2$$

Under this scheme, the likelihood $[Y \mid r, \theta]$ is given by

$$[Y \mid r, \theta] = \prod_{i=1}^{n} [Y_i \mid r, \lambda_1, \lambda_2]$$

$$\propto \lambda_1^{\sum_{i=1}^{r} Y_i} \lambda_2^{\sum_{i=r+1}^{n} Y_i} \exp\{-r\lambda_1 - (n-r)\lambda_2\}$$

Thus, for two different assumptions relating to sampling distributions that we regard as particularly relevant to image processing problems, we have derived expressions for the likelihood function necessary for the evaluation of the changepoint posterior distribution defined by (2.3) and (2.7). We now consider various specifications for the prior distributions that appear in these equations, namely $[\theta \mid \psi]$ and $[r]$. We first consider choices for the continuous

parameters θ appearing in the likelihood.

(2.3.2) Forms for $[\theta \mid \psi]$.

For each of the likelihoods we propose a selection of prior forms $[\theta \mid \psi]$ chosen so as to reflect the quantitative and qualitative nature of our prior beliefs. In general, we suggest the use of "conjugate" prior distributions (priors that combine with the likelihood so that the prior and posterior distributions for the parameter concerned take the same functional form) for convenience, but as before this restriction is not necessary. To obtain a representation of "prior ignorance", we consider limiting cases of (conjugate, proper) informative prior distributions (that is, via limits of elements of ψ), resulting in (improper) non-informative prior distributions.

(2.3.2.1) $[Y_i \mid r, \theta]$ Normal

We consider three cases :

(1.a) θ known

(1.b) τ or (τ_1, τ_2) known, θ or (θ_1, θ_2) unknown

(1.c) θ unknown

Situations in which one of θ_1, θ_2 is presumed known, or in which one or both of θ_1, θ_2 is presumed known with τ or (τ_1, τ_2) unknown are regarded as unrealistic in our context, but could quite easily be accommodated into our reasonably flexible framework (see, for example, priors 1.1.4 and 1.1.6 in Appendix 1). Also, we do not consider "one-sided" prior assumptions (that is, assumptions of the form $\theta_i > \theta_j$), but such priors could be included (and indeed may be relevant in later applications). For example,

$$[\theta_1, \theta_2, \tau] = [\theta_2 \mid \theta_1, \tau][\theta_1]$$

$$[\theta_2 \mid \theta_1, \tau] = U(\theta_1, \theta_1 + \tau^{1/2})$$

$$[\theta_1] = N(\mu_1, \eta_1^{-1})$$

is one such prior.

Finally, we note that in case (1.c) above, specifically with τ or (τ_1, τ_2) unknown, we choose (θ_1, θ_2) *a priori* independent, but dependent on τ or (τ_1, τ_2). For example, for likelihood (A) with τ unknown, we choose $[\theta_1, \theta_2, \tau]$ so that

$$[\theta_1, \theta_2, \tau] = [\theta_1 \mid \tau][\theta_2 \mid \tau][\tau]$$

rather than

$$[\theta_1, \theta_2, \tau] = [\theta_1][\theta_2][\tau].$$

The second of these two possible priors (a form in which θ_1, θ_2 and τ are chosen *a priori* independent) is regarded as inappropriate, as it induces in (θ_1, θ_2) a lack of invariance to scale changes (as noted by Spiegelhalter and Smith (1982)) which is undesirable. In the equivalent case with τ known, however, we may clearly choose θ_1, θ_2 *a priori* independent.

(2.3.2.2) $[Y_i \mid r, \theta]$ Poisson

We consider two cases :

(2.a) θ known

(2.b) θ unknown

Again, in case (2.b), we consider conjugate (Gamma) priors for the unknown $\theta = (\lambda_1, \lambda_2)$, and their non-informative limits; other restrictions ("one-sided" priors etc.) are as above.

Appendix 1 contains a selection of posterior forms $[r \mid Y, \psi]$ derived for a range of choices for $[\theta \mid \psi]$, assuming a uniform prior for r . The general form of $[r \mid Y, \psi]$ is broadly the same over the range of priors, but a degree of sensitivity to prior input is exhibited.

All the forms of $[r \mid Y, \psi]$ in Appendix 1 are derived under the assumption that a "change" is known to occur (i.e. $1 \leq r \leq n-1$). The possibility of "no change" (i.e. $r = n$) is a straightforward extension of our formulation. In this case we consider $Y_1, \ldots, Y_n$ where

$$Y_1, \ldots, Y_n \sim [Y_i \mid \theta]$$

where $\theta = (\theta, \tau)$ or λ and we could consider both θ and τ, or τ, or neither known *a priori*, and assign priors accordingly. No difficulty arises in the evaluation of posterior probabilities in this case provided proper priors are used. However, in the prior ignorance case, if the non-informative limits of the proper priors used are improper, then we are faced with the problem of assigning the constants of proportionality (omitted from Appendix 1) which we feel should be different for the two models "change" and "no change" due to the difference in dimensionality between the two models. A possible solution to this problem via a multiplicative correction factor is discussed by Spiegelhalter and Smith (1982) and Booth and Smith (1982).

(2.3.3) Forms for $[r]$.

The discrete prior distribution for the changepoint parameter r, denoted $[r]$, will generally be taken to be uniform over the range $1 \leqslant r \leqslant n-1$,

$$[r] = \frac{1}{n-1} \qquad r = 1,\ldots.,n-1 .$$

If a "no change" possibility is to be entertained, $[r]$ will taken to be

$$[r] = \begin{cases} \dfrac{(1-p)}{n-1} & r = 1,\ldots.,n-1 \\ p & r = n \end{cases}$$

for some p $(0 \leqslant p \leqslant 1)$. The implications of each of these particular choices in the edge-detection context are discussed in more detail below.

Thus a simple scheme for tackling the edge-detection problem can be proposed. For a sequence of known length, we would evaluate $[r \mid Y, \psi]$ in the light of the image-formation process and prior beliefs, using the techniques described in section (2.3). Suppose y is the vector of entries in any given row or column of the image matrix. Then as a solution to the edge-detection problem, we would merely compute $[r \mid Y, \psi]$ for $Y = y$, and report the posterior modal value and position as the most likely edge-position and associated measure of uncertainty in that particular row or column. We would repeat this procedure over all or a fixed set of rows and columns in the image independently, and report the set of recorded edge-positions as the result of the analysis. (This scheme clearly ignores certain aspects of the edge-detection problem, i.e. spatial continuity of edges. These points are discussed later in subsequent chapters.)

Such a scheme is attractive for a number of reasons that we detail in section (2.6) below. First, we demonstrate the use of the scheme in the context of the simple edge-detection problem described above.

(2.4) Implementation of the edge-detection scheme.

We now seek to implement the scheme proposed above in the context of the simple edge of figure 3 and the image depicted in figure 4, under the same image-formation and noise assumptions. Consider the elements of row j, say, and let $Y_i \equiv Y_{ij}$. We noted that in (2.2) the conditional distribution $[Y_i \mid r, \theta]$, $i = 1,\ldots,n$, is Normal, due to the image-formation and noise assumptions, with parameters $\theta = (\theta_1, \theta_2, \sigma)$. If we assume that the noise terms in (2.1) are mutually independent for all cells in any row, then subsequently the variables

$Y_1, \ldots, Y_n$ are mutually independent, conditional on θ_1, θ_2, and thus $[Y \mid r, \theta]$ is given by

$$[Y \mid r, \theta] = \prod_{i=1}^{n} [Y_i \mid r, \theta] . \tag{2.8}$$

For convenience we reparameterise by replacing σ with τ, the precision, where $\tau = \frac{1}{\sigma^2}$, and hence

$$[Y \mid r, \theta] \propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{\tau}{2} \left[\sum_{i=1}^{r} (Y_i - \theta_1)^2 + \sum_{i=r+1}^{n} (Y_i - \theta_2)^2 \right] \right\} , \tag{2.9}$$

identical to likelihood (A) in section (2.3.1.1). Now, if the texture mean levels and noise precision are *a priori* unknown, our next task is to specify some form for $[\theta \mid \psi]$. For demonstration purposes, following Booth and Smith (1982), we choose a simple form of non-informative prior distribution. Let $[\theta \mid \psi]$ be given by

$$\begin{aligned} [\theta \mid \psi] &\equiv [\theta_1, \theta_2, \tau] \\ &= [\theta_1 \mid \tau] [\theta_2 \mid \tau] [\tau] \\ &= const \end{aligned} \tag{2.10}$$

for $-\infty < \theta_1, \theta_2 < \infty, \tau > 0$. This is a standard non-informative prior form, and can be regarded as the limit of a standard informative conjugate prior distribution having θ_1 and θ_2 *a priori* independent conditional on τ (Spiegelhalter and Smith (1980)), namely prior 8 in section 1.1.8 of Appendix 1. Finally, we specify $[r]$ to be uniform on the range $1 \leq r \leq n-1$. Combining (2.9) and (2.10) via (2.7), we obtain $[r \mid Y, \psi]$ as

$$[r \mid Y, \psi] \propto \{r(n-r)\}^{-1/2} \left\{ \sum_{i=1}^{r} (Y_i - \overline{Y_A})^2 + \sum_{i=r+1}^{n} (Y_i - \overline{Y_B})^2 \right\}^{-n/2} \tag{2.11}$$

where

$$\overline{Y_A} = \frac{1}{r} \sum_{i=1}^{r} Y_i$$

$$\overline{Y_B} = \frac{1}{n-r} \sum_{i=r+1}^{n} Y_i$$

Note that in this formulation of the edge-detection problem, we consider prior models allowing **exactly** one changepoint, and thus the valid range for r (under our definition) is

$1 \leqslant r \leqslant n-1$. Booth and Smith (1982) also consider a "no changepoint" alternative, which extends the valid range for r to $1 \leqslant r \leqslant n$, and induces the obvious minor change in (2.9). Here, in this particular example, we restrict our attention to one changepoint models, and the resulting posterior form (2.11), but discuss the implications of the "no changepoint" alternative model in chapter 3.

We now proceed with an implementation of the proposed edge-detection scheme based on posterior distribution (2.11) in an analysis of the image in figure 4 derived from simple edge true scene in figure 3.

(2.5) Edge-detection - results.

In the following analysis, the posterior density in (2.11) was evaluated for each row of the image in figure 4, and the position of the posterior mode recorded, along with the modal probability. The results of this analysis can be seen in figure 8(a)

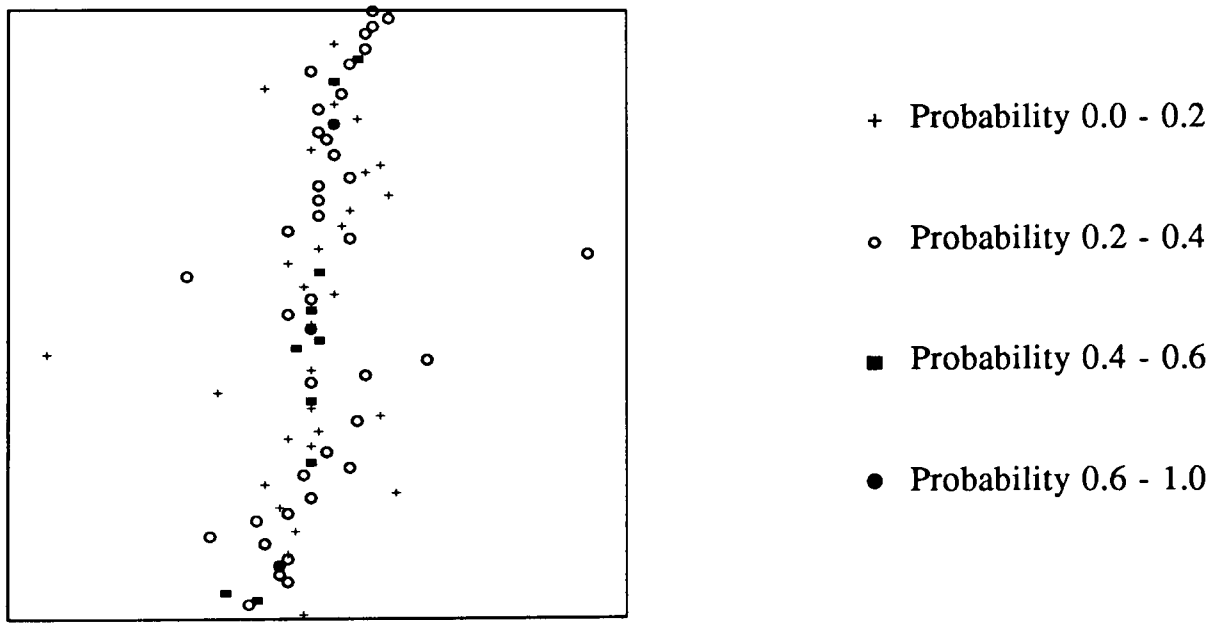


Fig 8(a) : Results of row analysis

The code for the symbols in figure 8(a) is also given. The complete analysis of the eighty rows of the image took around one second of CPU time. It is clear that much of the edge structure has been captured, i.e. many edge-point candidates arising as modes in the changepoint posterior density lie at or close to the true edge-point in the row concerned. In many cases, the results of a preliminary analysis such as this will be sufficiently accurate to enable the subsequent supervised or unsupervised processing techniques to proceed more efficiently - we can easily discern edge-regions as opposed to texture-regions, visually or automatically, allowing for more straightforward segmentation. However, as a representation of the edge itself, figure 8(a) is inaccurate due to the presence of serious edge misclassifications or "outliers". It is possible to remove these outliers using ideas of spatial continuity of the edge, that is, via our

interpretation of the edge itself as continuous in S_θ. We discuss this issue in greater detail in chapter 4. For the moment, we consider a simple technique for the removal of such misclassifications.

In the discretised version of the true scene, we would expect edge points in rows and columns to lie close to other edge-points in the adjacent rows and columns. Similarly, we would expect accurate edge-point classifications resulting from an edge-detection analysis to lie in close proximity to each other. Thus any "isolated" candidate points can be regarded as misclassifications, with the term isolated to be defined in some suitable fashion. A possible simple "smoothing" technique (in the sense that isolated candidate points disrupt our interpretation of an edge as being locally continuous at all points on its length) is to centre a small window at each candidate edge-point, and count the number of other edge-points falling within that window. The candidate point can then be accepted as an edge-point or disregarded as a misclassification on the basis of the number of adjacent edge-points. Such a technique was used to smooth the raw results and produce the sets of points depicted in figure 8(b) and (c).

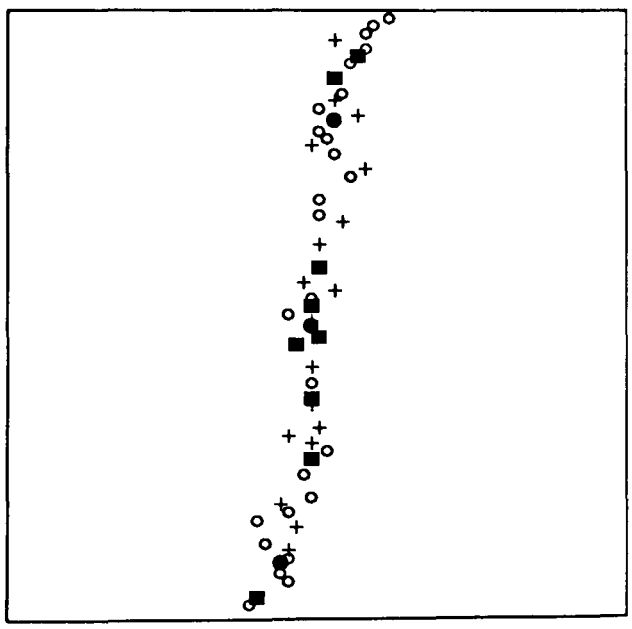


Fig 8(b) : 2 pts per window

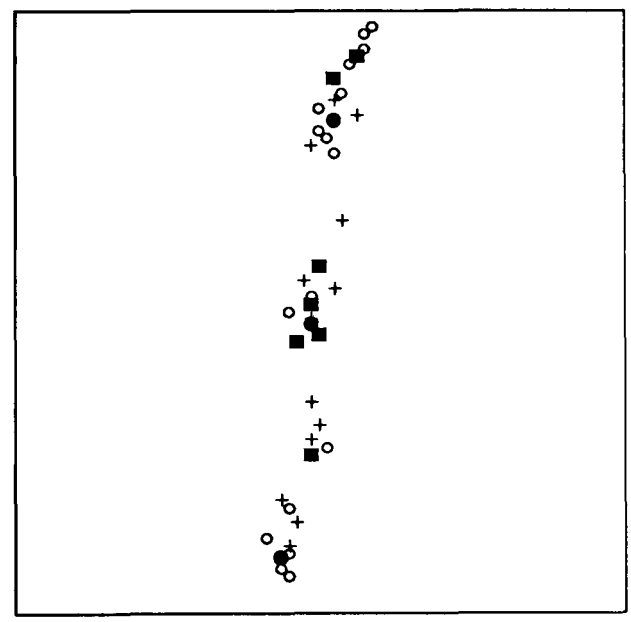


Fig 8(c) : 3 pts per window

In figure 8(b) a 7 x 7 pixel window was used in conjunction with an acceptance criterion of two points per window. For figure 8(c) the criterion was altered to three. In both cases, the additional CPU time involved in the smoothing procedure was of the order of 0.3 seconds. Thus the total processing time to produce figures 8(b) and (c) from the image was of the order of 1.3 seconds. Many of the misclassified points have been removed.

(2.6) Conclusions.

For this simple example, and despite the relatively high level of noise-corruption, the changepoint technique for edge-detection has performed both efficiently and effectively in its

least sophisticated form and under some fairly limiting assumptions. The performance of the technique for lower noise levels is demonstrated in figure 9. Figures 9(a) - (d) depict the results of row changepoint analysis for each of the eighty rows of the image in Figure 4 corrupted by Gaussian white-noise of differing variances producing Signal-Noise ratios (S.N.R.) 1.5, 2.0, 2.5, and 3.0 respectively. The results shown are "unsmoothed" (in the sense defined above), and again the analysis in each case took of the order of one second. Note the low number of edge-point misclassifications in Figures 9(c) and (d).

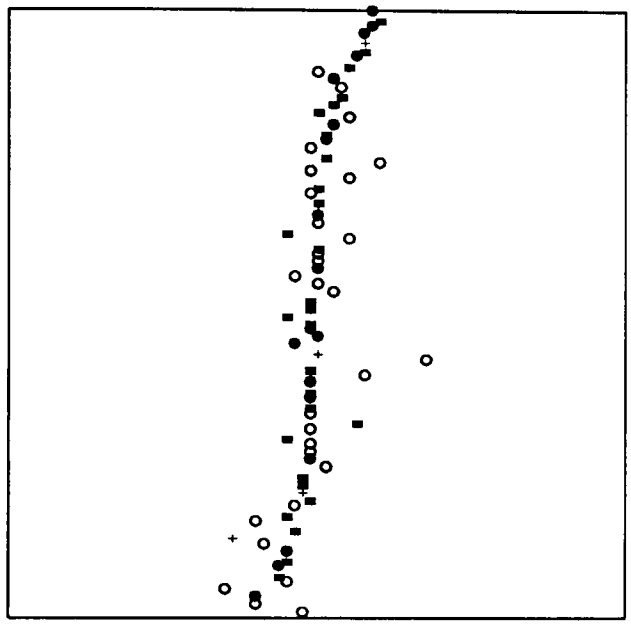


Fig 9(a) : S.N.R. 1.5

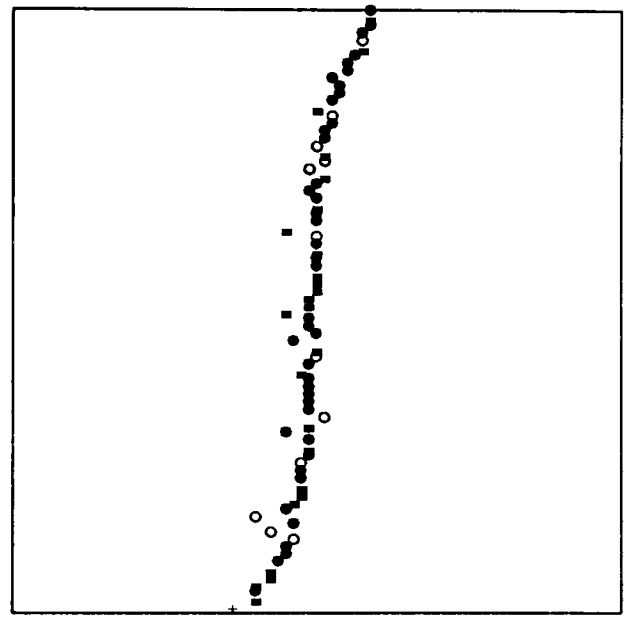


Fig 9(b) : S.N.R. 2.0

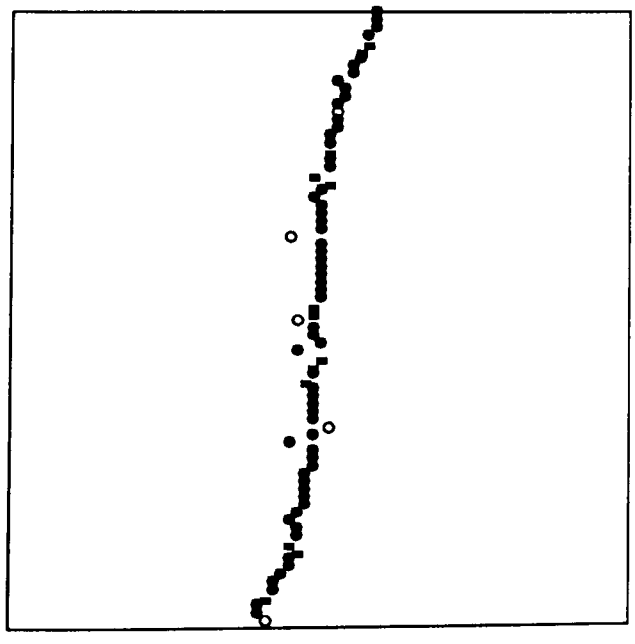


Fig 9(c) : S.N.R. 2.5

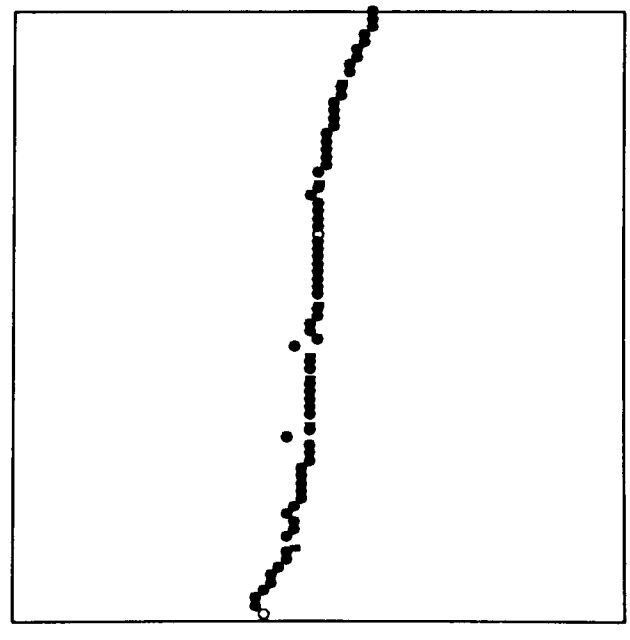


Fig 9(d) : S.N.R. 3.0

Changepoint techniques seem attractive and preferable to localised methods for several reasons. First, as mentioned above, the changepoint approach seems to reflect more adequately the nature of the edge-detection problem. Secondly, the localised methods - differencing,

filtering, convolution, local averaging - although intuitively reasonable to some extent are, in fact, generally quite arbitrary. The Bayesian changepoint approach at least has a basis in statistical decision theory. Thirdly, the localised methods generally depend heavily on expert input of parameters - threshold, window-width etc. - usually arrived at through detailed prior knowledge of the true scene and image. For the changepoint technique, as we have seen, at most only very general form of prior knowledge is required. Allied to the last two points, the localised methods return a real number at each cell and rely on thresholding to point up edge-regions, with no measure of uncertainty attached. The changepoint technique returns the most probable edge-position in the row concerned, in light of the data in that row and prior assumptions, with its associated probability. Finally, and perhaps most importantly, the changepoint technique out-performs the simple localised methods at comparable Signal-Noise ratios, as illustrated by a simple example, the results of which are depicted in figures 9(e)-(g).

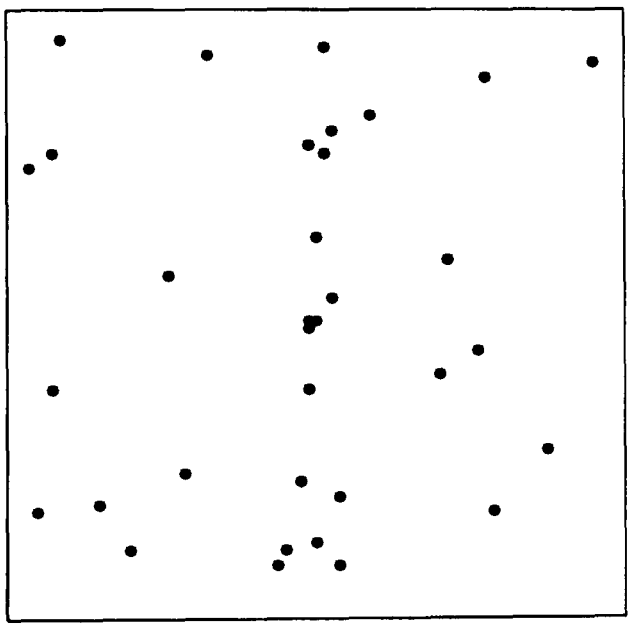


Fig 9(e) : threshold 5.0

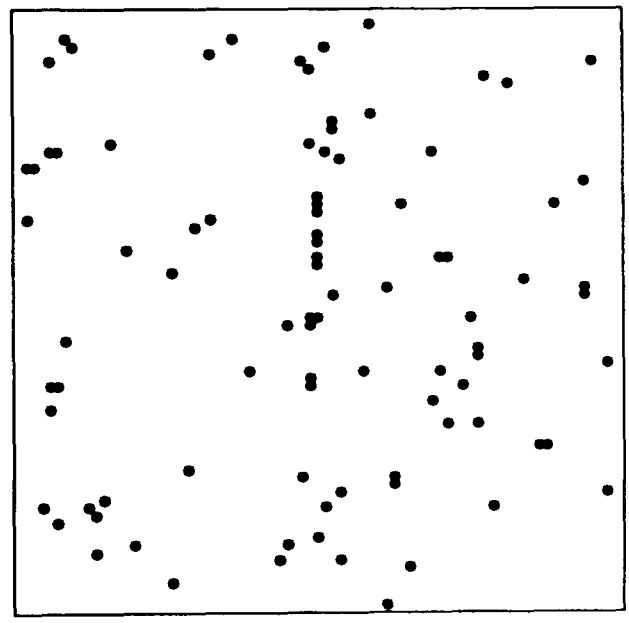


Fig 9(f) : threshold 4.5

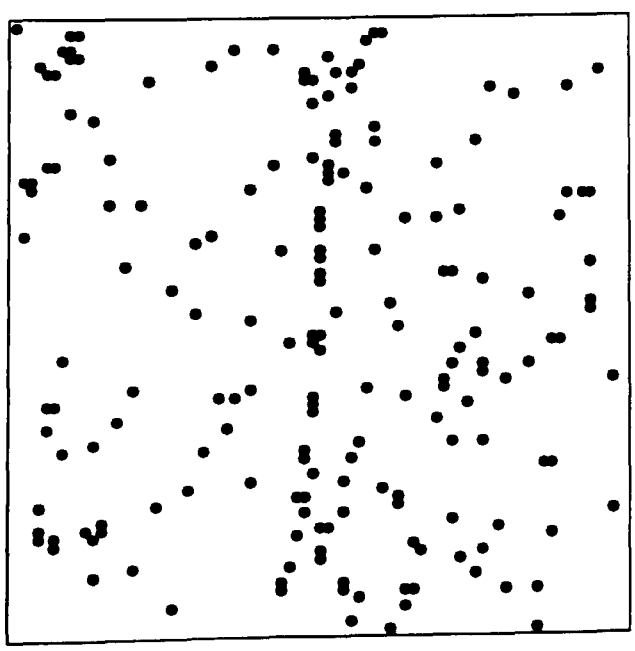


Fig 9(g) : threshold 4.0

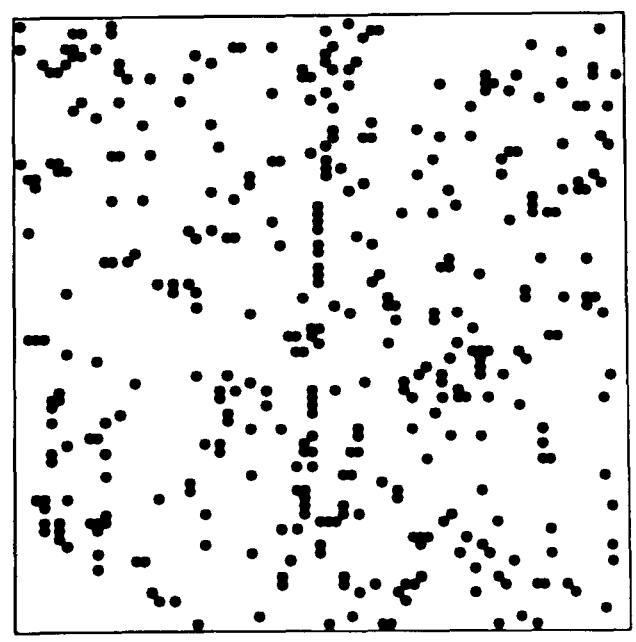


Fig 9(h) : threshold 3.5

Figures 9(e)-(g) depict the result of a simple localised edge-detection method, namely first differencing in two perpendicular directions, with threshold of acceptance ranging from 5.0 to 3.5. (i.e. we take first-order differences along the rows and columns of the image , evaluate the edge-magnitude at each cell as the square-root of the sum of the squares of these differences, and plot all points for which this magnitude is greater than an arbitrarily chosen threshold value.) The image was precisely the same as that used for figure 9(d), with a Signal-Noise ratio of 3.0. The results are clearly inferior to those obtained using changepoint analysis on the same image, and the processing time (0.5 seconds), although shorter, was not an adequate compensation. Also, the results obtained are not sufficiently robust to the choice of the (arbitrarily assigned) threshold value. We choose this example of localised edge-detection methods (not merely because of its inadequacy to deal with the edge-detection problem) because it compares in terms of processing time and prior knowledge of the true scene with the changepoint technique. More sophisticated (but still localised) methods compare unfavourably in terms of processing time.

(2.7) Extension of ideas.

We now seek to extend the above ideas concerning edge-detection via changepoint analysis in three general directions:

(1) More complex true scenes.

The simple example above demonstrated adequately the use of changepoint techniques in edge-detection. However, although it captured the nature of the edge-detection problem exactly (locating a discontinuity in some aspect of the image arising at the boundary between two non-localised features) it dealt with an idealised true scene. More realistic true scenes would involve convex objects, multiple regions, patterns, "thin" features etc.. We examine the performance of the changepoint techniques in each of these areas in chapter 3.

(2) Exploitation of spatial continuity.

As observed previously, the analysis of the simple image in Figure 2 did not take into account the fact that the edge in the true scene was spatially continuous, i.e. adjacent rows of the image were treated completely separately and independently. It would be reasonable to assume that, in light of the progress made generally in statistical image-processing, the introduction of the notion of local dependence and spatial continuity at the prior stage of one step of the procedure would improve results. We seek to adapt the changepoint technique in this way in chapter 4.

(3) Variation of image-formation and noise processes.

In our initial example, we assumed a simple linear form for the image-formation process, and that the noise process corrupted each pixel in the true scene identically and independently with Gaussian white-noise. This again is an idealised situation, and in chapter 5 we seek to extend the changepoint technique, specifically via choices of forms for prior distributions for the unknown parameters and functional dependencies, to handle more general situations.

Prior to this, however, we generalise the analysis of the true scene in figure 3 in two more straightforward ways. First, given our knowledge of the true scene concerning the general orientation of the edge with respect to the usual coordinate axes, a row analysis only seemed necessary. Practically we would have no such knowledge, and a column analysis would also be needed. Clearly, analysis of the image in any two orthogonal directions would suffice, in this sense, but for the moment, for convenience, we restrict attention to row plus column analyses, termed a "full" analysis.

Secondly, and in light of the previous discussion, it is desirable to incorporate a "no changepoint" or "no edge" alternative into the analysis. This task is straightforward. Under the alternative, we have $Y_1,....,Y_n$ identically distributed, and thus $[Y \mid r, \theta] \equiv [Y \mid \theta]$. For our initial example, therefore, with the Y_i's independently Normally distributed, this implies that (2.9) is replaced by

$$[Y \mid \theta] \propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{\tau}{2} \left[\sum_{i=1}^{n} (Y_i - \theta)^2 \right] \right\} \tag{2.12}$$

where now $\theta = [\theta, \tau]$. We proceed and specify $[\theta \mid \psi]$ as before but, as mentioned previously, we must be aware of the difference in dimensionality between the "changepoint" and "no changepoint" models when specifying improper prior forms. Such difficulties are avoided if proper prior distributions are specified. The careful use of (2.12), in conjunction with the specification of a prior for r of the second form as presented in section (2.3.3), will allow for the no changepoint model to be admitted. This should help in the removal of the misclassifications that appear in figure 8(a). The removal (or non-detection) of such points, or "false edges", is a familiar problem in classical edge-detection.

Chapter 3 : Analysis of Complex True Scenes.

The changepoint formulation of the edge-detection problem described in chapter 2 related to the analysis of images derived from true scenes in which the edge took the form of a single smooth curve. In this chapter, we seek to extend the formulation and adapt and improve the single changepoint technique so that more complex true scenes may be analysed in a similar fashion. The next natural class of true scenes that must be considered is that where the single simple edge is replaced by a single closed curve, so that S_θ is comprised of precisely two texture regions. We shall see that various amendments to our original implementation of the edge-detection scheme are necessary. The analysis of images derived from this class of true scenes is discussed in sections (3.1) - (3.5). Another class of true scenes of interest are multiple region or composite true scenes containing more than two texture varieties. A further extension of the single changepoint formulation is necessary for the analysis of such true scenes, and this extension and other issues are discussed in sections (3.6) and (3.7).

We begin by considering a very simple class of two-texture true scenes, which, despite their straightforward nature, allow us to illustrate the extension of our changepoint-based edge-detection technique.

(3.1) Convex object true scenes - circle.

We first consider a simple convex object, namely a circle, lying completely within the region S_θ. Figures 10(a) and (b) depict such a true scene and an image derived from the true scene and the image-formation and noise processes of the previous chapter.

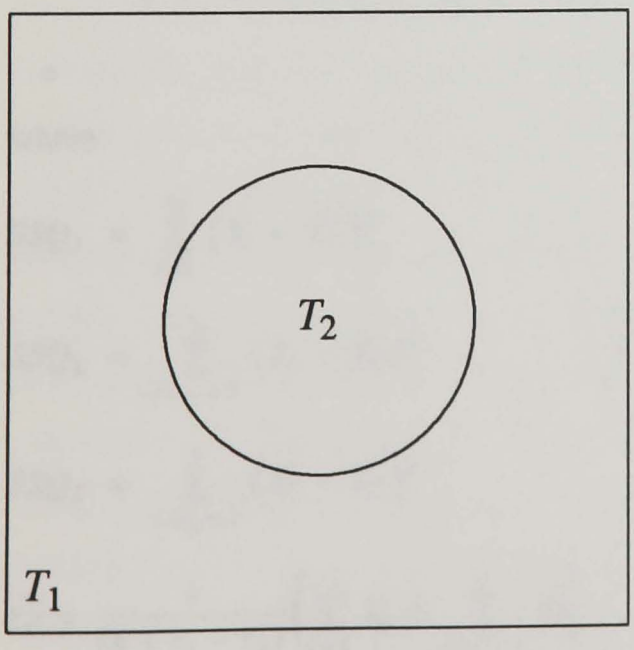


Fig 10(a) : true scene

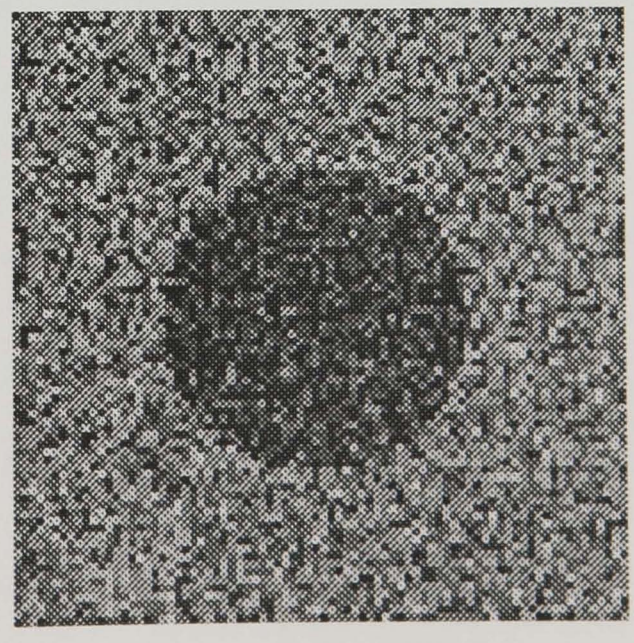


Fig 10(b) : image

This is a familiar test image: see for example Peli and Malah (1982). Two features are

immediately apparent. First, many rows and columns in the true scene contain no edge. Secondly, the remaining rows and columns contain two edges, or points at which pixels in different texture regions are adjacent. This feature will of course be present in all true scenes that contain single convex objects, and thus due to the topological differences from our initial example, the edge-detection problem is fundamentally different.

As we saw in section (2.3), it is possible to generalise the single-changepoint formulation to a k changepoint one. Again consider $Y = (Y_1, \ldots, Y_n)$, a sequence of random variables. Let $r_1, \ldots, r_k$ be the unknown positions of the k changepoints, and θ and ψ be as previously defined. Then the joint posterior density of the variables $r_1, \ldots, r_k$ is given by

$$\begin{aligned} [r_1, \ldots, r_k \mid Y, \psi] &\propto [Y \mid r_1, \ldots, r_k, \psi]\,[r_1, \ldots, r_k] \\ &= \int [Y \mid r_1, \ldots, r_k, \theta]\,[\theta \mid \psi]\,[r_1, \ldots, r_k] \end{aligned} \tag{3.1}$$

and we can make inference about the positions of the unknown changepoints via this posterior distribution. In the edge-detection context, $k = 2$ for single convex objects. Also, note that the data elements indexed by 1 to r_1 and r_2+1 to n are identically distributed conditional on θ. For example, under the same image-formation and noise processes as in (2.1), we have that $Y_1, \ldots, Y_{r_1+1}$ and $Y_{r_2+1}, \ldots, Y_n$ are distributed as $N(\theta_1, \sigma^2)$, and that $Y_{r_1+1}, \ldots, Y_{r_2}$ are distributed as $N(\theta_2, \sigma^2)$, with the Y_i's independent. In this case, the joint posterior distribution for (r_1, r_2) conditional on Y under the **exactly** two changepoint model using the non-informative prior specification (2.10) for $\theta = (\theta_1, \theta_2, \sigma)$ is thus given by

$$[r_1, r_2 \mid Y, \psi] \propto \{(r_2 - r_1)(n + r_1 - r_2)\}^{-1/2} \{SSQ_1 + SSQ_2 + SSQ_3\}^{-n/2} \tag{3.2}$$

where

$$SSQ_1 = \sum_{i=1}^{r_1} (Y_i - \overline{Y_C})^2$$

$$SSQ_2 = \sum_{i=r_1+1}^{r_2} (Y_i - \overline{Y_D})^2$$

$$SSQ_3 = \sum_{i=r_2+1}^{n} (Y_i - \overline{Y_C})^2$$

$$\overline{Y_C} = \frac{1}{(n + r_1 - r_2)} \left(\sum_{i=1}^{r_1} Y_i + \sum_{i=r_2+1}^{n} Y_i \right)$$

$$\overline{Y_D} = \frac{1}{(r_2 - r_1)} \sum_{i=r_1+1}^{r_2} Y_i$$

We could evaluate the posterior distribution (3.2) for all pairs (r_1, r_2) and locate the joint posterior mode. Again, a no changepoint alternative can be considered. Note also that this formulation restricts attention to locating **exactly** two changepoints and thus, in this its simplest form, cannot cope with one changepoint sequences. However, with a little care, this problem can be overcome by letting r_2, without loss of generality presumed greater than r_1, equal n (here again we must recognise that the no, one and two changepoint models are models of different dimension). If we have sufficient prior knowledge of the true scene (i.e that it entirely contains a convex object) we need not entertain the one changepoint alternative.

Thus we may accommodate more complex structures than the simple edge of our original example. However, in the absence of relatively detailed prior knowledge of the true scene, evaluation of the changepoint posterior probabilities for a sequence under the hypothesis of greater than one changepoint is undesirable, principally due to the amount of computation involved. For example, rough calculations indicate that the amount of computation required for evaluation of probabilities for the one changepoint model increases linearly with n, whereas the amount of computation for the two changepoint model increases with n^2 (and, similarly, with n^k for the k changepoint model). Clearly, this is prohibitive, and as one motivation for the development of edge-detection routines is that they should operate, at least in part, as pre-processing operations, they should not entail large amounts of computation. We shall return to this theme on many occasions throughout this thesis. In the light of the above considerations, we now develop other exact and approximate methods for the multiple changepoint/edge detection problem.

(3.2) Approximation in multiple-changepoint models.

First, it is important to attempt to understand the precise nature of the Bayesian changepoint detection technique. Consider again the single changepoint posterior density, to be evaluated for a sequence Y. For each r, consider the "left" sub-sequence $Y_L = Y_1, \ldots, Y_r$, and the "right" sub-sequence $Y_R = Y_{r+1}, \ldots, Y_n$. To evaluate the posterior probability for r, we presume that (1) the elements in the left sub-sequence are identically distributed, (2) the elements in the right sub-sequence are identically distributed, and (3) the distributions involved in (1) and (2) are different. We integrate the likelihoods in (1) and (2) with respect to unknown population parameters, to obtain the marginal distributions $[Y_L \mid \psi]$ and $[Y_R \mid \psi]$. Given a realisation $y = (y_L, y_R)$ we would expect these marginal distributions to attain their maximum when $r = r^*$, the true changepoint position, as away from r^* neither (1) or (2) will be accurate. Thus (in expectation at least) the technique will "always" identify the true changepoint. Now consider a two changepoint sequence, with changepoints at (r_1^*, r_2^*), $1 < r_1^* < r_2^* < n$, and the behaviour of the changepoint posterior distribution under the one changepoint hypothesis. Again, a "high" posterior probability will result when (1), (2), and (3) hold together. For the two changepoint sequence, however, one of (1) and (2) will always

be inaccurate, to a greater or lesser degree, as measured (in some way) by the marginal probability attained. But, on considering the behaviour in the vicinity of r_1^* and r_2^*, we may expect a localised mode in the posterior distribution at $r = r_1^*$ and $r = r_2^*$. More superficially, r_1^* and r_2^* mark abrupt changes in the nature of (sampling and marginal) probability distributions, and thus we might expect both to be detected by the one changepoint posterior distribution. We shall discuss these points in more detail below, after investigating the behaviour of the one changepoint posterior distribution under an incorrect model specification by means of a simulation study.

Thus, we may expect the single changepoint posterior calculation to assist in the identification of changepoint positions in a two or more changepoint sequence in two ways. First, we might expect that the mode of the distribution should frequently lie at one of the true changepoint positions. Secondly, we might also expect local modes at or near both of the two of the true changepoint positions. In practice, the results are encouraging. Figure 11 depicts the results of 1000 simulations of two changepoint sequences of total length 80 at a fixed Signal-Noise ratio of 3.0, with average posterior probability under the one changepoint hypothesis and prior assumptions leading to (2.11) plotted on the vertical scale. The positions of the two changepoints were chosen to be symmetrical about the sequence mid-point, and the inter-changepoint distance was decreased over the series from (a) to (d).

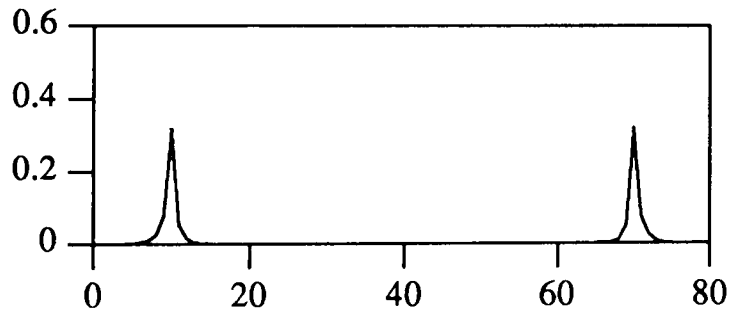


Fig 11(a) : Changepoints at 10, 70

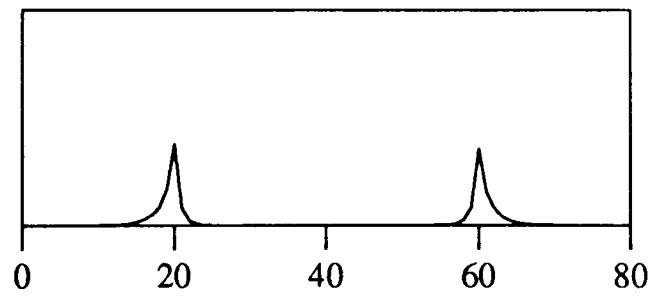


Fig 11(b) : Changepoints at 20, 60

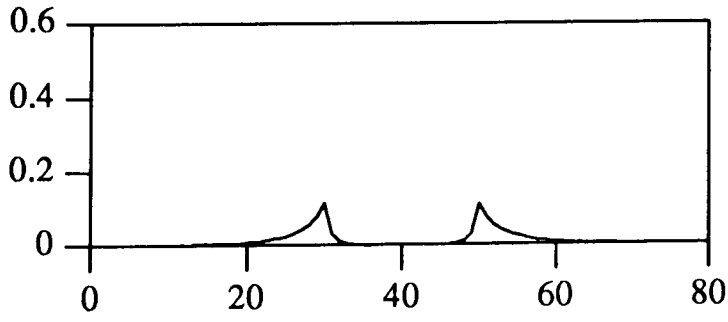


Fig 11(c) : Changepoints at 30, 50

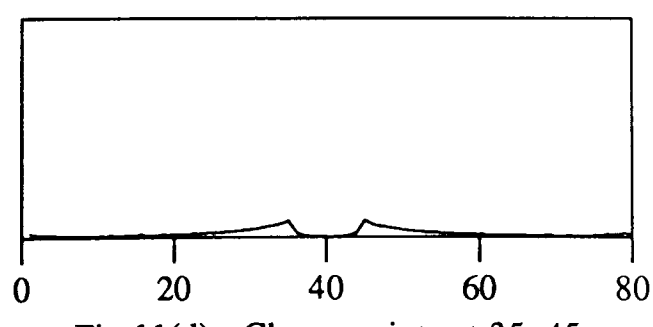


Fig 11(d) : Changepoints at 35, 45

It is clear that, in these idealised situations, the single changepoint posterior assists in identification in two changepoint sequences. The results are most instructive when the inter-changepoint distance is large, as would have been predicted in the light of the above discussion. The behaviour of the single changepoint posterior at lower Signal-Noise ratios is depicted in figure 12. The changepoint positions were fixed at 20 and 60, and the Signal-Noise

ratio decreased from 2.5 to 1.0 in intervals of 0.5 over the series (a) - (d).

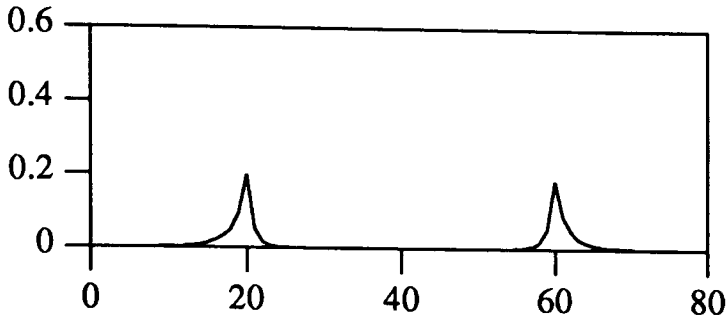


Fig 12(a) : S.N.R 2.5

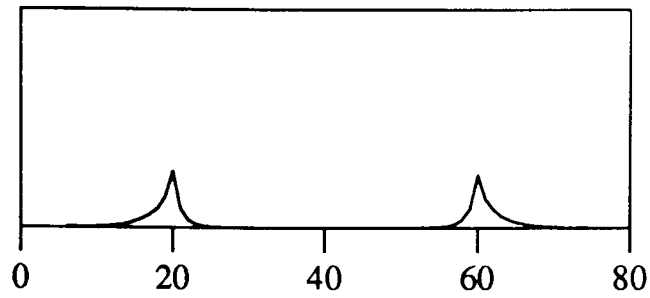


Fig 12(b) : S.N.R 2.0

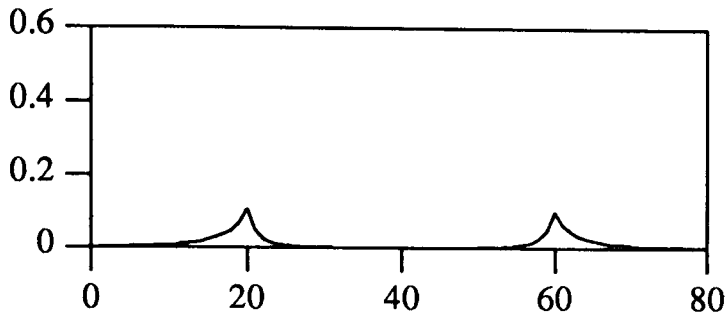


Fig 12(c) : S.N.R 1.5

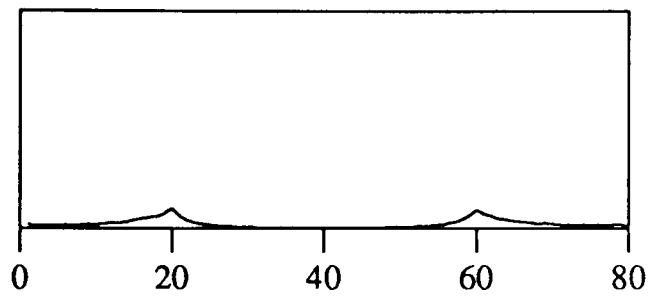


Fig 12(d) : S.N.R 1.0

Note the decrease in modal values and the increase in spread of probability as Signal-Noise ratio decreases. This is in line with our previous experience with changepoint posterior probabilities for one changepoint sequences.

Finally, we introduce asymmetry into the changepoint positions. 1000 simulations were carried out for various asymmetric combinations of r_1^* and r_2^* at fixed Signal-Noise level 3.0. The results are shown in figure 13.

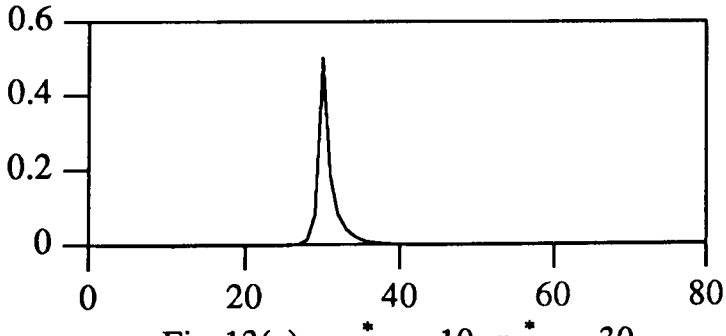


Fig 13(a) : $r_1^* = 10, r_2^* = 30$

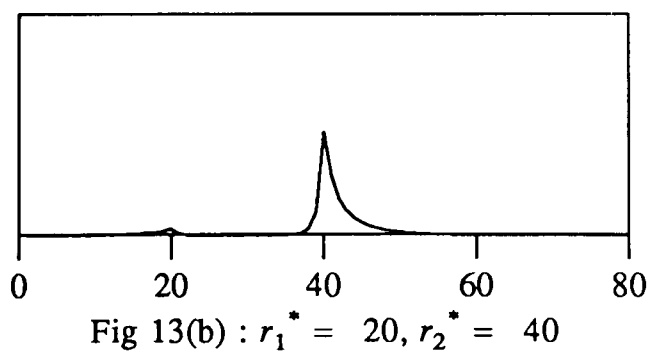


Fig 13(b) : $r_1^* = 20, r_2^* = 40$

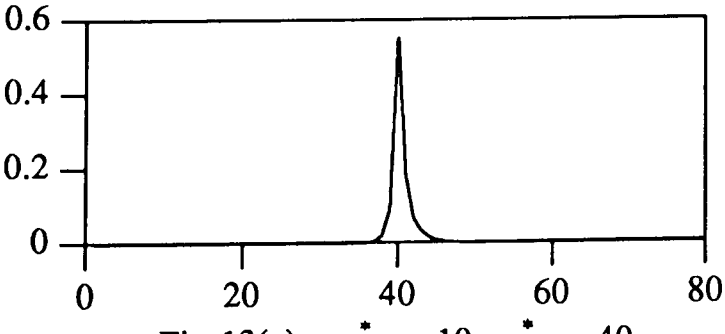


Fig 13(c) : $r_1^* = 10, r_2^* = 40$

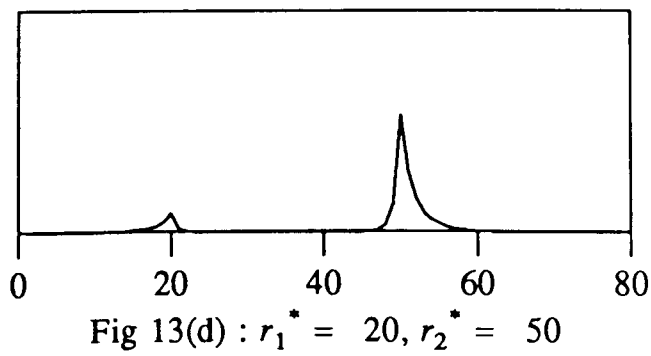


Fig 13(d) : $r_1^* = 20, r_2^* = 50$

Figure 13 illustrates a problem. Whereas (b) and (d) clearly depict posterior distributions with two distinct modes, (a) and (c) depict seemingly unimodal distributions. All four distributions were generated with relation to two changepoint sequences. In light of these

experimental results, it seems plausible that the one changepoint posterior distribution will deal adequately with two changepoint sequences in many cases where the inter-changepoint distance is large, in the sense that we may associate modes in the distribution with true changepoint positions. We now attempt some formal justification.

(3.2.1) General investigation of the one changepoint approximation.

Consider a two changepoint sequence Y with changepoints at r_1^* and r_2^*, and the posterior distribution $[r \mid Y, \psi]$ derived under a one changepoint hypothesis. Consider the ratio of the posterior probabilities at r and r_1^*. From (2.3) we have

$$\frac{[r \mid Y, \psi]}{[r_1^* \mid Y, \psi]} = \frac{[Y \mid r, \psi]\,[r]}{[Y \mid r_1^*, \psi]\,[r_1^*]}$$

$$= \frac{[Y_1, \ldots, Y_r \mid \psi_1]\,[Y_{r+1}, \ldots, Y_n \mid \psi_2]}{[Y_1, \ldots, Y_{r_1^*} \mid \psi_1]\,[Y_{r_1^*+1}, \ldots, Y_n \mid \psi_2]} \qquad (3.3)$$

assuming equal prior probabilities for r and r_1^*, and that beliefs about (θ_1, θ_2) are *a priori* independent, with $\psi = (\psi_1, \psi_2)$. For $r < r_1^*$, we may factorise the numerator as

$$[Y_1, \ldots, Y_r \mid \psi_1]\,[Y_{r+1}, \ldots, Y_{r_1^*} \mid \psi_2, Y_{r_1^*+1}, \ldots, Y_n]\,[Y_{r_1^*+1}, \ldots, Y_n \mid \psi_2]$$

and the denominator as

$$[Y_1, \ldots, Y_r \mid \psi_1]\,[Y_{r+1}, \ldots, Y_{r_1^*} \mid \psi_1, Y_1, \ldots, Y_r]\,[Y_{r_1^*+1}, \ldots, Y_n \mid \psi_2]$$

Thus, from (3.3),

$$\frac{[r \mid Y, \psi]}{[r_1^* \mid Y, \psi]} = \frac{[Y_{r+1}, \ldots, Y_{r_1^*} \mid \psi_2, Y_{r_1^*+1}, \ldots, Y_n]}{[Y_{r+1}, \ldots, Y_{r_1^*} \mid \psi_1, Y_1, \ldots, Y_r]} \qquad (3.4)$$

An interpretation of (3.4) is as follows. The numerator is a measure of how well we could "predict" $Y_{r+1}, \ldots, Y_{r_1^*}$ from $Y_{r_1^*+1}, \ldots, Y_n$, whereas the denominator is a measure of how well we could "predict" $Y_{r+1} \ldots, Y_{r_1^*}$ from $Y_1, \ldots, Y_r$. Clearly, therefore, in expectation this ratio will be less than one for all $r < r_1^*$, given the distribution of the elements of Y. Now consider $r > r_1^*$. The numerator in (3.3) may be factorised as

$$[Y_1, \ldots, Y_{r_1^*} \mid \psi_1]\,[Y_{r_1^*+1}, \ldots, Y_r \mid \psi_1, Y_1, \ldots, Y_{r_1^*}]\,[Y_{r+1}, \ldots, Y_n \mid \psi_2]$$

and the denominator may be factorised as

$$[Y_1,\ldots,Y_{r_1^*}\mid\psi_1][Y_{r_1^*+1},\ldots,Y_r\mid\psi_2,Y_{r+1},\ldots,Y_n][Y_{r+1},\ldots,Y_n\mid\psi_2]$$

and thus, again, we re-express (3.3) as

$$\frac{[r\mid Y,\psi]}{[r_1^*\mid Y,\psi]} = \frac{[Y_{r_1^*+1},\ldots,Y_r\mid\psi_1,Y_1,\ldots,Y_{r_1^*}]}{[Y_{r_1^*+1},\ldots,Y_r\mid\psi_2,Y_{r+1},\ldots,Y_n]} \tag{3.5}$$

We may interpret (3.5) in a similar way, and again conclude that, provided r is near to r_1^*, the denominator will be larger than the numerator, and the ratio less than one. Hence we would expect a local maximum at $r = r_1^*$. A similar argument can be applied with respect to r_2^*. Thus it is reasonable to expect modes in the posterior distribution at the true changepoints.

Now consider the case where beliefs about (θ_1,θ_2) are *a priori* dependent, or indeed we assume *a priori* that θ_1 and θ_2 have a common element, as in section (2.4). Consider the ratio in (3.5). This now becomes

$$\frac{[r\mid Y,\psi]}{[r_1^*\mid Y,\psi]} = \frac{[Y_1,\ldots,Y_r\mid\psi][Y_{r+1},\ldots,Y_n\mid\psi,Y_1,\ldots,Y_r]}{[Y_1,\ldots,Y_{r_1^*}\mid\psi][Y_{r_1^*+1},\ldots,Y_n\mid\psi,Y_1,\ldots,Y_{r_1^*}]} \tag{3.6}$$

Again we examine the behaviour of this ratio in the vicinity of r_1^*. Consider $r < r_1^*$. We may factorise the second term in the numerator as

$$[Y_{r+1},\ldots,Y_{r_1^*}\mid\psi,Y_1,\ldots,Y_r,Y_{r_1^*+1},\ldots,Y_n][Y_{r_1^*+1},\ldots,Y_n\mid\psi,Y_1,\ldots,Y_r]$$

and the first term in the denominator as

$$[Y_1,\ldots,Y_r\mid\psi][Y_{r+1},\ldots,Y_{r_1^*}\mid\psi,Y_1,\ldots,Y_r].$$

Cancelling the term $[Y_1,\ldots,Y_r\mid\psi]$, we obtain

$$\frac{[r\mid Y,\psi]}{[r_1^*\mid Y,\psi]} = \frac{[Y_{r+1},\ldots,Y_{r_1^*}\mid\psi,Y_1,\ldots,Y_r,Y_{r_1^*+1},\ldots,Y_n]}{[Y_{r+1},\ldots,Y_{r_1^*}\mid\psi,Y_1,\ldots,Y_r]} \cdot \frac{[Y_{r_1^*+1},\ldots,Y_n\mid\psi,Y_1,\ldots,Y_r]}{[Y_{r_1^*+1},\ldots,Y_n\mid\psi,Y_1,\ldots,Y_{r_1^*}]} \tag{3.7}$$

and can interpret this expression in the same way as above. We would expect the first term in (3.7) to be less than one if r_1^* and $n - r_1^*$ were "large enough", due to the "corrupting" presence of $Y_{r_1^*+1},\ldots,Y_n$ in the numerator. Also, we would expect the second term in (3.7) to be approximately equal to one, again provided $n - r_1^*$ was large enough. Thus we would again

expect the ratio to be less than one. Now consider $r > r_1^*$. We now factorise the first term in the numerator of (3.6) as

$$[Y_1,\ldots, Y_{r_1^*} \mid \psi]\,[Y_{r_1^*+1},\ldots, Y_r \mid \psi, Y_1,\ldots, Y_{r_1^*}]$$

and the second term in the denominator

$$[Y_{r_1^*+1},\ldots, Y_r \mid \psi, Y_1,\ldots, Y_{r_1^*}, Y_{r+1},\ldots, Y_n]\,[Y_{r+1},\ldots, Y_n \mid \psi, Y_1,\ldots, Y_{r_1^*}]\ .$$

Cancelling the term $[Y_1,\ldots, Y_{r_1^*} \mid \psi]$, we obtain

$$\frac{[r \mid Y, \psi]}{[r_1^* \mid Y, \psi]} = \frac{[Y_{r_1^*+1},\ldots, Y_r \mid \psi, Y_1,\ldots, Y_{r_1^*}]}{[Y_{r_1^*+1},\ldots, Y_r \mid \psi, Y_1,\ldots, Y_{r_1^*}, Y_{r+1},\ldots, Y_n]} \cdot \frac{[Y_{r+1},\ldots, Y_n \mid \psi, Y_1,\ldots, Y_r]}{[Y_{r+1},\ldots, Y_n \mid \psi, Y_1,\ldots, Y_{r_1^*}]}\ . \tag{3.8}$$

Again, we expect the first term to be less than one and the second approximately to be one, provided r_1^* and $n - r_1^*$ are large enough. Thus the ratio of posterior probabilities is less than one for all prospective changepoint positions in the vicinity of the true changepoint, and so we would expect a mode in the posterior distribution at the true changepoint r_1^*. Similar arguments lead us to expect another (local) mode at r_2^*. We conclude therefore that, in many cases, for two changepoint sequences, local modes in one changepoint posterior distributions will be good indicators of true changepoint positions. In the special case where $r_1^* = n - r_2^*$ and the data sequence exhibits symmetry, as in figures 11 and 12, it is also easy to see that $[r \mid Y, \psi]$ takes the same value (in expectation) at r_1^* and r_2^*, provided that the prior specification for the unknown elements of θ is exchangeable, by simple re-ordering of the subscripts of the Y_i.

Thus in the general case it appears that analysis of a two changepoint data sequence under a one changepoint modelling assumption produces potentially useful results. More specifically, in the case of the posterior distribution (2.11) we can obtain some more rigorous results.

(3.2.2) Investigation of the one changepoint approximations under normality.

We consider specifically the posterior distribution (2.11) derived under the a non-informative prior specification for θ when n is relatively large. The sum of squares term dominates the behaviour of $[r \mid Y, \psi]$ in this case, and consequently the maximum value of $[r \mid Y, \psi]$ occurs when $\sum_{i=1}^{r} (Y_i - \overline{Y_A})^2 + \sum_{i=r+1}^{n} (Y_i - \overline{Y_B})^2$ is a minimum (resulting in an

estimate of r identical to the maximum-likelihood estimate for the same sampling distribution). We now study the behaviour of this sum of squares in expectation when the distribution of Y is known to be that of a two changepoint sequence.

Without loss of generality (or after some suitable transformation) we assume that, for some α, Y is distributed such that

$$Y_i \sim \begin{cases} N(0,1) & 1 \leq i \leq r_1^* \\ N(\alpha,1) & r_1^* + 1 \leq i \leq r_2^* \\ N(0,1) & r_2^* + 1 \leq i \leq n \end{cases}$$

where as before r_1^* and r_2^* are the true changepoint positions. Now, $\sum_{i=1}^{r} (Y_i - \overline{Y_A})^2 + \sum_{i=r+1}^{n} (Y_i - \overline{Y_B})^2$ can be expressed in the form $Y^T Q Y$, where

$$Q = \begin{bmatrix} A_r & 0 \\ 0 & A_{n-r} \end{bmatrix}$$

where

$$[A_p]_{ij} = \begin{cases} 1 - \dfrac{1}{p} & i = j \\ \dfrac{-1}{p} & i \neq j \end{cases} .$$

Now suppose $\mathrm{E}[Y] = \mu$ and $\mathrm{V}[Y] = \Sigma$. Then, by a well known result,

$$E[Y^T Q Y] = \mu^T Q \mu + tr(Q\Sigma) . \tag{3.9}$$

In this case, where $\Sigma = I_n$,

$$tr(Q\Sigma) = n-2 \tag{3.10}$$

and, after some algebra,

$$\mu^T Q \mu = \sum_{i=1}^{r} (\mu_i - \overline{\mu_A})^2 + \sum_{i=r+1}^{n} (\mu_i - \overline{\mu_B})^2 = \lambda_r , \tag{3.11}$$

say, where $\overline{\mu_A} = \dfrac{1}{r} \sum_{i=1}^{r} \mu_i$ and $\overline{\mu_B} = \dfrac{1}{(n-r)} \sum_{i=r+1}^{n} \mu_i$. We observe at this point that under the distributional assumptions made above, $\sum_{i=1}^{r} (Y_i - \overline{Y_A})^2 + \sum_{i=r+1}^{n} (Y_i - \overline{Y_B})^2$ has a non-central chi-squared distribution with non-centrality parameter λ_r.

Hence, in the light of (3.9) to (3.11), the sum of squares is minimised in expectation when λ_r is minimised. After some algebra, it can be shown that

$$\lambda_r = \begin{cases} (r_2^* - r_1^*)\left(1 - \dfrac{r_2^* - r_1^*}{n-r}\right)\alpha^2 & 1 \leq i \leq r_1^* \\ \left[r_1^*\left(1 - \dfrac{r_1^*}{r}\right) + (n - r_2^*)\left(1 - \dfrac{n - r_2^*}{n-r}\right)\right]\alpha^2 & r_1^* \leq i \leq r_2^* \\ (r_2^* - r_1^*)\left(1 - \dfrac{r_2^* - r_1^*}{r}\right)\alpha^2 & r_2^* \leq i \leq n \end{cases} \qquad (3.12)$$

We now consider the behaviour of λ_r over three separate ranges.

(1) $1 \leq r \leq r_1^*$.

For increasing r, it is clear that λ_r is monotonically decreasing for r in this range, and hence the minimum is attained at $r = r_1^*$.

(2) $r_1^* \leq r \leq r_2^*$

First consider the difference $\lambda_{r_1^*+1} - \lambda_{r_1^*}$. It is easy to see that

$$\lambda_{r_1^*+1} - \lambda_{r_1^*} = \left[\frac{r_1^{*2}}{r_1^*(r_1^* + 1)} - \frac{(n - r_2^*)^2}{(n - r_1^*)(n - r_1^* - 1)}\right]\alpha^2 \ . \qquad (3.13)$$

Provided r_1^* is relatively large, the first term in the bracket is approximately equal to one. Also, provided $(n - r_1^*)$ is relatively large compared to $(n - r_2^*)$, ($r_2^* - r_1^*$ is large) the second term is appreciably less than one. Thus this difference is greater than zero, and $\lambda_{r_1^*+1} > \lambda_{r_1^*}$. It can be shown in a similar fashion that $\lambda_{r+1} > \lambda_r$ for all r in the vicinity of r_1^*, and thus that λ_r is locally minimised at r_1^* for r in this range. Note that the magnitude of the difference $\lambda_{r_1^*+1} - \lambda_{r_1^*}$ is dependent on α^2.

Now consider the difference $\lambda_{r_2^*-1} - \lambda_{r_2^*}$. Again, it is easy to see that

$$\lambda_{r_2^*-1} - \lambda_{r_2^*} = \left[\frac{(n - r_2^*)^2}{(n - r_2^*)(n - r_2^* + 1)} - \frac{r_1^{*2}}{r_2^*(r_2^* - 1)}\right]\alpha^2 \ . \qquad (3.14)$$

Making the same approximations as above, the first term is approximately one and the second less than one, again making the difference greater than zero, and $\lambda_{r_2^*-1} > \lambda_{r_2^*}$. It can be shown that $\lambda_{r-1} > \lambda_r$ for all r in the vicinity of r_2^* and thus that λ_r is locally minimised at r_2^* for r in this range. Again note the influence of α^2 on the magnitude of this difference. The exact behaviour of λ_r for r over the whole of this range may be investigated in the same

way - we would expect some form of quadratic behaviour.

(3) $r_2^* \leqslant r \leqslant n$

For increasing r, it is clear that λ_r is monotonically increasing for r in this range, and hence the minimum is attained at $r = r_2^*$.

We have shown that λ_r is locally minimised at r_1^* and r_2^*. But more aspects become apparent on further inspection. First, if we take r_1^* and r_2^* such that $r_1^* = n - r_2^*$ (symmetric) then the behaviour of λ_r is symmetric - hence the behaviour of $[r \mid Y, \psi]$ depicted in figures 9 and 10. Secondly, if we vary the inter-changepoint distance $r_2^* - r_1^*$, then it is obvious from the form of λ_r that the minima at r_1^* and r_2^* will be less marked as the inter-changepoint distance decreases - hence the behaviour depicted in figure 9. Thirdly, we have noted the role played by α in the above, namely, as α^2 decreases the minima at r_1^* and r_2^* will be less marked. For this example, given the sampling distribution, α can be equated with Signal-Noise ratio - hence the behaviour depicted in figure 10. Finally, we note that

$$\lambda_{r_1^*} - \lambda_{r_2^*} = (r_2^* - r_1^*)^2 \left(\frac{1}{n - r_2^*} - \frac{1}{r_1^*} \right) \alpha^2 \qquad (3.15)$$

so that if $n - r_2^* > r_1^*$ then $\lambda_{r_1^*} > \lambda_{r_2^*}$, and vice-versa - hence the behaviour depicted in figure 11. Also, it is clear that the magnitude of $\lambda_{r_1^*} - \lambda_{r_2^*}$ varies as r_1^* and r_2^* vary, for fixed $r_2^* - r_1^*$ - hence the difference between figures 11(a) and (b), and figures 11(c) and (d). Thus we can adequately explain and understand the behaviour of the posterior distribution (2.11) in expectation, for large n.

We conclude this investigation by studying the behaviour of λ_r under a more general form for the distribution of Y. Suppose now that the magnitude of the change in mean-level at r_1^* is not equal to the change in mean-level at r_2^*, i.e.

$$Y_i \sim \begin{cases} N(0,1) & 1 \leqslant i \leqslant r_1^* \\ N(\alpha,1) & r_1^* + 1 \leqslant i \leqslant r_2^* \\ N(\beta,1) & r_2^* + 1 \leqslant i \leqslant n \end{cases}$$

for some α, β. Note that if $\beta = \alpha$ (or $\alpha = 0$) then there is effectively no changepoint at r_2^* (or r_1^*), and we revert simply to a one changepoint sequence. Also, if $\beta = 0$ we revert to the case above.

Using (3.9) to (3.11), and after some algebra, it can be shown that

$$\lambda_r = (r_2^* - r_1^*)\left(1 - \frac{r_2^* - r}{n-r}\right)(\alpha-\beta)^2 + \left(\frac{r_1^* - r}{n-r}\right)[(r_2^* - r_1^*)\alpha^2 + (n - r_2^*)\beta^2]$$

$$(1 \leqslant r \leqslant r_1^*)$$

$$\lambda_r = (n - r_2^*)\left(1 - \frac{n - r_2^*}{n-r}\right)(\alpha-\beta)^2 + r_1^*\left(1 - \frac{r_1^*}{r}\right)\alpha^2 \qquad (r_1^* \leqslant r \leqslant r_2^*)$$

$$\lambda_r = (r_2^* - r_1^*)\left(1 - \frac{r_2^*}{r}\right)(\alpha-\beta)^2 + \frac{r_1^*}{r}[(r_2^* - r_1^*)\alpha^2 + (r - r_2^*)\beta^2]$$

$$(r_2^* \leqslant r \leqslant n)$$

Clearly, the behaviour of λ_r is less straightforward in this case. However it can be shown that λ_r is monotonically decreasing for $1 \leqslant r \leqslant r_1^*$, monotonically increasing for $r_2^* \leqslant r \leqslant n$, provided r_1^*, and $r_2^* - r_1^*$ are relatively large compared to n, and locally minimised at r_1^* and r_2^* provided $(\alpha-\beta)^2$ is small compared to α^2. Note that, in particular,

$$\lambda_{r_1^*} - \lambda_{r_2^*} = (r_2^* - r_1^*)\left[\frac{n - r_2^*}{n - r_1^*}(\alpha-\beta)^2 - \frac{r_1^*}{r_2^*}\alpha^2\right] \qquad (3.16)$$

and so we would expect posterior modal values of (2.11) to vary with $\alpha^2, (\alpha-\beta)^2, r_1^*$ and r_2^*. Note also, as suggested above, that by setting $\beta = \alpha$, we may derive expectation results for the posterior distribution (2.11) under the assumption of one changepoint. Finally, note that we can infer only the qualitative behaviour of (2.11) in expectation from the results above, due to the non-linear relationship between (2.11) and $\sum_{i=1}^{r}(Y_i - \overline{Y_A})^2 + \sum_{i=r+1}^{n}(Y_i - \overline{Y_B})^2$ - however as this relationship involves logarithmic functions, it is reasonable that we should derive our results concerning λ_r using difference methods.

(3.2.3) Conclusions.

We have seen that the one changepoint posterior distribution (2.11) will often provide an adequate means of analysis for a two changepoint sequence; that is, the mode of (2.11) coincides with one of the true changepoint positions. We have also seen that (2.11) is generally bimodal in these circumstances, with the modes corresponding to the true changepoint positions. Thus we might feel justified in trying to locate and record the pair of modal positions for such sequences. However, figure 13 demonstrates that this strategy may not be easily implemented. There, due to the asymmetry of the sequences, one mode dominates the other, and so search techniques may be subverted. Also, as we have tried to formulate the

changepoint/edge detection problem using decision-theoretic ideas, it is difficult to justify a two mode search in that framework (i.e. we choose the mode of the one changepoint posterior distribution as our estimator as it minimises Bayes risk with respect to a pre-specified loss function - it is not easy to see how the choice of localised modes can be justified in this way). Thus, in general, we merely locate and record posterior modal position and probability.

We now have some practical experience and theoretical understanding of the behaviour of the one changepoint posterior distribution given a realisation from a two changepoint sequence. Most importantly, we have seen that we may associate posterior modes with true changepoint positions. This is of great importance in the edge-detection context, as it implies that we may analyse more complex true scenes and images such as those in figure 10 in exactly the same way that we analysed the simple example of figure 3 (i.e. using one changepoint posterior distributions and recording the position of the posterior mode for each row and column), thus keeping computational expense to a minimum.

(3.3) Analysis of circle true scenes.

We now proceed to analyse an image derived from the true scene in Figure 10(a). For demonstration purposes, θ_1 was taken as 0.0, θ_2 was taken as 3.0, and the image-formation process was identical to that in equation (2.1), with $\sigma^2 = 1.0$ (hence a relatively large Signal-Noise ratio of 3.0) The results of the analysis are depicted in figure 14.

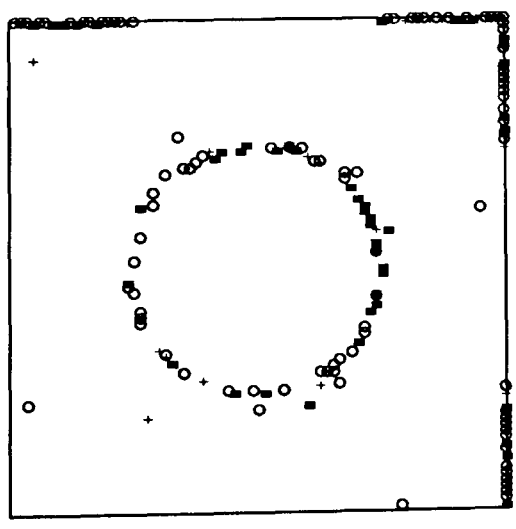


Fig 14(a) : One changepoint posterior

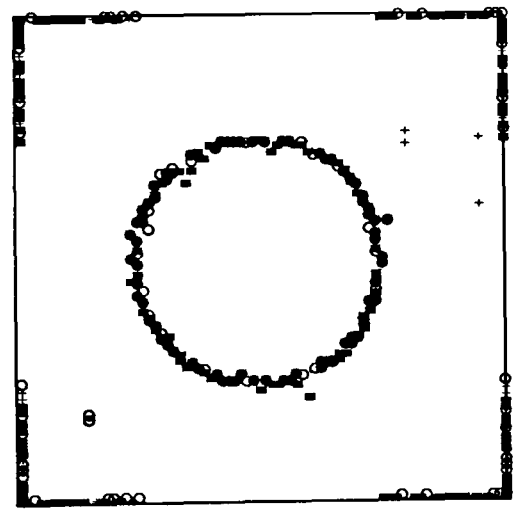


Fig 14(b) : Two changepoint posterior

Figure 14(a) depicts the raw results of full analysis using the one changepoint posterior distribution (2.11), incorporating positive probability of no changepoint/edge (plotted as changepoint at end of row/column). It is clear that much of the circle structure has been captured, and edge regions are clearly discernible. The analysis took of the order of two seconds. Figure 14(b) depicts the raw results of full analysis using the two changepoint posterior

distribution (3.2), again incorporating a positive probability of no changepoint/edge (plotted now as changepoints at beginning and end of row/column). The edge is again located, with posterior probabilities for pairs of changepoints being higher than those for single changepoints in figure 14(a) as we would expect. However, in the context of the edge-detection problem, the results are essentially equivalent - we have successfully located edge regions and regions of homogeneity in both cases. The analysis involved in the production of figure 14(b) involved of the order of two minutes of processing time. Thus, for the edge-detection problem, the one changepoint posterior technique is clearly preferable in this case, due to the considerable saving in processing time (a factor of around sixty). In the light of figure 12, however, we may expect the one changepoint technique to be of less use at lower Signal-Noise ratios. Figure 15 depicts the results of full one changepoint analysis on the circle true scene for Signal-Noise ratios decreasing from 2.5 to 1.0 in the series (a) to (d).

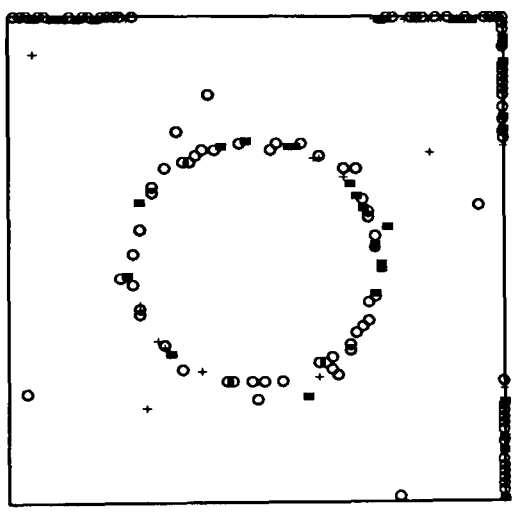


Fig 15(a) : S.N.R. 2.5

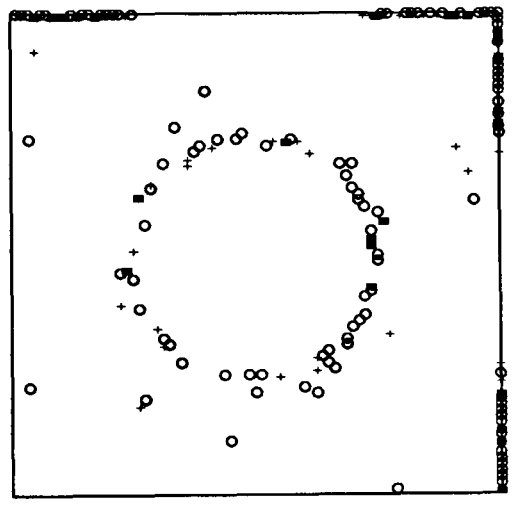


Fig 15(b) : S.N.R. 2.0

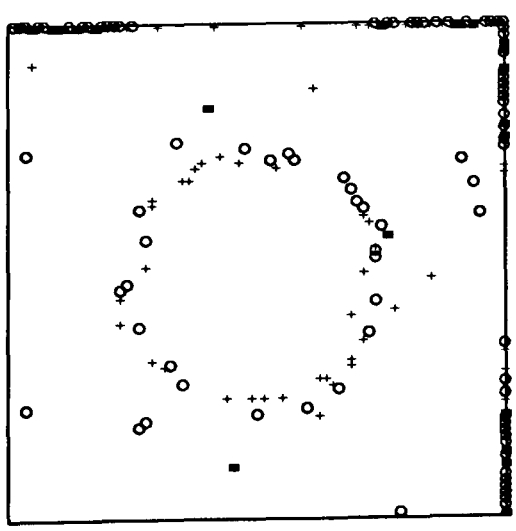


Fig 15(c) : S.N.R. 1.5

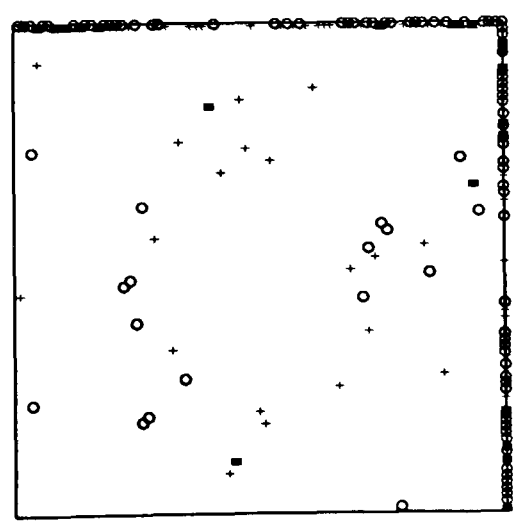


Fig 15(d) : S.N.R. 1.0

Quite surprisingly, the one changepoint "approximation" technique gives adequate results for Signal-Noise ratios as low as 1.5. However, the true scene involved is a favourable one in that it contains a large, symmetrically situated object. Figures 16 and 17 depict the results of

analyses when the true scene is potentially less favourable.

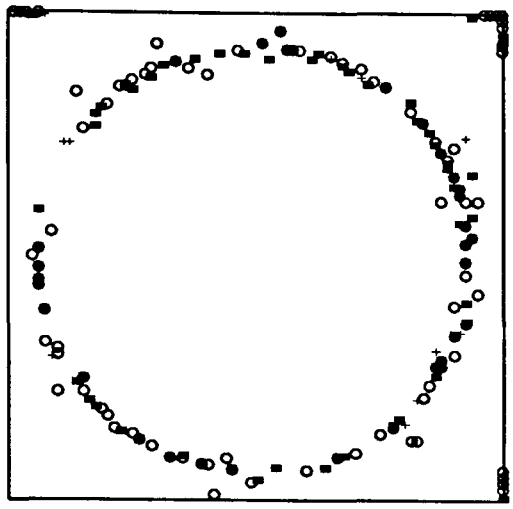


Fig 16(a) : radius 35.0

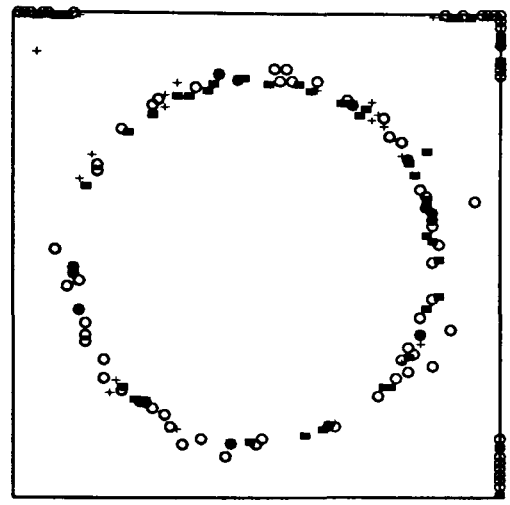


Fig 16(b) : radius 30.0

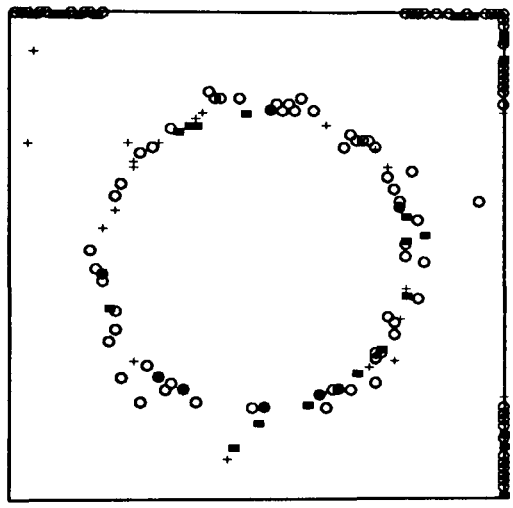


Fig 16(c) : radius 25.0

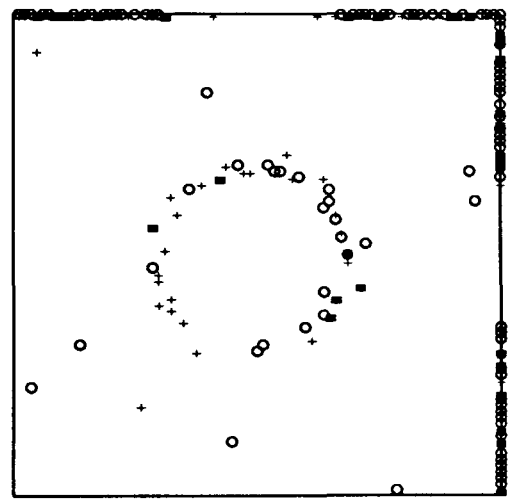


Fig 16(d) : radius 15.0

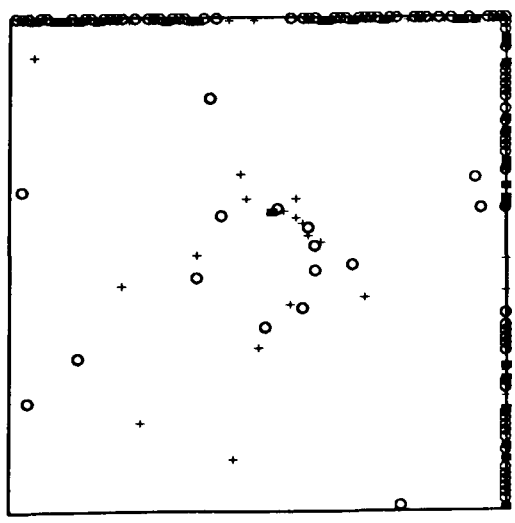


Fig 16(e) : radius 10.0

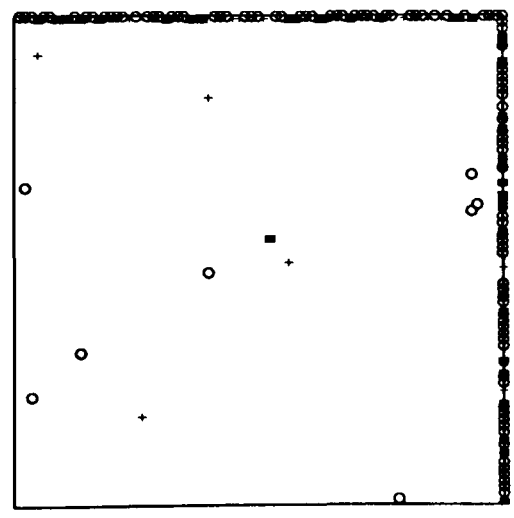


Fig 16(f) : radius 5.0

For the sequence in figure 16(a)-(f), symmetry was preserved but the radius of the circle was varied at fixed Signal-Noise 2.0. The results seem adequate except in the case of figure 16(f).

We shall develop special strategies to deal with small objects in section (3.5). Figure 17 depicts the results of one changepoint analyses of an image derived from a true scene containing a circle of radius 20.0 displaced from the centre of S_θ with Signal-Noise ratio fixed at 2.0.

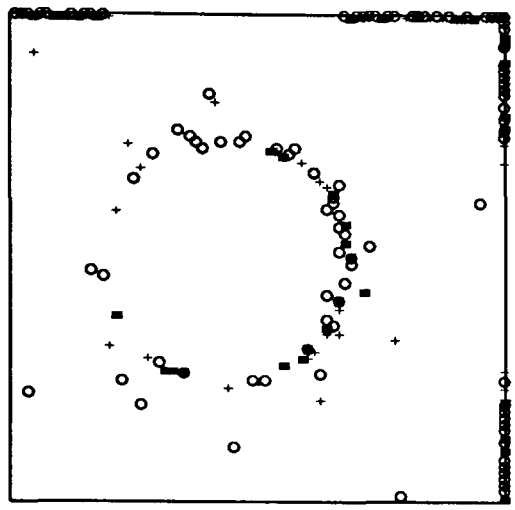


Fig 17(a) : centre (35,40)

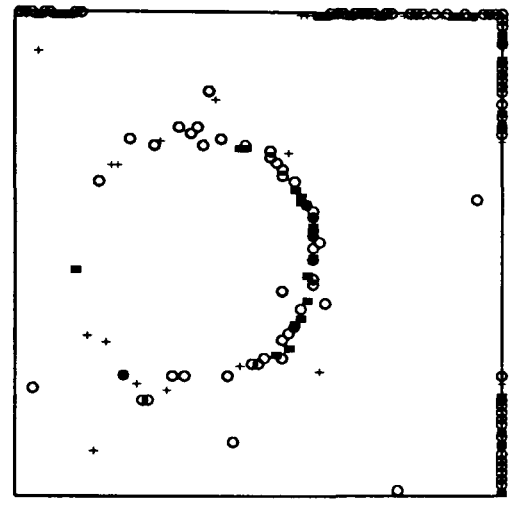


Fig 17(b) : centre (30,40)

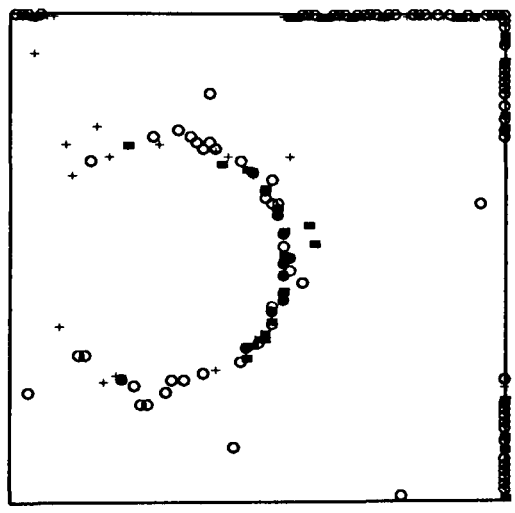


Fig 17(c) : centre (25,40)

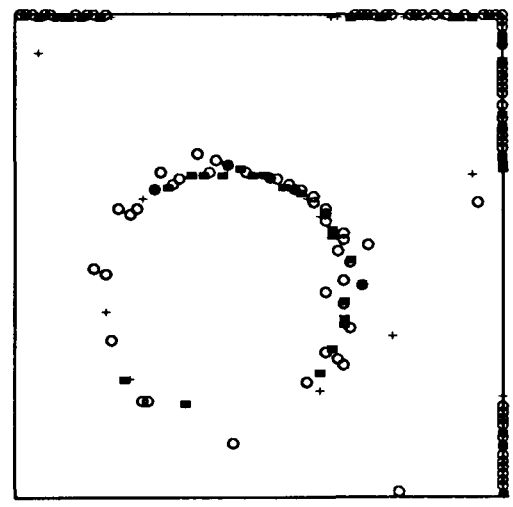


Fig 17(d) : centre (35,35)

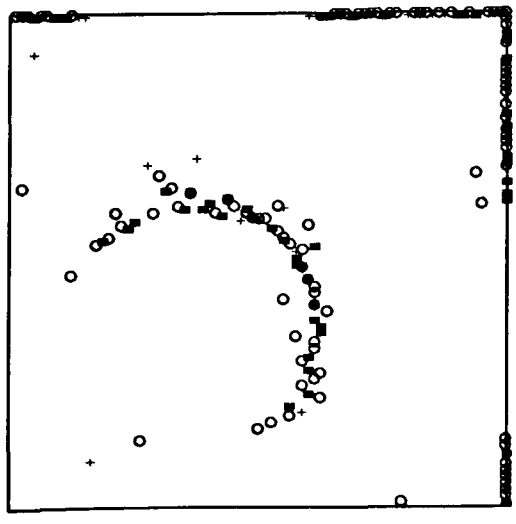


Fig 17(e) : centre (30,30)

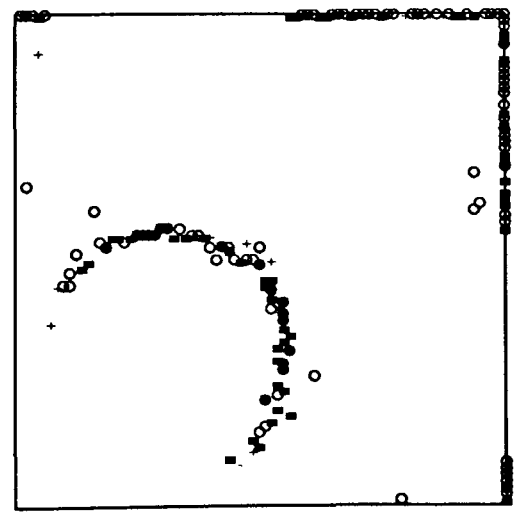


Fig 17(f) : centre (25,25)

These results confirm that for any row in which two edges/changepoints are asymmetrically situated, we detect one (that one nearer the middle of the row in question) with greater frequency. However, for these particular examples, the edge-detection problem has been tackled with some success - we have captured a good deal (if not all) of the circle structure, and have certainly located edge and non-edge regions (note the number of rows/columns correctly classified as having no edge/changepoint). We now propose some simple techniques which further improve upon the results in figures 15 to 17.

(3.3.1) Analysis of other projections.

In the above, we analysed each row and column of the image data matrix using changepoint techniques and combined the results. However we could equally well choose to analyse the image in any two (perpendicular or other) directions, and we would generally expect comparable results. Now for any single convex object true scene, any planar projection through the image data will contain either two or no edges, and thus there must exist a set of optimal projections for the edge-detection problem (in the sense discussed above, for instance with inter-edge distances maximised). In practice we would not have the necessary information about object position and orientation to make use of these optimal projections, although an adaptive analysis technique would potentially be able to choose interesting projections on the basis of results already obtained. In any case, we could augment our preliminary full analysis of the image by analysis of other projections.

Recall figure 15(d). At a Signal-Noise ratio of 1.0, our previous analysis did not adequately solve the edge-detection problem for the image concerned. Now consider figure 18(a), which depicts the results obtained as from the previous analysis plus the results of a full analysis using the pair of perpendicular directions making an angle of 45° with the rows and columns of the image (termed a "cross" analysis). For comparison, figure 18(b) depicts the results of a full analysis using a two changepoint posterior distribution.

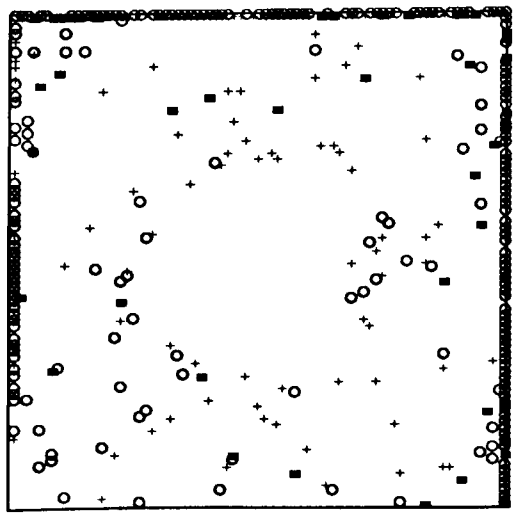


Fig 18(a) : full + cross analyses

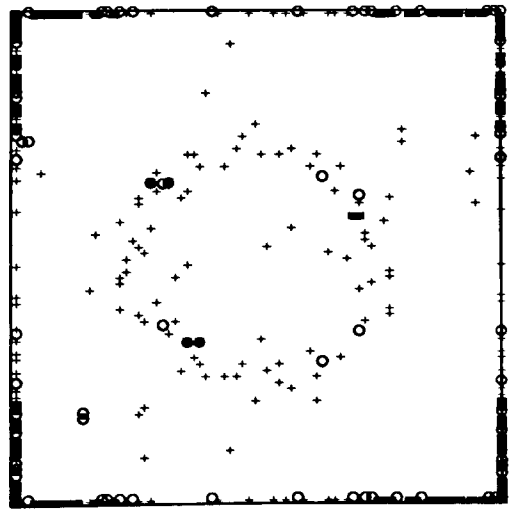


Fig 18(b) : two changepoint analysis

Despite the presence of many obvious edge-misclassifications in figure 18(a), we note an improvement in the results in the sense that we now can more clearly discern the central region of homogeneity. The analysis took the order of four seconds. Figure 18(b) depicts the results of analysis based on a two changepoint posterior distribution. As we would expect the results are more satisfying, but again the results took of the order of two minutes to produce, an unacceptable amount of processing time compared with the small addition to processing time needed to carry out extra one changepoint analyses along secondary projections. (Note - the precise choices of secondary projections for this particular true scene are not important due to the scenes symmetric nature. It is possible to choose optimally for any scene. For convenience here we choose diagonals in the image matrix.)

(3.3.2) Binary segmentation.

As we have seen, our proposed method associates modes in the one changepoint posterior distribution with the true position of a changepoint, independently of whether the underlying sequence actually had one or two (or possibly more) changepoints. Thus a possible strategy for the detection of multiple changepoints is to compute the posterior probabilities and locate the posterior mode $\hat{r}$ for the each row/column sequence Y as usual, and then to repeat the procedure for both of the sub-sequences $Y_L = (Y_1, \ldots, Y_{\hat{r}})$ and $Y_R = (Y_{\hat{r}+1}, \ldots, Y_n)$ independently, locating the posterior mode in each case, unless originally $\hat{r} = n$. We term this technique binary segmentation, and note its implicit use in many fields (e.g. search-type algorithms). Clearly, in the general case we could iterate this segmentation until each segment has the posterior mode at its end (indicating no changepoint in each sub-sequence). However for the edge-detection problem in the analysis of single convex object true scenes, we need only segment at most once for each row/column.

We propose this technique chiefly to assist in the analysis of true scenes involving some form of asymmetry, as in figure 17, and for which some of the underlying structure is captured by the standard analysis. Clearly, if the standard analysis produces poor results, then the binary segmentation will be of little additional use as it depends largely on the accurate detection of one of the changepoints for its initial step. Hence we proceed to analyse the true scenes underlying figure 17 using the binary segmentation technique. Recall that in these true scenes the Signal-Noise ratio was fixed at 2.0 and the degree and nature of the asymmetry varied in the series (a) to (f) The one changepoint posterior distribution (2.11) was used at each stage, on the entire row/column sequence initially and then on the two resulting subsequences (Note - this is to some extent incoherent from a Bayesian perspective since, in deriving (2.11), we specified non-informative prior distribution for the unknown parameters of the sampling distribution. After the initial analysis of the row/column data sequence, strictly, we should be able to specify informative prior distributions for these parameters and carry out

the analysis of the sub-sequences using a different posterior distribution. However, for convenience, we restrict attention to analysis using (2.11) at each stage.). The results of the analysis are depicted in figure 19.

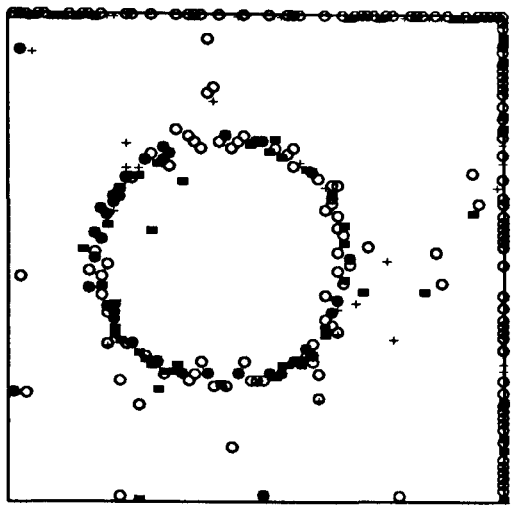


Fig 19(a) : centre (35,40)

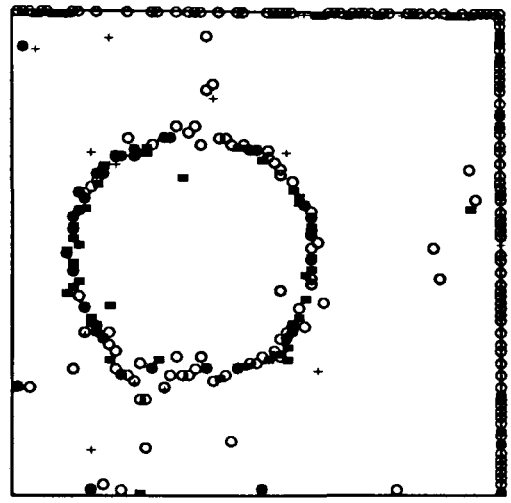


Fig 19(b) : centre (30,40)

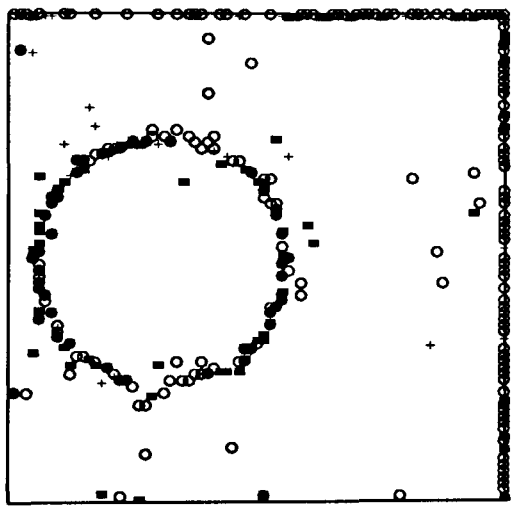


Fig 19(c) : centre (25,40)

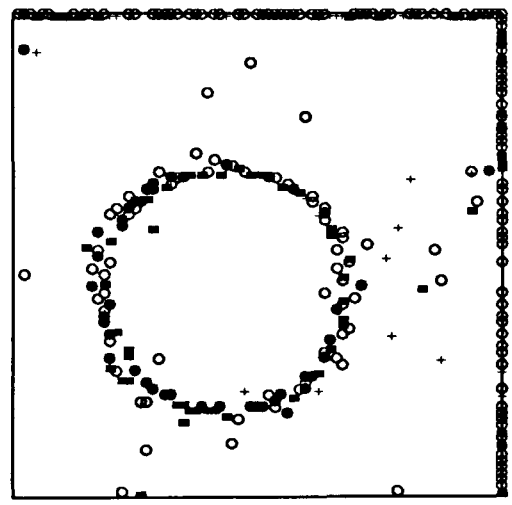


Fig 19(d) : centre (35,35)

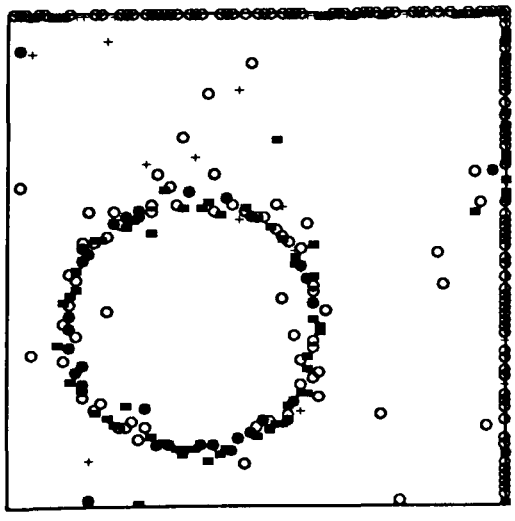


Fig 19(e) : centre (30,30)

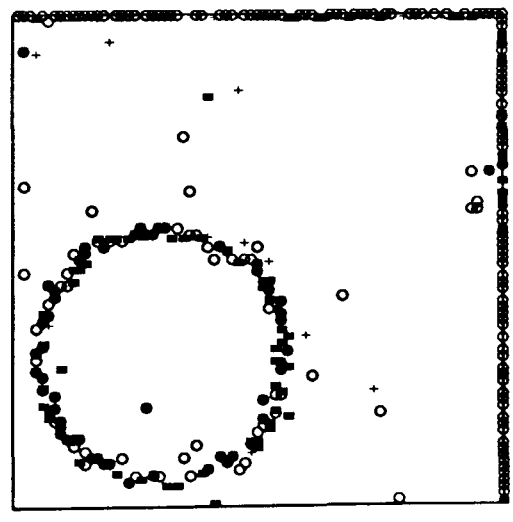


Fig 19(f) : centre (25,25)

It is clear that, in each case, the circle structure has been captured almost entirely. The analysis in each case took the order of 2.7 seconds of processing time, only a minor increase from the previous full analyses, and certainly quite acceptable. The only possible problem is that there appear to be a relatively high number of edge-misclassifications compared with the results we have previously regarded as adequate. There are two reasons for this. First, in the full analyses used to produce, say, figures 14 to 16, we recorded 160 posterior modes, one for each eighty rows and columns. For figure 19, this number rose to something over 300 in each of (a) to (f). Thus it is not surprising that the results appear more "noisy". Secondly, for figures 14 to 16 each data sequence had length eighty, whereas for figure 17 the lengths of the data (sub-)sequences were variable, theoretically having average length forty. Intuitively (and in light of the discussion of the behaviour of (2.11) in expectation above) we would expect to locate the true changepoint position with higher probability for longer data sequences. In the same vein, in figure 19, no indication is given as to the length of the (sub-)sequence from which any particular recorded changepoint resulted (but, as before, one of four symbols was used to indicate the magnitude of the modal posterior probability). Strictly, we should indicate the dependence of the changepoint posterior distribution on sequence length explicitly by writing $[r \mid Y, n, \psi]$ instead of $[r \mid Y, \psi]$, and it is not immediately clear that we may regard posterior probabilities resulting from sequences of different lengths as equivalent. Recall (3.15). We saw there that the difference in the expected value of the sum of squares was given by

$$\lambda_{r_1^*} - \lambda_{r_2^*} = (r_2^* - r_1^*)^2 \left(\frac{1}{n - r_2^*} - \frac{1}{r_1^*} \right) \alpha^2$$

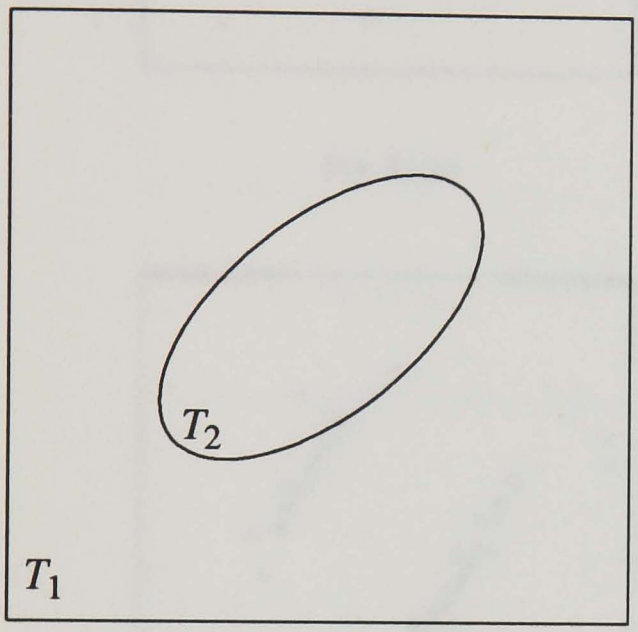
Now, it is easy to see that increasing n, r_1^*, and r_2^* by a factor of k, say, will induce an increase in the magnitude of this difference by a factor of k , unless of course we have symmetry when the difference remains 0. Hence altering n subject to these conditions will alter the resulting posterior probabilities at r_1^* and r_2^* relative to each other in non-symmetric cases.

For the edge-detection problem, these matters are somewhat irrelevant at this point, as we have located a set of candidate edge-points and could legitimately proceed to make inference from them treating each equally. Any subsequent analysis that took into account their associated measure of uncertainty (the modal posterior probability $p = \max_r [r \mid Y, \psi]$, where the associated minimum Bayes risk for the 0-1 loss function is $1 - p$) would, however, be regarded as more satisfying. Thus we return to the implications of inference based on sequences of different length at a later stage.

We now proceed to study an obvious generalisation of the class of circle true scenes.

(3.4) Convex object true scenes - ellipse.

For our second convex object true scene example we consider an ellipse lying within region S_θ. This is a more complex true scene than the circle of figure 10: for the circle, we could vary three parameters in the production of the true scene - two location parameters plus radius, whereas for the ellipse we could vary five - two location, lengths of major and minor axes, and orientation. Figures 20(a) and (b) depict an ellipse true scene and image derived from it using the image-formation process (2.1).



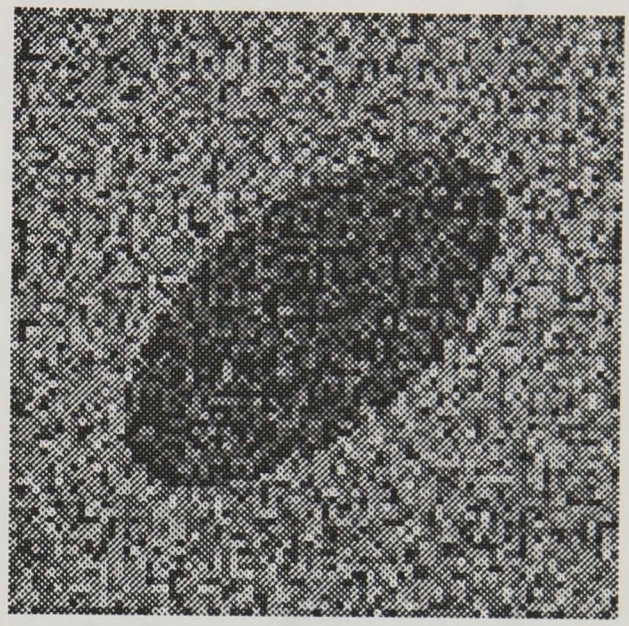


Fig 20(a) : true scene

Fig 20(b) : image

We proceed to study the performance of the techniques developed above with respect to the ellipse true scene. First, we study the effect of orientation on the ability of the various analytic techniques to capture object structure. Fixing location (p,q) as the centre of S_θ (40,40), and major and minor axis lengths (a,b) as (30,15), and a Signal-Noise ratio of 2.0, different true scenes were obtained by varying the angle of orientation α, measured from the positive x-axis in the usual way. The results are depicted in figure 21. In figure 21(a) and (b) $\alpha = 0$, in (c) and (d) $\alpha = \frac{\pi}{3}$, and in (e) and (f) $\alpha = \frac{2\pi}{3}$. Figures 21(a), (c), and (e) depict the results of full analysis. Although some of the ellipse structure is captured in each case, with partial edges detected, the complete underlying structure is not captured. This is as we would have predicted from the discussion above, and is due to the small inter-changepoint distance in (a) and the inherent asymmetry in (c) and (e). Figures 21(b), (d) and (f) depict the results of binary segmentation. As above, these results are more satisfying. Practically the whole ellipse edge has been detected in each case, and there are relatively few misclassified points. These results confirm that the binary segmentation technique is of considerable use in the analysis of images derived from convex object true scenes.

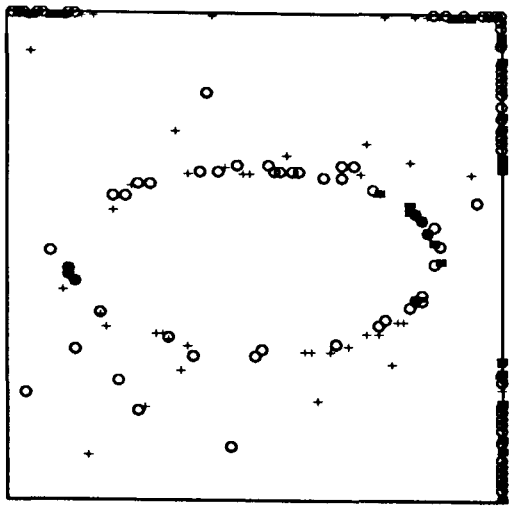


Fig 21(a)

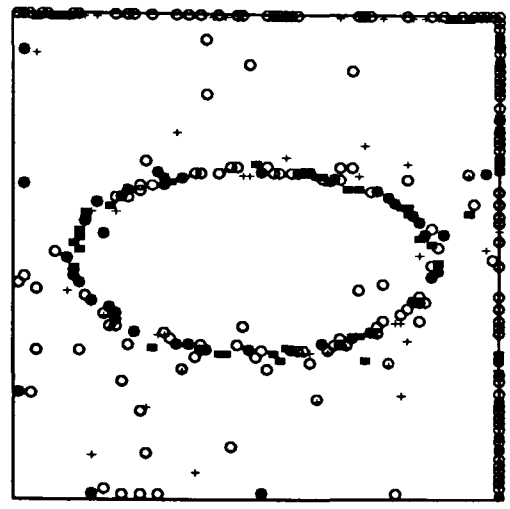


Fig 21(b)

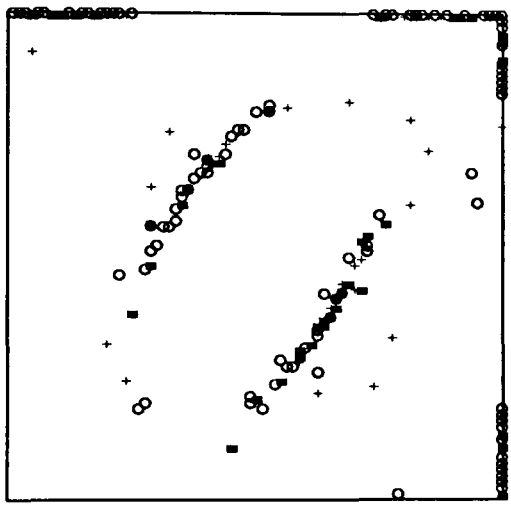


Fig 21(c)

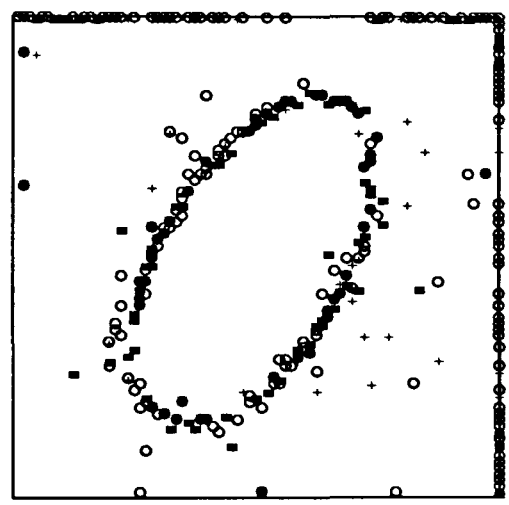


Fig 21(d)

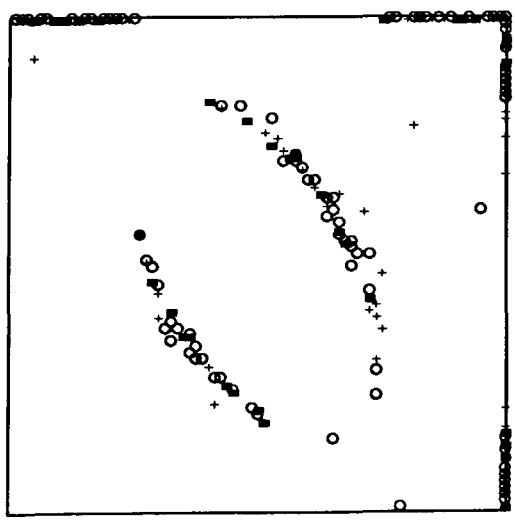


Fig 21(e)

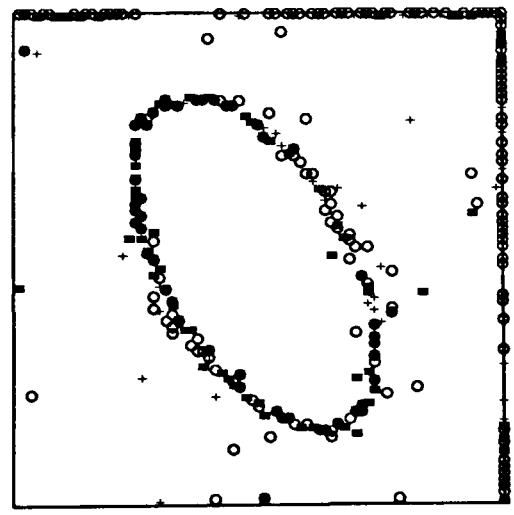


Fig 21(f)

Next, we study the behaviour of the technique when the scale parameters a and b are varied. As in the previous example, the Signal-Noise ratio was fixed at 2.0, and the angle of orientation α was fixed at $\frac{\pi}{6}$. First, a and b were varied with the ratio $b/a = e$ (ellipticity)

fixed equal to 0.5, as in the true scene underlying figure 21. The results of the two types of analysis discussed previously are depicted in figure 22.

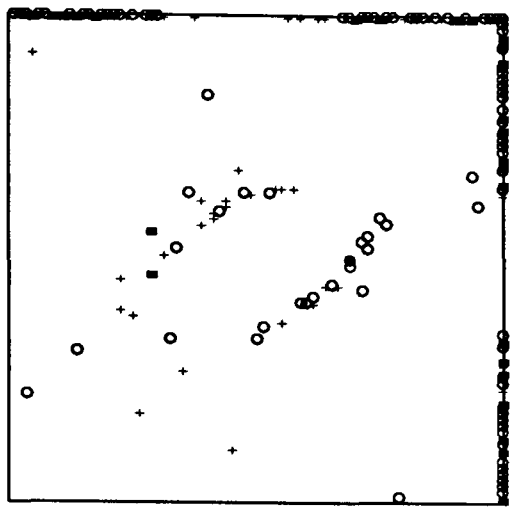


Fig 22(a)

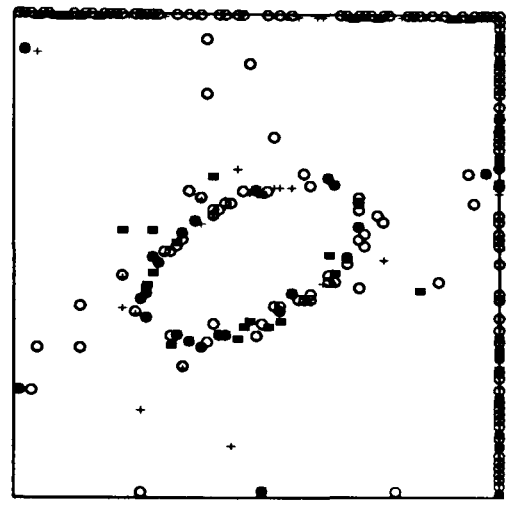


Fig 22(b)

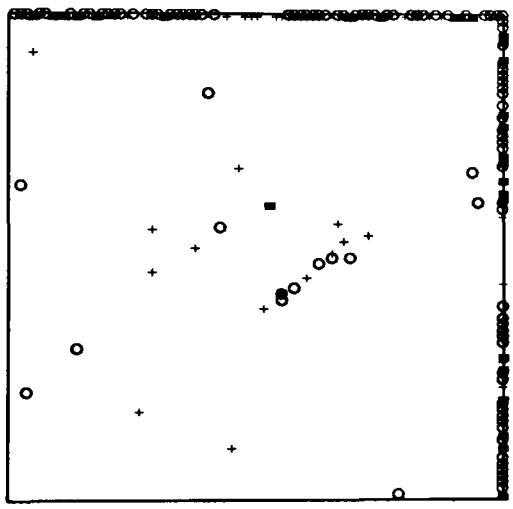


Fig 22(c)

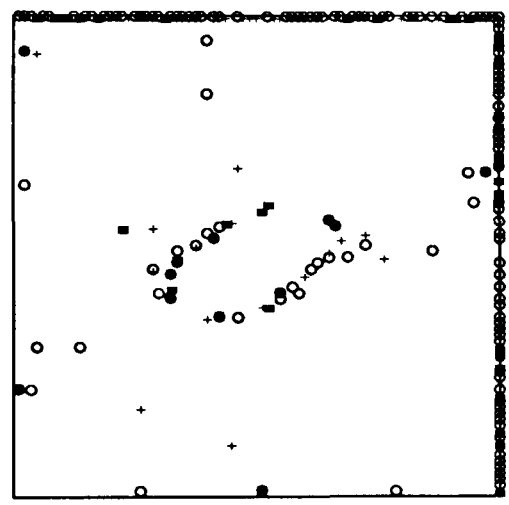


Fig 22(d)

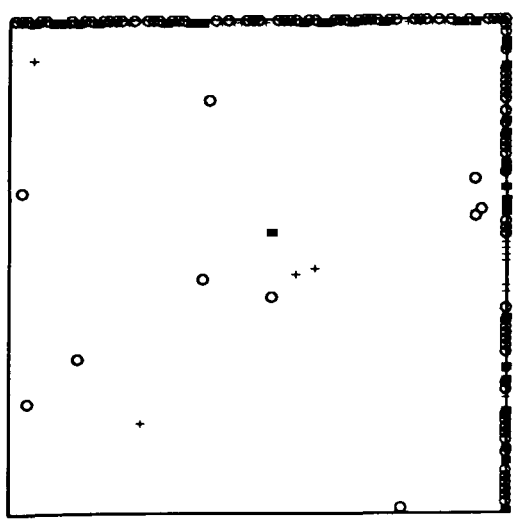


Fig 22(e)

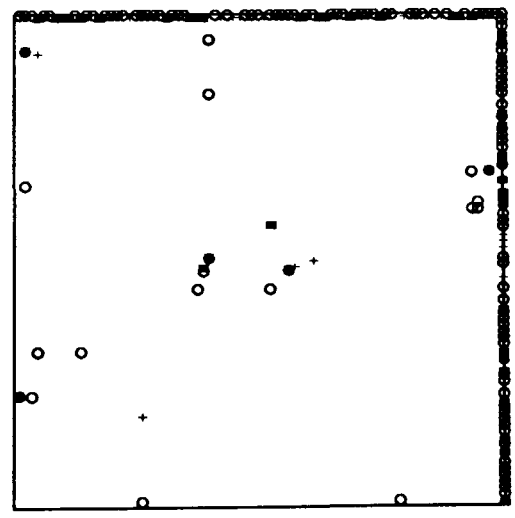


Fig 22(f)

Figures 22(a) and (b) relate to an ellipse with a = 20, figures (c) and (d) to an ellipse with a =

15, and figures (e) and (f) to an ellipse with a = 10, with e = 0.5 in each case. The results show quite clearly that the binary segmentation technique provides the most satisfactory results, and that techniques based purely on the single changepoint posterior distribution give poor results for very small ellipses at this order of Signal-Noise ratio. This is an entirely understandable phenomenon which is explained fully below.

It is interesting at this point to compare these results with those obtained by using the two changepoint posterior distribution (3.2), depicted in figure 23.

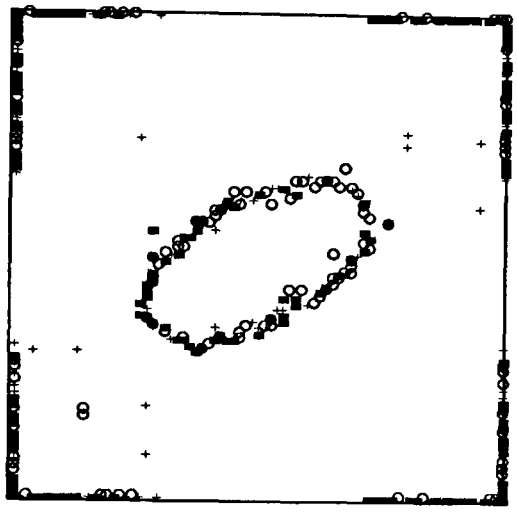


Fig 23(a) : a = 20

Fig 23(b) : a = 15

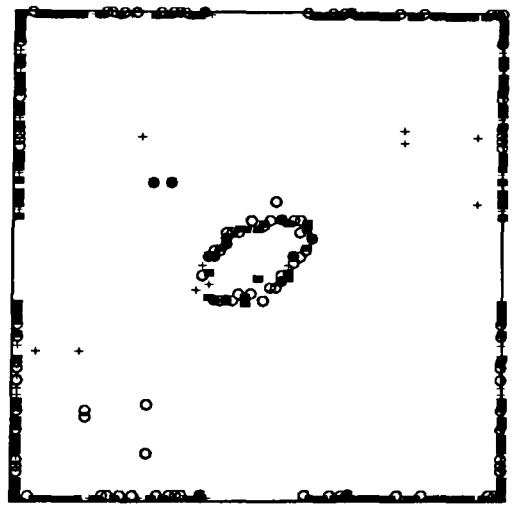


Fig 23(c) : a = 10

These results are by far superior to those depicted in figure22, but again the processing time involved (around two minutes) is prohibitive. In the case of (a) and (b) here, the one changepoint "approximate" method gives adequate results, but in (c) the two changepoint method is preferable. Hence, as noted above, it seems likely that we must adapt the one changepoint technique in some way in the case where the inter-changepoint distance is small i.e. when we have prior knowledge that the convex object is small. We discuss this after first noting the behaviour of the techniques for ellipses having larger ellipticity, because, as we

shall see, similar problems are encountered.

Consider ellipses of fixed area where e is allowed to vary. For demonstration purposes, we fix $ab = 200$, but allow e to vary between 0.75 and 0.01. The Signal-Noise ratio was fixed at 2.0, and the binary segmentation technique used. The results of the analysis are depicted in figure 24.

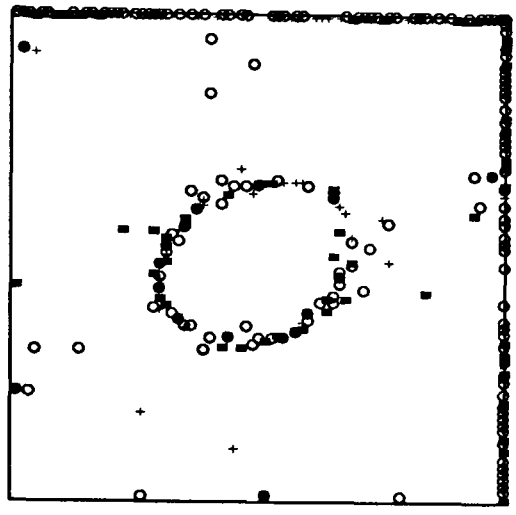


Fig 24(a) : e = 0.75

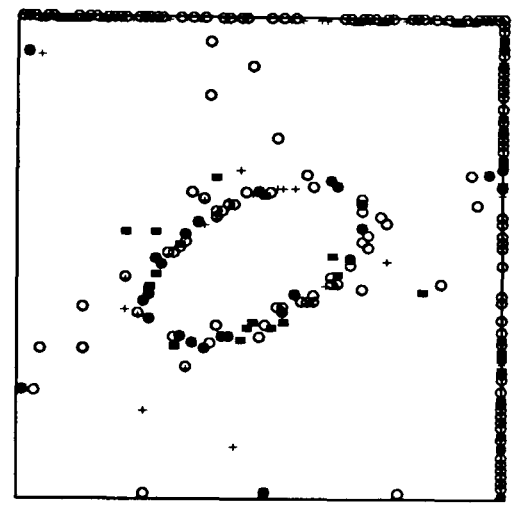


Fig 24(b) : e = 0.5

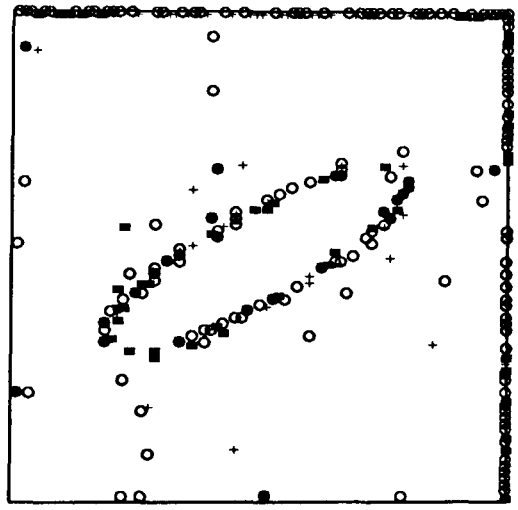


Fig 24(c) : e = 0.25

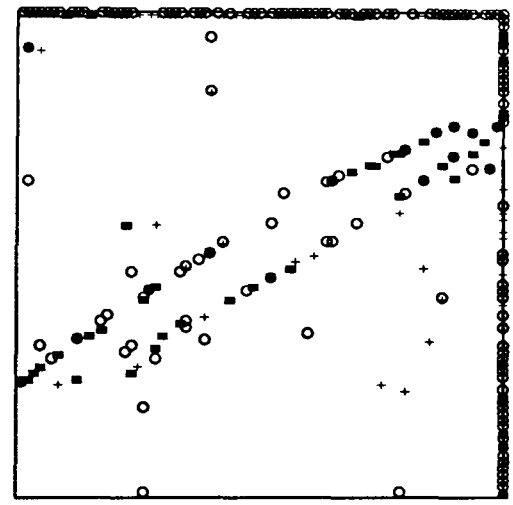


Fig 24(d) : e = 0.1

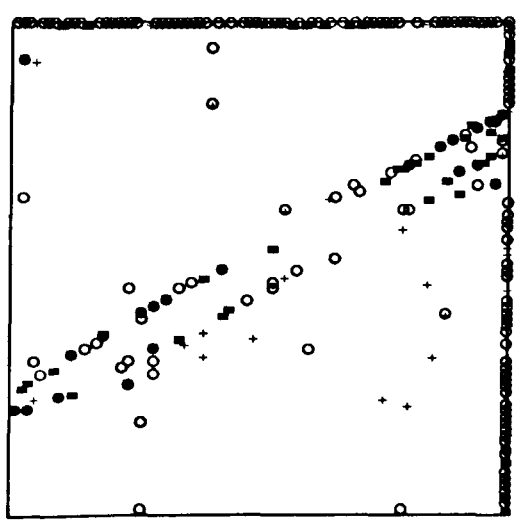


Fig 24(e) : e = 0.05

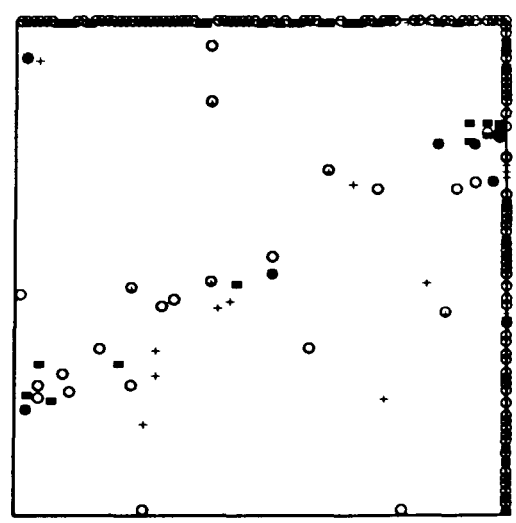


Fig 24(f) : e = 0.01

These results are remarkably good, even when the degree of ellipticity is high (the results are adequate for these true scenes for the other two methods discussed above also). However, in the most extreme case here, figure 24(f), the structure is lost due to the small inter-changepoint distance. It is clear that we must adapt our formulation of the edge-detection problem so that images derived from true scenes containing small objects may be detected.

(3.5) Adaptation of changepoint formulation for small objects.

As we remarked above, the changepoint technique we have introduced is inadequate for such problems. Recall, therefore, our motivation for the choice of the changepoint technique as a solution to the edge-detection problem. We saw in a previous section how the detection of a simple edge between two larger homogeneous texture regions under simple image-formation assumptions was equivalent to the statistical problem of changepoint analysis. Now, clearly, in the two edge case where the edges are close together in any particular row/column, the data sequence transects a small convex object, a much more pleasing statistical analogy is that of outlier detection - we seek to detect a small number of (consecutive) pixel values seemingly having raised level relative to the background. Thus we reformulate the problem as follows - under the image-formation process (2.1), it is now clear that, for some $r\,(1 \leqslant r \leqslant n)$ and some integer k (k small) the conditional distribution $[Y_i \mid r, \theta]$ is $N(\theta_1, \sigma^2)$ for $1 \leqslant i \leqslant r$ and $r+k+1 \leqslant i \leqslant n$, and is $N(\theta_2, \sigma^2)$ for $r+1 \leqslant i \leqslant r+k$, where k is regarded as another parameter of the system, and is taken to represent the "width" of the object in the row/column in question. In practice k will often be unknown, and thus strictly we should specify a prior distribution for it and carry out a full Bayesian analysis, and so this is merely a reparameterisation of the two changepoint case above which we reject due to computational limitations. But in this form, a simplification of the problem is obvious. We could legitimately restrict the valid range of k to reflect our prior knowledge of the true scene (that it contains a small object) by specifying a prior probability of zero for values of k outside that range, and proceed to compute the joint posterior distribution $[r, k \mid Y, \psi]$. Alternatively, we could fix k to have some value thought *a priori* to be smaller than the object width, proceed and compute $[r \mid Y, k, \psi]$ and then report the mode of this posterior distribution. This alternative cannot strictly be regarded as an edge-detection technique, as if k is markedly smaller than the true object width then the posterior mode will frequently occur at positions internal to the object, but away from the object boundary, so we move into the domain of object- rather than edge-detection. However, for small objects, edge-points and internal points are essentially equivalent, and we detect the background region of homogeneity correctly, so that much of the problem is solved.

We have already examined the behaviour of the two changepoint posterior distribution (3.2), and it is not particularly instructive to study the behaviour of a restricted version, other

than to note the obvious improvement in amount of processing time needed. We concentrate, therefore, on the k fixed alternative. Under the same prior assumptions that we used to derive (2.11), the posterior distribution for r is easily seen to be given by

$$[r \mid Y, k, \psi] \propto \{k(n-k)\}^{-1/2} \left\{ \sum_{i \in O_k'} (Y_i - \overline{Y_{O_k'}})^2 + \sum_{i \in O_k} (Y_i - \overline{Y_{O_k}})^2 \right\}^{-n/2} \qquad (3.17)$$

where $O_k' = \{1,..,r,r+k+1,..,n\}$, $O_k = \{r+1,..,r+k\}$,

$\overline{Y_{O_k}} = \frac{1}{k} \sum_{i \in O_k} Y_i$, and $\overline{Y_{O_k'}} = \frac{1}{(n-k)} \sum_{i \in O_k'} Y_i$.

We now study the behaviour of (3.17) for a typical "small object" example. Figure 25 depicts the results of 1000 simulations of a two changepoint sequence with changepoints at 47 and 53 and Signal-Noise ratio 2.0 for various prior assumptions and corresponding posterior distributions. As in figures 11 to 13 in section (3.2), the average posterior probability was plotted on the vertical scale.

Figure 25(a) depicts the results obtained when using the usual one changepoint posterior distribution (2.11). Despite the fact that the posterior mode (neglecting the ends of the sequence) occurs at one of the true changepoint positions, and localised modes are associated with both of the changepoints, this "expectation" result leads us to believe that, in practice, for any single sequence, posterior modes will not be related to true changepoint positions due to the presence of random fluctuations in the data sequence (the "expectation" of the posterior distribution (2.11) with respect to the distribution of such a sequence is practically uniform). Figures 25(b) to (d) depict the results obtained when the posterior distribution (3.17) is used for various choices of k.

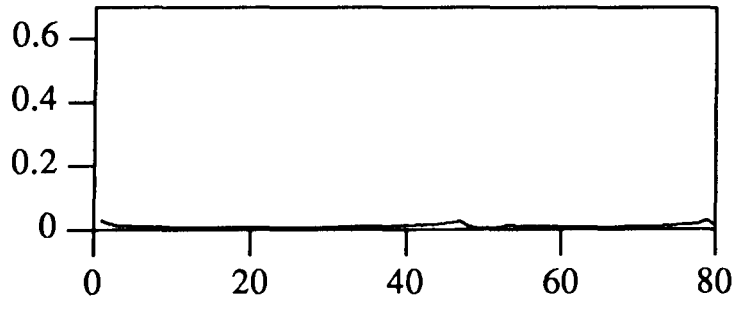


Fig 25(a) : distribution (2.11)

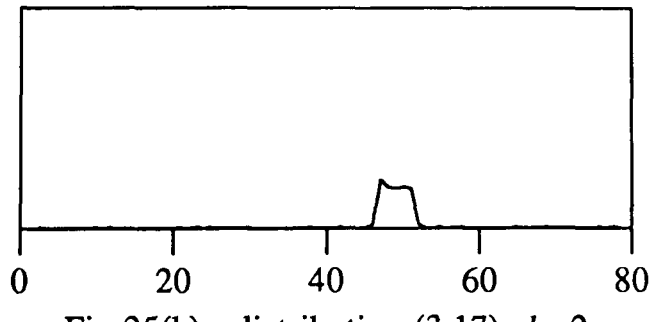


Fig 25(b) : distribution (3.17), k =2

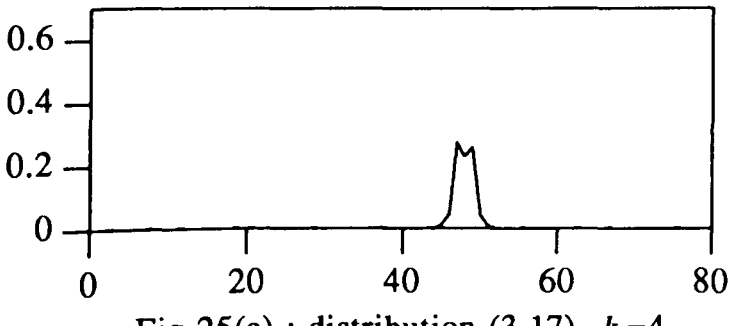


Fig 25(c) : distribution (3.17), k =4

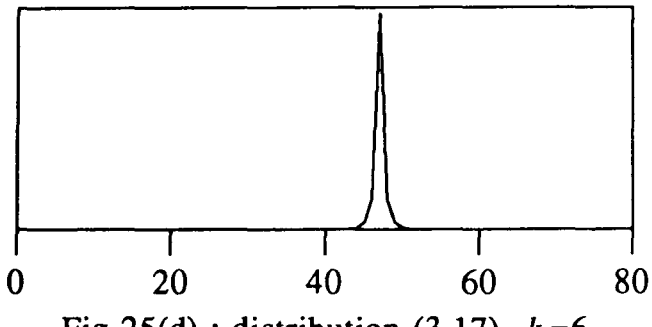


Fig 25(d) : distribution (3.17), k =6

The results in (b) - (d) are more satisfactory than in (a) as in each case the region in which the changepoints occur is detected with much greater frequency. Note that, provided the actual inter-changepoint distance is small, we can specify k to be smaller than that distance and still obtain adequate results (figures 25(b) and (c)). Another important feature is that the computation involved in the evaluation of the probabilities in (3.17) is essentially equivalent to that required to evaluate (2.11). Thus we would expect a marked improvement on the processing time required for distribution (3.2).

We now proceed to analyse the true scene underlying figures 22(e) and 23(c) using the posterior distribution (3.17). Nominally, we choose $k=3$. Recall that the ellipse had dimensions (10,5), and that a Signal-Noise ratio of 2.0 was imposed. The results of the three techniques for such an image are depicted in figure 26. Figures 26(a) and (c) are the results obtained previously by using posterior distributions (2.11) (timing 1.8s) and (3.2) (timing 120s), respectively. Clearly, (a) is inadequate and (c) excellent. Figure 26(b) depicts the results obtained when (3.17) is used. Despite the presence of misclassified points, the results are a marked improvement on those in (a). The processing time required to produce these results was 1.6 seconds. Thus it would seem preferable to use (3.17) (and smooth) rather than using (3.2).

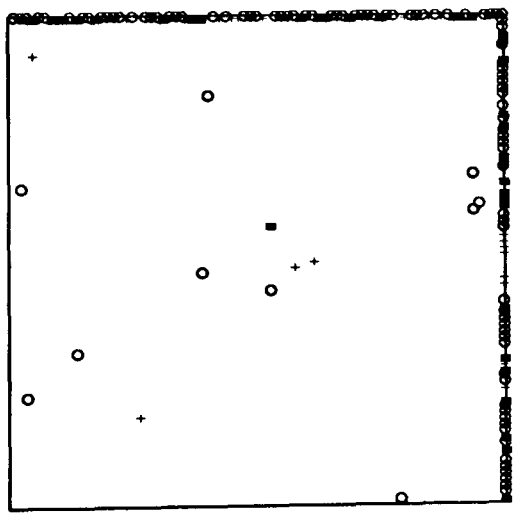


Fig 26(a) : distribution (2.11)

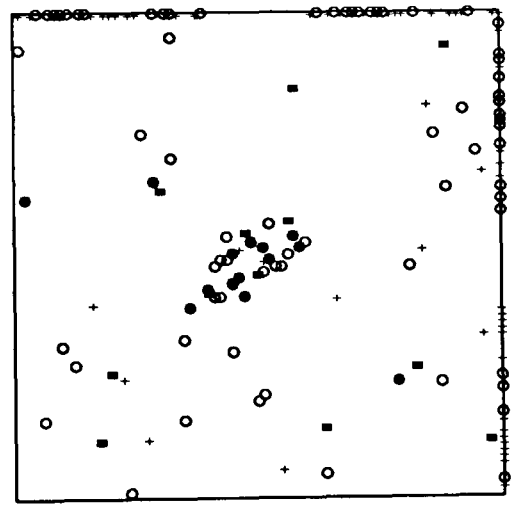


Fig 26(b) : distribution (3.17), $k=3$

Fig 26(c) : distribution (3.2)

We now demonstrate the robustness of this detection technique to choices of k. Figure 27 depicts the results obtained when the true scene underlying figure 24(f) was analysed using the technique based on (3.17) for various choices of k.

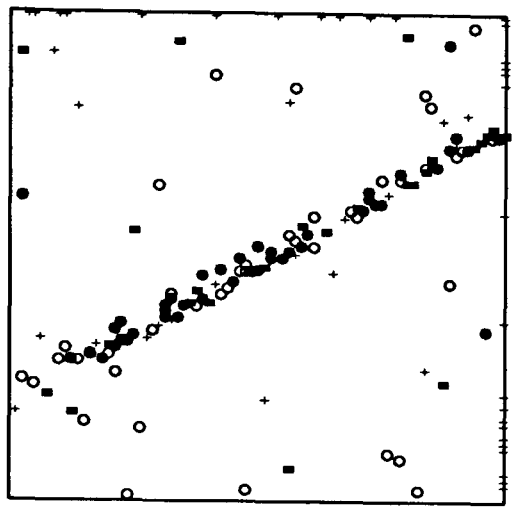


Fig 27(a) : $k = 2$

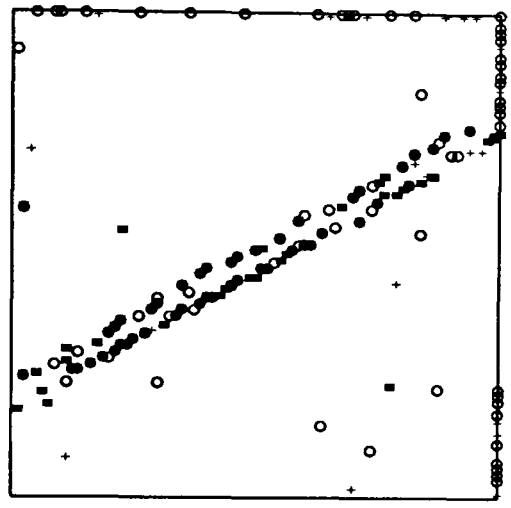


Fig 27(b) : $k = 4$

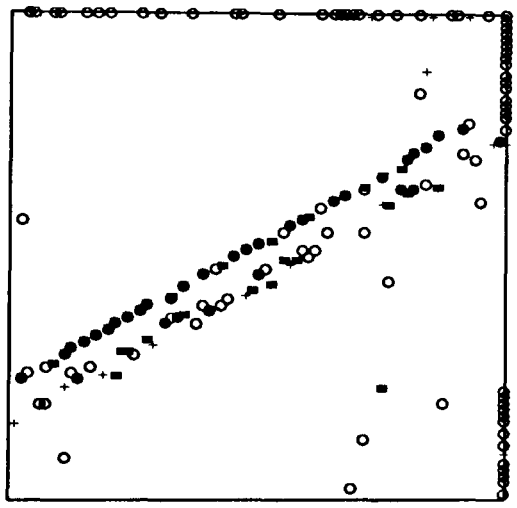


Fig 27(c) : $k = 6$

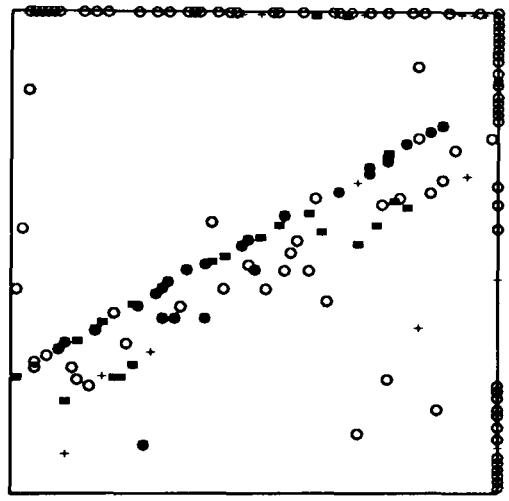


Fig 27(d) : $k = 8$

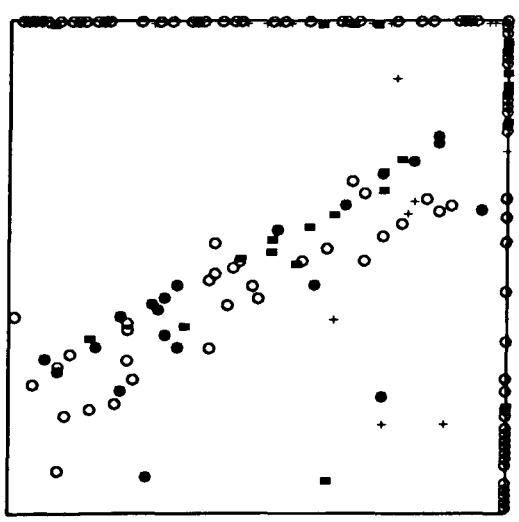


Fig 27(e) : $k = 10$

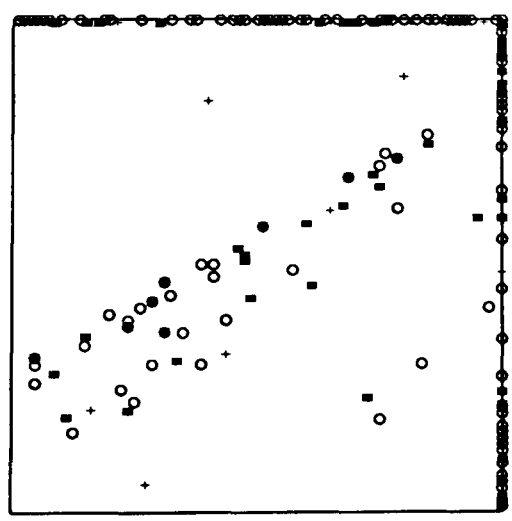


Fig 27(f) : $k = 12$

It is clear that, in each case, the results are more satisfying than those depicted in figure 24(f), even when the chosen value for k is considerably larger than the true inter-changepoint distance in any row/column. The results are most impressive when the chosen value for k is of the same order as this distance, as we would expect. Thus, if we have sufficient (but not unrealistically specific) prior knowledge concerning object size, it would seem preferable to use the technique based on (3.17).

We return to the use of (3.17) in chapter 6 in our discussion of multiple object detection problems.

Note : we acknowledge above the similarity between the changepoint detection problem when the inter-changepoint distance is small, and the detection of outliers in a given set of observations. Consider a standard Bayesian approach to the outlier problem (see, for example, Pettit and Smith (1983)). Specifically, we wish to identify a subset of observations that arise from distributions having different parameters from those of the majority. Consider the case where we wish to detect k successive observations of this nature. Then we may write the joint distribution of the variables concerned $Y = (Y_1, \ldots, Y_n)$ conditional on the unknown parameters of the sampling distribution $\theta = (\theta_1, \theta_2)$ as

$$[Y \mid \theta] = \sum_r \pi_r [Y \mid \theta]_r$$

i.e. a mixture distribution, where $[Y \mid \theta]_r$ is the joint conditional distribution assuming that the variables $(Y_r, \ldots, Y_{r+k})$ are the outliers, and π_r is the prior probability that this is in fact correct. Now it is easily seen that, provided the prior $[\theta]$ is independent of r (in the sense of the above),

$$[\theta \mid Y] = \sum_r \pi_r^* [\theta \mid Y]_r$$

where $[\theta \mid Y]_r$ is the posterior distribution corresponding to $[Y \mid \theta]_r$ and the specified prior, and π_r^* is given by

$$\pi_r^* = \frac{\pi_r \int [Y \mid \theta]_r [\theta]}{\sum_r \pi_r \int [Y \mid \theta]_r [\theta]} \quad .$$

Inference is subsequently made via π_r^*, the posterior probability that $(Y_r, \ldots, Y_{r+k})$ are the outliers. It is easy to see that under the same distributional and prior assumptions, π_r^* and $[r \mid Y, k, \psi]$, and indeed the two methods, are identical.

We have developed various techniques for the solution of the edge-detection problem for convex object true scenes based around variations of the Bayesian changepoint identification idea. Further illustrations and examples of the use of such techniques may be found in Appendix 2. We now proceed to develop a further technique specifically designed to solve the changepoint problem arising from the analysis of more complex true scenes, motivated by the need for accuracy and efficiency in processing.

(3.6) Alternative approach to Bayesian multiple changepoint detection.

As we have seen above, in the simple one changepoint case, exact inference concerning changepoint position is straightforward and computationally feasible. In the two changepoint case, exact inference is also possible, but the amount of computation involved in evaluating posterior probabilities is prohibitive. We have shown that in this case approximate inference based on a one changepoint posterior distribution is computationally efficient and adequate in many circumstances. However, in the k ($k > 2$) changepoint case, exact inference is impractical, and approximate inference via simpler changepoint assumptions is inadequate. Now, in the analysis of complex true scenes, it is probable that solution of the edge-detection problem using the changepoint based techniques suggested above will require that exact or approximate inference of some sort is not only possible but computationally feasible. This motivates the search for a changepoint identification technique which retains some of the points of those described above (intuitively appealing, decision-based, non arbitrary etc.) but does not involve the need for excessive computation. Before turning to this, however, we discuss other various aspects of the identification problem.

In all of the above, for example in the derivation of (2.11) and (3.2), when we are concerned with changepoint identification, we propose the number of changepoints for the sequence concerned, evaluate the joint posterior distribution and report the joint posterior mode on the basis of a realisation of the sequence. Now, it is equally as valid to evaluate the marginal posterior distribution and report the marginal mode for each changepoint individually - this is the equivalent solution to the same decision problem when using a different (but still reasonable) loss function to that used previously. However, the only access we have to the marginal posterior distribution for each individual changepoint in the above framework is via the joint posterior distribution (by marginalising in the usual way), which we have seen to be expensive to compute, even when the number of changepoints is relatively small. If we could compute the marginal distributions by another method and report the marginal modes, then we may be able to lessen the computational load relative to that of the method discussed previously.

Consider the sequence of random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ represented in figure 30, assumed to have k changepoints $(r_1, \ldots, r_k)$ of unknown position but where k is presumed

known. Let $r_0 = 0$, and $r_{k+1} = n$. Let the sampling distribution of $Y_j = (Y_{r_{j-1}+1},\ldots, Y_{r_j})$ have parameters θ_j, and $(Y_{r_{j-1}+1},\ldots, Y_{r_j})$ be conditionally independent given θ_j, $j = 1,\ldots, k+1$. Finally, let $(Y_1,\ldots, Y_{k+1})$ be conditionally independent given $\theta = (\theta_1,\ldots, \theta_{k+1})$.

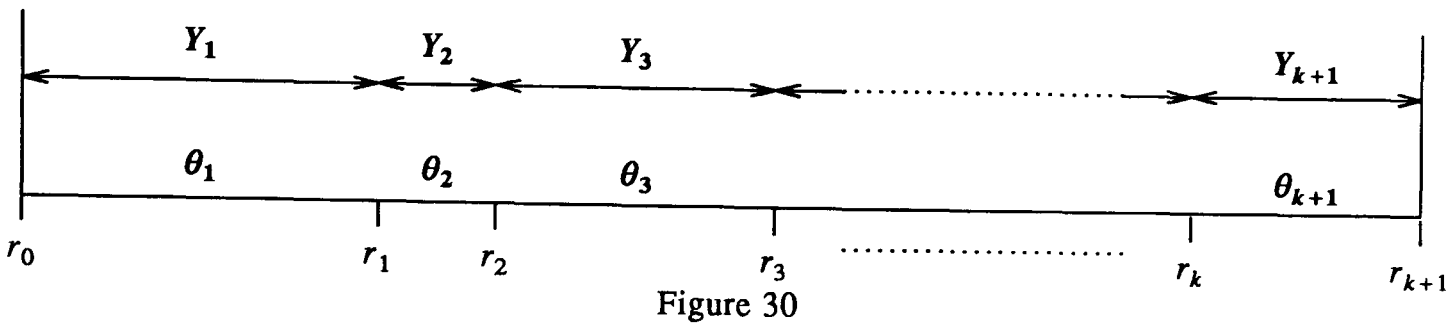


Figure 30

From the conditional independence assumptions, via an equation equivalent to (2.7), we may write the marginal posterior distribution of r_j conditional on $(r_0,\ldots, r_{j-1}, r_{j+1},\ldots, r_{k+1})$, as

$$\begin{aligned}[r_j \mid r_0,\ldots, r_{j-1}, r_{j+1},\ldots, r_{k+1}, Y, \psi] &\equiv [r_j \mid r_{j-1}, r_{j+1}, Y, \psi] \\ &\propto \int [Y_j, Y_{j+1} \mid r_j, \theta_j, \theta_{j+1}]\,[\theta_j, \theta_{j+1} \mid \psi]\,[r_j \mid r_{j-1}, r_{j+1}] \\ &= \int \prod_{i=r_{j-1}+1}^{r_j} [Y_i \mid \theta_j] \prod_{i=r_j+1}^{r_{j+1}} [Y_i \mid \theta_{j+1}]\,[\theta_j, \theta_{j+1} \mid \psi] \\ &\quad .[r_j \mid r_{j-1}, r_{j+1}] \quad . \quad (3.18)\end{aligned}$$

It is clear that in this formulation of the multiple changepoint problem, which is identical to our original one, the marginal posterior distribution for r_j conditional on the $k-1$ changepoints depends only on r_{j-1} and r_{j+1}, for $j = 1,\ldots, k$. But conditional on r_{j-1} and r_{j+1}, the marginal posterior distribution for r_j is identical to the usual one changepoint posterior distribution given by (2.7) evaluated for the sub-sequence (Y_j, Y_{j+1}), with the valid range of r_j being restricted to $r_{j-1}+1,\ldots, r_{j+1}$. Thus we may write down $[r_j \mid r_{j-1}, r_{j+1}, Y, \psi]$ using standard and familiar techniques, for $j = 1,\ldots, k$.

We now have the set of discrete, univariate posterior distributions $[r_j \mid r_{j-1}, r_{j+1}, Y, \psi]$ for $j = 1,\ldots, k$. However, we wish to make inference on the basis of the unconditional marginal distributions $[r_j \mid Y, \psi]$. The crucial link is made via the simulation technique known as stochastic substitution, and in particular the method for the evaluation of joint or marginal distributions from such conditional distributions that was described in chapter 1 section (1.4.1.2), the Gibbs Sampler, introduced initially by Geman and Geman (1984), and developed further for more familiar statistical problems by Gelfand *et al.* (1989). Recall that, for the collection of random variables of interest, where the corresponding collection of full conditional distributions are completely specified and straightforward to sample from, the Gibbs Sampler algorithm proceeds as follows. Given an arbitrary starting value for each of the

variables, sample repeatedly from each of the full conditional distributions in turn, where the "current" value (i.e. that most recently obtained by the sampling procedure) for each of the conditioning variables is used in place of the true value in the functional form of the full conditional distribution. It can be shown (Geman and Geman (1984), appendices) that, subject to regularity conditions, as the number of iterations, t say, increases, the t'th simulated value for variable j tends in distribution to a realisation from the unconditional marginal distribution of variable j, and that the t'th set of simulated values for all variables jointly tends in distribution to a realisation from the joint distribution.

In the multiple changepoint problem, therefore, the solution to the problem of computing marginal posterior distributions on the changepoint positions is straightforward. We write down $[r_j \mid r_{j-1}, r_{j+1}, Y, \psi]$ for $j = 1, \ldots, k$ by the usual techniques, and choose starting values $(r_{10}, \ldots, r_{j0})$. Then we generate a value r_{11} from the discrete posterior distribution $[r_1 \mid r_0, r_{20}, Y, \psi]$ by c.d.f inversion, then a value r_{21} from $[r_2 \mid r_{11}, r_{30}, Y, \psi]$ etc. until "convergence" (as yet to be defined) at or for a predetermined number of iterations t, resulting in a set of sample values $(r_{1t}, \ldots, r_{jt})$. We then obtain an estimate of the unconditional marginal distribution $[r_j \mid Y, \psi]$ for each j by combining the probabilities in $[r_j \mid r_{j-1t}, r_{j+1t}, Y, \psi]$ additively and averaging (the discrete analogue to the finite mixture density estimator of the marginal density in Gelfand and Smith (1990)). Note that in this formulation we have integrated out the parameters of secondary interest, namely $\theta = (\theta_1, \ldots, \theta_{k+1})$. If the θ_j's were of interest, and we wanted to calculate the marginal posterior densities $[\theta_j \mid Y, \psi]$, we could extend the Gibbs Sampler by including the $k+1$ conditional posterior densities

$$[\theta_j \mid r_0, \ldots, r_{k+1}, \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_{k+1}, Y, \psi]$$

in the sampling cycle described above. It is easily seen that this conditional posterior density for θ_j simplifies to

$$[\theta_j \mid r_{j-1}, r_j, Y_j, \psi] \propto \int \prod_{i=r_{j-1}+1}^{r_j} [Y_i \mid \theta_j] [\theta_j \mid \psi]$$

We could now simulate $2k+1$ observations per iteration, as opposed to k previously, so we might expect processing time to increase, at least by a factor of two, and by a greater factor if the functional form of $[\theta_j \mid r_{j-1}, r_j, Y_j, \psi]$ is difficult to sample from. Also convergence may be more difficult to discern for the marginal posterior densities for the continuous parameters θ_j. Thus we concentrate here solely on the changepoint posterior marginals.

The above analysis suggests that we may use the Gibbs Sampler algorithm to compute the k changepoint posterior marginal distributions and hence derive estimates of the unknown changepoint positions. However, to implement the algorithm, we must overcome two

difficulties. First, we must be able to decide, in some sense, when convergence has occurred. Secondly, to lessen total time to convergence of the algorithm, it is important that an adequate set of starting values $(r_{10}, \ldots, r_{k0})$ are chosen. We address each of these problems in turn.

There exist no formal criteria for the assessment of convergence of the Gibbs Sampler algorithm. Gelfand *et al.* (1988) describe graphical convergence diagnostics based upon the stability of the marginal posterior density estimates over a range of numbers of iterations. One possibility is that the algorithm is diagnosed as converged if the spline-smoothed curves representing the density estimates at successive multiples of 10 iterations are indistinguishable by eye. This might be a reasonable diagnostic for many purposes, as it represents an overall distributional comparison. The analogue in the case of our discrete marginal posterior distributions would be to assess visually histogram similarity at 10 iterations apart, which would presumably be as easy to discern, but perhaps subject to a higher degree of fluctuation. However, we can improve on this informal procedure in the discrete case by using some measure of distance between distributions (i.e. merely a summation over the finite and discrete range of each variable of some suitably chosen function of two successive marginal probability estimates) and some stopping criterion - for instance, we could regard convergence to have occurred when the total squared distance between successive estimates of the marginal distributions is less than some constant for each of the changepoint variables - this option is not generally readily available in the continuous case. We note this possible approach, but reject it on the grounds of inefficiency in favour of another convergence diagnostic technique in the spirit of those used by Geman and Geman (1984) and Ripley (1988), namely via what we broadly term "summary statistics", an alternative rejected by Gelfand *et al.* principally due to its inefficiency, but also presumably due to their different objectives (their interest being in reporting the posterior density as a whole). In the image processing context, Geman and Geman assess convergence of the Gibbs Sampler (there used in conjunction with the simulated annealing technique) in terms of number of pixel "flips"; that is, changes in (the relevant) posterior modal estimates. Ripley assesses convergence by monitoring (in the annealing process) the magnitude of the energy function in the exponent of the (Gibbs) posterior distribution, a quantity related to modal posterior probability. These seem intuitively more appealing and relevant approaches to our problem, as they specifically relate to the quantity of interest (in our case these would be modes in the marginal posterior distributions) and are potentially more expediently implemented. Thus, for the moment at least, we regard convergence as having occurred when the position of the mode of each marginal changepoint posterior distribution between iterations a fixed number apart has stabilised.

Therefore, a possible scheme for implementation of the Gibbs Sampler algorithm in the multiple changepoint case is as follows. Given the set of starting values $(r_{10}, \ldots, r_{k0})$, sample once from each of the conditional posterior distributions $[r_j \mid r_{j-1}, r_{j+1}, Y, \psi]$ by c.d.f.

inversion and then iterate the procedure, setting the conditioning variables r_j equal to the most recent value of r_{jt} obtained. After t_0 such iterations, obtain values $(r_{1t_0}, \ldots, r_{kt_0})$. Repeat this procedure m times, and obtain m *i.i.d.* replicates, $(r_{1t_0}^{(i)}, \ldots, r_{kt_0}^{(i)})$ for $i = 1, \ldots, m$. Compute the marginal posterior distribution estimates after t_0 iterations, $[r_j \mid Y, \psi]_{t_0}$ by summing the individual probabilities in the distributions $[r_j \mid r_{j-1t_0}^{(i)}, r_{j+1t_0}^{(i)}, Y, \psi]$ over i, element by element, and dividing each element by m, i.e.

$$[r_j \mid Y, \psi]_{t_0} = \frac{1}{m} \sum_{i=1}^{m} [r_j \mid r_{j-1t_0}^{(i)}, r_{j+1t_0}^{(i)}, Y, \psi]$$

or

$$\Pr(r_j = r \mid Y, \psi)_{t_0} = \frac{1}{m} \sum_{i=1}^{m} \Pr(r_j = r \mid r_{j-1t_0}^{(i)}, r_{j+1t_0}^{(i)}, Y, \psi)$$

the discrete analogue of the mixture density estimate in Gelfand *et al.* Locate the mode of each posterior distribution estimate, $[r_j \mid Y, \psi]_{t_0}$, the vector of those marginal modes being denoted by $(\hat{r}_{1t_0}, \ldots, \hat{r}_{kt_0})$. Then, using the replicates $(r_{1t_0}^{(i)}, \ldots, r_{kt_0}^{(i)})$ as starting values, sample again from the conditional posterior distributions until the t_1'th iteration is complete, $t_1 = 2t_0$, thus producing a new set of m *i.i.d.* replicates $(r_{1t_1}^{(i)}, \ldots, r_{kt_1}^{(i)})$. Compute the estimates $[r_j \mid Y, \psi]_{t_1}$ in the same way as above, and again locate their modes $(\hat{r}_{1t_1}, \ldots, \hat{r}_{kt_1})$. If the successive modal positions coincide, or $\hat{r}_{jt_0} = \hat{r}_{jt_1}$, for each j, we deem convergence to have occurred, otherwise we repeat the three steps of the procedure (calculate and sample from the full conditional distributions for another t_0 iterations and m replications, compute marginal distribution estimates, locate modes) for $t_2, t_3, \ldots$, where $t_j = (j+1)t_0$, until convergence after t_c iterations, when $\hat{r}_{jt_{c-1}} = \hat{r}_{jt_c}$ for all j.

It might seem at first sight that the amount of computation involved in such a scheme is potentially very large. Using the values proposed by Gelfand *et al.* of $t_0 = 10$ and $m = 50$, we must evaluate each of the k conditional posterior distributions $[r_j \mid r_{j-1}, r_{j+1}, Y, \psi]$ a total of $t_0 m = 500$ times between comparisons of modal positions, and in the process generate one variate from each distribution. The c.d.f. inversion technique is a "one-for-one" sampling scheme, and therefore we must generate a total of $t_0 mk = 500k$ variates uniformly from $(0, 1)$ during the same period. However, as we shall see, in many cases we are able to choose t_0 and m to be considerably smaller than those values proposed above, and still obtain adequate results.

The remaining problem concerning the implementation of the Gibbs Sampler algorithm is that of choosing the set of starting values $(r_{10}, \ldots, r_{k0})$. We might naively choose k values uniformly spaced in, or randomly chosen and suitably ordered from, the set $\{1, \ldots, n-1\}$.

Using this approach, however, it is clear that because of the properties of the conditional distribution $[r_j \mid r_{j-1}, r_{j+1}, \boldsymbol{Y}, \psi]$ we have to some extent studied in a previous section, if two or more of the true changepoint positions lie between r_{j-1} and r_{j+1}, the algorithm will be subverted unless great care is taken in the procedure used when updating the values of the conditioning variable. An alternative choice for the set of starting values would be to choose each of the r_{j0} equal to $r_{k+1} = n$ for $j = 1, \ldots, k$. This would add an extra element in that we might expect the set $(r_{11}, \ldots, r_{k1})$ to correspond quite closely to the true changepoint positions, due to the nature of the conditional posterior distributions.

We now present an illustrative example of the use of the Gibbs Sampler algorithm in the multiple changepoint problem.

(3.6.1) Approximation of changepoint marginal posterior distributions, $k = 2$.

Consider a sequence $\boldsymbol{Y}$ having $k = 2$ changepoints at unknown positions represented as random variables r_1, r_2. We wish to make inference about these positions marginally on the basis of a realisation $\boldsymbol{y}$ of $\boldsymbol{Y}$, and certain prior assumptions. Specifically, assume that $(Y_1, \ldots, Y_{r_1})$ are identically distributed as $N(\theta_1, \sigma^2)$, $(Y_{r_1+1}, \ldots, Y_{r_2})$ are identically distributed as $N(\theta_2, \sigma^2)$, and $(Y_{r_2+1}, \ldots, Y_n)$ are identically distributed as $N(\theta_3, \sigma^2)$. Assuming also the Y_i's to be conditionally independent given $\theta = (\theta_1, \theta_2, \theta_3, \sigma)$, and that θ is unknown and that the form of $[\theta_j, \theta_{j+1}, \sigma]$ is identical to (2.10), it is easily seen that if we again define $r_0 = 0$ and $r_{k+1} \equiv r_3 = n$, and choose $[r_j \mid r_{j-1}, r_{j+1}]$ to be uniform, the form of the conditional posterior distribution $[r_j \mid r_{j-1}, r_{j+1}, \boldsymbol{Y}, \psi]$ is identical to (2.11), i.e.

$$[r_j \mid r_{j-1}, r_{j+1}, \boldsymbol{Y}, \psi] \propto \{(r_j - r_{j-1})(r_{j+1} - r_j)\}^{-1/2} \left\{ \sum_{i=r_{j-1}+1}^{r_j} (Y_i - \overline{Y_{Aj}})^2 + \sum_{i=r_j+1}^{r_{j+1}} (Y_i - \overline{Y_{Bj}})^2 \right\}^{-n_j/2} \quad (3.19)$$

where

$$\overline{Y_{Aj}} = \frac{1}{(r_j - r_{j-1})} \sum_{i=r_{j-1}+1}^{r_j} Y_i$$

$$\overline{Y_{Bj}} = \frac{1}{(r_{j+1} - r_j)} \sum_{i=r_j+1}^{r_{j+1}} Y_i$$

$$n_j = (r_{j+1} - r_{j-1})$$

and r_j takes values in $\{r_{j-1} + 1, \ldots, r_{j+1} - 1\}$. If we allow a non-zero prior probability of "no changepoint" in $\{r_{j-1} + 1, \ldots, r_{j+1}\}$, we make the relevant adjustment to (1.24), and choose $[r_j \mid r_{j-1}, r_{j+1}]$ to be of the form

$$[r_j \mid r_{j-1}, r_{j+1}] = \begin{cases} \dfrac{(1-p)}{(r_{j+1} - r_{j-1} - 1)} & r_j = r_{j-1}+1, \ldots, r_{j+1}-1 \\ p & r_j = r_{j+1} \end{cases} \quad (3.20)$$

for some p ($0 \leq p \leq 1$). However, were we to choose p non-zero then we would be allowing the possibility of coincident changepoint positions on any single iteration, and hence effectively a reduction of k on subsequent iterations, which we feel may disrupt the Gibbs Sampler algorithm unduly. For the moment we choose $p = 0$. As suggested above we choose (r_{10}, r_{20}) to equal n. We now proceed to compare the marginal posterior distributions for r_1 and r_2 obtained, the processing time involved, and the number of iterations to convergence using the Gibbs Sampler with various choices of t_0 and m, for several pairs (r_1^*, r_2^*) of true changepoint positions and Signal-Noise ratios.

First, we study the effect that varying t_0 and m has on processing time. The entries in table 1 are the amounts of processing time required for the Gibbs Sampler to converge for various pairs of choices (t_0, m) using the stable-mode convergence criterion averaged over 200 runs. The sequences of length 80 generated had true changepoints $r_1^* = 24$ and $r_2^* = 56$, $\theta_1 = \theta_3$, and a Signal-Noise ratio of 2.0.

		t_0				
		1	2	3	5	10
m	1	0.1178	0.1530	0.1846	0.2519	0.4274
	3	0.2518	0.3609	0.4603	0.6748	1.1828
	5	0.3956	0.5604	0.7301	1.0770	1.9492
	10	0.7277	1.0798	1.4280	2.1357	3.9022
	20	1.4256	2.0974	2.7737	4.1068	7.7784

Table 1

On inspection of the the timings in table 1, it is clear that computation time increases more quickly with m than with t_0 - this is encouraging, as m relates primarily to the adequacy of the estimate of the marginal distribution, which is not our chief concern, and only acts as a secondary factor in relation to assessment of rate of convergence (i.e. through the convergence diagnostic, posterior modal position, which we might expect to be fairly stable even for small m in the majority of cases). It should also be noted that, for this particular admittedly

straightforward example the algorithm was diagnosed as converged on iteration $2t_0$ in practically all of the 200 repetitions of the analysis.

Thus it only remains to compare the results obtained using the Gibbs Sampler with those obtained from an "exact" analysis using (3.2). Figures 31(a) and (b) depict the exact marginal posterior distributions of r_1 and r_2 respectively, evaluated using (3.2) and marginalisation, averaged over 200 runs. Figure 31(c) and (d) depict the "approximate" marginal posterior distributions obtained using Gibbs Sampler for various choices of t_0 and m.

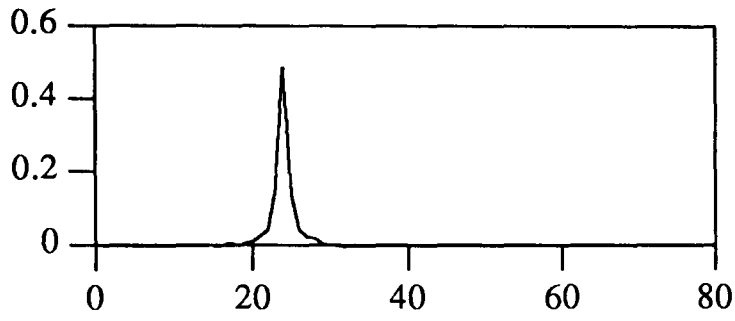


Fig 31(a) : Exact margin of r_1

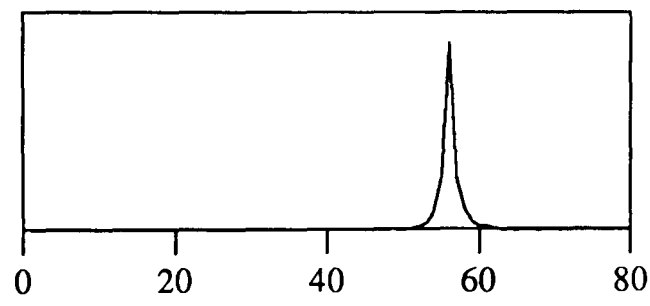


Fig 31(b) : Exact margin of r_2

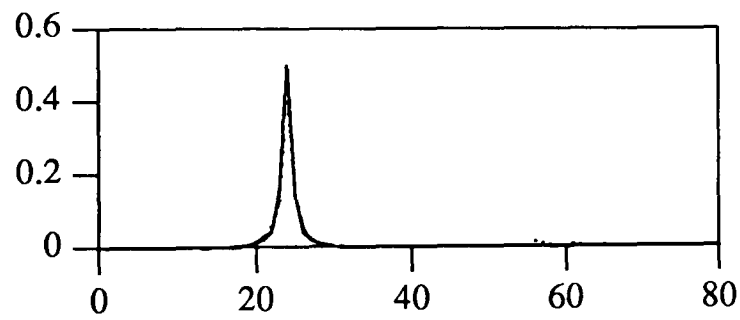


Fig 31(c) : Approximate margin of r_1

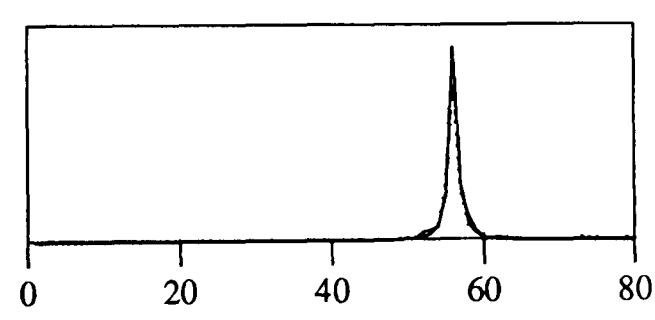


Fig 31(d) : Approximate margin of r_2

The solid, dashed and dotted lines in (c) and (d) correspond to (10,20), (3,5) and (1,1) as choices for (t_0, m) respectively. The three are virtually indistinguishable, and all approximate the true marginal posterior distribution more than adequately. This illustrates the considerable potential of the Gibbs Sampler technique for the multiple changepoint problem.

The evaluation of the exact posterior probabilities in figures 31(a) and (b) took an average of 0.5355 seconds over the 200 runs. Thus, in many cases for this particular example, the Gibbs Sampler gives adequate results in a shorter time than the exact analysis. We now repeat the analysis with $r_2^* = 30$ at the same Signal-Noise ratio and under the same prior assumptions. The results are depicted in figure 32.

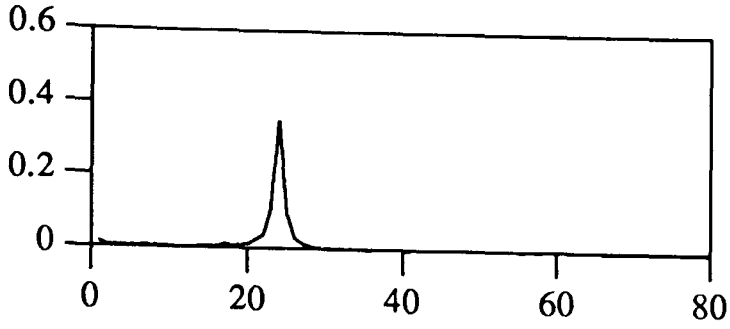


Fig 32(a) : Exact margin of r_1

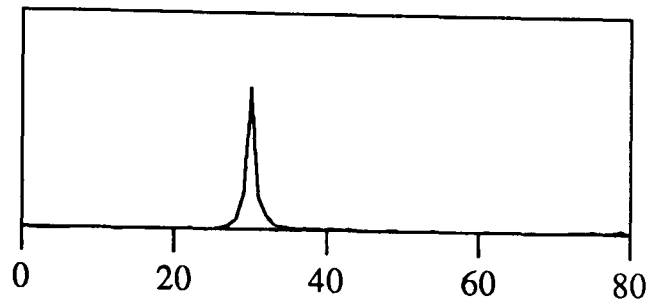


Fig 32(b) : Exact margin of r_2

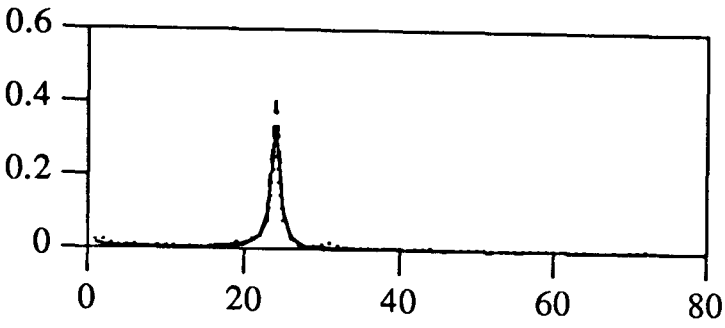


Fig 32(c) : Approximate margin of r_1

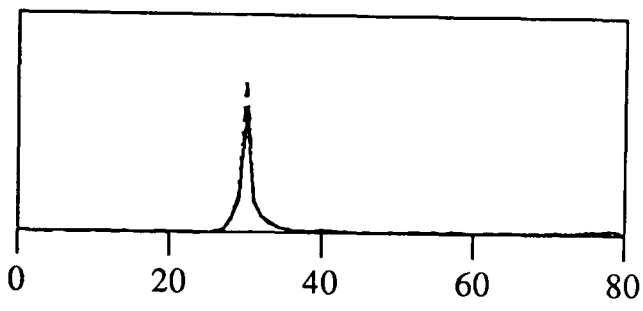


Fig 32(d) : Approximate margin of r_2

In figure 32(c) and (d), the solid, dashed and dotted lines in (c) and (d) correspond to (5,3), (5,1) and (1,1) as choices for (t_0, m) respectively. Again, it is clear that in each case the estimate of the marginal posterior distribution approximates the exact marginal posterior distribution well. The relative timings involved were comparable to those in table 1, and the latter two cases were appreciably faster than the evaluation of the exact margins. Again, for this example, convergence was generally perceived to have occurred after $2t_0$ iterations in the majority of cases.

We now study the effect that altering Signal-Noise ratio has on the efficiency of the Gibbs Sampler in producing estimates of marginal distributions and modes. Fixing $r_1^* = 24$ and $r_2^* = 40$, we vary the Signal-Noise ratio between 2.0 and 1.0 in decrements of 0.5. The results are depicted in the series figure 33 to figure 35. In each case, (a) and (b) depict the exact marginal distributions of r_1 and r_2, respectively, obtained via (3.2), and (c) and (d) depict the approximate marginal distributions obtained using the Gibbs Sampler, with the pair (t_0, m) chosen as (5,1) for demonstration purposes.

The results are clearly impressive (in the sense that the approximation is excellent) even at low Signal-Noise ratios. It is interesting to note the rate of convergence in each case. In figure 33(c) and (d), with S.N.R equal to 2.0, convergence was diagnosed on average after 11.12 iterations (approximately $2t_0$) and the average processing time was 0.2501 seconds. In figure 34(c) and (d), S.N.R. equal to 1.5, the corresponding averages were 14.65 ($3t_0$) and 0.3199 seconds, and in figure 35(c) and (d), the averages were 26.67 ($5t_0$) and 0.5471 seconds respectively. Hence, as we might have predicted, the rate of convergence decreases with decreasing Signal-Noise ratio. In the latter case, the processing time involved when using the Gibbs Sampler was comparable with that involved when evaluating the marginal posterior

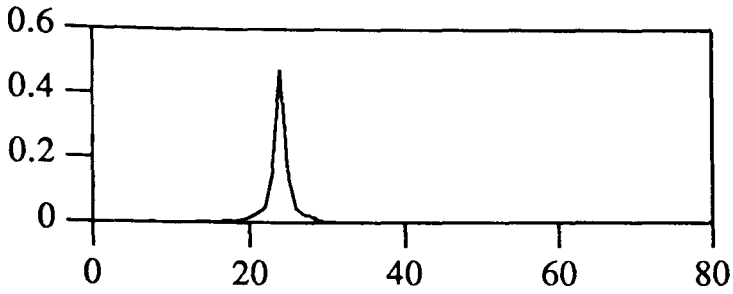


Fig 33(a) : exact

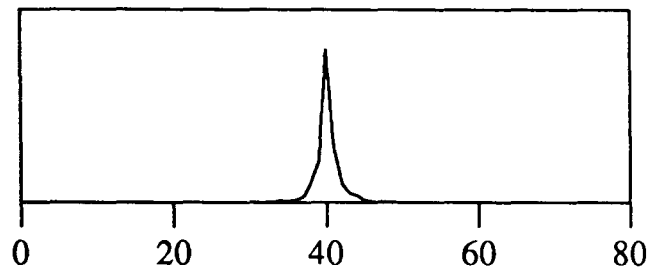


Fig 33(b) : exact

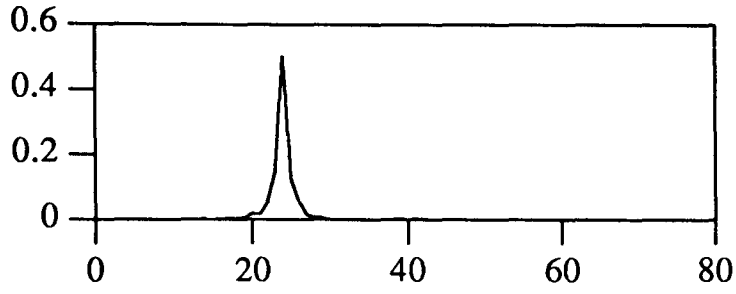


Fig 33(c) : approximate

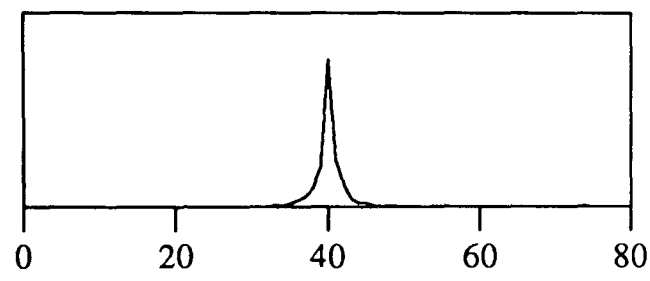


Fig 33(d) : approximate

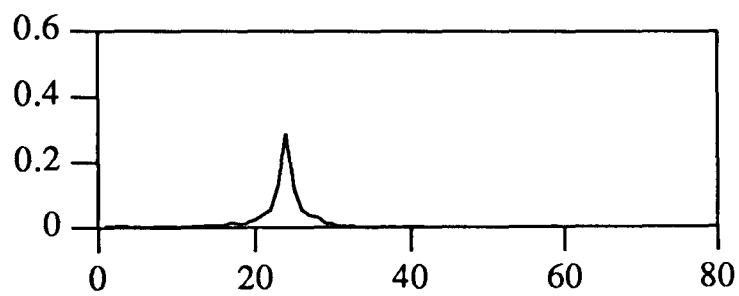


Fig 34(a) : exact

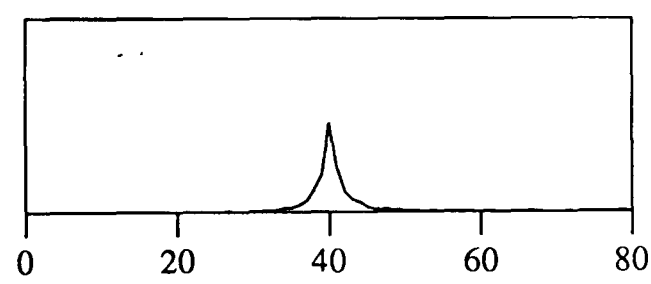


Fig 34(b) : exact

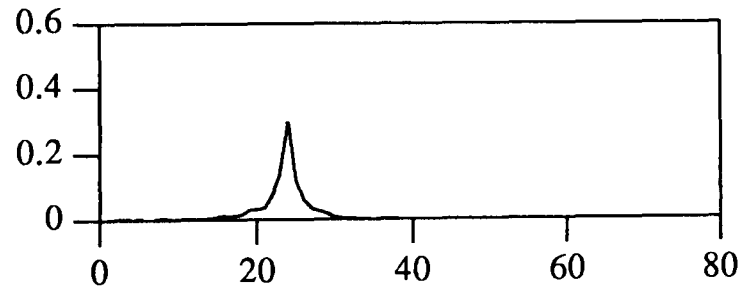


Fig 34(c) : approximate

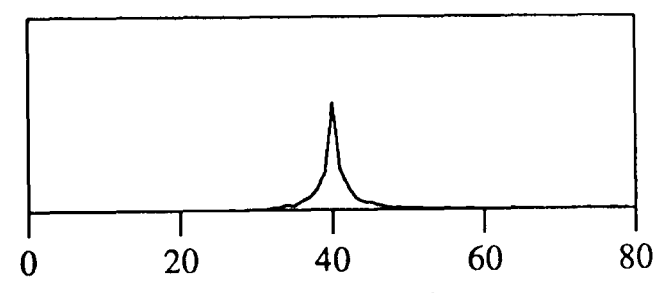


Fig 34(d) : approximate

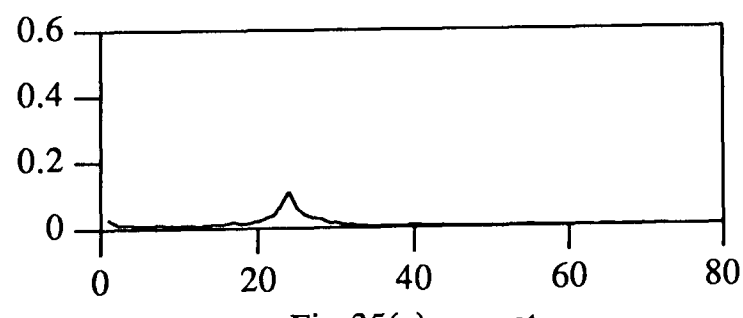


Fig 35(a) : exact

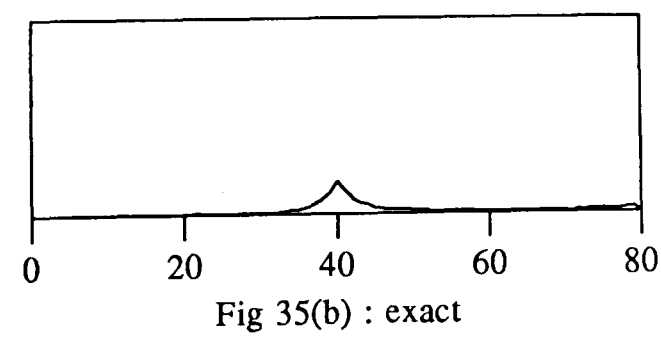


Fig 35(b) : exact

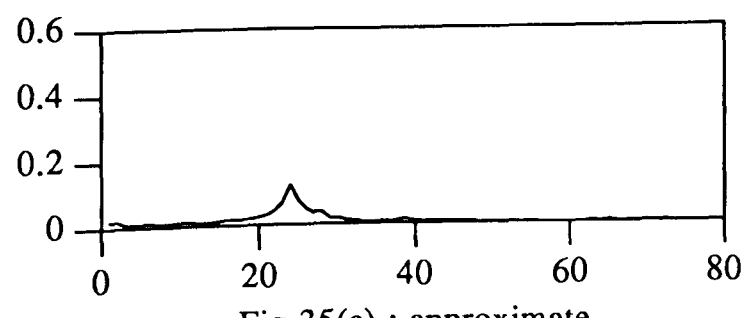


Fig 35(c) : approximate

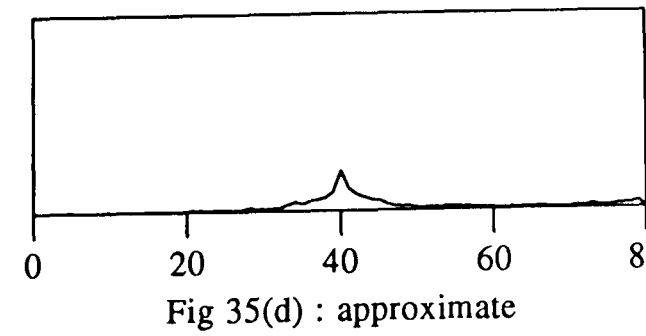


Fig 35(d) : approximate

distributions exactly, which conflicts with our motivation for using the Gibbs Sampler in the multiple changepoint problem (potential computation time reduction in the image-processing context). We need not be unduly worried, however, as we shall see below that the saving incurred for larger k is appreciable.

(3.6.2) Approximation of changepoint marginal posterior distributions, $k = 3$.

We now investigate the behaviour of the Gibbs Sampler for larger k. Consider the case $k = 3$. Using the same prior specifications as above, it is again straightforward to evaluate the conditional posterior distributions, and each takes precisely the same form as (3.19). The Gibbs Sampler algorithm then requires that we should sample and update the conditioning variables iteratively to convergence. We might intuitively expect processing time to increase linearly with k. For demonstration and comparative purposes, we investigate the performance of the Gibbs Sampler on Normal sequences having changepoints at 24, 40, and 66, with corresponding mean levels 0.0, 2.0, 4.0, and 1.0. The margins obtained via the Gibbs Sampler (averaged over 200 runs) and the exact margins obtained using the three changepoint equivalent to (3.2) are depicted in Figure 36.

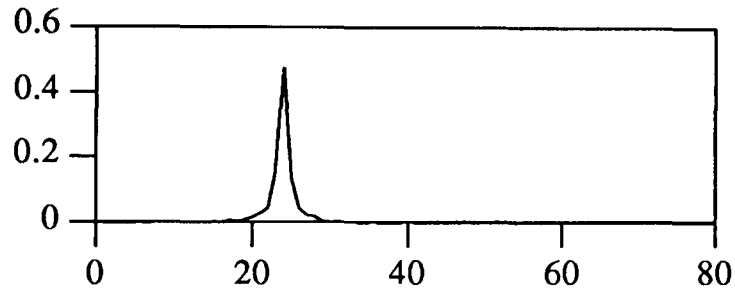


Fig 36(a) : Exact margin of r_1

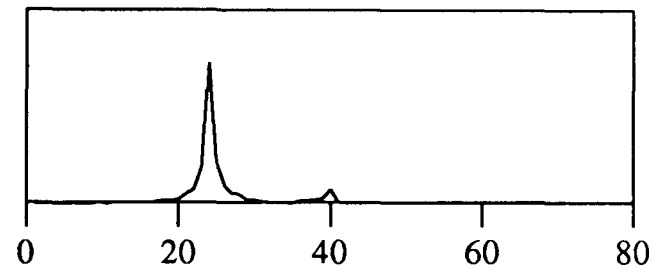


Fig 36(b) : Approximate margin of r_1

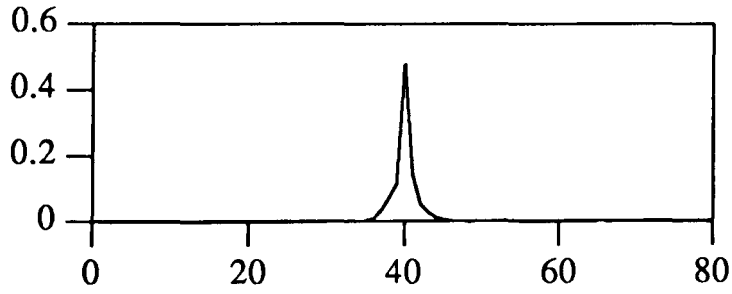


Fig 36(c) : Exact margin of r_2

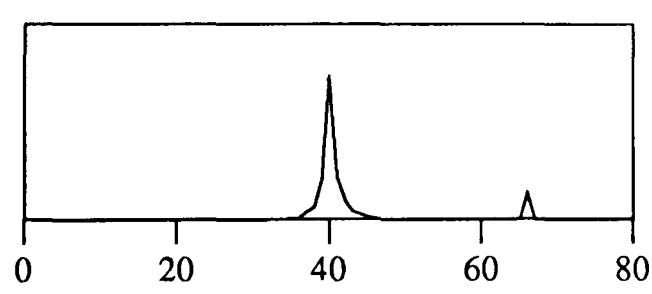


Fig 36(d) : Approximate margin of r_2

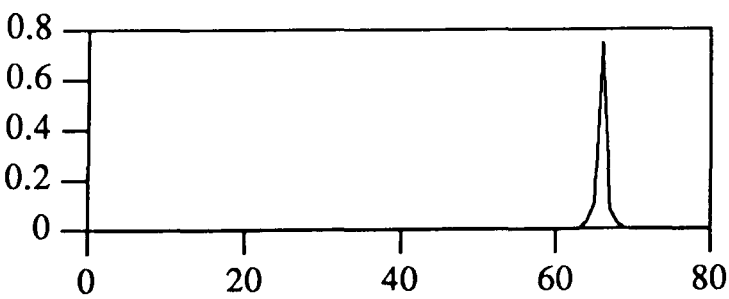


Fig 36(e) : Exact margin of r_3

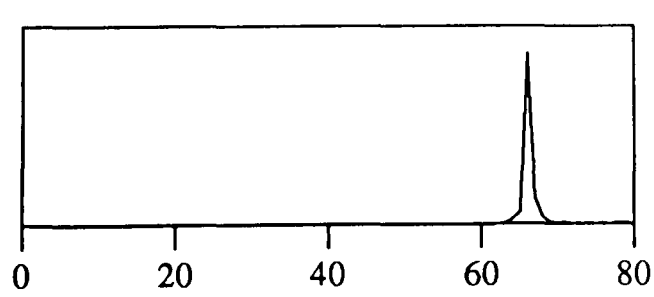


Fig 36(f) : Approximate margin of r_3

The results of the Gibbs Sampler depicted in Figure 36(b), (d), and (f) were obtained using $t_0 = 20$ and $m = 1$ to show the potential of the algorithm. It is clear that the approximate margin is very close to the exact margin in each case. For these values of t_0 and m, the

average processing time was 0.9386 seconds, and the average number of iterations needed for the algorithm to converge was 45.20 ($2t_0$). The exact analysis required an average of 17.40 seconds of processing time for calculation of the margins. Thus we now have a 17-fold time-saving in calculation of the margins in the three changepoint case, compared to approximately a 3-fold saving in the two changepoint case.

It is interesting to study the results obtained using the Gibbs Sampler in the three-changepoint case when the value of t_0 is varied. Figure 37 depicts the r_2 margin resulting from a Gibbs Sampler analysis carried out with t_0 taking the values 15, 10, 5, 1 in the series (a) to (d), with convergence being assessed by modal position stability in the usual way.

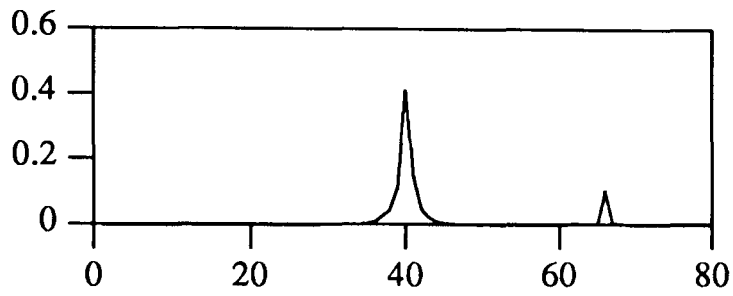


Fig 37(a) : t_0 = 15

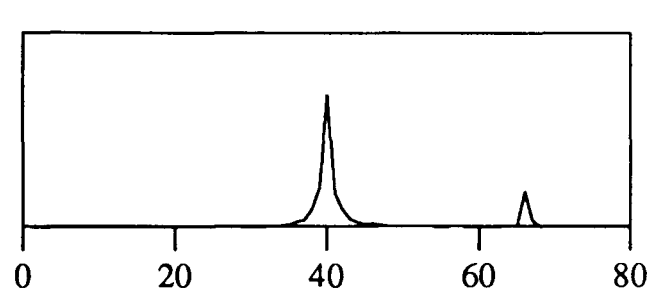


Fig 37(b) : t_0 = 10

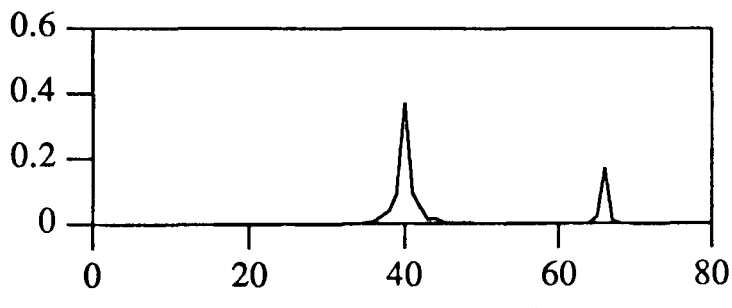


Fig 37(c) : t_0 = 5

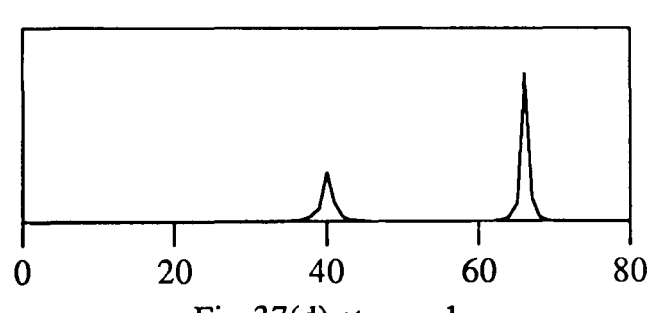


Fig 37(d) : t_0 = 1

These results illustrate one minor drawback of our implementation of the Gibbs Sampler to evaluate the marginal distributions. We must take great care over the assessment of convergence, since when using the same convergence diagnostic as above with a range of values of t_0, the resulting marginal distribution estimates are markedly different. Choosing t_0 large enough practically ensures convergence to the "correct" distribution but increases processing time. One possible alternative implementation would involve choosing t_0 large initially, completing the sampling sampling cycle t_0 times and evaluating the marginal distributions and modal positions, and then re-calculating the distributions and modes after every subsequent or alternate iteration, assessing convergence in the usual way. This would hopefully ensure that the Gibbs Sampler had "settled down" to the correct values of the changepoint positions before any assessment of convergence is made. Results of an analysis using this alternative scheme - choosing t_0 = 20 initially then inspecting modal positions after each subsequent iteration - are depicted in figure 36. Again, only the r_2 margin is shown.

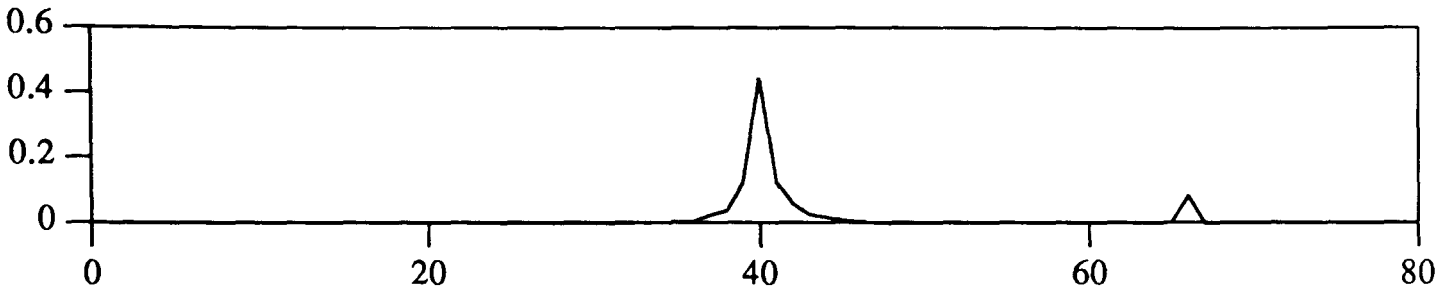


Fig 38(a) : Approximate margin of r_2, $t_0 = 20$

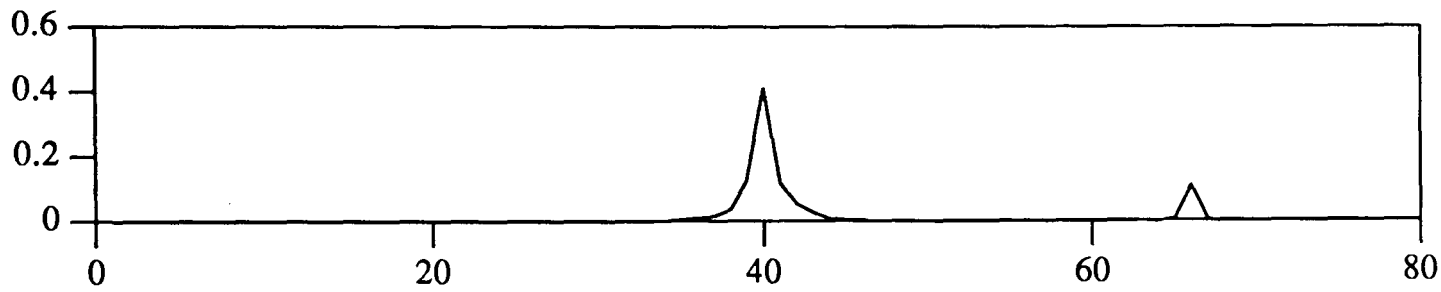


Fig 38(b) : Approximate margin of r_2, alternative scheme

It is clear that (a) and (b) are very similar. The processing time involved in producing (a) and (b) was 0.9386 and 0.5117 seconds respectively. Thus we have practically halved the amount of processing time needed by using the alternative scheme, and achieved comparable results. Further processing time reductions may be achieved in a similar way.

Two other points should be noted. First, as mentioned previously, we could assess convergence via a distance measure between successive marginal distribution estimates. This would presumably help to eliminate problems associated with the choice of t_0 and provide some notion of rate of convergence. However, use of this method would present its own difficulties. In addition to those mentioned above concerning computation time, we would be forced to introduce scales of distance, and stopping-rules etc. which would further complicate practical implementation. Secondly, in the context of the edge-detection problem, as our interest is principally in marginal posterior modal positions, we need not be overly concerned with the nature of the distribution estimates themselves provided that we gain sufficient and accurate information from a Gibbs Sampler analysis to enable us to report the posterior modal positions. This implies that, provided we are not primarily concerned with modal posterior probability for each margin, we can afford to choose t_0 smaller than is necessary in other problems. For example, in the case depicted in figure 37, we gain information about the position of both the second and third changepoints from the r_2 margin (much in the same way that we could have made inference about two changepoint positions when using the one changepoint posterior distribution (2.11) in a previous section). The precise behaviour seen in figure 37 is understandable given our experience with changepoint sequences when we recall that the magnitude of the mean-level change at r_3^* is 3.0, compared with 2.0 at r_2^*, so we might expect a "short-term" convergence effect in the iterative procedure to r_3^*.

Finally, before we demonstrate the application of this sampling-based method in the edge-detection problem, we propose a simplification/approximation to the full Gibbs Sampler algorithm with the intention of further reducing processing time. Recall the set of full conditional posterior distributions $[r_j \mid r_{j-1}, r_{j+1}, Y, \psi]$, $j = 1, \ldots, k$, and the set of starting values $(r_{10}, \ldots, r_{k0})$. Then the simplified algorithm proceeds as follows. Compute $[r_1 \mid r_0, r_{20}, Y, \psi]$, and instead of sampling from this distribution, choose the value r_{11} to be that at which the distribution is maximised. Then choose r_{21} to be that value which maximises $[r_2 \mid r_{11}, r_{30}, Y, \psi]$ etc. until the modal positions stabilise. It is reasonable to expect the set $(r_{1t_c}, \ldots, r_{kt_c})$ to coincide with the actual changepoint positions after convergence at iteration t_c in many straightforward instances. This approximation to the full Gibbs Sampler procedure follows the I.C.M. approximation to the maximum probability estimates in image segmentation problems discussed in section (1.4.3.1) of chapter 1. It also has links with the binary segmentation technique proposed in section (3.3.2).

We now demonstrate the use of this approximation to the Gibbs Sampler in the two and three changepoint cases. First, we investigate its behaviour for two changepoint sequences identical to those in figure 33, i.e. with the change in mean-level at r_2^* equal in magnitude but opposite in sign to that at r_1^*. We choose $r_1^* = 24$ and $r_2^* = 50$, and a Signal-Noise ratio of 2.0, and make the same prior assumptions and hence use the same functional form (3.19) for the full conditional posterior distributions. The results of 1000 repetitions of the analysis are presented in figure 39.

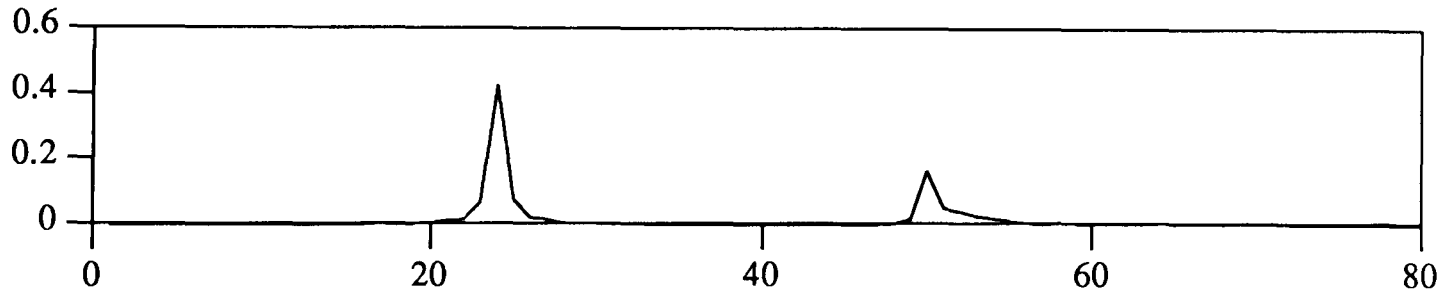


Fig 39(a) : Distribution of mode , r_1 margin

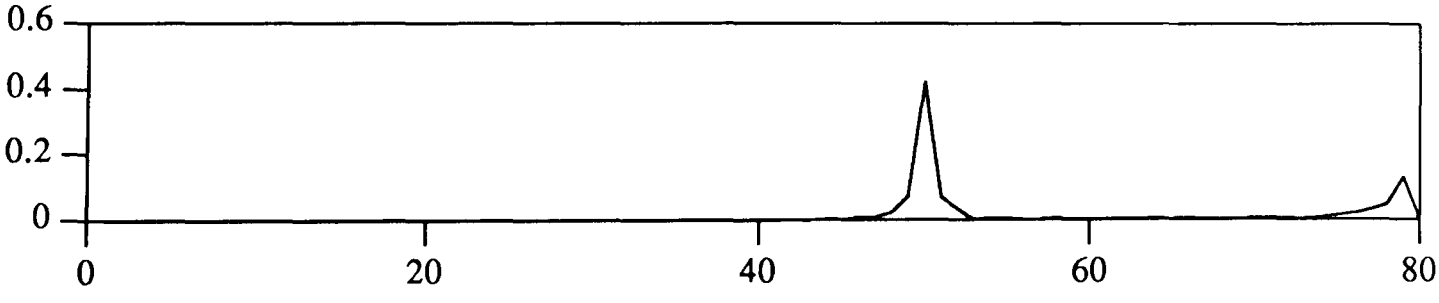


Fig 39(b) : Distribution of mode , r_2 margin

The proportion of occurrences of each point being detected as a mode at convergence are plotted on the vertical scale (hence figure 37 should not strictly be thought of as depicting the marginal probability distributions of r_1 and r_2, but as expected frequency distributions of the modal positions of the marginal distributions of r_1 and r_2). It is clear from figure 39 that this

approximate method provides adequate results in the two changepoint detection case. The average processing time involved in producing these results was 0.0823 seconds, and the algorithm had generally converged before the fourth iteration. This represents a time saving of at least one-third compared with the timings in table 1.

We now investigate the three changepoint case, in particular sequences identical to those represented in figure 36, with true changepoints at 24, 40, and 66, and mean-levels 0.0, 2.0, 4.0 and 1.0, and a common variance of 1.0. Again 1000 repetitions of the analysis were carried out. The relative expected frequencies are depicted in figure 40.

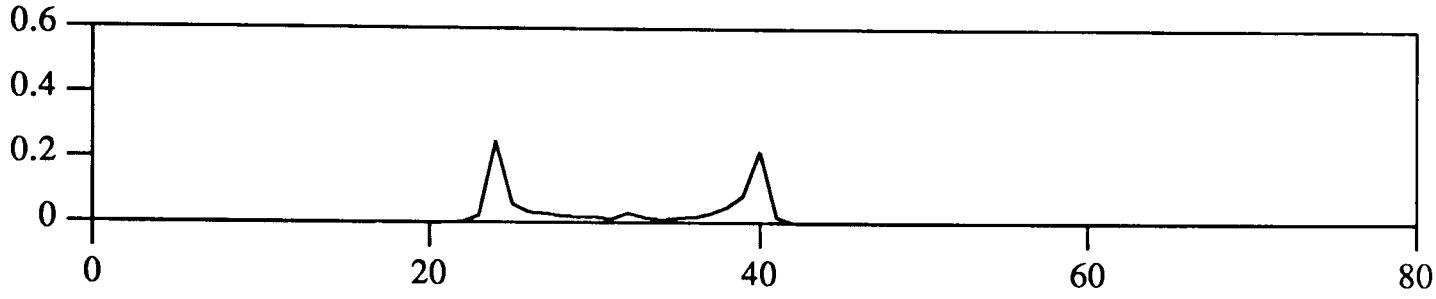


Fig 40(a) : Distribution of mode, r_1 margin

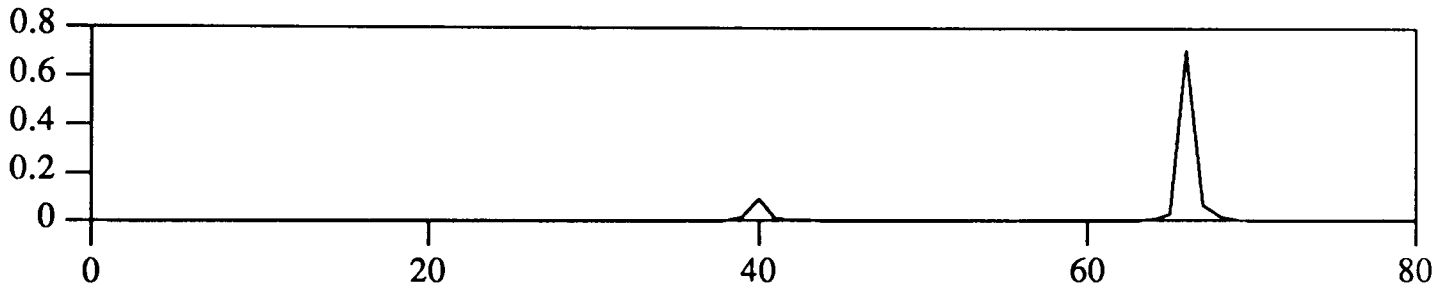


Fig 40(b) : Distribution of mode, r_2 margin

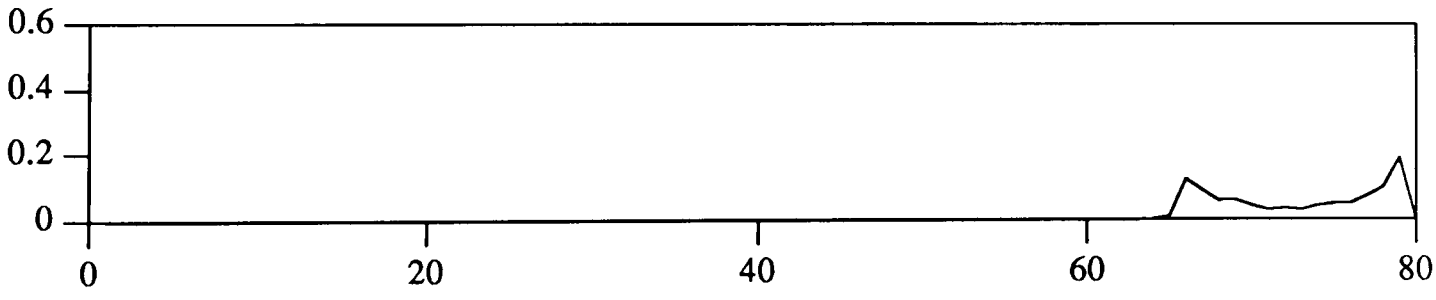


Fig 40(c) : Distribution of mode, r_3 margin

These results are less adequate than those obtained previously, due presumably to the fact that the Signal-Noise ratio at r_3^* is appreciably larger than at either r_1^* or r_2^*, and hence we again get a "short-term" convergence effect. However, we still gain some information as to the true changepoint positions by combining the results from the three margins in some way. The average processing time involved in this case was 0.0998 seconds.

Note: throughout the above we have assumed the number of changepoints k in any analysis to be known. This is of course practically unrealistic. However, it can be verified that if k is mis-specified relative to the true nature of the data sequence, then results at convergence are broadly satisfactory, that is, we gain sufficient information from the margins individually and

jointly to be able to infer the positions of the true changepoints. Figure 41 depicts the margins obtained from a Gibbs Sampler analysis of a two changepoint sequence with $r_1^* = 24$ and $r_2^* = 40$ and a Signal-Noise ratio of 2.0 (i.e. precisely as in figure 33), but with k chosen to be three.

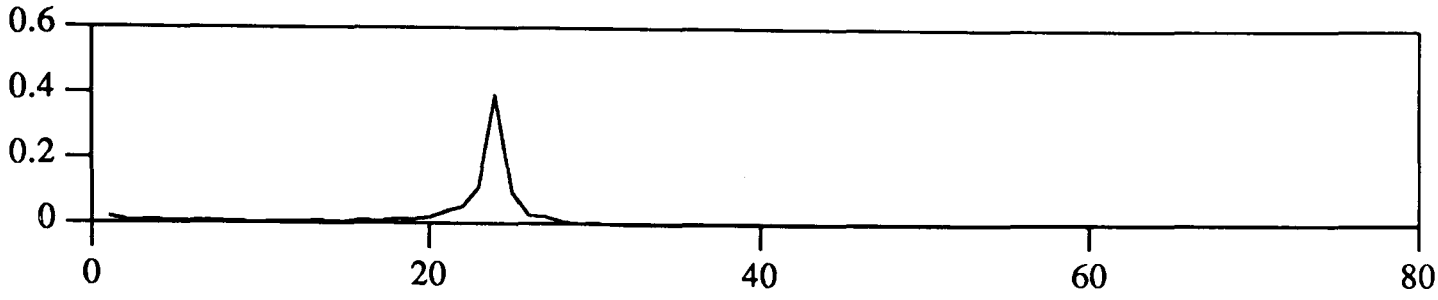


Fig 41(a) : r_1 margin

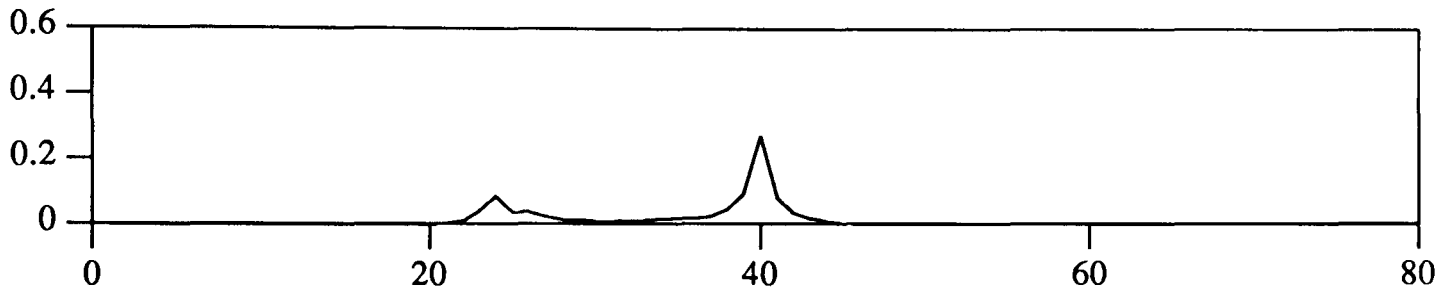


Fig 41(b) : r_2 margin

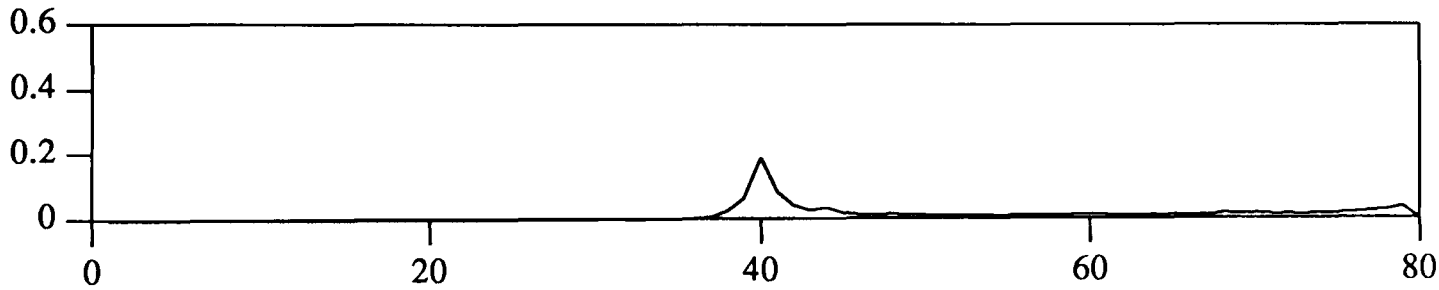


Fig 41(c) : r_3 margin

It is clear that the two changepoints are detected accurately as modes in the marginal distributions. Similar results are obtained if k is specified to be less than the actual number of changepoints in the sequence (much in the same way that the one changepoint posterior margin (2.11) provides useful information when used to analyse two changepoint sequences). However, we generally consider it advisable to specify k larger rather than smaller in the case where the actual number of changepoints is truly unknown, as the results from particular margins would then be re-enforced by the results from others.

(3.6.3) Edge-detection analysis using marginal approximation.

We now conclude this section by demonstrating the use of these approximate techniques for multiple changepoint detection with specific reference to the edge-detection problem. First, however, we make a further comment concerning the implementation of the Gibbs Sampler-based "full" changepoint analysis of an image.

We noted above the potential difficulties encountered in choosing the set of values needed to initialise the Gibbs Sampler algorithm. Now, the schemes we have suggested above for the analysis of an image have involved row-by-row and column-by-column based changepoint distribution computations with each row/column being treated independently. However, we would generally expect the edges in the true-scene to be spatially continuous. We return to this very important piece of prior knowledge in more detail at a later stage, but here we merely consider its implications for the implementation of a Gibbs Sampler-based edge-detection procedure. Consider a true-scene and image to be analysed using Gibbs Sampler techniques, with each row/column assumed to contain at most k edges/changepoints. Assuming, for example, the row analysis to be carried out in the same order as in all of the previous schemes (i.e. beginning at row 1 and with the analysis of row j immediately following the analysis of row $i-1$ for $i = 2, \ldots, n$) then it is clear that, having specified the necessary initial values for row $i-1$ and completed a Gibbs Sampler analysis of the data in that row, with k edge-point candidates selected as coinciding with modes in the k changepoint marginal posterior distributions, then because of our qualitative prior knowledge concerning spatial continuity a sensible choice for the initial values needed for the Gibbs Sampler analysis of row i would be those k values recorded as the modes for row $i-1$. This is an intuitively reasonable choice for two reasons. First, if the j'th mode in row $i-1$ corresponded to a true edge-point, then because of the assumed spatial continuity of the edge, it is likely that row i would also contain an edge-point in the vicinity of that mode, and hence we might expect an improvement in the rate of convergence of the Gibbs Sampler. Secondly, even if the j'th mode in row $i-1$ did not correspond to an edge-point but rather to an outlying or extremely noise-corrupted value, it is unlikely that row i would also contain a similarly outlying value in the vicinity of this mode, and thus due to the stochastic nature of the "updating" step of the Gibbs Sampler it is unlikely that the algorithm would converge to a point near to it (more generally, the choice of the k modes from row $i-1$ as starting values for row i can only theoretically improve the rate of convergence).

We now proceed with a demonstration of the use of the Gibbs Sampler algorithm in the analysis of a simple image. We study the results obtained from an analysis of the circle true-scene and image in figure 10. We saw above that, after making certain prior assumptions related to the nature of the image-formation process and true-scene parameters the one and two changepoint posterior distributions (2.11) and (3.2) both dealt adequately with the edge-detection problem when the Signal-Noise ratio was relatively large (> 2.0), but that the one changepoint approximation was less adequate for lower values. Figure 42(a) depicts the results of a full analysis using the one changepoint posterior distribution (2.11). Figure 42(c) depicts the results of a full analysis using the using a combination of (2.11) and the two changepoint posterior distribution (3.2). Each row/column was first analysed using (2.11) with the "no

changepoint" possibility having a positive probability, and unless the mode of the distribution was found to indicate a "no changepoint" decision (i.e. unless the mode occurred at $r = n$) then that row/column was re-analysed using (3.2), and for comparison purposes the individual margins for each changepoint were computed and their modal position and value recorded. Figure 42(b) depicts the results of a similar analysis but where the margins for the individual changepoints were computed using the Gibbs Sampler techniques discussed above, with the stable mode convergence criterion and the method for choosing initial values for the iterative step of the algorithm on the basis of the results of the previous analysis were both implemented. The values of t_0 and m were both nominally chosen to equal 1 for demonstration purposes, and k was chosen to be 2 (reasonable, if we know that the true scene contains a single convex object).

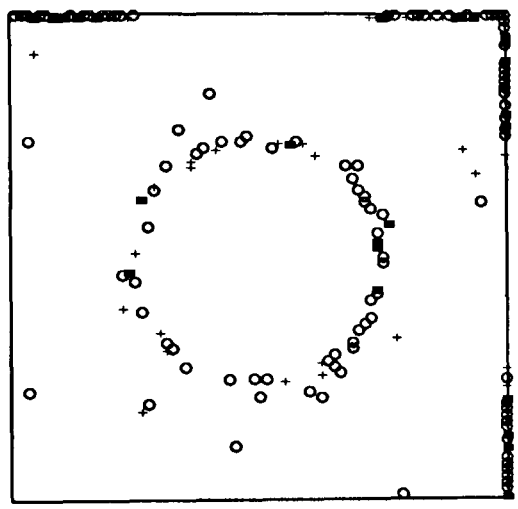


Fig 42(a) : r_1, exact

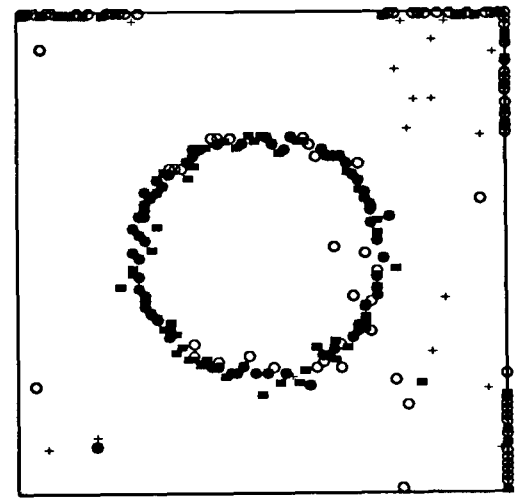


Fig 42(b) : r_1 , r_2, approximate

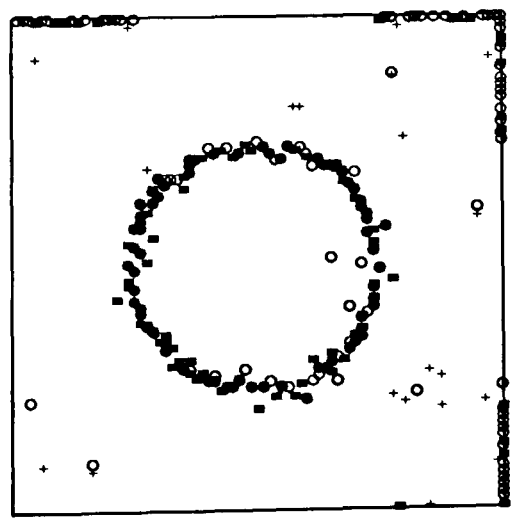


Fig 42(c) : r_1 , r_2, exact

It is clear that the results in each case are more than adequate. Recall, however, that the probabilities represented in figure 42(a) are really only an approximation to the true probabilities, as they arise from an incorrectly specified model (one instead of two changepoints per row and

column). The processing time involved in producing the results in figure 42(a) was 1.82 seconds. Now, on inspection of the results in figures 42(b) and (c), it is evident that the Gibbs Sampler-based technique has provided results which compare extremely favourably with the results arising from an exact analysis. The processing times involved in producing the results for figures 42(b) and (c) were 10.78 and 70.80 seconds respectively. Thus we have achieved at least a 7-fold time saving even in this straightforward case by using the approximate method. This is very encouraging when we consider that our primary motivation for introducing Gibbs Sampler-based techniques was to lessen computation time for multiple changepoint/edge true scenes and images.

It is of interest to study the results of the three analyses on the same true scene but where the Signal-Noise ratio in the image is decreased to 1.0. Figure 43 depicts the results of the analyses of such an image. Precisely the same schemes were used for (b) and (c) as in the example above.

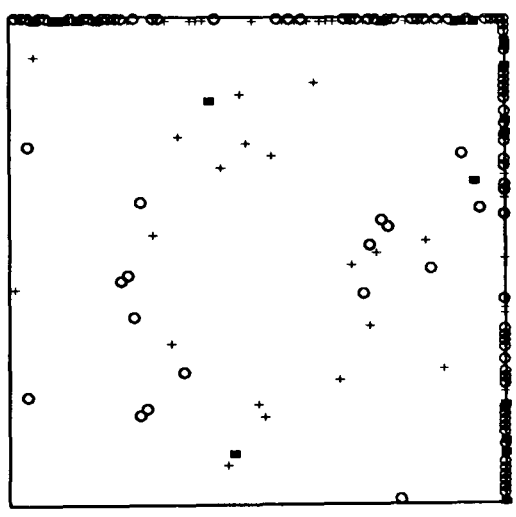


Fig 43(a) : r_1, exact

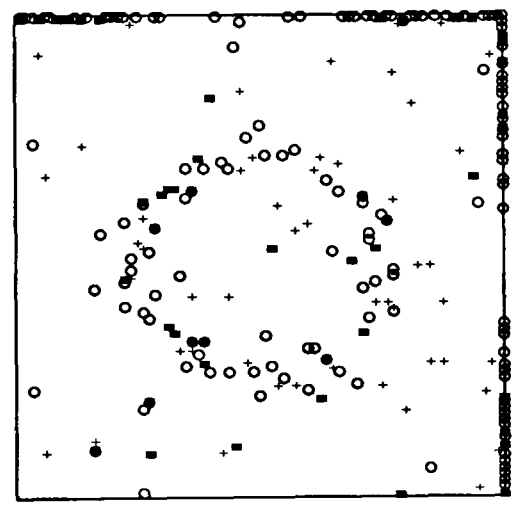


Fig 43(b) : r_1 , r_2, approximate

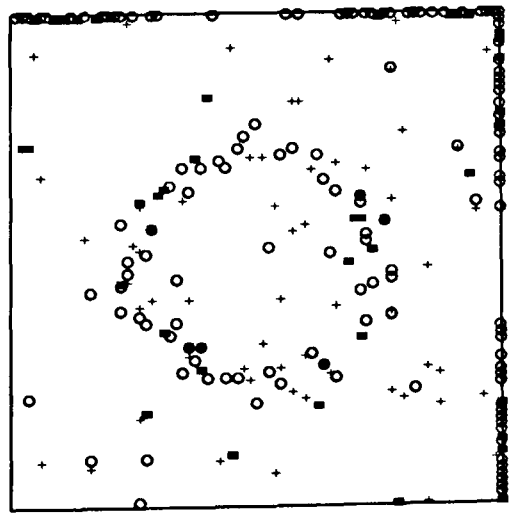


Fig 43(c) : r_1 , r_2, exact

In this, a more difficult case, the one changepoint technique does not provide adequate results.

However, again, the results in (b) and (c) compare favourably, and give a reasonable indication as to the position of the edge in the image. The processing times involved in the production of these results were 1.78, 10.44, and 55.88 seconds respectively, and so again the Gibbs Sampler-based technique seems preferable.

The results depicted in figures 42 and 43, in conjunction with the reduction of the amount processing time required are encouraging. The Gibbs Sampler technique for the approximation of marginal posterior distributions makes the changepoint-based edge-detection analysis of multi-region true scenes. We now proceed to demonstrate the use of this technique in the analysis of such true scenes.

(3.7) Analysis of multiple region true scenes.

We now turn to a more complex true-scene and investigate whether the Gibbs Sampler techniques cope adequately with the added complexity. Figures 44(a) and (b) represent an artificial true scene and image comprising a square and rectangle having different textures denoted θ_2 and θ_3 as before, on a background texture denoted θ_1.

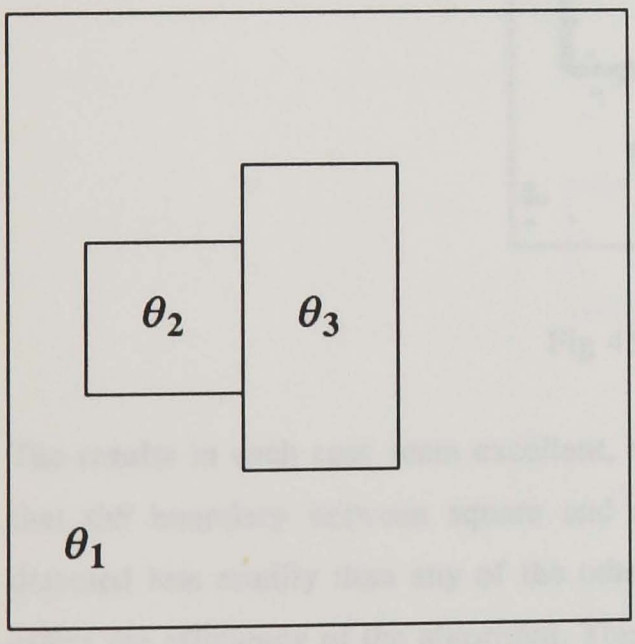


Fig 44(a) : true scene

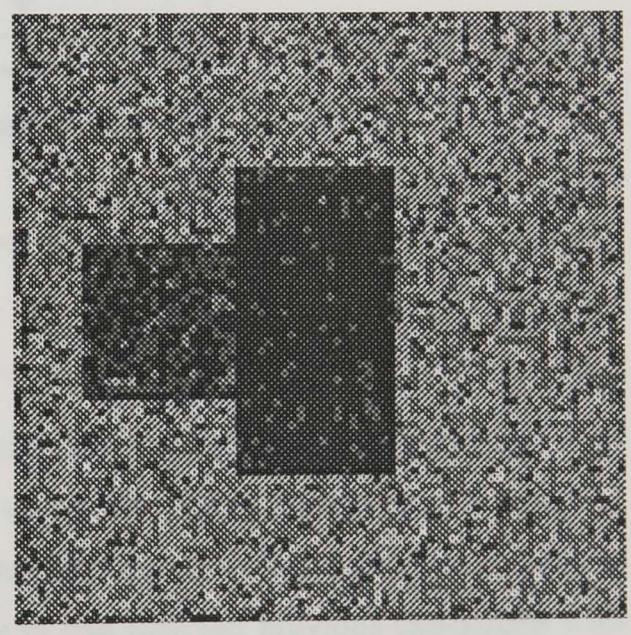


Fig 44(b) : image

For illustrative purposes, we assume the same form for the image-formation process as before with additive Gaussian white-noise of constant variance 0.5 corrupting each pixel independently, and choose mean-levels of 0.0, 2.0, and 3.0 for the three textures respectively. We proceed to analyse this image using Gibbs Sampler-based techniques. Figure 45 depicts the results obtained using the implementation of the Gibbs Sampler suggested above (preliminary "no changepoint" analysis, choice of initial values, convergence etc.) for various choices of the pair (t_0, m), with k chosen to be three in the analysis of each row/column (clearly this represents a mis-specification of k in some of the rows/columns).

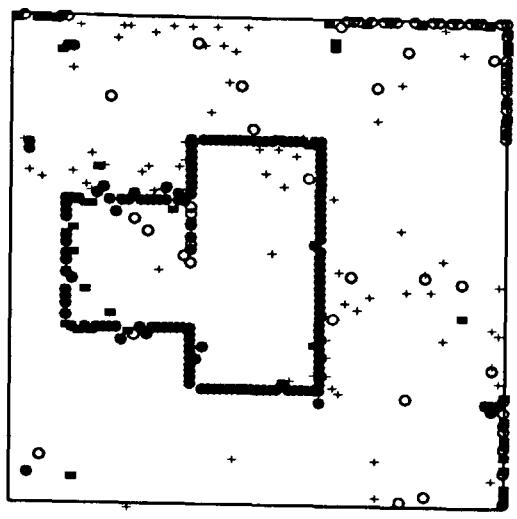


Fig 45(a) : $t_0 = 1$, $m = 1$

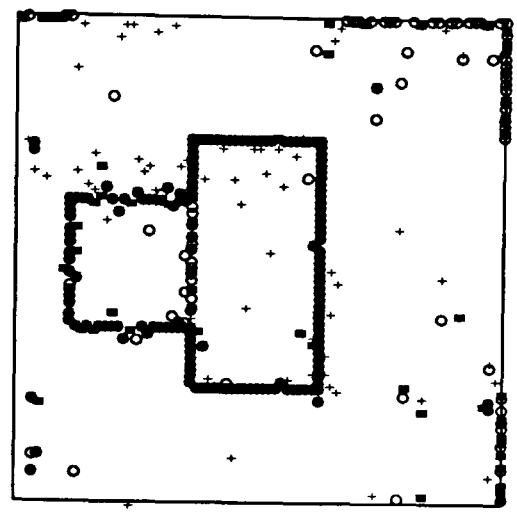


Fig 45(b) : $t_0 = 3$, $m = 1$

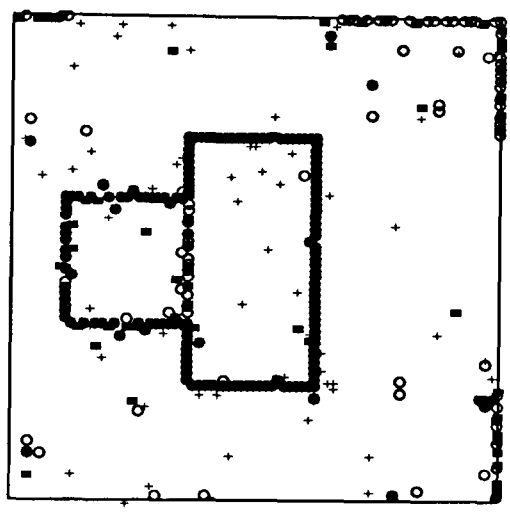


Fig 45(c) : $t_0 = 5$, $m = 1$

The results in each case seem excellent, except possibly for one incomplete edge in (a). Note that the boundary between square and rectangle, where the Signal-Noise ratio is 1.41, is detected less readily than any of the other edges, and that Signal-Noise ratio does in general effect the efficiency of the algorithm. This is entirely in line with our previous experience with the edge-detection problem. The timings involved in the production of the results were 12.50, 14.78 and 21.38 seconds for (a), (b), and (c) respectively. An exact three changepoint analysis of this true scene was not feasible due to the enormous amount of processing time required.

It is interesting at this stage to note the effect that our use of prior knowledge concerning edge continuity has on the convergence of the algorithm. Figure 46 depicts the results of an analysis identical to that above, but with the initial values for the Gibbs Sampler chosen independently of results from other rows/columns. These results are inferior in the sense that there appear to be more mis-classifications of edge-points, and one edge (that where the Signal-Noise ratio is at its lowest) remains virtually completely undetected - this was evident for larger values of (t_0, m) also. Also, the respective timings for (a), (b) and (c) in this case

were 18.38, 21.24, and 21.82, and so the rate of convergence is seemingly appreciably slower. Thus the importance of the use of spatial prior knowledge is clearly demonstrated, even when that prior knowledge is purely qualitative (i.e. we know merely that "edges are continuous", but need not quantify this statement in any way).

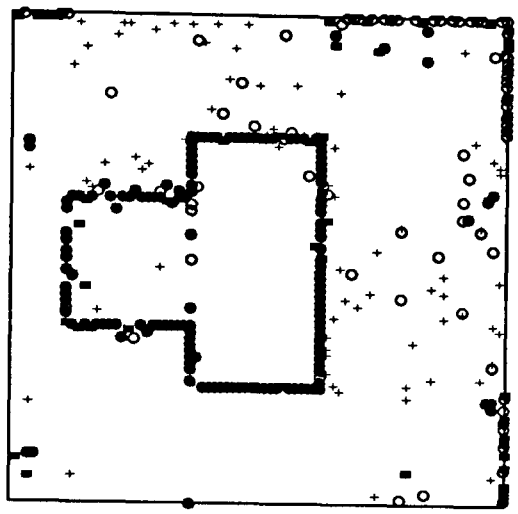


Fig 46(a) : $t_0 = 1$, $m = 1$

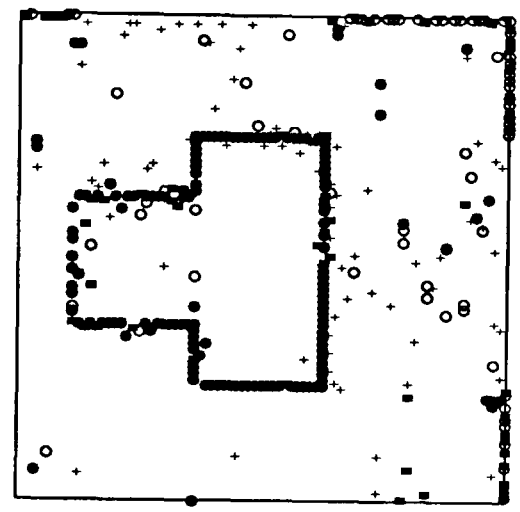


Fig 46(b) : $t_0 = 3$, $m = 1$

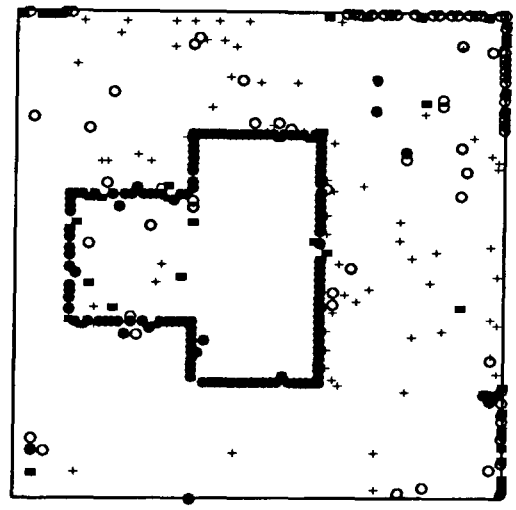


Fig 46(c) : $t_0 = 5$, $m = 1$

We now complicate the true scene further by effectively increasing the number of changepoints to be found in any particular row or column. Figure 47(a) and (b) represent an artificial true scene and image comprising two overlapping circles of the same texture θ_2 on a background texture θ_1, where the "intersection" of the circles has a different texture θ_3 created as some function of the θ_2's, taken here to be additive. We assume the same form for image-formation and noise processes, and choose mean levels 0.0 and 2.0 for θ_1 and θ_2, and hence by the additive assumption we induce θ_3 to be 4.0. Note that now the Signal-Noise ratio is 2.83 at each boundary (we choose these relatively large values in order to demonstrate the potential of the Gibbs Sampler technique and to allow comparison with results obtained using exact methods. The problem becomes difficult at lower levels of S.N.R. when using any

Again, the results seem generally good. However, several features are worthy of comment. In figure 48(a), the outline of the two circles has been detected, but the boundaries of the intersection have not - this is as we would have predicted, since $k = 2$ represents a mis-specification for several of the rows/columns. It would have been possible to inspect the margins obtained more closely for evidence of other changepoints, but this would have been time-consuming (and barely justifiable in the decision-theoretic sense). The processing time involved in the production of the results in figure 48(a) was 10.84 seconds (an exact two changepoint analysis of this image produced practically identical results in 96.82 seconds). Secondly, in figure 48(b), practically all of each of the boundaries has been detected. This is surprising as $k = 3$ represents a mis-specification for all of the rows/columns of the image. There is a degree of mis-classification of edge-points, but many of these correspond to lower modal probabilities than those recorded actually on the edges themselves. The processing time involved was 17.16 seconds. Finally, in figure 48(c), each of the boundaries has been fully detected, but there appears to be a larger number of mis-classifications. This is due to our recording four points for each row/column where the majority of rows/columns contain no or two edges. Again, the points actually on the boundaries seem to be recorded with higher probabilities The processing time involved here was 22.74 seconds. So, overall in this case, we might prefer the results in (b) to those in (a) and (c).

Finally, we investigate the performance of the Gibbs Sampler on a complex composite true scene. Figures 49(a) and (b) represent an artificial true scene and image comprising four objects - rectangle, square, ellipse, circle - of various textures and their intersections. For this example, the rectangle, circle and ellipse were chosen to have the same texture θ_2, the square to have texture θ_3, square/rectangle intersection to have texture θ_4 created additively from θ_2 and θ_3, with the background being texture θ_1.

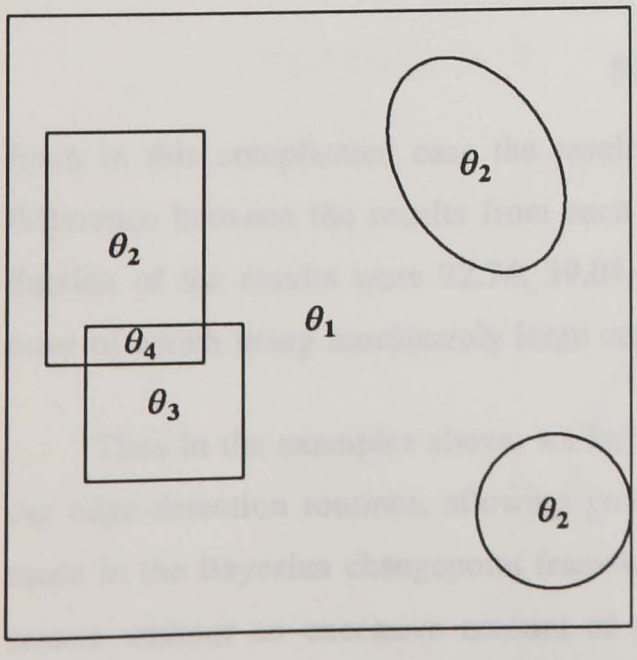


Fig 49(a) : true scene

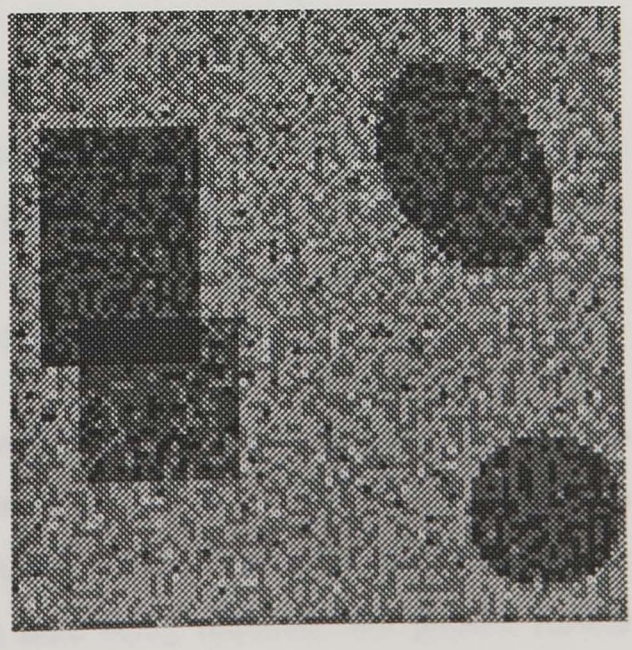


Fig 49(b) : image

We assume the same form for image-formation and noise processes, and choose mean levels 0.0, 2.0, and 1.5 for θ_1, θ_2, and θ_3, inducing θ_4 to be 3.5, and thus there are a range of Signal-Noise ratios in the image. We carry out an identical Gibbs Sampler analysis to that above, with k fixed and equal to 4, and m equal to 1. Figure 50 depicts the results obtained for a range of values of t_0.

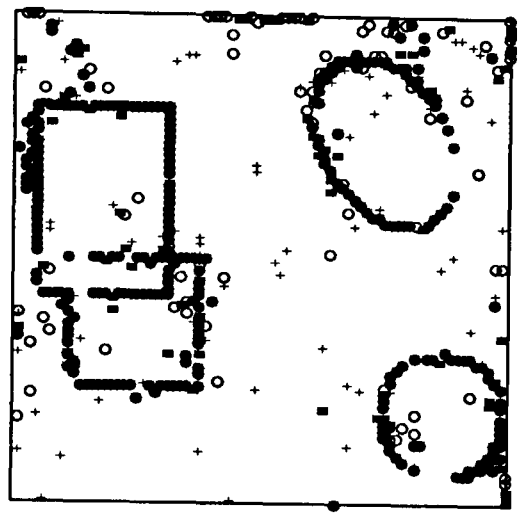


Fig 50(a) : t_0 = 1

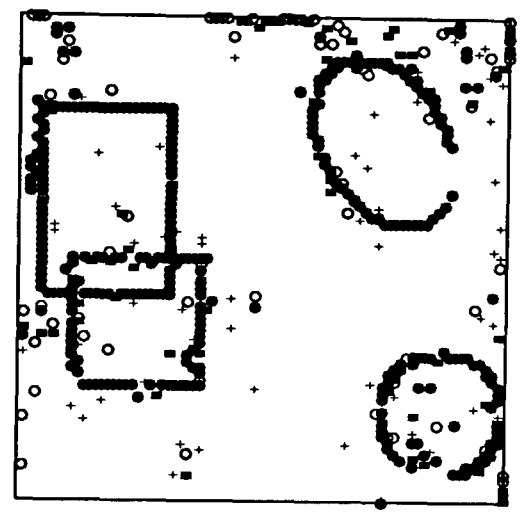


Fig 50(b) : t_0 = 10

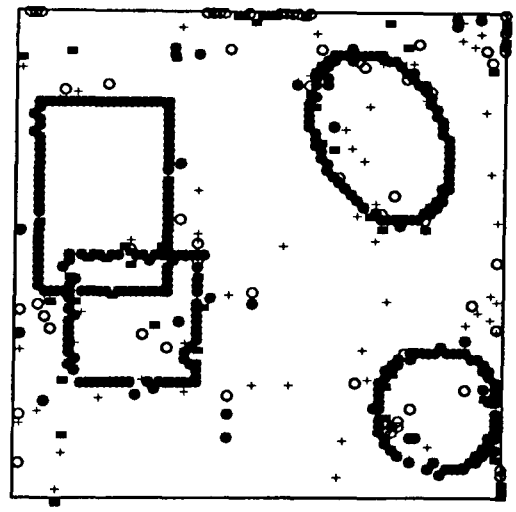


Fig 50(c) : t_0 = 20

Even in this complicated case the results are encouraging. Note that there is visually little difference between the results from each analysis. The processing times involved in the production of the results were 22.74, 39.04 and 69.66 seconds for (a), (b) and (c) respectively, none of which being inordinately large considering the relative complexity of the true scene.

Thus in the examples above, we have seen how the Gibbs Sampler algorithm is useful to our edge-detection routines, allowing good approximate inferences about edge positions to be made in the Bayesian changepoint framework for images derived from relatively complex true scenes without an excessive amount of processing time being required. We present further applications of the Gibbs Sampler algorithm later in this thesis.

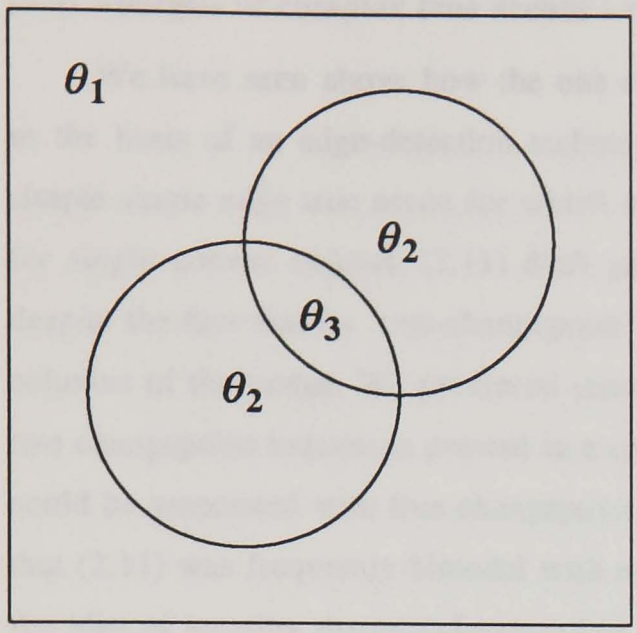


Fig 47(a) : true scene

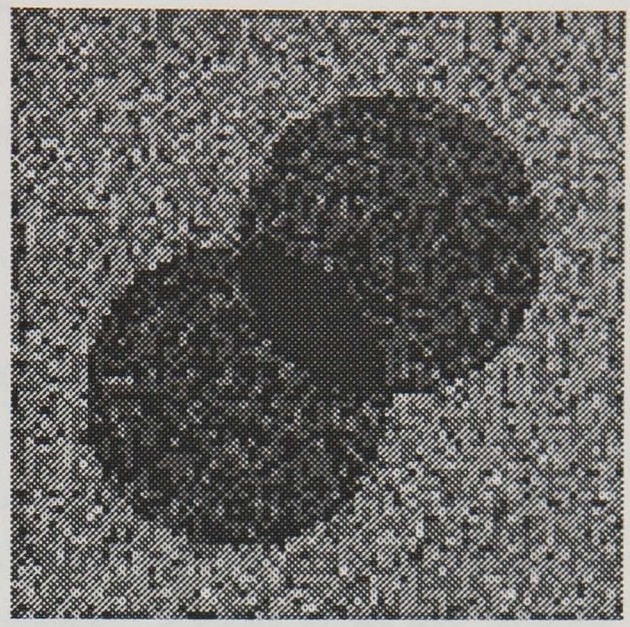


Fig 47(b) : image

technique). We proceed to carry out a Gibbs Sampler analysis under exactly the same prior assumptions and using the same implementation as above, with t_0 and m both equal to 1. We study the behaviour for different choices of k. The results are depicted in figure 48.

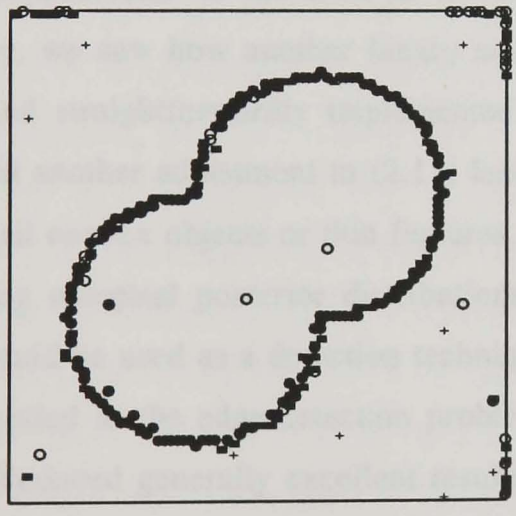


Fig 48(a) : $k = 2$

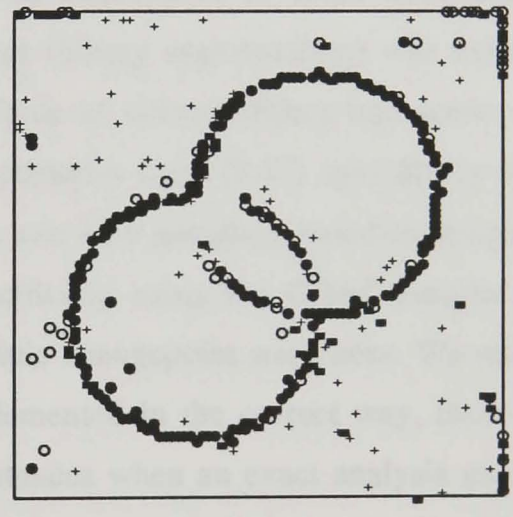


Fig 48(b) : $k = 3$

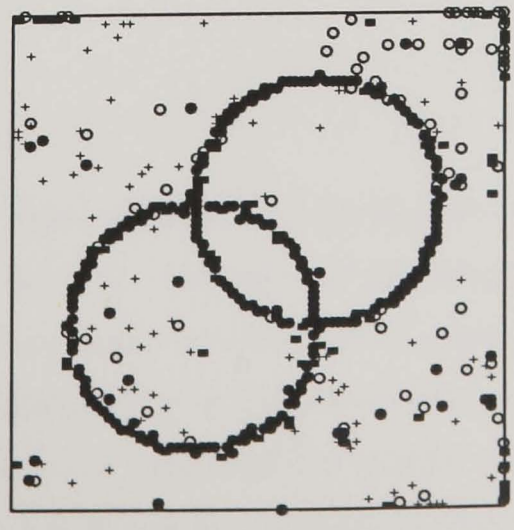


Fig 48(c) : $k = 4$

(3.8) Analysis of complex true scenes - conclusions.

We have seen above how the one changepoint posterior distribution (2.11) can be used as the basis of an edge-detection technique for more complex true scenes as well as for the simple single edge true scene for which it was originally designed. We saw, for example, that for single convex objects, (2.11) dealt adequately with a surprisingly large number of cases, despite the fact that its "one-changepoint" presumption was incorrect for each of the rows and columns of the image. We presented some theoretical justification for the use of (2.11) for the two changepoint sequences present in a convex object true scene, and saw that modes in (2.11) could be associated with true changepoint positions in a two changepoint sequence, and indeed that (2.11) was frequently bimodal with modes at both true changepoint positions. We rejected the idea of locating the two changepoints using "approximate" inference in this way, however, on both practical and theoretical grounds. We then saw that exact inference was possible using the natural two changepoint extension to (2.11), namely (3.2), but that the resulting edge-detection technique involved an unacceptable amount of computation. We therefore discussed adjustments to a "full" (row/column) analysis using (2.11) suitable for the analysis of convex object true scenes. First, we saw how the analysis of projections in the image data other than those perpendicular to the axes of the true scene was informative in some circumstances. Secondly, we saw how another binary search technique (binary segmentation) was extremely useful and straightforwardly implemented in the analysis of convex object true scenes. We discussed another adjustment to (2.11) leading to the posterior form (3.17) specifically to deal with small convex objects or thin features. Finally, we saw how sampling based techniques for evaluating marginal posterior distributions, more specifically using the Gibbs Sampler algorithm, could be used as a detection technique for multiple changepoint sequences. We saw that when applied to the edge-detection problem and implemented in the correct way, these techniques produced generally excellent results in circumstances when an exact analysis using the multiple changepoint extension to (2.11) would not have been feasible due to the immense amount of computation required.

Chapter 4 : Spatial Dependence and Edge Continuity.

As we have seen in chapters 2 and 3, ideas from changepoint analysis can be used as the basis of a decision-based technique for the solution of the important problem of edge-detection. We have seen that we may identify edges in the image with modes in changepoint posterior probability distributions. However, we have typically carried out the above analyses processing each row/column in isolation, and treated the results obtained independently (the exceptions to this being our use of "smoothing" in the production of figures 8(b) and (c), and the development of the implementation of the Gibbs Sampler-based detection methods). Now, while this may be justified on the grounds of simplicity and expediency, it clearly ignores one crucial aspect of our prior knowledge of the nature of the true scene, namely the presence of localised spatial dependence. The use of local dependence priors (Gibbs/M.R.F) for true scenes in statistical image processing is well documented in recent years, and is widely regarded as being of fundamental importance. For the edge-detection problem, the concepts of spatial dependence in and continuity of the true scene is important in two ways. First, we would expect pixels in non-edge regions to exhibit spatial dependence in the usual way, and consequently non-edge portions in any one particular row should correspond spatially to non-edge portions in adjacent rows. Secondly, we would expect edges in the true scene to be spatially continuous, and thus edge-points in any row should correspond to similarly positioned edge-points in adjacent rows. This suggests that the changepoint based techniques developed above may be amended to incorporate local dependence ideas in two ways, either (1) via the prior form for texture parameters (assuming some joint prior distribution for the pixel elements of adjacent rows/columns in the true scene), or (2) via the form of the prior distribution for changepoint position in adjacent rows/columns (or indeed via a combination of (1) and (2)). However, we must ultimately balance the advantages that may be gained from these considerations against the additional processing demand needed for their implementation.

We now proceed and attempt to adapt the changepoint detection techniques using the ideas discussed above.

(4.1) Localised pixel dependence.

We attempt to incorporate local dependencies into the form of our prior distribution for the discretised pixel version of the true scene for use in changepoint based edge-detection techniques. In keeping with the notation of previous chapters $\theta = \{\theta_{ij}, i,j = 1,\ldots,n\}$ represent the random variable corresponding to the discretised $n \times n$ pixel version of the true scene S_θ, and $Y = \{Y_{ij}, i,j = 1,\ldots,n\}$ represent the random variable corresponding to the observed image derived from θ. Let θ_i and Y_i represent a single row/column taken from the

true scene and image respectively. Now recall the definition of the Markov Random Field joint distribution in section (1.4.1.1) of chapter 1 related here specifically to θ. In addition to a positivity condition, the M.R.F. is characterised by the following relation - for each (i,j), the conditional distribution of θ_{ij} given $\theta_{(ij)} = \{\theta_{kl} ; k,l = 1,\ldots,n , k \neq i , l \neq j\}$ can be written

$$[\theta_{ij} \mid \theta_{(ij)}] \equiv [\theta_{ij} \mid \theta_{\partial ij}] \tag{4.1}$$

where $\theta_{\partial ij}$ represents the pixels in some suitably defined neighbourhood of (i,j), usually taken to be a subset of pixels in S_θ in the vicinity of (i,j).

In the edge-detection problem, our interest lies in changepoint posterior distributions given the data in rows/columns of the image. To allow the evaluation of these distributions, we must specify prior distributions for the true scene parameters, and previously we have chosen particularly simple forms. We now concentrate on more complex choices using the conditioning property (4.1). We consider two alternatives.

(4.1.1) Introduction of pixel dependence: method 1.

Consider the joint posterior changepoint distribution of the single changepoints $(r_{j-1}, r_j, r_{j+1}) = r_J$ in each of the three adjacent rows $j-1, j, j+1$ given the data in those rows $(Y_{j-1}, Y_j, Y_{j+1}) = Y_J$. If we denote the true scene pixel values in rows $j-1, j, j+1$, $(\theta_{j-1}, \theta_j, \theta_{j+1})$ by θ_J then it is clear that

$$\begin{aligned} [r_J \mid Y_J, \psi] &\propto [Y_J \mid r_J, \psi]\,[r_J] \\ &= \int [Y_J \mid r_J, \theta_J]\,[\theta_J \mid r_J, \psi]\,[r_J] \end{aligned} \tag{4.2}$$

It remains to specify forms for each term in (4.2). The first term is merely the likelihood, and under the same conditional independence assumptions as in the previous section can be written

$$[Y_J \mid r_J, \theta_J] = \prod_{(i,j) \in S_J} [Y_{ij} \mid \theta_{ij},]$$

where S_J denotes the pixels of S in rows $j-1, j, j+1$. The third term in (4.2) is the joint prior on changepoint position in the three rows, and can be chosen to be uniform, or chosen so as to reflect the spatial continuity of the edge - this is discussed in section (4.2).

Thus it remains to specify a form for $[\theta_J \mid r_J, \psi]$, the joint prior for true scene values in the three rows. In this section we choose this prior to be of the Gibbs/M.R.F. form (4.1), and may select, for example, from any of the forms presented in Besag (1974,1986). We now provide an illustrative example. Suppose that the image formation process corrupted each pixel in the true scene independently and identically with Gaussian white-noise as before, so that

$$[Y_{ij} \mid \theta_{ij}] = N(\theta_{ij}, \sigma^2) \tag{4.3}$$

and suppose that σ^2 is known. Then a suitable choice for the distribution of θ_J might be based on the conjugate auto-normal version of (4.1), i.e.

$$\begin{aligned}[\theta_{ij} \mid \theta_{\partial ij}] &\equiv N\Big(\mu_{ij} + \sum_{(k,l) \in \partial ij} \beta_{kl}^{ij}\theta_{kl} \, , \lambda_{ij}\Big) \\ &\propto \exp\left\{ \frac{-1}{2\lambda_{ij}}\Big(\theta_{ij} - \mu_{ij} - \sum_{(k,l) \in \partial ij} \beta_{kl}^{ij}\theta_{kl} \Big)^2 \right\} \end{aligned} \tag{4.4}$$

We thus introduce a further stage into the hierarchy of the modelling of the image data (we choose the prior mean, variance and interaction parameters $\mu_{ij}, \lambda_{ij}, \beta_{kl}^{ij}$ to be the same across each texture but different between textures). It is now clear that the prior for θ_J takes the form of a multivariate Normal distribution

$$\begin{aligned}[\theta_J \mid \psi] &\equiv N(\mu_J, Q_J^{-1}) \\ &\propto \exp\left\{ -\frac{1}{2}(\theta_J - \mu_J)^T Q_J (\theta_J - \mu_J) \right\} \end{aligned} \tag{4.5}$$

where μ_J is the $3n \times 1$ vector of prior means, and Q_J is the $3n \times 3n$ matrix with diagonal entries $\frac{1}{\lambda_{ij}}$ and off-diagonal entries $-\frac{\beta_{kl}^{ij}}{\lambda_{ij}}$ - we choose the parameters to ensure that Q_J is symmetric and positive-definite. We may naturally extend this prior to $[Y_J \mid r_J, \theta_J]$, reflecting the changepoint positions in the three rows by careful choice of μ_{r_j} and by ensuring that β_{kl}^{ij} is non-zero if and only if pixels (i,j) and (k,l) are neighbours and are not separated by changepoints forming an edge in the three rows. We denote the resulting Normal prior conditional on r_J by

$$[\theta_J \mid r_J, \psi] \equiv N(\mu_{r_j}, Q_{r_j}^{-1}) \quad . \tag{4.6}$$

Now, from (4.3) it is clear that

$$[Y_J \mid \theta_J] \equiv N(\theta_J, \sigma^2 I)$$

and thus by (4.6) and a standard result of Lindley and Smith (1972)

$$\begin{aligned}[Y_J \mid r_J, \psi] &= \int [Y_J \mid r_J, \theta_J] [\theta_J \mid r_J, \psi] \\ &\equiv N(\mu_{r_j}, \Sigma_{r_j}) \end{aligned} \tag{4.7}$$

where $\Sigma_{r_j} = \sigma^2 I + Q_{r_j}^{-1}$.

Thus to compute the probabilities in (4.2) we must evaluate the multivariate Normal probability in (4.7), i.e. the quantity

$$(2\pi)^{\frac{-3n}{2}} |\Sigma_{r_j}|^{\frac{-3n}{2}} \exp\left\{ -\frac{1}{2}(Y_J - \mu_{r_j})^T \Sigma_{r_j}^{-1} (Y_J - \mu_{r_j}) \right\}. \tag{4.8}$$

We now investigate the nature of Σ_{r_j} and its inverse. It is clear that

$$\begin{aligned} \Sigma_{r_j} &= \sigma^2 I + Q_{r_j}^{-1} \\ &= Q_{r_j}^{-1}(\sigma^2 Q_{r_j} + I) \end{aligned}$$

and hence

$$\Sigma_{r_j}^{-1} = (\sigma^2 Q_{r_j} + I)^{-1} Q_{r_j} \tag{4.9}$$

so that interest turns to inverting $\sigma^2 Q_{r_j} + I$. First, we examine the precise form of Q_{r_j} in the changepoint context. Assuming that $\lambda_{ij} = \lambda_1$ and $\beta_{kl}^{ij} = \beta_1$ if pixels (i,j) and (k,l) are in texture region 1, and $\lambda_{ij} = \lambda_2$ and $\beta_{kl}^{ij} = \beta_2$ if pixels (i,j) and (k,l) are in texture region 2, and under a second order nearest neighbour system (where the eight neighbours of pixel (i,j) are $\{(k,l) : k = i, i \pm 1, l = j, j \pm 1, (k,l) \neq (i,j)\}$), Q_{r_j} can be written

$$Q_{r_j} = \begin{bmatrix} A_{r_{j-1}} & B_{r_{j-1}r_j} & 0 \\ B^T_{r_{j-1}r_j} & A_{r_j} & B_{r_j r_{j+1}} \\ 0 & B^T_{r_j r_{j+1}} & A_{r_{j+1}} \end{bmatrix} \tag{4.10}$$

where the first row and column block refer to θ_{j-1}, the second to θ_j and the third to θ_{j+1}, the off-diagonal blocks representing the interactions between adjacent rows - note that rows $j-1$ and $j+1$ are *a priori* independent. A_r is the $n \times n$ matrix given by

$$A_r = \begin{bmatrix} C_r^1 & 0 \\ 0 & C_{n-r}^2 \end{bmatrix} \tag{4.11}$$

and C_k^i is the $k \times k$ tri-diagonal matrix matrix with diagonal elements $\frac{1}{\lambda_i}$ and off-diagonal elements $-\frac{\beta_i}{\lambda_i}$ for $i = 1, 2$. Now consider $B_{r_{j-1}r_j}$. We make the restriction that changepoint positions in adjacent rows can only differ by one pixel to reflect the spatial continuity of the edge

over the three rows. Thus we consider only three cases.

First, if $r_{j-1} = r_j - 1$, then $B_{r_{j-1}r_j}$ can be written

$$B_{r_{j-1}r_j} = \begin{bmatrix} D^1_{r_j-1} & 0 & 0 \\ u^{1T}_{r_j-1} & 0 & 0 \\ 0 & v^2_{n-r_j} & D^2_{n-r_j} \end{bmatrix} \tag{4.12}$$

where D^i_k is the $k \times k$ tri-diagonal matrix with all non-zero elements equal to $-\frac{\beta_i}{\lambda_i}$ for $i = 1, 2$, u^i_k is the $k \times 1$ vector with only the k'th element non-zero and equal to $-\frac{\beta_i}{\lambda_i}$, and v^i_k is the $k \times 1$ vector with only the first element non-zero and equal to $-\frac{\beta_i}{\lambda_i}$ for $i = 1, 2$, and 0 is the zero matrix of appropriate dimension.

Secondly, if $r_{j-1} = r_j$ then $B_{r_{j-1}r_j}$ can be written

$$B_{r_{j-1}r_j} = \begin{bmatrix} D^1_{r_j} & 0 \\ 0 & D^2_{n-r_j} \end{bmatrix} \tag{4.13}$$

Finally, if $r_{j-1} = r_j + 1$, then $B_{r_{j-1}r_j}$ can be written

$$B_{r_{j-1}r_j} = \begin{bmatrix} \mathrm{D}^1_{r_j} & u^1_{r_j} & 0 \\ 0 & 0 & v^{2T}_{n-r_j-1} \\ 0 & 0 & D^2_{n-r_j-1} \end{bmatrix} \tag{4.14}$$

Thus, using (4.10) - (4.14), we can write down Q_{r_j} and hence $\sigma^2 Q_{r_j} + I$ explicitly - the latter being identical in form to Q_{r_j} but with the diagonal elements of C^i_k replaced by $\frac{\sigma^2}{\lambda_i} + 1$, and all other non-zero elements of Q_{r_j} replaced by $-\frac{\sigma^2\beta_i}{\lambda_i}$, for $i = 1, 2$.

It remains to invert $\sigma^2 Q_{r_j} + I$ before computation of the quantity in (4.8). This is a solvable problem, as we may evaluate the determinant using a recurrence relation, and hence immediately write down by inspection the inverse for this tri-diagonal matrix, or alternately use an iterative method for successive values of r over the complete range. However, it is clear that the amount of computation involved in the remaining calculation is considerable in either case (the evaluation of the determinant factor and the sum of squares in (4.8) would be time consuming for large n even if the form of Σ_{r_j} were relatively straightforward, which it

certainly is not in this situation).

We may simplify the problem of inverting Σ_{r_J} by making certain assumptions. We could revert to the simpler situation in which Q_{r_J} is taken to be independent of r_J, and the changepoint structure is only reflected in choice of prior mean vector. Then Q_{r_J} would take the form

$$Q_{r_J} = \begin{bmatrix} A_n & B_n & 0 \\ B^T{}_n & A_n & B_n \\ 0 & B^T{}_n & A_n \end{bmatrix} \tag{4.15}$$

where A_n is the $n \times n$ tri-diagonal matrix with diagonal elements $\frac{1}{\lambda}$ and non-zero off-diagonal elements $-\frac{\beta}{\lambda}$, and B_n is the $n \times n$ tri-diagonal matrix with all non zero elements equal to $-\frac{\beta}{\lambda}$. Under this assumption, evaluation of (4.8) is more straightforward as the $\Sigma_{r_J}^{-1}$ term is constant for all choices of r_J and thus need be evaluated only once in any implementation. Inversion of Σ_{r_J} here is possible, but again complex - the presence of the $\sigma^2 I$ term makes the inversion procedure non-trivial. This suggests an alternative simplification - set σ^2 equal to zero originally so as to make $\Sigma_{r_J} = Q_{r_J}{}^{-1}$ and thus the evaluation of (4.8) straightforward, as clearly then $\Sigma_{r_J}^{-1} = Q_{r_J}$ which we have straightforwardly specified above. Setting σ^2 equal to zero is merely a reparameterisation of the problem, which basically introduces local dependence into the first (data) stage of the modelling hierarchy, i.e. the distribution of Y_J given the true scene parameters replaces conditional independence with a covariance structure given by $Q_{r_J}{}^{-1}$, i.e.

$$[\, Y_J \mid r_J, \mu, \psi_Y \,] = N(A_{r_J}\mu, Q_{r_J}{}^{-1})$$

where $\mu = (\mu_1, \mu_2)$ refers to the mean level parameters, and $\psi_Y = (\lambda_1, \lambda_2, \beta_1, \beta_2)$ are the known parameters in the dispersion matrix Q_{r_J}. However, for uniformity of notation and preservation of the hierarchical structure, we continue to regard the covariance structure as being introduced at the first prior stage (note that this is essentially equivalent to the assumptions made in the previous section when we considered the rows/columns independently, and regarded the textures as being homogeneous). It is clear from (4.1) precisely how the introduction of the dependence structure acts much in the way of classical noise-reduction methods in non-statistical image-processing, by use of local averaging.

Hence we may use the above simplifications to aid in the evaluation of the posterior probabilities. However, we consider techniques based upon, for instance, (4.2) and (4.8), as being practically inappropriate due to the large amount of computation required. For example,

informal calculations suggest that the evaluation of the sum of squares in (4.8) takes the order of 20 times longer than standard analysis using the usual one changepoint posterior distribution (2.11). Also, difficulties may arise over the specification of the prior parameters in (4.4). It is widely known that choice of these hyperparameters is critical in standard M.R.F. based statistical image reconstruction. Finally, although the extension of the formulation to multiple changepoint problems is possible, its implementation is certainly not straightforward, and it is not clear how the one changepoint spatial posterior distribution would behave in a two changepoint context.

In light of the above discussion, we consider a second technique based upon a different changepoint posterior distribution in an effort to remove some of the difficulties mentioned.

(4.1.2) Introduction of pixel dependence: method 2.

Consider the conditional posterior distribution on changepoint position in row j given the data Y_j in that row and the true scene pixel values in rows $j-1$ and $j-1$, $[r \mid Y_j, \theta_{j-1}, \theta_{j+1}, \psi]$. Then it is clear that

$$\begin{aligned}[r \mid Y_j, \theta_{j-1}, \theta_{j+1}, \psi] &\propto [Y_j \mid r, \theta_{j-1}, \theta_{j+1}, \psi]\,[r \mid \theta_{j-1}, \theta_{j+1}] \\ &= \int [Y_j \mid r, \theta_{j-1}, \theta_j, \theta_{j+1}]\,[\theta_j \mid r, \theta_{j-1}, \theta_{j+1}, \psi] \\ &\quad .\,[r \mid \theta_{j-1}, \theta_{j+1}] \qquad (4.16)\end{aligned}$$

Under the usual conditional independence assumptions we have a further simplification

$$\begin{aligned}[Y_j \mid r, \theta_{j-1}, \theta_j, \theta_{j+1}] &\equiv [Y_j \mid r, \theta_j] \\ &= \prod_{i=1}^{n} [Y_{ij} \mid \theta_{ij}]\end{aligned}$$

It is clear that now the dimensionality of the problem has been reduced from $3n$ to n and thus we might expect the amount of processing time required to be reduced by a factor of three compared to the above technique.

After specifying forms for $[Y_{ij} \mid \theta_{ij}]$, $[\theta_j \mid r, \theta_{j-1}, \theta_{j+1}, \psi]$ and $[r \mid \theta_{j-1}, \theta_{j+1}]$, we may evaluate the functional form of (4.16). We might then proceed by substituting estimates of θ_{j-1} and θ_{j+1} based on Y_{j-1} and Y_{j+1} into this functional form to enable us to compute the posterior probabilities. We consider this solution with respect to the illustrative example described above.

Suppose, as before, that the Y_{ij} are independently Normally distributed conditional on the θ_{ij}, specifically $[Y_{ij} \mid \theta_{ij}] \equiv N(\theta_{ij}, \sigma^2)$, or $[Y_j \mid r, \theta_j] \equiv N(\theta_j, \sigma^2 I_n)$. Now, we must specify some prior distribution for θ_j conditional on $\theta_{j-1}, \theta_{j+1}$ and r. We choose this prior to

be of the form of (4.4), i.e.

$$[\theta_{ij} \mid r, \theta_{j-1}, \theta_{j+1}] \propto \exp\left\{ -\frac{1}{2\lambda_{ij}} (\theta_{ij} - \mu_{ij} - \beta^{ij}_{i-1j}\theta_{i-1j} - \beta^{ij}_{i+1j}\theta_{i+1j} - T_{\partial ij})^2 \right\} \quad (4.17)$$

where $T_{\partial ij} = \sum_{(k,l)} \beta^{ij}_{kl}\theta_{kl}$ and the summation runs over the neighbours of (i,j), but with $l \neq j$, the mean parameters μ_{ij} chosen to reflect the nature of the changepoint sequence, and the variance and interaction parameters $\lambda_{ij}, \beta^{ij}_{kl}$ are chosen to be equal for simplicity. This prior then takes the form of a multivariate Normal distribution of dimension n, i.e.

$$\begin{aligned} [\theta_j \mid r, \theta_{j-1}, \theta_{j+1}] &\equiv N(A_r\mu + T, Q_r^{-1}) \\ &\propto |Q_r|^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}(Y_j - A_r\mu - T_j)^T Q_r (Y_j - A_r\mu - T_j) \right\} \end{aligned} \quad (4.18)$$

where $\mu = [\mu_1, \mu_2]^T$ are the mean levels of the two textures, A_r is the $n \times 2$ matrix reflecting the changepoint position in row j, $T_j = [T_{\partial 1j}, \ldots, T_{\partial nj}]^T$, and Q_r is the $n \times n$ interaction matrix which takes the form of (4.11)

$$Q_r = \begin{bmatrix} C_r & 0 \\ 0 & C_{n-r} \end{bmatrix} \quad (4.19)$$

where as before C_k is the $k \times k$ tri-diagonal matrix with diagonal elements $\frac{1}{\lambda}$ and non-zero off-diagonal elements $-\frac{\beta}{\lambda}$. Again, using the result of Lindley and Smith, it is clear that

$$\begin{aligned} [Y_j \mid r, \theta_{j-1}, \theta_{j+1}] &= \int [Y_j \mid r, \theta_{j-1}, \theta_j, \theta_{j+1}][\theta_j \mid r, \theta_{j-1}, \theta_{j+1}, \psi] \\ &\equiv N(A_r\mu + T, \Sigma_r) \end{aligned} \quad (4.20)$$

where $\Sigma_r = \sigma^2 I + Q_r^{-1}$ and we may evaluate the posterior probabilities in (4.16). As before, inversion of Σ_r as in (4.9) is complex, but feasible due to the simple form of Q_r. Recall that, by (4.9),

$$\Sigma_r^{-1} = (\sigma^2 Q_r + I)^{-1} Q_r \quad (4.21)$$

but here Q_r is block tri-diagonal, and thus we may write $(\sigma^2 Q_r + I)^{-1} = P_r$, where

$$P_r = \begin{bmatrix} E_r^{-1} & 0 \\ 0 & E_{n-r}^{-1} \end{bmatrix} \tag{4.22}$$

and E_k is the $k \times k$ tri-diagonal matrix with diagonal elements $\frac{\sigma^2}{\lambda} + 1$ and non-zero off-diagonal elements $-\frac{\sigma^2\beta}{\lambda}$. Note that the extension to the case where λ and β are different for different textures at the prior stage is straightforward. Hence, from (4.21) and (4.22), we have

$$\begin{aligned} \Sigma_r^{-1} &= P_r Q_r \\ &= \begin{bmatrix} E_r^{-1} C_r & 0 \\ 0 & E_{n-r}^{-1} C_{n-r} \end{bmatrix} \end{aligned} \tag{4.23}$$

Unfortunately, this form for the inverse of Σ_r involves a large number of non-zero terms, and hence to lessen the amount of computation required we must again set σ^2 equal to zero, equivalent to the assumption of homogeneous textures exhibiting a localised dependence structure, or alternatively, dependence at the data stage.

Finally, before implementing such a scheme, we make one further simplifying assumption by specifying the prior means to both equal zero, and letting the true values in adjacent rows solely govern the texture mean levels in the changepoint analysis of row j. This removes the need for specifying values for μ_1 and μ_2, or parameters in any subsequent stage of the hierarchy. The only parameters we must specify now are the variance and interaction parameters, which is pleasing as we concentrate here on the effect of the introduction of the dependence structure.

Taking into account all of the above simplifications, we proceed with an evaluation of the conditional posterior distribution $[r \mid Y_j, \theta_{j-1}, \theta_{j+1}, \psi]$, but with Y_{j-1} and Y_{j+1} used as estimates of θ_{j-1} and θ_{j+1}, and assuming a uniform prior for r, i.e.

$$[r \mid Y_j, \hat{\theta}_{j-1}, \hat{\theta}_{j+1}, \psi] \propto |Q_r|^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}(Y_j - \hat{T}_j)^T Q_r (Y_j - \hat{T}_j) \right\} \tag{4.24}$$

where $\hat{T}_j = [\hat{T}_{\partial 1j}, \ldots, \hat{T}_{\partial nj}]^T$ and $\hat{T}_{\partial ij} = \sum_{(k,l)} \beta Y_{kl}$, the summation running over the neighbours of (i,j) but with $l \neq j$.

Thus, in (4.24), we have another possible changepoint posterior distribution of interest. However, we again encounter problems in practice. Ultimately, we do have to choose values for λ and β - to do so without previous experience or unrealistically detailed knowledge of the true scene is complex. We might try an analysis of the image concerned using a range of

values for the hyperparameters and compare results over the range, but this is unsatisfactory. In light of this and the other factors mentioned above (we have made several very restrictive assumptions), we reject the practical implementation of these techniques as a solution to the edge-detection problem. Instead, we concentrate on more naive techniques which attempt to catch the flavour of some of the more complex ideas discussed but require less in the way of prior input.

We now investigate some naive techniques for improvement of the changepoint posterior based techniques for solution of the edge-detection problem. We saw above how the formal introduction of ideas of spatial continuity corresponded roughly to noise-reduction techniques and local averaging in non-statistical image-processing. We now study the effect of such noise-reduction techniques for some of the true scenes and images studied in the previous section.

(4.1.3) Introduction of pixel dependence: naive methods.

First, consider the simple edge true scene in Figure 3 on p. 25. We saw how the standard changepoint technique based on the posterior distribution in (2.11) was ineffective at low Signal-Noise ratios. Thus we compare the results obtained using (2.11) on the original image with those obtained by pre-processing the image using local averaging over a small neighbourhood (taken here to be second order nearest neighbour) and carrying out an analysis using (2.11). It is clear that such an averaging procedure induces a correlation in the data in the pre-processed image, which we note but subsequently ignore for ease of processing. Also, local averaging and subsequent changepoint analysis of this nature is strongly related to (4.17) above, but with the precise dependence structure somewhat altered. The results of the two row only analyses are depicted in figure 51.

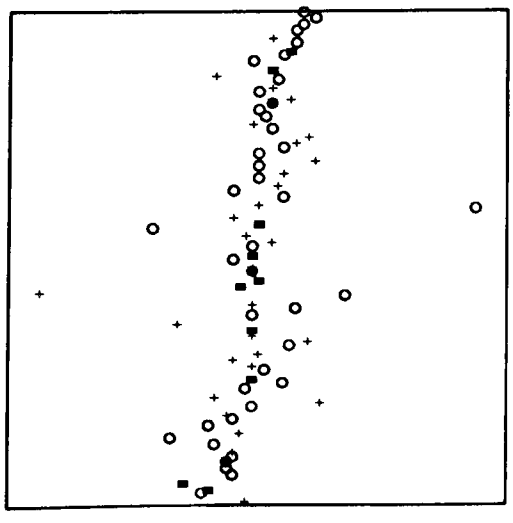


Fig 51(a) : standard analysis

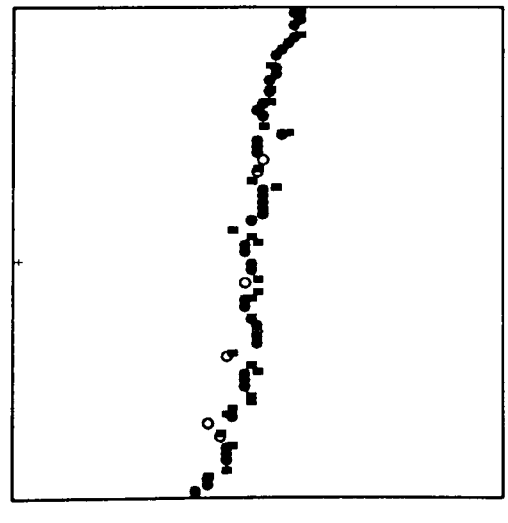


Fig 51(b) : analysis of pre-processed data

Recall that, in this example, the Signal-Noise ratio was fixed as 1.0. It is clear that such

simple pre-processing is of considerable use and importance. In this context, local averaging effectively increases Signal-Noise ratio at the boundary, and also lessens the impact of any outlying or extreme noise-corrupted values in the image data. We would expect also that it would "blur" the edge, but as can be seen from figure 51, this is not necessarily the case. The processing times involved in producing the results in figures 51(a) and (b) were 0.92 and 3.64 seconds respectively (with pre-processing time included in the latter case) for the row analyses.

We now carry out a full analyses on the same true scene as that in figure 10(a) on p. 44, namely the circle, again with the Signal-Noise ratio fixed and equal to 1.0. The results are depicted in figure 52.

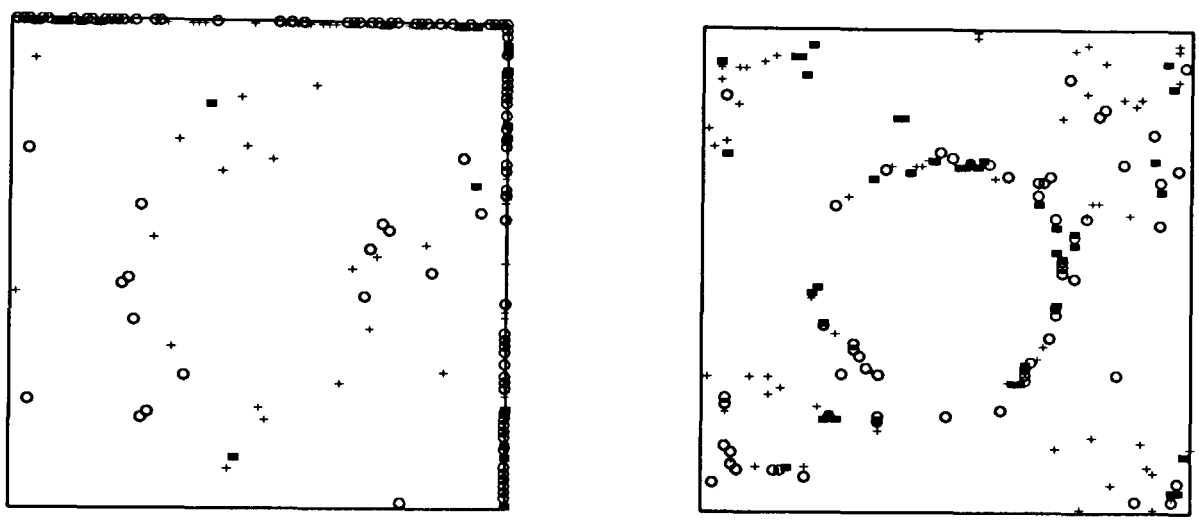


Fig 52(a) : standard analysis Fig 52(b) : analysis of pre-processed data

Again, the pre-processing of the image data gives rise to more satisfying results - we can more readily discern the nature of the true scene in (b) than in (a). The processing times involved here were 1.82 and 5.78 seconds for (a) and (b) respectively.

Thus we have seen how very naive ideas concerning spatial continuity can actually improve the performance of the changepoint based techniques introduced previously. We now suggest other similar simple ideas. For example, prior to evaluating the changepoint posterior distribution conditional on the data in the row/column concerned, we must specify a form for the distribution of the unknown parameters of the textures, $[\theta \mid \psi]$. We saw above how conditional priors dependent on true scene pixel values in adjacent rows of the form $[\theta_j \mid r, \theta_{j-1}, \theta_{j+1}]$ could also be used. A natural extension would be to use a prior for θ_j which was dependent in some way on the observed data rather than the true scene values in adjacent rows, i.e. we might choose the hyperparameters for the prior for θ_j as being (deterministically) related to the data in adjacent rows. Consider, for example, a Normal changepoint sequence we might select an informative conjugate Normal/Chi-squared prior for the unknown parameters, and choose hyperparameters on the basis of results in adjacent rows.

Also, it is evident from our original formulation of the changepoint problem that we may derive posterior distributions for the true scene parameters in each row, and that these posterior distributions in the Normal case take the form of mixtures of Normal/Chi-squared distributions, which may then be used to derive prior distributions for for adjacent rows.

It should be noted that these latter ideas are fundamentally different to our original ideas about the use of local dependence in priors for the true scene in the edge-detection problem, in the sense that they merely use (*a posteriori*) inferences from single rows to aid in the analysis of others, rather than utilising any underlying prior structure. Thus, strictly, we might expect the latter ideas to be of little assistance for true scenes corrupted by high levels of additive noise. This point is demonstrated by figure 53. The results depicted in figure 53(a) and (b) were obtained by row analysis of the single edge true scene in our previous example, where the Signal-Noise ratio at the boundary in the image is 1.0. Figure (a) depicts the results obtained when changepoint posterior (2.11) is used, where all of the prior distributions specified are non-informative, whereas (b) depicts the results obtained when using a posterior distribution derived by assuming a degenerate prior distribution i.e. where all of the texture parameters are known and correctly specified.

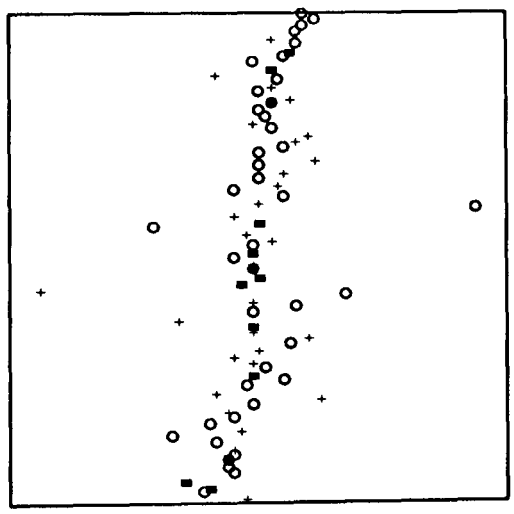


Fig 53(a) : Non-informative priors

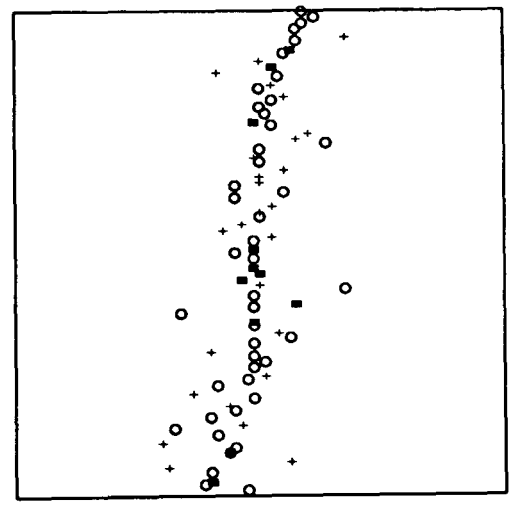


Fig 53(b) : All parameters known

The results are very similar in terms of accuracy in detection of the edge (presumably due to the large amount of data available in each row). Thus as we indicated above, use of knowledge concerning spatial dependencies in the fashion mentioned above without use of a true scene dependence structure is of little use in this example (we obtain similar results when completely ignorant of the true scene parameters to those obtained when we know them precisely) compared with the improvement obtained when using dependence-based local averaging ideas as in figure 53(b). Thus it seems that introducing some form of local averaging procedure into the changepoint-based edge-detection analysis is the most profitable way of incorporating aspects of localised pixel dependence.

(4.2) Edge continuity.

We now attempt to incorporate some notion of edge continuity into our changepoint formulation of the edge-detection problem. Above, we saw how to amend our choice of prior for true scene parameters in each row by introducing a joint structure for the true scene parameters in adjacent rows using Gibbs/M.R.F. type ideas. Here we try to amend our choice of prior for changepoint position.

(4.2.1) Two row joint prior specification.

Consider the discrete, univariate prior distribution on changepoint position in row j, $[r_j]$ for use in the evaluation of the changepoint posterior distribution. Previously, we have taken this prior to be uniform (a discrete, non-informative prior). We now study other possible choices. For example, it is clear that, using the notation introduced above,

$$\begin{aligned} [r_j] &= \sum_{r_{j-1}} [r_j, r_{j-1}] \\ &= \sum_{r_{j-1}} [r_j \mid r_{j-1}]\,[r_{j-1}] \end{aligned} \tag{4.25}$$

and hence by choice of $[r_j, r_{j-1}]$ and marginalisation we may be able to introduce some idea of edge continuity - unfortunately, (4.25) is of little practical use in this form, as if we are *a priori* ignorant of changepoint position in any row, then all quantities in (4.25) will ultimately be uniform. However, it does lead to a more feasible proposal. Recall that in our full row analyses we would begin at row 1, proceed to row 2 etc. until row n had been analysed, treating each row and its corresponding changepoint posterior distribution independently. Now consider the changepoint prior for row j conditional on the data in row $j-1$, Y_{j-1}, denoted $[r_j \mid Y_{j-1}]$. Then under the usual conditional independence assumptions

$$\begin{aligned} [r_j \mid Y_{j-1}] &= \sum_{r_{j-1}} [r_j, r_{j-1} \mid Y_{j-1}] \\ &= \sum_{r_{j-1}} [r_j \mid r_{j-1}]\,[r_{j-1} \mid Y_{j-1}] \end{aligned} \tag{4.26}$$

where the second term in the latter expression is merely the posterior distribution from row $j-1$ which is known from analysis of the previous row. Thus (4.26) defines an iterative scheme through which we can encourage edge continuity. Before implementation we note several factors. First, as is clear from (4.26), we must specify the distribution $[r_j \mid r_{j-1}]$. This could be chosen to take several forms, i.e.

$$[r_j \mid r_{j-1}] = \begin{cases} \dfrac{p}{W+1} & |r_j - r_{j-1}| \leqslant W \\ \dfrac{1-p}{n-W-1} & \text{otherwise} \end{cases} \tag{4.27}$$

for some W and p, that is, constant over a symmetric interval containing r_{j-1}. Another possibility for the choice of $[r_j \mid r_{j-1}]$ would take the form

$$[r_j \mid r_{j-1}] \propto \exp\{-\lambda|r_j - r_{j-1}|\} \qquad |r_j - r_{j-1}| \leqslant W \tag{4.28}$$

for some W and λ. Figure 54 shows the effect of introducing such an updating scheme for $[r_j \mid r_{j-1}]$. Figure 54(a) depicts the results of standard row analysis of the single edge true scene-based with Signal-Noise ratio equal to 1.0 using (2.11), and figure 54(b) depicts the results when (2.11) is implemented in conjunction with (4.28) as the marginal distribution of Y_j conditional on r, with $\lambda = 2.0$ and $W = 4$ for demonstration purposes.

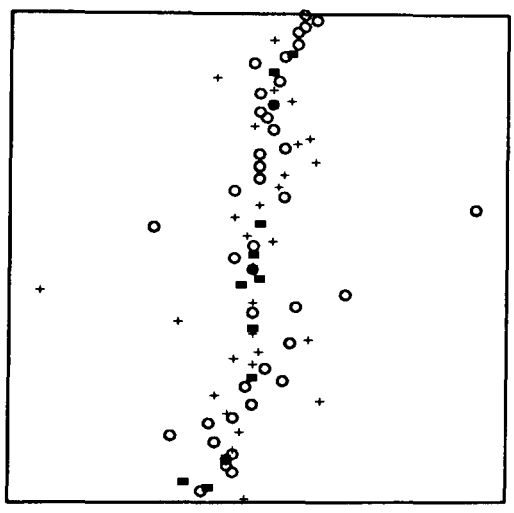


Fig 54(a) : Standard analysis

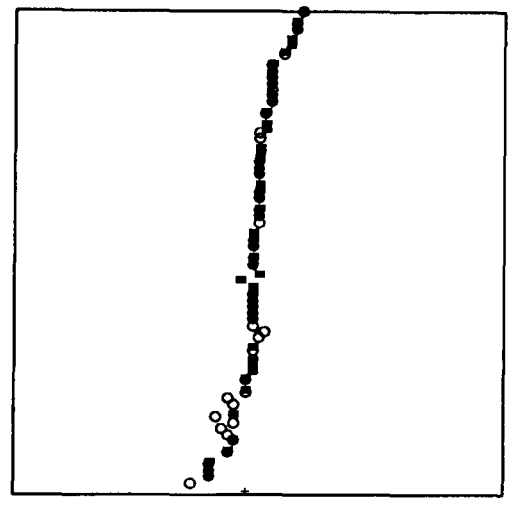


Fig 54(b) : Implementation of (4.28)

The improvement is remarkable. However, the processing time involved in the production of (b) was 16.6 seconds which is considerably more than that required for the standard analysis. It it also interesting at this point to study the effect that different choices of λ have on the resulting set posterior modes. Figure 55(a), (b) and (c) correspond to choices of $\lambda = 1.5$, 1.0 and 0.5 respectively.

It is clear that the choice of λ, although not crucial, does effect the final results. Figure 55 seems to indicate that λ should be chosen to be large, but as we shall see below, this is not always advisable. We also note this stage that the choice of W is important, but that because of our knowledge of the continuous nature of edges in the true scene, is much more straightforward.

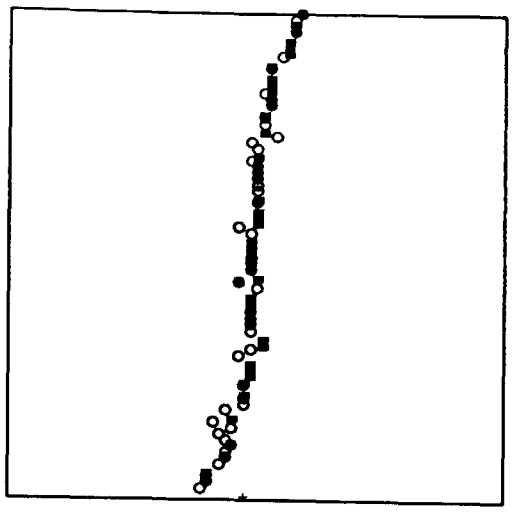


Fig 55(a) : $\lambda = 1.5$

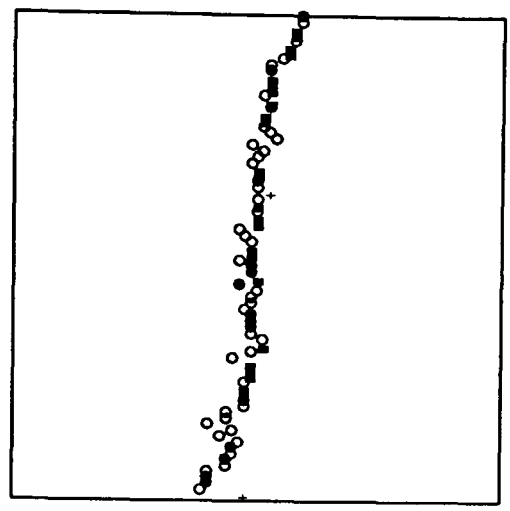


Fig 55(b) : $\lambda = 1.0$

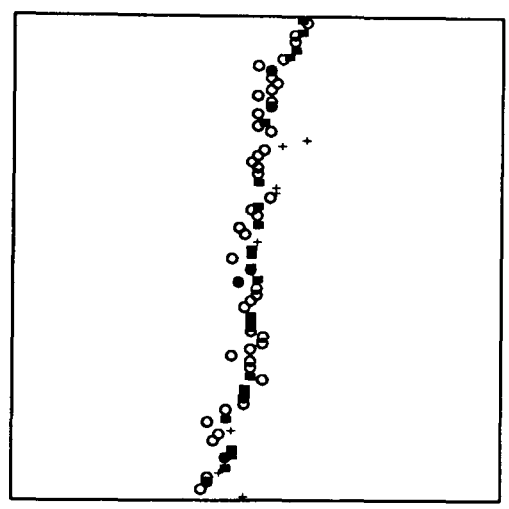


Fig 55(c) : $\lambda = 0.5$

The second factor we note in (4.26) is that in this precise form, $[r_{j-1} \mid Y_{j-1}]$ depends on the posterior distribution $[r_{j-2} \mid Y_{j-2}]$ through $[r_{j-1} \mid Y_{j-2}]$ etc., so that there is some form of relation between the distribution for row j and the distributions for all preceding rows. This has little influence if, as in the single edge true scene above, all changepoints occur at approximately the same position in each row i.e analysis is being carried out in a direction perpendicular to a reasonably straight edge. Generally, however, it may be regarded as undesirable. Fortunately, this property is easily removed, as the changepoint prior acts multiplicatively at each stage and thus its effect can be removed by division . For example, if we require that only adjacent rows should be related, then from (4.26)

$$[r_j \mid Y_{j-1}] = \sum_{r_{j-1}} [r_j \mid r_{j-1}] [r_{j-1} \mid Y_{j-1}]$$

$$\propto \sum_{r_{j-1}} [r_j \mid r_{j-1}] [Y_{j-1} \mid r_{j-1}] [r_{j-1} \mid Y_{j-2}] \qquad (4.29)$$

and so the effect of row $j-2$ on the posterior distribution in row j can be removed by multiplying the posterior probability for each possible realisation of r_j by the corresponding posterior probability for that realisation for row $j-2$. From (4.29) it is clear that this is equivalent to evaluating $[r_{j-1} \mid Y_{j-1}]$ with a uniform prior distribution for r_{j-1} for use in the evaluation of $[r_j \mid Y_{j-1}]$. The extension to allow a relationship between rows a larger distance apart is straightforward. We now study the effect that removing the influence of distant rows in such a manner has on the set of results obtained. First, we examine the effect in the analysis of the single edge true scene. Figure 56(a) depicts the results obtained using a standard implementation of (4.28) reproduced from figure 54(b) for comparison. Figure 56(b) depicts the results obtained when the scheme for the removal of long distance effects discussed above is used. The Signal-Noise ratio at the boundary was again 1.0, and λ and W were nominally chosen to be 2.0 and 4 respectively.

It is clear that, in this case, adjustment of (4.28) to procure the removal of long distance effects is undesirable, as the edge is less well-defined in (b) than in (a) - this is as we would have predicted from the above discussion, as here the row analysis is being carried out in a direction perpendicular to a virtually straight edge, and so information about changepoint position in any row will be relevant to the changepoint position in distant rows. However, two positive aspects can be noted. First, the results in figure 55(b) are more satisfying than those in figure 53(a) from a standard analysis, in the sense that the spread of detected edge-points is smaller, making the edge itself easier to discern. Secondly, the processing time involved in the production of the results in figure 55(b) was 10.9 seconds, representing a time saving over the standard implementation of (4.28) by a factor of a third. We may wish to trade accuracy of results for reductions in processing time at some later stage.

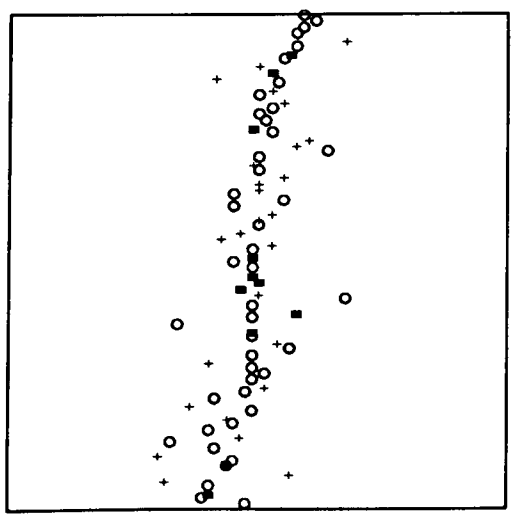


Fig 56(a) : Standard implementation (4.28)

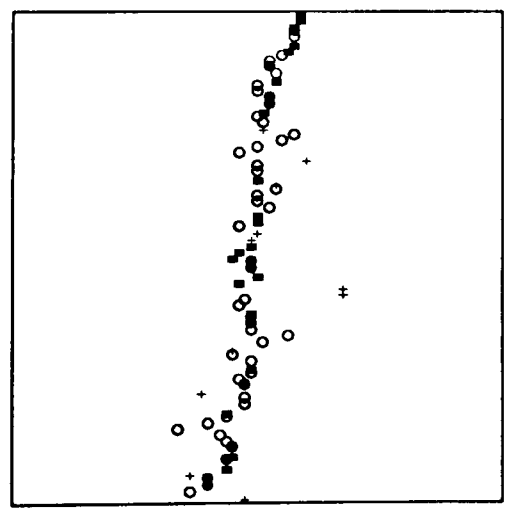


Fig 55(b) : Adjusted implementation (4.28)

We now study the effect of adjusting (4.28) in the way described above when the edge in the true scene has a different orientation. Figure 57 depicts the results of row analysis of a

single edge true scene where the edge is not perpendicular to the direction of analysis. Figure 57(a) depicts the results obtained using the standard implementation, (b) the adjusted implementation. The Signal-Noise ratio was again 1.0, and λ was chosen to be 2.0, as this gave the most satisfying results previously.

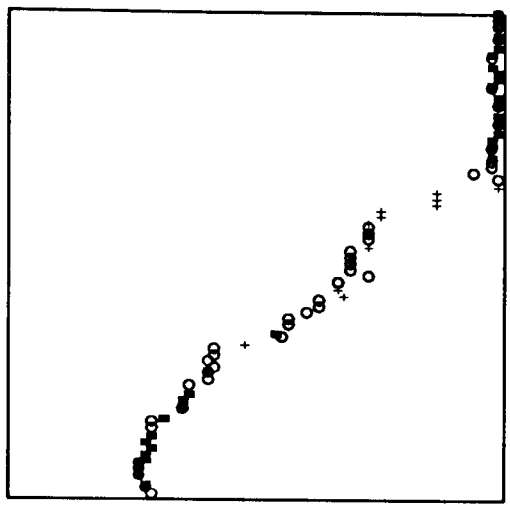


Fig 57(a) : Standard implementation (4,28)

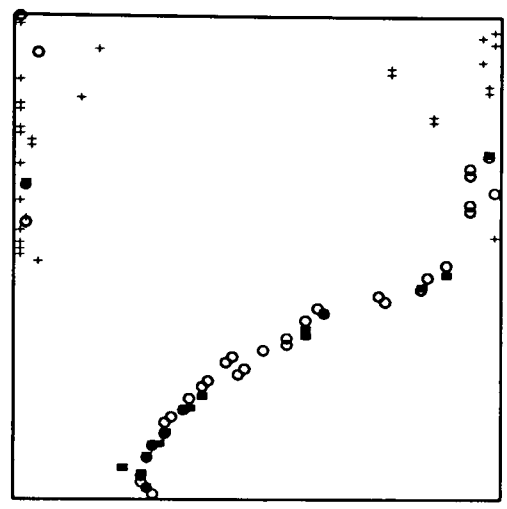


Fig 57(b) : Adjusted implementation (28)

It is clear that there is a marked upward trend in the detected edge in (a) compared to that in (b) - this is again what we would have predicted from our understanding of the standard iterative scheme, as the changepoint posterior distributions for rows several pixels apart are strongly related. Figure 57(b) reflects the actual location and nature of the edge much more accurately. Figure 58 depicts the results of the same analysis repeated with λ now chosen to be 1.0.

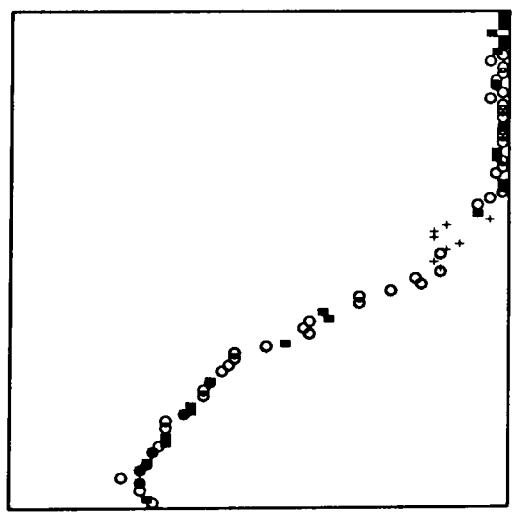


Fig 58(a) : Standard implementation (4,28)

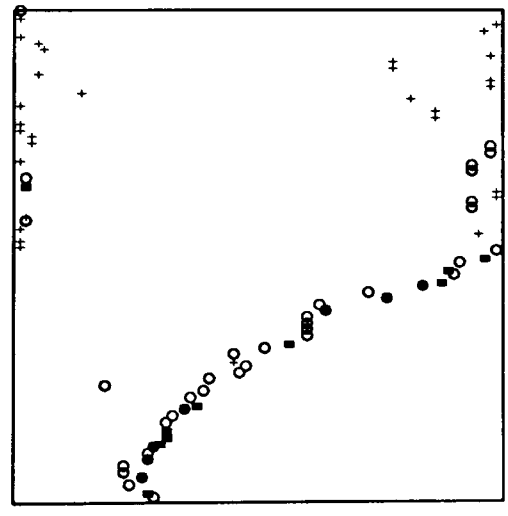


Fig 58(b) : Adjusted implementation (4,28)

Now the difference between (a) and (b) is less marked, indicating that choosing λ large is not necessarily optimal. The analyses involved in the production of figures 57 and 58(b) were again appreciably faster than those for 57 and 58(a).

As a final remark, we note that a full (row and column) analysis of edge true scenes such as that in figures 57 and 58 would have overcome the problems mentioned above - we could have chosen λ to be relatively small and still obtained accurate results - but this would of course increase processing time. Also, use of other forms for $[r_j \mid r_{j-1}]$, i.e. such as that in (4.27), could be designed so as to reduce the effect that distant preceding rows have on those subsequently processed.

The final feature we note in (4.26) is that there is an asymmetry in the processing and updating scheme, i.e. we compute the changepoint posterior distribution for row j dependent on the results from rows $j-1, j-2$ etc. which lie wholly "to the left of" or "below" row j. This does not reflect our prior knowledge of the edge in the true scene, which would indicate some form of symmetry in the influence that adjacent rows have on the posterior distribution in row j (we might have some joint belief *a priori* concerning changepoint positions in row j and rows $j-1, j+1$ for example). We now attempt to adapt (4.26) to incorporate such prior knowledge.

(4.2.2) Three row joint prior specification.

Consider the changepoint prior for row j conditional on the data in rows $j-1$ and $j+1$ denoted by $[r_j \mid Y_{j-1}, Y_{j+1}]$. Then under the usual conditional independence assumptions, an equivalent expression to (4.26) is

$$[r_j \mid Y_{j-1}, Y_{j+1}] = = \sum_{r_{j-1} r_{j+1}} [r_{j-1}, r_j, r_{j+1} \mid Y_{j-1}, Y_{j+1}]$$

$$= \sum_{r_{j-1} r_{j+1}} [r_j \mid r_{j-1}, r_{j+1}] [r_{j-1}, r_{j+1} \mid Y_{j-1}, Y_{j+1}] \qquad (4.30)$$

If we make the additional assumption that in this scheme the posterior distributions from rows $j-1$ and $j+1$ are independent in the processing of row j (in the way described above) then

$$[r_j \mid Y_{j-1}, Y_{j+1}] = \sum_{r_{j-1} r_{j+1}} [r_j \mid r_{j-1}, r_{j+1}] [r_{j-1} \mid Y_{j-1}] [r_{j+1} \mid Y_{j+1}]$$

where the posterior distributions $[r_{j-1} \mid Y_{j-1}]$ and $[r_{j+1} \mid Y_{j+1}]$ are computed using uniform priors for changepoint position (this ensures that this scheme can be used in the analysis of whole images without the induced inter-row relationship being present. If we had not made the independence assumption above, the amount of computation required would be excessively large). We alter notation slightly at this stage - we write the posterior distribution for changepoint r conditional on data Y as $[r \mid Y]_U$ if a uniform prior distribution for r is used. Hence (4.30) becomes

$$[r_j \mid Y_{j-1}, Y_{j+1}] = \sum_{r_{j-1} r_{j+1}} [r_j \mid r_{j-1}, r_{j+1}] [r_{j-1} \mid Y_{j-1}]_U [r_{j+1} \mid Y_{j+1}]_U \qquad (4.31)$$

Note from (4.31) that as we can arrive at, say, $[r_j \mid Y_j]_U$ from $[r_j \mid Y_{j-1}, Y_{j+1}]$ simply by multiplying by an appropriate factor, the bulk of the computation (i.e. the evaluation of the marginal distribution of data conditional on changepoint position but unconditional of other unknown parameters) need only be carried out once for each row in a complete analysis, thus keeping computational costs to a minimum. However, it is also clear from (4.31) that evaluation of $[r_j \mid Y_{j-1}, Y_{j+1}]$ in this way involves a double summation over r_{j-1} and r_{j+1} and thus we would expect an n-fold increase in processing time compared to the evaluation of (4.26). Thus the use of (4.31) in this form is not practicable, and we seek simplifications. For example, we might try forming the univariate distribution on $\{1, \ldots, n-1\}$, $[r_{j-1j+1} \mid Y_{j-1}, Y_{j+1}]$ given by

$$[r_{j-1j+1} \mid Y_{j-1}, Y_{j+1}]_U = [r_{j-1} \mid Y_{j-1}]_U [r_{j+1} \mid Y_{j+1}]_U \qquad (4.32)$$

(representing a coincident changepoint position in rows $j-1$ and $j+1$) and then evaluating the posterior distribution

$$[r_j \mid Y_{j-1}, Y_{j+1}] = \sum_{r_{j-1j+1}} [r_j \mid r_{j-1j+1}] [r_{j-1j+1} \mid Y_{j-1}, Y_{j+1}]_U \qquad (4.33)$$

The amount of computation involved in (4.33) is thus of the same order as that required for (4.26). We could extend the definition in (4.32) to

$$[r_{j-1j+1} \mid Y_{j-1}, Y_{j+1}]_U = \sum_{|r_{j+1} - r_{j-1}| \leq N} [r_{j-1} \mid Y_{j-1}]_U [r_{j+1} \mid Y_{j+1}]_U \qquad (4.34)$$

for small N, but this would increase processing time.

As we have seen, an increase in processing time seems to be inevitable for all of the above techniques. The additional burden is due principally to the marginalisation procedure necessary after specifying *a priori* the joint probability structure for changepoint positions in adjacent rows. In light of this, we now take a fundamentally different approach in an attempt to incorporate prior knowledge concerning edge continuity into the changepoint based edge-detection techniques described above.

We saw in the derivation of (4.26) and (4.30) how marginalisation of the joint prior distribution conditional on data in neighbouring rows could be used to derive useful prior distributions for changepoint positions in single rows. We saw the effect of incorporating *a*

posteriori inferences from adjacent rows. Here, we propose that these posterior inferences (i.e. the changepoint posterior distributions from neighbouring rows) be used directly to derive the single row priors, without the need for the time-consuming marginalisation procedure. We discuss two possible techniques of this type.

(4.2.3) Obtaining prior for r_j via $[r_{j-1} \mid Y_{j-1}]$ and $[r_{j+1} \mid Y_{j+1}]$.

First, we consider choosing the prior distribution for changepoint position in row j conditional on the data in rows $j-1$ and $j+1$, as in (4.31). We might then derive this prior directly as a function of the posterior distributions in the neighbouring rows computed using uniform priors for changepoint positions, denoted $[r_{j-1} \mid Y_{j-1}]_U$ and $[r_{j+1} \mid Y_{j+1}]_U$ respectively in the above notation. For instance, we might combine these two distributions pointwise additively, i.e. so that

$$\Pr(r_j = r \mid Y_{j-1}, Y_{j+1}) = \Pr(r_{j-1} = r \mid Y_{j-1})_U + \Pr(r_{j+1} = r \mid Y_{j+1})_U - \Pr(r_{j-1} = r \mid Y_{j-1})_U \Pr(r_{j+1} = r \mid Y_{j+1})_U , \quad (4.35)$$

or multiplicatively, so that

$$\Pr(r_j = r \mid Y_{j-1}, Y_{j+1}) = \Pr(r_{j-1} = r \mid Y_{j-1})_U \Pr(r_{j+1} = r \mid Y_{j+1})_U \quad (4.36)$$

for $r = 1, \ldots, n-1$, assuming, as above, that the changepoint variables r_{j-1} and r_{j+1} are independent in the derivation of $[r_j \mid Y_{j-1}, Y_{j+1}]$. Evaluation of the priors in this way would induce an n fold reduction in processing time compared to (4.26). Note that we regard the symmetric (in $j-1$ and $j+1$) forms of (4.35) and (4.36) as essential here to minimise the disruptive effect of isolated outlying or extreme data values in the image on the changepoint posterior distributions. We also regard the additive form (4.35) to be more intuitively reasonable for use in the changepoint/edge-detection context. We now study the effect that, for example, (4.35) has on the results obtained using an otherwise standard analysis. Figure 59 depicts the results obtained of an analysis using tha adjusted version of (4.28) discussed above, and the modification in (4.35), where the single edge true scene concerned was corrupted to produce an image with Signal-Noise ratio equal to 1.0 at the boundary.

It is clear that, for this image, the results from the adjusted version of (4.28) and an implementation of (4.35) are very similar, both being an improvement on the results obtained by a standard analysis. Crucially, however, the processing times involved in the production of the results in (a) and (b) were 10.12 and 1.16 seconds respectively. Therefore, we have achieved of the order of ten-fold reduction in processing time by using the modification

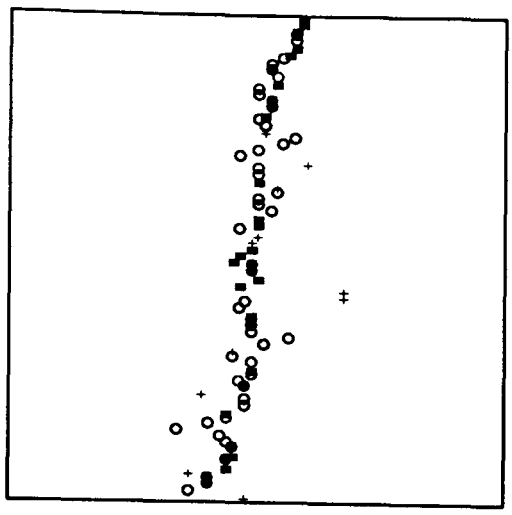


Fig 59(a) : adjusted (4.28)

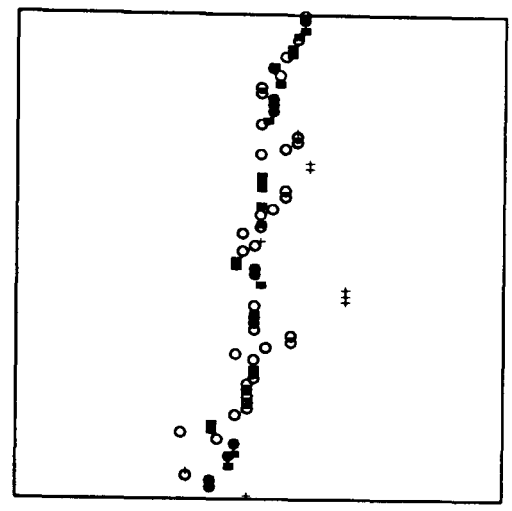


Fig 59(b) : (4.35)

(4.35) compared to the adjusted version of (4.28). The improvement is also noticeable in a full row/column analysis of the image underlying figures 57 and 58, where again the Signal-Noise ratio was 1.0 the results of which are depicted in figure 60.

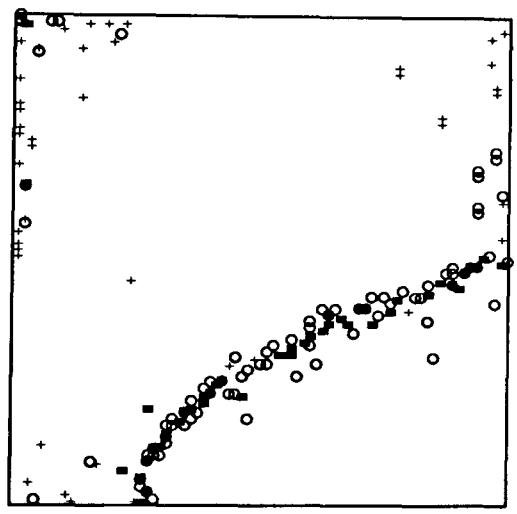


Fig 60(a) : adjusted (4.28)

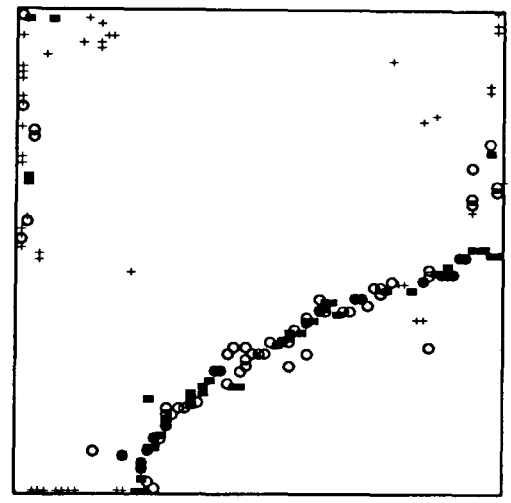


Fig 60(b) : (4.35)

The results are again comparable and the processing times involved were 22.32 and 2.24 seconds for (a) and (b) respectively. Thus again the implementation of (4.35) is a factor of 10 quicker than the adjusted version of (4.28).

So far in this section, we have only studied modifications to the one changepoint posterior based technique for edge-detection. We saw in a previous section, however, that the one changepoint posterior could be used to make approximate inference in more complicated multi-changepoint situations corresponding to more complex (convex object) true scenes. For completeness, we include here one such approximate analysis, that of the circle in figure 10, with the modifications concerning edge continuity included. Figure 61 depicts the results obtained from an analysis of a centrally positioned circle with Signal-Noise ratio equal to 1.0 by each of the three techniques represented in figures 59 and 60. In figure 61(a), the image

was analysed with the "exactly one changepoint" version of (2.11), and in figure 61(b) λ was chosen to be 1.0, and W was set equal to 4.

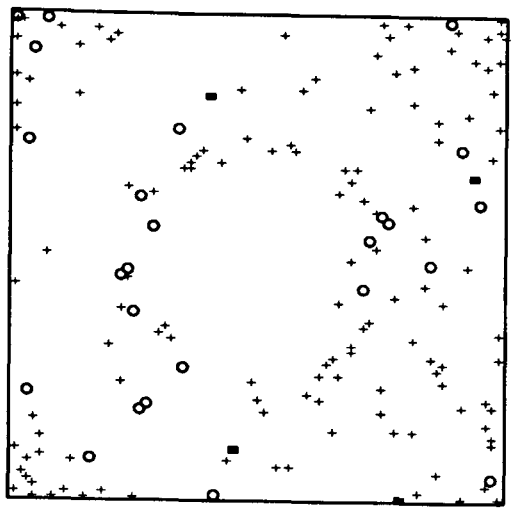


Fig 61(a) : standard analysis

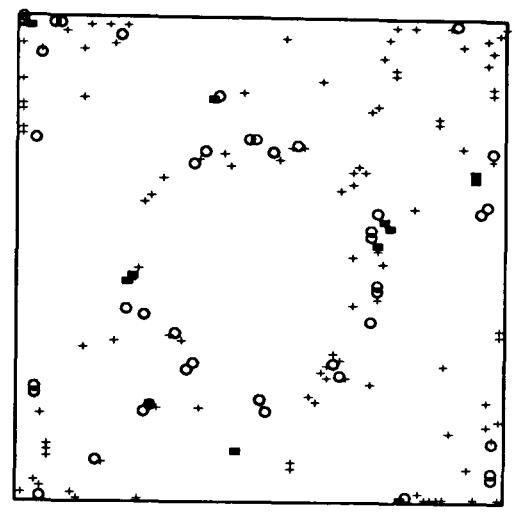


Fig 61(b) : adjusted (4.28)

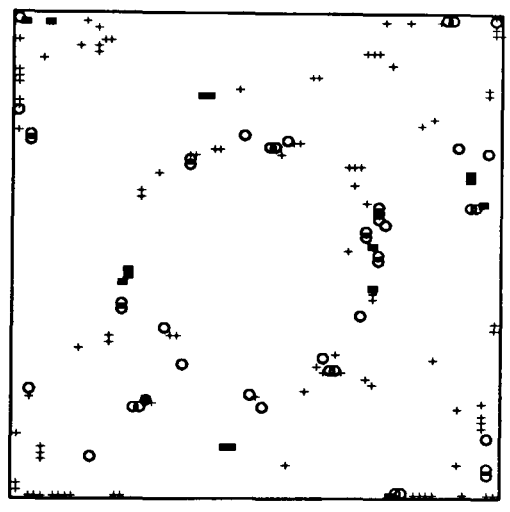


Fig 61(c) : (4.35)

The results in (b) and (c) do seem to be an improvement on those in (a), but not as great an improvement as for the single edge true scene above. It should be noted that everything that we have derived using prior knowledge of changepoint position over adjacent rows in the single changepoint case can be reproduced in multiple changepoint case if necessary, but would naturally involve further increases in processing time.

Finally in this section, we propose one further technique by which the prior distribution for r_j may be obtained from the posterior distributions in rows $j-1$ and $j+1$. Consider the changepoint prior distribution for row j conditional on the posterior estimates of changepoint positions (i.e. posterior modes) for rows $j-1$ and $j+1$ derived using uniform priors for changepoint position and denoted $[r_j \mid \hat{r}_{j-1}, \hat{r}_{j+1}]$. Priors of this form are of interest here as in the edge-detection problem we would like the edge-point estimates in neighbouring rows to be in close proximity to each other. Also, no marginalisation procedure is necessary and

therefore we might expect processing techniques based on such priors to be reasonably time efficient. Again the bulk of the calculation, the evaluation of marginal row data distributions, need be carried out once only for each row. We might specify $[r_j \mid \hat{r}_{j-1}, \hat{r}_{j+1}]$ to be of the form

$$[r_j \mid \hat{r}_{j-1}, \hat{r}_{j+1}] = \begin{cases} K\exp\{-\lambda|r_j - M_j|\} & |r_j - M_j| \leq W_1 \\ 0 & \text{otherwise} \end{cases} \tag{4.37}$$

where $M_j = (\hat{r}_{j-1} + \hat{r}_{j+1})/2$ and K is a normalising constant. We might introduce a further feature where $[r_j \mid \hat{r}_{j-1}, \hat{r}_{j+1}]$ is chosen to be uniform on the range between $\hat{r}_{j-1}$ and $\hat{r}_{j+1}$ and zero elsewhere if $|\hat{r}_{j-1} - \hat{r}_{j+1}| > W_2$ as this would indicate some unwanted spatial discrepancy between adjacent edge-points. By choosing λ to be zero in (4.37) we obtain a uniform distribution over the range $\{M_j - W_1, \ldots, M_j + W_1\}$. Again, choices of W_1 and W_2 can be made with reference to ideas about edge continuity. Many priors such as (4.37) may be specified.

We now investigate the effect of priors such as (4.37). Figure 62 depicts the results of three analyses of the familiar single edge true scene. Figures 62(a) depicts the results obtained from a row analysis using the modification in (4.35) respectively, whereas (b) depicts the results obtained when (4.37) is implemented, with $\lambda = 3.0$, $W_1 = 4$ and $W_2 = 6$. Again, the Signal-Noise ratio at the boundary was 1.0.

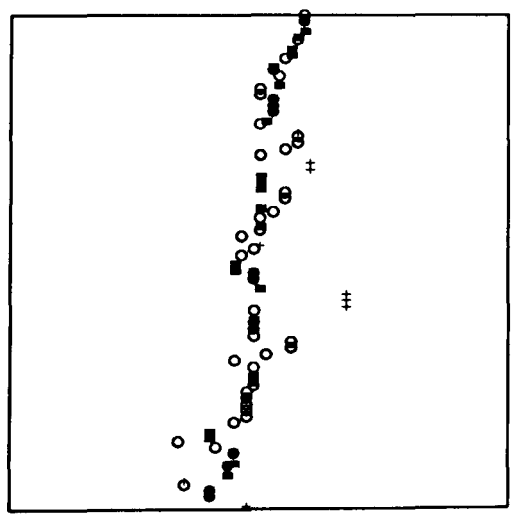


Fig 62(a) : (4.35)

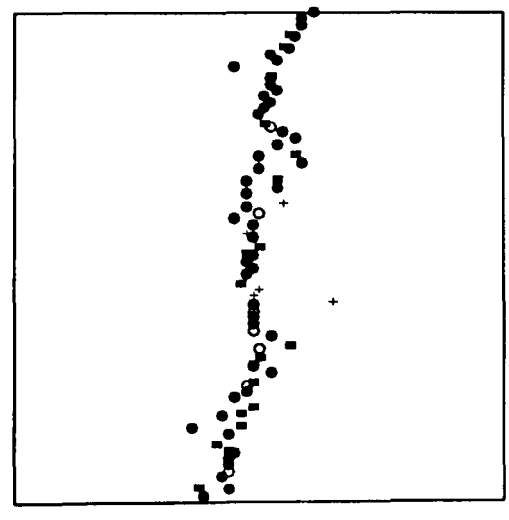


Fig 62(b) : (4.37)

It is clear that the results in (b) are comparable with those in (a) as a representation of the edge, in terms of the degree of continuity exhibited. The processing time involved in the production of the results in (b) was 1.24 seconds, marginally slower than for a standard analysis, and comparable to an analysis using the modification (4.35). Thus it seems that using priors such as that in (4.37) has advantages, as well as being perhaps more intuitively

appealing than, say, (4.35) or (4.36).

It is interesting to see how altering λ affects the results obtained. Figure 63 depicts the results obtained when λ was chosen to be 0.0 and 10.0 in (a) and (b) respectively. The same values for W_1 and W_2 were used as for figure 62.

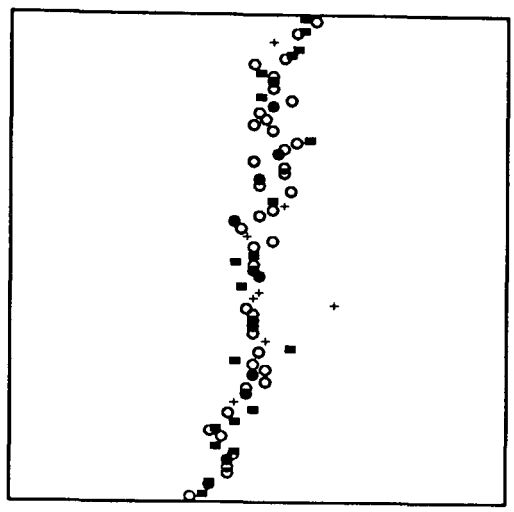


Fig 63(a) : λ = 0.0

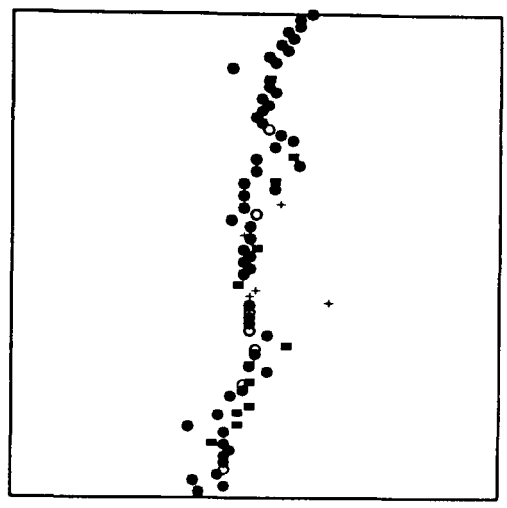


Fig 63(b) : λ = 10.0

The results are broadly similar. Thus it seems that this prior is less sensitive to choices of λ than was, say, (4.28). This must be regarded as a positive aspect, as we would need our posterior inferences to be to some degree robust to a range of prior specifications - recall that λ is merely a hyperparameter relating to the (conditional) structure of changepoints in adjacent rows.

For completeness, we include the full analysis of the single edge true scene/image in figure 57 using the prior in (4.37). Figure 64(b) depicts the results obtained when using (4.37) with λ = 3.0, whereas (b) depicts the results from a standard analysis using (4.35) Recall that the Signal-Noise ratio for the image concerned was 1.0. Again the results seem satisfactory, and the processing times for (a) and (b) were again comparable.

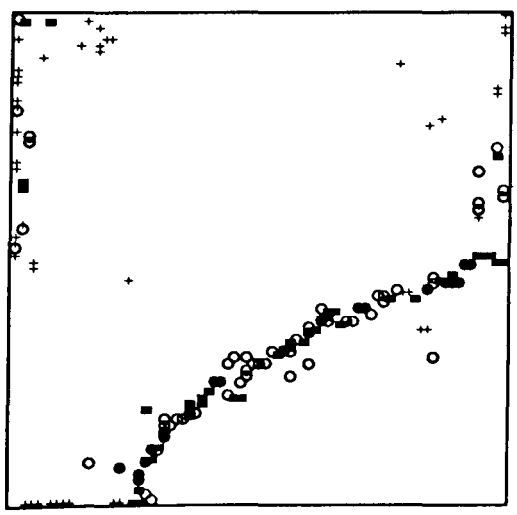


Fig 64(a) : (4.35)

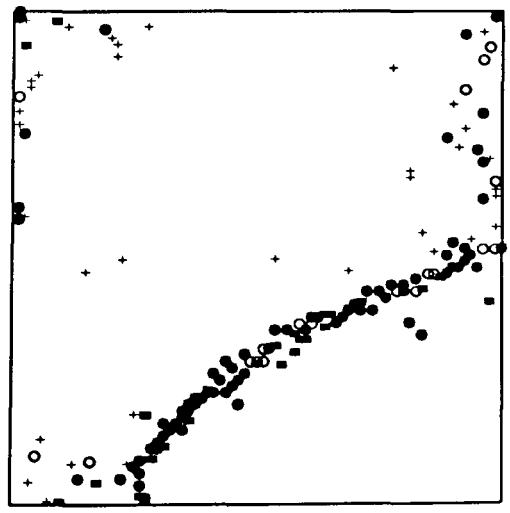


Fig 64(b) : (4.37)

Thus it seems that priors of the form of (4.37) compare favourably with the other ideas and techniques we have seen. One potential misgiving we may have is that using priors conditional on posterior modal positions in neighbouring rows may not be sufficiently effective for approximate inference in multiple changepoint cases. Somewhat surprisingly, however, priors of the form of (4.37) seem to perform quite as adequately as the other techniques discussed above. This is demonstrated by figure 65, where (a) depicts the results of a standard analysis, and (b) the results of an analysis with (4.37) implemented, with $\lambda = 3.0$.

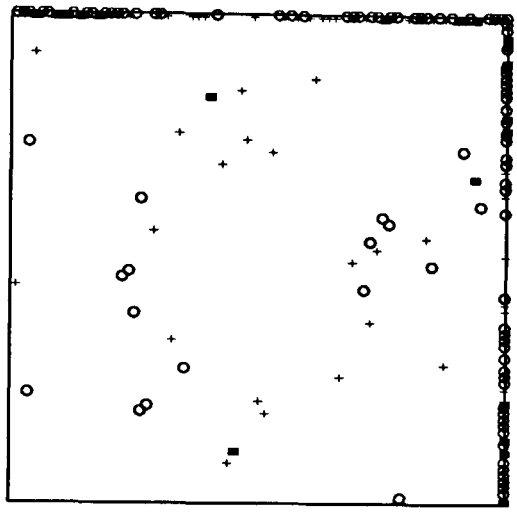


Fig 65(a) : standard analysis

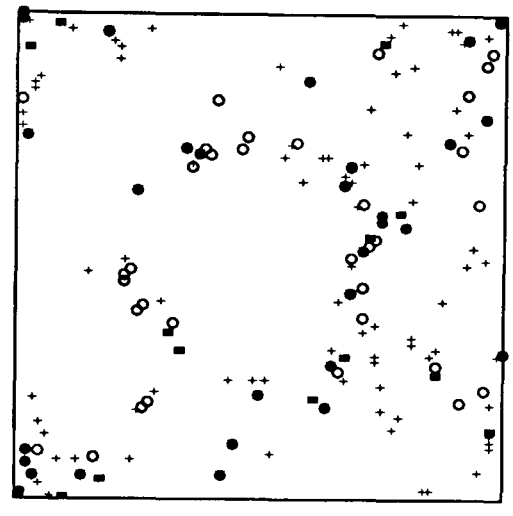


Fig 65(b) : (4.37)

(4.4) Post-processing.

As we have seen above, it is possible to incorporate ideas of spatial continuity into changepoint based techniques via priors for unknown parameters. In precisely the same way, given a set of edge-point candidates with their associated posterior probabilities resulting from a changepoint based analysis of an image, we would like to be able to use these ideas in order to remove isolated and therefore probably mis-classified points from this set. We have already used a simple "smoothing" technique of this nature to enhance the appearance of a set of results, and more importantly, we shall see later when we study parametric edge reconstruction techniques that simple estimation procedures (e.g. least-squares estimation) are extremely sensitive to the presence of outlying mis-classified points. Hence we now proceed to study briefly some simple post-processing techniques.

(4.3.1) Naive post-processing techniques.

First, and most simply, we could accept or reject an edge-point candidate on the basis of its associated posterior probability alone, i.e. accept the point if the probability is greater than some pre-fixed threshold, reject otherwise. This is somewhat of an ad-hoc technique, but is perhaps more acceptable than the other threshold based criteria mentioned previously as the quantity of interest is a posterior probability rather than, for example, some arbitrary intensity

level. Also, such a technique can be justified in terms of the Bayesian decision problem concerning changepoint identification - recall that the Bayesian estimate of changepoint position is the posterior modal position corresponding to the minimised Bayes risk under a specified loss function. However, in practice, this technique is inefficient for removing isolated edge-points, as are others purely based on posterior probabilities.

Secondly, we could use only our knowledge of edge continuity to remove mis-classified points, i.e. an edge-point distant from any other recorded edge-points is by definition a mis-classification. This idea forms the basis for the smoothing technique we saw above. There, the number of edge-points recorded as modes in the changepoint posterior distributions lying in a small sub-grid centred at each individual edge-point in turn was counted, and if that number was again greater than a pre-fixed threshold the individual point was accepted. This technique, although easily implemented and reasonably effective, is difficult to justify in a decision-theoretic setting. It clearly captures the essence of the problem, but requires a degree of input (setting sub-grid size and choosing the threshold number of pixels) which might be regarded as too detailed, even though we appeal to the ideas of spatial continuity etc. discussed in detail previously. Fortunately, in many of the simple cases we have studied (the single edge true scene processed by the standard changepoint technique, the convex object true scene processed using binary segmentation, complex true scenes processed using Gibbs Sampler based methods all at high (> 1.0) Signal-Noise ratios and the improved results obtained using spatial priors) the simple smoothing technique is reasonably robust to changes in grid-size and threshold value as the edge itself is boldly delineated compared with the mis-classified points which are spread relatively diversely. We also reiterate that for the purpose of keeping processing time to a minimum, we may have to compromise and use such simple and intuitively appealing but basically ad-hoc techniques, provided that the results are robust to input parameters.

Finally, we consider a simple iterative scheme based on the standard edge-detection techniques developed above (we regard these as post-processing operations as they are implemented subsequent to and dependent on the results of the initial analysis). Buck *et al.* (1988) describe an iterative technique based on changepoint analytic methods for the segmentation of a two texture true scene, altering prior values for the texture parameters on the basis of the results after each iteration. Here, in the edge-detection context, we consider altering the changepoint prior between iterations in the same way.

(4.3.2) An iterative post-processing scheme.

Consider first the row analysis of a true scene using changepoint techniques but with each row being treated independently from all others, and using uniform priors on changepoint position, as in our initial examples in chapter 2. After the row analysis is complete, we have obtained a posterior probability for each pixel in each row of the image, namely the

probability that the changepoint in that row occurs at that pixel, conditional on the data in that row. Let $E^{(0)}$ be the $n \times n$ matrix whose entries are the n^2 posterior probabilities . Now consider the matrix $P^{(0)}$ whose elements are formed from the elements of $E^{(0)}$ by some local operation. In particular, consider the case where the (i,j)'th element of $P^{(0)}$ is given by

$$\left[P^{(0)}\right]_{ij} = \frac{1}{K_j} \sum_{(k,l) \in \partial ij} \left[E^{(0)}\right]_{kl} \tag{4.38}$$

i.e. a local average, where K_j is a scaling constant to ensure that the elements of row j sum to 1. The next step in the iterative procedure is to form the matrix $E^{(1)}$ where

$$\left[E^{(1)}\right]_{ij} = \left[E^{(0)}\right]_{ij} \left[P^{(0)}\right]_{ij} . \tag{4.39}$$

It is then clear that the elements of $E^{(1)}$ are merely the row posterior probabilities on changepoint position, where the probabilities are evaluated using priors for changepoint position defined by the rows of $P^{(0)}$ - recall that these prior probabilities update the posterior probabilities obtained using uniform priors in a simple multiplicative fashion. To obtain estimates of changepoint positions in we then normalise and locate the maximum row by row in the usual way. We now repeat this procedure and form $P^{(1)}$ from $E^{(1)}$ via (4.38), and then $E^{(2)}$ from $E^{(0)}$ and $P^{(1)}$ via (4.39), and normalise and locate the row modes etc. until the positions of these modes stabilise. The amount of computation required to implement such a scheme would not appear to be overly large, as merely simple local averaging operations are necessary in addition to the usual probability calculations and maximisation routines.

We now illustrate the use of the iterative scheme defined by (4.38) and (4.39) on images derived from the simple edge and circle true scenes. Figure 66 depicts the results obtained after each of the first three iterations of the scheme when a row analysis using the posterior distribution (2.11) is carried out, the Signal-Noise ratio at the boundary being 1.0.

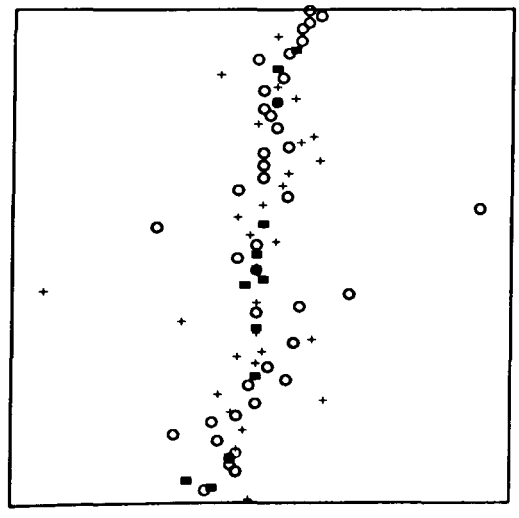


Fig 66(a) : Raw results

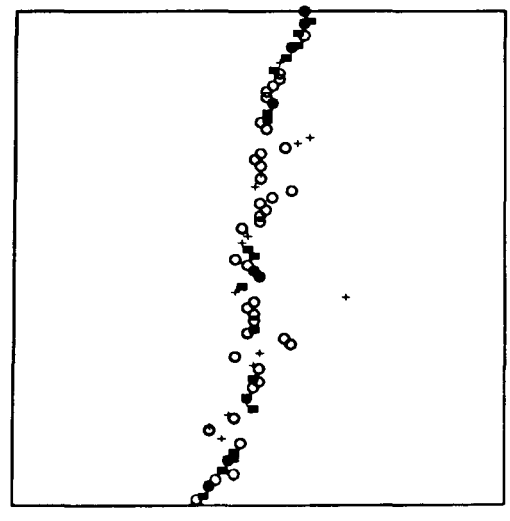


Fig 66(b) : First iteration

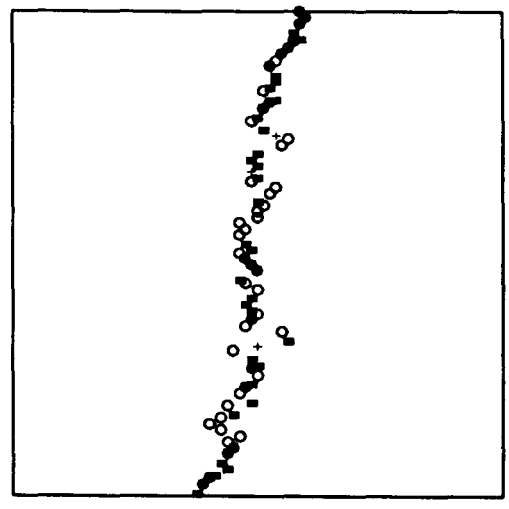


Fig 66(c) : Second iteration

There is a marked improvement between (a) and (b) in terms of the results being a representation of the edge. The additional processing time needed was 1.3 seconds per iteration. Figure 67 depicts the results obtained after each of the first three iterations of the scheme for a full analysis again based on the posterior distribution (2.11). The Signal-Noise ratio involved was again 1.0.

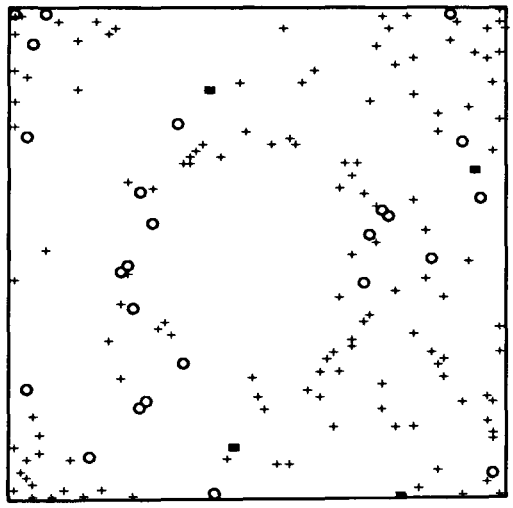


Fig 67(a) : Raw results

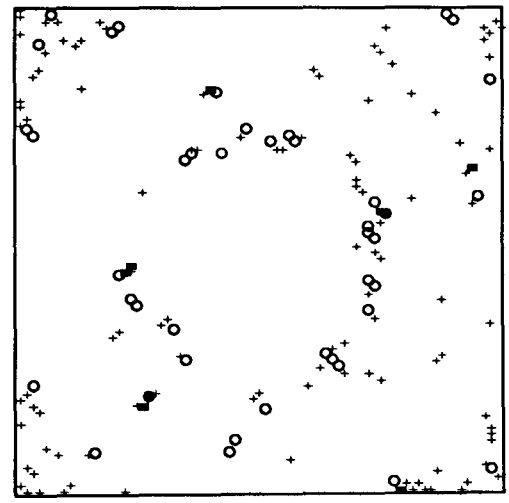


Fig 67(b) : First iteration

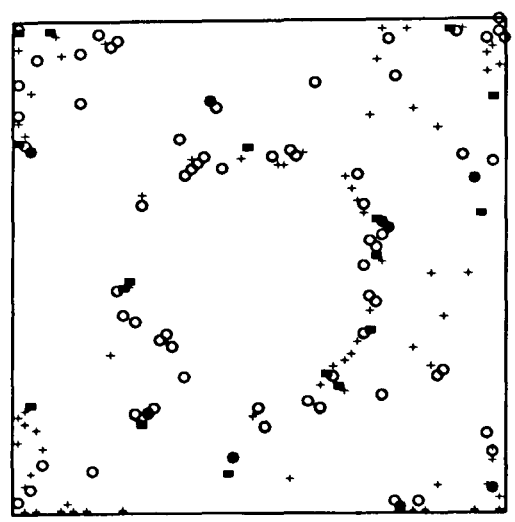


Fig 67(c) : Second iteration

Again there is a clear improvement after each iteration, with the edge being located with increasing accuracy. The additional processing time here was 2.6 seconds. Thus, in both cases, the procedure leads to more satisfying results. Naturally, we encounter the usual difficulties associated with such iterative schemes which we have no completely satisfactory solutions other than simple intuitive ones. For instance, we again merely assess "convergence" of the algorithm via modal position stability etc. without fully understanding the nature and behaviour of the results at intermediate stages. However, in practice, more than three iterations are rarely needed, and so we content ourselves with accepting the results at this stage.

It is easy to develop other simple iterative schemes of this nature which can be used for the post-processing of results from changepoint based analyses.

(4.5) Spatial dependence and edge continuity - conclusions.

We have seen in this section how to incorporate spatial prior knowledge into our changepoint based analytic techniques, with the intent of removing isolated or mis-classified edge-points. We first attempted to use Gibbs-type priors for the true scene pixel parameters, and saw that the amount of computation involved proved to be restrictive. However, approximate versions of these types of priors were successfully and efficiently used as the basis of noise-reduction algorithms for pre-processing of the image data. We then developed several special forms for changepoint prior distributions in an attempt to encourage edge-continuity. Finally, we discussed post-processing schemes, and suggested one particular iterative procedure. The majority of examples presented concerned the analysis of single edge true scenes using the one changepoint posterior distribution (2.11), with the obvious extension to the multiple changepoint case being mentioned.

Chapter 5 : Variation of Image–Formation Process.

We have seen how under a simple linear form for the image-formation process, the edge-detection problem may be straightforwardly formulated to be equivalent to an extensively studied version of the statistical problem of changepoint identification - that in which each pixel in the discretised version of the true scene is corrupted identically, independently and additively with Gaussian white-noise in the formation of the image. This a model commonly used in image processing (see, for example, Hansen and Elliot (1982)), specifically used in the study of remote sensing and satellite data. However, it is clearly limited in its applicability, and we now seek more general models, and attempt to apply the ideas we have seen previously concerning changepoint analysis and identification to these models.

(5.1) Mathematical representation of the image-formation process.

We begin by studying the standard formulation of the image-formation process described by Geman and Geman (1984), and initially follow their notation. Let F denote the true scene as represented by the discretised (pixel) intensity process, and let N represent the noise-process. Then the observed (degraded) intensity image G can be written

$$G = \psi(\phi(H(F)), N) \tag{5.1}$$

where H affects local-averaging or "blurring" on F (and thus corresponds to the point-spread function in classical image processing), ϕ is some (possibly non-linear) transformation function, and ψ represents a combination function, used to incorporate the noise-process into the the degradation model. A (not unrealistic) simplification of (5.1) is made by assuming that N corrupts each pixel independently. Thus (5.1) can be written

$$G = \phi(H(F)) * N \tag{5.2}$$

where * represents a (usually additive or multiplicative) combination function so that now N acts (functionally) independently of F. If we consider each pixel in isolation, then for pixel (i, j), (5.2) is equivalent to

$$G_{ij} = \phi\Big(\sum_{(k,l)} H(i-k, j-l) F_{kl} \Big) * N_{ij} \tag{5.3}$$

where the summation runs over pixels in the vicinity of (i, j). Geman and Geman also make

additional assumptions concerning the nature of F and N, and introduce the idea of a true scene line process, L (corresponding to edge positions in the true scene specified relative to pixel positions), which is not transformed by the image-formation process. In the segmentation problem, interest then centres on specifying prior forms for the true scene F which take the form of Gibbs/M.R.F. distributions. Now, in the notation adopted above, we regard the transformed version of F by H and ϕ as being merely equivalent to another process θ (i.e. $\phi(H(F)) \equiv \theta$), and replacing N by ε, and G by Y, we merely have that $Y = \theta * \varepsilon$, or

$$Y_{ij} = \theta_{ij} * \varepsilon_{ij} \tag{5.4}$$

as implied in (2.1). Thus it is clear how the conditional independence assumptions concerning the Y_{ij}'s is justified. The formulation (5.4) is equivalent to that used by Besag (1986) and others for solution of the segmentation problem. Note that, in this notation, (5.1) may be re-written simply as $Y = f(\theta, \varepsilon)$.

Thus, despite the complexity of the physical nature of the image-formation process described in (5.1), (5.2) and (5.3), the final form as given in (5.4) is relatively straightforward, and we need only concentrate on specifying forms for the noise-process (error structure). We then face the problem of making inference about the unknown θ (or F) and L conditional on the realisation Y (or G) through (5.4) and statistical decision theory via Bayes theorem, and so we must also choose prior distributions for these unknown parameters. The methods of solution of the segmentation and edge-detection problems are then crucially different. In the segmentation problem, as we have basically one observation per unknown parameter (i.e. one realisation derived from the true scene value at each pixel), we must specify complex spatial priors so as to maximise the influence of our relevant prior knowledge concerning local dependence. In the edge-detection problem, however, we have vastly fewer numbers of unknown parameters of interest (i.e. edge positions in rows and columns), and so we can afford to use less complex prior forms for these parameters and the (nuisance) parameters (used to describe texture characteristics) in our analysis (we marginalise the problem from interest in (θ, L) to interest solely in L by integration). Indeed, we have seen in the examples above that, for our changepoint based techniques, we may even process individual rows and columns independently in many cases and still obtain useful results. It is these simplifying measures which allow less time consuming processing algorithms to be devised.

We now study different simple choices for the function *. First, we revert to our initial assumption, where * is taken to be additive.

(5.2) Additive noise corruption.

We saw in our initial examples in chapters 2 and 3 how the simple linear image-formation process in (2.1)

$$Y_{ij} = \theta_{ij} + \varepsilon_{ij} \tag{5.5}$$

played an important role in our changepoint based edge-detection techniques, and in statistical image-processing in general. We chose to investigate the situation where the ε_{ij} were independent and identically distributed Gaussian white-noise variables, and the model for the true scene in which the θ_{ij} were chosen to be equal across textures. We now seek other interesting choices of the image-formation process. Recall that, in our solution of the edge-detection problem using changepoint analytic techniques, one necessary step was to calculate the functional form of the likelihood $[\, Y \mid r\,, \theta\,]$ i.e. the conditional distribution of the data in each row given the true scene parameters and edge/changepoint position in that row. Thus we restrict attention to those image-formation processes for which the calculation of $[\, Y \mid r\,, \theta\,]$ in product form is straightforward.

(5.2.1) ε_{ij} Normally distributed.

Consider first the two texture image derived from a single edge true scene such as that in figure 3, with characteristic parameters $(\theta_1\,, \theta_2\,)$. First, we consider the case which we have to some extent studied previously, where the error terms ε_{ij} are Normally distributed and the textures are presumed homogeneous - thus we write $\theta_i \equiv \theta_i,\ i = 1\,, 2$. Then there are clearly three reasonable models we may use for the true scene and error structure for use in the edge-detection techniques -

(1) Common variance in error terms across textures.

(2) Common texture mean.

(3) Different texture means, different error variances between textures.

(1) we have already studied, (2) and (3) are other possible models, perhaps arising from image-formation processes having different physical attributes. Note that in (1) the difference in distribution of the elements in observed image Y is purely due to our assumptions concerning true scene pixel parameters, in (2) the difference is purely due to different error assumptions, and in (3) the difference is due to a combination of both. Also, here we only consider models inducing within-texture homogeneity in terms of distribution of the observed image Y - this restriction is entirely reasonable (and ultimately necessary, although it is sometimes introduced at higher levels of the hierarchy in other areas of the statistical modelling of images).

In each of these cases, where the error terms ε_{ij} are presumed independent, the observed data Y_{ij} are clearly conditionally independent given the true scene parameters, and thus $[Y \mid r, \theta]$ may be formed easily for data in each row/column, with the dependence on edge/changepoint position being of utmost importance in this likelihood. It now remains to specify prior distributions for the unknown parameters in $[Y \mid r, \theta]$, that is, the texture parameters and changepoint position. In the examples we have seen, we have used non-informative prior distributions for the continuous parameters, which can be viewed as limiting forms of conjugate prior distributions. We chiefly restrict attention here to such conjugate prior forms because of their analytic tractability which allows for less time consuming processing.

The resulting changepoint posterior forms under a range of prior assumptions (certain parameters known, dependent priors etc.) can be found in Appendix 1. We concentrate principally on case (1) above, for the reasons discussed, and also because it represents the most widely studied situation in the changepoint literature. We include this appendix for completeness, but present no examples of the use of the range of posterior distributions in the edge-detection context as we feel that we have sufficient knowledge and experience of the behaviour of such distributions. We note that amount of computation and thus processing time required increases linearly with sequence length n for all of the one changepoint posterior distributions included, and thus overall processing times for the analysis of images should be comparable.

It is interesting to study the results obtained under an incorrect model specification, for example, under assumption of common variance for the error terms when in fact the different textures are corrupted by different levels of noise (we might regard this as a technique for detecting changes in underlying mean level even if we suspect that there is also a change in variance). First, we repeat our simulation experiments to study the behaviour of the "common variance" changepoint posterior distribution in such a situation. Figure 68 depicts the posterior distributions obtained when calculated via (2.11) for sequences in which the change in mean level is from 0.0 to 1.5, but where there is also a change in standard deviation of the error terms from an initial value of 1.0. Figures (a) and (b) depict the results obtained when there is a decrease in error standard deviation, whereas (c) and (d) correspond to an increase. The actual changepoint position was 32 in a sequence of length 80. As before, the posterior distributions shown are obtained by averaging over 1000 replications, and thus can be regarded as expectation results (expectation taken with respect to the data distribution). Two features are apparent. First, the posterior mode in each case corresponds precisely to the actual changepoint position. This is encouraging, as it indicates that the common variance posterior distribution is of use even when it represents a mis-specification. Secondly, the modal value of the distribution decreases and its variance increases as the error variance increases. Further

experimentation indicates that the distribution becomes approximately uniform in appearance when the standard deviation change is of the order of +1.5. Thus we might expect the mis-specification to be of importance when the error terms have these orders of magnitude for their variance.

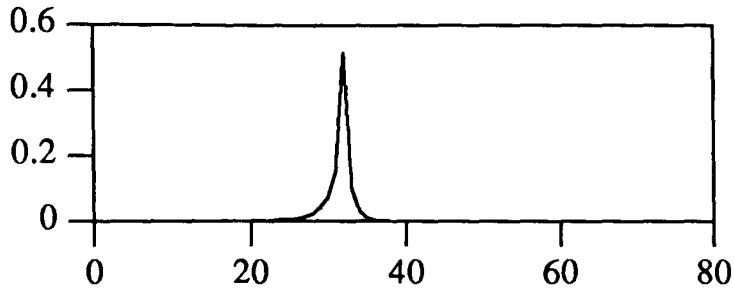


Fig 68(a) : Std. dev. change -0.5

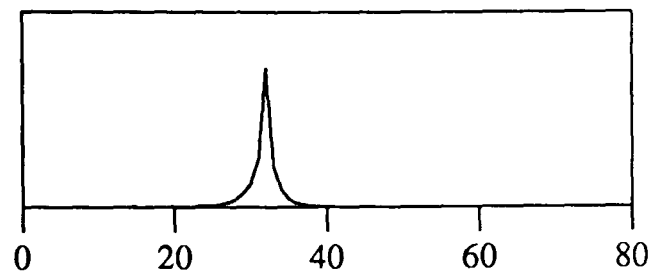


Fig 68(b) : Std. dev. change -0.25

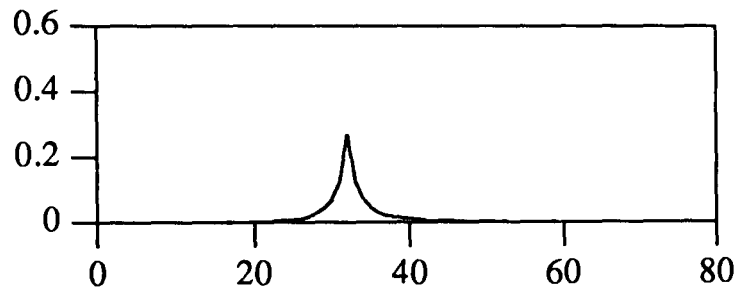


Fig 68(c) : Std. dev. change 0.25

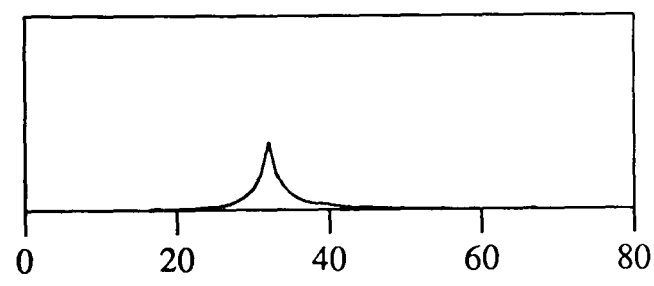


Fig 68(d) : Std. dev. change 0.5

It is also interesting to compare the (expected) posterior distributions obtained for two changepoint sequences as in figures 11, 12, and 13 using (2.11) when there is again a different error structure for each texture. Figure 69 depict the "expected" (i.e. averaged over 1000 runs) posterior distributions calculated using (2.11) in which the change in mean level is 2.0, and a change in variance that increases from (a) to (d). The error variance for the "outer" textures is 1.0. For demonstration purposes, we choose the actual changepoint positions to be 25 and 55.

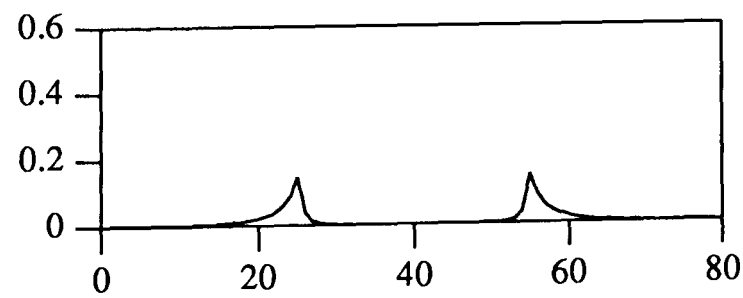


Fig 69(a) : Std. dev. change -0.5

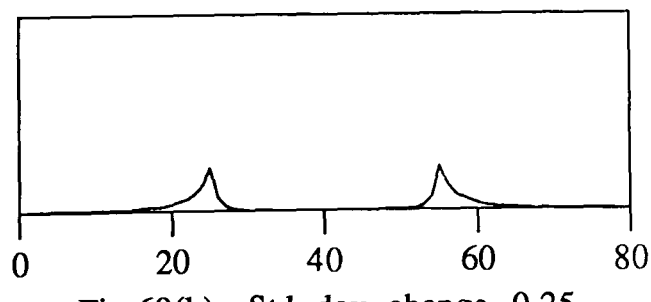


Fig 69(b) : Std. dev. change -0.25

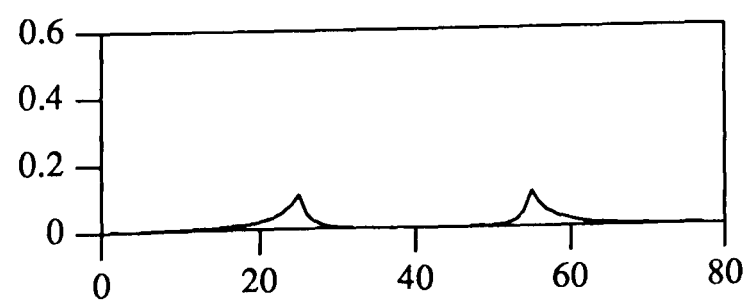


Fig 69(c) : Std. dev. change 0.25

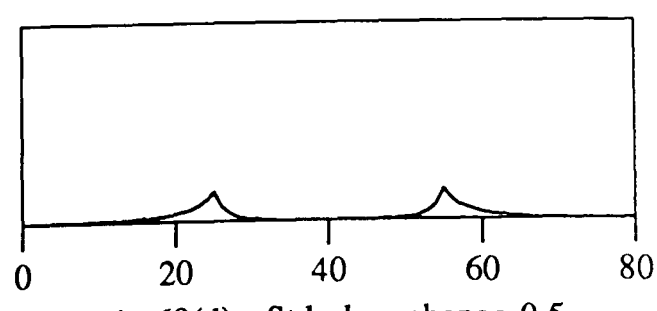


Fig 69(d) : Std. dev. change 0.5

We note that the features present above are also present here. So again mis-specification of the

model in this way frequently allows adequate results to be obtained. Note also that the exact behaviour of (2.11) when actually it represents such a mis-specification in light of the true data distribution could have been predicted from results analogous to those in section (3.2) of chapter 3.

Finally, we inspect the results obtained when the posterior distribution (2.11) is used as the basis of an edge-detection technique in situations when it in fact represents a mis-specification relative to the actual image-formation process. Figure 70 depicts the results of a row analysis of the single edge true scene where the change in mean level was from 0.0 to 1.0, and the initial variance was 1.0. This series of results is exactly line with what we would have predicted form the posterior distributions depicted in figure 68. Figure 71 depicts the results obtained by a similar analysis of the circle true scene with background of mean level 0.0 and circle mean level 2.0, with background error variance 1.0 and varying circle error variance. Again, the results are in line with the expected posterior distributions depicted in figure 69. Thus overall, it seems that if we make a common variance assumption even when this represents a mis-specification, the results we obtain for the edge-detection problem are adequate.

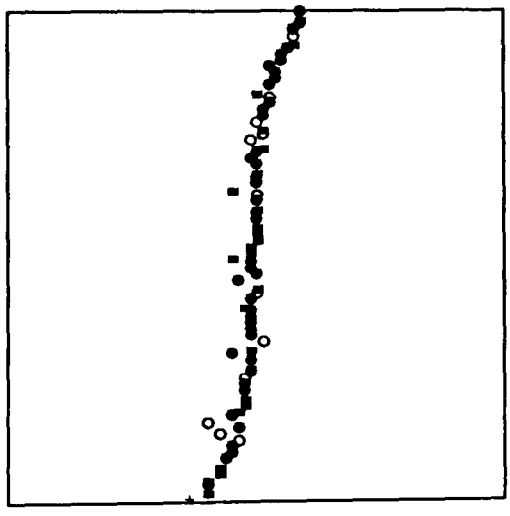


Fig 70(a) : Std. dev. change -0.5

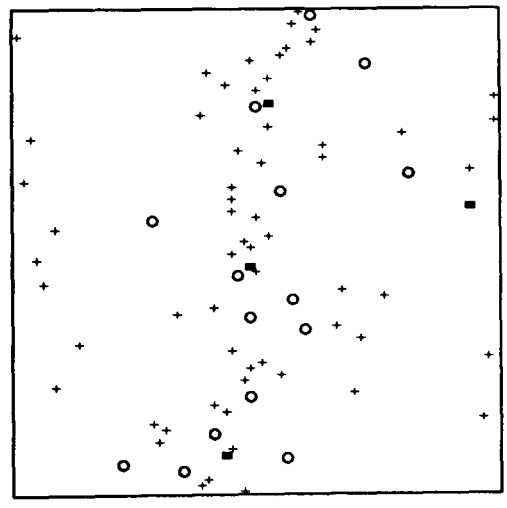


Fig 70(b) : Std. dev. change 0.5

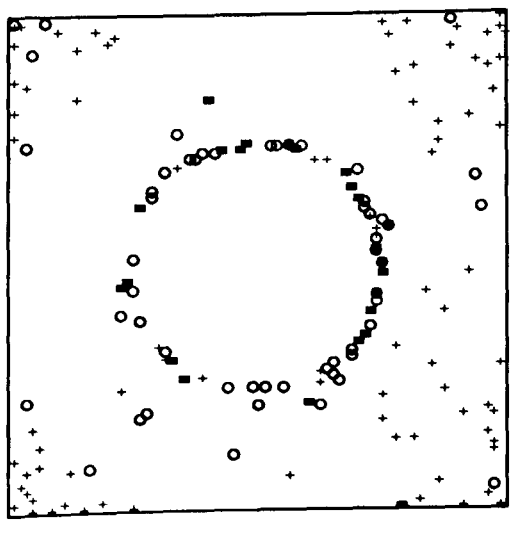


Fig 71(a) : Std. dev. change -0.5

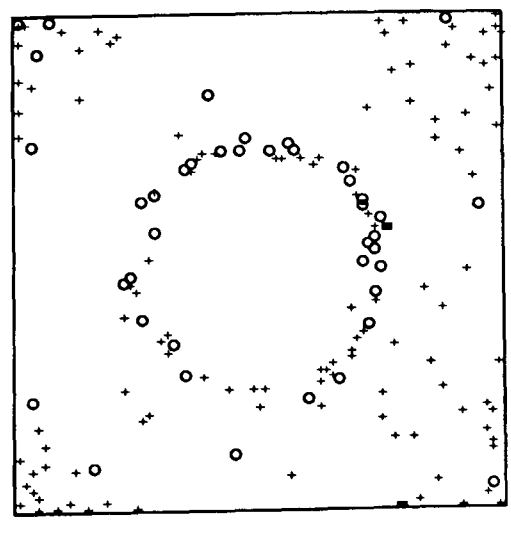


Fig 71(b) : Std. dev. change 0.5

It is possible to suggest other additive error structures (e.g. t-distributed errors) for modelling continuous data. For the modelling of discrete data, we may again suggest additive error structures applicable in the image analysis context. For example, consider a binary true scene (pixels taking values 0 or 1) as in the single edge and single convex object true scenes we have seen above, corrupted with additive binary noise, that is pixels in the true scene have their value "flipped" stochastically with constant probability p (the noise corruption level is evaluated as $100\,\% \times p$). Again, if we assume that the error terms are independent as in our previous examples, then the elements in the observed binary image are conditionally independent. This is a commonly assumed form for of image-formation process (see, for example Peli and Malah (1982), Greig *et al.* (1989)) and a binary image corrupted in such a way with varying p is depicted in figure 72. It is also clear that we may discretise the images consisting of continuous valued pixels into binary images by thresholding etc. We shall see below precisely how to evaluate the changepoint posterior distribution for use in the edge-detection problem for such images.

(5.2.2) Changepoint identification for binary sequences.

Consider the standard additive image formation model discussed above (as in (5.5)) of the form

$$Y_{ij} \;=\; \theta_{ij} \oplus \varepsilon_{ij} \tag{5.6}$$

where θ_{ij} takes the values 0 or 1, where $\oplus$ represents addition modulo 2, and where ε_{ij} takes the values 0 or 1 with probabilities $1-p$ and p respectively, p being an *a priori* unknown parameter in the model. It is clear that this describes precisely the additive binary noise model of the previous paragraph. If we again assume that the error terms are independent, the observed values Y_{ij} are conditionally independent given the true scene values θ_{ij}. Now, consider as before a single row, j say, taken from the image data. In the edge-detection problem, we wish to detect the point representing the boundary at which texture 1 and texture 2 meet, in light of the data in row j. As we have seen this is equivalent to the *a posteriori* identification of the position r of a changepoint for the data sequence under the same modelling assumptions. Now, for the model in (5.6) and the subsequent assumptions, it can be seen that the conditional distribution of data elements in row j is given by

$$\left[\, Y_{ij} \,|\, r, \theta_{ij}, p \,\right] \;\equiv\; \begin{cases} \textit{Bernoulli}(p) & i \;=\; 1, \ldots, r \\ \textit{Bernoulli}(1-p) & i \;=\; r+1, \ldots, n \end{cases} \tag{5.7}$$

where if we allow p to take values on the whole of $(0,1)$, (5.7) reflects our *a priori* indifference as to whether texture 1 "precedes" texture 2 in the underlying true scene. Suppressing the

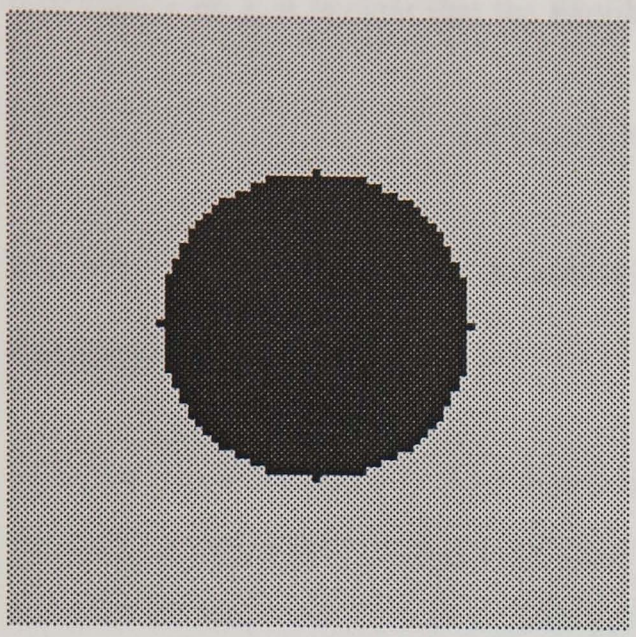


Fig 72(a) : $p = 0$

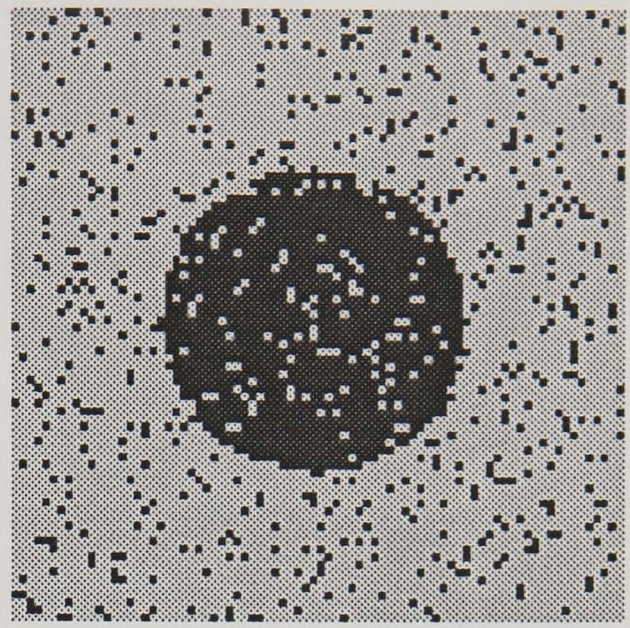


Fig 72(b) : $p = 0.1$

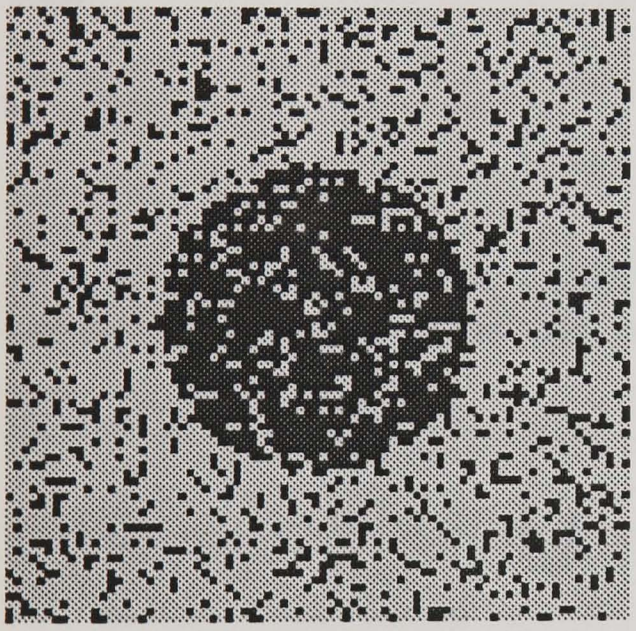


Fig 72(c) : $p = 0.2$

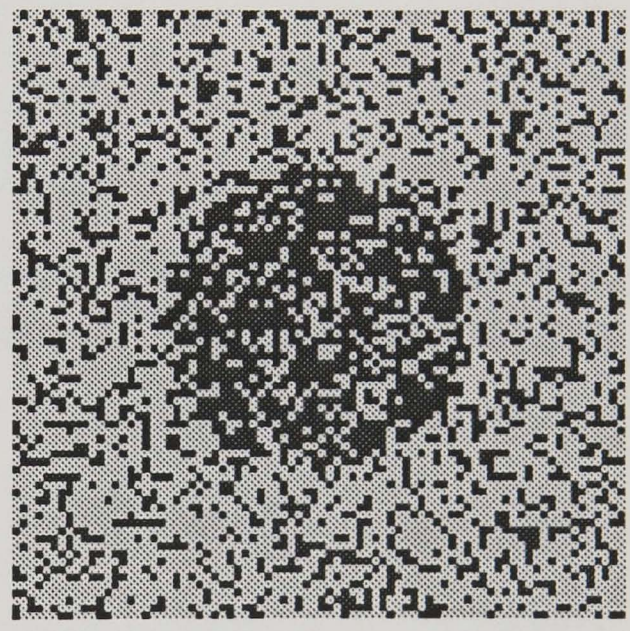


Fig 72(d) : $p = 0.3$

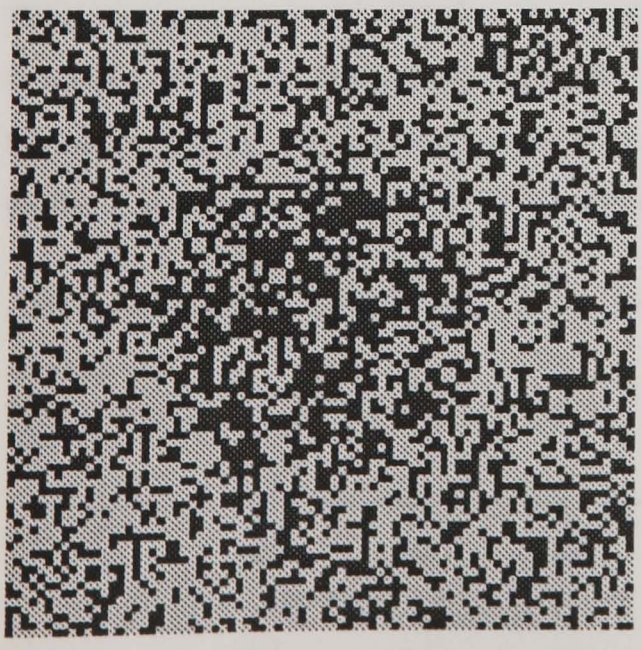


Fig 72(e) : $p = 0.4$

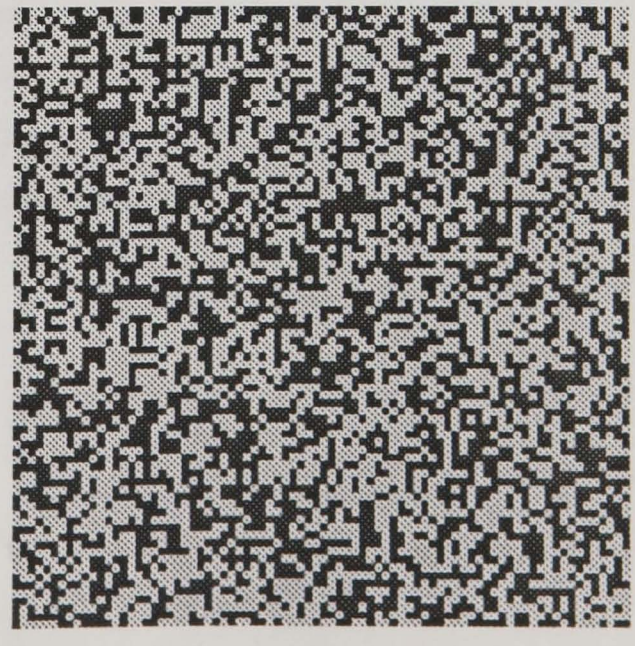


Fig 72(f) : $p = 0.5$

dependence on j, it is clear that the likelihood $[Y \mid r,p]$ is given by

$$[Y \mid r,p] = p^{n-r+S_r}(1-p)^{r-S_r} \tag{5.8}$$

where $S_r = \sum_{i=1}^{r} Y_i - \sum_{i=r+1}^{n} Y_i$. Assuming a uniform prior for r over $1,\ldots,n-1$, the posterior distribution of r is given by

$$[r \mid Y,\psi] \propto \int [Y \mid r,p]\,[p \mid \psi] \tag{5.9}$$

and thus we must specify a prior distribution for the unknown parameter p. We might choose this prior to be informative (a conjugate Beta distribution for example), or non-informative for which we have several candidates. The non-informative prior distribution we choose is of the form

$$[p \mid \psi] \propto \{p(1-p)\}^{-\frac{1}{2}} \tag{5.10}$$

for p in $(0,1)$ and zero elsewhere, which is intuitively pleasing because it is a proper prior, and can be derived using several different logical arguments. Hence from (5.8) to (5.10) we have

$$[r \mid Y,\psi] \propto \Gamma\left(n-r+S_r+\frac{1}{2}\right)\Gamma\left(r-S_r+\frac{1}{2}\right) \tag{5.11}$$

which can be evaluated using the appropriate NAG library routine.

We now study the behaviour of the posterior distribution for r in various circumstances. First, we investigate the different forms obtained when p (corresponding to Signal-Noise ratio) is varied. Figure 73 depicts the "expected" (in the sense discussed previously) posterior distributions obtained using (5.11) when p is increased from 0.1 to 0.4. The actual changepoint position was again chosen to be 24 in a sequence of length 80.

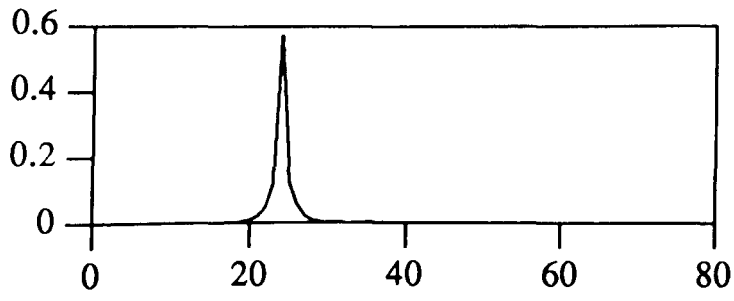


Fig 73(a) : $p = 0.1$

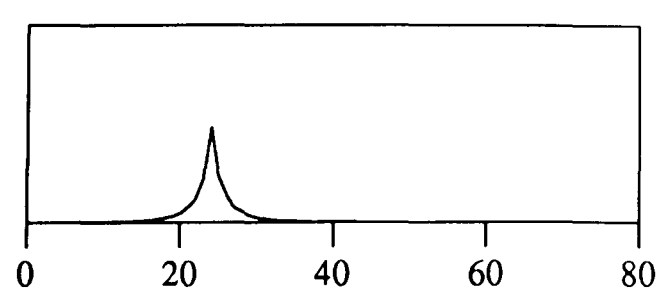


Fig 73(b) : $p = 0.2$

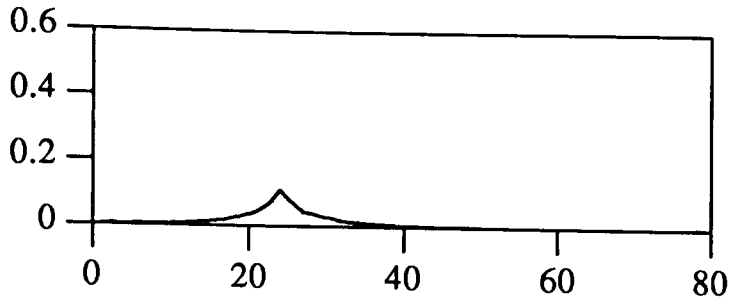


Fig 73(c) : $p = 0.3$

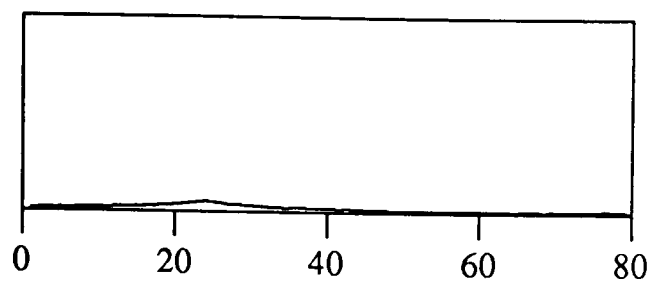


Fig 73(d) : $p = 0.4$

As we would have predicted, the posterior mode corresponds to the actual changepoint position, and the modal probability value decreases as the error rate increases. Note that, in (d), where the noise corruption is severe the changepoint is still detected as the mode of the posterior distribution, albeit barely discernible from the other values in the distribution. Clearly this behaviour will be mirrored for values of p increasing above 0.5. When $p = 0.5$ (representing a lack of any underlying structure) and $p = 0$ or 1 (representing a row consisting of one texture only) the following posterior distributions are obtained.

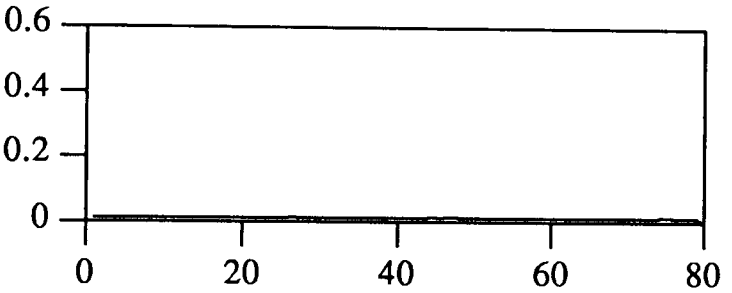


Fig 73(e) : $p = 0.5$

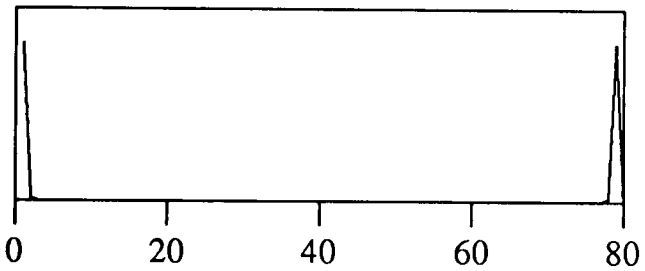


Fig 73(f) : $p = 0$ or 1

Figure 73(e) represents a uniform distribution, and (f) a distribution for which practically all of the probability lies at the ends of the sequence. Both of these results are entirely reasonable given the actual distribution of the data.

We now study the behaviour of (5.11) when the data results from a two changepoint sequence. Figure 74 depicts the expected posterior distributions obtained for the range of values of p above, where the two changepoints at 24 and 56 in the sequence.

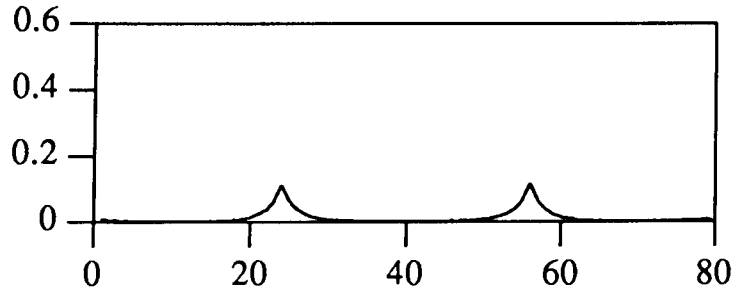


Fig 74(a) : $p = 0.1$

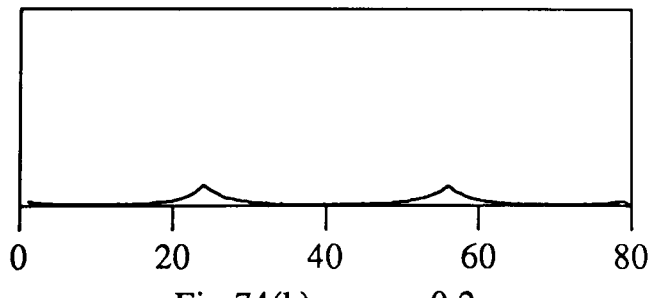


Fig 74(b) : $p = 0.2$

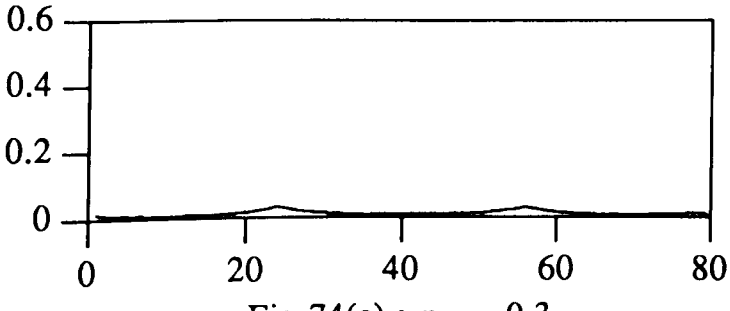


Fig 74(c) : $p = 0.3$

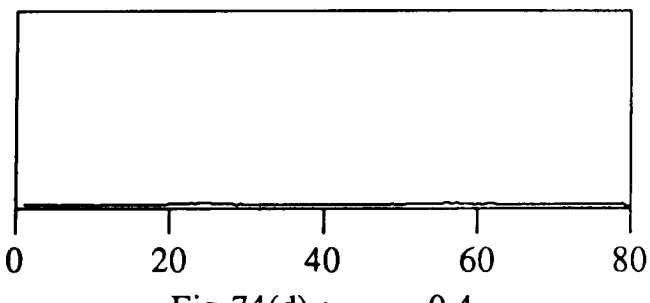


Fig 74(d) : $p = 0.4$

The results depicted in figure 74 are in line with our previous experience with symmetric sequences in that the resulting expected posterior distributions are also symmetric, but with the returned modal probabilities being less than those returned for one changepoint sequences having the same degree of noise-corruption, and with these modal probabilities decreasing as the degree of noise-corruption increased. We discover, on further experimentation, that, as before, the modal probabilities increase with inter-changepoint distance. For sequences with changepoints asymmetrically positioned, we again find that the resulting expected posterior distributions are also asymmetric, with modal probability increasing inversely with distance from the middle of the sequence. The precise expected behaviour of (5.11) under varying distributional assumptions could be investigated in the same way that we investigated (2.11) in section (3.2.2) in chapter 3. Crucially, we observe that we may associate modes in the posterior distribution with actual changepoint positions as before - this could have been predicted from the informal discussion in section (3.2.1).

We now study the results obtained from a row analysis of single edge true scene-based images with varying degrees of noise-corruption. Figure 75 depicts the results obtained using (5.11) independently on each row where the images concerned had been derived from the true scene using $p = 0.1$, 0.2 , and 0.3 in (a), (b) and (c) respectively.

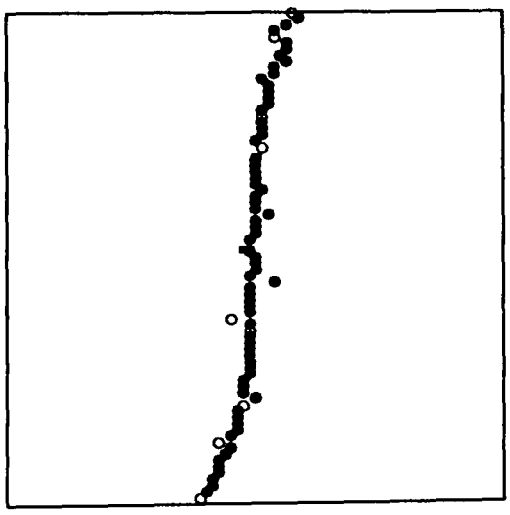


Fig 75(a) : $p = 0.1$

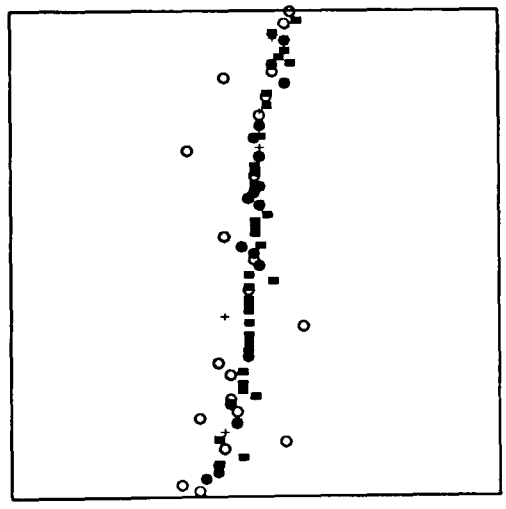


Fig 75(b) : $p = 0.2$

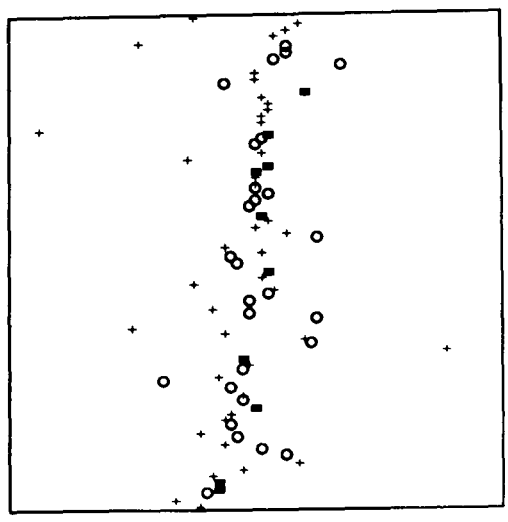


Fig 75(c) : $p = 0.3$

The changepoint technique seems to have coped adequately in each of these situations, except possibly in (c), where there appear to be quite a large number of edge mis-classifications even though the edge has been broadly detected correctly. Further experimentation showed that $p = 0.3$ proved to be an approximate upper limit on when (5.11) was effective for detecting the edge. The analyses in figure 75 required an average of 2.00 seconds processing time, approximately twice as much as the analogous analyses based on (2.11). This increase was largely due to repeated calls of NAG routine S14ABF to evaluate the (log) Gamma function terms in (5.11).

We now turn to an analysis of images derived from the circle true scene. Figure 76 depicts the results obtained from a full analysis in each case, where the same values of p are used as in the previous example.

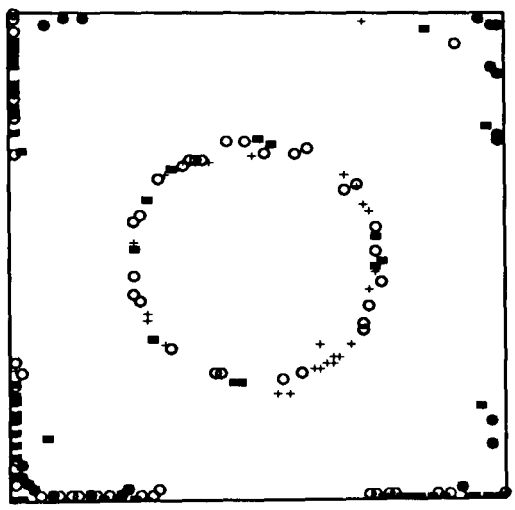


Fig 76(a) : $p = 0.1$

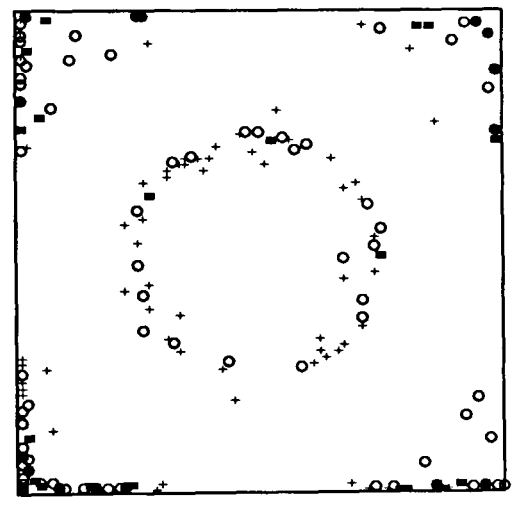


Fig 76(b) : $p = 0.2$

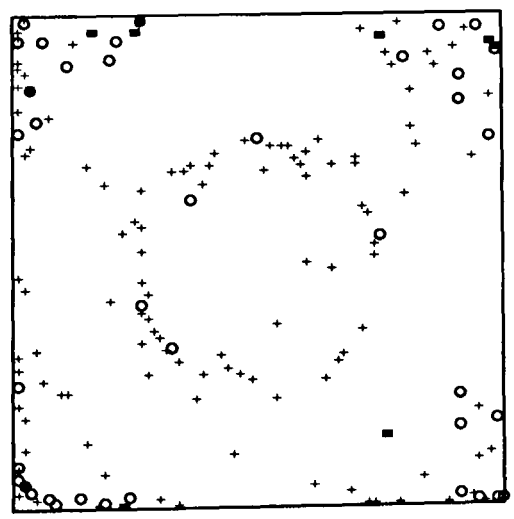


Fig 76(c) : $p = 0.3$

In each case the circle structure has been captured, with the number of edge mis-classifications increasing with p. The amount of processing time required here was 3.90 seconds for each analysis.

Analysis of asymmetric circle true scene images gave results broadly similar to those depicted in figure 17, that is, with one portion of the edge being detected only but with higher returned modal probabilities. Analysis of other (e.g. convex object) images is naturally quite straightforward. It is also possible to repeat the other developments of analytic technique that were made to the simple analysis using (2.11) (e.g. binary segmentation techniques for convex object true scenes, Gibbs Sampler-based techniques for complex true scenes, spatial ideas etc.).

Finally, we note that, in the analysis of the single edge true scene and continuous data under a normality assumption using (2.11), then for a Noise ratio $1/\sigma$ and without loss of generality the mean levels for the textures were taken to be 0 and 1, then assuming prior ignorance as to the allocation of mean levels to textures, the expected mis-classification or error rate, denoted p for reasons we shall see below is given by

$$p = 1 - \Phi\left(\frac{0.5}{\sigma}\right) \tag{5.12}$$

where $\Phi(.)$ is the unit Normal distribution function. Clearly p is the probability that a pixel would be mis-classified if the image was segmented using threshold 0.5. Thus p here can be regarded as being equivalent in some sense to the p in (5.7) to (5.10). This gives a means of comparing the efficiency of (2.11) and (5.11) on roughly equal terms. Note especially that $\sigma = 1$ corresponds roughly to $p = 0.31$, and thus we see by comparing figure 8(a) with figure 75(c) that analysis using (2.11) (and the continuum in the data) is more efficient.

This completes our study of additive error structures. We note, but do not provide examples here, that a similar approach could be adopted for multiplicative error structures, or indeed that we could regard the multiplicative noise to be acting additively after a suitable (log) transformation. However, the additive (Gaussian) noise-model is assumed in the large majority of practical applications of statistically based image analysis techniques (presumably for reasons of analytic tractability above all others). We now present one further image-formation model of a different variety to those discussed previously in this section.

(5.3) Data arising from Poisson sources.

Consider the image-formation process where the observed value at pixel (i,j), Y_{ij}, is a random variable having a Poisson distribution with expected value θ_{ij}, i.e.

$$[Y_{ij} \mid \theta_{ij}] \propto \theta_{ij}^{Y_{ij}} e^{-\theta_{ij}} \tag{5.13}$$

Such a model is applicable in the analysis of "count-based" image data, with the count arising possibly from the monitoring of radioactive materials in a medical environment. Note that it does not fit easily into the Signal * Noise form that we have seen in the earlier parts of this section, but nevertheless does represent a valid image model (in the segmentation problem we regard the θ_{ij} as unobservables about which we make *a posteriori* inferences after making *a priori* assumptions and observing data Y_{ij}). Again, consider a single row j taken from the image data, and the application of the single changepoint technique for solution of the edge-detection problem. Assuming homogeneity of textures (i.e. θ_{ij} is constant for all pixels in each texture) it is clear that

$$[Y_{ij} \mid r, \theta] \equiv \begin{cases} Poisson(\theta_1) & i = 1, \ldots, r \\ Poisson(\theta_2) & i = r+1, \ldots, n \end{cases} \tag{5.14}$$

Thus the likelihood is straightforwardly formed through (5.14). Again, we regard the pair (θ_1, θ_2) as unknown parameters for which we must specify a prior distribution. We could choose an informative conjugate (i.e. Gamma) prior for each of θ_1 and θ_2 independently - the precise details of this and the resulting posterior form are given in part two of Appendix 1. Here, we concentrate on the non-informative limit of this conjugate prior, given by

$$[\theta_1, \theta_2 \mid \psi] \propto \frac{1}{\theta_1^{\frac{1}{2}} \theta_2^{\frac{1}{2}}} \tag{5.15}$$

the choice of which can again be justified using several different logical arguments. From (5.14) and (5.15) , and assuming a uniform prior distribution for r, therefore, we obtain

$$[r \mid Y, \psi] \propto \frac{\Gamma\left(\sum_{i=1}^{r} Y_i + \frac{1}{2}\right)\Gamma\left(\sum_{i=r+1}^{n} Y_i + \frac{1}{2}\right)}{r^{\sum_{i=1}^{r} Y_i + \frac{1}{2}} (n-r)^{\sum_{i=r+1}^{n} Y_i + \frac{1}{2}}} \tag{5.16}$$

which again merely involves simple terms, and can be evaluated straightforwardly using NAG library routines S14AAF or S14ABF.

(5.3.1) Behaviour of changepoint posterior distribution - Poisson sequences.

We now study the behaviour of this posterior distribution for various values of θ_1 and θ_2, first in the one changepoint and then the two changepoint case. Figure 77 depicts the expected posterior obtained when θ_1 has value nominally fixed as 2.0, and θ_2 is allowed to vary (we shall study later if we are able to interchange θ_1 and θ_2 and obtain identical distributions). The actual changepoint position in each case was again taken to be 24 in the sequence

of length 80.

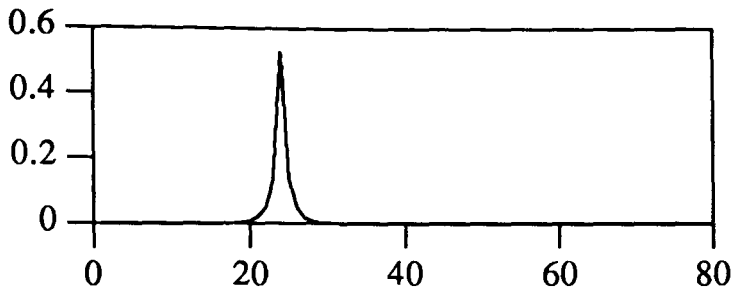


Fig 77(a) : θ_2 = 6.0

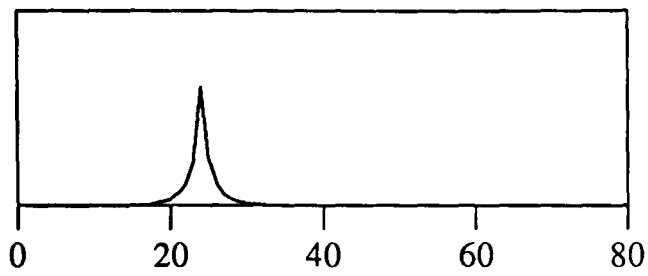


Fig 77(b) : θ_2 = 5.0

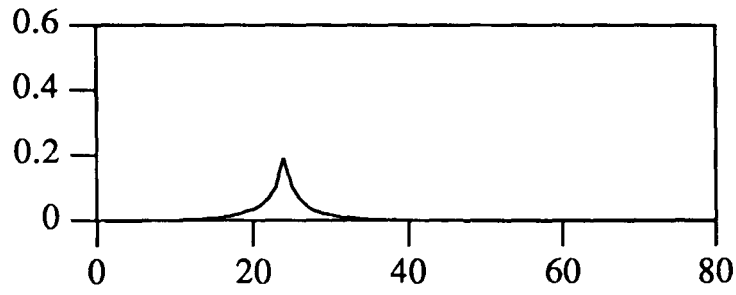


Fig 77(c) : θ_2 = 4.0

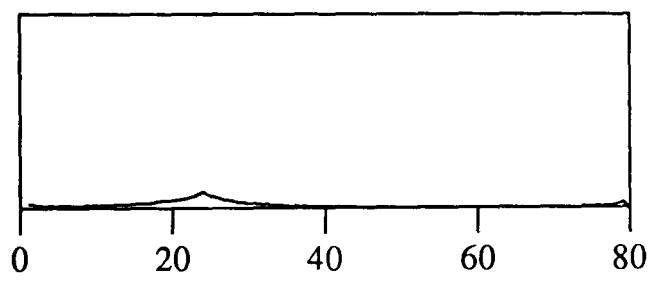


Fig 77(d) : θ_2 = 3.0

As we would expect of from our previous experience of changepoint sequences and Bayesian identification techniques, the changepoint is accurately detected in expectation, with modal probability increasing with θ_2. In this case the posterior distributions appear to be symmetric about the mode in its vicinity.

We now study the behaviour of (5.16) in the two changepoint case. Figure 78 depicts the results obtained for the symmetric changepoint case (changepoints again at 24 and 56) for the same values of θ_1 and θ_2 as above.

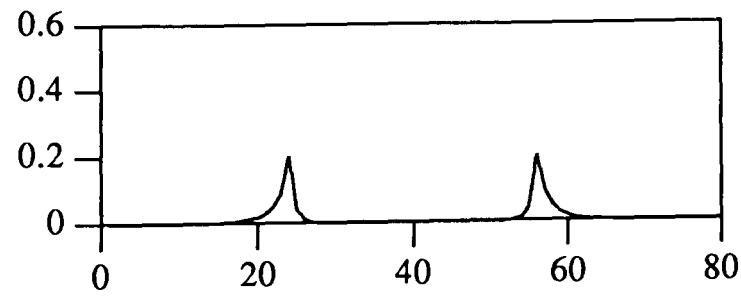


Fig 78(a) : θ_2 = 6.0

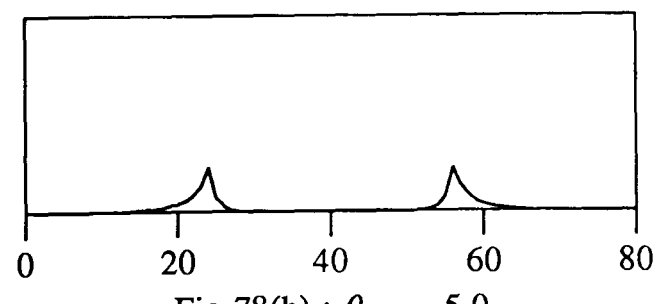


Fig 78(b) : θ_2 = 5.0

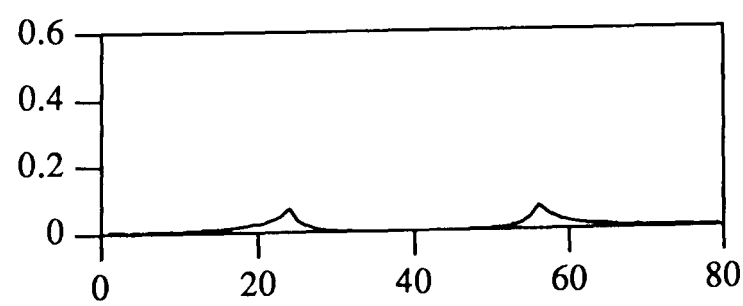


Fig 78(c) : θ_2 = 4.0

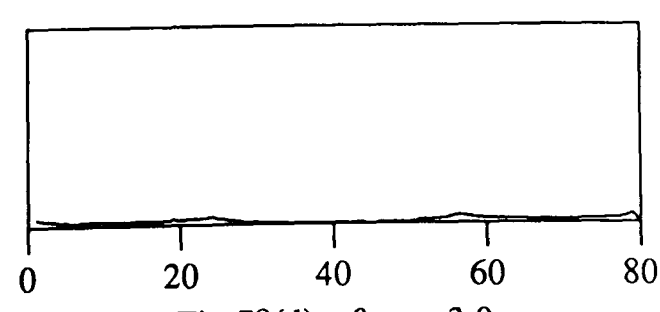


Fig 78(d) : θ_2 = 3.0

These posterior distributions have identical features to the posterior distributions we have seen previously, and the results in the asymmetric cases are broadly similar. We again omit a full study of the expected behaviour of this posterior form, but we do however note two points

apparent on further experimentation with different values of θ_1 and θ_2 in the two changepoint case. First, we discover that if we interchange θ_1 and θ_2 then the resulting posterior distribution is altered - this point is made in figure 79, with the modes in (a) taking slightly greater values than those in (b), and the distribution in (a) as a whole having a slightly smaller variance than that in (b).

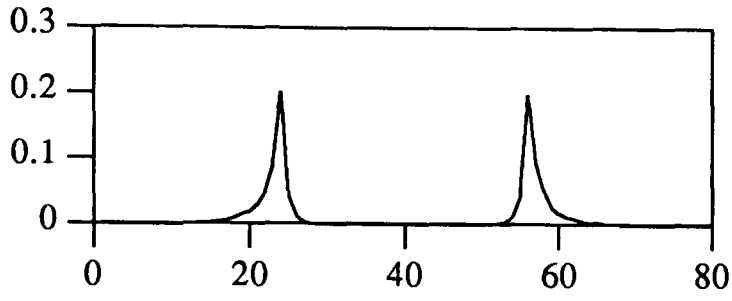


Fig 79(a) : $\theta_1 = 2.0, \theta_2 = 6.0$

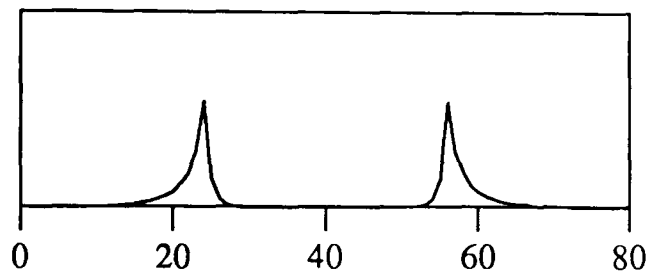


Fig 79(b) : $\theta_1 = 6.0, \theta_2 = 2.0$

The difference between the distributions is more marked when the changepoints are asymmetrically situated. Secondly, it is clear from its form that (2.11) will be invariant to changes of location in the data when the scale is fixed, i.e. the posterior distribution obtained for a sequence where the true mean levels are 0.0 and 1.0 is the same as that obtained when the true mean levels are α and $\alpha + 1$, for any value of α provided the variance remains constant. There is no such simple additive (figure 80(a), constant mean difference) or multiplicative (figure 80(b), constant ratio of means) based relationship for (5.16).

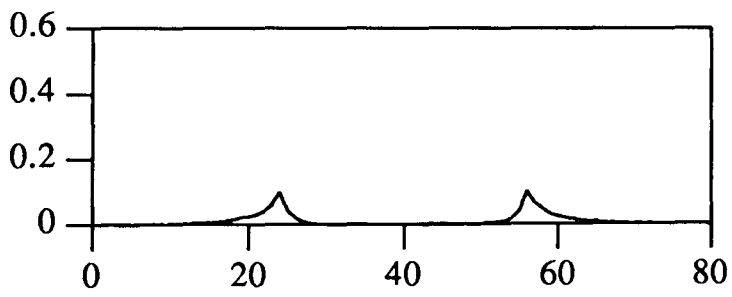


Fig 80(a) : $\theta_1 = 6.0, \theta_2 = 10.0$

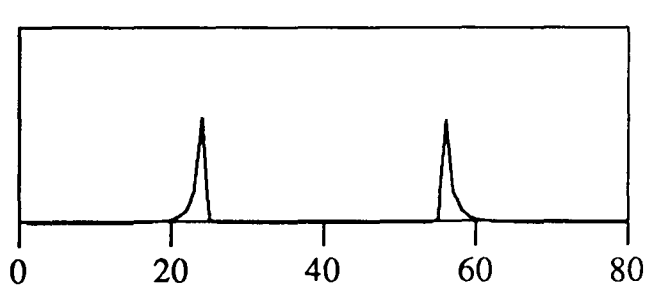


Fig 80(b) : $\theta_1 = 6.0, \theta_2 = 18.0$

Neither of the points above is surprising given first the relatively complex nature of (5.16), and also the inherent differences (e.g. variance-mean relationship) between the Poisson distributed variables and the other variables we have studied. Important though it is that we may still associate modes in the posterior distributions with actual changepoint positions in the two changepoint case, it is clear that we must overcome the difficulties that these differences bring. Previously, we could have made simple data transformations (e.g. normalisation) without affecting the posterior probabilities obtained - with Poisson distributed data this is not the case. We now attempt to solve this problem by means of another form of data transformation before proceeding to investigate the analysis of Poisson source based images.

As indicated, the complications that arose above were due principally to the mean-variance inter-relationship, i.e. for a Poisson variable Y,

$$E[Y] = \theta \quad , \quad V[Y] = \theta$$

for some $\theta > 0$. Now, it can be shown (see, for example Rao (1973), pp. 426-7) that for large θ, the variable $Z = \sqrt{Y}$ is approximately normally distributed, with

$$E[Z] = \sqrt{\theta} \quad , \quad V[Z] = 0.25$$

i.e. asymptotically, the variance of Z is independent of θ. In the changepoint/edge-detection context, therefore, we might expect to be able to make the square-root transformation of the data to normality in order to avoid some of the difficulties mentioned above. Also, we might expect the resulting posterior forms to be more straightforward than (5.16) - recall that in the changepoint problem, we regard the mean levels θ_1 and θ_2 as unknown (nuisance) parameters which we can remove by integration over suitable prior measures, and in the normal case this generally produces attractive changepoint posterior distributions.

(5.3.2) Square-root transformation of Poisson data.

We now study how making the square-root transformation on the data effects the behaviour of our Bayesian changepoint identification techniques. First, we derive the one changepoint posterior distribution in the case equivalent to (2.1) and the subsequent derivation of (2.11) when the error-term variance is known. Consider a sequence of normally distributed variables Y assumed to have a single changepoint at an *a priori* unknown position, with the two subsequences having *a priori* unknown mean levels θ_1 and θ_2 and known variance σ^2 (see in particular section 1.1 of Appendix 1). In order to evaluate the posterior distribution for changepoint position r, we must specify a prior distribution for the pair (θ_1, θ_2). Several such prior forms are suggested in section 1.1, Appendix 1. The precise form we choose here is the non-informative limit of independent conjugate normal priors for each parameter, described in 1.1.2 of the appendix, this reducing merely to the uniform measure on $(-\infty, \infty)$ (clearly an improper prior, so that we must take care if we wish to incorporate a no changepoint alternative into the formulation as before). Using standard techniques, and substituting $\tau = \frac{1}{\sigma^2}$, we obtain

$$[r \mid Y, \psi] \propto \{r(n-r)\}^{-1/2} \exp\left\{ -\frac{\tau}{2} \left[\sum_{i=1}^{r} (Y_i - \overline{Y_A})^2 + \sum_{i=r+1}^{n} (Y_i - \overline{Y_B})^2 \right] \right\} \qquad (5.17)$$

where again $\overline{Y_A} = \frac{1}{r} \sum_{i=1}^{r} Y_i$ and $\overline{Y_B} = \frac{1}{(n-r)} \sum_{i=r+1}^{n} Y_i$.

We note immediately from the form of (5.17) the presence of the sum of squares term in the exponent, and hence that for large n the qualitative behaviour of this posterior distribution in expectation with respect to modal positions will be identical to that which we considered in section (3.2.2) of chapter 3 (i.e. that the posterior distribution will be maximised when the sum of squares is minimised, and two localised modes will be obtained if the distribution of Y is altered to that of a two changepoint sequence). We note also that, as for (2.11), the evaluation of (5.17) involves merely simple functions of sufficient statistics, and thus should be relatively straightforward. Both of these points are positive aspects when we consider the relative merits of evaluating (5.16) Y is Poisson and evaluating (5.17) when Y is subjected to a square-root transformation to approximate normality as described above, as we saw that the form of (5.16) did not lend itself easily to an analytic study of its behaviour, and that its evaluation required the evaluation of special functions via NAG library routines.

Before studying the adequacy of the normal approximation specifically in the changepoint case, we first investigate the behaviour of (5.17) when the data Y is actually exactly normally distributed. Without loss of generality, we consider the case $\sigma^2 = \tau = 1.0$, and various choices for θ_1 and θ_2, as it is clear from its form that posterior distributions obtained via (5.17) will be invariant to location changes in the distribution of Y. This is a very important feature enabling us to evaluate the changes in the posterior distribution obtained when the Signal-Noise ratio (or in this case the difference in means) is changed.

In the one changepoint case, the posterior distributions obtained are as we would have predicted, with the mode corresponding to the actual changepoint position. The behaviour over a range of values of Signal-Noise ratio was also similar to that for (2.11). Both of these features are entirely reasonable given our experience with Bayesian changepoint techniques, and more specifically the study we made of the properties of sums of squares of normally distributed variables in section (3.2.2). In the two changepoint case, we again discover that the posterior distributions obtained via (5.17) are frequently bimodal, with the modes corresponding to the actual changepoint positions. In addition, all the features concerning symmetry and asymmetry that we have noted previously when we have used one changepoint posterior distributions to make approximate inferences for two changepoint sequences were present in the results obtained using (5.17). We also note that in all of these results we were able to interchange mean levels θ_1 and θ_2 and obtain identical posterior distributions.

It is interesting to briefly compare the results obtained in the two changepoint case when using (2.11) and (5.17). For demonstration purposes, we choose an asymmetric sequence (changepoints at 24 and 65) where the mean change is large (2.0). Figure 81 depicts the two expected posterior distributions obtained.

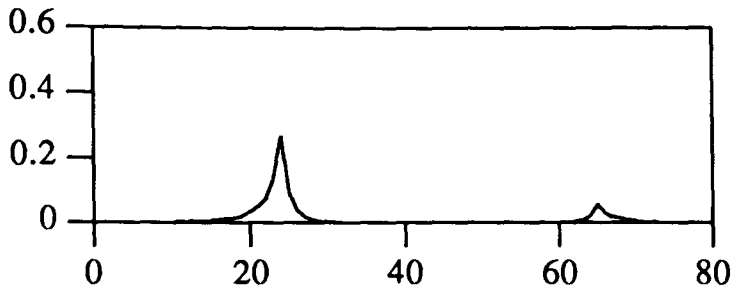


Fig 81(a) : (2.11)

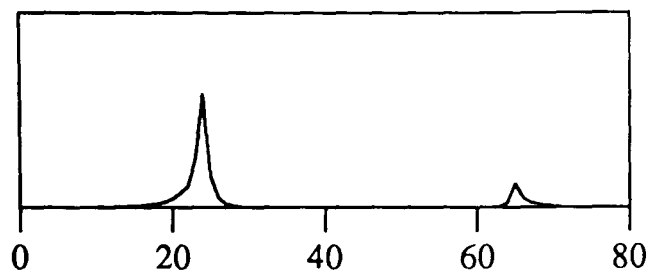


Fig 81(b) : (5.17)

The modal probabilities for (5.17) are greater than those for (2.11). This is reasonable, as clearly we instilled a greater degree of structure and a lesser degree of ignorance using our prior knowledge of σ^2. This difference in behaviour could of course be deduced from the functional forms of (2.11) and (5.17).

We now investigate the consequences of making the square-root transformation to approximate normality for Poisson sequences in the changepoint context. First, recall that in the derivation of (5.17) we specified a non-informative prior distribution for each of the mean levels. We note that the particular prior that we have used corresponds exactly to a non-informative prior specification for the mean levels in the Poisson distributed case (i.e. ignorance of θ is equivalent to ignorance of $\sqrt{\theta}$). Had we had some quantitative prior knowledge of the mean levels in the Poisson case, then we would have had to taken a degree of care over the choice of prior distributions after the transformation to approximate normality.

Consider then the transformation of Poisson sequences to approximate Normal sequences, and the subsequent posterior distributions obtained. Consider first the one changepoint case. Figure 82 depicts the posterior distributions obtained for a typical (and representative) sequence where the changepoint was at 30, and the mean change was from 8 to 12 in the original Poisson sequence. Figure 82(a) depicts the expected posterior distribution when using (5.16), (b) when using (5.17).

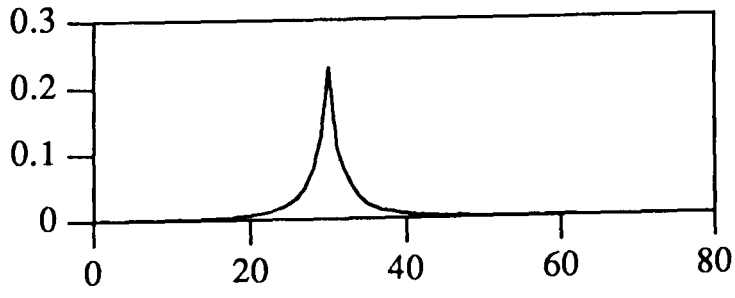


Fig 82(a) : (5.16)

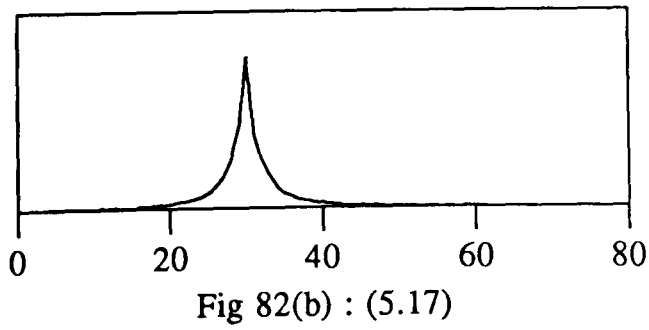


Fig 82(b) : (5.17)

The two distributions are indistinguishable. Thus it seems that in the one changepoint case the square-root transformation to normality does not have a disruptive effect on the resulting posterior distributions. However, we note again that interchanging θ_1 and θ_2 does alter the resulting modal values to some small degree, and thus we now acknowledge this as a general and slightly negative feature of of our changepoint detection methods (and consequently whenever

we refer to θ_1 and θ_2 from here onwards, we assume that $\theta_1 < \theta_2$). However, when non-informative priors for the mean levels are used the overall difference between the posterior distributions is negligible, with the posterior modal positions unchanged (presumably why this feature did not seem especially relevant in our previous analyses). Thus this feature is of limited importance for the context in which we work. Note that this feature is of course not present for the additive binary noise case we studied in the derivation of (5.11), where the distributions of all the data variables were controlled by a single parameter p.

Consider now this transformation of Poisson variables in the two changepoint case. Figure 83 depicts the resulting posterior distributions calculated via (5.16) and (5.17) for another typical and representative sequence where the actual changepoints were at 20 and 50, and again the mean change was from 8 to 12 in the original Poisson sequence.

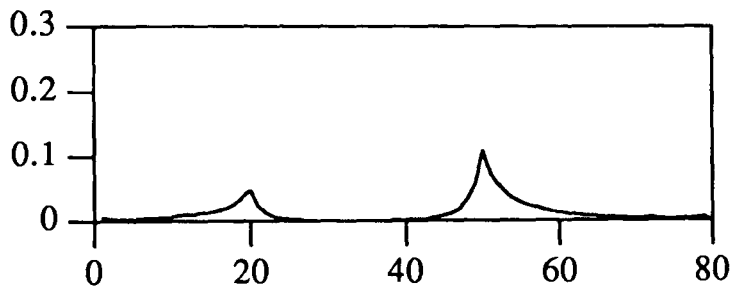


Fig 83(a) : (5.16)

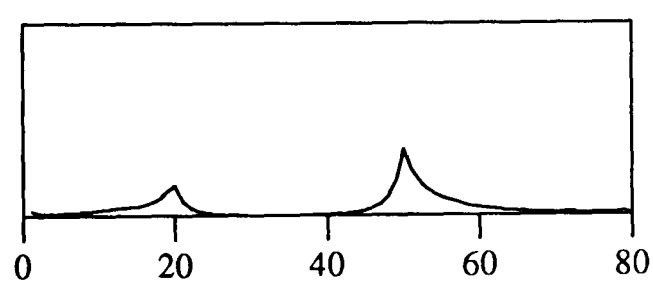


Fig 83(b) : (5.17)

Again, the distributions are virtually indistinguishable. Thus we feel reasonably satisfied that the Normal approximation that we have made is adequate for the derivation of changepoint posterior distributions.

In light of the results that we have seen above, we might now feel confident in using the square-root transformation of the Poisson data sequences to approximate normality for two reasons. First, the resulting posterior form (5.17) is more easily computed than (5.16), not requiring the evaluation of special functions, and hence would induce a saving in computation time in the context of the edge-detection routines we have developed. Secondly, the functional form of (5.17) is more analytically tractable than that of (5.16), and hence we may apply all of the theory we investigated earlier. Finally, and perhaps most importantly, the form of the transformation gives us insight into how altering the difference in mean levels in the Poisson sequence (analogous to altering Signal-Noise ratio for Normal sequences) precisely effects the resulting changepoint posterior distributions - whereas before we saw that in the Normal case, (2.11) was invariant to location changes that preserved the difference in mean levels, we now realise that in the Poisson case, (5.16) is (approximately) invariant to location changes that preserve the difference in the square-roots of the mean levels. This is confirmed by the posterior distributions depicted in figure 84, the two changepoint sequence having changepoints at 24 and 65, and where the mean levels were changed in such a fashion, with the difference in their square-roots being kept constant at 0.712788.

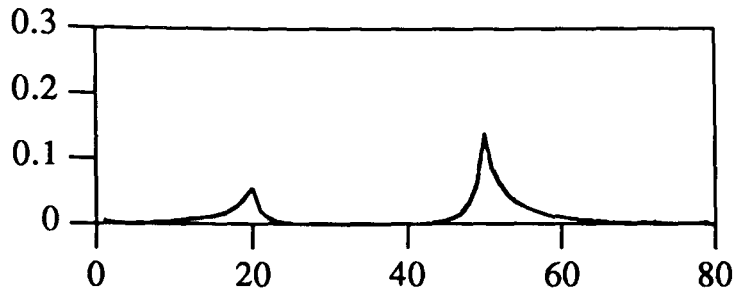


Fig 84(a) : $\theta_1 = 6.0$, $\theta_2 = 10.0$

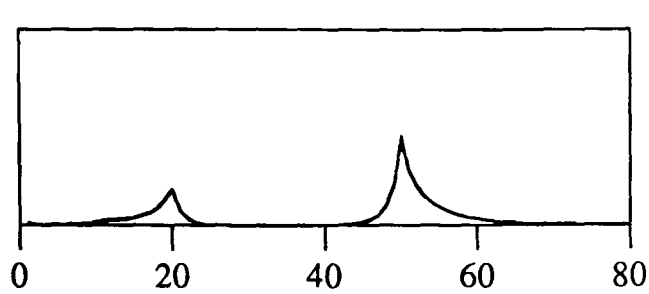


Fig 84(b) : $\theta_1 = 10.0$, $\theta_2 = 15.0161$

Again, these distributions are virtually indistinguishable, and our thoughts above are confirmed. The results are repeated if the square-root transformation is made prior to evaluation of the posterior distribution. The distributions for the transformed variables are shown on figure 84 as dashed lines, but are practically coincident with the solid lines representing the original distributions at every point.

Finally in this section we study the use of changepoint based edge-detection techniques when the image data arises from Poisson sources.

First, we consider the analysis of single edge true scene-based images. Figure 85 depicts the results obtained using (5.16) on un-transformed Poisson distributed data where θ_1 was fixed equal to 4.0, and θ_2 was allowed to vary.

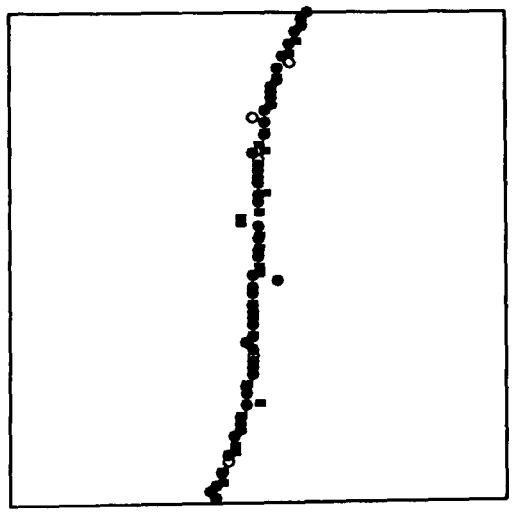


Fig 85(a) : $\theta_2 = 10.0$

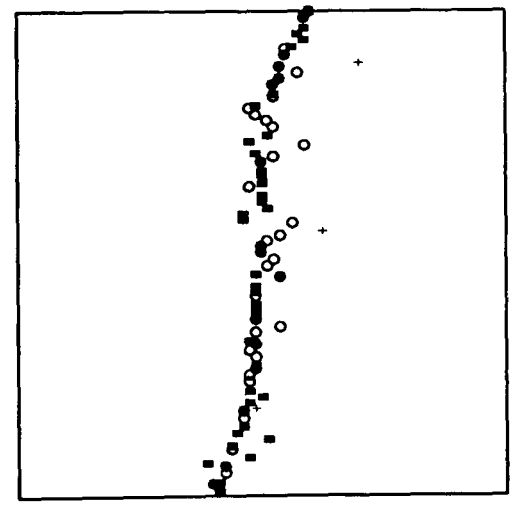


Fig 85(b) : $\theta_2 = 8.0$

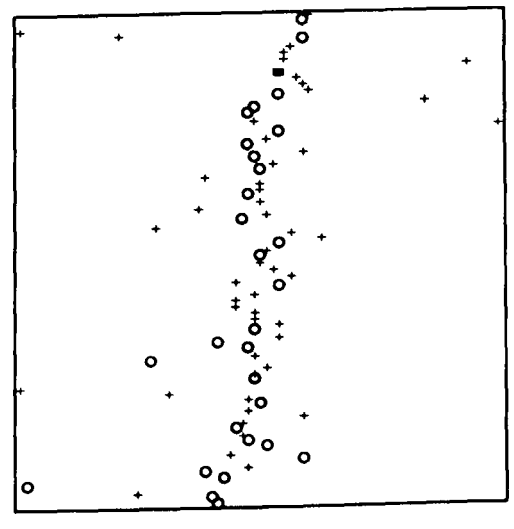


Fig 85(c) : $\theta_2 = 6.0$

Again, the changepoint technique has coped adequately in each of these cases. The analyses involved in the production of the results in figure 85 required an average of 2.48 seconds of processing time. Recall that the analyses based on (2.11) took of the order of one second - the increase here is again due to the repeated calls of NAG routine S14ABF to evaluate the log Gamma function needed in the evaluation of (5.16).

Now we consider an analysis of the data after it has been transformed using a square-root transformation. Note also that after this initial transformation, we make make a further location transformation of each data element by some constant, taken here to be the data mean over all the image, and still preserve the effective Signal-Noise ratio. This is sometimes necessary purely for computational convenience. Note also, that the differences in mean level in figure 85(a), (b), and (c) correspond approximately to Signal-Noise ratios of 2.34, 1.65, and 0.89 respectively in the Normal approximation, given a known variance of 0.25. This gives us some perspective of the performance of the changepoint techniques under these distributional assumptions relative to the performance of, for instance, (2.11). Figure 86 depicts the results obtained using (5.17) on the transformed data.

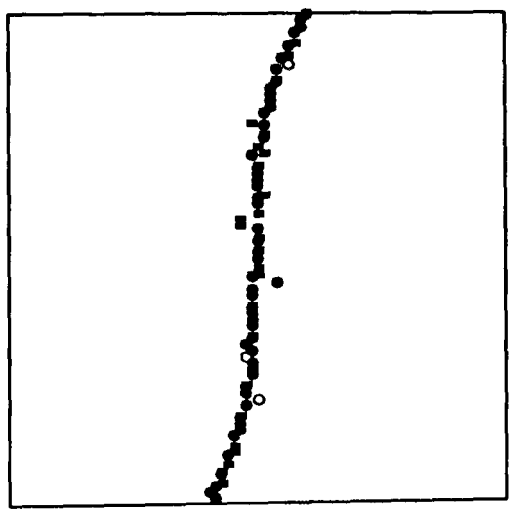


Fig 86(a) : θ_2 = 10.0

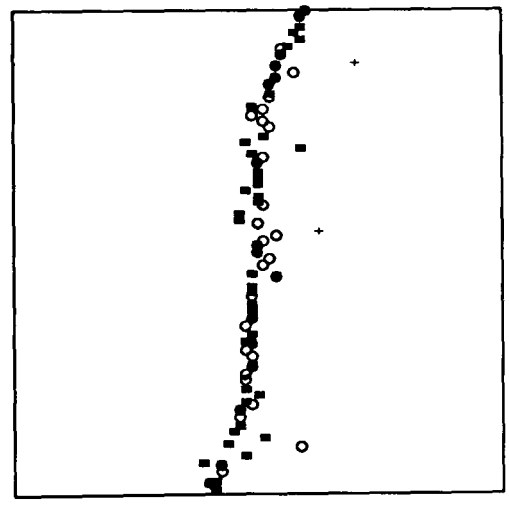


Fig 86(b) : θ_2 = 8.0

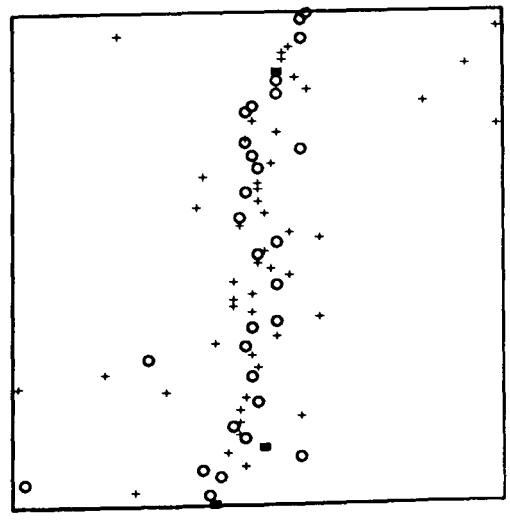


Fig 86(c) : θ_2 = 6.0

These results are virtually indistinguishable from those in figure 85, re-enforcing our faith in the Normal approximation for Poisson distributed data. Importantly, however, the analyses of these images using (5.17) required an average of only 1.18 seconds of processing time. This represents a saving of over a half compared to the analysis of the identical data set, which we must regard as important given our primary objective of efficient and accurate processing. Finally, we consider the analysis of circle true scenes and images produced under the assumption of Poisson sources. Figure 87 depicts the results obtained using (5.16) on the untransformed data, where again θ_1 was fixed at 4.0 and θ_2 varied.

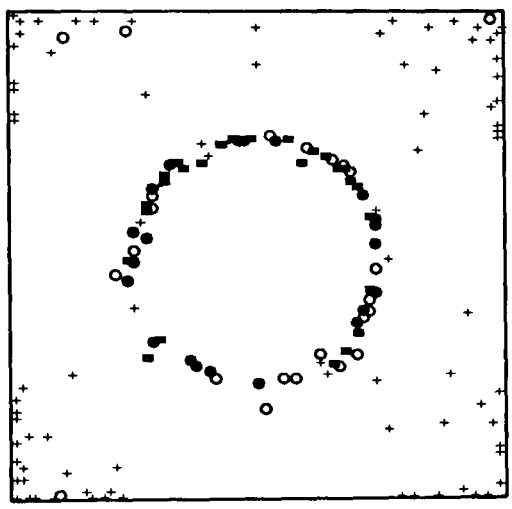


Fig 87(a) : θ_2 = 10.0

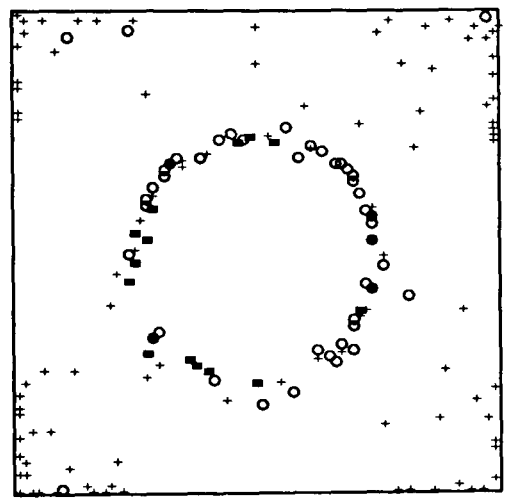


Fig 87(b) : θ_2 = 8.0

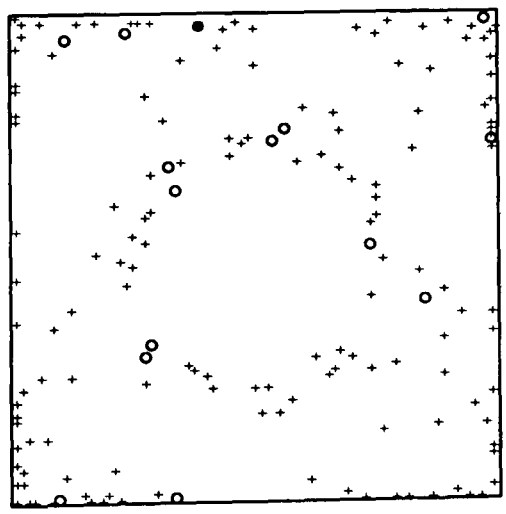


Fig 87(c) : θ_2 = 6.0

The circle structure is captured satisfactorily in (a) and (b), but somewhat less so in (c). However, we can clearly perceive the central region of homogeneity, and so the results overall seem to be adequate. The analysis required an average of 4.56 seconds in each case. Now we consider an analysis using (5.17) on the square-root transformed data, again with the shift in location by the overall data mean implemented. The results are depicted in figure 88.

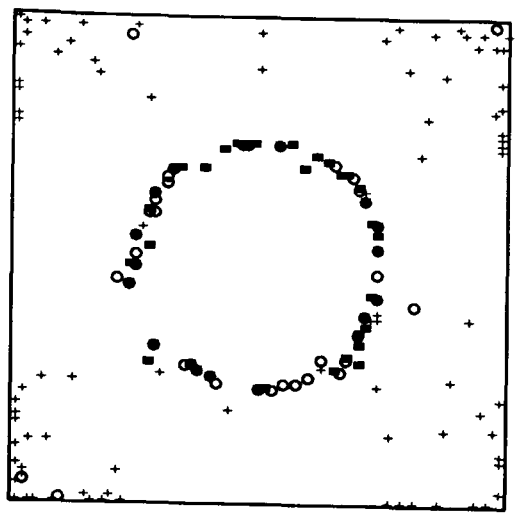


Fig 88(a) : $\theta_2 = 10.0$

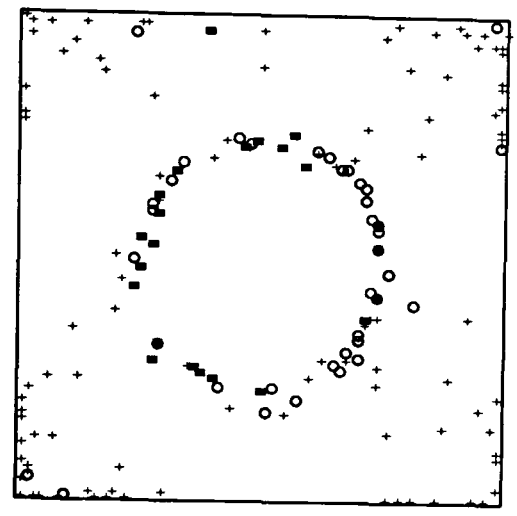


Fig 88(b) : $\theta_2 = 8.0$

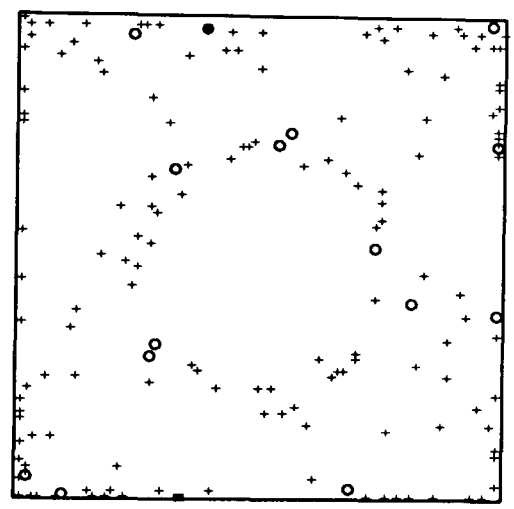


Fig 88(c) : $\theta_2 = 6.0$

Again the two sets of results are indistinguishable. The analysis of this data using (5.17) required an average of 2.30 seconds of processing time, again a saving of around a half compared to the analysis of the un-transformed data.

Analysis of asymmetric true scene images gave results that were in line with what we have seen previously for (2.11) and (5.11) with one portion of the edge being detected. Again, we may develop further techniques for edge-detection based on (5.16) or (5.17) in the same way that we did previously to cope with convex object and more complex true scenes.

(3.4) Variation of image-formation process - conclusions.

We have seen in this section how the changepoint posterior distributions used in the edge-detection techniques we have developed can be adjusted to cope with a range of image-formation models. First, in the additive Gaussian noise case, we saw how only relatively minor adjustments to our initial formulation of the edge-detection problem (two textures, different mean levels but common variance) were required to deal with slightly different image

models (i.e. different texture means, variances). We also studied the results obtained when the changepoint posterior distribution was derived under an incorrect model specification, and saw that the mis-specified model often produced adequate results. Next we studied a different additive noise model, namely that where a binary image is corrupted by additive binary noise. We derived a changepoint posterior distribution in the usual way, (5.11), studied its behaviour in expectation for various degrees of corruption, and then studied its adequacy as a solution of the edge-detection problem. Principally, we saw that it gave acceptable results, but required larger amounts of processing time than the techniques previously studied. Also, we deduced that it was generally less efficient than techniques based on an underlying normality assumption. Finally, we studied an image-formation process which was different in nature from any previously mentioned, namely where the data arise from Poisson sources. We derived a changepoint posterior distribution, (5.16), and studied its behaviour in expectation, but saw that its behaviour for over a range of texture means was not predictable, and that its evaluation involved evaluation of complex functions. For these reasons, we discussed the effect of a square-root transformation of the data to approximate normality. We saw that the equivalent changepoint problem for the transformed data involved the derivation of a posterior distribution under a known variance condition. The behaviour of this posterior distribution, (5.17), was then studied in expectation, and its adequacy as some sort of approximation to (5.16) was found to be satisfactory. The nature of the transformation also gave insight into the behaviour of (5.16) for different values of the underlying mean levels. Finally, (5.16) and (5.17) were used as the bases of edge-detection techniques for un-transformed and transformed Poisson data images respectively. The results obtained were found to be comparable and adequate in each case, but it was discovered that the processing time required when using (5.17) was only half that required when using (5.16). Thus we concluded that the transformation of Poisson data to approximate normality was advisable in every sense.

Chapter 6 : Edge–Reconstruction and Object Detection.

In the preceding chapters, we have seen how, given an observed image derived from an *a priori* unknown true scene, Bayesian changepoint analytic techniques may be used as the basis of a solution to the important problem of edge-detection. Using such methods, we have obtained sets of candidate edge-points and associated probabilities as modal positions and values in changepoint posterior distributions. Using ideas about local spatial relationships and our interpretation of the posterior probabilities themselves, we have removed some of the edge-points, regarding them as mis-classifications. Thus the first stage of our processing of the image is complete. We now concentrate on the subsequent stages of processing, and attempt to incorporate the simple techniques that we have developed into the more complex and sophisticated techniques that are necessary for successful solution of the problems that arise in the area of the statistical analysis of images. First, we discuss our primary objectives in this area.

The approach that we take to image analysis problems will be, of course, ultimately dependent on the underlying decision problem. For instance, in the segmentation problem, we wish to classify pixels into texture classes on the basis of a limited amount of data (commonly one observation per unknown parameter) and certain qualitative (and possibly quantitative) prior knowledge of the true scene. This is a relatively straightforward statistical decision problem to formulate, but a relatively difficult one to solve practically, and thus we attempt to simplify the problem using what we would regard in this context as pre-processing techniques such as edge-detection. However, if we merely wish to report the positions of regions of discontinuity in the image, or the position, dimensions and orientation of objects in the image, then the underlying decision problem, and hence our approach to its solution, is quite different. We shall see in our final chapter how our changepoint techniques can be used directly to aid in the solution of segmentation-type problems. In this chapter we concentrate on two related and straightforward problems (again perhaps to be regarded as preparatory techniques to be carried out prior to segmentation), both of which follow naturally from the changepoint based edge-detection results that we have already obtained. First, we investigate simple edge-reconstruction techniques, that is, given a set of edge-points we seek routines that will represent the edge by means of a smooth curve in some coherent manner, either for visual enhancement or for analytic purposes. We discuss both parametric and non-parametric techniques. Secondly, we develop object detection routines so that, given a set of edge-points, we may draw inferences concerning the location, dimensions and orientation of objects in the true scene. We examine particularly the case of objects with edges having a certain parametric form, namely that of an ellipse. We also study the related "tank-spotting" problem, where we wish to discover the location of a (possibly unknown) number of small objects in the true

scene, and where the true scene itself changes with time but the object positions in successive "frames" are closely related. Again, throughout all of our discussion, we must reflect the need for time-efficient processing, as, at this stage, we may still be able to trade this against the need for a high level of accuracy.

(6.1) Edge-reconstruction.

We now seek to develop techniques that will allow texture region boundaries and object edges to be reconstructed from the set of edge-point candidates obtained via the changepoint based detection routines that we have seen in previous sections. We attempt to achieve reconstruction by implementation of simple and therefore hopefully time-efficient statistical estimation schemes, and hence represent the edge as a smooth curve in the plane. We mention both parametric (e.g. least-squares etc.) and non-parametric (e.g. spline-based) estimation methods and compare their relative merits. First, we introduce some necessary notation.

Let $E_S = \{e_i : i = 1, \ldots, P\}$ denote the set of P pixels in S that have been detected as edge-points in a changepoint analysis of the image, and let the point e_i have (real) coordinates (x_i, y_i) in a coordinate system where the axes are parallel to the edges of the rectangular region S. For the moment, we treat each of the elements of E_S equally, making no reference to either their associated posterior probability or the direction in which they were detected as edge-points (i.e. in the row or column concerned). We note the relevance of each of these points at a later stage. Also, we regard the elements of E_S as having already been post-processed to some degree, possibly using spatial or probabilistic ideas to exclude clear misclassifications. Despite this, P is still generally rather large (of the order of the number of rows and columns of the image). We proceed to consider various possible techniques for reconstructing an edge or edges from the set E_S.

We first concentrate on the representation of a single edge in the true scene as a smooth parametric curve. Our first approach is to fit a simple linear statistical model to the location data (the coordinates of the elements of E_S).

(6.1.1) Single edge representation via polynomial regression.

Consider a simple polynomial regression model for coordinate variables (X, Y) in the coordinate system having axes parallel to the edges of region S of, for definiteness, Y on X, of dimension $k+1$, i.e. where

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \ \ldots \ + \alpha_k X^k + \varepsilon \qquad (6.1)$$

where ε is regarded as an error term. The problem of representing the edge now merely reduces to choosing the parameters $\alpha = (\alpha_0, \ldots, \alpha_k)$ (and indeed k itself) on the basis of the

set of (observed) coordinates (x_i, y_i) of the elements of E_S, and some reasonable and sensible criterion.

Clearly, if we regard ε as some random quantity, then (6.1) is merely a simple form of linear statistical model for the location data, which has been extensively studied, and thus we are able to select from a wide range of techniques and criteria in order to solve the edge-representation problem (for instance, for a Bayesian formulation of the analysis of the linear model, we mention Lindley and Smith (1972) as an important reference). For our purposes, however, given our constraints on processing time, we seek a simple formulation that leads to an easily implemented method of solution. One such intuitively plausible solution would involve choosing α to minimise the distance D_y, where

$$D_y = \sum_{i=1}^{K} (y_i - \hat{y}_i)^2 \tag{6.2}$$

and $\hat{y}_i$ is the fitted y coordinate obtained from the model in (6.1) and x_i for a particular choice of α. Let $d_{y_i}^2 = (y_i - \hat{y}_i)^2$. Viewing (6.2) purely as a distance measure, it is clear that choosing α in this way (using a "least-squares" type criterion) is intuitively reasonable . Also, it is clear that it is equivalent to using a maximum-likelihood criterion under assumption that the error terms in (6.1) are independent and identically normally distributed. This latter point is somewhat less appealing in these circumstances than that purely of a distance based criterion, but perhaps more statistically satisfactory - recall that in this context, the standard maximum-likelihood estimates are identical to Bayesian estimates resulting from a specific (non-informative) choice of prior distribution. We also note several other important features. First, in place of, or in addition to, (6.1), we might consider the equivalent polynomial regression model of X on Y given by

$$X = \beta_0 + \beta_1 Y + \beta_2 Y^2 + \quad \ldots \quad + \beta_k Y^k + \varepsilon \tag{6.3}$$

where again $\beta = (\beta_0, \ldots, \beta_k)$ are regarded as *a priori* unknown parameters, and choose those parameters to minimise D_x, where

$$D_x = \sum_{i=1}^{K} (x_i - \hat{x}_i)^2 \tag{6.4}$$

and, as above, we write $d_{x_i}^2 = (x_i - \hat{x}_i)^2$. Clearly, we would expect (6.1) to provide a better fit for broadly "horizontal" edges and (6.3) a better fit for "vertical" edges in the true scene. Practically, we might fit both (6.1) and (6.3), and assess relative goodness of fit between the

two with respect to the relative magnitudes of the minimum values of D_y and D_x attained. Secondly, we note that the minimisations of D_y and D_x are readily available, using for instance NAG library routine F04ANF and associated routines, and that generally these minimisations are not computationally costly even when k is relatively large. Thus straightforwardly we may fit polynomials of high order to obtain an effective representation of the edge. Thirdly, we might also consider choosing the coefficient parameters of the model by minimisation of the distance measured as the shortest straight line distance between the fitted and observed points (a hybrid of (6.2) and (6.4)). However, the formulation of this type of solution is less straightforward and more complex to implement. Finally, as we noted above, there is an equivalence between this type of least-squares type distance criterion and maximum-likelihood estimation under a normality assumption. We may thus exploit this relationship to obtain measures of uncertainty (standard errors etc.) if so required, and to introduce other aspects (weighting of points, sensitivity considerations etc.) into the formulation. We study some of these points and other necessary modifications in a later section.

We now investigate the reconstructions obtained using the simple polynomial regression technique described above, where the set E_S represents a typical set of results obtained by changepoint analysis of an image derived from various single edge true scenes. First, recall the familiar single edge of figure 3 in chapter 2. We have seen in figures 8 and 9 how the results obtained using changepoint based techniques vary with Signal-Noise ratio. On the basis of the results depicted there, now study the various smooth curve representations obtained using (6.1) and (6.3) above. In each case, for demonstration purposes we attempt to fit cubic, quartic and quintic polynomials to the data and note the relative residual sums of squares (6.2) and (6.4) obtained - we can compare these directly as a measure of relative (but not absolute) goodness of fit. Figure 89 depicts the edge-reconstructions obtained after an initial row analysis of images derived from the single edge true scene and varying degrees of noise-corruption. In each case, the solid, dashed, and dotted lines represent the fitted cubic, quartic and quintic curves obtained by regressing X on Y (i.e. via (6.3)), the data (the set E_S) consisting of the row modal positions of one changepoint posterior distributions evaluated using (2.11) under the relevant prior assumptions. The edge-point data was spatially unsmoothed, but those points having associated posterior probability less than a pre-fixed threshold (in this case 0.2, where the common prior probability for each potential changepoint position in the sequence of length 80 is 0.0127) are omitted from the subsequent analysis - we do not regard this step as unreasonable or impractical.

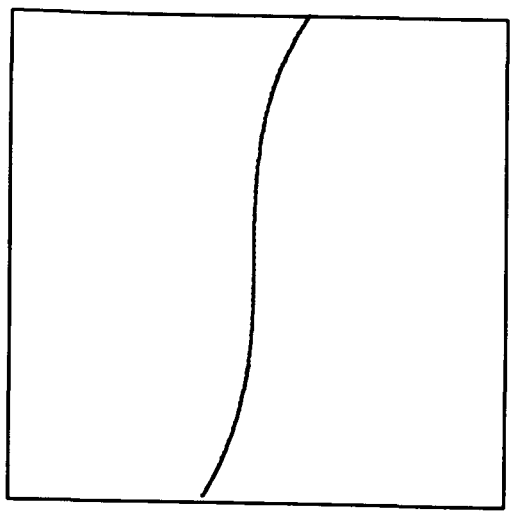


Fig 89(a) : S.N.R. = 2.0

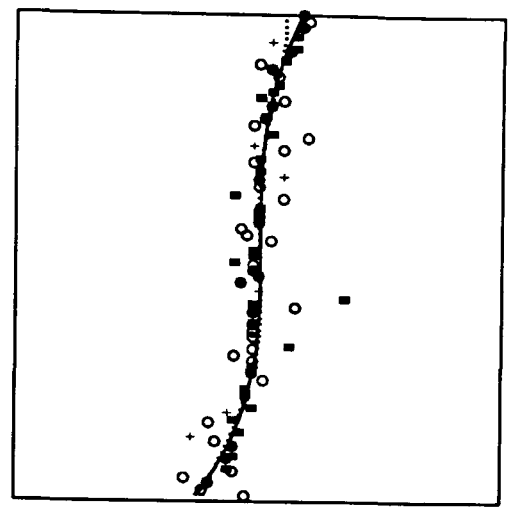


Fig 89(b) : S.N.R. = 1.4

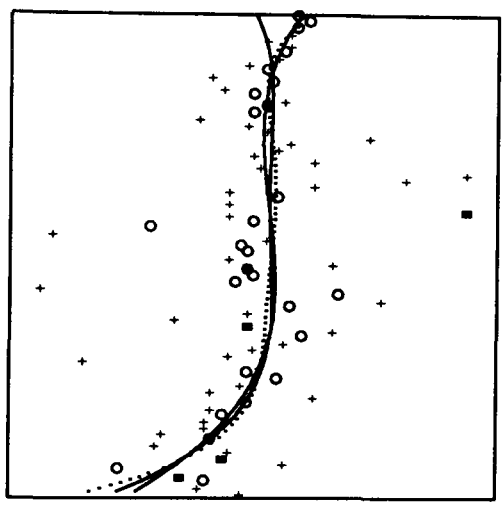


Fig 89(c) : S.N.R. = 0.8

It is clear from figure 89 that at high Signal-Noise ratios the changepoint results and consequently the fits obtained are quite adequate - the edge representations resulting from the three fitted polynomials are indistinguishable (the edge-points themselves are omitted from figure 89 (a)). At lower Signal-Noise ratios, the changepoint modes are more diversely spread over the image, and hence the resulting curve differ to a greater degree, and are less adequate as representations of the edge - this effect can be seen in (c), where each of the curves, although fair summaries of the cloud of points, differ radically from the actual edge in the lower part of scene. Also, and as would be expected from figure 89, the residual sums of squares for the curves fitted to the results in (a) were considerably smaller than those for the curves in (c), but in each case there was no significant improvement in fit over the increasingly complex models. We stress at this point that our objectives here are purely representational rather than explanatory, and so we are indifferent to precisely which model, curve and set of estimated parameters we choose, given comparable residual sums of squares. We must also guard against "overfitting" the data points - our interpretation of the nature of edges in the true scene

suggests that, very generally, they should be able to be represented adequately by a low order polynomial in this way. Thus we should be able to choose k in (6.1) or (6.3) to be no more than, say, five or six. With the large number of points in E_S, K of the order of n for an $n \times n$ image, to fit polynomials of this order should be sufficient. Evaluation of the least-squares polynomials using NAG routine F04ANF required of the order of 0.01 seconds of processing time, negligible compared to, say, the computation of the changepoint posterior distributions.

We now compare the curves and fits obtained for edges having other orientations in the true scene. Figure 90 depicts the raw changepoint results from the analysis of a true scene containing a single edge broadly making an angle of $\frac{3\pi}{8}$, $\frac{\pi}{4}$, and $\frac{\pi}{8}$ in (a), (b), and (c) respectively, where the image is derived under a Signal-Noise ratio of 1.0. In each case, the edge itself was not generated using a polynomial form as in (6.1) or (6.3). Again, the analysis was based on posterior distribution (2.11).

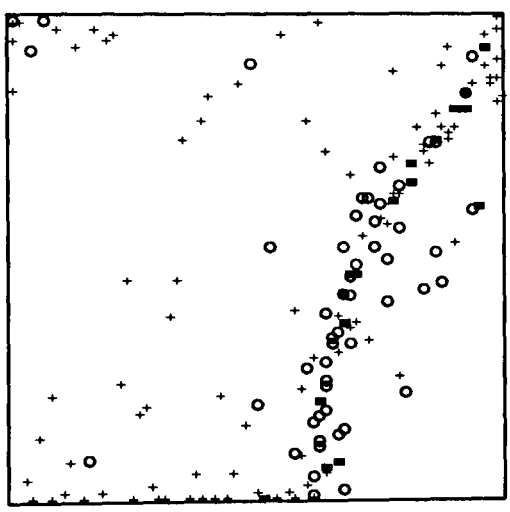


Fig 90(a) : $3\pi/8$

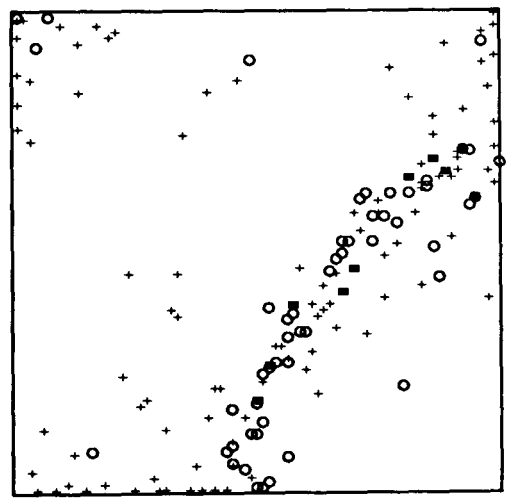


Fig 90(b) : $\pi/4$

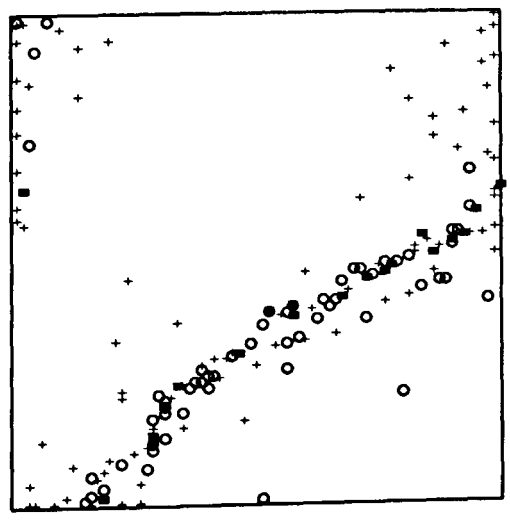


Fig 90(c) : $\pi/8$

These results seem to exhibit the same broad features, that is, general appearance, apparent number of mis-classifications, degree of localisation, and it is clear that simple smoothing

techniques could be used to improve their appearance. We now seek to fit polynomial curves to these sets of points as above to obtain a representation of the edge in each case. Figure 91 depicts the curves obtained when, for demonstration purposes, quintic polynomial forms are chosen. In each of (a), (b), and (c), the solid line represents the regression of Y on X, the dotted line the regression of X on Y. In this instance, the changepoint results were post-smoothed using a 7×7 window - the remaining points after post-smoothing are included in figure 91.

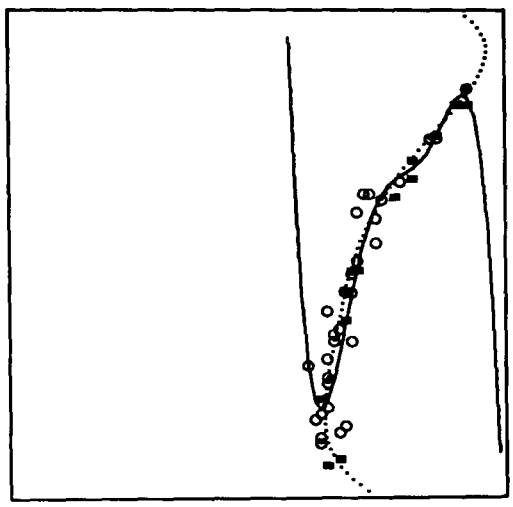


Fig 91(a) : $3\pi/8$

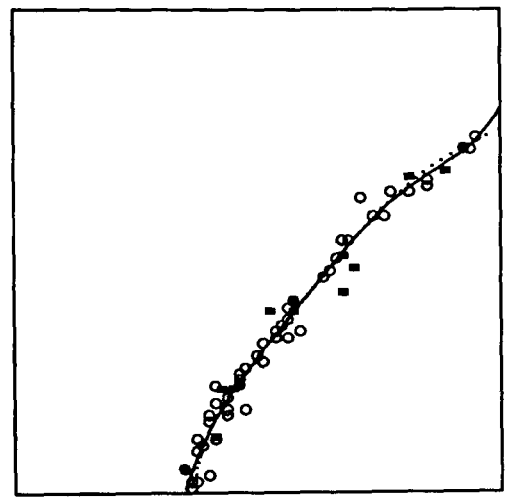


Fig 91(b) : $\pi/4$

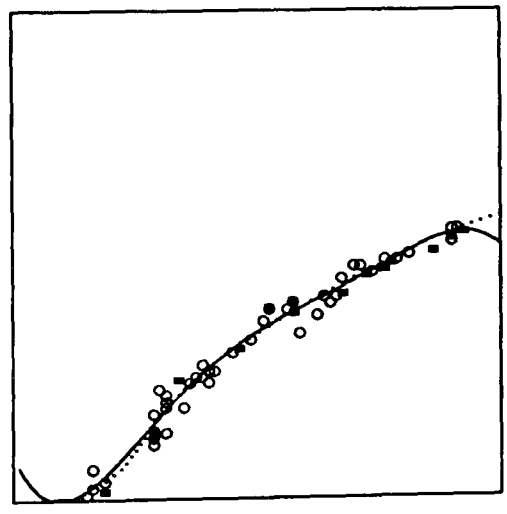


Fig 91(c) : $\pi/8$

It can be seen that each of the curves is an adequate representation of the edge structure implied in each of the data sets (in (a), we would ignore any parts of the fitted curve distant from all elements of the data set). It is interesting to note the residual sums of squares for each individual regression. In figure 91(a), The regression of Y on X resulted in a sum of squares of 1605.049, whereas the regression of X on Y resulted in a sum of squares of only 101.789. Thus we would clearly choose the latter curve as a better representation of the data points. In (b), the respective sums of squares were 413.459 and 217.877, and thus we would still choose the latter curve, but would be more indifferent between the two. In (c), the

respective sums of squares were 158.163 and 282.875, and here so we would choose the former curve. This is as we predicted above, with quality of fit in each case being dependent on edge orientation.

Thus we have seen that, frequently, single, simple edges in the true scene can be detected using changepoint analytic techniques, and represented adequately using curves derived from polynomial regression models. Crucially, these procedures can be implemented efficiently using simple and widely available routines. We now note some further points for consideration that can be seen to be relevant in this area.

(6.1.1.1) Weighting of points.

We noted above that it would be intuitively reasonable to incorporate the modal posterior probability associated with each of the recorded edge-points into our estimation/representation procedures. Also, we noted that classification of pixel as an edge-point was more credible if several other pixels in its immediate vicinity were also classified as edge-points - this formed the basis for our simple smoothing algorithm. Both of these ideas may be used in the edge representation problem by means of weighting of the data points. This would involve first forming a set of non-negative weights, $\{w_1, \ldots, w_K\}$ say, one for each element of E_S, so that w_i was large if the modal probability for point e_i was large and/or if e_i was in proximity a large number of other elements of E_S. Then, recalling the equivalence between least-squares criteria and maximum-likelihood estimation under normality, we may derive weighted least-squares estimates using standard theory. However, a precise technique for relating the modal probability and proximity to other edge-points, two quantities vastly different in nature, effectively and coherently is not immediately obvious. Also, evaluation of such weighted least-squares estimates is generally less straightforward and more computationally demanding. Routines are available, e.g. NAG library routine E02ADF, for solution of general weighted least-squares curve fitting problems, but they also involve a relatively complex implementation.

(6.1.1.2) Robustness and influence.

It is well known that least-squares criteria such as those discussed above are sensitive to outlying values, that is, inferences made from a data set containing one or more spurious observations may differ greatly from inferences made from a similar data set with the spurious observations removed. This is, naturally, a cause for some concern in the edge representation context, as here, although we might hope to remove the majority of mis-classified or outlying edge-points using the simple smoothing technique, it is clear that serious mis-classifications will occur and possibly adversely affect the estimation of the regression parameters. We thus might seek, first, to make the estimation procedure robust to the effect of such outlying edge-

points, and second, detect outlying points with reference to their degree of influence on the procedures.

Several measures can be taken to robustify the least-squares based estimation procedure in, for example, (6.1) and (6.3). For instance, we might replace D_y by D_y', where

$$D_y' = \sum_{i=1}^{K} d_{y_i}' \tag{6.5}$$

and d_{y_i}' is some distance function defined for the i'th observed and fitted y-values, and then obtain estimates for the parameters by minimisation of D_y' Interest here, therefore, clearly centres on specifying forms for d_{y_i}', and several such forms have been proposed. First, we might choose

$$d_{y_i}' = |y_i - \hat{y}_i| \tag{6.6}$$

the absolute difference in observed and fitted y-values. This is equivalent to an assumption that the error terms in the model have a double-exponential (and therefore heavy-tailed) distribution, and corresponds to the so-called L_1-norm procedure. The L_1-norm solution for an over-determined system of simultaneous equations can be evaluated using NAG library routine E02GAF - this solution is relatively analytically complex. Second, we might choose

$$d_{y_i}' = \begin{cases} (y_i - \hat{y}_i)^2 & |y_i - \hat{y}_i| \leq h \\ |y_i - \hat{y}_i| & \text{otherwise} \end{cases} \tag{6.7}$$

which is equivalent to an assumption that the error terms in the model have a "Huber" distribution, i.e. loosely, that the "smaller" errors have a Normal distribution and the "larger" errors have a double-exponential distribution, with the threshold between the two given by constant h. Again, this induces the error distribution to be heavy-tailed. Finally, we might choose

$$d_{y_i}' = \begin{cases} (y_i - \hat{y}_i)^2 & |y_i - \hat{y}_i| \leq h \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

which is approximately equivalent to an assumption that the error terms have a t distribution with number of degrees of freedom relating to choice of h. Thus again, this choice corresponds to choosing a heavy-tailed distribution for the error terms. Minimisation of (6.7) and (6.8) is possible using routines in the E04 section of the NAG library. This generally becomes quite involved and complex, and thus requires careful implementation.

Each of the proposed robustness measures can be viewed within the Bayesian framework; see, for example Smith (1983). Consider the common statistical problem of making inference about an *a priori* unknown parameter θ conditional on data Y. In the Bayesian methodology, we make inference via the posterior distribution $[\theta \mid Y]$ given by

$$[\theta \mid Y] \propto [Y \mid \theta][\theta]$$

in the usual way. Taking logs and differentiating partially with respect to θ, we obtain

$$I_{\theta|Y} = I_{Y|\theta} + I_{\theta} \tag{6.9}$$

where $I_{\theta|Y} = -\frac{\partial}{\partial\theta}\log[\theta \mid Y]$, $I_{Y|\theta} = -\frac{\partial}{\partial\theta}\log[Y \mid \theta]$, and $I_{\theta} = -\frac{\partial}{\partial\theta}\log[\theta]$ - we refer to the terms in (6.9) as "influence functions". It is thus clear through (6.9) how assumptions concerning the likelihood function, specifically the error structure imposed on the data in the Signal + Noise model, can be related on a linear scale in the inferential context, and how varying the choice of error structure can lead to robust estimation procedures. Consider the choices for d_{y_i}' that we have already discussed - $d_{y_i}^2$ originally, then the functions in (6.6), (6.7) and (6.8), and the corresponding distributional assumptions of Normal, double-exponential, Huber, and approximate t-distributed errors. The influence functions for each of these forms are depicted in figure 92, with residual $\varepsilon_i = y_i - \hat{y}_i$ plotted as abscissa and corresponding influence plotted as ordinate.

In (a), the influence of ε_i increases linearly with its magnitude. In (b), the influence of ε_i is constant. In (c), the influence of ε_i varies linearly with its magnitude when $|\varepsilon_i|$ is less than h and is constant otherwise. In (d), the influence of ε_i varies linearly with its magnitude when $|\varepsilon_i|$ is less than h and is zero otherwise. Thus in the latter three case, we limit the influence on our inferences about the unknown parameters of the system of points for which ε_i is large. By such methods, therefore, we would hope to make the estimation procedures robust to the presence of outlying values. However, in a practical context, "tuning" the procedure (choosing values of h) might prove difficult, and, as mentioned previously, the necessary minimisation routines are computationally expensive, and also involve other difficulties (starting values, natural constraints to be specified). Thus, we feel that rather than attempt to robustify the least-squares estimation procedure, we should remove possible outlying or misclassified points using the smoothing technique we have already seen, or other such simple techniques.

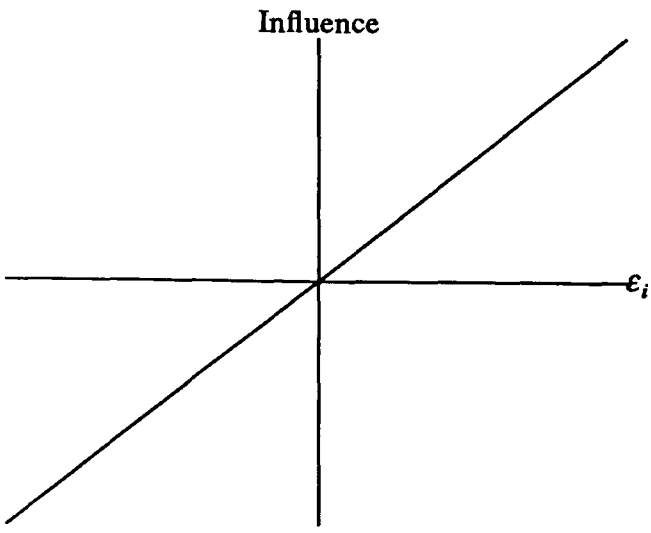


Fig 92(a) : $d_{y_i}{}^2$

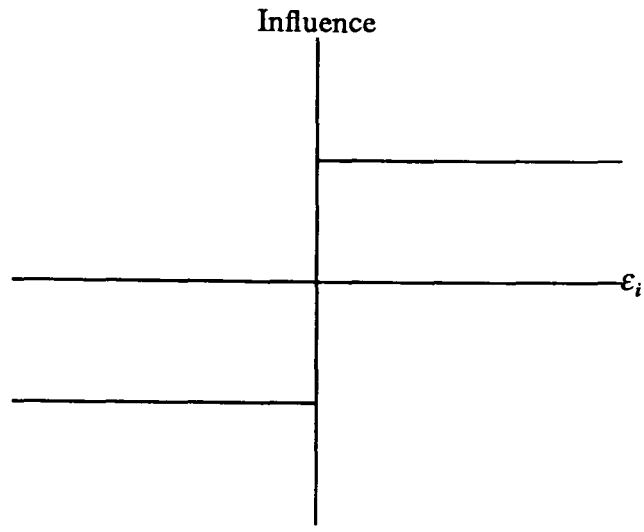


Fig 92(b) : (6.6)

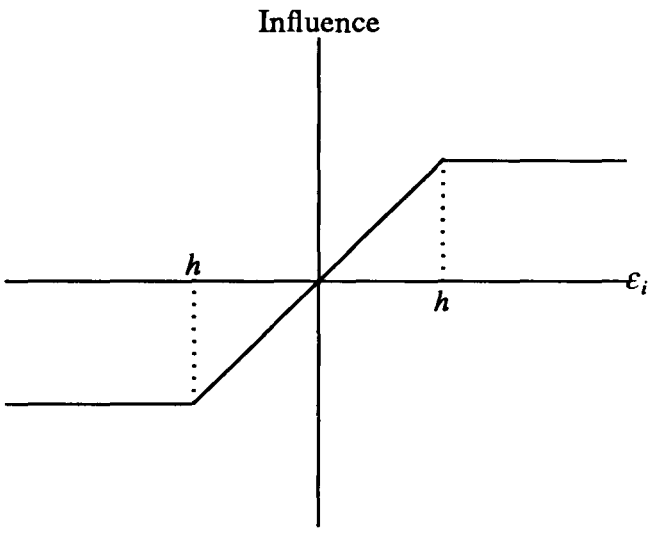


Fig 92(c) : (6.7)

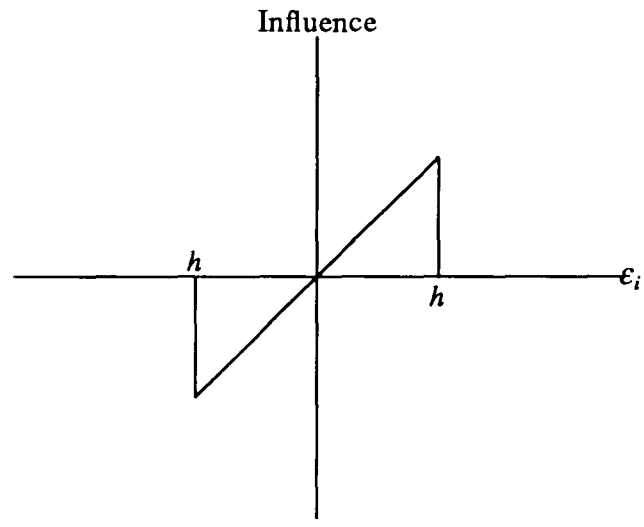


Fig 92(c) : (6.8)

A topic that is related to the robustness considerations that we have mentioned above is that of identification of outliers in a data set (rather than seeking to limit their effect on *a posteriori* inferences). In a Bayesian framework, for instance, we might assess the importance of a point with respect to posterior inferences about the unknown parameter(s) of interest with reference to its influence on the relevant posterior or predictive distribution - see, for example, Johnson and Geisser (1982), Pettit and Smith (1983) for a discussion and a comparison with various non-Bayesian measures concerned with identification of outliers in the Normal linear model. This technique for outlier detection is also similar in nature to non-Bayesian cross-validation type techniques. We might view, therefore, such techniques as alternatives to the simple smoothing technique implemented above. However, such methods are known to be time consuming in their implementation, and thus we are again inclined against them in favour of simpler methods. We shall mention the importance of influence and robustness in more detail in a later section in connection with a least-squares estimation problem for a particular family of closed curves.

Thus we have seen above that, after suitable smoothing etc. of the results arising from changepoint analysis of an image, single simple edges may generally be represented adequately using parametric regression models, and that for these simple models the additional computation required is minimal. However, we more commonly have to analyse images derived from true scenes containing multiple or composite edges, or closed curves representing the edges of objects. The simple polynomial regression models we have described above are inadequate for modelling and hence reconstructing these types of edge, and thus we might feel it necessary to introduce other, more sophisticated reconstruction techniques. We now discuss a familiar non-parametric regression-type approach commonly used when attempting solution to the curve-fitting problem.

(6.2) Curve-fitting via spline-functions

Consider first the problem of reconstructing a simple edge from the results of a changepoint based analysis of an image. Given the set of recorded and accepted edge-points E_S and their coordinates, (x_i, y_i), we seek as a solution to the edge-reconstruction problem some (continuous) functional relationship between the x and y coordinates for values of x in (some region of) S, not merely in the function at the data points themselves, but also in the some sort of extrapolation in the interval on the x axis between two adjacent values of x_i. The common statistical model used in this data representation problem can be written

$$Y_i = f(X_i) + \varepsilon_i \tag{6.10}$$

where our objective is to make inference about the function f - more specifically, we seek some estimate $\hat{f}$ of f. For convenience, we shall assume that the observed x_i are ordered $(x_1 \leqslant x_2 \leqslant \ldots \leqslant x_P)$, and that the error terms ε_i are independent zero-mean variables.

We could, of course, list any number of criteria for the choice of $\hat{f}$. We could, for instance, choose a least-squares criteria as above, and hence choose $\hat{f}$ so that the residual sum of squares is minimised. Without any further constraints, the parametric solution to such a problem is of the form of a polynomial function of order $P + 1$, similar to those in the models described above, and the The non-parametric solution is effectively any function that interpolates the data points. Such solutions, however, are generally intuitively unsatisfying, and are especially so in the edge-reconstruction context - we interpret edges in a true scene to be broadly smooth curves with possibly occasional sharp corners, rather than curves exhibiting a high degree of localised variability. For this reason, it has been widely suggested that instead of merely minimising the residual sum of squares, we should rather minimise some function involving the sum of squares and some penalty function used to penalise any curve exhibiting rapid local variation. One such penalty function which has been commonly used takes the

form

$$\int \{f''(x)\}^2 dx$$

which is combined additively with the residual sum of squares to produce a "modified" or "penalised" sum of squares - such a choice for the penalty function has a physical interpretation as the strain energy in a wire when that wire takes the same shape as f. Thus the problem of producing an estimate $\hat{f}$ reduces to finding that function which minimises the modified sum of squares $D^m{}_y$ given by

$$D^m{}_y = \sum_{i=1}^{P} (Y_i - f(x_i))^2 + \beta \int \{f''(x)\}^2 dx \qquad (6.11)$$

where β is regarded as a "smoothing" parameter, and is used to trade-off fidelity to the data (as measured by the first term) against the smoothness or localised variability of f (as measured by the second term). Such a procedure can be viewed as a penalised-likelihood approach (under the assumption that the error terms in (6.10) are normally distributed), and can also be justified in a Bayesian framework (see, for example, Kimeldorf and Wabha (1970)).

The solution to the minimisation problem in (6.11) can be shown to be a piecewise cubic polynomial having continuous first and second derivatives at the data points. Such a curve is termed a cubic spline, and represents a flexible solution to the non-parametric regression problem that we have described above. Spline functions are widely used as representational devices in data analysis, image processing and computer graphics. There exists an extensive literature on splines in a numerical analysis context, and a less extensive one relating to statistical analysis. Three important references are Silverman (1985) which contains a statistical grounding and theoretical justification of Bayesian and non-Bayesian aspects of the spline function in data analysis problems, further references relating to the statistical foundation of the subject, and univariate and bivariate examples, De Boor (1978), which describes the problem in a numerical analytic context, and providing a number of straightforward and efficient Fortran programs for implementations, and Bartels *et al.* (1987), which contains a more contemporary but less rigorous numerical analytic view, geared especially to computer graphical representations, and investigating extensions of the ideas to multi-dimensions and generalised spline curves of higher degree. Other important theoretical and practical application references are detailed within these three. Specifically in the context of computer vision, Shahraray and Anderson (1989) used the formulation in (6.11) to reconstruct contours in images. Crucially, it can be shown that the computational burden involved in evaluating f is surprisingly small, in fact going up only linearly with the number of data points. Thus, it seems that spline

functions provide an ideal form of solution to our edge-reconstruction problem, not only in the case of simple single edges, but also for more complex composite edges. In the case of such composite true scenes, spline smoothing problem is clearly slightly different to both the single edge or single closed curve problems. The most important features in composite scenes are those points at which smooth curves meet with a discontinuity in derivative. Consider, for instance, the reconstruction of a rectangle with sharp corners from edge-point data. Clearly, the standard cubic spline curve cannot deal easily with such a problem, as we have seen that cubic splines always have continuous first derivative. This difficulty can be overcome, however, with a little care - for example, we might use an interpolating spline ($\beta = 0$) through a subset of E_S, the elements of which are chosen as representing the positions where the smooth curves meet. We discuss composite true scenes and the related reconstruction problems at greater length at a later stage.

There are, however, a number of negative aspects associated with implementing a spline based curve fitting algorithm. First, we have the problem of choosing the value for the smoothing parameter β. This task may be performed automatically using a standard or generalised cross-validation technique (for details and comments, see Silverman (1985) and discussion), but this appears to add considerably to the computational burden, and although intuitively reasonable is still somewhat arbitrary. Also, we cannot readily regard β as a hyper-parameter that we may choose at our discretion. Secondly, and related to the first point, our sole goodness-of-fit statistic is the value of the modified sum of squares in (6.11) which, as indicated previously, is ultimately of little use in assessing adequacy of the curve as a representation of the relationships in the data. Thirdly, and perhaps most importantly, by their very non-parametric nature, spline curves can not be regarded as anything but representational - there is no explanatory aspect, nor is their any scope for model elaboration or simplification - and indeed we might argue that because of the scale difference between edge and pixel width, indicating that the inter-ordinate distance is small compared with the size of the edge itself, the smoothed changepoint results themselves provide as adequate a representation of the edge as the spline smoothed curve in many cases. Therefore, spline based curve-fitting techniques seem incongruous in relation to our chief interpretation of the image analysis problem - we are principally and ultimately concerned with a pixel-by-pixel segmentation of the true scene into regions of like or homogeneity at some discretised level. We have tolerated the representational polynomials above because such a segmentation could be readily deduced from them, but for piecewise polynomials such a segmentation is by no means as straightforward. It is for this reason that we regard spline based curves as representations of edges in the true scene, despite their many positive aspects, as beyond the scope of this thesis. We shall now subsequently consider only techniques that have some form of explanatory quality, or aid in solution to the segmentation problem.

We now concentrate on the edge-reconstruction problem for true scenes containing single convex objects. Now, it is clear that the polynomial regression models proposed above are inappropriate, and indeed not sufficiently sophisticated to reconstruct accurately or even adequately the closed curve defining the object boundary. However, the ideas introduced above of minimum distance/least-squares criteria are still appealing. Thus, initially, we concentrate on attempting to fit other parametric models to the data set E_S using criteria similar in nature to (6.2) and (6.4). Note that, in solving the edge-reconstruction using parametric models, we also partially solve the object detection and recognition problem discussed previously (we gain information as to the location, dimensions and orientation of the object). This would not as readily be the case if we were to use non-parametric models. We now proceed and attempt to implement the least-squares techniques discussed above for a particular class of convex objects whose boundary can be described using a particular parametric form.

(6.3) Edge-reconstruction for elliptical objects.

Consider an ellipse in the Cartesian plane having centre (p,q), with major and minor semi-axes a and b, and orientated so that its major axis make an angle of α with the positive x-axis. Such an ellipse defined by these five parameters is depicted in figure 93.

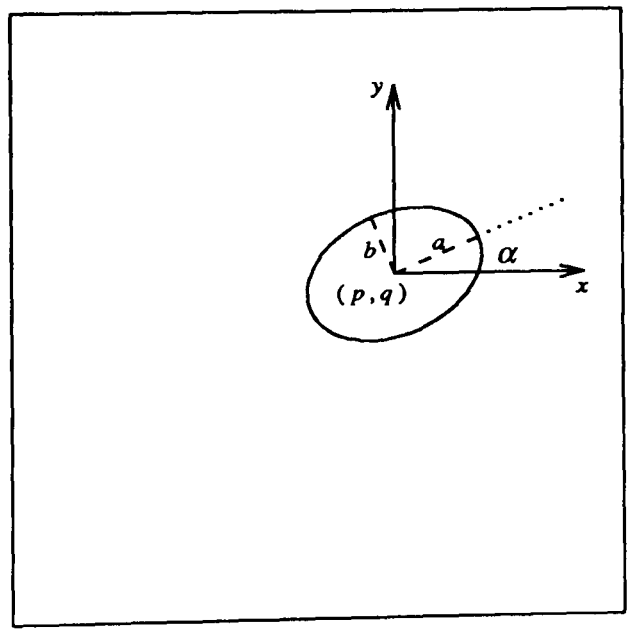


Fig 93 : Ellipse

A point (x,y) in Cartesian coordinates lying on this ellipse therefore satisfies

$$\frac{((x-p)\cos\alpha + (y-q)\sin\alpha)^2}{a^2} + \frac{(-(x-p)\sin\alpha + (y-q)\cos\alpha)^2}{b^2} = 1. \quad (6.12)$$

We regard the ellipse to have a sufficiently flexible parametric form as to justify our attention (and venture to suggest that it be taken very broadly as representative of the class of all simple

convex objects).

Our objective in the edge-reconstruction problem is thus to estimate the five parameters the ellipse on the basis of the elements of E_S, the edge-points resulting from changepoint based analysis of the image. As discussed previously, it is reasonable to solve this estimation problem using some minimum-distance/least-squares criterion. However, several points are relevant here that were not relevant above in our study of polynomial regression and least-squares estimation in the context of simple edges. First, we felt that in the majority of cases, at least one of (6.2) or (6.4) would be acceptable as the quantity for minimisation in the above due to the nature of the edge. Here, where the edge is in fact a closed curve, neither seems appropriate, and we might instead prefer to concentrate on the minimum perpendicular distance criterion. Secondly, we were able to obtain adequate representations for the edge in our previous analysis using simple linear models. It is clear that this is not immediately possible here, as (6.12) is plainly non-linear in each of the five parameters, and thus we might be interested in formulating a reparameterisation of the problem. These and other points are discussed in detail below.

The problem of fitting conic sections to data has been studied previously by Bookstein (1979), who used a conic splining technique. Other related problems have been discussed by Ripley and Rasson (1977), and Moore (1984), who studied the problem of reconstructing convex sets where typically the data consisted of a set of coordinates in the plane, and knowledge as to whether each point was internal or external to that set. For our purposes, the most relevant formulation and solution to the ellipse reconstruction problem was described by Forbes (1987). It is this formulation that we follow most closely.

(6.3.1) Ellipse reconstruction - standard formulation.

First we consider of problem of fitting an ellipse through the elements of E_S so that the sum of shortest distances between point and ellipse over all points is minimised, that is we seek to minimise

$$D = \sum_{i=1}^{K} d_i^2 \tag{6.13}$$

where $d_i^2 = (x_i - x_{iC})^2 + (y_i - y_{iC})^2$ and (x_{iC}, y_{iC}) are the coordinates of the point e_{iC} on the ellipse nearest to point e_i, for each i (i.e. e_{iC} lies at the intersection of the ellipse and the normal to ellipse passing through e_i). Our first task, therefore, is to determine these coordinates. We may parameterise a point (x, y) on the ellipse by

$$x = a\cos\alpha\cos\theta - b\sin\alpha\sin\theta + p$$

$$y = a\sin\alpha\cos\theta + b\cos\alpha\sin\theta + q$$

and hence, after some algebra, it can be shown the minimum value of d_i^2 occurs when θ satisfies

$$a(x\cos\alpha + y\sin\alpha)\sin\theta + b(x\sin\alpha - y\cos\alpha)\cos\theta + (b^2 - a^2)\sin\theta\cos\theta = 0. \qquad (6.14)$$

It is not straightforward to solve (6.14) for θ analytically. Forbes suggests including the determination of the coordinates of the points e_{iC} (through the angles θ_i defined by (6.14)) for $i = 1, \ldots, K$ into the overall optimisation procedure, but this increases the dimensionality of the problem by K. Thus, we reject this minimum distance ellipse for the moment, and turn our attention to other possibilities.

One minimum distance criterion of interest is the minimum radial distance (i.e. rather than measuring the shortest distance to the ellipse for each point e_i, we measure the distance along the line radial from the centre of the ellipse through e_i). This may be regarded as an approximation to the minimum perpendicular shortest distance, and the approximation will clearly be more adequate for points for lying near the major and minor axes of the ellipse. Using this approximation is advantageous, as we may now obtain the individual distance terms more readily - we may immediately write down the point-to-curve distance in this case without the need for solution of equations such as (6.14). For each i, the point on the ellipse of interest is defined (in our alternative parameterisation) by θ_i, where

$$\tan(\theta_i + \alpha) = \frac{y - q}{x - p}$$

or

$$\tan\theta_i = \frac{(y - q)\cos\alpha - (x - p)\sin\alpha}{(y - q)\sin\alpha + (x - p)\cos\alpha}$$

from which we may straightforwardly evaluate the minimum radial distance for each point e_i in terms of (p, q, a, b, α). Forbes discusses a correction of the minimal radial distance criterion by evaluating the angle between the radial vector $(x - p, y - q)^T$ and the normal vector $\boldsymbol{n}$ given in this case by

$$\boldsymbol{n} = \begin{pmatrix} (x - p)(a\cos^2\alpha + b\sin^2\alpha) + (y - q)(b - a)\cos\alpha\sin\alpha \\ (x - p)(b - a)\cos\alpha\sin\alpha + (y - q)(b\cos^2\alpha + a\sin^2\alpha) \end{pmatrix}$$

for a point (x,y) on the ellipse in the above parameterisation which can be used to give some indication of by how much the radial distance overestimates the true (perpendicular) distance between each point e_i and the ellipse.

Forbes also describes a further approximate least-squares technique based on a linearised version of the problem (that we shall discuss in detail below) and using a weighting scheme for each of the residuals.

Solution of the ellipse reconstruction problem for this parameterisation and via these exact and approximate least-squares techniques, therefore, requires the use of some numerical minimisation routine, and thus may be relatively complicated to implement. Prior to attempting such solutions, therefore, we seek a simpler formulation. It is clear that, in the above parameterisation, the potential problems arise due to the non-linearity of, for instance, (6.12) and (6.13) in the ellipse parameters. We thus seek a formulation in which this non-linearity is replaced by linearity (we have seen in the previous section, parameter estimation procedures etc. are much more straightforward to implement for models in which the parameters appear linearly, and also that we may easily incorporate simple ideas of robustness and influence into such (least-squares based) procedures for linear models for data). To this end, we now discuss a linearised version of the ellipse reconstruction problem.

(6.3.2) Linear least-squares ellipse reconstruction

It is well known that the equation of a general conic section can be expressed in Cartesian coordinates in the form

$$c_5x^2 + c_4y^2 + c_3xy + c_2x + c_1y + c_0 = 0. \qquad (6.15)$$

For an ellipse we have the additional condition that $c_3^2 < 4c_5c_4$ (with similar constraints for circles, parabolas etc.). Clearly, (6.13) defines a six parameter model, whereas the ellipse reconstruction problem involves only a five parameter model. Forbes discusses a suitable constraint in order to define a well determined and numerically stable model which we now describe.

We seek some simple linear constraint on the parameters of (6.15). It is inadvisable merely to set one of $(c_0, \ldots, c_5)$ equal to a non-zero constant and divide through (6.15), as in reality that parameter might in fact be equal to zero for the ellipse concerned. However equating coefficients of terms in x^2, y^2 etc. in (6.12) and (6.15), it is apparent that the constants in (6.15) are given by

$$c_5 = \frac{\cos^2\alpha}{a^2} + \frac{\sin^2\alpha}{b^2}$$

$$c_4 = \frac{\cos^2\alpha}{b^2} + \frac{\sin^2\alpha}{a^2}$$

$$c_3 = 2\cos\alpha\sin\alpha\left(\frac{1}{a^2} - \frac{1}{b^2}\right)$$

$$c_2 = -2c_5p - c_3q$$

$$c_1 = -2c_4q - c_3p$$

$$c_0 = c_5p^2 + c_4q^2 + c_3pq - 1.$$

It is immediately clear that both c_4 and c_5 are strictly non-negative. Given *a priori* ignorance of the actual parameter values, we note that setting $c_4 + c_5$ equal to a constant is preferable to setting either c_4 or c_5 to a constant individually (as we have no knowledge as to which of these options, with the parameters appearing asymmetrically in each, is actually more accurate) - clearly $c_4 + c_5$ can be reduced to $(a^2 + b^2)/a^2b^2$.

In light of the above discussion, therefore, we may re-express (6.15) as

$$(x^2 + y^2) - \gamma_4(x^2 - y^2) - 2\gamma_3xy - \gamma_2x - \gamma_1y - \gamma_0 = 0 \qquad (6.16)$$

after dividing each term by $c_4 + c_5$. Equating coefficients in (6.12) and (6.16), we now have that

$$\gamma_4 = \frac{a^2 - b^2}{a^2 + b^2}\cos2\alpha$$

$$\gamma_3 = \frac{a^2 - b^2}{a^2 + b^2}\sin2\alpha$$

$$\gamma_2 = 2p(1 - \gamma_4) - 2q\gamma_3$$

$$\gamma_1 = 2q(1 + \gamma_4) - 2p\gamma_3$$

$$\gamma_0 = 2\frac{a^2b^2}{a^2 + b^2} - \frac{p\gamma_2}{2} - \frac{q\gamma_1}{2}$$

enabling us to relate parameters in this our linear parameterisation $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ to our original parameterisation (p, q, a, b, α) of the ellipse. Forbes notes the geometric significance of γ_3 and γ_4 and their relation to the positions of the foci of the ellipse. It is possible, after some algebra, to invert the relations above, and consequently, if we define eccentricity

$e = b/a$, from Forbes (section 2.4) we have that

$$\tan 2\alpha = \frac{\gamma_3}{\gamma_4}$$

$$e = \left(\frac{1-\gamma_5}{1+\gamma_5}\right)^{\frac{1}{2}}$$

$$p = \frac{((1+\gamma_4)\gamma_2 + \gamma_3\gamma_1)}{2\gamma_6}$$

$$q = \frac{((1-\gamma_4)\gamma_1 + \gamma_3\gamma_2)}{2\gamma_6}$$

$$a = \left(\frac{(2\gamma_0 + p\gamma_2 + q\gamma_1)}{2(1-\gamma_5)}\right)^{\frac{1}{2}}$$

$$b = ea$$

with γ_5 and γ_6 given by

$$\gamma_5 = (\gamma_3^2 + \gamma_4^2)^{\frac{1}{2}} \quad \text{and} \quad \gamma_6 = 1 - \gamma_5^2$$

and α is determined exactly given γ_3 and γ_4 by the fact that $\gamma_4 \geq 0 \Rightarrow |\alpha| \leq \frac{\pi}{4}$, and $\gamma_3 \geq 0 \Rightarrow \alpha \geq 0$.

Thus, in (6.16), we have a version of the general ellipse equation that is linear in the set of ellipse parameters. Thus, as in our reconstruction of simple edges using polynomial regression, we may now simply fit a linear model of the form of (6.16) to the data in E_S in order to reconstruct the ellipse, that is, given the coordinates (x_i, y_i) of e_i, we find the least-squares solution to the over-determined system of equations given by

$$\varepsilon_i = (x_i^2 + y_i^2) - \gamma_4(x_i^2 - y_i^2) - 2\gamma_3 x_i y_i - \gamma_2 x_i - \gamma_1 y_i - \gamma_0 = 0 \qquad (6.17)$$

for $i = 1, \ldots, K$. It is interesting to note the precise quantity measured by ε_i. On inspection of the form of (6.15), Forbes (section 2.3) discovers that the i'th residual error ε_i can be written as $-2e/\pi(1+e^2)$ times the difference between the area of the best fit ellipse (as defined by the system (6.17)) and the ellipse with the same centre, orientation and eccentricity and that passes through (x_i, y_i). Thus, whereas previously we have interpreted the least-squares parameter estimates to be those for which the sum of squares of distances in the plane is minimised, here we interpret the estimates to be those for which the sum of squares of the differences in areas of ellipses is minimised. This is quite a radical change in interpretation, although a perfectly legitimate one. For the general edge-reconstruction problem, therefore,

our new formulation of this problem and its subsequent interpretation may seem inappropriate. However, in the specific context of ellipse reconstruction, which is inextricably linked to the object detection problem, we regard the new formulation to be rather more sensible, as "objects" are viewed expressly as space-filling entities in two dimensions (and we have (inadvertently) derived a solution relating to minimum area residuals which actually reflects the nature of the ellipse reconstruction/detection problem). Thus, the use of (6.15) and the usual least-squares criterion is reasonable. Recall, also, that such an approach to parameter estimation is equivalent to maximum-likelihood estimation in the Normal linear model, and consequently we can use robustifying measures as discussed in section (6.1.1.2) to improve the estimation procedure.

Forbes (section 2.5) proceeds to examine in detail the numerical stability of the model proposed in (6.16), and his principal conclusions are as follows. The model is "well-conditioned" (in the numerical analytic sense) if e is large, or if α is near 0 or $\pi/2$. For small e ($e < 0.1$, an extremely eccentric ellipse), it is "numerically advantageous" to rotate the data prior to implementation of the estimation procedure so that the major axis of the ellipse is parallel to the x-axis. Forbes suggests that this can be done automatically by first fitting a straight line to the data, and then rotating by the angle that the straight line made with the x-axis. In our experience, this procedure often gives unsatisfactory results when automated, and is only generally useful when implemented manually. It should be noted, however, for our applications and for the objects for which our changepoint techniques have been developed, a value of $e < 0.1$ can be regarded as untypical. Forbes also suggests that the data should be translated by their centroid so that the centre of the linear least-squares ellipse lies close the origin during the estimation procedure, and subsequently translated back to its true location before the estimates are reported. Again, in practice, we find this procedure unreliable and generally unhelpful.

Forbes (section (2.6)) also proceeds with details of how (classically) to estimate the variances of the parameters of interest (p, q, a, b, α) via the variance of $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ - each of the parameters of interest can be regarded as a function, $z = z(\gamma)$ say, of γ, and consequently the variance of z can be written

$$\sigma_z^2 = \sigma^2 \nabla z^T (J^T J)^{-1} \nabla z \tag{6.18}$$

where σ is the root mean square residual resulting from the fit of (6.17), J is the design matrix for the model (6.17), and ∇ is the gradient operator. Whilst it must be noted that such a classical statistical procedure conflicts with our general Bayesian framework, such a procedure can be used to give us qualitative (if not quantitative) insight into the nature of estimates arising from a least-squares based analysis of the linear model (6.27). For example, Forbes considers

the limiting cases as e approaches 0 and 1, and examines the resulting behaviour of the parameter variances. He discovers that the estimate of the variance of α increases with e - this is entirely reasonable, as e tending to 1 corresponds to an ellipse becoming circular, when of course α is a redundant parameter. He also notes that the variance of each of the location parameters p and q depends on e and α, and that with e small, α near 0 induces a large variance for p and a small variance for q, whereas for α near $\pi/2$ the converse is true. Again, the behaviour in each case is intuitively reasonable. Finally, from variance calculations, Forbes establishes that the variance of a increases with e, and also that for a very eccentric ellipse, it is important to have sufficient data points at the extreme values on the major axis to accurately determine a.

As indicated above, we might regard quantitative inferences based on such estimation procedures as theoretically unsound - we have used classically (maximum-likelihood) based estimation techniques that lie uneasily with the Bayesian methodology that we have previously adopted. However, for our purposes and in the context of a relatively straightforward (and, crucially, linear) model, where the number of data points is relatively large, such procedures generally suffice. We shall see at a later stage how estimates of variances can be useful as model diagnostics, and as a basis for a technique for the removal of of outlying points.

Forbes suggests one further approximate method, in which each of the points in the linear least-squares problem described by (6.16) are weighted by a factor related to the difference between the residual error given in (6.17) and the true (shortest) distance to the ellipse. The approximate method thus requires the numerical solution of a non-linear least-squares problem that can be obtained using an iterative linear least-squares algorithm. As mentioned above, such procedures are often difficult to implement and time-consuming. Thus, for the moment, we concentrate solely on the linear formulation and the model in (6.16) and (6.17).

We now turn to the actual implementation of the estimation procedures described above. We begin with an investigation of the parameter estimates obtained for simulated edge-point data, as it is important that we have a thorough understanding of the adequacy of the procedures before we turn to the analysis of data arising from the edge-detection routines described in previous sections.

(6.3.3) Analysis of simulated edge-point data.

Our interest in this section lies in the performance of the estimation schemes described above, principally the linear least-squares scheme given in (6.16) and (6.17), under various conditions. For instance, we certainly wish to investigate (quantitatively) the estimates obtained for different combinations of the parameters (p, q, a, b, α), and the effect of varying the degree of "noise-corruption" in the data (recall that the least-squares approach is equivalent to maximum-likelihood estimation in the case where the error-terms are normally

distributed - we have interest in the effect that altering the variance of this Normal distribution has on the resulting estimates). Also, Forbes concentrates on the situation where the data is uniformly distributed around the edge of the ellipse - we have interest in the case where the data are not distributed in this way, specifically in the case where the data are restricted to one section of the edge. Finally, we are interested in the robustness of the formulation to outlying values (points lying away from the edge of the ellipse). As we shall see, each of these areas is important in the edge-detection context. First, we discuss how precisely to introduce randomness into the simulated edge-point data given the values of (p, q, a, b, α).

It is clear that if the number of edge-points K is greater than or equal to 5, and each of the points lies precisely on the ellipse, then each of the schemes will reconstruct the ellipse exactly - this case is of little theoretical or practical interest. However, if the points are displaced in some (random) way then consequently we would expect some disparity between the true ellipse parameters and the resulting estimates of these parameters, with model adequacy or otherwise being partly judged with respect to the degree of this disparity in the usual way. We may introduce randomness into the edge-point data in several ways. Recall that the linear formulation subsequently requires an assumption of normality in the error terms ε given by

$$\varepsilon = (x^2 + y^2) - \gamma_4(x^2 - y^2) - 2\gamma_3 xy - \gamma_2 x - \gamma_1 y - \gamma_0 \sim N(0, \sigma_\varepsilon{}^2)$$

where, in the context of our inferences about (p, q, a, b, α), σ_ε is an unknown nuisance parameter. Thus, in order to introduce randomness in a way so that our model is a correctly specified, we might generate a value of ε for some fixed value of σ_ε, and replace γ_0 by $\gamma_0 - \varepsilon$ in (6.16). Now it is clear from the inversion formulae that map $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ to (p, q, a, b, α) that as γ_0 is present only in the forms for a and b, it is only these two parameters that will be effected by any alteration to γ_0. Thus to generate an edge-point randomly with Normal errors, we may simply replace γ_0 by $\gamma_0 - \varepsilon$ in the formula for a and b, and obtain a^* and b^* and then generate a point (x, y) using

$$\begin{aligned} x &= a^* \cos\alpha \cos\theta - b^* \sin\alpha \sin\theta + p \\ y &= a^* \sin\alpha \cos\theta + b^* \cos\alpha \sin\theta + q \end{aligned} \qquad (6.17)$$

for some value of θ (to produce data uniformly distributed around the ellipse, we simply choose the K values of θ to be $\pi/K, 2\pi/K, \ldots, 2\pi$). Under such a scheme, the model in (6.17) is correctly specified. Figure 94 depicts 80 edge-points uniformly distributed in this way, generated using three different values of σ_ε. The actual parameter vector for the ellipse from which the data was generated was $(p, q, a, b, \alpha) = (40, 40, 20, 10, 0.5)$.

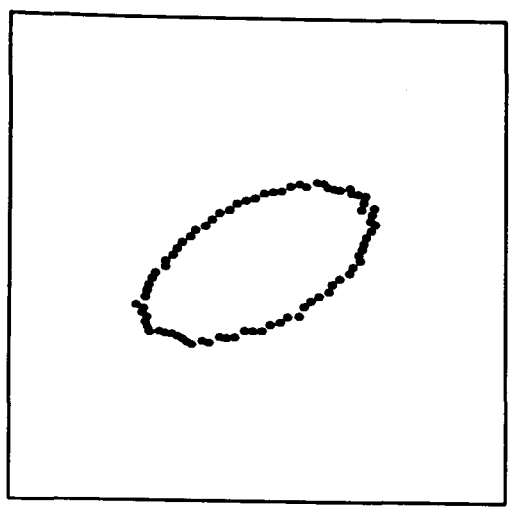


Fig 94(a) : σ_ε = 10.0

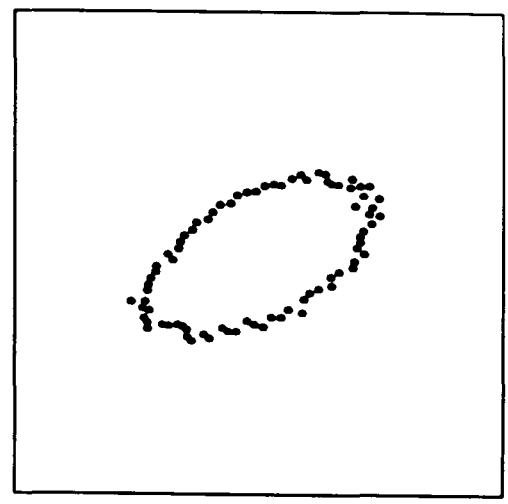


Fig 94(b) : σ_ε = 20.0

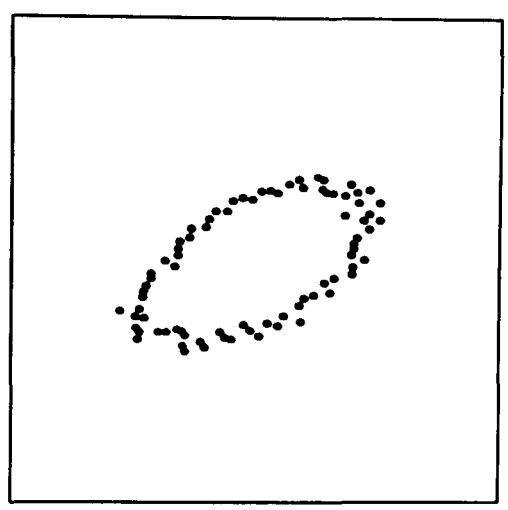


Fig 94(c) : σ_ε = 30.0

In each case, for this particular ellipse, the randomness of the points is most visually evident at the ends of the major axis. These pictures are reminiscent of many we have seen in previous sections. However, the technique that we have used to induce randomness in the points is not intuitively appealing as a simulation technique for points resulting from changepoint based edge-detection analysis - recall that the error terms here relate to a measurement of area, whereas, in actuality, due to the nature of the changepoint technique, we might expect them to relate to some measurement of distance. Another similar feature is that, in light of the nature of the changepoint posterior distributions that we have studied, we would expect the detected edge-points to lie externally (i.e. more distant from the centre of the ellipse than the edge itself) to a (raised-level) object on a (lower-level) background - here we have internal and external points. In fact, we have little knowledge of the true distribution of detected points resulting from changepoint analysis. In spite of these remarks, simulating points in the way described is useful in order to allow us to study to performance of the estimation procedures.

We could, of course, simulate random points in many other ways - for instance, we could include additive errors in x and y in (6.17), or in any of the ellipse parameters etc..

However, only for our original scheme does (6.14) represent a correct model specification, and so it is on this scheme that we concentrate.

We now investigate the estimates obtained using the linear model (6.14) and the simulated data. For demonstration purposes, we study the $(40,40,20,10,0.5)$ ellipse above for various values of the error standard deviation σ_ε, with now 20 data points being uniformly distributed on the edge of the ellipse. The least-squares solution to the over-determined system of equations is obtained using NAG library routine F04ANF, and standard errors for the ellipse parameters are obtained using (6.18). The parameter estimates, standard errors and mean square errors are presented in table 2.

	σ_ε				
	10	20	30	40	50
p	40.073 (0.106)	40.147 (0.213)	40.223 (0.322)	40.301 (0.434)	40.382 (0.553)
q	39.972 (0.076)	39.949 (0.154)	39.931 (0.232)	39.920 (0.313)	39.915 (0.396)
a	20.375 (0.084)	20.785 (0.167)	21.228 (0.252)	21.700 (0.339)	22.200 (0.429)
b	10.012 (0.171)	10.007 (0.323)	9.989 (0.457)	9.961 (0.574)	9.925 (0.676)
α	0.508 (0.001)	0.516 (0.002)	0.523 (0.003)	0.530 (0.004)	0.535 (0.005)
m.s.e	5.89	11.65	17.29	22.81	28.22

Table 2

Clearly, the estimation procedure is producing adequate results even for relatively large values of σ_ε. Two features are evident. First, the standard errors associated with estimates of p and b are markedly larger than for the other estimates. This is inherent in the linear formulation that we have adopted, and bears out the observations that we have made previously. Secondly the estimate of α differs from the true parameter value by over 10% for $\sigma_\varepsilon = 50.0$, despite the small standard error. This is a reason for mild concern, but the true and reconstructed ellipses actually only differ to a very minor degree, and are virtually indistinguishable visually, as can be seen from figure 95. From a representational aspect, therefore, the estimation procedure is sufficiently accurate. In terms of explanation, important in the area of object detection, the location and dimension parameters are estimated adequately. Thus, we have reasons to be satisfied to a degree with the procedure.

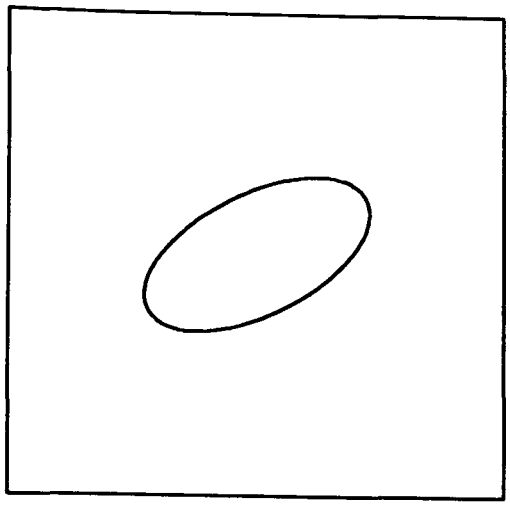


Fig 95(a) : true ellipse

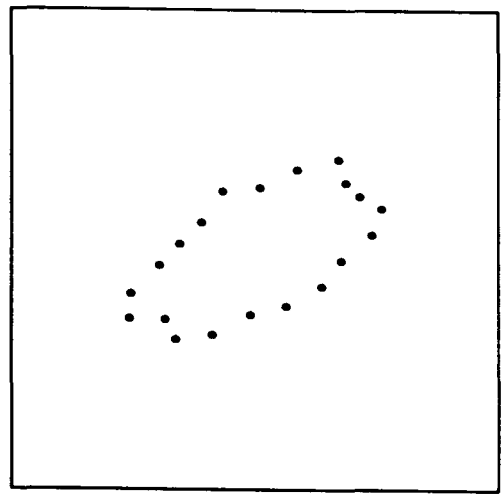


Fig 95(b) : simulated edge-points

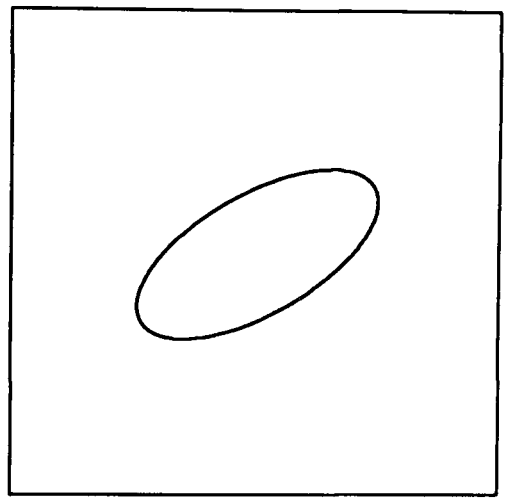


Fig 95(c) : reconstructed ellipse

On closer inspection of table 2, and further experimentation with larger values of σ_ε, other trends become apparent. For instance, the estimates of parameters a and e (defined naively as the ratio of the estimate of b to the estimate of a) respectively increase and decrease monotonically with σ_ε - this appears to be true for ellipses of general dimension and orientation. We do not discuss the reasons for such trends here - see Forbes (sections 2.5 and 2.6) for a full discussion of the behaviour of the model and estimates. We merely comment that such a feature is typical of least-squares estimation procedures, which, as we mentioned above, are not robust to the presence of outliers, and thus potentially unreliable in situations where the data is subject to high-levels of noise-corruption. Thus we might take steps to robustify the procedure, such as those described in the previous section. However, we also notice that the estimates of the parameters p and q are usually accurately determined by the procedure for these data sets - we shall see later that the presence of outliers having a different origin (spurious points not resulting from corruption of the ellipse itself) does, unfortunately, adversely affect the estimates of these parameters also.

We now investigate the adequacy of the estimation procedures when the data is not uniformly distributed around the edge of the ellipse. We could choose to investigate the case where the set of (x,y) values are simulated using (6.19) and a set of θ values randomly (rather than deterministically) chosen in the $(0,2\pi)$ interval. We prefer, however, to concentrate on situations more relevant to our own particular interests. For example, we saw in figure 15 that commonly only one section or arc of the object edge is detected using changepoint techniques. Also, in figure 19, we saw that it is possible to detect two disconnected arcs of the object edge, and these arcs may coincide with the major or minor axes - we have already noted the need for sufficient data in these regions. It is straightforward to simulate data for each of these situations.

We begin with the single arc problem. First, we consider the case when the points cover precisely half of the perimeter of the ellipse. Figure 96 depicts three such sets of points (with one arc in the direction of the major and minor axis covered in (a) and (c), and an intermediate case in (b)), and the resulting ellipse reconstructions. The true ellipse parameters used were identical to those above. Again, 20 data points were generated, σ_ε was nominally chosen to be 20.0, and the linear least-squares model (6.16) was fitted.

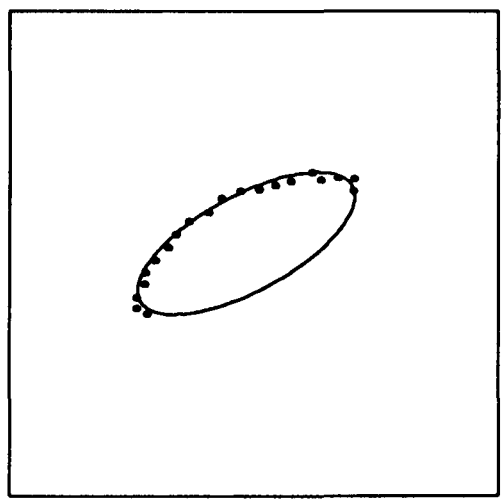


Fig 96(a) : major axis

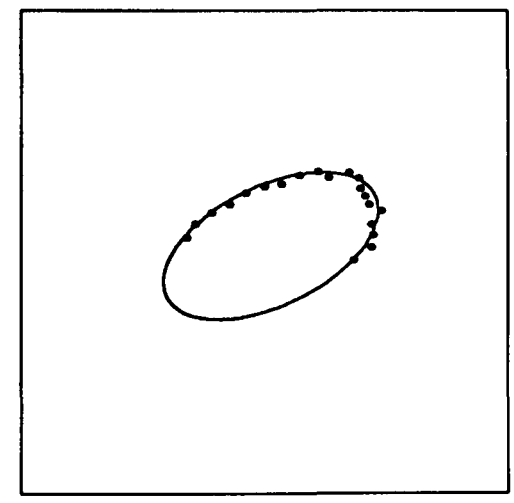


Fig 96(b) : intermediate

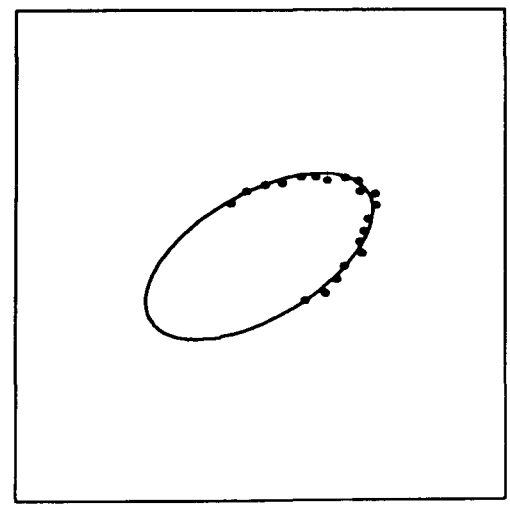


Fig 96(c) : minor axis

The parameter estimates, standard and mean square errors presented in table 3.

	(a)	(b)	(c)
p	38.745 (0.325)	40.921 (0.334)	39.844 (0.403)
q	41.733 (0.277)	41.044 (0.346)	39.798 (0.261)
a	19.759 (0.509)	19.027 (0.372)	20.624 (0.362)
b	7.863 (0.443)	9.845 (0.773)	10.240 (0.254)
α	0.499 (0.001)	0.455 (0.004)	0.525 (0.002)
m.s.e	11.25	12.83	12.83

Table 3

The most noticeable difference between the three reconstructed ellipses lies in the estimates for the b parameter, which varies by as much as 30% relative to the true parameter value. Inspection of the estimates and standard errors reveals an unsurprising picture - we estimate the length of the major axis more accurately in (a), the minor axis in (c), and the standard errors are generally larger in (b). The estimation procedure is performing as we would have predicted. Note also that the mean square error for the fit in (a) is smaller than for either (b) or (c) - thus, as we would generally regard (a) as an inferior reconstruction to both (b) and (c) relative to the true parameter values, we must ensure that we take great care over the inferences we draw from such goodness-of-fit statistics.

We now investigate the situation where the data points extend over a smaller arc of the perimeter of the ellipse. For demonstration purposes, we consider an example where this arc partially includes one end of the minor axis of the ellipse, the parameters of which are identical to those used in the previous example. For this example, σ_ϵ was chosen to be 20.0, and 20 simulated points were generated. The results of fitting the linear least-squares model to each of three data sets (corresponding to arcs covering 2/5, 1/3 and 2/7 of the ellipse perimeter respectively) are depicted in figure 97.

These results are generally unsurprising - note how the location of the reconstructed ellipse differs increasingly from the true location as the fraction of the perimeter over which the data points extend decreases, and how the estimate of parameter a is generally quite good. For smaller fractional coverage, estimates of parameters a and b became unobtainable as the points tended to collinearity. In the edge-reconstruction/object-detection context, therefore, we might only feel it necessary or of interest to fit an ellipse through any edge-point data set if we

have a sufficient number of points in two perpendicular directions.

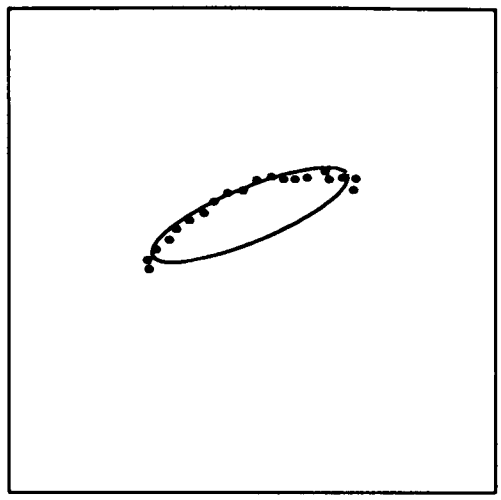


Fig 97(a) : 2/5 coverage

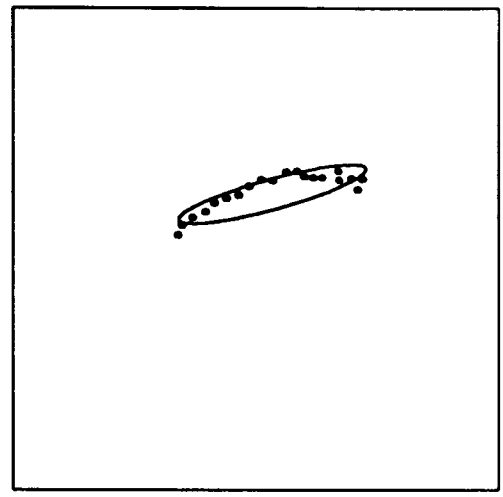


Fig 97(b) : 1/3 coverage

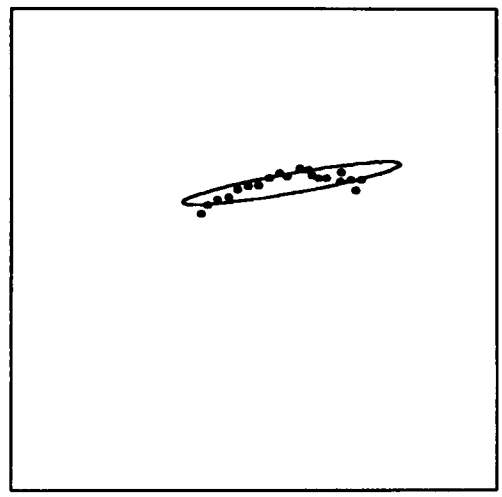


Fig 97(c) : 2/7 coverage

Finally, we investigate the situation in which the data points extend over two disconnected arcs of the ellipse perimeter. Using the identical ellipse the the one used in previous examples, with $\sigma_\varepsilon = 20.0$, 30 data points were generated so to cover two (diametrically opposed) arcs, with a total angular coverage of half the ellipse perimeter. The data points are depicted in figure 98.

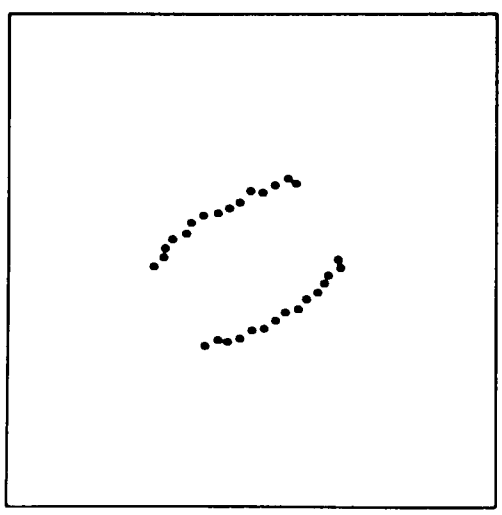


Fig 98(a) : major axes

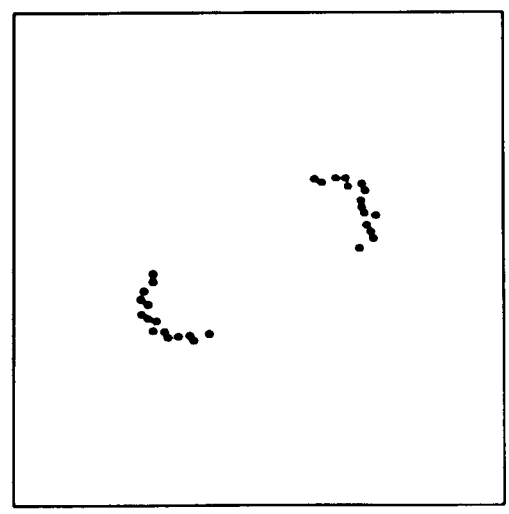


Fig 98(b) : minor axes

These data sets are included for the following reason - no fit of an ellipse was possible using the identical NAG routine for any of the three, with one or other or both of the parameters a and b being impossible to estimate (or being improbably large). The reasons for this are evident - again, we must obtain sufficient data in the two relevant perpendicular directions.

We have attempted to explore and expose the limitations and fallibilities of our ellipse reconstruction technique. We conclude that it is essential to have data points covering large arcs of the ellipse perimeter to ensure that the ellipse is adequately reconstructed - if this is the case, then the linear least-squares formulation returns acceptably accurate estimates. This indicates that it is absolutely necessary for such a reconstruction to be possible that a two changepoint posterior distribution or one of the modifications of the single changepoint posterior forms that we discussed should be used, rather than merely the single changepoint form itself. Fortunately, when using the two changepoint form or, for instance, binary segmentation at a sufficiently high Signal-Noise ratio, it is generally possible to obtain edge-points over large arcs of the perimeter of the object - see, for example, figure 19 on p. 62.

Finally in our discussion of the analysis of simulated edge-point data relating to ellipses, we examine the behaviour of the linear least-squares formulation when presented not only with noisy data arising from approximately accurate edge-point classifications, but with other, spurious and outlying data points. Such data could arise from insufficient smoothing of the results of a changepoint analysis of the image, or indeed as a result of the noise present in the image itself. We now discuss the effect of introducing such spurious edge-points, and possible techniques for reducing their impact.

(6.3.4) Removal of spurious edge-points.

First, we consider the introduction of a single spurious point, $s = (x_s, y_s)$ say, into the data set. For demonstration purposes, we generate 15 edge-points placed uniformly around the perimeter of the familiar $(40, 40, 20, 10, 0.5)$ ellipse, with $\sigma_\varepsilon = 20.0$, and nominally choose s to be at $(15, 10)$. Figure 99(a) depicts the entire data set, and (b) and (c) depict the resulting reconstructed ellipses with s omitted and included respectively.

The effect of including the single point is immediately apparent - the ellipse has been elongated and translated toward the lower left-hand corner of S. The point parameter estimates represented in (b) and (c) were $(40.29, 40.03, 19.88, 10.40, 0.53)$ and $(36.79, 36.56, 31.04, 8.95, 0.70)$ respectively. The respective mean square errors were 11.42 and 67.08. Again, the lack of robustness of the least-squares estimation technique is demonstrated. Further examples with other outliers are depicted in figure 99(d), (e) and (f).

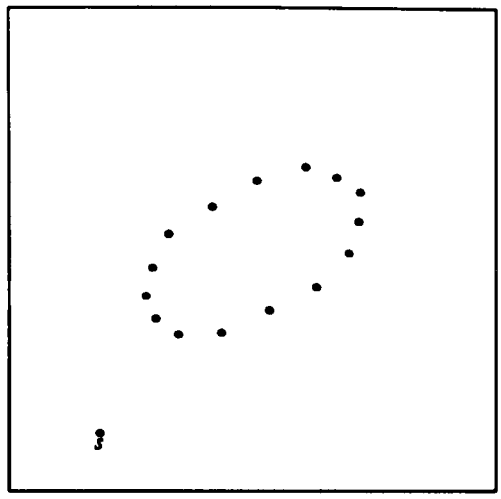


Fig 99(a) : data set

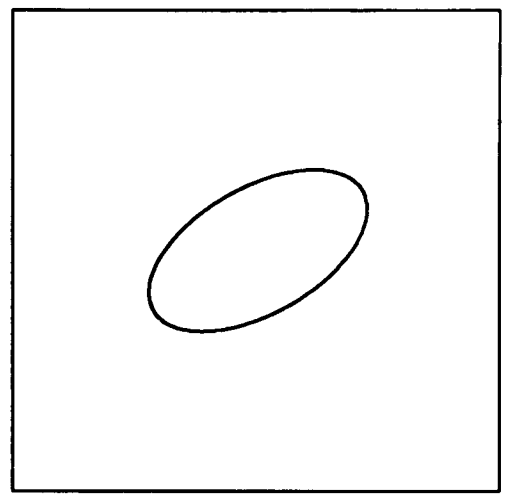


Fig 99(b) : s omitted

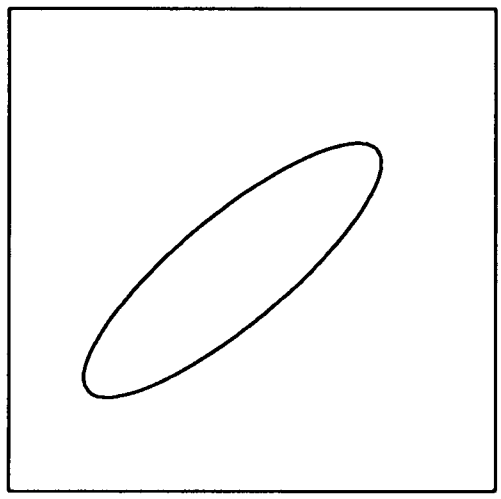


Fig 99(c) : s included

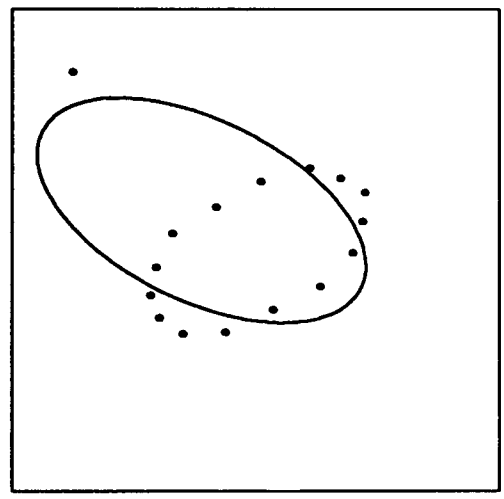


Fig 99(d) : $s = (10, 70)$

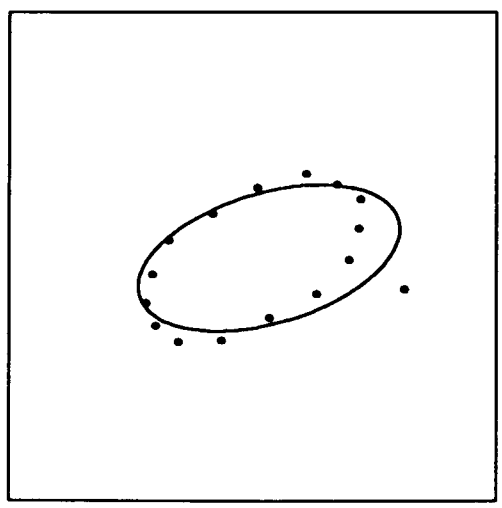


Fig 99(e) : $s = (65, 35)$

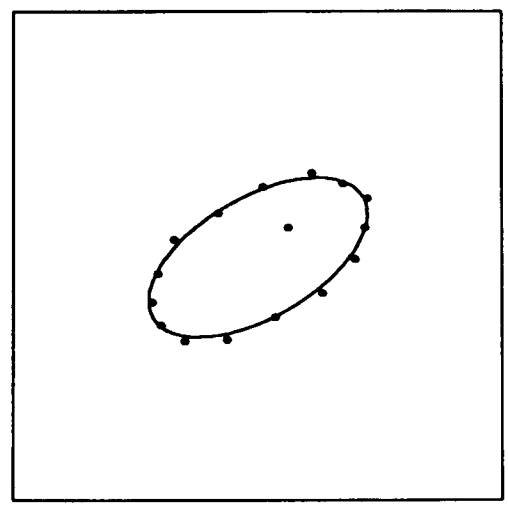


Fig 99(f) : $s = (45, 45)$

These examples illustrate the degree to which various outlying values can affect the reconstructed ellipse, and it is evident that the presence of spurious elements in the data set is detrimental to the least-squares estimation procedure. We now discuss three approaches intended to overcome the problems posed by the presence of such spurious data points.

First we could ensure that adequate smoothing is carried out, using the probabilistic and spatial ideas we have discussed at length previously. To automate this process, to make it independent of visual supervision (one of our goals), and still reject all outlying/mis-classified points is a relatively relatively complex task, however - we could never feel completely secure that the automated procedure has performed sufficiently adequately.

Secondly, as indicated on several occasions above, we could take steps to robustify the least-squares procedure. For example, instead of using a least-squares criterion, we could use the L_1-norm or minimum absolute distance criterion. It is possible to fit the linear model given in (6.16) and using an L_1-norm procedure automatically using NAG library routine E02GAF. Figure 100 depicts the ellipses reconstructed from the data sets in figure 99(d), (e), and (f) under an L_1-norm criterion.

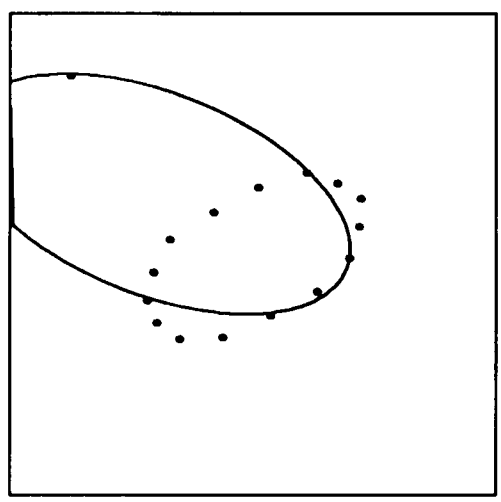


Fig 100(a) : s = (10,70)

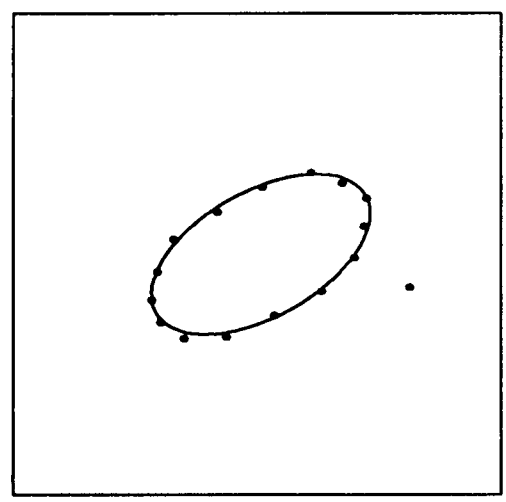


Fig 100(b) : s = (65,35)

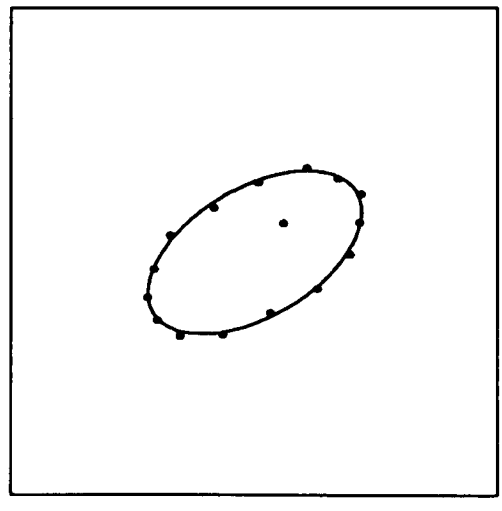


Fig 100(c) : s = (45,45)

We have obtained an improved fit in case (b) (this was so for the data set in figure 99(a) also), and the fit in case (c) is as adequate as the previous one. However, the reconstructed ellipse in case (a) is still utterly different to the true ellipse. It is relatively straightforward to construct other such examples, even using single spurious data points. It should be noted that if

the number of "legitimate" data points is increased the L_1-norm solution can eventually coincide with the true ellipse - in case (a) here, 30 points are sufficient. This is not the case for the least-squares solution, which will always be influenced by spurious data points, albeit to a lesser degree as the number of legitimate points increases - in case (a), the ellipse based on 30 legitimate data points and the single outlier is still markedly different from the true ellipse. The other robustifying techniques (using (6.7) and (6.8)) may also be implemented, but in our experience this proves to be laborious, difficult to automate, and usually ineffective.

One other point is important here. Although we may try automatically to smooth the data and subsequently obtain robust estimates from them, we still have no diagnostic (other than our own visual assessment) that would allow us to judge whether the reconstructed ellipse is an adequate representation of the data - we have already noted the unreliability of using mean square errors to assess absolute goodness-of-fit - or whether any spurious and outlying points remain in the data set. The third of our approaches attempts to address each of these problems using techniques for outlier identification. We discuss one classically-based procedure and one Bayesian technique for the removal of outlying edge-points. First, we must introduce a geometric algorithm that is of assistance in their implementation.

(6.3.4.1) Convex-hull peeling.

Consider the nature of outlier detection problems in the ellipse reconstruction context. Generally, we have a large number of points scattered around the edge itself, and rather a smaller number of spurious points distant from the edge. One standard form of analysis, namely cross-validation, requires that each data point in turn be omitted from the data set to assess its importance (as measured by some (generally classically based) inferential quantity - mean square error, for example). One major criticism of this approach, however, is that it is extremely time-consuming - of order $n!$ for a set of n points if the number of outlying points is unknown. In the context of ellipse reconstruction, experience informs us that spurious points are most detrimental when they lie externally to and distant from the ellipse - clearly, therefore, we ought try and remove these "external" points as early as possible. This provides our motivation for the use of the technique of convex hull peeling.

Convex-hull peeling proceeds as follows. Consider a set P_0 of points in the plane (we consider the 2-D case explicitly - the extension to spaces in other dimensions is straightforward). Fit a relevant model to the entire data set. Form the convex hull h_1 of the elements of P_0, and then form a subset H_1 of P_0, the elements of which being those elements in P_0 lying on the boundary of h_1. Then assess the individual importance of each element $e_{H_1 i}$ of H_1 to P_0 by evaluating some diagnostic quantity derived from the model obtained after fitting $P_0 \setminus e_{H_1 i}$. After all of the elements of H_1 have been addressed, form a new set of points P_1 given by $P_1 = P_0 \setminus H_1$. Repeat the above procedure, by forming h_2 and H_2 defined accordingly.

Such a scheme allows us to assess the relative importance of points in the data set individually from the most "external" inwards. We suggest two possible diagnostic quantities. Using a classical approach, we could compare the effect of omitting each point on overall mean square error, or on individual parameter estimates and standard errors. From a Bayesian standpoint, we could examine the degree to which omitting each point alters the posterior distribution for the set of ellipse parameters, for example using the ideas of Pettit and Smith (1983). In either case, it should become apparent which points are more important than others, or when an adequate fit is perceived. Other details will become evident when we provide examples below.

Clearly, for us to be able to implement such an approach, we must readily be able to form the convex hull of any given set of points. This is a non-trivial task, but fortunately there is the widely available algorithm of Green and Silverman (1979) devised precisely for this purpose which can be easily and efficiently implemented. Indeed, the related work of Silverman and Titterington (1981) on minimum covering ellipses is also potentially of use, and we shall see one application of this work at a later stage. The precise details of the convex hull peeling algorithm are given in the 1979 paper - here we merely concentrate on application to outlier removal.

Consider the data set in figures 99(e) and 100(a), with single outlier at $s = (10, 70)$. We now attempt to apply the ideas discussed above and the convex hull peeling algorithm in an attempt to obtain an adequate fit. Figure 101 shows how the results develop - figure 101(a) depicts the data itself, and (b) the boundary of the convex hull of the points. The points are numbered in the order they are to be omitted - the ordering is established by the Green-Silverman algorithm, and takes the points in an anti-clockwise direction, starting at the point with largest x-coordinate. Our technique for outlier detection proceeds as follows. For each of the points 1 to 11, we obtain estimates and standard error for the ellipse parameters and the corresponding mean square errors by fitting the linear model (6.16) using a least-squares criterion to the remaining 15 data points. The mean square error arising from the fit of the complete data set was 184.83.

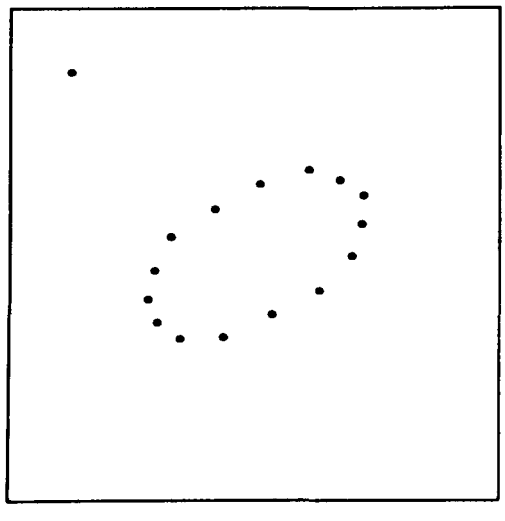


Fig 101(a) : data set

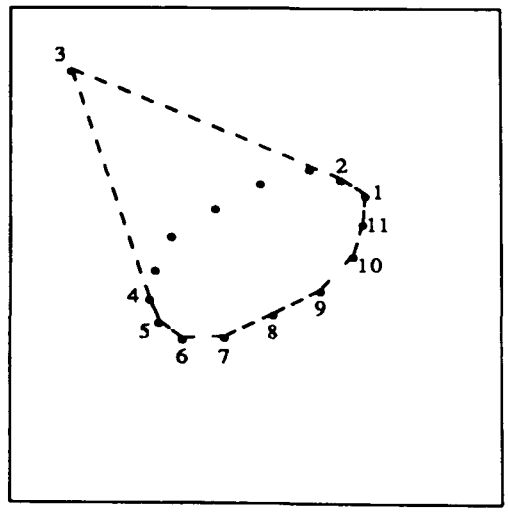


Fig 101(b) : boundary of convex hull

For demonstration purposes, we adopt a classical approach, and assess model adequacy purely via the value of the mean square error obtained. The mean square errors to the nearest integer are displayed in table 4.

Point omitted										
1	2	3	4	5	6	7	8	9	10	11
180	185	11	190	182	176	188	189	187	190	190

Table 4

Clearly, omitting point 3 changes the mean square error radically, whereas omitting the other points individually has little effect. Thus we have successfully identified the outlier as the point which most significantly alters the goodness of fit of the model. Once point 3 is omitted, the mean square error and parameter stabilise, and no other points are diagnosed as outliers on subsequent iterations (all fifteen remaining points lie on the boundary of the convex hull on subsequent iterations, and the mean square error is not changed radically from the 11.42 value quoted above). Figure 102 (c) depicts the reconstructed ellipse using the remaining fifteen points.

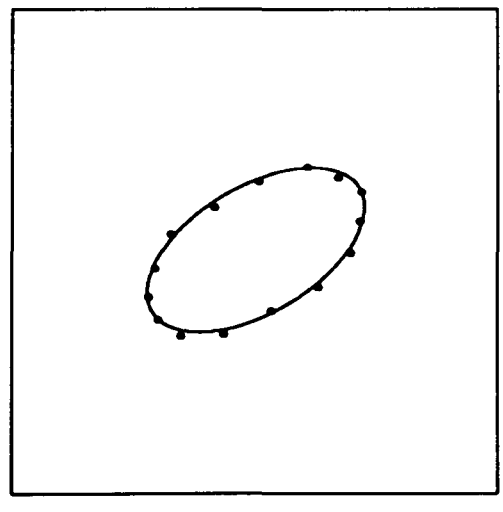


Fig 101(c) : remaining points and fit

This data set might be claimed to be relatively untypical, with the spurious data point being quite contradictory in its position relative to the remaining data, and hence clearly very influential. In situations when the spurious point is reasonably close to the rest of the data, the technique might not perform as adequately. However, on experimentation with various different positions for s, we find that the technique actually detects the outlying point (as the single point that changes most radically the mean square error of the fitted model) in situations when s is in close proximity to the rest of the data. Some examples are given in figure 102.

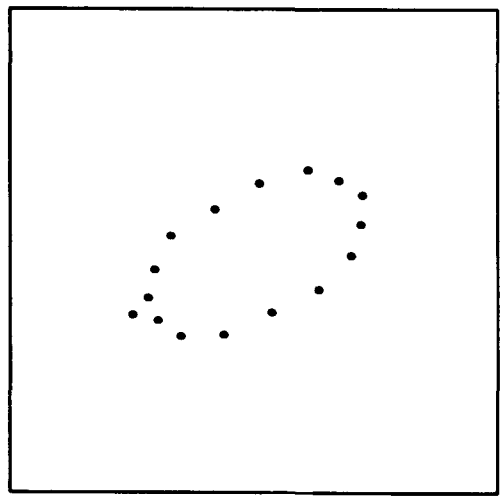


Fig 102(a) : $s = (20,30)$

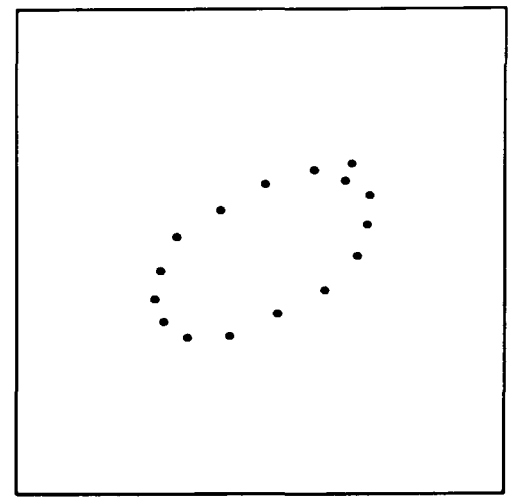


Fig 102(b) : $s = (55,55)$

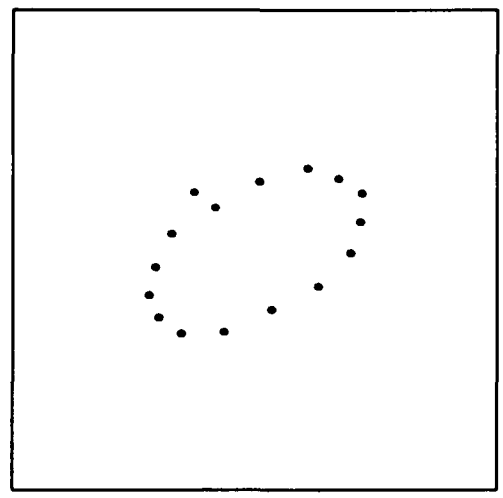


Fig 102(c) : $s = (30,50)$

In each case, the spurious point lies close to the others. The mean square errors resulting from a fit of the linear model (6.14) were, respectively, 15.41, 15.57 and 29.27. After using the convex hull peeling technique and inspecting mean square errors, the spurious points were identified on the first iteration, with the relevant mean square errors being reduced to the 11.42 value, and with the fit not being improved significantly by the removal of other points on subsequent iterations (here we do not use the term "significant" in the statistical sense - although clearly we could impose such an interpretation onto the difference in mean square error values, and indeed shortly will do so - but merely as a qualitative statement). We note in passing that for each of the data sets depicted in figure 102, the parameter estimates obtained were actually quite adequate before the spurious point was removed - however, for these examples, we are in the privileged position of knowing the true ellipse parameter values, and this of course is practically unrealistic. We might adopt the attitude to continue removing points until some facet of the resulting fit, for example mean square error, stabilises. Clearly, however, this approach is not completely satisfactory, as although generally the spurious point is accurately identified, the re-fitting of the model on subsequent iterations is time-consuming.

In the examples given above, we have merely concentrated on the case of a single outlier. We now give one further brief example to investigate the performance of the detection technique in the multiple outlier case, after making some relevant comments. The nature of the new problem is clearly sightly different to that of the old. Previously we geared the identification technique to the single outlier problem by omitting each point on the convex hull individually in turn and subsequently replacing it. In the multiple outlier case, a more appealing approach would be to discard a point completely (never re-include it at any later stage of the ellipse fitting analysis) once it is identified as an outlier. Using this method the total amount of computation required should be lessened. Such an approach is easily implemented with a minor adjustment to the algorithm described above. Our decision to label a point as an outlier or otherwise must now be made "on-line" rather than retrospectively, but this merely calls for a minor adjustment to our original approach.

We use such an approach in the analysis of the data set in figure 103, which contains three spurious points at (60,10), (35,60), and (10,50) amidst the 15 legitimate points lying uniformly around the perimeter of the ellipse, corrupted using $\sigma_\varepsilon = 20.0$. A fit of the linear-model under a least-squares criterion resulted in point parameter estimates of (42.84, 37.47, 31.47, 14.66, −0.60), and a mean square error of 190.36.

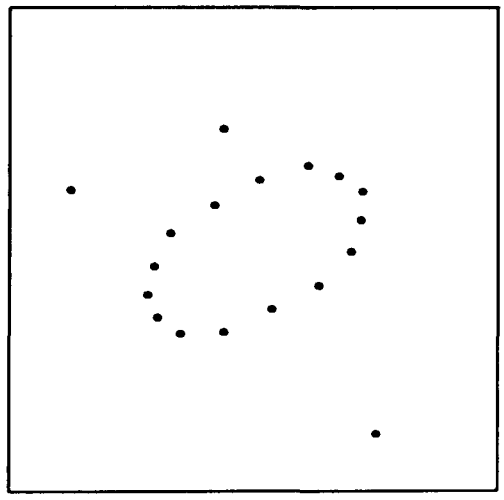


Fig 103(a) : data set

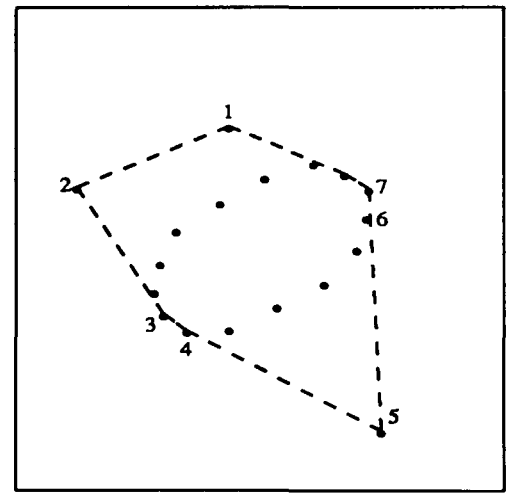


Fig 103(b) : 1st iteration

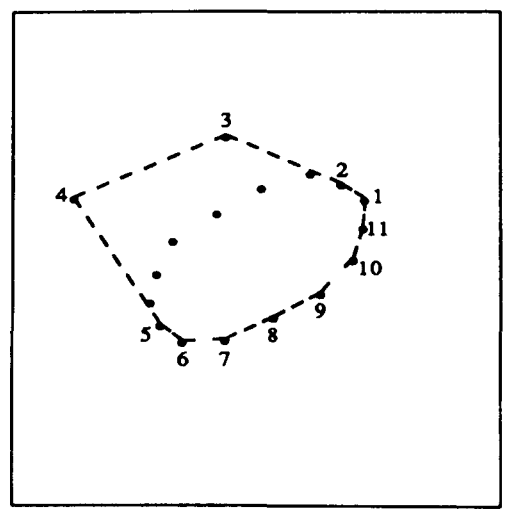


Fig 103(c) : 2nd iteration

Using the convex hull peeling technique, omitting each point in turn and re-fitting the model, on the first iteration we obtain the mean square errors contained in table 5. It is clear that the fit is improved most radically when point 5 is omitted, and thus as usual proceed with the second iteration as depicted in figure 103(c), and subsequently remove all the outlying points. However, we also note that the fit is significantly improved when point 2 is omitted (again we use the term "significantly" in a qualitative sense) - thus it might be advantageous to diagnose point 2 an outlier and subsequently remove it from the analysis.

	Point omitted - 1st iteration						
	1	2	3	4	5	6	7
m.s.e.	191.91	171.89	189.08	190.33	121.15	184.69	187.65

Table 5

The amended scheme thus proceeds as follows - using the convex hull boundary in figure 103(b), we begin to peel points until we note that omitting point 2 reduces the mean square error to 171.89, which we regard as a significant reduction. We consequently discard point 2, and continue peeling points on the same convex hull, now comparing the new mean square errors to this value. This continues until we reach point 5, which, when omitted, effects a further reduction of the mean square error to 77.07. We consequently discard point 5 also. No other points are discarded on the first iteration. On the second iteration, the mean square error is only reduced drastically when point 3 is omitted, when the mean square is reduced to the 11.42 value we obtained earlier. Thus we obtained identical conclusions in fewer iterations (and fewer fits of the model) than when using the previous approach.

The techniques described above are straightforward to implement in a supervised setting. However it is our desire that the such techniques should be automated, and thus we must precisely define the way in which we decide to discard points. We require a notion of statistical significance to relate to our goodness-of-fit statistic (this seems more relevant here in the automatic case than in the supervised case). Problems relating to robustness and influence for regression models such as that in (6.16) have been widely studied in a classical statistical framework, and many test statistics and diagnostics have been proposed; see, for example, Cook and Weisberg (1982) for a comprehensive survey. It is clear that simple criteria for the discarding of points in the edge-point data set automatically can be introduced.

(6.3.4.2) Bayesian detection of influential observations.

We might regard the outlier detection problem to be almost equivalent (contained within) the following broader problem. Given a set of data points resulting from an informative experiment concerning a parameter of interest, locate elements in that set which most affect the inferences that we make. We might consequently wish to identify these important or "influential" points as possible outliers. In a Bayesian setting, one sensible measure of the influence of a point (or, more generally, subset of points) can be derived in relation to the change in the posterior distributions for the parameters in the proposed model when that point is omitted. For instance, Pettitt and Smith (1983) suggest a Bayesian influence measure for inference about the mean parameters in Normal linear models based on the Kullback-Liebler distance between the respective posterior distributions. They show that, under a non-informative joint prior specification for the mean parameter θ and with measurement error variance σ^2 known, the influence measure for the i'th observed value s_i can be written (Pettitt and Smith, equation (3.11))

$$I(\{i\}) = \frac{1}{2}\frac{v_i}{(1-v_i)}\left[v_i + (2-v_i)\,t_i^2\right] \tag{6.20}$$

where $v_i = \boldsymbol{j}_i^T(\boldsymbol{J}^T\boldsymbol{J})^{-1}\boldsymbol{j}_i$, and t_i is given by

$$t_i = \frac{(s_i - \boldsymbol{j}_i^T\hat{\theta})}{[\sigma^2(1-v_i)]^{1/2}}$$

where $\hat{\theta}$ is the usual least-squares estimate of θ, and $\boldsymbol{j}_i^T$ is the i'th row of the design matrix $\boldsymbol{J}$. The value of t_i, the "Studentised residual" for point i, is commonly used as an intuitive measure of "outlyingness".

Thus, in (6.20), we have the basis of a technique for outlier detection, the details of which similar to the non-Bayesian technique described above. Now, instead of evaluating mean square errors etc. and discarding points on the convex hull which improve the global fit of the model, we now measure the influence of individual points (or sets of points) on the convex hull using (6.20) (or the equivalent formula) and discard points accordingly. It is clear, however, that to evaluate (6.20), we must know σ precisely. Pettitt and Smith note the robustness of (6.20) (in terms of the ordering of possible outliers) to changes in σ, and this is our experience also. Thus, subsequently, we set $\sigma = 1.0$ for demonstration purposes.

For one brief illustration, we will use the multiple outlier data set in figure 103. Table 6 contains the value of the Bayesian influence measure in (6.20) ($\times 10^{-5}$) evaluated for the points on the convex hull in the first iteration. It is clear that point 5 is the most influential

point and thus is correctly identified as an outlier. On the next two iterations, the other two spurious points are correctly identified.

	Point omitted - 1st iteration						
	1	2	3	4	5	6	7
m.s.e.	0.021	0.178	0.032	0.033	10.22	0.246	0.096

Table 6

Again, this technique is reasonably straightforward to implement in a supervised fashion, but more complicated to implement automatically. Here though, (6.20) actually measures some distance between posterior distributions, and thus we might feel happier in pre-fixing some lower threshold below which we do not seek to reject any further points (although we might actually require this threshold might to depend on the value of σ).

We have discussed various models, estimation procedures, robustness measures etc. for simulated edge-point data, intended to investigate performance in the reconstruction of edges in single convex object true scenes. We now present some examples where the data set is actually derived from the changepoint based analysis of such scenes.

(6.3.5) Analysis of true edge-point data.

We begin by making a three relevant comments. First, as we saw in section (3.3.2) of chapter 3, binary segmentation seems the most sensible technique to use for the analysis of single convex objects, and the problem of dealing with edge-points arising as posterior modes for sequences of different length is avoided as we (can) treat each recorded point equally. Secondly, if we use the binary segmentation technique then, except for ellipses having eccentricity near zero, we are generally assured of obtaining edge-points that are distributed completely around the perimeter of the ellipse. Thirdly, because of the difference in order of magnitude between pixel and object scales (for the type of objects in which we are interested), we obtain a large number of correctly recorded edge-points relative to the number of outlying or mis-classified points. Thus, we might regard the simple smoothing of results as adequate for removal of such points, rather than using the other techniques described above.

Our first example is based on the results of analysis of an image derived from a true scene in which the ellipse is identical to that in our simulated examples, that is defined by the values $(40, 40, 20, 10, 0.5)$ for the various parameters. Under assumption of an image-formation process identical to that in (2.1), the image was derived from the true scene with Signal-Noise ratio 1.5, and then was analysed using the binary segmentation technique based

on the one changepoint posterior distribution (2.11). The raw results of a full analysis, and the smoothed results using a simple smoother (chosen nominally, without reference to the recorded points) are depicted in figure 104(a) and (b) respectively. Figure 104(c) depicts the reconstructed (solid) and actual (dotted) ellipses. The reconstruction was obtained via a fit of the linear model (6.16) under a least-squares criterion.

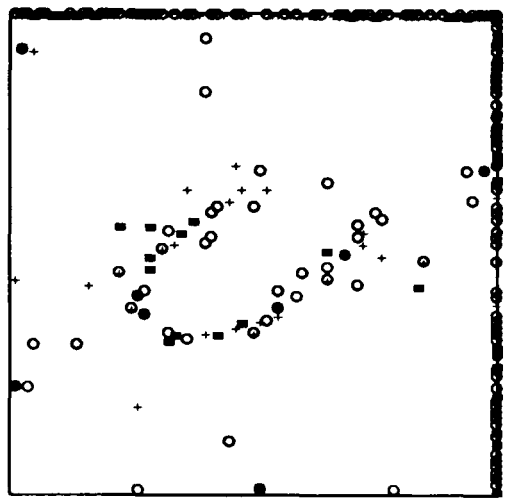


Fig 104(a) : raw results

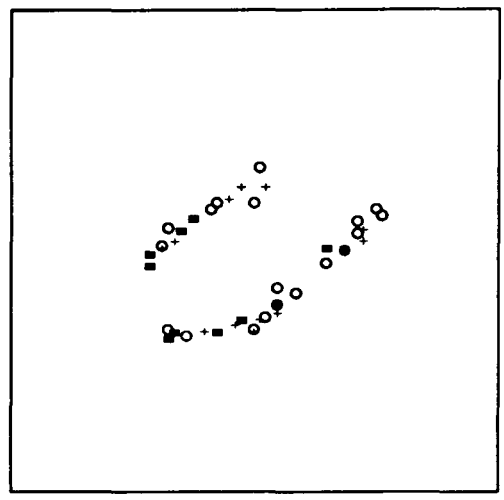


Fig 104(b) : smoothed results

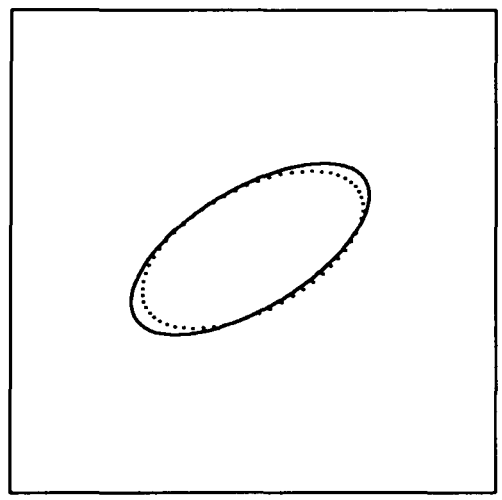


Fig 104(c) : reconstructed ellipses

The parameter estimates (standard errors) obtained from the least-squares fit were $(39.56\,(0.72), 40.21\,(0.54), 22.10\,(0.47), 9.86\,(0.63), 0.54\,(0.005))$ for the vector of five parameters (p, q, a, b, α) respectively, and the mean square error was 35.68 for the 28 points fitted. Clearly, the reconstructed ellipse is slightly elongated (due to the lack of points at the ends of the ellipse on he major axis), but otherwise the reconstruction seems adequate.

Our second example is based on the analysis of an image derived from a true scene with an ellipse defined by the parameters $(20, 45, 15, 15, 0)$, i.e. a circle. Figure 105 depicts the results of a full binary segmentation analysis and the subsequent ellipse reconstructions. The Signal-Noise ratio in the image was 1.5.

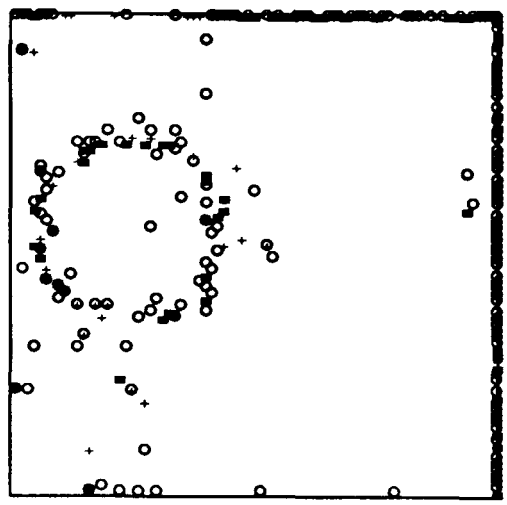


Fig 105(a) : raw results

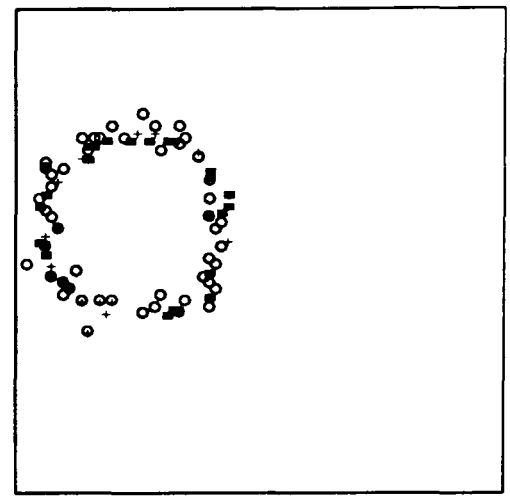


Fig 105(b) : smoothed results

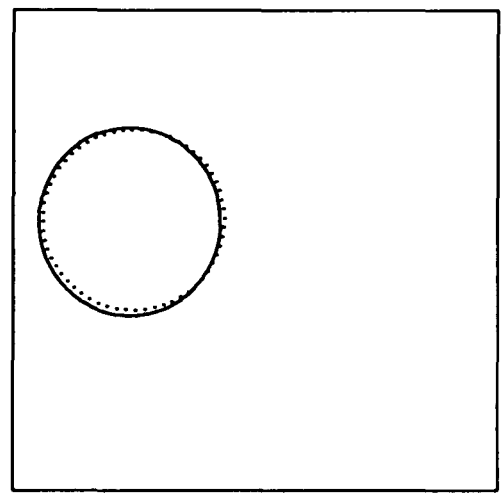


Fig 105(c) : reconstructed ellipses

The parameter estimates and standard errors obtained by a fit of the linear model under a least-squares criterion were

$$(p, q, a, b, \alpha) = (19.30\,(0.23), 44.72\,(0.24), 15.68\,(0.17), 15.01\,(0.83), 1.64\,(0.089)) .$$

The reconstruction is again quite adequate, with the reconstructed ellipse resembling the true one very closely.

Our final example is based on the analysis of an image derived from a true scene with an ellipse defined by the parameters (40, 40, 50, 5, 0.7), i.e. eccentricity 0.1, reasonably extreme. Figure 106 depicts the results of a full binary segmentation analysis and the subsequent ellipse reconstruction. The Signal-Noise ratio in this case was 2.0.

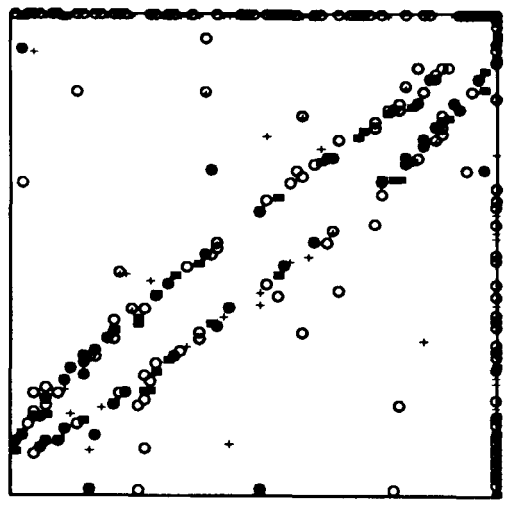


Fig 106(a) : raw results

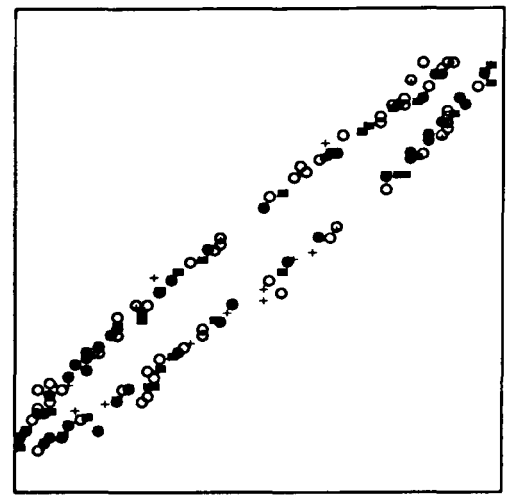


Fig 106(b) : smoothed results

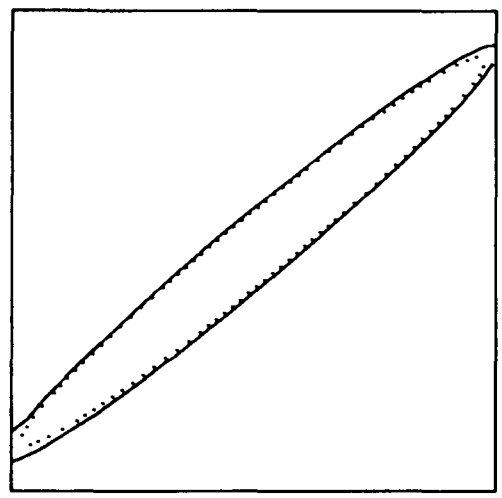


Fig 106(c) : reconstructed ellipses

For this data, the parameter estimates and standard errors were as follows -

$$(p, q, a, b, \alpha) = (39.60\,(0.83), 39.64\,(0.71), 53.46\,(0.67), 5.42\,(0.05), 0.70\,(0.0001))$$

The true and reconstructed ellipses are virtually indistinguishable, although again the most noticeable discrepancy between the two lies in the lengths of the major axes - the estimate of a is again larger than the true value. The reason for this general trait in the reconstructed ellipses is that, for ellipses of even only mild eccentricity, the changepoints in any given row are close together in regions of the ellipse near the ends of the major axis. We might attempt to overcome the discrepancy between true and estimated parameter values by means of some empirically determined discount factor, or by ensuring that the analysis is carried along a pair of perpendicular directions parallel to the axes of the ellipse (and thus accurately determining the position of edge-points at the ends of the ellipse).

The typical examples above confirm that the ellipse estimation procedures perform adequately on edge-point data arising from changepoint based analysis of simulated images derived from single convex object true scenes, where the objects themselves are regarded as ellipses, and of reasonable dimensions as to make a standard changepoint based analysis applicable. Finally in this section we concentrate specifically on a problem in object detection in which we see that it is convenient to assume that the objects can be regarded as ellipses, but are small in relation to the size of region S, and where there may exist more than one object in any given true scene.

(6.4) Multiple object detection - The Tank Spotting problem.

In the majority of practical examples that we have seen above, we have been interested in the detection of localised boundaries between large-scale, homogeneous texture regions, occurring in simple or composite true scenes. We now study a problem of a different nature. Consider the case where the true scene contains a (possibly unknown) number of small (relative to the dimensions S_θ) convex objects. Then the related inference or detection problem might involve discovery of the number, locations and dimensions of these objects from a noise-corrupted version of the true scene. Further, the true scene might arise as one "frame" in a "film" of moving objects, and thus we might also wish to discover the associated velocities. Such problems commonly arise in medical imaging - tumour detection from Gamma camera images - or in a military environment - detection of land vehicles from satellite photographs. It is from the latter of these that the name of this particular inference problem - "Tank Spotting" - is derived (albeit a misleadingly offensive label for a largely defensive technique). Figure 107 depicts a typical Tank Spotting image containing three objects centred at pixels $(20,60)$, $(30,40)$ and $(70,10)$ in the 80×80 grid.

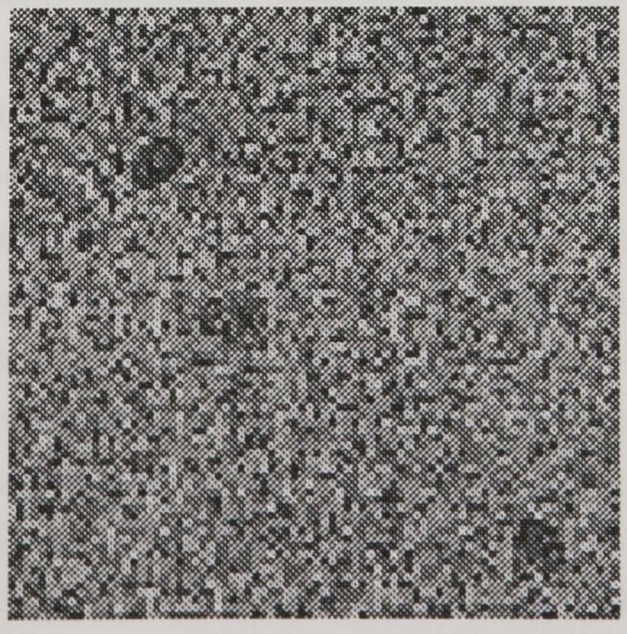


Fig 107 : multi-object image

Clearly, the nature of the problem indicates that the standard edge-detection techniques we have developed are not applicable here. Recall, however, the discussion that led to the derivation of the posterior form (3.17). There, we had interest in a two changepoint sequence with the inter-changepoint distance k small and presumed known, and where the outer segments of the sequence were derived from the same texture region (and hence contained elements that were identically distributed). Such a formulation is clearly of interest in the Tank Spotting problem, where the objects (tanks) themselves are small in relation to the size of the background (battlefield) region, and we have a degree of prior knowledge as to their dimensions.

Thus we propose to use (3.17) as the basis for a solution to the Tank Spotting problem. First, we make several comments. We noted after the simulation study of the behaviour of (3.17) in figure 25 that it was not necessary to know the value of k exactly, and that it was permissible to specify a value for k smaller than the true value and still obtain useful information from the posterior distribution. Furthermore, if we choose k in this "minimal" sense, then figure 25 informs us that the posterior mode will occur with (approximately) equal frequency across a narrow band near to (immediately following in the data sequence) the true first changepoint position - intuitively, (3.17) will return with equal probability any point for which the next $k-1$ pixels in that row are contained within the object. Thus, we are specifically in the object detection (rather than edge-detection) domain, and this should be reflected in subsequent inferential procedures. Finally, and related to the second point, when smoothing the results using the simple smoothing technique described previously, we might discard recorded points more readily than in the edge-detection problem, as we would expect a higher detected point density within an object than in the vicinity of an edge.

We now present some examples to demonstrate the implementation of an object detection technique based on posterior distribution (3.17). First, some notes on the nature of the test images. We shall investigate images derived from single and multiple object true scenes, and generally we will be interested in situations where the objects are "bright" or "hot" compared with the background (i.e. the images exhibit a relatively high Signal-Noise ratio). Also, for ease of automatic image generation, we use elliptical objects in our test images - rectangles, hand-drawn figures etc. would plainly be as adequate. The image generation procedure will thus be as follows. For each object, we pre-fix a non-extreme value for the eccentricity e of the ellipse , say 2/3, and a value for minor axis length b. We then merely choose a location (p,q) randomly in S, and an orientation α randomly in $(0,2\pi)$, and generate a suitably distributed variate as a pixel value for each pixel within the object. In the multiple object case we might like to make each object identical in dimensions. We shall use the image-formation process in (2.1) with a common noise variance, and usually induce the same Signal-Noise ratio in each object. In the multiple object case we might like to make each object identical in

dimensions and only different to each other in location and orientation.

(6.4.1) Single object detection.

We begin with an example in which the true scene contains only one object - we shall see later how such cases are of interest , for instance , in tumour detection in medical imaging. Figure 108(a) depicts an image derived from a single object true scene , where the object , an ellipse , is located at (14.12 , 42.10) (chosen automatically) , and has dimensions (9.0 , 6.0) and angle of orientation 4.48. The Signal-Noise ratio is 2.0. Figure 108(b) depicts the results of a full changepoint analysis based on the posterior distribution (3.17) , with k chosen to be 4. Figure 108(c) depicts the points remaining after use of a simple smoother - a 7×7 grid , with an acceptance criterion of 4 points per grid (we denote this (3 , 4) - 3 pixels either side of the central one - rather than (7 , 4)).

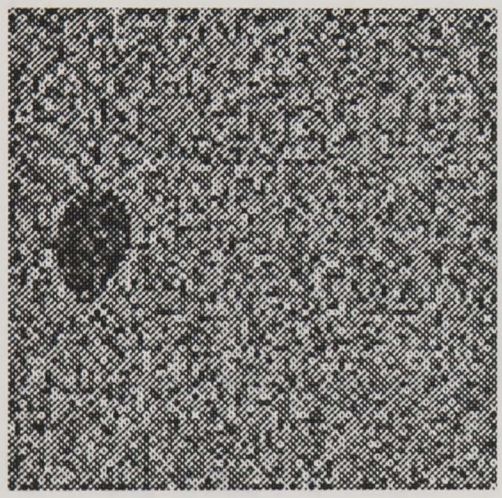


Fig 108(a) : image

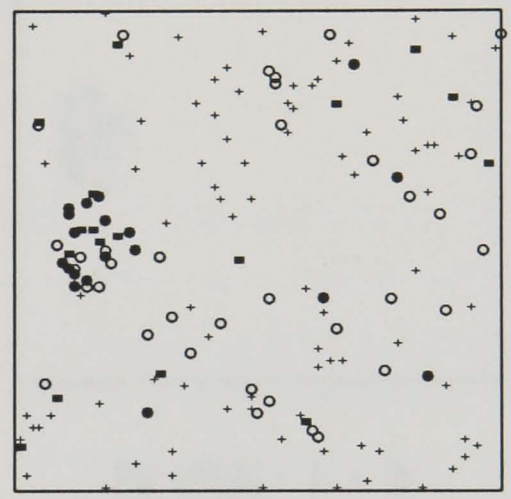


Fig 108(b) : raw results

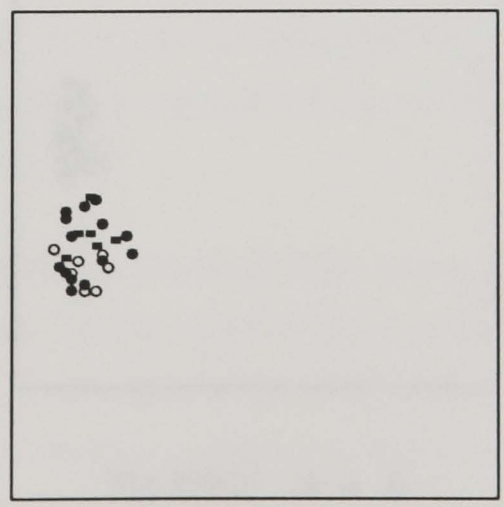


Fig 108(c) : smoothed results

In (b), there is a broad spread of recorded points across S_Y. However, there is only one region in which the density of recorded points is high, that being the region of the object. Note also that the posterior probabilities associated with recorded points in this region are higher

than for points elsewhere. The simple smoother effectively removes the large number of spurious and mis-classified points, leaving figure 108(c) a fair representation of the position and dimensions of the tumour.

The initial results , therefore , seem adequate. However , we might feel it necessary to demonstrate the robustness of the final results to the choices of smoothing window and criterion , and the choice for k. We have discussed choices of the post-processing (smoothing) parameters above - loosely , we try to accept high densities of points as objects. We now demonstrate the effect that altering k has on the final results. Figure 109 depicts the results obtained from an analysis of the single object true scene in figure 108(a) when k is chosen to take values 2 , 6 , and 8. The results have been smoothed using a (3 , 4) smoother.

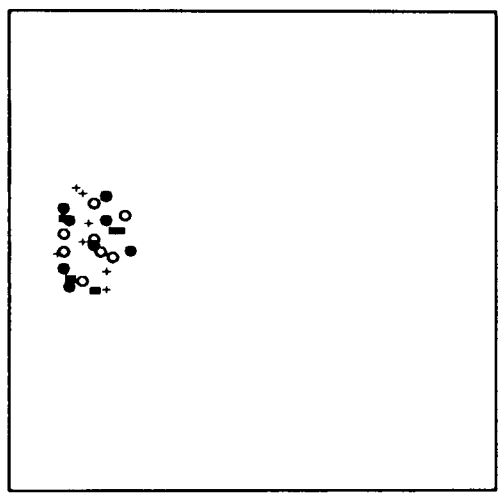


Fig 109(a) : $k = 2$

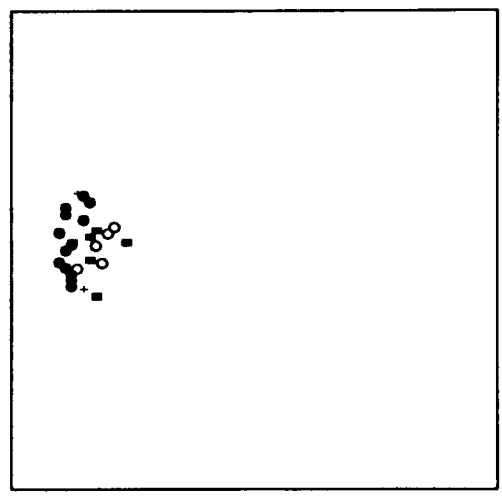


Fig 109(b) : $k = 6$

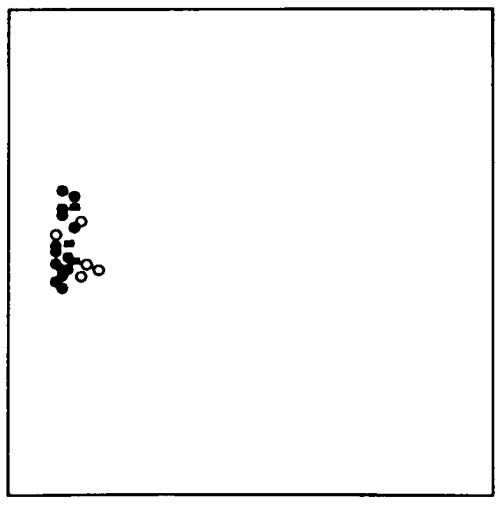


Fig 109(c) : $k = 8$

The results are broadly similar to each other, and to those in figure 109(c), and so we conclude that the technique is fairly robust to choices of k, at least in terms of the visual appearance of the results. Theoretically, however, we still regard it as preferable to choose k in the minimal sense described above.

Having obtained the results from changepoint analysis and post-smoothing, we are now faced with making inference about the location, dimensions and orientation of the object. Previously, we have used statistical methods to reconstruct, say, the edges of ellipses from sets of edge-points. Here, however, the problem is somewhat different - we are presented with a set of points, some of which we are prepared to admit as being internal to the object in question - and thus the edge-reconstruction assumptions and techniques do not seem particularly relevant. We could adopt one of several naive approaches in an attempt to draw some form of approximate inference. For example, we could report the centroid of the recorded data as an estimate of the location of the object. Less adequately, we could report the dimensions of the smallest rectangle with sides parallel to the coordinate axes that contains all of the data points as estimates of the dimensions of the object, or attempt to fit a linear regression model through the data to obtain an estimate of the orientation of the data. We now discuss a slightly more sophisticated technique based on a geometric algorithm devised by Silverman and Titterington (1981).

(6.4.1.1) Minimum covering ellipses.

Consider the set E_S of recorded and accepted points in S. Then one problem of interest (and relevance to the object detection problem) is to construct the ellipse in the plane of minimum area covering all of the elements of E_S. In p dimensions, this problem is known as the Minimum Ellipsoid problem, and is familiar in a statistical context, relating to problems in the areas of design and outlier detection. For the planar case, Silverman and Titterington develop an efficient algorithm to compute the minimum covering ellipse for a given data set. They prove that either $3, 4, \text{or} 5$ elements of E_S only lie on this minimum ellipse, and as clearly these points must lie on the convex hull of all points in E_S, the amount of computation required is limited irrespective of the number of elements in E_S. Precise details of the algorithm, and the geometric and statistical arguments are given in Silverman and Titterington (1981).

The use of the idea of a minimum covering ellipse seems appropriate in the context of object detection - we wish to make inference about location, dimensions and orientation from a set of points known to be internal to some convex object of *a priori* unknown shape, and, as mentioned previously, we deem it adequate to assume an ellipse to be an approximation to the shape of the object. Importantly, the minimum ellipse algorithm provides the basis for an automatic and efficient technique for object detection. Thus, we now proceed to attempt an implementation of the algorithm for the sets of results in figures 108 and 109.

In light of the comments made above concerning the choice of k, we shall attempt object reconstructions for the data sets corresponding to $k = 2, 4, \text{and} 6$. Figure 110 depicts the three data sets, and the reconstructed ellipses.

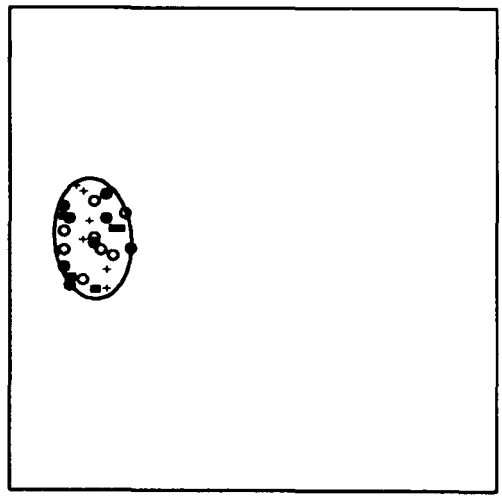


Fig 110(a) : $k = 2$

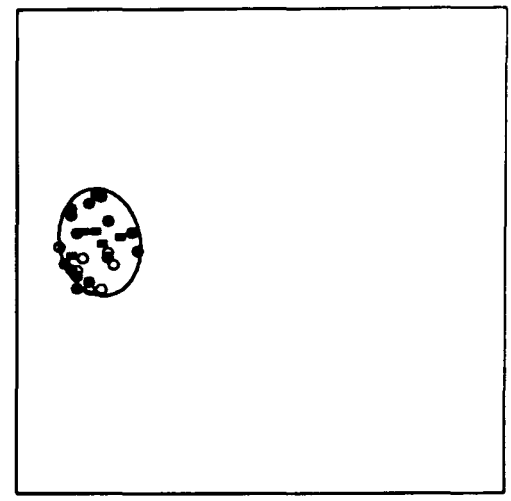


Fig 110(b) : $k = 4$

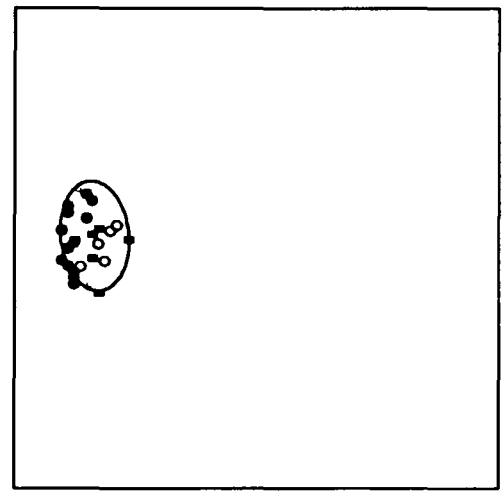


Fig 110(c) : $k = 6$

Clearly, although they differ slightly in shape, the reconstructed ellipses are similar. The "estimated" parameters (although we do not estimate these parameters as such) of the minimum covering ellipses in the three cases are presented in table 7.

	$k = 2$	$k = 4$	$k = 6$
p	14.67	13.63	13.33
q	41.67	41.59	42.00
a	9.96	8.84	9.27
b	6.33	6.65	5.64
α	1.65	1.75	1.67

Table 7

Further examples of Tank Spotting for single objects can be found in the images appendix. Generally, the technique works adequately for Signal-Noise ratios as low as 1.0, and is fairly robust to changes in the chosen value for k and nature of the smoother used.

The timings involved in the production of the results at various stages for the examples above are as follows. The changepoint analysis using (3.17) took, on average, around 1.7 seconds of processing time. The smoothing procedure and reconstruction of the ellipse using the Silverman-Titterington algorithm took an average total time of around 1.9 seconds. Thus, the total processing time involved in the detection of the object was around 3.6 seconds. Furthermore, the entire procedure can be readily automated. The only slight problem that might occur in an automatic implementation concerns the particular choice of smoother and the removal of isolated pairs, triples etc. of points, although hopefully this can be tackled by the point-density arguments and an iterated smoothing approach.

(6.4.2) Multiple object detection.

We now turn to the multiple object detection problem. The difficulties arising from this new problem are two-fold. Not only do we wish to make inference about the locations, dimensions and orientation of each of the objects, as discussed, we also wish to be able to distinguish between objects, and hence enumerate them. We might feel at this stage that merely using (3.17) as the basis of a detection method for multiple object images is insufficient, but note that in a previous example, a one changepoint posterior distribution was of use in a more complex situation. Thus, for the moment, we proceed with (3.17) rather than formulate anything more intricate.

It is possible to solve the problem of being able to distinguish between objects by introducing a labelling scheme into the smoothing procedure - that is, if a point is accepted under the point-density criterion, then it and all the points in its immediate vicinity are labelled with a number or "type", unless any of these points have been labelled previously as another type in which case the new point is labelled with this old type. This procedure adds negligibly to the amount of processing already required.

We now present examples of the performance of our algorithm in the multiple object detection problem. Figure 111(a) depicts an image derived from a true scene containing three tanks located at (58.31,75.10), (22.19,39.74), and (76.83,25.62) - again these locations were chosen randomly in S - and various orientations. The tanks are identical having dimensions (3.0,4.5), and the image formation process is such that the intensity for each tank compared to the background corresponds to a Signal-Noise ratio of 2.0. Figure 111(b) depicts the results of a full analysis using (3.17), with k chosen to be 2 (with the benefit of a degree of prior knowledge as to the dimensions of the tanks). Figure 111(c) depicts the results after smoothing with a (3,4) smoother, and the ellipses reconstructed using the Silverman-

Titterington algorithm.

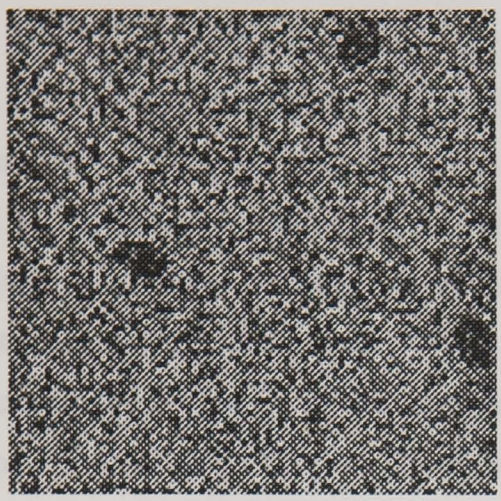


Fig 111(a) : image

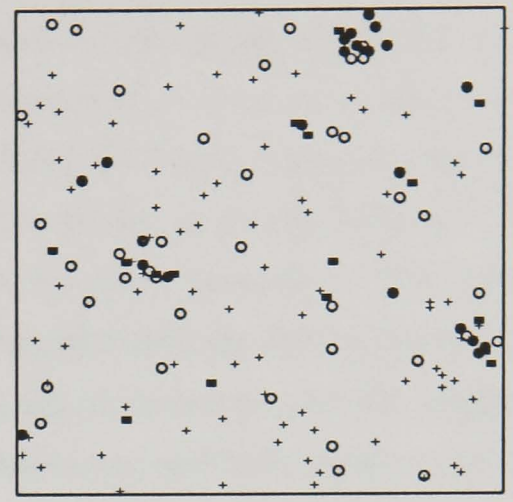


Fig 111(b) : raw results

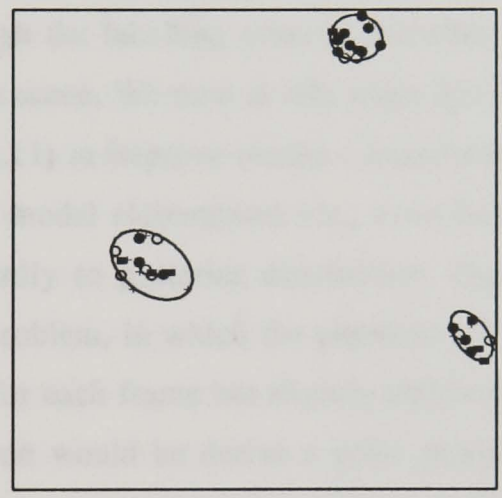


Fig 111(c) : reconstruction

The parameter values for the ellipses reconstructed from the smoothed and labelled changepoint data are presented in table 8.

	Tank 1	Tank 2	Tank 3
p	56.93	23.04	75.96
q	75.34	37.58	25.64
a	4.32	7.43	4.72
b	3.81	4.82	2.64
α	-0.14	-0.59	2.17

Table 8

The location parameters for the reconstructed ellipses are all accurate to within around 3% of the true scene location parameters, whereas the dimension parameters all differ to a larger degree. These features are intuitively reasonable given the nature of our object detection technique - as mentioned previously, we would expect the mode of (3.17) to occur within a narrow band centred on the true "changepoint" position, and thus we might expect the reconstructed ellipses to be located correctly but be increased in size relative to the true ellipses. This suggests the use of some multiplicative discount factor for these parameters. The orientation parameters for the reconstructed ellipses are generally unreliable for making inference about the orientations of ellipses in the true scene - again, this is connected with the nature of the posterior distribution (3.17). We also note that the results were generally robust to changes in k and the type of smoother used.

Thus we have developed a fully automatic technique for the detection of multiple objects in the Tank Spotting problem. Not only do we obtain information related to location and dimension, we also have (through the labelling scheme described above) the ability to count the number of objects in the true scene. We note at this stage that all the modifications that we made to posterior distribution (2.11) to improve results - introduction of spatial continuity into the prior specification, different model elaborations etc., even binary segmentation methods - could be made as straightforwardly to posterior distribution. Especially relevant here is the "frame of film" Tank Spotting problem, in which the positions of the tanks are rapidly changing, being regarded as constant for each frame but slightly different between frames. Clearly, a sensible approach in this situation would be derive a prior distribution over the pixels in S_θ (and consequently in each of the rows and columns of the image) for the positions of the objects through their recorded positions in the previous frame. This should assist greatly in tracking the tanks. We also note that (3.17) is in fact a special case of a two changepoint distribution, and we could extend the formulation and calculate further calculate further posterior distributions to be able to detect multiple objects (equivalently , these distributions are merely special cases of the multiple changepoint distributions that we have seen previously. Naturally , we would encounter the usual problems concerned with computational demand , but we could overcome this by replacing exact inference via the posterior distributions by approximate inference via the Gibbs Sampler based technique described in a previous section.

This concludes our study of edge-reconstruction and object detection problems for simple true scenes involving single edges and objects. Finally in this section, we discuss briefly the difficulties associated with reconstruction problems relating to more complex composite true scenes.

(6.5) Edge-reconstruction for complex true scenes.

Until this point, we have regarded edges in the true scene as simple, smooth and possibly closed single curves. For complex true scenes, however, our interpretation of the edges they contain is slightly different. More specifically, we regard the edges as being only piecewise simple or smooth, having discontinuities in first derivative - see, for example, the edges in the true scenes depicted in figures 44, 47, and 49. Thus, in the processing of results derived from such true scenes, we might seek to treat the reconstruction problem in this way, that is, reproduce the piecewise portions of the edges using the techniques discussed previously, and them "seam" the portions together. There are, of course, difficulties with such an approach, perhaps the most obvious one being determining automatically (in an unsupervised fashion) the seaming points. It is possible to suggest intuitively reasonable techniques - first differencing as an approximate gradient calculation, or localised curve fitting for small sets of neighbouring points - but the problem generally remains a stumbling block to automated reconstruction. A natural solution technique for the piecewise edges in a composite true scene is, of course, to use the spline techniques discussed at length previously, but as we mentioned then there are also analogous problems associated with such an approach. However, recall the precise nature of our decision problem, our ultimate goal - to achieve a discretised, pixel-by-pixel segmentation of the image into homogeneous regions. It is not clear how any piecewise reconstruction of edges from a set of edge-points derived from a composite true scene, a purely representational procedure, could aid in solution of the segmentation problem. Thus we regard such reconstructions as being beyond the scope of this thesis.

(6.6) Edge-reconstruction and object detection - conclusions.

In this section we have discussed stages of processing subsequent to an initial edge-detection analysis using changepoint analytic methods. We began by studying various procedures designed for representational purposes, that is, to visually enhance the resulting edge-point data. We studied various aspects of a least-squares approach in polynomial regression, and discussed the merits and otherwise of spline smoothing in data representation, but omitted any detailed discussion as it was felt not to be relevant to the specific decision problems covered in this thesis. Inherent in this was the fact that we now principally regard edge-detection as a pre-processing technique, an aid in the solution of the segmentation problem, rather than as a final solution of ultimate interest. However, specifically in the context of object detection, we still deem the edge-reconstruction problem important if it in any way assists in the gathering of information concerning the location, dimensions and orientation of objects in the true scene. To this end, we discussed at length the problems associated with the reconstruction of the edge of elliptical objects (taken as representative of the class of convex

objects), making use in particular of a linearised ellipse model and least-squares estimation procedures. We also discussed robustness and outlying points, described a familiar and potentially automatic technique for robust ellipse reconstruction based on convex hull peeling. We then discussed the related Tank Spotting problem, and adapted a changepoint technique discussed in a previous section to help in solution of the multiple object detection problem, and called on a minimum covering ellipse algorithm to help in making inferences about the objects. Finally, we noted that edge-reconstruction for complex true scenes was possible using spline based methods, but of little interest compared to the more important segmentation problem.

Chapter 7 : Image Segmentation and Pixel Classification.

In previous sections, we have seen the development of techniques for edge-detection in image processing resulting from the formulation of the problem as an exercise in statistical changepoint identification. We also saw how the results of the edge-detection analysis could be processed to allow inference in the related areas of edge-reconstruction and object detection. We eventually concluded that, despite the intuitive, theoretical and practical importance of learning about areas of discontinuity in the true scene, edge-detection *per se* should perhaps only be regarded as a pre-processing procedure, and not a primary and ultimate objective. Our principal aim is to achieve a segmentation of some discretised version of the true scene into homogeneous regions, a pixel-by-pixel classification into a (finite) number of distinct divisions or "colours" (the object detection and Tank Spotting problems may be regarded as being of a slightly different nature to the segmentation problem, or merely as special cases of it). In this final section, we attempt to integrate the edge-detection techniques that we have studied in previous sections with existing methods of solution to segmentation-type problems.

As detailed in the introduction, we may view the segmentation problem in several subtly different ways. First, we may regard it as an estimation problem, usually in a Bayesian framework - estimate the true scene pixel values via suitable posterior distributions and appropriate loss functions (M.A.P. estimation via annealing (Geman and Geman) etc.). Secondly, we may regard it as a probabilistic classification or discrimination problem, again usually in a Bayesian framework - assign labels to pixels using (discrete) posterior probabilities, generally without spatial considerations. Thirdly, we may regard it as a non-probabilistic classification problem, and consider solutions that may or may not be statistical in nature, or statistical solutions that are intuitively reasonable but informal (for example, Besag's ICM approach). Clearly, there are relationships between these three broad categories, but we shall continue to view them to be largely distinct. The presence of edges, discontinuities in the true scene, is a disruptive feature for each of these approaches. Consider, for example, the problem of obtaining initial estimates for the parameters of the image-formation process - this is an important task in each of the techniques mentioned above. One common suggestion relates to the use of "training data" - data gathered from some other source allowing posterior and predictive distributions, or naive estimates for these parameters to be computed. Unfortunately, in practice, this "other source" is not available, and we are forced gather the necessary information from the un-processed image itself. In an unsupervised setting, this task is by no means straightforward, as we need to be sure that the locations from which we extract the training data - for example, taking the form of 7×7 grids of pixel values - are internal to regions of homogeneity in the true scene. Clearly, this is only possible after an initial edge-detection routine has been used, and the edge-points post-processed, so that edge regions may be delineated. We

shall see such an approach implemented in relation to various classification schemes for various true scenes and images later. This typical example illustrates the importance of edge-detection as a pre-processing procedure. It also re-enforces our impression that edge-reconstruction purely as a technique for visual enhancement is of little use in an analytic context - we need only be able to distinguish edge-regions from texture regions at the pre-processing stage, and this can be achieved practically as well when using the smoothed edge-point data as when using an "estimate" of the edge reconstructed from them.

Therefore, in light of the above discussion, we investigate how our changepoint-based edge-detection routines can be introduced into a general approach to image segmentation. We have two principal objectives. First, we desire that pre-processing should be included in a straightforward way that, as in all our previous analyses, does not incur large computational costs. Secondly, we desire that it be introduced in an unsupervised fashion, or at least that the need for manual intervention be kept to a minimum. We shall see later how these objectives are achieved in the context of sophisticated classification, estimation and discrimination techniques. We begin by developing naive classification procedures using the results obtained directly from our edge-detection routines and smoothing procedures.

(7.1) Naive classification via changepoint analysis.

Given the set of post-smoothed and acceptable edge-points E_S, we wish to construct a segmented version of the discretised true scene. For simple true scenes, we may naively achieve the segmentation with reference only to the position of each pixel in relation to the approximate detected edge (i.e "left", "right" of, or "above", "below" a simple edge, internal or external to the closed curve representing the edge of an object), with the actual pixel value in the image having no bearing on the classification of that pixel. Intuitively, such purely spatial classification techniques might be of use when little is known about the nature of the image-formation or noise-processes, or as initial steps prior to more sophisticated processing. We now give a simple illustrative example of a technique of this nature relating to the analysis of single edge true scenes, and subsequently shall see a similar technique developed for convex object true scenes.

(7.1.1) Single edge example.

Recall the set of edge-points depicted in figure 9(b) on p. 40 in chapter 2, and reproduced in figure 112(a). These edge-points arise as modes in changepoint posterior distributions used in a row analysis of an image for which the Signal-Noise ratio is 2.0. The points clearly visually delineate the position of the edge. Consequently, it is visually straightforward to segment the image into its two constituent regions - those sets of pixels to the left and to the right of the edge. This segmentation may be achieved in an automated sense as follows -

for each row, label each pixel to the left (recall that, without loss of generality, we defined pixel 1 in a row to be at the left-hand end of that row) of the detected changepoint as "0", and each to the right as "1", with labelling being independent between rows. Figure 112(b) depicts the resulting segmentation using the naive technique described above. For comparison purposes, figure 112(c) depicts a two-texture segmentation of the image using a simple (non-spatial) maximum-likelihood classification rule, with texture means known.

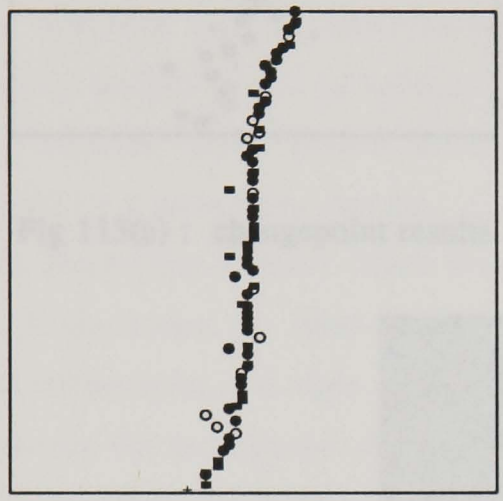


Fig 112(a) : changepoint results

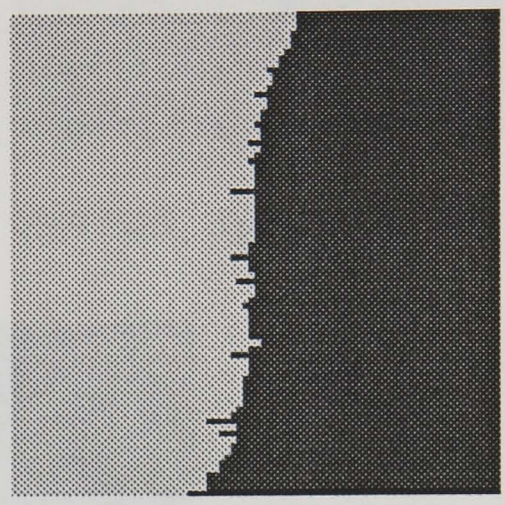


Fig 112(b) : naive classification

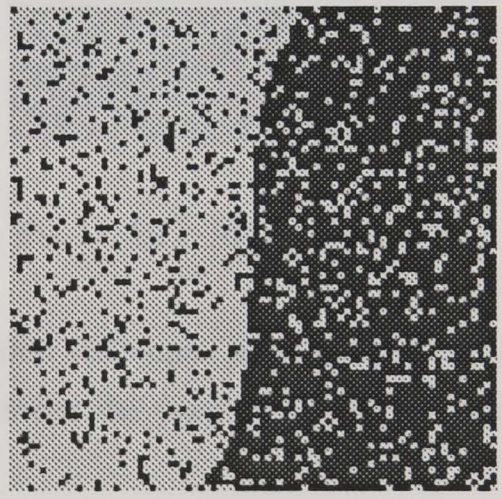


Fig 112(c) : M.L. classification

Clearly, the naive classification rule classifies a higher percentage of pixels correctly - with the maximum-likelihood rule we expect an error-rate of around 16% in the best possible case (when the threshold for texture segmentation is optimally chosen), and this clearly exceeds the error rate in figure 112(b). We might expect difficulties to arise for images having a lower intrinsic Signal-Noise ratio - the adequacy of the changepoint results and the subsequent segmentations will obviously decrease. Figure 113 depicts the results of an analysis of the identical image to that in figure 112, but where the Signal-Noise ratio has been decreased to 1.0.

Fig 113(a) : changepoint results

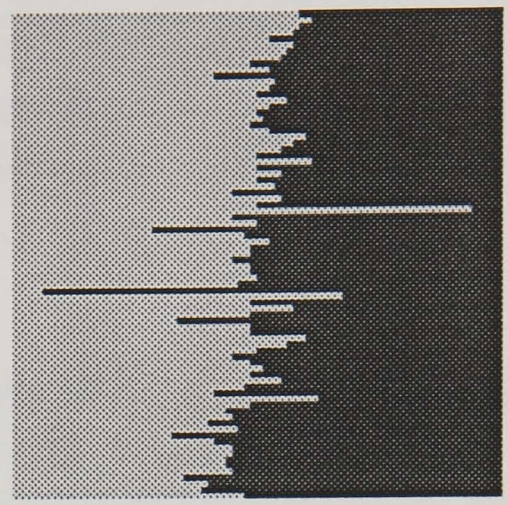


Fig 113(b) : naive classification

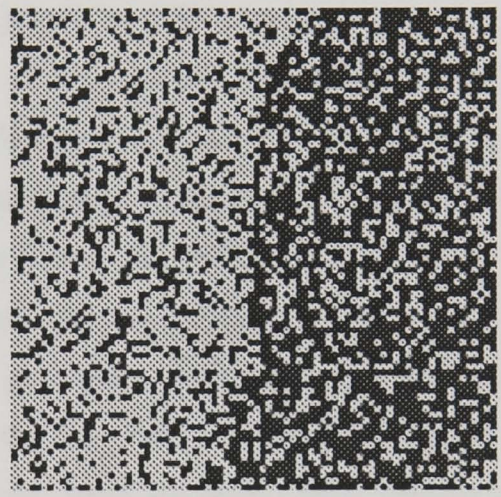


Fig 113(c) : M.L. classification

Remarkably, the naive segmentation technique, although producing imperfect results, is still reasonably adequate, and out-performs the maximum-likelihood classifier which now exhibits an error rate of almost 31%.

The comparison between the naive and maximum-likelihood classification techniques on these terms is not strictly a fair one - the former reflects both value and spatial location of the pixel in the image, whereas the latter is specifically non-spatial. However, the maximum-likelihood segmentation is often used as an initial stage in more sophisticated processing procedures, and hence we might regard the naive technique as an alternative that requires less in the way of prior input and, incidentally, processing time (requiring approximately 1.5 seconds).

Thus, for this simple example, the naive classification technique seems to produce adequate segmentations. However, the specific true scene that we have studied, where the edge is generally perpendicular to the direction of analysis, is particularly easy to segment purely on the basis of the results of a row analysis of the image. For true scenes in which the edge has a

more general angle of orientation, more care must be taken over the way in which pixels are classified. Fortunately, only a minor change to the segmentation algorithm is required to allow column based pixel classification to be incorporated into the complete classification algorithm. We thus have a simple and automatic classification procedure for use with single edge true scenes.

(7.1.2) Convex object example.

We now seek a similar naive classification scheme for single convex object true scenes. Our previous notions of a pixel being to the "left" or "right" of an edge are replaced by "internal" and "external" pixels (object and background pixels), and hence the naive classification routine will be a little more complex. A possible labelling procedure analogous to the above, therefore, proceeds as follows. Given the set of acceptable edge-points E_S resulting from a full analysis of the image, we label pixels in each row such that those pixels left of the extreme left-hand changepoint and right of the extreme right-hand changepoint are labelled "0", and those between the two extreme changepoints "1", and then repeat the process for each column. We adopt this approach as the geometry of the object indicates that each row and column in the true scene will contain either two or no changepoints, and although in practice the actual number of detected changepoints may differ from these two values, the use of extreme changepoints in the labelling procedure seems an intuitively sensible and conservative approach. The positions of the the two extreme changepoints are discovered in an efficient manner by tracking along the pixel sequence from either end in turn. One possible difficulty with such an approach arises when row and column classifications for any particular pixel conflict (i.e. a pixel is labelled "0" by row and "1" by column, for example). This difficulty may be overcome by taking the mean classification for that pixel over row and column, and subsequently attach a label of "0" if this mean is 0 or 0.5, or "1" if it is 1.0 - this naive approach will suffice for our current purposes, as it can be implemented without the need for any further intricate labelling algorithm. Alternatively, of course, we might choose to classify via row or column only, and thus avoid problems concerned with such conflicting classification - for convex objects whose edges are accurately detected, this approach is valid but less conservative, intuitively mis-classifying pixels at a higher rate.

Figure 114 provides an illustration of the use of the naive classification algorithm for single convex object true scenes. Figure 114(a) depicts an image derived from a true scene containing a single ellipse using the usual image-formation process (2.1) to produce a Signal-Noise ratio of 2.0. Figure 114(b) depicts the unsmoothed results of a full analysis using the binary segmentation version (specifically designed for such images) of changepoint analysis based on posterior distribution (2.11). Figure 114(c) depicts the segmentation resulting from the naive classification (jointly via row and column classifications) procedure described above after the changepoint results were smoothed using a simple $(3,2)$ smoother.

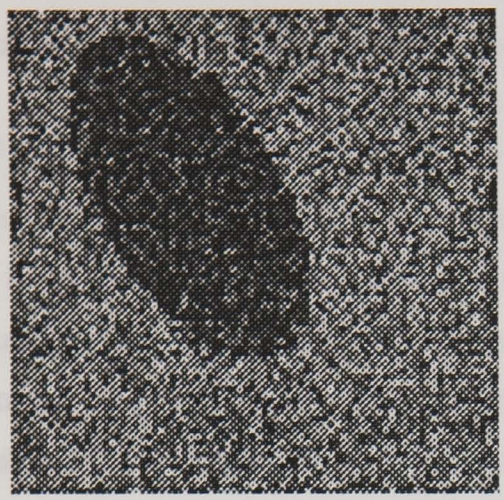


Fig 114(a) : image

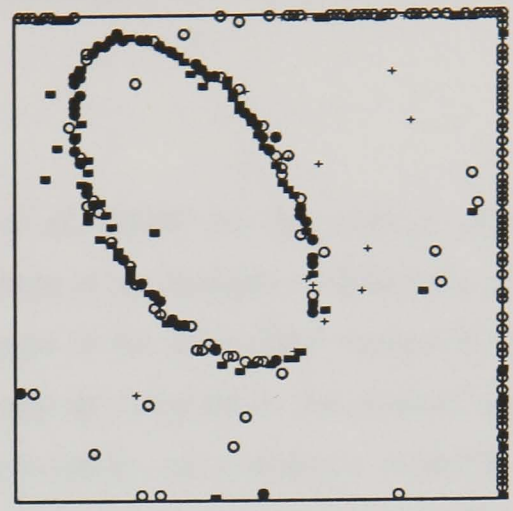


Fig 114(b) : changepoint results

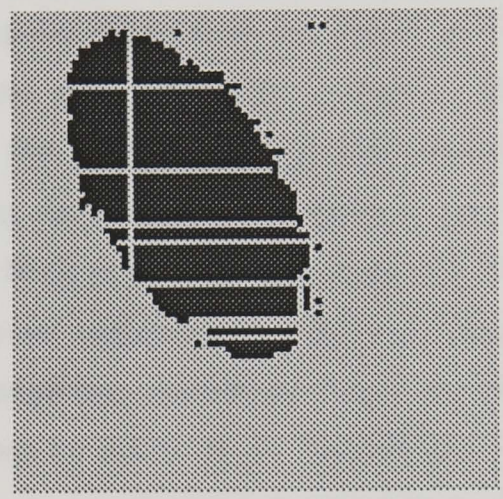


Fig 114(c) : naive classification

The segmentation using the naive classification method seems to be adequate. Again, the segmentation was achieved in an unsupervised setting in a matter of seconds.

It is possible to suggest many other such simple classification techniques. Importantly, the ideas can be extended with care to be of use for more complex and composite true scenes. Recall our study of edge-detection based on multiple changepoint posterior distributions, and the Gibbs Sampler technique for approximating such distributions. It is easy to see how, with careful labelling of pixels, and combination of row and column classifications etc., naive segmentations may be achieved. Also, using standard or approximate changepoint analysis, it is possible to derive posterior distributions for each of the unknown texture parameters in addition to changepoint positions. Such a procedure may clearly be of use in attempting to achieve a segmentation of the image, but would add considerably to the amount of, processing time required.

Each of the naive techniques described above classifies pixels according to their position relative to changepoint positions in row or column on the basis of a changepoint posterior

distribution. It is possible to incorporate these naive techniques into an iterative and adaptive classification scheme which we now describe.

(7.2) Iterative changepoint classification.

The following scheme was proposed by Buck *et al.* (1988) for the analysis of images derived from simple true scenes, and is a typical example of an iterative classification scheme. Consider, for example, the case where rows and columns in the discretised version of the true scene contain at most two edge-points - clearly this may be extended to the general case of k edge-points per row with care, but as we have seen previously, exact analysis is computationally infeasible for $k \geq 3$. Also, assume that the image-formation process is as in (2.1), and that our model of the image is of homogeneous regions in S_θ. Then in the usual way we may write down the form of two changepoint joint posterior distribution incorporating all one changepoint and the no changepoint alternatives by means of the obvious parameter constraints. For convenience now and later, we specify conjugate Normal-Gamma priors for the texture parameters, and for the first iteration the limiting non-informative versions of these conjugate priors are used. Using the posterior distributions, we may classify pixels by row and column as described above. Buck *et al.* consider the special "object on background" case, in which there are only two texture regions having distinct characteristics, and thus seek to achieve a segmentation of the image into two classes "high" (1) and "low" (0). They combine the row and column classifications by means of an "intersection principle" similar to the one we proposed above in which pixels are classified 1 only if they are classified 1 in both row and column analyses. Thus, after the first iteration, we have a segmentation of the entire image. The adaptive and iterative stage proceeds as follows. Given the initial segmentation, we alter the hyperparameters in the prior for the texture parameters in any given row or column in accordance with estimates for the texture parameters derived from the segmentation elsewhere in the image - the updating of the hyperparameters in this way is straightforward. We then proceed with another full analysis and complete another segmentation. This procedure is iterated until no changes in pixel classification are observed between iterations.

This procedure is an intuitively reasonable attempt to achieve a segmentation via changepoint analytic techniques. However, the algorithm as it stands has no formal theoretical justification, and consequently little is known concerning its convergence properties. Also, as the use of two changepoint posterior distributions is practically essential, computational requirements limit the scope of use of the algorithm to images having a small number of rows and columns. Unfortunately, the changepoint detection methods that we have developed become less reliable as the length of the data sequence decreases. Despite these negative factors, Buck *et al.* report that satisfactory segmentation results can be obtained.

The adaptive and iterative aspects of this algorithm are similar in nature to aspects of the Gibbs Sampler based detection techniques for image segmentation we discussed in chapter 1. In the adaptive sense, we would include the full conditional posterior distributions for the texture parameters into the sampling cycle, and not integrate the likelihood, leaving full conditional versions of the changepoint posterior distributions, and would then sample iteratively until convergence. We would usually treat each row and column independently, but could possibly incorporate with care global interpretations for the texture parameters, taking into account the pixel classification in the "current" segmentation (using either a synchronous or asynchronous pixel updating method). The classification technique described above contains no stochastic element in either its pixel value or texture parameter updating procedures.

Thus above we have described classification schemes based purely on the results the edge-detection routines using changepoint analysis. The schemes are to some extent arbitrary, and take little account of actual image pixel values. A simple and natural extension would be to use the segmentations achieved by such schemes as initial segmentations in more formal probabilistic classification or estimation schemes, as suggested previously. The advantage of using the naive classification schemes initially is that it can be done in an unsupervised manner - recall the problems associated with choosing suitable discrimination parameters without prior knowledge of content of the true scene. We now present some examples to illustrate how the simple schemes can be incorporated into probabilistic classification routines. We begin with an example of a simple discriminatory approach.

(7.3) Simple probabilistic classification.

Consider the simplest possible pixel classification scheme that classifies a pixel to a texture with respect to that textures posterior probability given the pixel image value - that is, classify pixel i to texture T_j if

$$\Pr(T_j \mid Y_i) > \Pr(T_k \mid Y_i) \tag{7.1}$$

for all other textures T_k. In the two texture case under this rule, therefore, we would classify pixel i to texture T_0 ($i \in T_0$) if

$$\Pr(T_0 \mid Y_i) > \Pr(T_1 \mid Y_i)$$

and to texture T_1 otherwise. By Bayes theorem, cancelling constants in the usual way, this condition is equivalent to

$$\Pr(Y_i \mid T_0)\Pr(T_0) > \Pr(Y_i \mid T_1)\Pr(T_1) \qquad \Rightarrow i \in T_0 \tag{7.2}$$

where $\Pr(Y_i \mid T_j)$ is the likelihood of observing Y_i conditional on pixel i being in texture j, and $\Pr(T_j)$ is the prior probability that pixel i is in texture j, for $j = 0,1$. Under the standard assumptions concerning the image-formation process (i.e. that the error terms are normally distributed) it is clear that, for pixel i in texture j,

$$\Pr(Y_i \mid T_j) = \left(\frac{1}{2\pi\sigma_j^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_j^2}(Y_i - \mu_j)^2\right\}$$

where μ_j is the mean for texture j, and σ_j is the standard deviation of the noise terms corrupting texture j. For the moment we shall assume that σ_j need not be equal for each texture, and that in fact μ_j and σ_j are parameters at our disposal. The simplest version of this classification scheme therefore proceeds as follows. We choose pairs (μ_j, σ_j), and prior probabilities $\Pr(T_j)$ for $j = 0,1$, evaluate the terms in (7.2) and allocate pixels to textures accordingly. We might choose (μ_j, σ_j) arbitrarily, or using simple techniques (inspection of an appropriate grey-level histogram), or using training data - choice of these parameters is not straightforward in an unsupervised setting. Also, in the case of prior ignorance of the true scene, we are virtually forced to specify the $\Pr(T_j)$ to be equal - in the two texture case, $\Pr(T_0) = \Pr(T_1) = 0.5$ - and revert to a classical discrimination rule, equivalent to the maximum-likelihood rule we saw earlier. Calculation of expected error rates are straightforward for such schemes.

Thus, the most significant difficulty with the implementation of such a scheme lies in the specification of the parameters (μ_j, σ_j) - choice of these parameters in the case of *a priori* ignorance of the true scene image-formation process is practically impossible, and is also not straightforward when training data is available. However, using the naive classification schemes described above, the difficulty may be overcome in an unsupervised setting.

(7.3.1) Probabilistic classification - simple example.

Consider, for example, the analysis of an image derived from a single convex object true scene. We may repeat the edge-detection analysis (as in figure 114(b)) and, on the basis of the results obtained, achieve a naive segmentation in the way described above (figure 114(c)) by labelling all "external" points 1, all "external" points 0, and all points for which the row and column classifications conflict 0.5. We then proceed and produce a binary segmentation of the image into two textures using the probabilistic discrimination techniques described by (7.1) and (7.2) in the following way. For each texture T_0 and T_1, we obtain estimates for mean and variance parameters using those pixel values labelled 0 an 1 in the naive segmentation respectively - generally, as the amount of data relating to each texture is large, the usual maximum-likelihood estimates will suffice. Further, having produced such estimates, we may specify the prior texture probabilities $\Pr(T_j)$ for $j = 0,1$ in accordance with the initial segmentation - for example, we might choose $\Pr(T_0)$ to be 0.75 if the naive classification was 0, 0.5 if the naive

classification was 0.5, and 0.25 if the naive classification was 1. A simple scheme such as this one merely provides an automatic implementation for the familiar classification technique, incorporating some low level and qualitative information via the choices of the $\Pr(T_j)$.

We now present a simple example to demonstrate the use of the automatic classification technique described above. Figure 115(a) depicts an image derived from a true scene containing a single ellipse centred at $(60,20)$ of dimensions $(15.0,7.5)$ and angle of orientation 2.1, corrupted so that the Signal-Noise ratio is 1.5 - means 1.5 and 0.0 for object and background respectively, with a noise variance of 1.0. Figure 115(b) depicts the segmentation achieved using the naive classification technique derived from the results of a full changepoint analysis using posterior distribution (2.11). Figure 115(c) depicts the segmentation achieved using the simple probabilistic classification technique with the prior probabilities and parameters in (7.1) and (7.2) chosen automatically. In this instance, we assume that variances are different between textures - it is possible to derive the analogous form to (7.1) under a common variance assumption, using a pooled sample estimator.

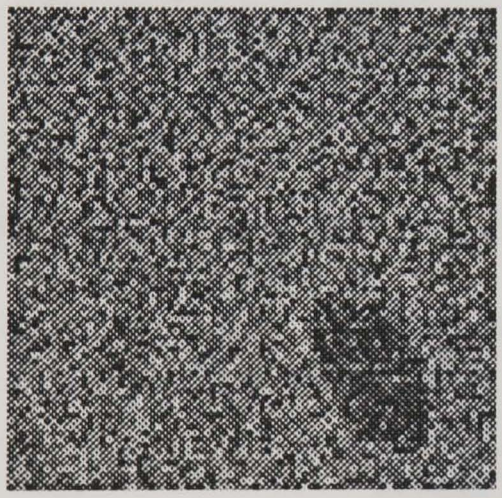


Fig 115(a) : image

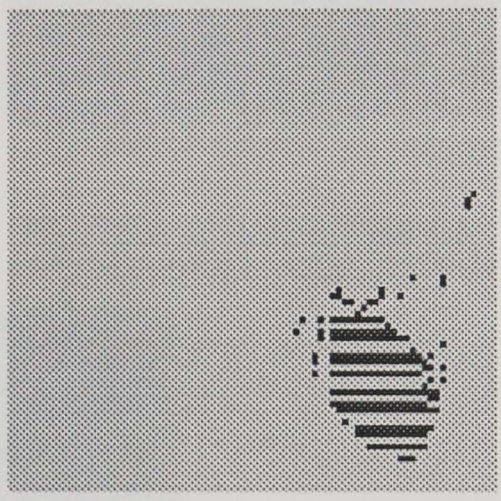


Fig 115(b) : naive classification

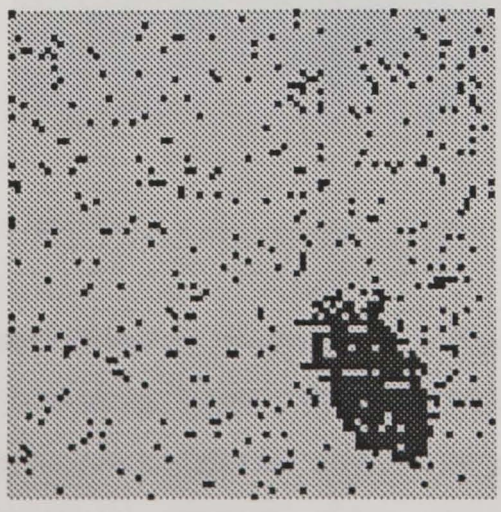


Fig 115(c) : classification using (7.2)

The estimates for the mean and variance pair (μ_j,σ_j) for textures 0 and 1 were

$(-0.005, 1.003)$ and $(1.321, 1.148)$ respectively, with, initially, 5610 pixels being classified 0 and 191 pixels being classified 1 - the remaining 599 pixels were indeterminately classified as 0.5. For comparison purposes, figure 116 depicts the segmentations obtained using a supervised version of the simple classification scheme for various choices of input parameters under a common variance assumption and equal prior texture probabilities. Figure 116(a) depicts the segmentation obtained when the means from above are used, (b) where the texture means are correctly (i.e. exactly as in the image) specified, and (c) where the means are incorrectly specified.

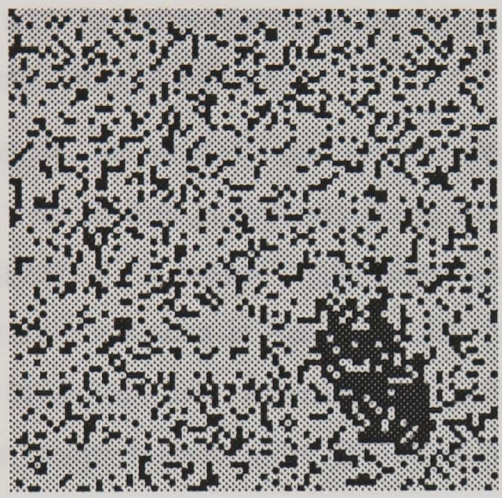


Fig 116(a) : means -0.005 and 1.321

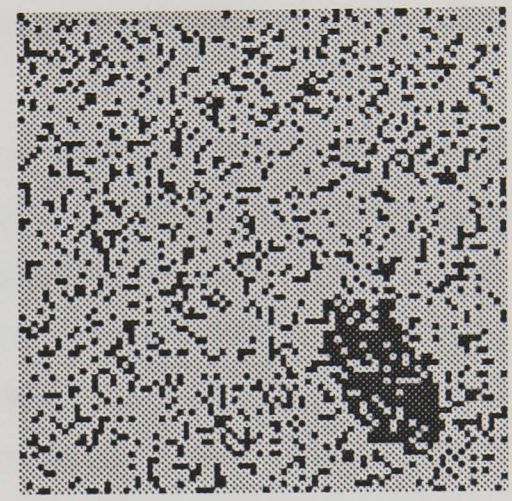


Fig 116(b) : means 0.0 and 1.5

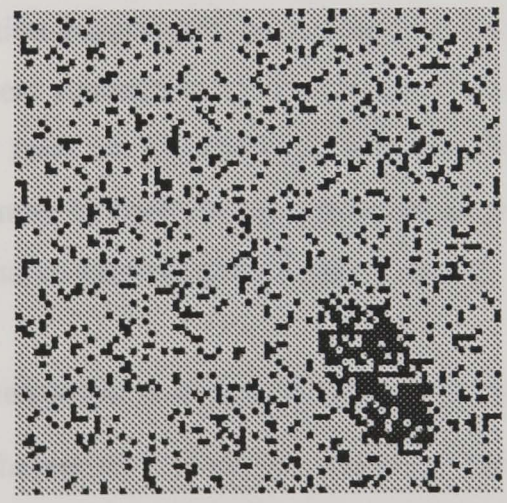


Fig 116(c) : means 0.0 and 2.0

Of the four probabilistic binary segmentations of the image in figure115(a), it is clear that the pixel mis-classification rate is lowest in figure 115(c) in which some degree of spatial knowledge derived from the results of changepoint analysis is incorporated. A direct comparison of figures 115(c) and 116(a) (in which the texture means are specified identically) illustrates the effect that specifying unequal prior texture probabilities has on the final segmentation. We also note at this point that the segmentation achieved via the naive classification technique (figure 115(b)) is in fact superior to each of the other four segmentations - this is

relevant if we merely seek an initial approximate segmentation for use in further routines, but as the naive classification scheme is rather *ad hoc*, having no formal justification, the segmentation does not bear fair comparison with the segmentations achieved using formal probabilistic techniques.

The total processing time required for the unsupervised production of the segmentation in figure 115(c) was of the order of 5.0 seconds. This is comparable with times for other non-iterative segmentation algorithms, and is by no means prohibitive. Thus we have improved on a simple probabilistic classification scheme by incorporating results from edge-detection analysis and the naive classification approach in an unsupervised manner without incurring any great computational burden.

Clearly, the probabilistic scheme described above could be implemented in an iterative fashion, with texture parameters being updated subsequent to each successive segmentation, as in the so-called "K-means" algorithm (see, for example, Hartigan(1975)). The problem with such iterative schemes is that generally there exists little formal theory relating to the nature of their behaviour and assessment of their convergence. The Gibbs Sampler iterative algorithm for image segmentation discussed previously can be regarded as being superior to such schemes, because although convergence may again be difficult to discern, it is at least assured by the theorems proved by Geman and Geman. A natural and obvious extension to the usual Gibbs Sampler algorithm that merely concentrates on pixel classification (via global or marginal estimation of true classification parameters) would be one that incorporated some facility for updating the texture parameters subsequent to each successive segmented in some coherent and regulated fashion, in some form of simultaneous scheme. We now proceed to develop such an iterative scheme which may be used for segmentation after an initial segmentation has been achieved using one of the naive classification routines above.

(7.4) Simultaneous image segmentation and parameter estimation.

As described in the first chapter of this thesis, maximum posterior probability estimates of true scene pixel classifications under standard prior assumptions (Markov random field) may be obtained via stochastic relaxation (Gibbs Sampler) and optimisation (simulated annealing) in the following way - sample iteratively from the "full" conditional posterior distributions of the true scene classification parameters until convergence. Consider, for definiteness, the case where the image-formation process is as in (2.1) with with the noise terms independently normally distributed, but where the noise variance is different for each texture, and suppose that the number of textures is known to be K. Suppose also that we specify the loss function in the Bayesian estimation problem in such a way that the estimates of the true scene classification parameters correspond to the modes of the marginal posterior distributions for those parameters - the marginal posterior mode (M.P.M.) estimation scheme. Then the Gibbs

Sampler algorithm under these circumstances proceeds as follows. The full conditional posterior distribution of the true scene classification parameter for pixel (i,j) given the true scene classification elsewhere $\boldsymbol{\theta}_{(ij)}$, denoted $[\theta_{ij} \mid \boldsymbol{Y}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta}_{(ij)}]$ (a discrete distribution, with θ_{ij} taking values on $\{0,1,\ldots,K-1\}$) is given by

$$\begin{aligned}[\theta_{ij} \mid \boldsymbol{Y}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta}_{(ij)}] &= [\theta_{ij} \mid Y_{ij}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta}_{\partial ij}] \\ &\propto [Y_{ij} \mid \theta_{ij}, \boldsymbol{\mu}, \boldsymbol{\tau}]\,[\theta_{ij} \mid \boldsymbol{\theta}_{\partial ij}] \end{aligned} \tag{7.3}$$

where $\boldsymbol{\theta}_{\partial ij}$ is the true scene classification in pixels neighbouring (i,j), $\boldsymbol{\mu} = [\mu_0, \mu_1, \ldots, \mu_{K-1}]^T$ and $\boldsymbol{\tau} = [\tau_0, \tau_1, \ldots, \tau_{K-1}]^T$ are the $(K \times 1)$ vectors of texture means and precisions, and where $[Y_{ij} \mid \theta_{ij}, \boldsymbol{\mu}, \boldsymbol{\tau}]$ depends θ_{ij} in such a way that if $\theta_{ij} = k$ then

$$\begin{aligned}[Y_{ij} \mid \theta_{ij}, \boldsymbol{\mu}, \boldsymbol{\tau}] &\equiv [Y_{ij} \mid \mu_k, \tau_k] \\ &\propto \tau_k^{\frac{1}{2}} \exp\left\{ -\frac{\tau_k}{2}(Y_{ij} - \mu_k)^2 \right\}. \end{aligned} \tag{7.4}$$

In the usual way, $[\theta_{ij} \mid \boldsymbol{\theta}_{\partial ij}]$ takes the form of a Gibbs distribution so as to reflect qualitatively our prior knowledge of the spatial dependence structure in the true scene classification. Consider the simplest possible form for $[\theta_{ij} \mid \boldsymbol{\theta}_{\partial ij}]$, where

$$[\theta_{ij} \mid \boldsymbol{\theta}_{\partial ij}] \propto \exp\{\beta \#_{ij}^{()}\} \tag{7.5}$$

and where, for each k, $\#_{ij}^{(k)}$ is the number of elements of the vector $\boldsymbol{\theta}_{\partial ij}$ equal to k - this is equivalent to an assumption that the only contributions to the joint prior distribution for pixel classifications over the entire true scene, $[\boldsymbol{\theta}]$, are made pairwise by neighbouring pixels in the same texture. Such a prior distribution is commonly used as a simple model for the discretised true scene (see, for example Besag(1986)). The hyperparameter β in (7.5) is a parameter at our disposal, and may be chosen in a number of ways, as noted in the introductory section. Thus, combining (7.4) and (7.5) via (7.3), it is clear that the elements of the discrete distribution $[\theta_{ij} \mid \boldsymbol{Y}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta}_{(ij)}]$ are given by

$$\Pr(\theta_{ij} = k \mid \boldsymbol{Y}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta}_{(ij)}) \propto \tau_k^{\frac{1}{2}} \exp\left\{ -\frac{\tau_k}{2}(Y_{ij} - \mu_k)^2 + \beta \#_{ij}^{(k)} \right\} \tag{7.6}$$

for $k = 0, 1, \ldots K-1$. Thus, standard implementation of the Gibbs Sampler algorithm in conjunction with the annealing procedure for optimisation involves sampling iteratively from the distribution given by (7.6) raised to the power ρ - the annealing control parameter - after

normalisation, and varying ρ according to some pre-calculated schedule. The optimisation process necessary for the evaluation of the posterior modal values is thus time-consuming - convergence is required for each ρ and across all values of ρ in the schedule. The parameters μ and τ in this implementation can be regarded as known initially and fixed, but in many instances parameter updating schemes are used subsequent to each segmentation of the image being completed. Once such scheme suggested by Besag (1974) and implemented by other authors is that of maximum pseudo-likelihood estimation. This approach is related to the standard maximum-likelihood, but differs in the sense that the function of the parameters to be maximised is not a true likelihood, but the product of conditionally dependent likelihoods - see, for example, Lakshmanan and Derin (1989) for such one such implementation of a recursive "Adaptive Segmentation Algorithm" similar in nature to the EM algorithm for use in M.A.P. estimation, and related convergence results, and a discussion of other techniques. Such an approach seems neither as elegant nor as intuitively pleasing as that underlying the Gibbs Sampler algorithm, and is derived principally out of convenience.

However, consider the following minor adjustment to the standard Gibbs Sampler approach to M.P.M. estimation described above. In addition to the $n \times n$ full conditional posterior distributions $[\theta_{ij} \mid Y, \mu, \tau, \theta_{(ij)}]$, we may write down in the same way the full conditional posterior distributions for μ and τ, or equivalently, the K pairs of full conditional posterior distributions for μ_k and τ_k, $k = 0, 1, \ldots, K-1$. Consider, for example, the full conditional posterior distribution for μ_k given the data Y, the true scene pixel classification θ, the vector of mean parameters for the other textures $\mu_{(k)}$, and the vector of texture precision parameters τ, denoted $[\mu_k \mid Y, \mu_{(k)}, \tau, \theta]$. Now, and importantly, given the image-formation assumptions, it is clear that conditional on the true scene classification θ, the μ_k are independent of each other, and are also independent of the precision parameters relating to other textures, and also that for each k, μ_k is dependent only on pixel values in the image derived from pixels in the true scene classified to texture T_k, denoted Y_k. Therefore, we may simply write

$$[\mu_k \mid Y, \mu_{(k)}, \tau, \theta] \equiv [\mu_k \mid Y_k, \tau_k] \tag{7.7}$$

as the form for the full conditional posterior distribution for μ_k. In the usual way we rewrite $[\mu_k \mid Y_k, \tau_k]$ via Bayes theorem and our independence assumption

$$\begin{aligned} [\mu_k \mid Y_k, \tau_k] &\propto [Y_k \mid \mu_k, \tau_k][\mu_k] \\ &= \prod_{(i,j) \in T_k} [Y_{ij} \mid \mu_k, \tau_k][\mu_k] \end{aligned} \tag{7.8}$$

where $[\mu_k]$ is our prior distribution for μ_k. Under our assumption of normally distributed errors in the image-formation process, it is clear that if n_k denotes the number of elements of Y_k then the first term in (7.8) may be written

$$[Y_k \mid \mu_k, \tau_k] = \tau_k^{\frac{n_k}{2}} \exp\left\{ -\frac{\tau_k}{2} \sum_{(i,j) \in T_k} (Y_{ij} - \mu_k)^2 \right\}. \tag{7.9}$$

Now, for convenience, we may choose $[\mu_k]$ to be of the standard conjugate (normal) form for this likelihood, that is that $[\mu_k] \equiv N(\phi_k, \gamma_k^{-1})$ say. Thus, from (7.8) and the standard procedure, it is easy to show that

$$[\mu_k \mid Y_k, \tau_k] \equiv N\left(\frac{\tau_k n_k \overline{Y_k} + \gamma_k \phi_k}{\tau_k n_k + \gamma_k}, \frac{1}{\tau_k n_k + \gamma_k} \right) \tag{7.10}$$

where $\overline{Y_k} = \frac{1}{n_k} \sum_{(i,j) \in T_k} Y_{ij}$

Thus we have derived a form for the full conditional posterior distribution for μ_k. In a similar way we may derive the equivalent posterior distribution for τ_k, which we denote $[\tau_k \mid Y, \mu_k, \tau_{(k)}, \theta]$. Identical arguments to those leading to (7.7) allow us to write

$$[\tau_k \mid Y, \mu_k, \tau_{(k)}, \theta] \equiv [\tau_k \mid Y_k, \mu_k] \tag{7.11}$$

and consequently

$$\begin{aligned} [\tau_k \mid Y_k, \mu_k] &\propto [Y_k \mid \mu_k, \tau_k][\tau_k] \\ &= \prod_{(i,j) \in T_k} [Y_{ij} \mid \mu_k, \tau_k][\tau_k] \end{aligned} \tag{7.12}$$

where $[\tau_k]$ is our prior distribution for τ_k. Again, if we choose a standard conjugate (Gamma) form for $[\tau_k]$, say $[\tau_k] \equiv Ga(\alpha_k, \delta_k)$, then it easy to show that

$$[\tau_k \mid Y_k, \mu_k] \equiv Ga\left(\alpha_k + \frac{n_k}{2}, \delta_k + \frac{S_k}{2} \right) \tag{7.13}$$

where $S_k = \frac{1}{n_k} \sum_{(i,j) \in T_k} (Y_{ij} - \mu_k)^2$.

Thus in total we have the forms of $n^2 + 2K$ full conditional posterior distributions, namely $[\theta_{ij} \mid Y, \mu, \tau, \theta_{(ij)}]$ for $i, j = 1, \ldots, n$, and $[\mu_k \mid Y, \mu_{(k)}, \tau, \theta]$ and

$[\tau_k \mid Y, \mu_k, \tau_{(k)}, \theta]$ for $k = 0, 1, \ldots, K-1$, and in the example we have described above, we need to specify $2K + 1$ prior parameters, namely those relating to prior beliefs concerning texture parameters, $\phi_k, \gamma_k, \alpha_k, \delta_k$ for $k = 0, 1, \ldots, K-1$, and the parameter β. For convenience, we continue to suppress via our notation the functional dependence of each of the conditional distributions on these hyperparameters. We now propose the following amendment to the standard Gibbs Sampler approach to M.P.M. estimation in an attempt to include some form of adaptive parameter updating scheme into the segmentation procedure.

Recall the work of Gelfand *et al.* (1989) on the approximation of marginal posterior densities in "difficult" Normal models, and our work on the approximation of changepoint posterior distributions, via the Gibbs Sampler which has been discussed at length in previous sections. There, after sampling iteratively from the set of full conditional distributions until convergence is diagnosed over a number of replications, density/distribution estimates were formed. More specifically, for each replication, at convergence, the sampled values from each conditional form are regarded as essentially sampled values from the required marginal forms.

Our proposal of an amendment of the Gibbs Sampler algorithm for use in the image segmentation problem is as follows. We propose simply that on each iteration, we should sample from the $2K$ conditional posteriors $[\mu_k \mid Y, \mu_{(k)}, \tau, \theta]$ and $[\tau_k \mid Y, \mu_k, \tau_{(k)}, \theta]$, $k = 0, 1, \ldots, K-1$, in addition to the n^2 conditional posteriors $[\theta_{ij} \mid Y_{ij}, \mu, \tau, \theta_{\partial ij}]$, $i, j = 1, \ldots, n$, required for the standard analysis. For example, suppose that after iteration t, we have most recently obtained values $\theta_{ijt}, \mu_{kt}, \tau_{kt}$ for each of the parameters of interest (with the extra subscript denoting the number of the most recent iteration in the obvious way). Then on the next iteration, we sample again from the $[\theta_{ij} \mid Y_{ij}, \mu, \tau, \theta_{\partial ij}]$ with the relevant values from $\theta_{ijt}, \mu_{kt}, \tau_{kt}$ substituted into the posterior form as the conditioning variables, and also from the $[\mu_k \mid Y, \mu_{(k)}, \tau, \theta]$ and $[\tau_k \mid Y, \mu_k, \tau_{(k)}, \theta]$ in the same way. We choose here to update the most recent or "current" values asynchronously i.e. as soon as we have sampled a replacement value from the corresponding conditional posterior distribution, principally for computational convenience. Intuitively, the values obtained for the texture parameters will gradually become closer and closer to the true parameter values, and, crucially, the theorems proved by Geman and Geman insure that at convergence the sampled values are in fact variates from the true marginal posteriors, for both true scene pixel classification and texture parameters.

Thus, for a practical implementation of this amended form of the Gibbs Sampler algorithm for the evaluation of marginal posterior modes, we need consider, first, the specification of starting values and prior parameters, and, second, the assessment of convergence and the reporting of posterior estimates. We address these points in turn.

(7.4.1) Specification of hyperparameters.

The parameters $\phi_k, \gamma_k, \alpha_k, \delta_k$ for $k = 0, 1, \ldots, K-1$ and β are parameters in prior distributions, and as such we are at liberty to choose them in any way we see fit. However, "sensible" choices for these parameters will clearly improve the rate of convergence of the algorithm, and thus it is in our interest to choose them accurately. We have seen above how naive classification schemes may be used to derive initial parameter estimates and thus it is possible to specify the prior texture parameters in accordance with these estimates in an automatic procedure. Having specified the precise forms of the prior distributions, the choice of starting values for the Gibbs Sampler algorithm - for the texture means and precisions, we choose the initial values as the modes of the prior distributions, and for the true scene classification parameters, we use the familiar maximum-likelihood classification rule derived from the prior distributions for each individual pixel. The prior specification of β is somewhat more complicated. Whereas we have sufficient experience with normal models to allow us specify parameters in normal priors in a sensible way, we have relatively little experience of specifying parameters in Gibbs priors. For example, in the particular case above, it is widely known from theoretical work in statistical Physics on the Ising model, and from the work of, for instance, Besag, that the choice of β must be made with care to avoid critical and super-critical behaviour of the random field concerned. Typically, choosing β too large will result in convergence to segmented images of containing one texture only - for instance, for convex object true scenes, pixels classified to the object texture are gradually "peeled away" from the outside of the object until the object eventually disappears. Also, it is clear from the form of (7.6) that the choice of β should be related in some way to the values of texture precision parameters, indicating that some adaptive scheme for updating β between iterations on the basis of current parameter estimates elsewhere may be necessary. More generally, we should introduce a parameter β_k for each texture T_k, and also "between texture" pixel interaction parameters β_{kl} for pixels in textures T_k and T_l in the prior form in (7.5), to allow a more sophisticated prior model for the true scene classification. Thus the problems concerned with the specification of hyperparameters in Gibbs prior for the true scene compound each other rapidly. We can offer no further recommendation concerning the choice of β other than that described in previous sections, and for the moment content ourselves with the practical experience of others.

(7.4.2) Assessment of convergence.

As mentioned in our previous discussion of the Gibbs Sampler algorithm, assessment of convergence can often be a difficult task, and there exist no formal criteria that may be used as the basis of diagnostic tools. One commonly used approach in the M.P.M case is to compare the segmentations achieved after successive iterations, with convergence being diagnosed if the number of discordant pairs of pixels is small, say less than or equal to 5% of the pixels in the entire image. It is widely reported that during the progression of the algorithm, the

number of pixel "flips" between iterations rarely falls consistently to zero - it is easy to see why this is so, as when the Gibbs Sampler algorithm is used without an edge-process (Geman and Geman), the marginal posterior distributions for a pixel on or near the boundary of two textures will not allow easy classification into either texture, as the neighbourhood of that pixel contains significant numbers of pixels in the true scene classified to each. Thus, even near convergence, we might expect some fluctuation in the M.P.M classification of some pixels - we might feel because of this that merely counting the number of pixels changing classification, or the change in the number of pixels classified to each texture between iterations is insufficient. Despite this, we continue to use such criteria in an attempt to assess convergence because of their straightforwardness and ease of automatic implementation. Hopefully, at some later stage, we may be able also to use some of the ideas above concerning the location of pixels that change classification frequently to aid in the assessment of convergence. It may also be possible to incorporate some criteria involving the posterior distributions of the texture parameters into the overall assessment - for instance by tracking the posterior modal estimates of these parameters subsequent to each iteration - despite the fact that these parameters are not our primary concern. We now study precisely how the posterior estimates and distributions may be constructed in this context.

(7.4.3) Evaluation of modal estimates.

The task of locating the posterior modal estimates in standard implementations of M.P.M image segmentation routines is performed using simulated annealing as an optimisation tool. As mentioned above, this can be laborious and time consuming. In the work of Gelfand *et al.*, posterior estimates, densities and distributions are constructed as averages of the results of independent replications of the Gibbs Sampler analysis. Now consider the implications of such an approach in the image segmentation context. The notion of producing a number of independent replications of the analysis is daunting due to the amount of computation required - we must compute the exact forms of $n^2 + 2K$ posterior distributions, sample uniformly from $(0,1)$ n^2 times, and from Normal and Gamma distributions K times, on each iteration. Consequently, for the implementation described above with R replications, the marginal posterior distribution estimate for θ_{ij} is the average of R discrete K-valued distributions, the marginal posterior density estimate for μ_k is an equal-weighted mixture of R Normal densities of the form of (7.10), and the marginal posterior density estimate for τ_k is an equal-weighted mixture of R Gamma densities of the form of (7.13), for $k = 0,1,\ldots,K-1$, with the distributions and densities within each mixture being conditionally independent given $\mathbf{Y}$. The process of averaging over a number of replications is essentially used to reduce the variance of the density/distribution estimates. Thus, again the amount of computation required is large, especially for the (continuous) texture parameters, and the idea of producing replicate analyses is not appealing. Fortunately, it seems that this is not necessary in a large number of cases, for

the following reasons. On inspection of the marginal posterior distributions for the true scene classification parameters, we discover that, once the algorithm has converged (and often from a very early stage of the iterative procedure), these distributions are practically degenerate, that is, the posterior probability of one particular texture being almost 1.0, with all the other posterior probabilities being negligible. Only for pixels at or near texture boundaries in the true scene do the marginal distributions not exhibit this quality, as we might have predicted from the above discussion. Thus, generally, little is to be gained by performing replicate analyses, as our primary goal is to achieve a maximum probability segmentation of the image. Naturally, if we seek estimates for the texture parameters also, it is advantageous to perform replications. Consider the case of the marginal posterior density for μ_k. This posterior density is, as mentioned above an R-mixture of Normal densities that have approximately (i.e. in expectation) equal means and variances. Then the posterior variance for μ_k is approximately a factor of R times smaller than for each individual replication - hence performing replicate analyses is advisable. However, by the same argument, it is clear that the posterior modal estimate for μ_k will not be altered greatly by the formation of the R-mixture, and thus, if such a point estimate is sufficient (as it may well do given that the texture parameters are not of primary interest) then we need not be too concerned about performing only one or a small number of replicate Gibbs Sampler analyses.

(7.4.4) Merging of textures.

Finally, we discuss one further aspect of the implementation of the amended Gibbs Sampler algorithm, namely the choice of the number of textures K. Clearly, in the case where we are *a priori* completely ignorant of the nature of the true scene, specification of K is difficult, and therefore practically we might choose K large. Given a degree of prior knowledge (for instance from the results of an edge-detection analysis, we might be better able to choose K accurately. In either case, we would tend to choose K as representing some upper bound on the actual number of distinct textures that we believe to be present in the true scene. Consequently, we are faced with a second problem. If we specify K larger than the actual number of textures, then for convergence to a correct segmentation it will be necessary to merge two or more of the original textures into one at some stage of the analysis. This is a familiar technique in image processing (also known as "split and merge" or "region growing" etc. - see, for example, Cohen and Cooper (1987)) that is generally carried out with some small degree of statistical rigour, but rarely using a Bayesian methodology. In our amended Gibbs Sampler scheme, however, we immediately have available an intuitively reasonable texture merging technique that can be justified in a Bayesian framework. For each texture, we have an adequate approximation to the marginal posterior density for the texture mean and precision parameters. Thus if the pair of posterior distributions for the parameters of texture T_k are negligibly different to the pair of posterior densities for texture T_l, we may merge textures

T_k and T_l into one at that stage, form pooled sample posterior densities and estimates for the mean and precision parameters of the new texture, and proceed with the iterative part of the Gibbs Sampler algorithm, with the number of textures K decreased by one. We do not qualify the phrase "negligibly different" at this stage, but merely note the following several straightforward techniques for assessing the dissimilarity between two probability distributions. First, we could compare one or more summary statistics (modal position, mean, moments) of each distribution and merge the two textures if the elements of the two sets of statistics differed only by a small percentage. Alternatively, we could compare the densities as a whole using some distance measure - the Kullback-Liebler distance (for which the calculation is tractable for Normal and Gamma probability densities), the L_1-distance (total absolute distance) and the L_2-distance (total squared distance) are three possible alternative measures. Another simpler merge technique proceeds as follows. Given the modal posterior estimates for the texture parameters, $(\hat{\mu}_k, \hat{\tau}_k)$ and $(\hat{\mu}_l, \hat{\tau}_l)$ for textures T_k and T_l respectively, consider the two Normal distributions $N(\hat{\mu}_k, \hat{\tau}_k^{-1})$ and $N(\hat{\mu}_l, \hat{\tau}_l^{-1})$ - without loss of generality assume that $\hat{\mu}_k > \hat{\mu}_l$. We may then calculate the minimum probability of mis-classification p_c (the tail overlap between the two densities) achieved when the optimum threshold value, x^*, is used, where clearly x^* satisfies $\phi_k(x^*) = \phi_l(x^*)$ (with $\phi_k(.)$ and $\phi_l(.)$ being the two Normal density functions for k and l respectively). It is easily seen that

$$p_c = \frac{1}{2}\left(\Phi\left(\hat{\tau}_k^{\frac{1}{2}}(x^* - \hat{\mu}_k)\right) + 1 - \Phi\left(\hat{\tau}_l^{\frac{1}{2}}(x^* - \hat{\mu}_l)\right) \right) \qquad (7.14)$$

where $\Phi(.)$ is the unit Normal distribution function. Hence we may evaluate expected error rates for such an equal mixture of these distributions. We might then consider merging the textures if p_c is large, as this would indicate that, in the segmentation context, it would be difficult to differentiate between pixels in the image from the two textures. We note that each of the texture merging techniques described above would add to the amount of processing time required - after each iteration, each remaining texture would be compared pairwise with every other, with the textures being merged if necessary. However, this additional time would be minimal in comparison to the amount of time spent per iteration of the Gibbs Sampler algorithm itself. Another more subtle texture merging approach that avoids the problems encountered in the techniques described above, and is more in keeping with the general formulation is as follows. In deriving the full conditional posterior distributions for texture mean parameters given by (7.8), we specified a prior distribution $[\mu_k]$ for $k = 0, \ldots, K-1$, and regarded the parameters μ_k as *a priori* independent. To encourage or discourage texture merging, we could alter this prior specification so that the μ_k were *a priori* dependent (i.e. change $[\mu_k]$ to $[\mu_k \mid \mu_{(k)}]$) and repeat the original analysis. A simple prior conditional dependence structure, for example reflecting restricted or ordered ranges for the parameters, would not complicate

the sampling procedure overly, and would also not add appreciably to the amount of computation required. It is also possible to use other intuitively reasonable and simple merge techniques and criteria in addition to those described above.

Having noted each of the points above concerning the implementation of the amended Gibbs Sampler algorithm, we now proceed with several illustrative examples of its application for different images.

(7.5) M.P.M. segmentation using amended Gibbs Sampler - examples.

In the examples we present below, we shall be specifically concerned with the following points. First, we shall study the adequacy of the algorithm over a range of Signal-Noise ratios. Secondly, we shall investigate the robustness of the various estimates at convergence to prior specification - we require that the algorithm (and indeed the methodology from which it is derived) must perform adequately when these initial estimates and naive classifications are poor. Thirdly, we shall study various aspects of the convergence of the algorithm, and compare the efficiency of several of the diagnostics described above.

(7.5.1) Two texture true scene.

We begin with a simple two texture example where the segmentation problem is relatively straightforward. Figure 117(a) depicts an image derived from the familiar circle true scene discretised into an 80×80 pixel grid, where independent additive Gaussian noise terms of variance 1.0 are combined with each true scene pixel value, with texture mean levels of 0.0 and 2.0 for the background and object respectively, giving a Signal-Noise ratio of 2.0. Figure 117(b) depicts a binary segmentation of the image achieved using the usual maximum-likelihood criterion with prior texture means 0.0 and 2.0 under a common variance assumption - this represents an optimal choice of parameters in terms of producing a minimum expected error rate.

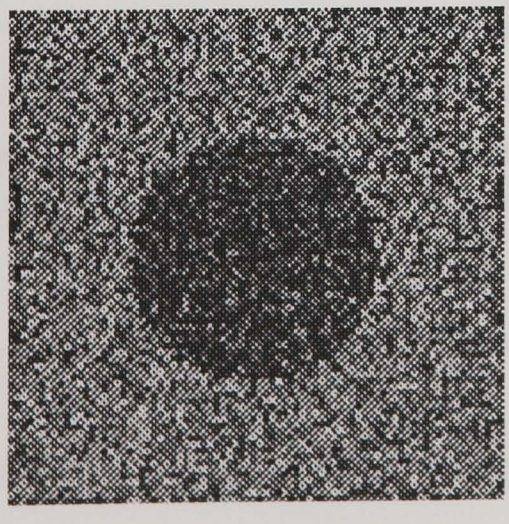


Fig 117(a) : image

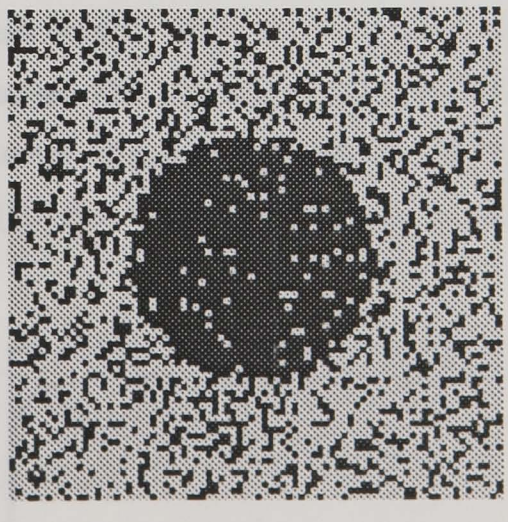


Fig 117(b) : two texture segmentation

It is clear that even in this relatively straightforward example, the non-spatial maximum-likelihood classifier does not produce an adequate segmentation. We now proceed with an implementation of the Gibbs Sampler algorithm in an attempt to achieve the M.P.M. solution to the segmentation problem.

In the two texture case, we must write down the forms of the full conditional posterior distributions and densities for each of the 6400 pixels in the true scene and the 4 texture parameters given in (7.6), (7.10), and (7.13) respectively (thus, essentially, we seek the solution to an estimation problem in 6404 parameters), and proceed to sample iteratively around the cycle until convergence. Prior to this, however, we must specify parameters in the prior distributions of the texture parameter, and also the interaction parameter β. In our example we choose these parameters in some arbitrary fashion, but we could easily have used the naive techniques discussed previously. For the texture means, we choose the mean, precision hyper-parameters in the Normal priors as $(0.0, 0.0001)$ and $(1.0, 0.0001)$ for textures T_0 (background) and T_1 (object) respectively - such a specification is sub-optimal in the sense described above, and reasonably "vague" (we could represent prior ignorance of the true texture mean levels by choosing prior precisions equal to 0.0) . For the texture precisions, we choose identical Gamma priors with hyperparameters $(2.0, 1.0)$ - thus, the prior modal position corresponds exactly to the true precision value 1.0, but still the prior information represented by these densities does not seem unrealistically detailed. For the moment, we choose β initially to take the value as proposed by Besag, namely 1.5, but we may have cause to review this choice at a later stage. We now present the results and segmentations obtained using this amended algorithm.

For demonstration purposes, we study the behaviour of a selection of the summary statistics mentioned above to obtain some idea of the nature of convergence for this relatively straightforward example. For the texture parameters, we track the posterior modal estimate in each case over the sequence of iterations, and for the true scene pixel classification parameters, we track the number of pixels allocated to each texture after each iteration, and the number of pixels changing classification on that iteration. The algorithm was allowed to run for 100 iterations in total. Figure 118 depicts the sequence of posterior modal estimates of the texture parameters for textures T_0 and T_1 plotted (on the vertical scale) against iteration number.

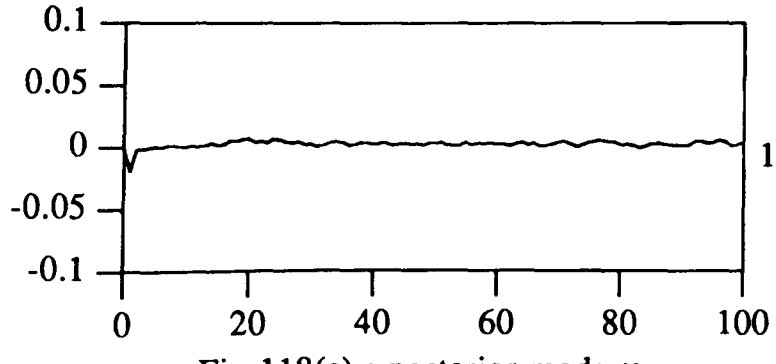


Fig 118(a) : posterior mode μ_0

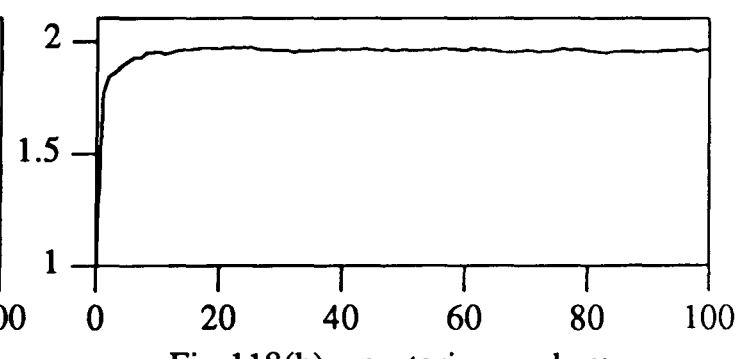


Fig 118(b) : posterior mode μ_1

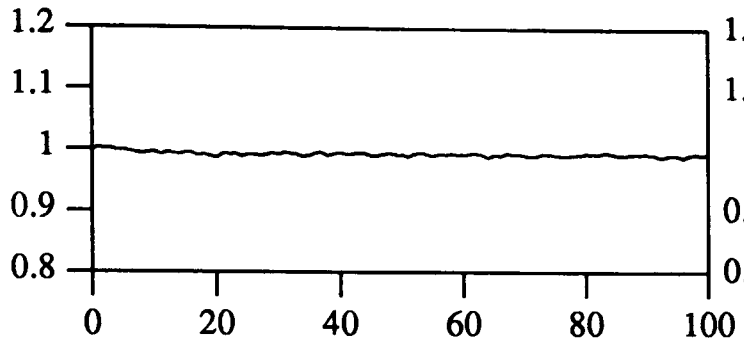


Fig 118(c) : posterior mode τ_0

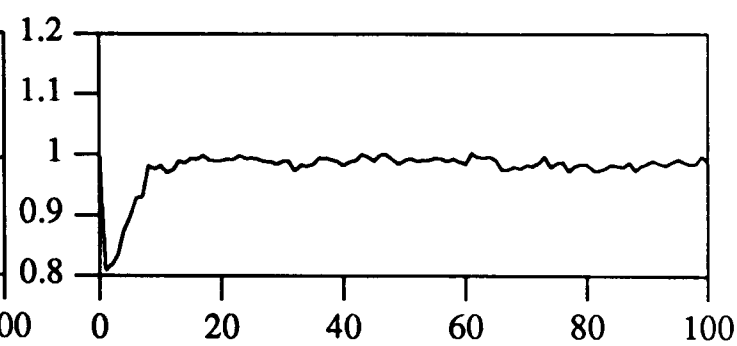


Fig 118(d) : posterior mode τ_1

It is clear that convergence of the parameters for T_0 occurs rapidly, the consequence of an accurate prior specification, and that convergence of the parameters for T_1 occurs within around 10 iterations. We adjudge convergence in this case by observing that the plots attain a constant horizontal level. It is also evident that the posterior modal estimate for parameter μ_1 is less than the true value of that parameter - around 1.96 as opposed to 2.0. This is no cause for concern, however, as it is again merely a consequence of the prior specification. The estimates obtained are generally excellent.

Figure 119 depicts plots of the sequences of numbers of T_0-allocated, n_0, and T_1-allocated, n_1, pixels in (a) and (b) respectively.

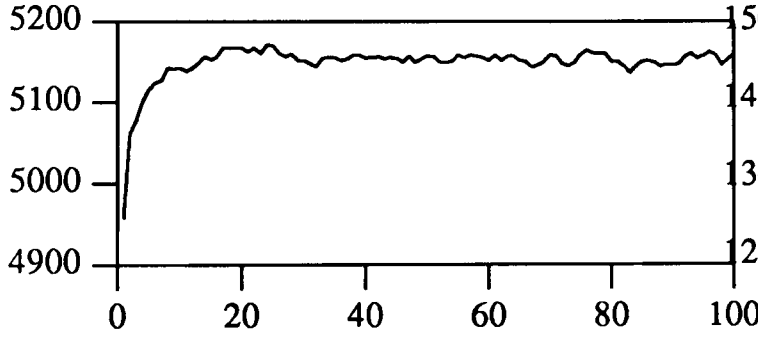


Fig 119(a) : n_0

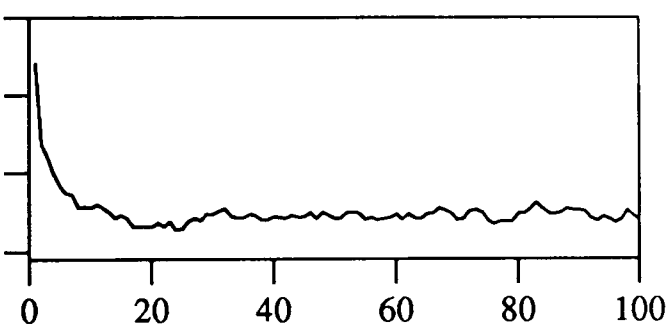


Fig 119(b) : n_1

The plots appear to stabilise at around 20 iterations, that is, significantly later than the plots in figure 118. Figure 120 depicts similar plots for the difference in the number of pixels allocated to each texture and the number of pixel flips at each iteration in (a) and (b) respectively.

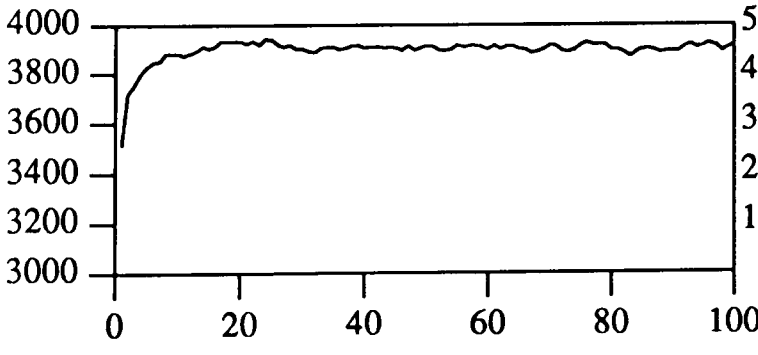


Fig 120(a) : $n_0 - n_1$

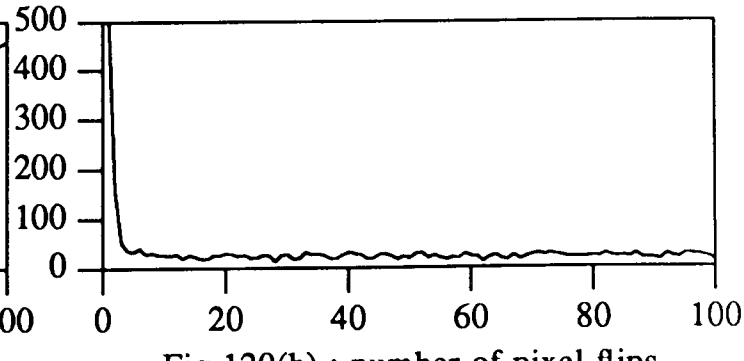


Fig 120(b) : number of pixel flips

The plot in (a) appears to stabilise again at around 20 iterations, whereas the plot in (b) stabilises much more rapidly, after only around 5 iterations. The contradictions implied by these plots demonstrate how difficult the assessment of convergence is for this algorithm.

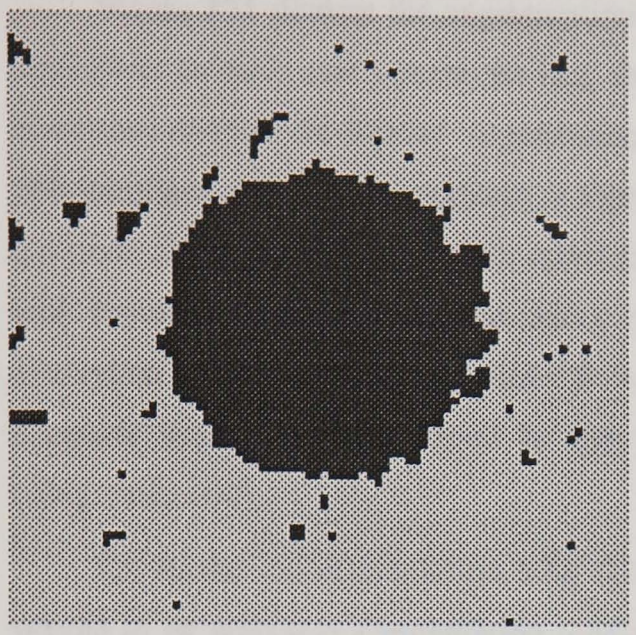


Fig 121(a) : 1 iteration

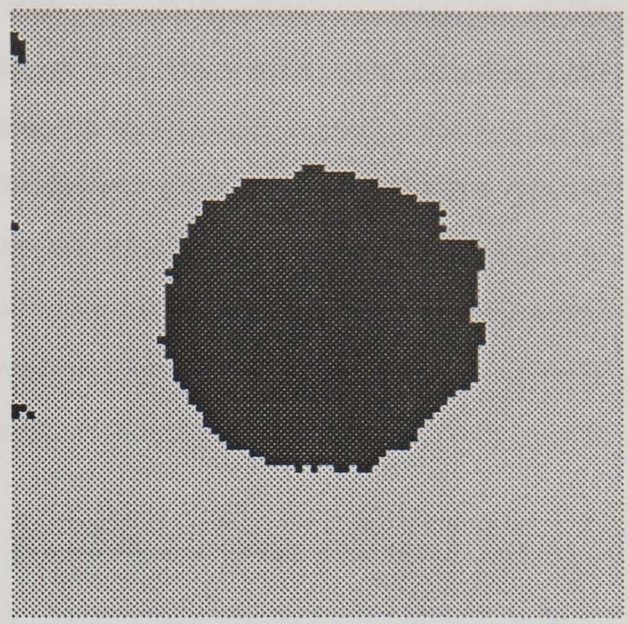


Fig 121(b) : 2 iterations

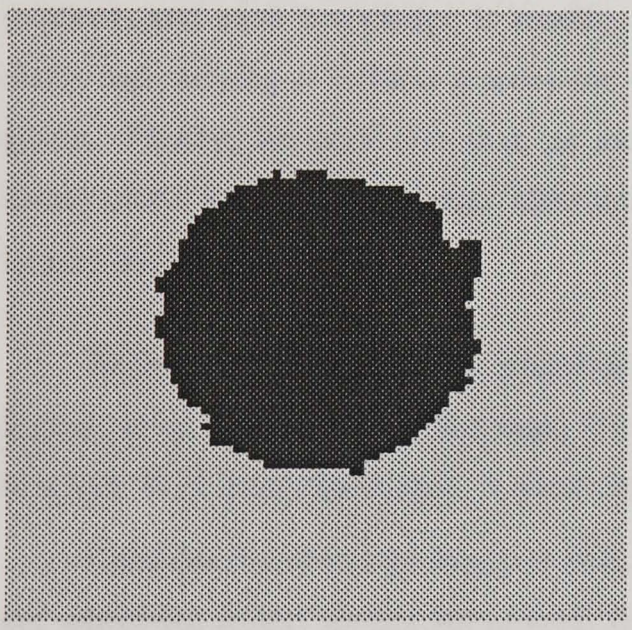


Fig 121(c) : 5 iterations

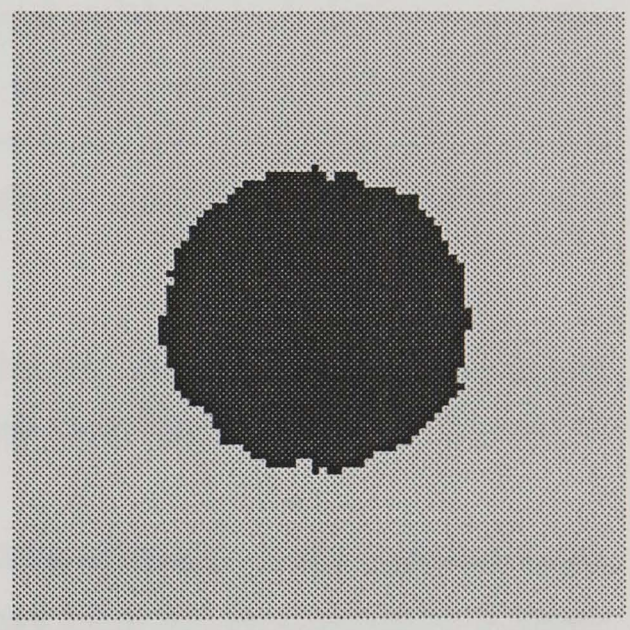


Fig 121(d) : 10 iterations

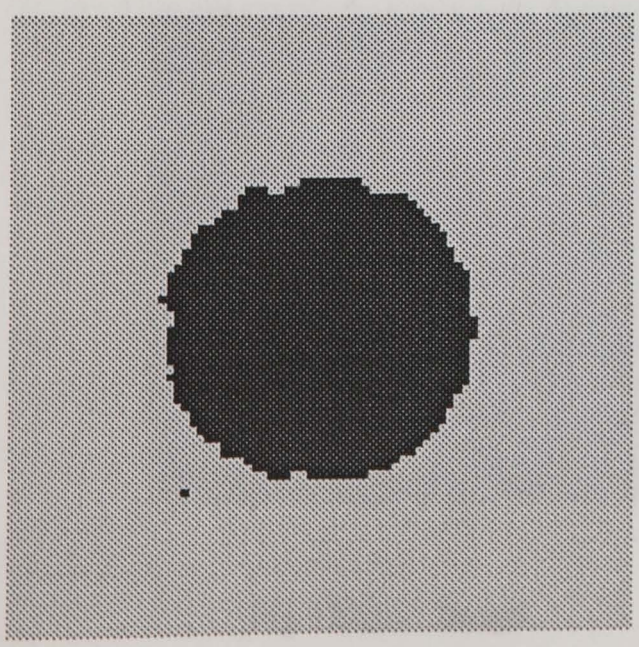


Fig 121(e) : 40 iterations

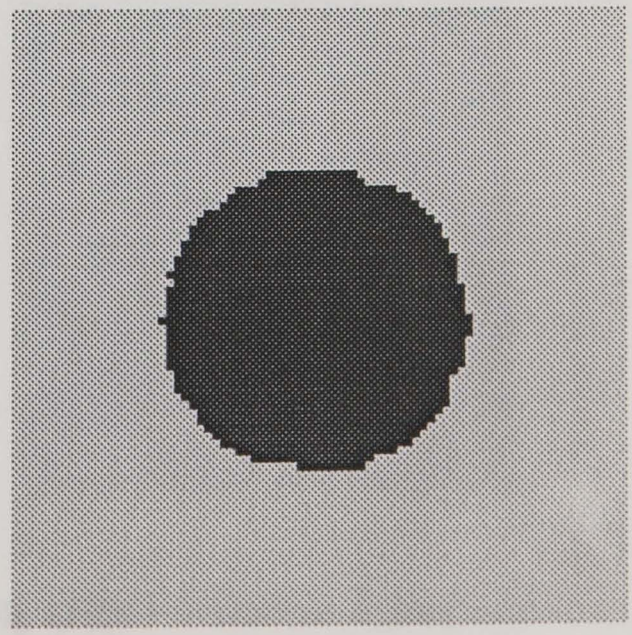


Fig 121(f) : 100 iterations

Figure 121 depicts the segmentations achieved after various numbers of iterations of the algorithm. In each case except the last, the representation is of the current (i.e. sampled) segmentation and not the M.P.M. segmentation at that iteration. In the last segmented image, after 100 iterations, the M.P.M. solution is represented, with the marginal posterior distributions for the pixel classification parameters approximated by the full conditional posterior distributions, with the conditioning parameters taking the values of those in the current segmentation (i.e. equivalent to number of replications $R = 1$). Convergence appears to occur rapidly, with an adequate segmentation achieved by the 10'th iteration (without the actual marginal modes being evaluated).

On further experimentation with this particular image, we found that the algorithm converged to an adequate segmentation (and a small error rate) for a wide range of prior parameters, provided that the maximum-likelihood classifier allocated sufficient pixels to each texture in the initial segmentation. Practically, this suggests that only very vague prior knowledge is required, and such knowledge is readily available from exploratory analysis. We note especially that varying β (within reason) for this image generally only alters the rate of convergence. Finally, we note also that, as mentioned above, the approximate marginal posterior distributions for the true scene classification parameters obtained via the algorithm were largely degenerate.

In the example above, the segmentation problem was relatively straightforward as the Signal-Noise ratio involved was high. We now proceed to investigate a more testing example in which the Signal-Noise ratio is somewhat lower.

Consider the identical image to that in figure 117(a) of derived from a circle true scene, but where the object texture mean-level is reduced from 2.0 to 0.5, hence producing a Signal-Noise ratio of 0.5. Figure 122(a) depicts such an image, and figure 122(b) depicts the optimal maximum-likelihood segmentation under a correct prior specification.

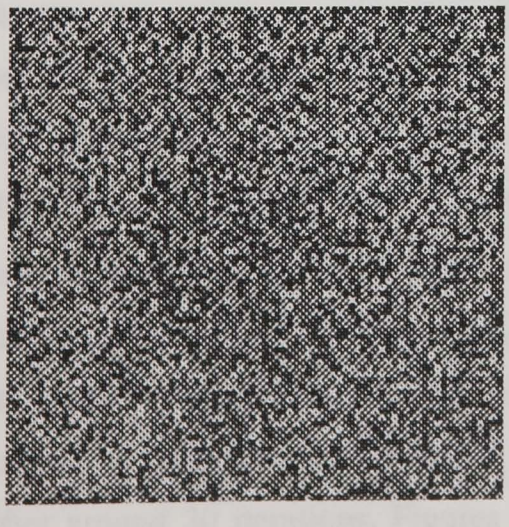


Fig 122(a) : image

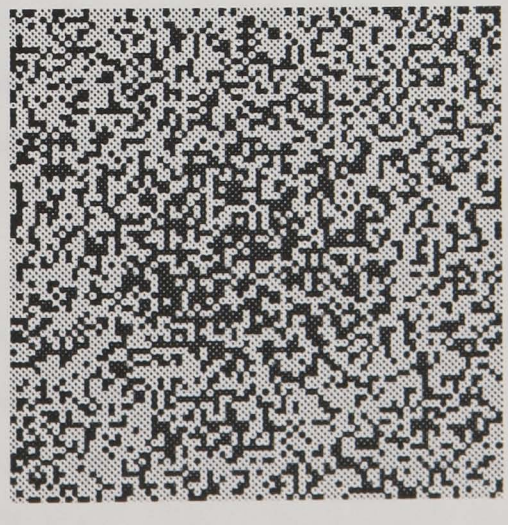


Fig 122(b) : two texture segmentation

Clearly this presents a very difficult segmentation problem, and a severe test of the Gibbs Sampler algorithm. On experimentation with images of this type, we find that several features that we did not need to consider previously assume greater importance. For example, the marginal posterior distribution estimates for the image in figure 117(a) were predominantly degenerate, and thus we were able to track the current (sampled) classification parameter values rather than the marginal modal estimates without any great loss in accuracy. For images such as that in figure 122(a), however, the marginal distribution estimates are not generally degenerate, and thus frequently the sampled and modal values for these distribution do not coincide. Hence, for definiteness, we now record the modal estimates instead of the "current" values, and use them in an identical way to that described above and demonstrated in figure 118, to aid in the diagnosis convergence Also, for images in which the Signal-Noise ratio is low, the segmentations obtained are extremely sensitive to the prior specification. We have already mentioned how the correct choice of β is important in all applications of the Gibbs Sampler algorithm, but that when the Signal-Noise ratio is high (and discrimination dominates smoothing), the margin for error in this choice is quite large. By the same argument, when the Signal-Noise ratio is low (and the discrimination is inconclusive), we would expect a correct choice of β to be crucial. Similarly, it is clear that the texture parameter prior distributions must be specified with some degree of adequacy, as in the example above, so that sufficient numbers of pixels are correctly classified in the initial segmentation.

Noting each of these features, we now investigate the performance of the Gibbs Sampler algorithm for the image in figure 122(a). To discover if any acceptable segmentations can be achieved such a case where the noise corruption is high, we first specify prior distributions for the texture parameters which correspond closely to their true values. For the texture means, we choose the mean and precision parameters in the Normal prior distributions to be $(0.0, 0.1)$ for μ_0 and $(0.6, 0.1)$ for μ_1, and for the texture precisions, we again choose identical Gamma priors with parameters $(2.0, 1.0)$ - this might be regarded as unrealistically specific as a practical example, but serves our demonstration purposes. We discuss the of β in more detail below, but for the moment, for reasons that will become apparent, we choose $\beta = 0.5$. Figure 123 depicts plots of the successive posterior modal estimates over 100 iterations of the subsequent implementation of the amended algorithm.

Here, the estimates of the texture T_0 parameters are well behaved, due again to the accuracy of the prior specification. The estimates for texture T_1, however, are less well behaved and more wildly fluctuating, due in part to the smaller "sample size" (the number of pixels in the true scene) for that texture. Generally, however, we would diagnose the algorithm as converged after around 20 iterations. Figures 124 and 125 depict plots of pixel counts and pixel flips.

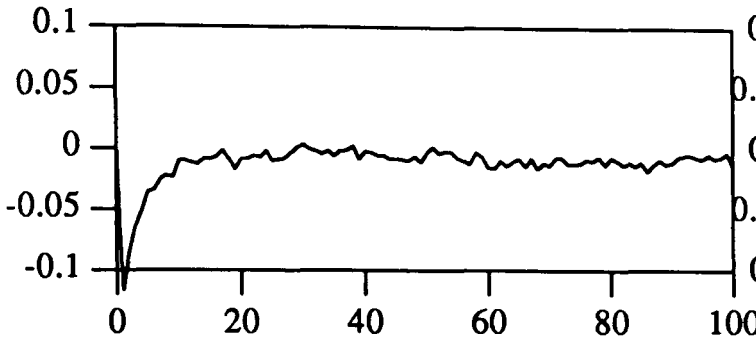


Fig 123(a) : posterior mode μ_0

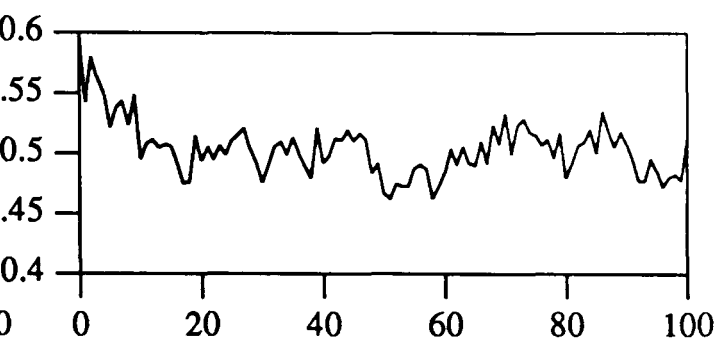


Fig 123(b) : posterior mode μ_1

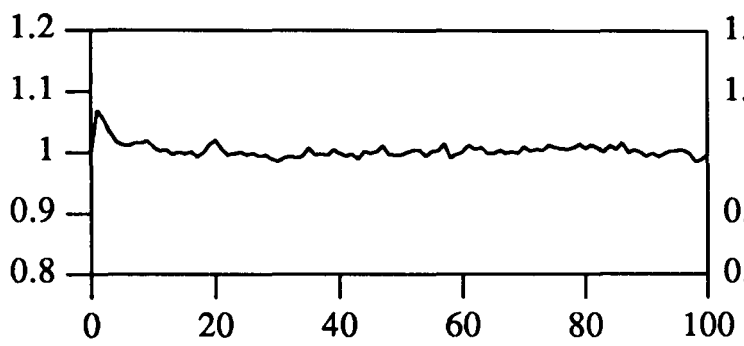


Fig 123(c) : posterior mode τ_0

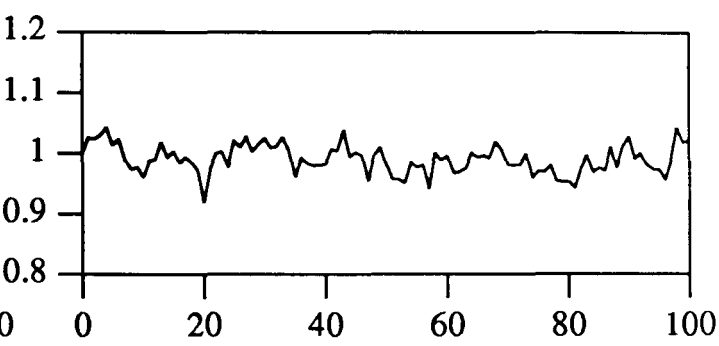


Fig 123(d) : posterior mode τ_1

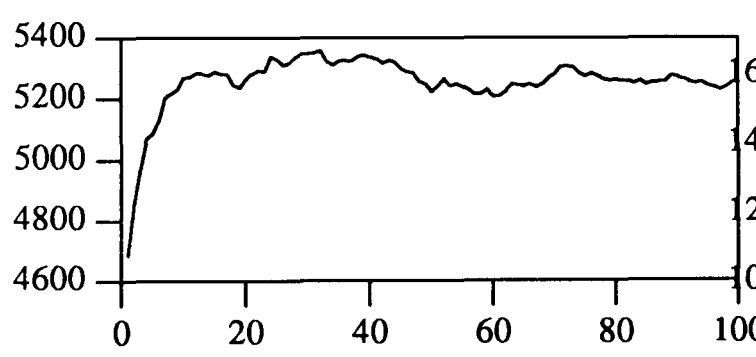


Fig 124(a) : n_0

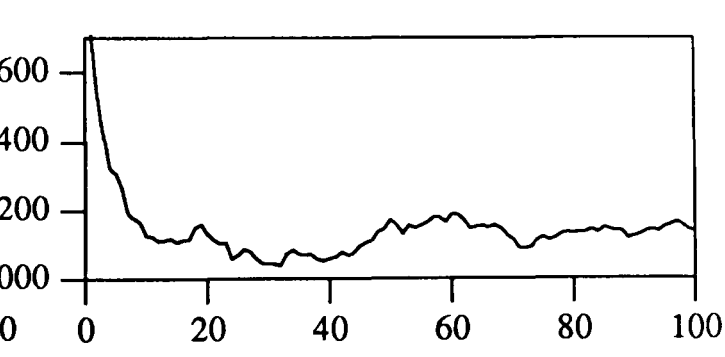


Fig 124(b) : n_1

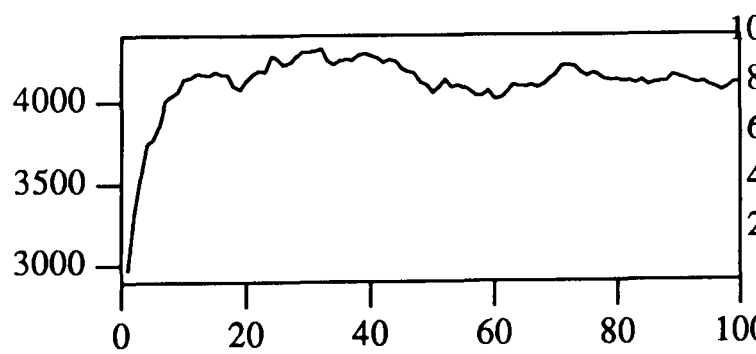


Fig 125(a) : $n_0 - n_1$

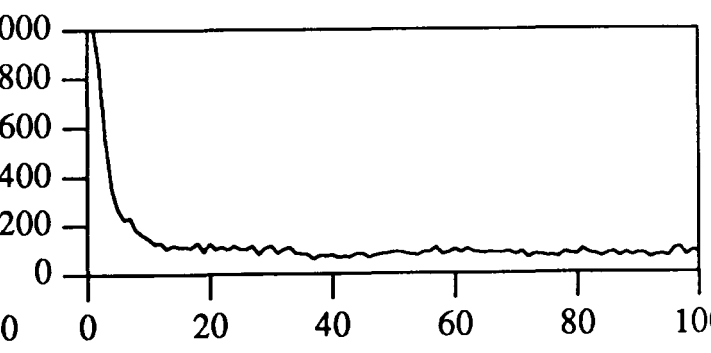


Fig 125(b) : number of pixel flips

Again, we draw the conclusion that the algorithm has converged after around 20 iterations. However, if we now inspect the two results at iteration 20 and iteration 100, we see that segmentations are in fact markedly different. Figure 126 depicts these two segmentations.

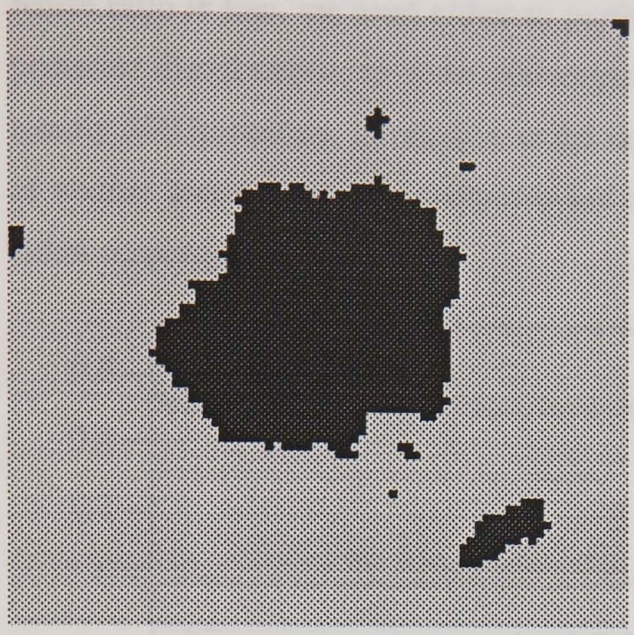


Fig 126(a) : 20 iterations

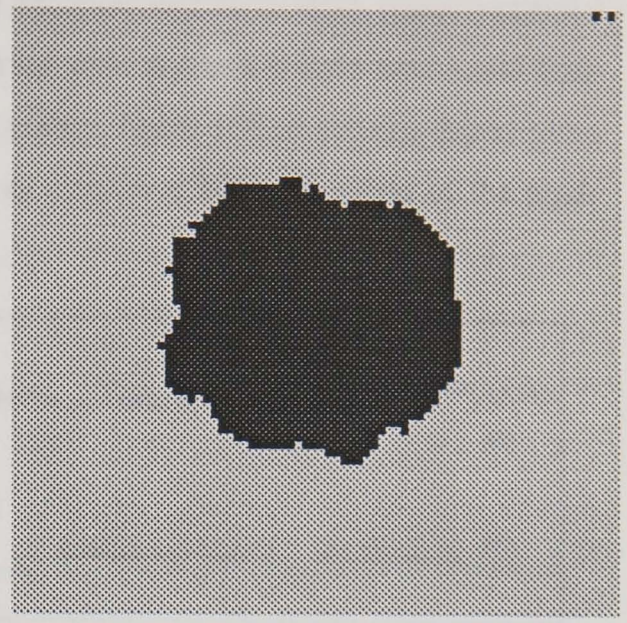


Fig 126(b) : 100 iterations

The total numbers of pixels allocated to the two textures in each of figure 126(a) and (b) are approximately equal, as are the respective texture parameter estimates. Clearly, however, the second segmentation is superior to the first. This illustrates the difficulty that the assessment of convergence of the iterative Gibbs Sampler procedure presents for images in which the Signal-Noise ratio is low, and the marginal posterior distributions for many of the classification parameters are not degenerate. Furthermore, consider figure 127(a) and (b), which depict the segmentations achieved after 40 and 60 iterations respectively.

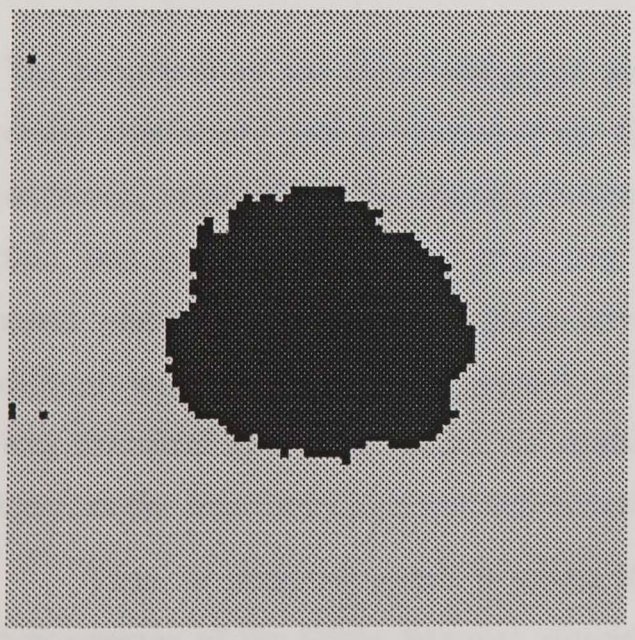


Fig 127(a) : 40 iterations

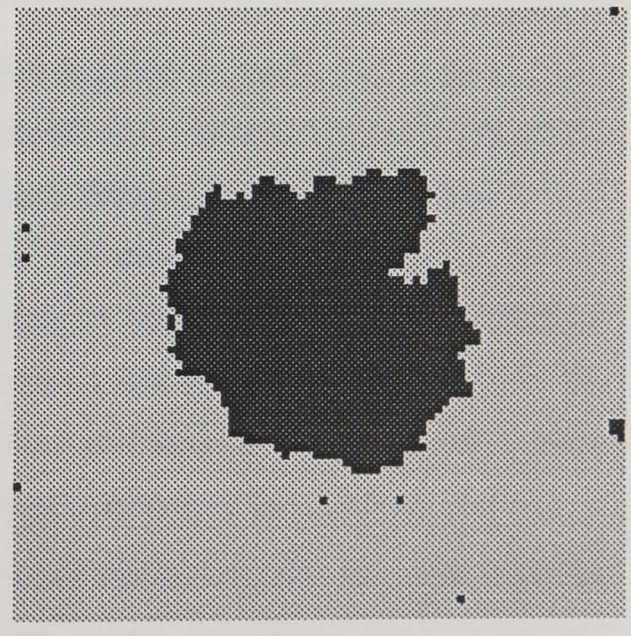


Fig 127(b) : 60 iterations

It is clear that, visually, the segmentation in (a) is more adequate than that in (b), despite the fact that it was achieved in a fewer number of iterations. This has profound implications for our ideas concerning convergence diagnostics, and suggests that for images which the Signal-

Noise ratio is low, we need either to perform replicate analyses and form more stable marginal posterior distributions, or view convergence as occurring over a much longer time-scale than merely 100 iterations. In relation to the latter of these points, figure 128(a) and (b) depict plots of texture mean parameter estimates and the quantity $n_0 - n_1$ over 1000 iterations of the algorithm respectively.

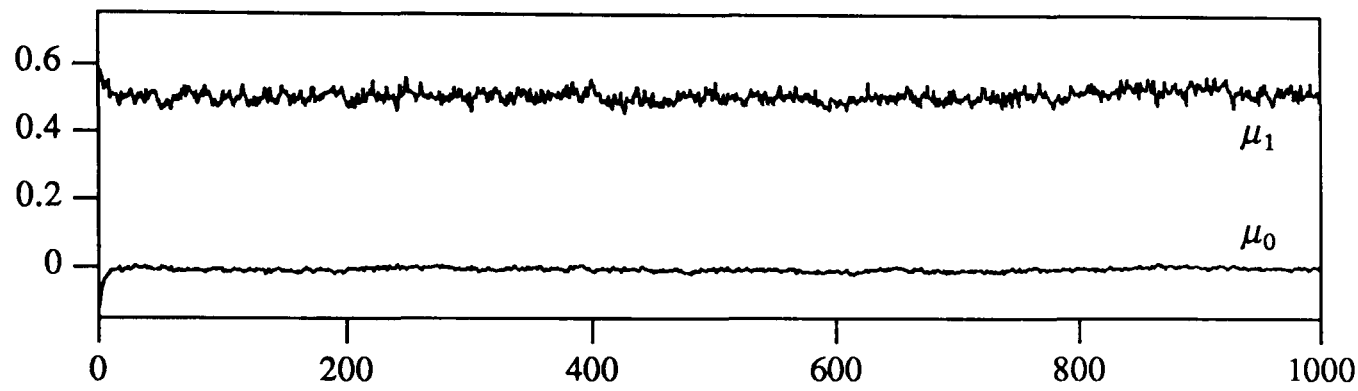


Fig 128(a) : texture mean parameter estimates

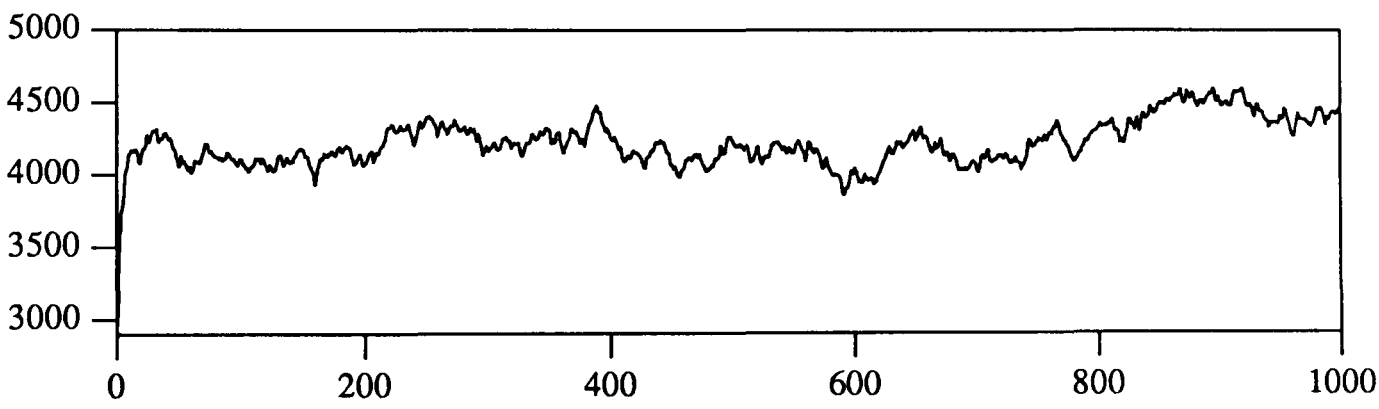


Fig 128(b) : $n_0 - n_1$

The behaviour of the texture mean parameter estimates depicted in figure 128(a) seems generally satisfactory over the extent of the 1000 iterations. However, toward the latter part of the sequence, value of $n_0 - n_1$ appears to undergo an upward change in value. On further inspection, this appears only to be a short-term phenomenon, but serves to point up the fact that in cases where the Signal-Noise ratio is low the assessment of convergence will be generally a difficult task. Clearly, were we to perform replicate analyses, and form density estimates in the way described by Gelfand *et al.*, we might expect that the task may be made moderately more easy, as the sequence of texture and classification parameter estimates will be more stable.

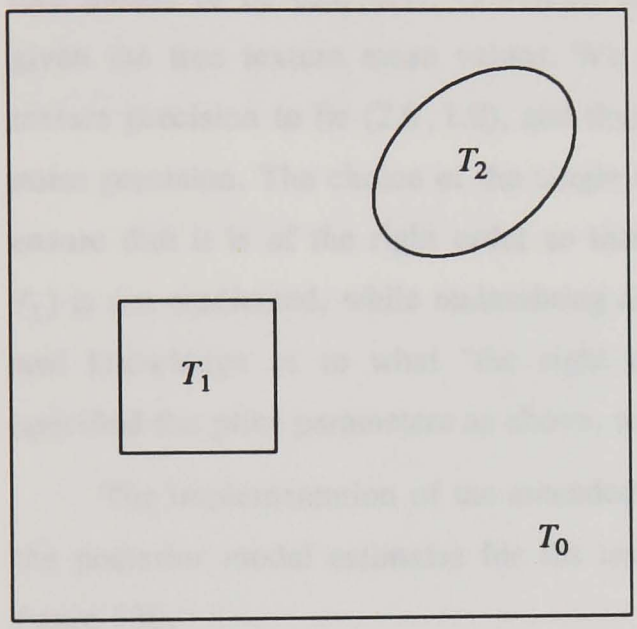
Thus, in practical terms, it seems that the problems concerned with the assessment of convergence of the Gibbs Sampler algorithm are, as yet, largely unresolved. Each of the criteria discussed above, along with others such as studying the relative positions of pixels that change classification between iterations frequently, or studying modal probability values, can be regarded as being of some use, but ultimately such assessments we still be made in some subjective fashion, perhaps only after a visual inspection of the segmentation obtained. This is an unsatisfactory situation, but until further investigation of the mathematical nature of the convergence of the algorithm has been carried out, we appear to have little alternative. We

console ourselves with the fact that, whenever the Signal-Noise ratio is sufficiently large (greater than 1.0 say, which we may discern from the posterior distributions for the texture means and precisions), we may run the algorithm for a fixed number of iterations, 50 say, then stop deterministically, and still obtain an adequate segmentation.

The problems discussed above arise predominantly through our lack of mathematical understanding of the global behaviour of the posterior field induced by our (Gibbs) prior specification and its interaction with the likelihood function that represents the relationship between the signal and the observed data at the pixel level. We mentioned briefly in the introductory chapter several ways in which, for simple examples, the parameters in the conditional prior distribution may be chosen to avoid critical or super-critical behaviour of the prior field globally. Such issues, however, are only of relevance to the dogmatic Bayesian, one who believes that his prior specification should reflect solely his prior opinions, without reference to how such a specification will interact with the data - here, we must be primarily concerned with the nature of the posterior field, as it is from this that we are to make inferences. Titterington (1986) suggests cross-validation as a possible non-Bayesian solution to this problem. We feel that such an adaptive approach is attractive, but unacceptable in this form. However, a adaptive solution (of sorts) to this problem may be as follows. Recall our amended Gibbs Sampler approach to the problem of simultaneous segmentation and parameter estimation. There, we merely included the relevant conditional posterior distributions for the texture parameters, given by (7.7) and (7.11), in the sampling cycle, and iterated to "convergence". Now, by simply adding another level to the hierarchy, we may regard the parameters in the conditional prior distributions as unknowns, write down a suitable form for their conditional (posterior) distributions, and include these posterior distributions in the sampling cycle also. Intuitively, we should thus obtain an adaptive scheme that prevents the posterior field behaving in a super-critical fashion. For example, in the simple two texture problem presented above, with interaction parameter β, we might write down the conditional distribution $[\beta \mid Y, \mu, \tau, \theta]$, and include it in the sampling cycle. Under certain simplifying assumptions, for instance that knowledge of the true scene parameter values dominates that of the data values, this conditional distribution may be simplified to $[\beta \mid \mu, \tau, \theta]$ and so on. We suggest that the functional form of $[\beta \mid \mu, \tau, \theta]$ should depend on θ through the number of like pairs of adjacent pixels or equivalently through the number of isolated pixels in some way in the current segmentation, in some way. We make such suggestions due to their intuitive appeal (we would like β to be large when the segmentation is "rough" and small when the segmentation is "smooth"), and because it is widely reported by other authors that the value of such smoothing parameters should be altered at various stages of the iterative procedure, but unfortunately we can offer no information as yet as to their effect on the convergence of the Gibbs Sampler algorithm.

(7.5.2) Multiple texture true scenes.

We now give one more brief example of the use of the amended Gibbs Sampler algorithm in a segmentation problem where the true scene is known to contain three textures, namely a background and two distinct objects. Figure 129(a) depicts the underlying true scene, where the three texture mean levels are 0.0, 1.5, and 2.5 for textures T_0, T_1 and T_2 respectively. Figure 129(b) depicts the image obtained when this true scene is corrupted by random, independent and additive zero-mean noise terms of precision 1.0, with the image discretised into an 80×80 grid of pixels.



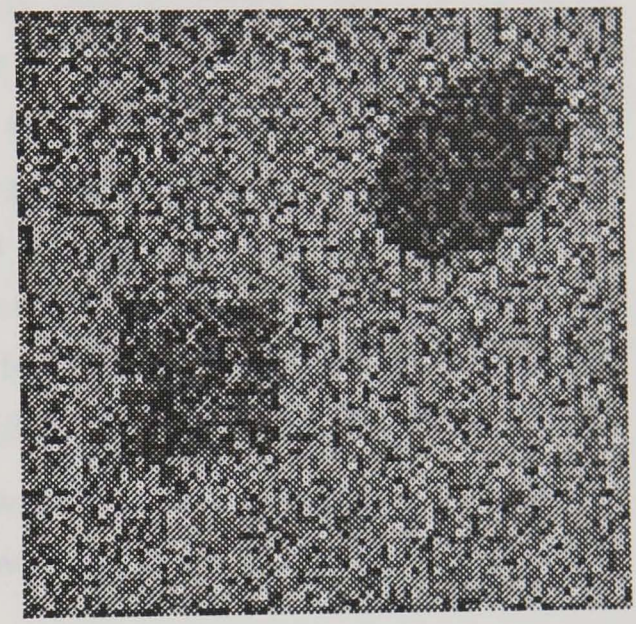


Fig 129(a) : true scene

Fig 129(b) : image

We feel that this is a relatively straightforward example, and yet is suitable to demonstrate the performance of the algorithm.

For this particular example, the Gibbs Sampler algorithm was implemented under a common texture precision assumption, that is, the three texture precision conditional posterior distributions given by (7.11) were replaced by a single conditional posterior distribution $[\tau \mid Y, \mu, \theta]$ for precision parameter τ. It is easily seen that, in place of (7.12), we have that

$$[\tau \mid Y, \mu, \theta] \propto \prod_{k=1}^{K} \prod_{(i,j) \in T_k} [Y_{ij} \mid \mu_k, \tau][\tau] \tag{7.14}$$

where $[\tau]$ is the prior distribution for τ. If we again choose a conjugate prior form for $[\tau]$, say $[\tau] \equiv \mathrm{Ga}(\alpha, \delta)$, then it is easy to show that

$$[\tau \mid Y, \mu, \theta] = \mathrm{Ga}\left(\alpha + \frac{n^2}{2}, \delta + \frac{SSQ}{2}\right) \tag{7.15}$$

where $SSQ = \frac{1}{n^2} \sum_{k=1}^{K} \sum_{(i,j) \in T_k} (Y_{ij} - \mu_k)^2$.

The sampling cycle now contains $n^2 + K + 1$ full conditional posterior distributions. We must now specify a total of six prior parameters for the texture means, and two for the texture precision. In this implementation, for convenience, we assume the prior distribution in (7.5), involving only one parameter, although we accept that it would be perhaps more satisfactory to specify a more complicated prior form in the three texture situation. For demonstration purposes, we choose the mean and precision parameters in the Normal priors for the texture means to be (0.0,0.1), (1.0,0.1), and (2.0,0.1). This represents a sub-optimal choice given the true texture mean values. We choose the parameters in the Gamma prior for the texture precision to be (2.0,1.0), and thus the prior mode corresponds precisely with the true noise precision. The choice of the single interaction parameter β needs care. We must try and ensure that it is of the right order so that the lower mean-valued texture region (in this case T_1) is not eradicated, while maintaining an adequate convergence rate. Naturally, we have no real knowledge as to what "the right order" is, but on previous experience, and having specified the prior parameters as above, we feel that $\beta = 1.0$ seems a sensible choice.

The implementation of the amended Gibbs Sampler algorithm produced the behaviour of the posterior modal estimates for the texture parameters over the 200 iterations depicted in figure 130.

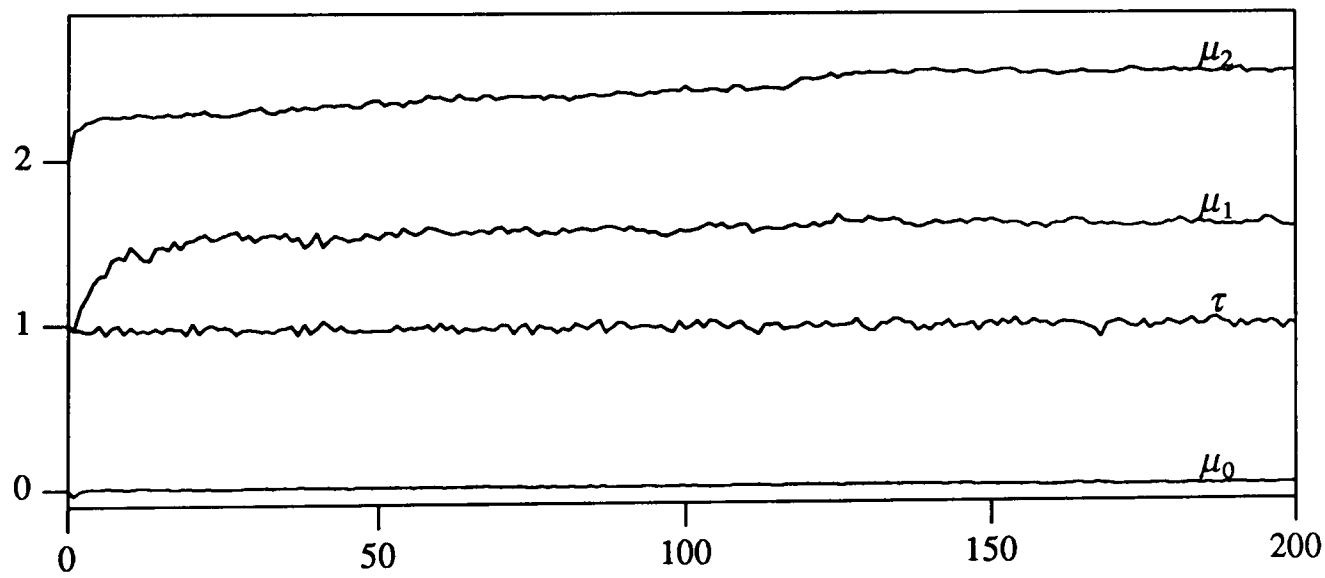


Fig 130(a) : texture parameter posterior estimates

It is clear that each of the parameters has converged before the 200 iterations are complete. One notable feature is the behaviour of the modal posterior estimate for the mean parameter for texture T_2. There appears to be an upward trend in the corresponding plot, until around 120 iterations, when the values of successive estimates stabilise.

The "current" segmentation was recorded after 25, 50, 75, 100, 150, and finally 200 iterations, and is depicted in figure 131. The sequence is extremely interesting. Throughout,

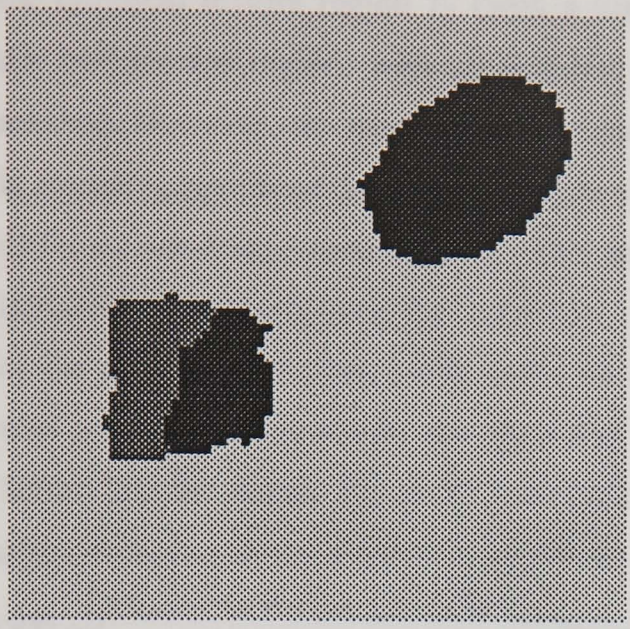


Fig 131(a) : 25 iterations

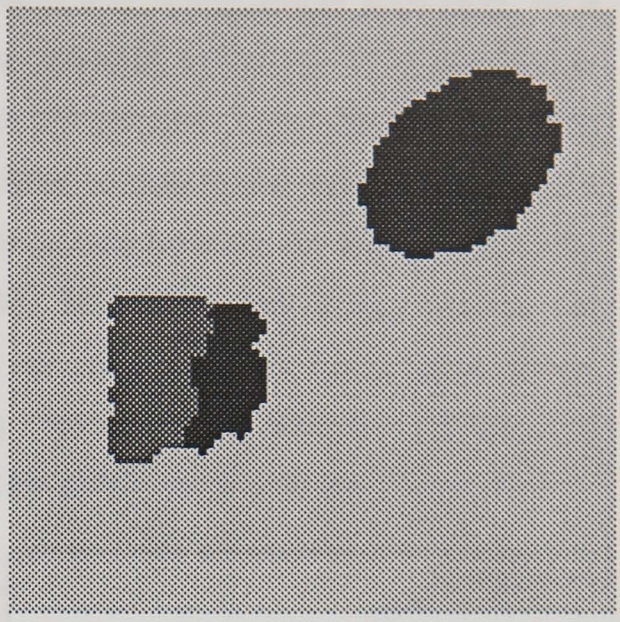


Fig 131(b) : 50 iterations

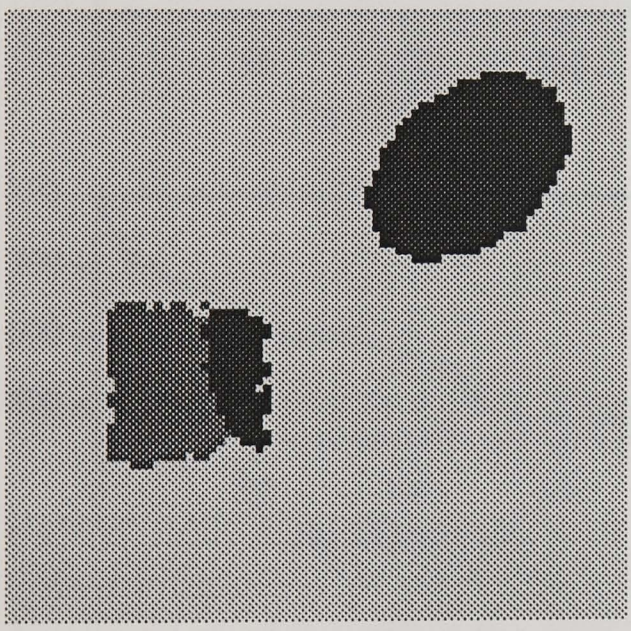


Fig 131(c) : 75 iterations

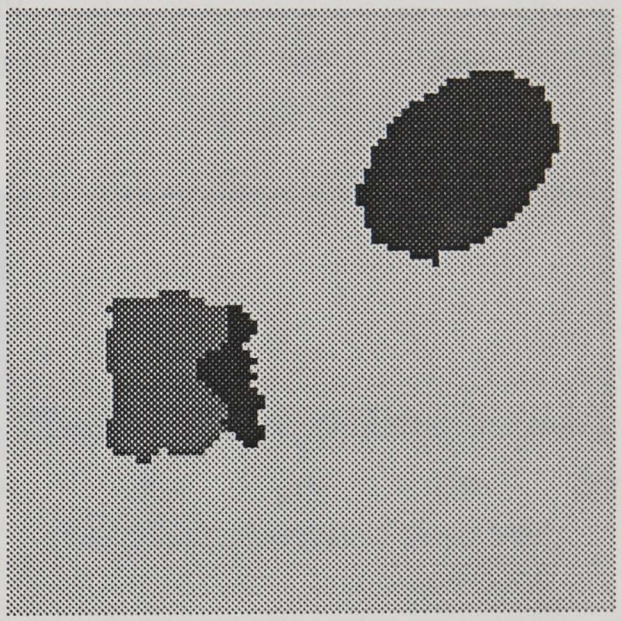


Fig 131(d) : 100 iterations

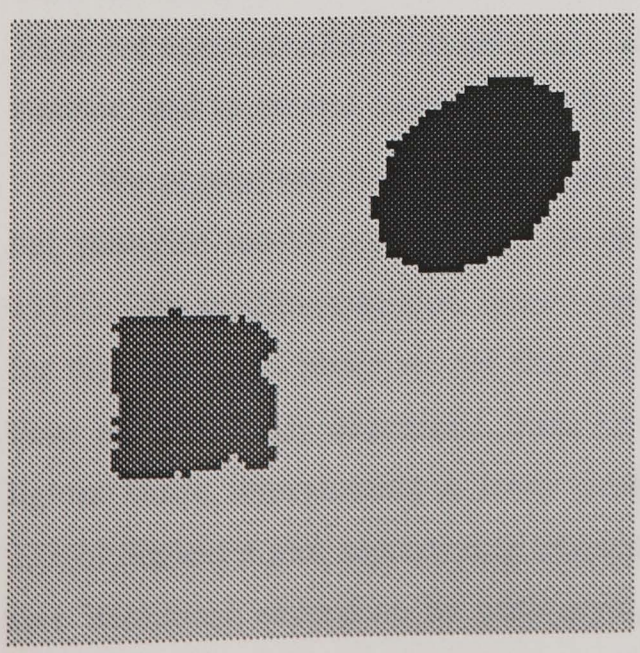


Fig 131(e) : 150 iterations

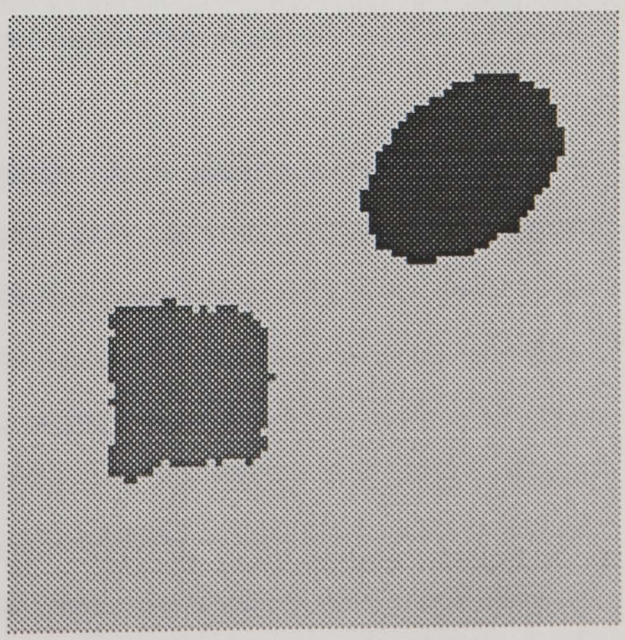


Fig 131(f) : 200 iterations

the elliptical texture region T_2 (with a Signal-Noise ratio 2.5) is reproduced accurately, as is the background region, with all isolated pixels being removed before the 25'th iteration is reached. However, the square texture region T_1 (where the effective Signal-Noise ratio is 1.5) is initially partitioned into two, with some pixels interior to that region being allocated to a different non-background texture. This partition persists until between the 100'th and 150'th iterations (thus explaining the behaviour in figure 130). We believe this to be a consequence of the prior specification, particularly the choice of a one parameter prior for the pixel allocation parameters. However, the algorithm appears to converge to an adequate segmentation before 200 iterations. Figure 132 depicts the plots of the pixel counts over the 200 iterations.

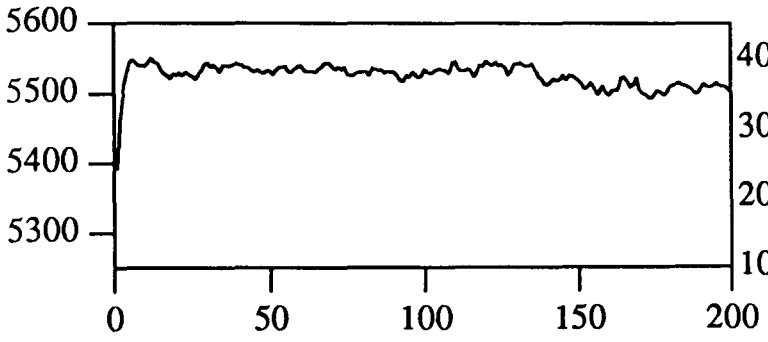


Fig 132(a) : n_0

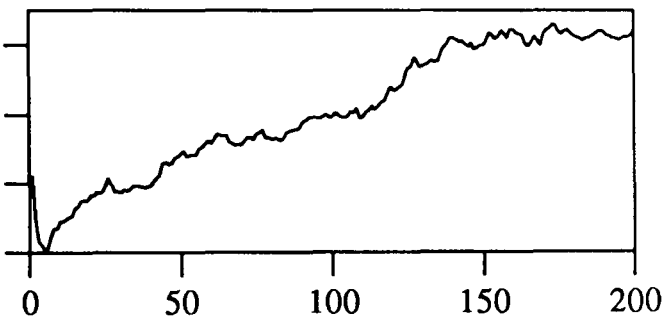


Fig 132(b) : n_1

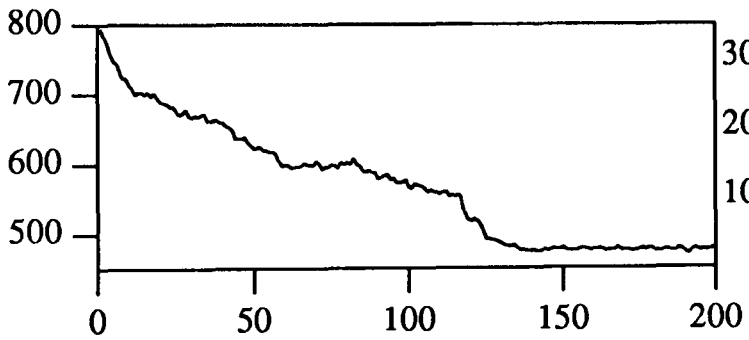


Fig 132(c) : $n2$

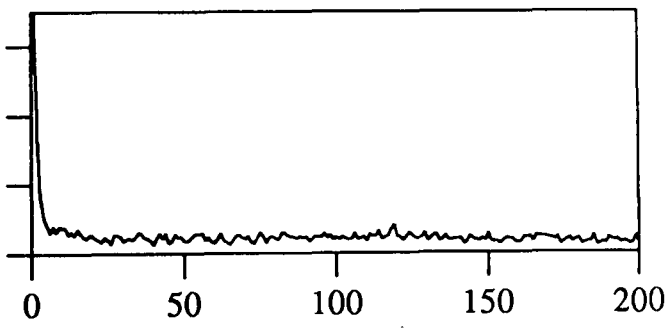


Fig 132(d) : number of pixel flips

Again, whereas the number of background pixels, n_0, remains approximately constant, the number of pixels allocated to textures T_1 and T_2 only stabilise after around 129 iterations. Note also that the recorded number of pixel flips stabilise rapidly, after around 25 iterations. Thus, we again conclude that the convergence should not merely be assessed via one summary statistic, and in general great care should be taken.

We present two final pixel images of interest. Figure 133(a) depicts a plot of the marginal posterior modal probabilities for the pixel classification parameters depicted in figure 131(f). The predominant dark colour represents a posterior probability of greater than 0.995 for that pixel, and hence we see that in the majority of cases, the marginal posterior distributions are practically degenerate. Note how the texture region boundaries are picked out in a lighter shade, representing a lower posterior probability for the pixels concerned. Figure 133(b) depicts a plot representing the number of flips (changes of allocation between successive iterations) undergone by each pixel, with a dark shade representing a high number of flips. Note here how again the texture boundaries are picked out, and note also that pixels

internal to texture T_2 changed allocation infrequently, whereas those internal to T_1 changed markedly more often, corresponding to the removal of the partition as described above. We believe plots such as these made at regular periods during the analysis may be of use in assessing the convergence of the algorithm.

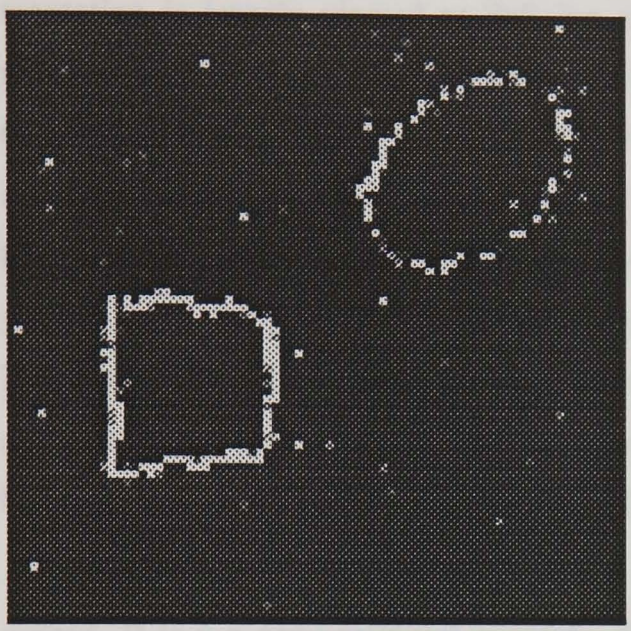


Fig 133(a) : posterior probability

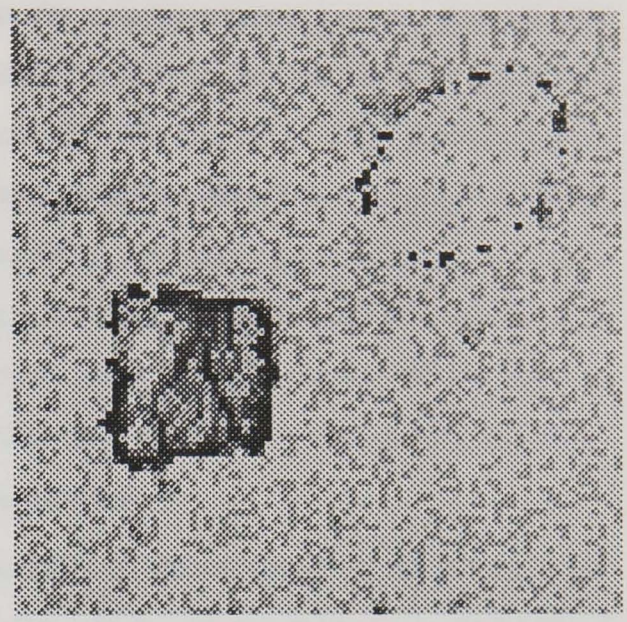


Fig 133(b) : number of flips

We note one interesting fact concerning the analysis of the image in figure 129(b). If the standard Gibbs Sampler algorithm is implemented with the same prior specification and parameters, but without the texture parameters being updated between iterations (either stochastically or using any other adaptive estimation technique), the segmentation obtained at convergence comprises only two textures, T_0 and T_2, with no pixels allocated to texture T_1. Thus, generally, we might regard our amended algorithm to be superior to the non-adaptive original version.

This completes the test examples that we present, and we conclude with some general remarks. We have studied examples of two and three texture true scenes, and investigated the adequacy of the amended Gibbs Sampler in each case. The extension to true scenes with larger numbers of textures is obvious, but with it goes the added complexity of the specification of the prior distributions for the texture parameters and the spatial prior for the pixel classification parameters, and also the increased numbers of statistics to use in the assessment of convergence. We also note here that for single texture true scenes the amended Gibbs Sampler produced adequate results (i.e. converged to a single texture segmentation), even when the prior texture means were chosen to be close together, and the amount of smoothing was minimal, independent of however many textures were initially surmised, and without the need for reference to any of the texture-merging techniques described above. Thus, although we have encountered several practical difficulties, we have reason to be generally satisfied with the performance of the amended Gibbs Sampler algorithm.

Given the emphasis of the earlier chapters, another natural extension to the Gibbs Sampler algorithm that has been suggested by many authors would be to incorporate knowledge or opinion concerning edge positions. The edge-process "dual" to the pixel-intensity process in the interpretation of the Markov Random Field prior forms in image segmentation is largely viewed as defining subtleties in the neighbourhood system, that is, two adjacent pixels are not regarded as neighbours if an edge-section is present between them (see Geman and Geman (1984) for a full discussion). Practically, the use of an edge-process within a Gibbs Sampler algorithm should help to preserve the boundaries between textures, and possibly even aid in improving the apparent rate of convergence, and our own preliminary investigations in in this area, and the work of other authors, seem to suggest that this is the case. Intuitively, our probabilistic formulation of the edge-detection problem as one in changepoint analysis seems to lend itself well to the use of an edge-process in this way, providing initial locations and corresponding probabilities for the edge-sections.

Finally in this chapter, we present a worked example to demonstrate the techniques discussed. It is a a semi-artificial example, the interest for which was motivated by the work of Ripley (1988, chapter 5), in which we shall demonstrate the implementation of the edge-detection routines and the use of the amended Gibbs Sampler algorithm for segmentation.

(7.6) Worked example - Ireland.

Figure 134(a) depicts a 64×64 pixel discretised version of a map of Ireland, reproduced from Ripley (1988, p102). It is used there as an example true scene to demonstrate the use annealing and related techniques. Figure 134(b) depicts an image derived from the binary true scene when Gaussian zero-mean error terms are added independently to each pixel - here, the two textures T_0 (light) and T_1 (dark) are homogeneous, with mean-levels 0.0 and 1.0 respectively, and the noise standard deviation was chosen to be 0.65, identical to that used by Ripley in his demonstration.

Fig 134(a) : true scene

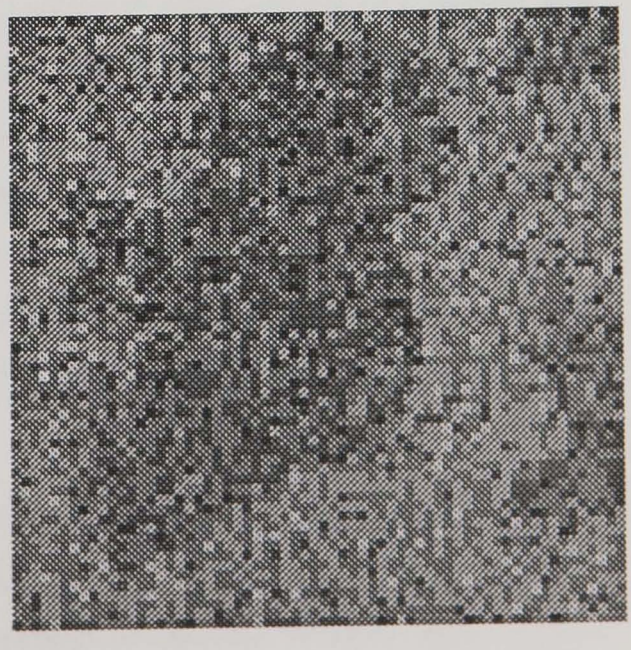


Fig 134(b) : image

This presents an interesting example on which to test the edge-detection and segmentation routines that we have developed, for a number of reasons. First, for solution of the edge-detection problem using a changepoint formulation, we should really be forced to evaluate multiple changepoint posterior distributions, as the nature of the true scene is complex. Secondly, the level of noise-corruption should render the edge-detection and segmentation problems relatively difficult. Thirdly, we have the results and comments of Ripley available for immediate comparison.

We shall assume that any pre-processing (for example, noise-reduction) has been carried out prior to any analysis that we shall perform. We begin as if in complete prior ignorance of the true scene, other than the knowledge that it contains two distinct textures, and the noise-terms are identically and independently normally distributed. Hence, first, we carry out an edge-detection analysis using changepoint techniques. For comparison, we analyse the image using posterior distributions derived under assumptions of one and two changepoints, and using the binary segmentation technique and the approximate inference procedures derived using the Gibbs Sampler approach for one, two, and three changepoint models, each of which we presented in chapter 3. We assume the same prior specification as in our previous analyses, that is, with the texture mean-levels and noise standard-deviation being regarded as unknown parameters about which we are *a priori* ignorant, and thus for which we specify a non-informative prior distribution. Figure 135 depicts the raw results of these various analyses. The posterior modal positions and values are represented using the symbols introduced in chapter 2.

The six plots in figure 135 result from the following changepoint-based analyses. Figure 135(a) depicts the results of a full analysis using the one changepoint posterior distribution (2.11). Figure 135(b) depicts the results of the binary segmentation type technique described in chapter 3, under the same prior specification as in the derivation of (2.11). Figure 135(c) depicts the results obtained when the two changepoint posterior distribution (3.2) is used to derive exact marginal distributions for the two changepoints. Figures 135 (d), (e) and (f) depict the results obtained when the Gibbs Sampler algorithm is used to obtain approximate marginal posterior distributions for one, two and three changepoints respectively, also as described in chapter 3, again under the same prior assumptions as in the derivation of (2.11). For demonstration purposes, the Gibbs Sampler algorithm was implemented using steps of 10 iterations for 1 replication, with convergence being assessed via stability of posterior modal positions. Each set of results presented is unsmoothed, apart from the removal of points on the frame of the region of interest.

The results are generally encouraging. Perhaps most surprisingly, the results derived under one changepoint models are adequate, with the eastern and southern coastlines being detected accurately. Note also how the approximate methods produce results equivalent to the

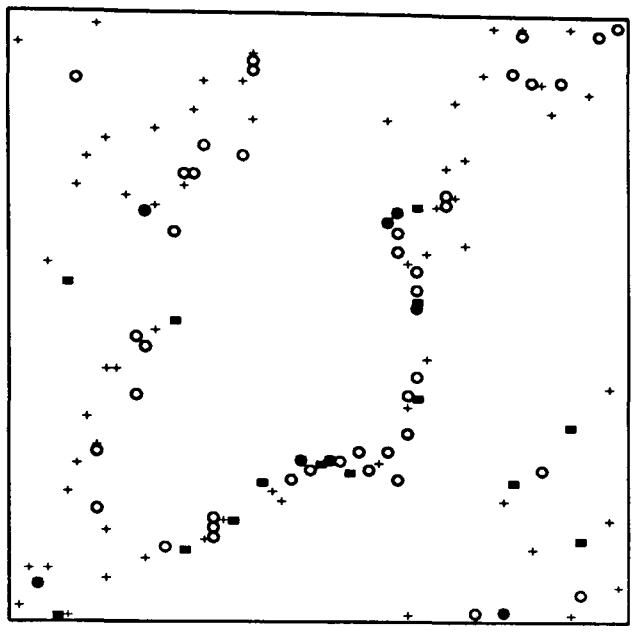


Fig 135(a) : one changepoint - exact

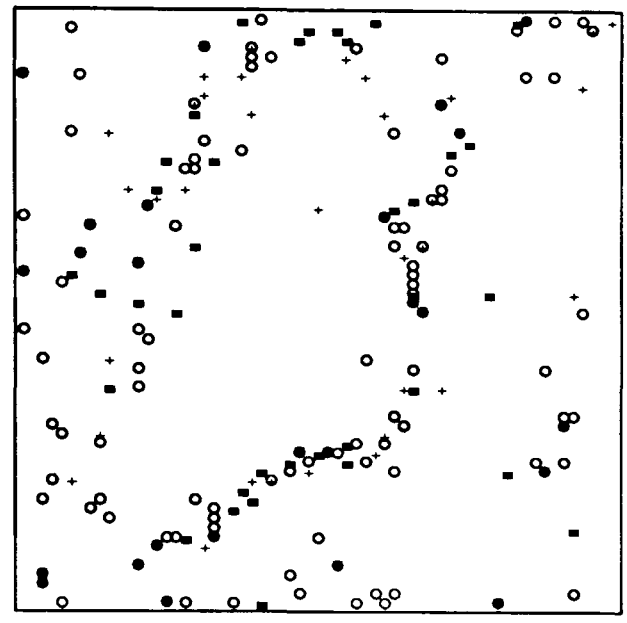


Fig 135(b) : binary segmentation

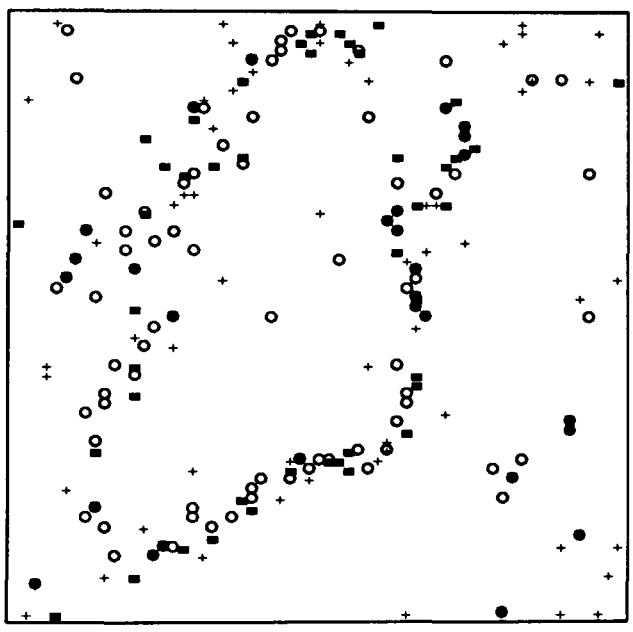


Fig 135(c) : two changepoint - exact

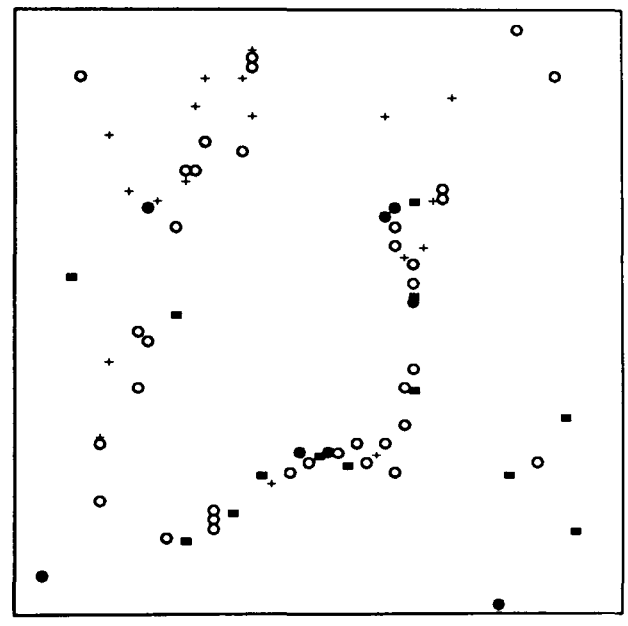


Fig 135(d) : one changepoint - approx.

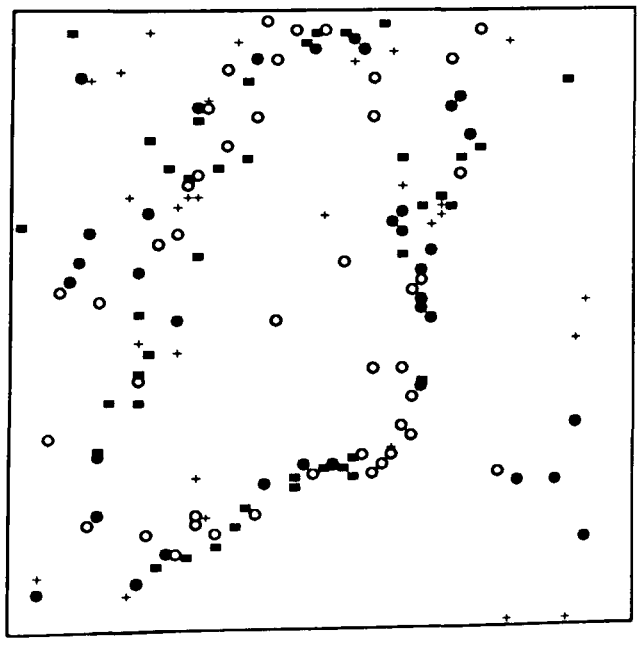


Fig 135(e) : two changepoint - approx.

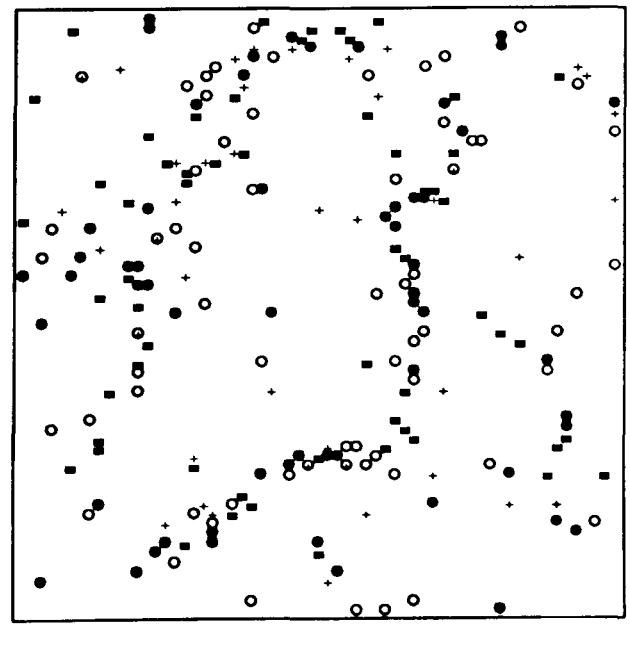


Fig 135(f) : three changepoint - approx.

corresponding exact methods. We note that simple smoothing of the results may visually improve their quality. The timings involved in the production of the results in figure 135(a) to (f) were 1.14, 1.90, 51.10, 9.88, 16.88, and 22.62 seconds respectively for the processing of the 128 rows and columns in the image. Thus, for the multiple changepoint models, use of the Gibbs Sampler approximate methods may lead to significant time savings without significant loss of accuracy.

We now proceed with an attempt to segment the noise-corrupted image. First, we obtain a segmentation automatically using the naive classification technique developed for convex object true scenes described earlier in this chapter. Using the results of the binary segmentation analysis depicted in figure 135(a), after suitable smoothing, each pixel is classified to texture T_1 if the row and column classifications agree that the pixel was "internal" to the object, and to texture T_0 otherwise. Figure 136(a) depicts the segmentation obtained by such a procedure.

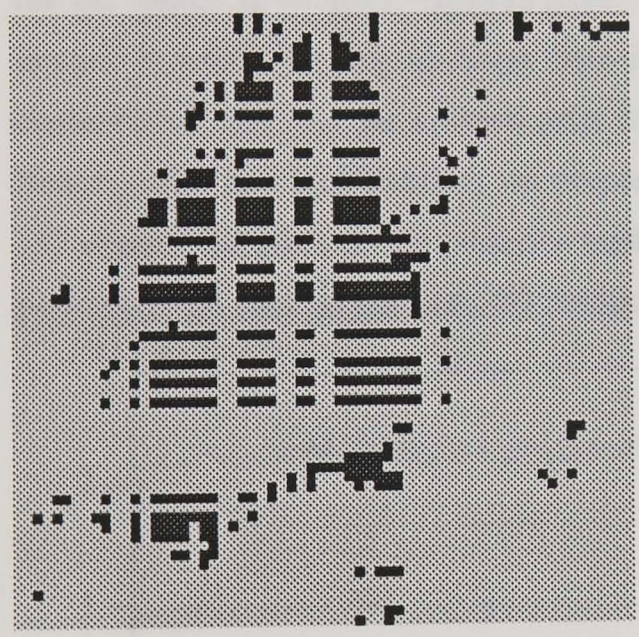


Fig 136(a) : naive classification

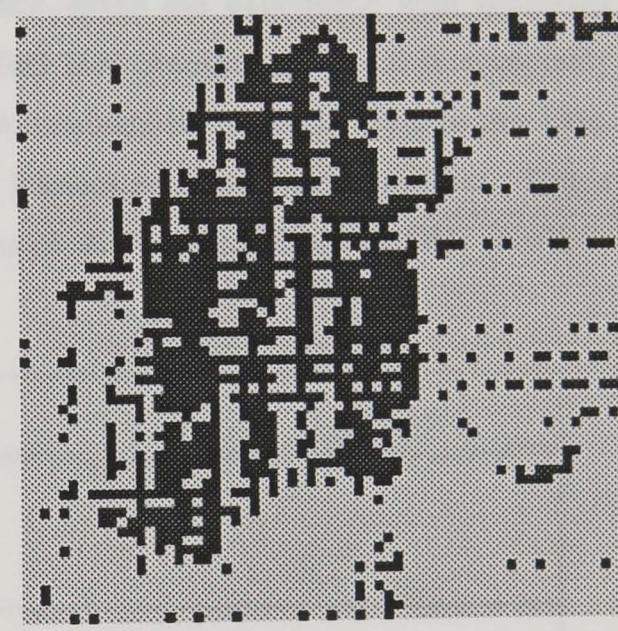


Fig 136(b) : probabilistic classification

From the segmentation in figure 136(a), we obtain naive estimates of the texture mean parameters and the noise variances. In this example, the estimates are 0.2413 and 0.8077 for the texture means, and 0.5860 and 0.6159 for the noise terms which, at this stage, we do not assume are identically distributed between textures. In total, 2231 pixels were naively allocated to texture T_0, and 439 were allocated to texture T_1. Figure 136(b) depicts the segmentation obtained when simple probabilistic classification is used, incorporating the prior knowledge concerning true classification for each pixel obtained during the edge-detection analysis, that is, we set the prior probability that any pixel belongs to texture T_0 or T_1 (nominally) to be 0.875 if the row and column classifications agree, and to 0.5 otherwise. Clearly, despite the sub-optimal choice of the texture parameters in the classification procedure, the segmentation obtained is generally quite adequate for use as an initial realisation in

subsequent more sophisticated analyses.

Finally for this example, we attempt to obtain an M.P.M. segmentation via the amended Gibbs Sampler algorithm described above. As usual, we must first specify the parameters in the prior distributions for the texture means and (now presumed common) noise precision, and the parameters in the Gibbs prior for the true scene. We may specify these parameters automatically in the following way. We may choose the prior means for the texture mean-level parameters to be equal to those obtained using the naive technique above, 0.2413 and 0.8077, and the prior precision for these parameters to be reasonably small, say, 0.001 in each case. For the noise precision, we could specify Gamma distribution with mode around $1.0/0.6 = 1.6667$, but here we revert to the $Ga(2.0, 1.0)$ specification used previously. Thus our prior specification for these parameters is again clearly sub-optimal, although quite adequate. We again choose a one parameter Gibbs prior with interaction parameter β. Now, given the results of our naive classification procedure, we have reason to believe that the effective Signal-Noise for the image is $(0.8077 - 0.2413)/0.6 \approx 1.0$. Hence, given our experience with simulated images in which the Signal-Noise ratio is 1.0, we feel that choosing $\beta = 1.0$ is acceptably conservative, and that the induced prior field will not dominate the posterior field in this case. Hence, using this specification, we proceed with an implementation of the algorithm. Figure 137 depicts the plots of the posterior modal estimates of the texture mean and noise precision parameters on each of the first 200 iterations.

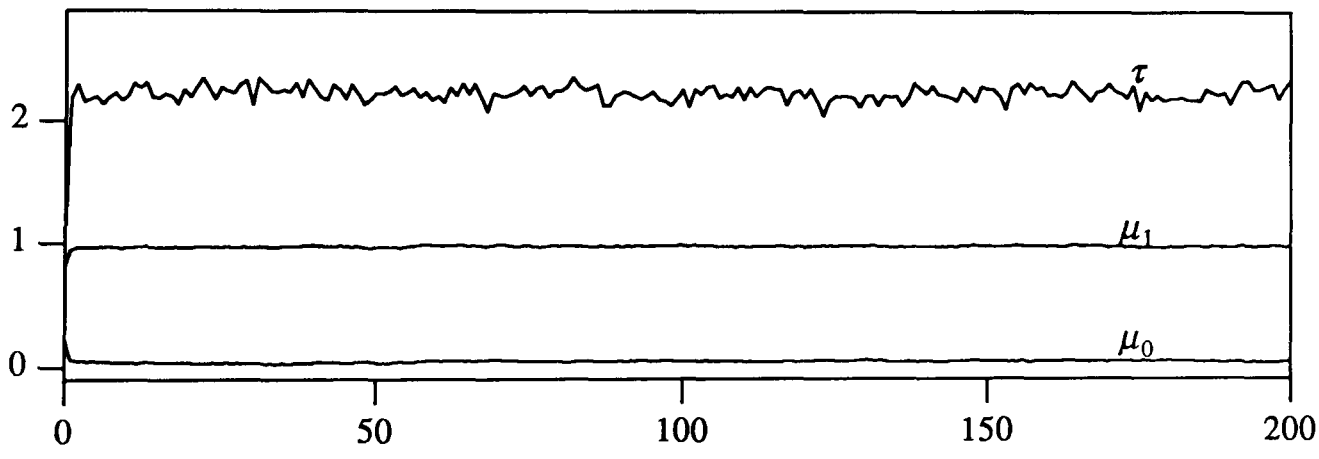


Fig 137(a) : texture parameter posterior estimates

The estimates generally seem well behaved, and on the basis of these plots we might infer that the algorithm has converged almost immediately. Figure 138 depicts the plots of the number of pixels allocated to the respective textures.

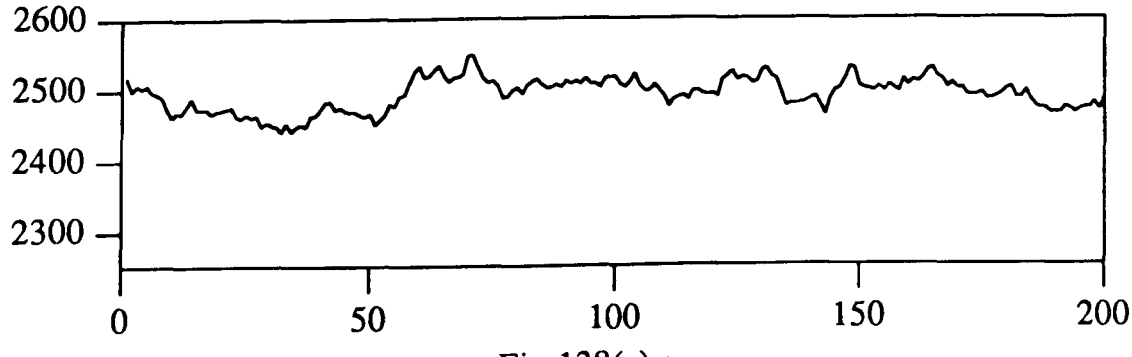


Fig 138(a) : n_0

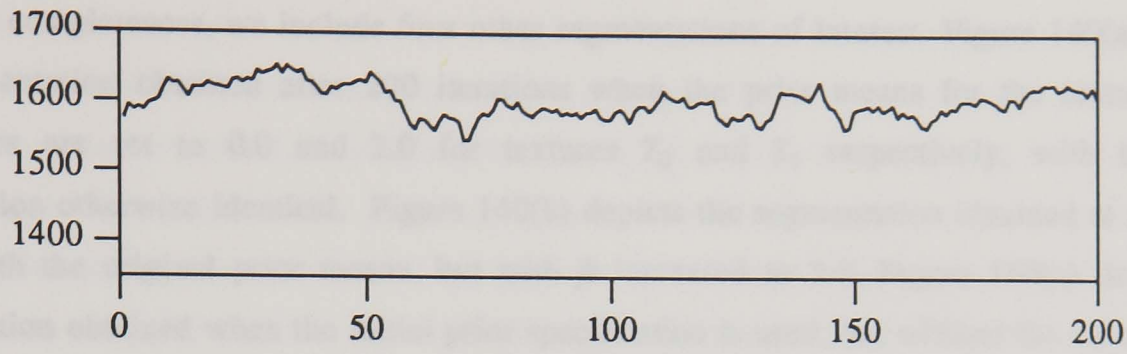


Fig 138(b) : n_1

Again, these plots appear to be relatively stable, after an abrupt change n_0 and n_1 at around the 50'th iteration, and so we are willing to accept that the algorithm has converged prior to, say, the 75'th iteration. Figure 139(a) and (b) depict the segmentations obtained after 75 and 200 iterations respectively.

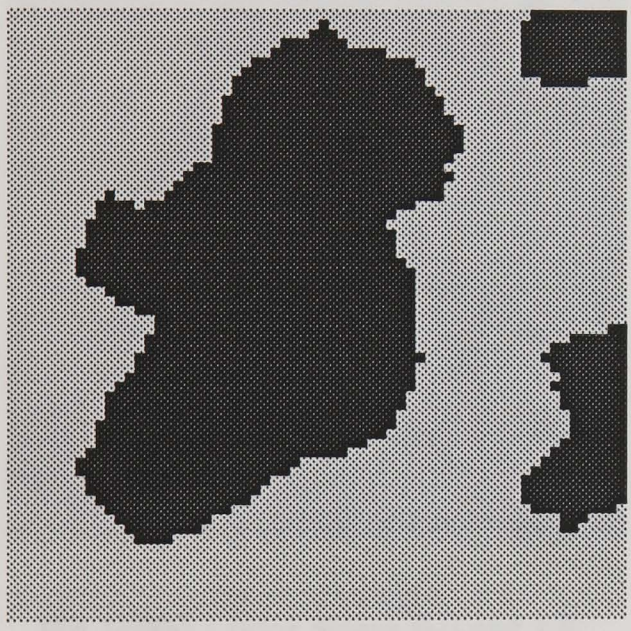


Fig 139(a) : 75 iterations

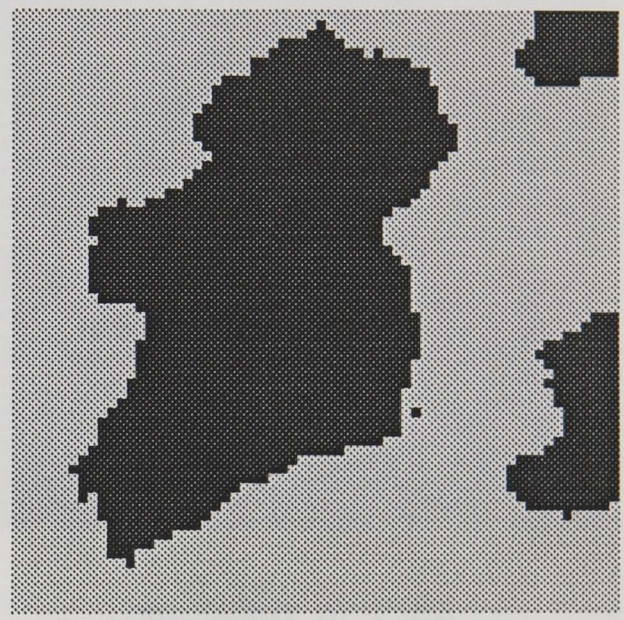


Fig 139(b) : 200 iterations

The two segmentations are broadly similar, but differ slightly, especially in the vicinity of the north-west coastline - with knowledge of the true scene we would regard the second segmentation as visually more satisfactory, but the percentage pixel difference is actually quite low (actually 1.688%). This confirms the segmentation problem as being very difficult - the human eye is able to perceive such small differences between images. Note that all small scale features in the true scene (the detail on the western coast, lakes, islands etc.) are not present in the segmented images. This is the penalty for making simple and global assumptions concerning the local nature of the true scene, as expressed via the Gibbs prior. It is generally suggested that this problem may be partly overcome by using more sophisticated priors (although, in practice, this may prove ineffective without sufficient prior knowledge of the true scene), or by discretising the true scene via a pixel grid of higher resolution than that used to discretise the image (which naturally increases the amount of computation needed).

For completeness, we include four other segmentations of interest. Figure 140(a) depicts the segmentation obtained after 200 iterations when the prior means for the texture mean parameters are set to 0.0 and 2.0 for textures T_0 and T_1 respectively, with the prior specification otherwise identical. Figure 140(b) depicts the segmentation obtained at the same stage, with the original prior means, but with β increased to 2.0. Figure 140(c) depicts the segmentation obtained when the initial prior specification is used, but without the values of the texture parameters being updated in the fashion prescribed by our amended version of the Gibbs Sampler algorithm. Figure 140(c) depicts the segmentation obtained by using Besag's I.C.M. classification technique, with the prior means specified as 0.2413 and 0.8077, and β equal to 1.0, after only 5 iterations.

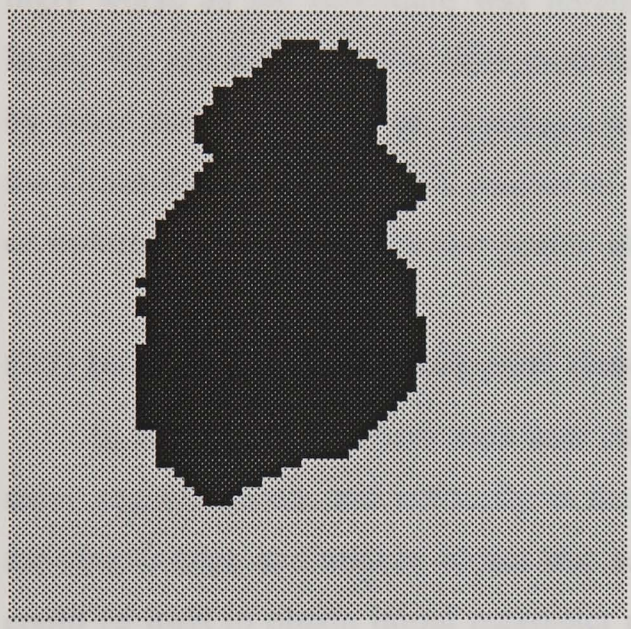


Fig 140(a) : prior means 0.0 and 2.0

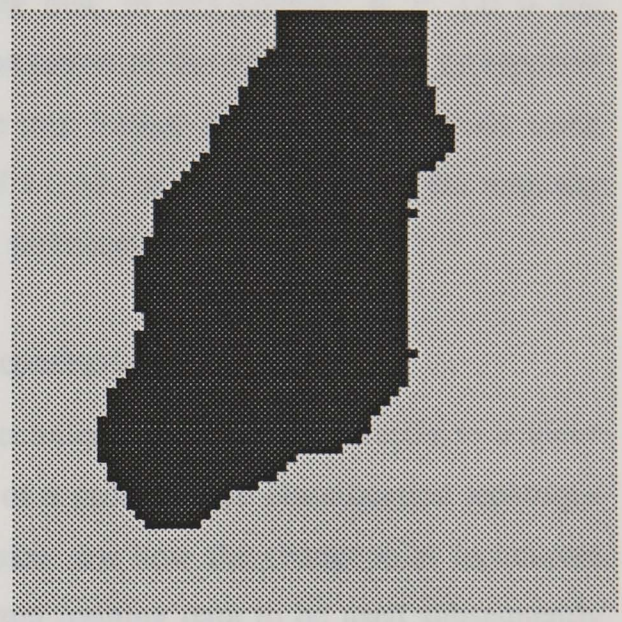


Fig 140(b) : β = 2.0

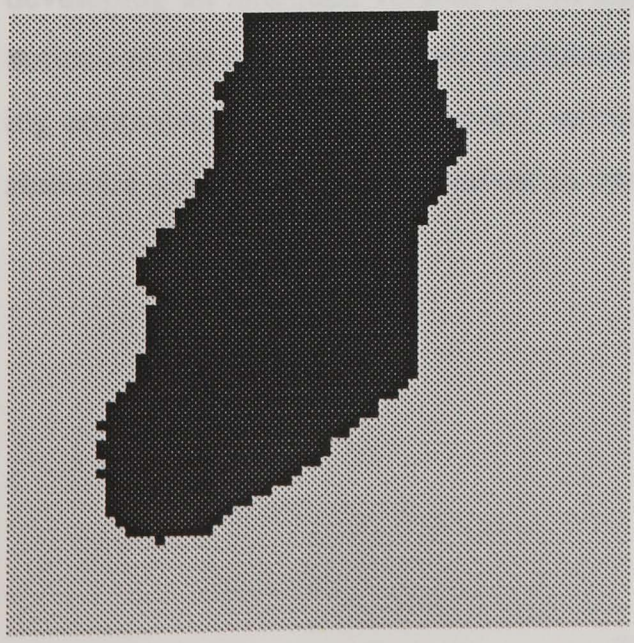


Fig 140(c) : non-adaptive

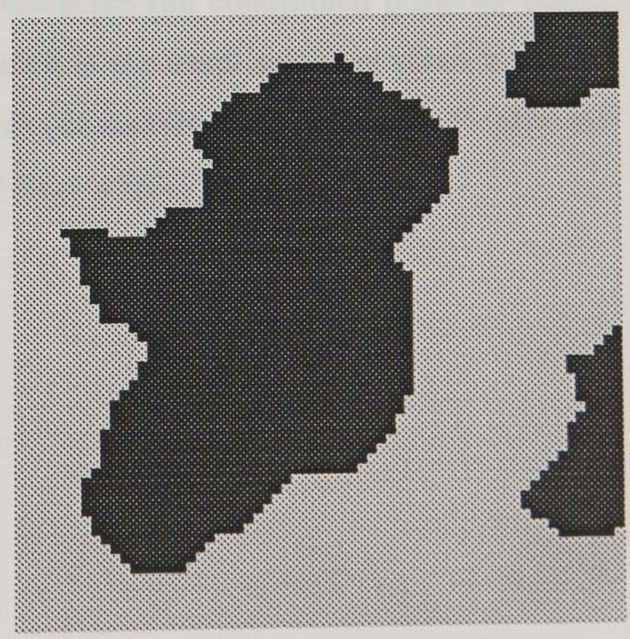


Fig 140(d) : I.C.M., β =1.0

Again, the reconstructions are broadly similar, but we notice several important features. The segmentation in (a) is clearly inadequate. This is largely due to the poor prior specification of the texture mean parameters. In reality, the algorithm had not converged at this stage, and did not converge for around another 200 iterations, after which the segmentation had improved a little relative to the true scene, even though the "Wales" and "Scotland" regions were still absent. Thus we infer that the prior specification can effect the rate of convergence of the algorithm adversely. The segmentations in (b) and (c) are also inadequate. In (b), the choice of β has led to the prior field dominating the posterior field, and in (c) the lack of updating of the parameters has had an appreciable negative effect. The segmentation achieved using I.C.M. depicted in (d) is remarkably good. However, the quality of this segmentation became increasingly worse as the procedure was iterated, indicating that, in this context, $\beta = 1.0$ is too large. We are also left with little probabilistic interpretation of the results. Thus, as reported by other authors, it may be advantageous to precede any formal probabilistic analysis using the Gibbs Sampler by a small number of iterations of I.C.M., in order to remove the majority of isolated pixels created by non-spatial classification techniques.

(7.7) Image segmentation and pixel classification - conclusions.

In this chapter we have seen how changepoint techniques can be used directly to achieve naive segmentations, and indirectly as the first stage of a probabilistic classification technique for simple edge and convex object true scenes. We also saw how the changepoint techniques could be used to obtain initial parameter estimates from training data to be used subsequently in more sophisticated segmentation schemes. We attempted to overcome the problem of coherent parameter estimation in one such scheme implemented via the Gibbs Sampler by developing an amended version of the algorithm in which the full conditional posterior distributions of the unknown texture parameters were included in the usual sampling scheme, allowing approximate posterior marginal estimates for these parameters to be obtained. Finally, we studied a worked example based on an image derived from a familiar true scene.

Appendix 1 : Posterior forms.

1. Normal.

1.1. Likelihood (A) - common variance , $\theta = (\theta_1, \theta_2, \tau)$

1.1.1 Prior 1 : $[\theta_1, \theta_2, \tau] = \text{constant.}$

All parameters known.

$$[r \mid Y, \psi] \propto \exp\left\{ -\frac{\tau}{2}\left[\sum_{i=1}^{r}(Y_i - \theta_1)^2 + \sum_{i=r+1}^{n}(Y_i - \theta_2)^2\right]\right\}$$

1.1.2 Prior 2 : $[\theta_1, \theta_2, \tau] = [\theta_1][\theta_2]$

τ known , independent Normal priors for θ_1, θ_2

$[\theta_1] = N(\mu_1, \eta_1^{-1})$

$[\theta_2] = N(\mu_2, \eta_2^{-1})$

$\psi = (\mu_1, \mu_2, \eta_1, \eta_2)$

$$[r \mid Y, \psi] \propto t_{1r}^{-1/2}\exp\left\{ -\frac{\tau}{2}[SSQ_r + t_{2r} + t_{3r}]\right\}$$

$$t_{1r} = (r\tau + \eta_1)((n-r)\tau + \eta_2)$$

$$t_{2r} = \frac{r\eta_1}{(r\tau + \eta_1)}(\bar{Y}_A - \mu_1)^2$$

$$t_{3r} = \frac{(n-r)\eta_2}{((n-r)\tau + \eta_2)}(\bar{Y}_B - \mu_2)^2$$

$$SSQ_r = \sum_{i=1}^{r}(Y_i - \bar{Y}_A)^2 + \sum_{i=r+1}^{n}(Y_i - \bar{Y}_B)^2$$

$$\bar{Y}_A = \frac{1}{r}\sum_{i=1}^{r} Y_i$$

$$\bar{Y}_B = \frac{1}{(n-r)}\sum_{i=r+1}^{n} Y_i$$

(Non-informative limit : $\eta_1, \eta_2 \to 0$)

1.1.3 Prior 3 : $[\theta_1, \theta_2, \tau] = [\theta_1][\theta_2]$

τ known , Uniform priors for θ_1, θ_2

$[\theta_1] = U(\delta_1, \varepsilon_1)$

$[\theta_2] = U(\delta_2, \varepsilon_2)$

$\psi = (\delta_1, \varepsilon_1, \delta_2, \varepsilon_2)$

$$[r \mid Y, \psi] \propto t_r^{-1/2} \exp\left\{ -\frac{\tau}{2} SSQ_r \right\} [\Phi_1(\varepsilon_1) - \Phi_1(\delta_1)][\Phi_2(\varepsilon_2) - \Phi_2(\delta_2)]$$

$t_r = r(n-r)$

$\Phi_1(.) = \Phi\left((r\tau)^{1/2}(. - \bar{Y}_A)\right)$

$\Phi_2(.) = \Phi\left(((n-r)\tau)^{1/2}(. - \bar{Y}_B)\right)$

$\Phi(.)$ – Unit Normal c.d.f

(Non-informative limit : $\varepsilon_1, \varepsilon_2 \to \infty, \delta_1, \delta_2 \to -\infty$)

1.1.4 Prior 4 : $[\theta_1, \theta_2, \tau] = [\theta_2 \mid \theta_1]$

θ_1, τ known , dependent Normal prior for θ_2 conditional on θ_1

(W.l.o.g assume $\theta_1 = 0$)

$[\theta_2 \mid \theta_1] = N(\theta_1, \eta_{12}^{-1})$

$\psi = (\eta_{12})$

$$[r \mid Y, \psi] \propto u_{1r}^{-1/2} \exp\left\{ -\frac{\tau}{2} [SSQ_{nr} + u_{2r}] \right\}$$

$u_{1r} = ((n-r)\tau + \eta_{12})$

$u_{2r} = \dfrac{(n-r)\eta_{12}}{u_{1r}} \bar{Y}_B^2$

$SSQ_{nr} = \sum_{i=1}^{r} Y_i^2 + \sum_{i=r+1}^{n} (Y_i - \bar{Y}_B)^2$

(Non-informative limit : $\eta_{12} \to 0$)

1.1.5 Prior 5 : $[\theta_1, \theta_2, \tau] = [\theta_2 \mid \theta_1][\theta_1]$

τ known , dependent Normal priors for (θ_1, θ_2)

$[\theta_2 \mid \theta_1] = N(\theta_1, \eta_{12}^{-1})$

$[\theta_1] = N(\mu_1, \eta_1^{-1})$

$\psi = (\mu_1, \eta_1, \eta_{12})$

$$[r \mid Y, \psi] \propto v_{1r}^{-1/2} \exp\left\{ -\frac{\tau}{2}[SSQ_r + v_{2r}] \right\} \exp\left\{ -\frac{v_{3r}}{2} \right\}$$

$$v_{1r} = (r\tau + K + \eta_1)((n-r)\tau + \eta_{12})$$

$$v_{2r} = \frac{rK}{r\tau + K}(\bar{Y}_A - \bar{Y}_B)^2$$

$$v_{3r} = \frac{(r\tau + K)\eta_1}{r\tau + K + \eta_1}(\mu_1 - \bar{Y}_C)^2$$

$$\bar{Y}_C = \frac{r\tau\bar{Y}_A + K\bar{Y}_B}{r\tau + K}$$

$$K = \frac{(n-r)\tau\eta_{12}}{(n-r)\tau + \eta_{12}}$$

(Non-informative limit : $\eta_1, \eta_{12} \to 0$)

1.1.6 Prior 6 : $[\theta_1, \theta_2, \tau] = [\theta_2 \mid \theta_1, \tau]$

θ_1, τ known , dependent Normal prior for θ_2 conditional on θ_1 and τ

(W.l.o.g assume $\theta_1 = 0$)

$[\theta_2 \mid \theta_1, \tau] = N(\theta_1, (\gamma\tau)^{-1})$

$\psi = (\gamma)$

$$[r \mid Y, \psi] \propto w_{1r}^{-1/2} \left\{ -\frac{\tau}{2}[SSQ_{nr} + w_{2r}] \right\}$$

$$w_{1r} = (n-r) + \gamma$$

$$w_{2r} = \frac{(n-r)\gamma}{(n-r) + \gamma}\bar{Y}_B^2$$

(Non-informative limit : $\gamma \to 0$)

1.1.7 Prior 7 : $[\theta_1, \theta_2, \tau] = [\theta_2 \mid \theta_1, \tau][\theta_1]$

τ known , dependent Normal priors for (θ_1, θ_2) conditional on τ

$[\theta_2 \mid \theta_1, \tau] = N(\theta_1, (\gamma\tau)^{-1})$

$[\theta_1] = N(\mu_1, \eta_1^{-1})$

$\psi = (\mu_1, \gamma, \eta_1)$

$$[r \mid Y, \psi] \propto a_{1r}^{-1/2}\left\{ -\frac{\tau}{2}[SSQ_r + a_{2r}]\right\}\left\{ -\frac{a_{3r}}{2}\right\}$$

$$a_{1r} = (r\tau + K' + \eta_1)((n-r) + \gamma)$$

$$a_{2r} = \frac{rK'}{r + K'}(\bar{Y}_A - \bar{Y}_B)^2$$

$$a_{3r} = \frac{r\tau + K'}{r\tau + K' + \eta_1}(\mu_1 - \bar{Y}_D)^2$$

$$\bar{Y}_D = \frac{r\tau\bar{Y}_A + K'\bar{Y}_B}{r\tau + K'}$$

$$K' = \frac{(n-r)\tau\gamma}{w_{1r}}$$

(Non-informative limit : $\eta_1, \gamma \to 0$)

1.1.8 Prior 8 : $[\theta_1, \theta_2, \tau] = [\theta_1 \mid \tau][\theta_2 \mid \tau][\tau]$

Independent Normal priors for (θ_1, θ_2) conditional on τ

Gamma prior for τ

$[\theta_1 \mid \tau] = N(\mu_1, \gamma_1\tau^{-1})$

$[\theta_2 \mid \tau] = N(\mu_2, \gamma_2\tau^{-1})$

$[\tau] = Ga(\alpha/2, \beta/2)$

$\psi = (\mu_1, \mu_2, \gamma_1, \gamma_2, \alpha, \beta)$

$$[r \mid Y, \psi] \propto b_{1r}^{-1/2}\{SSQ_r + b_{2r} + b_{3r} + \beta\}^{-\left(\frac{n+\alpha}{2}\right)}$$

$$b_{1r} = (r + \gamma_1)((n-r) + \gamma_2)$$

$$b_{2r} = \frac{r\gamma_1}{r + \gamma_1}(\mu_1 - \bar{Y}_A)^2$$

$$b_{3r} = \frac{(n-r)\gamma_2}{(n-r) + \gamma_2}(\mu_2 - \bar{Y}_B)^2$$

(Non-informative limit : $\gamma_1, \gamma_2, \alpha, \beta \to 0$)

1.2. Likelihood (B) - common mean , $\theta = (\theta, \tau_1, \tau_2)$

1.2.1 Prior 1 : $[\theta, \tau_1, \tau_2] =$ constant.

All parameters known.

$$[r \mid Y, \psi] \propto \tau_1^{\frac{r}{2}} \tau_2^{\frac{(n-r)}{2}} \exp\left\{ -\frac{\tau_1}{2} \sum_{i=1}^{r} (Y_i - \theta)^2 - \frac{\tau_2}{2} \sum_{i=r+1}^{n} (Y_i - \theta)^2 \right\}$$

1.2.2 Prior 2 : $[\theta, \tau_1, \tau_2] = [\theta]$

τ_1, τ_2 known Normal priors for θ

$[\theta] = N(\mu_1, \eta_1^{-1})$

$\psi = (\mu_1, \eta_1)$

$$[r \mid Y, \psi] \propto c_{1r}^{-1/2} \tau_1^{\frac{r}{2}} \tau_2^{\frac{(n-r)}{2}} \exp\left\{ -\frac{1}{2} [\tau_1 SSQ_{1r} + \tau_2 SSQ_{2r} + c_{2r} + c_{3r}] \right\}$$

$$c_{1r} = r\tau_1 + (n-r)\tau_2 + \eta_1$$

$$c_{2r} = \frac{r(n-r)\tau_1\tau_2}{r\tau_1 + (n-r)\tau_2} (\bar{Y}_A - \bar{Y}_B)^2$$

$$c_{3r} = \frac{r(n-r)\tau_1\tau_2\eta_1}{c_{1r}} (\mu - \bar{Y}_E)^2$$

$$SSQ_{1r} = \sum_{i=1}^{r} (Y_i - \bar{Y}_A)^2$$

$$SSQ_{2r} = \sum_{i=r+1}^{n} (Y_i - \bar{Y}_B)^2$$

$$\bar{Y}_E = \frac{r\tau_1\bar{Y}_A + (n-r)\tau_2\bar{Y}_B}{r\tau_1 + (n-r)\tau_2}$$

(Non-informative limit : $\eta_1 \to 0$)

1.3. Likelihood (C) - different means , variances $\theta = (\theta_1, \theta_2, \tau_1, \tau_2)$

1.3.1 Prior 1 : $[\theta_1, \theta_2, \tau_1, \tau_2] =$ constant.

All parameters known.

$$[r \mid Y, \psi] \propto \tau_1^{\frac{r}{2}} \tau_2^{\frac{(n-r)}{2}} \exp\left\{ -\tau_1 \sum_{i=1}^{r} (Y_i - \theta_1)^2 - \tau_2 \sum_{i=r+1}^{n} (Y_i - \theta_2)^2 \right\}$$

1.3.2 Prior 2 : $[\theta_1, \theta_2, \tau_1, \tau_2] = [\theta_1][\theta_2]$

τ_1, τ_2 known , independent Normal priors for θ_1, θ_2

$$[\theta_1] = N(\mu_1, \eta_1^{-1})$$

$$[\theta_2] = N(\mu_2, \eta_2^{-1})$$

$$\psi = (\mu_1, \mu_2, \eta_1, \eta_2)$$

$$[r \mid Y, \psi] \propto d_{1r}^{-1/2} \tau_1^{\frac{r}{2}} \tau_2^{\frac{(n-r)}{2}} \exp\left\{ -\frac{1}{2}[\tau_1 SSQ_{1r} + \tau_2 SSQ_{2r} + d_{2r} + d_{3r}] \right\}$$

$$d_{1r} = (r\tau_1 + \eta_1)((n-r)\tau_2 + \eta_2)$$

$$d_{2r} = \frac{r\tau_1\eta_1}{r\tau_1 + \eta_1}(\mu_1 - \bar{Y}_A)^2$$

$$d_{3r} = \frac{(n-r)\tau_2\eta_2}{(n-r)\tau_2 + \eta_2}(\mu_2 - \bar{Y}_B)^2$$

(Non-informative limit : $\eta_1, \eta_2 \to 0$)

1.3.3 Prior 3 : $[\theta_1, \theta_2, \tau_1, \tau_2] = [\theta_1 \mid \tau_1][\theta_2 \mid \tau_2]$

τ_1, τ_2 known .

Independent Normal priors for θ_1, θ_2 conditional on τ_1, τ_2

$$[\theta_1 \mid \tau_1] = N(\mu_1, (\gamma_1\tau_1)^{-1})$$

$$[\theta_2 \mid \tau_2] = N(\mu_2, (\gamma_2\tau_2)^{-1})$$

$$\psi = (\mu_1, \mu_2, \gamma_1, \gamma_2)$$

$$[r \mid Y, \psi] \propto e_{1r}^{-1/2} \tau_1^{\frac{(r-1)}{2}} \tau_2^{\frac{(n-r-1)}{2}} \exp\left\{ -\frac{1}{2}[\tau_1 SSQ_{1r} + \tau_2 SSQ_{2r} + e_{2r} + e_{3r}] \right\}$$

$$e_{1r} = (r + \gamma_1)((n-r) + \gamma_2)$$

$$e_{2r} = \frac{r\tau_1\gamma_1}{r + \gamma_1}(\mu_1 - \bar{Y}_A)^2$$

$$e_{3r} = \frac{(n-r)\tau_2\gamma_2}{(n-r) + \gamma_2}(\mu_2 - \bar{Y}_B)^2$$

(Non-informative limit : $\gamma_1, \gamma_2 \to 0$)

1.3.4 Prior 4 : $[\theta_1, \theta_2, \tau_1, \tau_2] = [\theta_1 \mid \tau_1][\theta_2 \mid \tau_2]$

All parameters unknown.

Independent Normal priors for θ_1, θ_2 conditional on τ_1, τ_2

Independent Gamma priors for τ_1, τ_2

$$[\theta_1 \mid \tau_1] = N(\mu_1, (\gamma_1 \tau_1)^{-1})$$

$$[\theta_2 \mid \tau_2] = N(\mu_2, (\gamma_2 \tau_2)^{-1})$$

$$[\tau_1] = Ga(\alpha_1/2, \beta_1/2)$$

$$[\tau_2] = Ga(\alpha_2/2, \beta_2/2)$$

$$\psi = (\mu_1, \mu_2, \gamma_1, \gamma_2, \alpha_1, \beta_1, \alpha_2, \beta_2)$$

$$[r \mid Y, \psi] \propto e_{1r}^{-1/2} \Gamma\left(\frac{r + \alpha_1 - 1}{2}\right) \Gamma\left(\frac{n - r + \alpha_2 - 1}{2}\right)$$

$$\{SSQ_{1r} + e_{2r} + \beta_1\}^{\frac{(r+\alpha_1-1)}{2}} \{SSQ_{2r} + e_{3r} + \beta_2\}^{\frac{(n-r+\alpha_2-1)}{2}}$$

(Non-informative limit : $\gamma_1, \gamma_2, \alpha_1, \beta_1, \alpha_2, \beta_2 \to 0$)

2. Poisson.

2.1.1 Prior 1 : $[\lambda_1, \lambda_2] =$ constant.

All parameters known.

$$[r \mid Y, \psi] \propto \lambda_1^{\sum_{i=1}^{r} Y_i} \lambda_2^{\sum_{i=r+1}^{n} Y_i} \exp\{-r\lambda_1 - (n-r)\lambda_2\}$$

2.1.2 Prior 2 : $[\lambda_1, \lambda_2] = [\lambda_1][\lambda_2]$

All parameters unknown.

Independent Gamma priors for λ_1, λ_2

$$[\lambda_1] = Ga(\alpha_1, \beta_1)$$

$$[\lambda_2] = Ga(\alpha_2, \beta_2)$$

$$\psi = (\alpha_1, \beta_1, \alpha_2, \beta_2)$$

$$[r \mid Y, \psi] \propto \frac{\Gamma\left(\alpha_1 + \sum_{i=1}^{r} Y_i + \frac{1}{2}\right) \Gamma\left(\alpha_2 + \sum_{i=r+1}^{n} Y_i + \frac{1}{2}\right)}{\{\beta_1 + r\}^{\alpha_1 + \sum_{i=1}^{r} Y_i + \frac{1}{2}} \{\beta_2 + n - r\}^{\alpha_2 + \sum_{i=r+1}^{n} Y_i + \frac{1}{2}}}$$

(Non-informative limit : $\alpha_1, \alpha_2, \beta_1, \beta_2 \to 0$)

Appendix 2 : Edge-detection - examples.

In this appendix, we present further examples of the use of the changepoint based edge-detection routines developed in previous chapters. In each case, part (a) depicts the true scene, and part (b) the image derived from the true scene under the image formation process in equation (2.1), where each pixel is corrupted independently and identically with additive zero-mean and normally distributed noise. For demonstration purposes, the images were generated to exhibit a Signal-Noise ratio of 1.5, and are displayed using a six level grey scale. In figure A2-1 through to A2-5, the results depicted in parts (c) to (f) arise as the result of the following analyses. Part (c) depicts the results of a full analysis with each row and column treated independently using the posterior distribution in (2.11) under a one changepoint assumption. Part (d) depicts the results using the binary segmentation technique described in section (3.3.2) of chapter 3, under similar assumptions. Part (e) depicts the results of a full analysis using the marginalised version of the joint posterior distribution in equation (3.2) under a two changepoint assumption. Part (f) depicts the results of a full analysis based on the marginal posterior approximation technique using the Gibbs Sampler algorithm discussed in section (3.2), with both number of iterations t_0 and number of replications m set equal to one. In each case, a non-informative specification for the texture mean-level and noise variance parameters was used. In addition, a probability of $p = 0.1$ was placed on the no changepoint alternative model. The amount of computation time required to produce each set of results is recorded in each case. In figure A2-6, part (c) depicts the results of an full analysis using (2.11), and parts (d), (e), and (f) depict the results obtained using the Gibbs Sampler approximation technique to compute the posterior marginal distributions under two, three, and four changepoint assumptions. Again, t_0 and m were set equal to 1 in each case. In figure A2-7, part (a) depicts a relatively simple three texture true scene, with the homogeneous textures T_0, T_1, and T_2 having mean-levels $0.0, 1.0,$ and 3.0 respectively, so that the Signal-Noise ratio at the T_0 / T_1 boundary is half that at T_1 / T_2 boundary. Parts (c) to (f) depict respectively the results of row analyses using (2.11), the binary segmentation technique, the marginalised version of (3.2), and the Gibbs Sampler based approximation to the posterior marginal distributions with t_0 and m equal to 1. Figure A2-8(a) depicts a chessboard type true scene with texture mean-levels 0.0 and 1.5, and part (b) depicts an image derived when the noise variance is 1.0. Again, parts (c) and (d) depict full analysis results obtained using (2.11) and the binary segmentation technique. Parts (e) and (f) depict the results of Gibbs Sampler based analyses under a three changepoint with (t_0, m) set equal to $(1, 1)$ and $(5, 3)$ respectively.

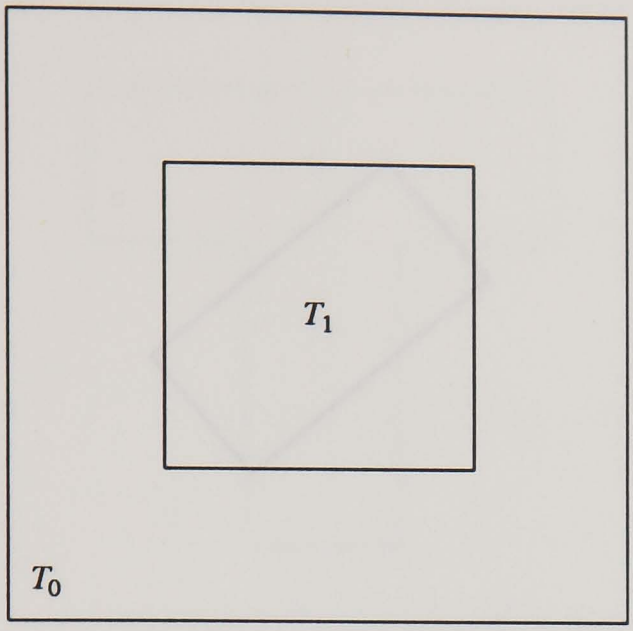


Fig A2-1(a) : true scene

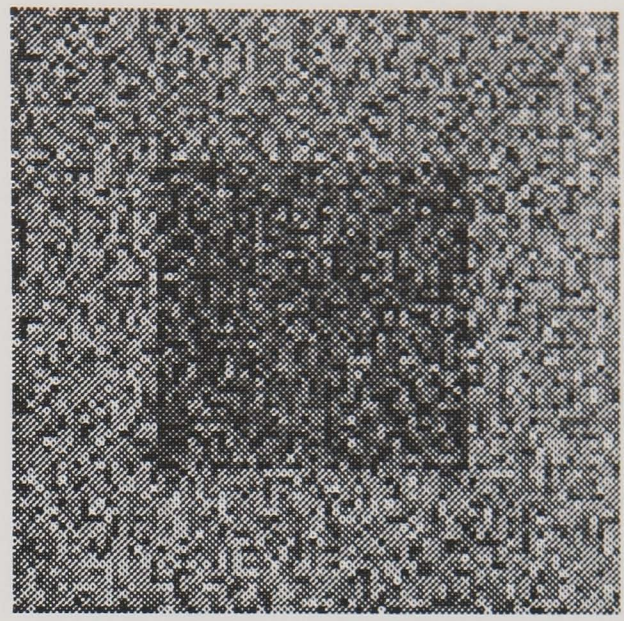


Fig A2-1(b) : image

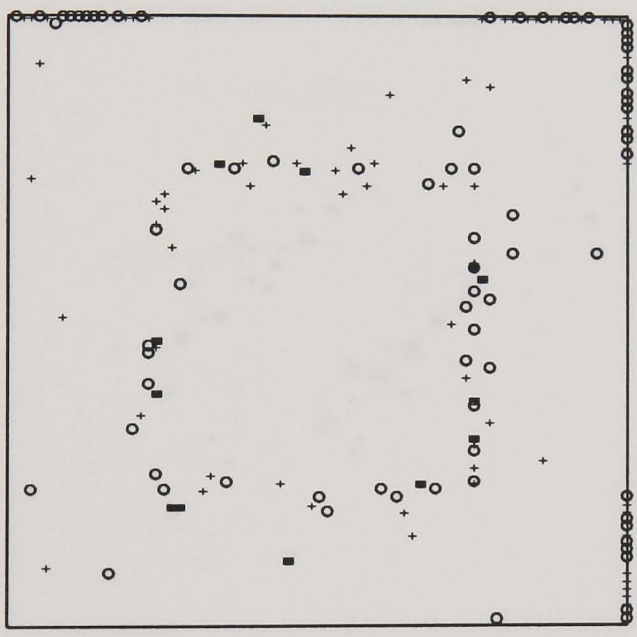


Fig A2-1(c) : 1.8 seconds

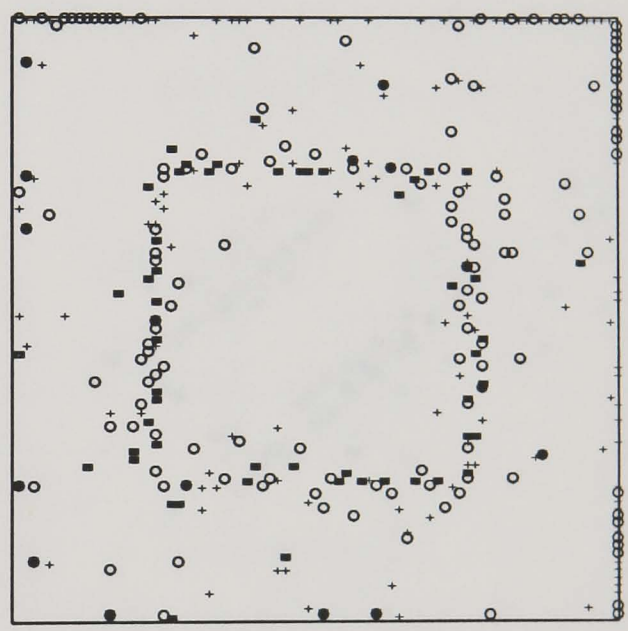


Fig A2-1(d) : 3.0 seconds

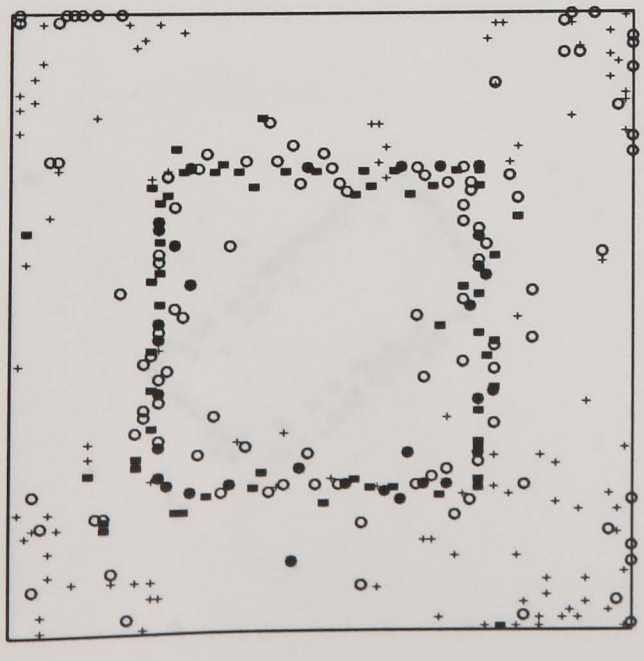


Fig A2-1(e) : 112.8 seconds

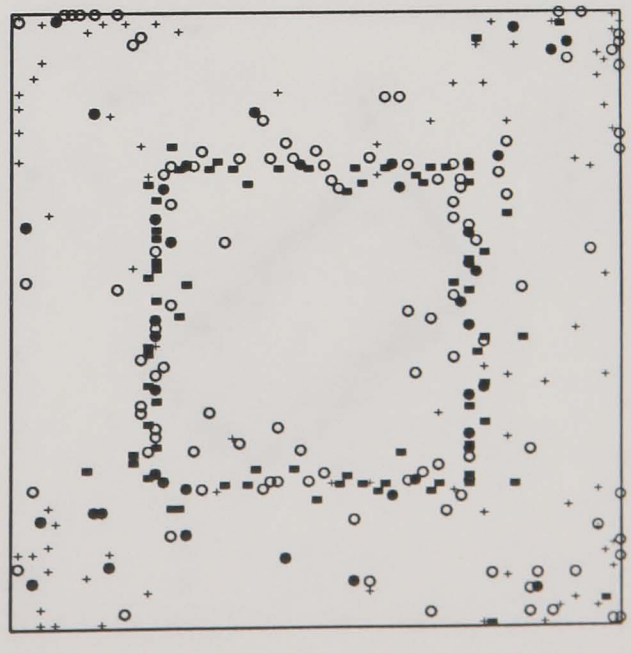


Fig A2-1(f) : 23.4 seconds

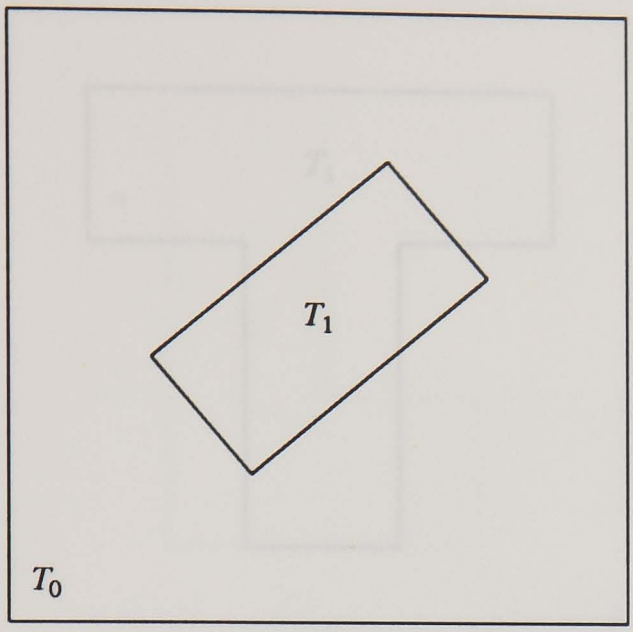


Fig A2-2(a) : true scene

Fig A2-2(b) : image

Fig A2-2(c) : 1.9 seconds

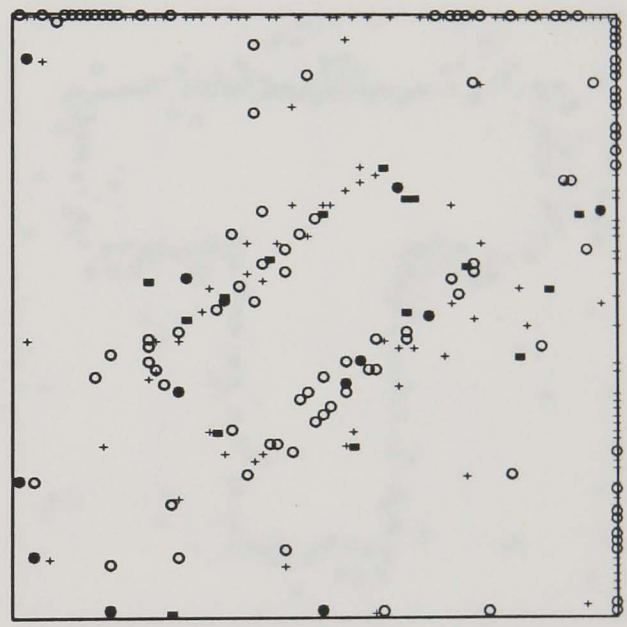


Fig A2-2(d) : 2.4 seconds

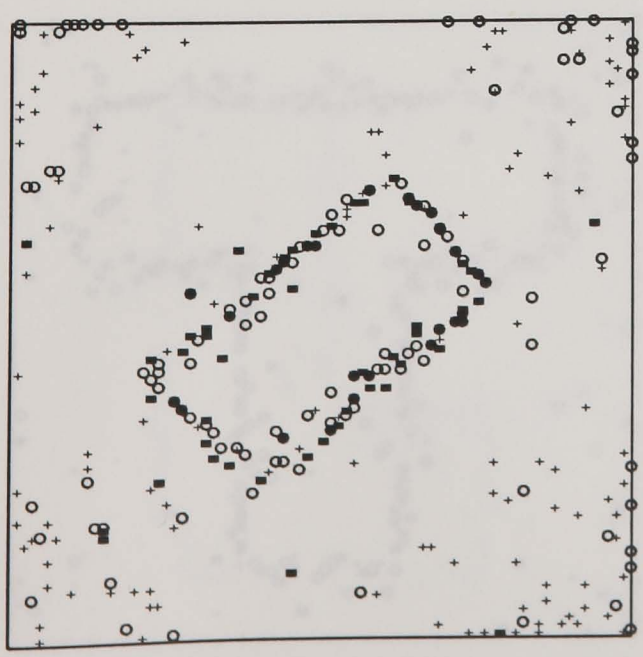


Fig A2-2(e) : 109.4 seconds

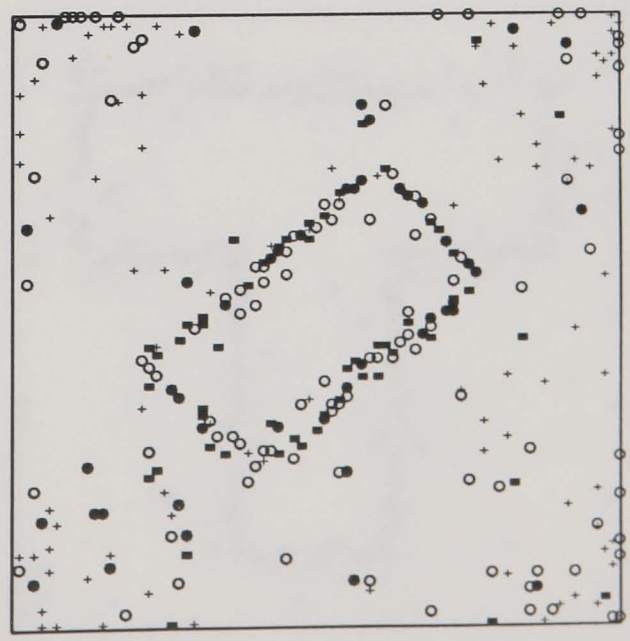


Fig A2-2(f) : 26.6 seconds

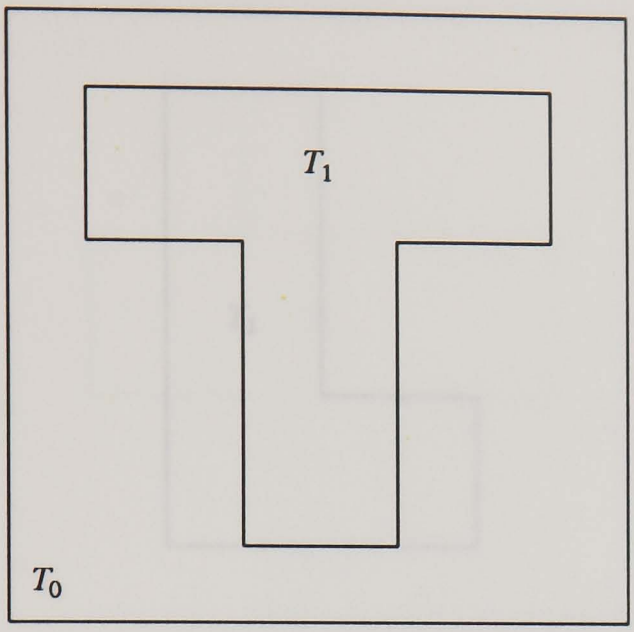


Fig A2-3(a) : true scene

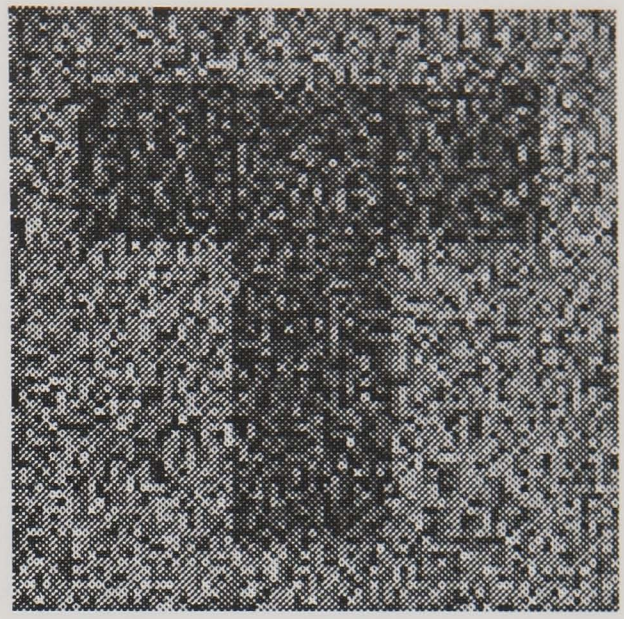


Fig A2-3(b) : image

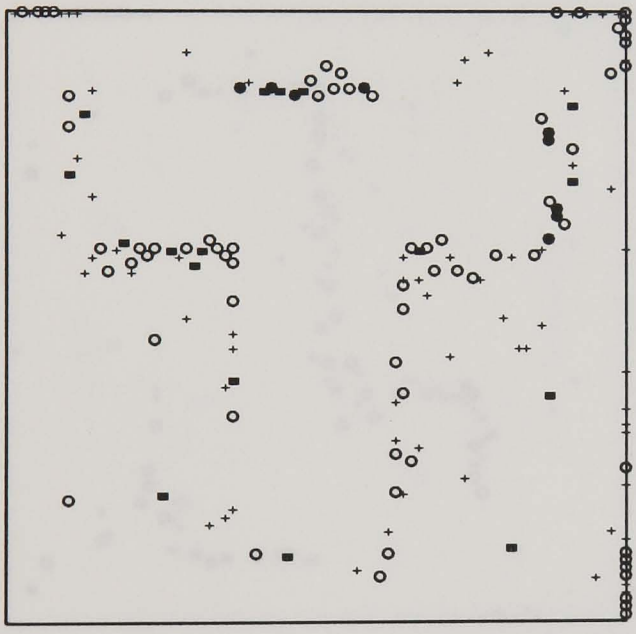


Fig A2-3(c) : 1.9 seconds

Fig A2-3(d) : 3.3 seconds

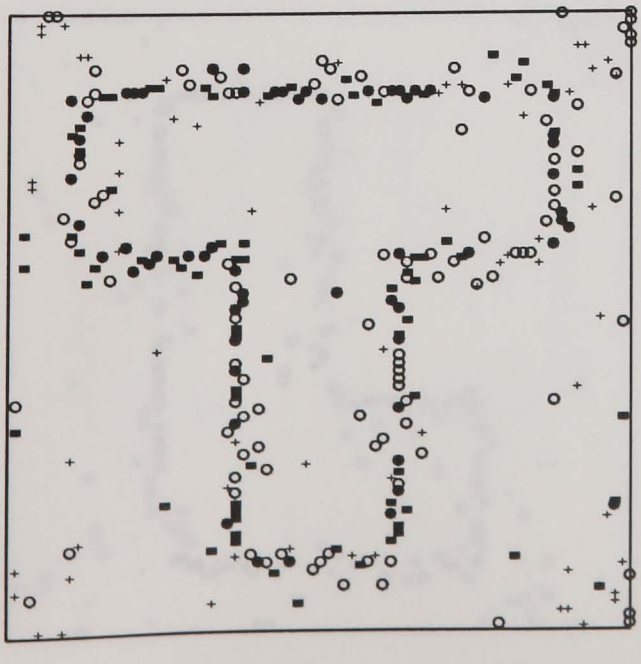


Fig A2-3(e) : 116.5 seconds

Fig A2-3(f) : 16.7 seconds

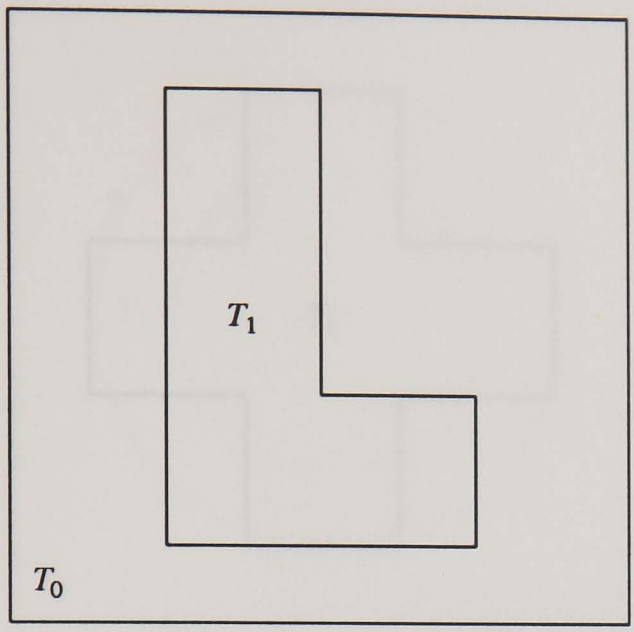


Fig A2-4(a) : true scene

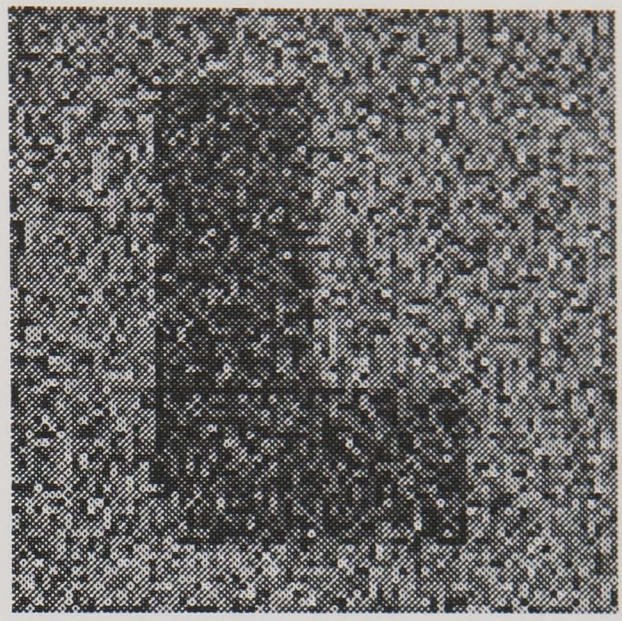


Fig A2-4(b) : image

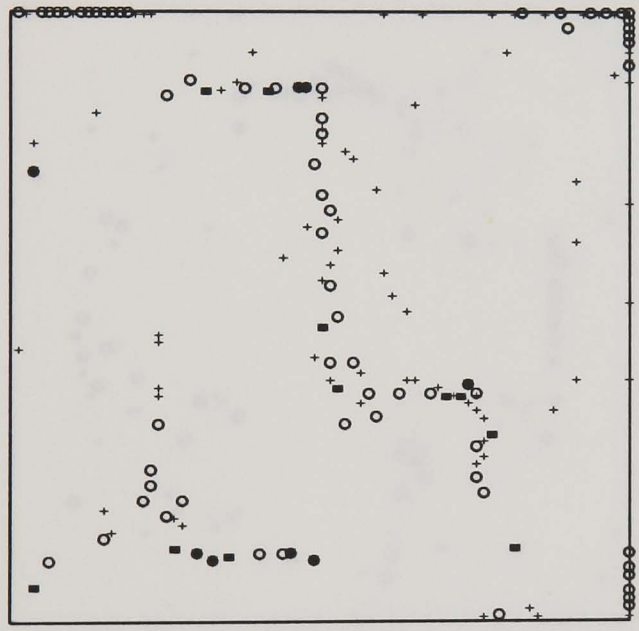


Fig A2-4(c) : 1.8 seconds

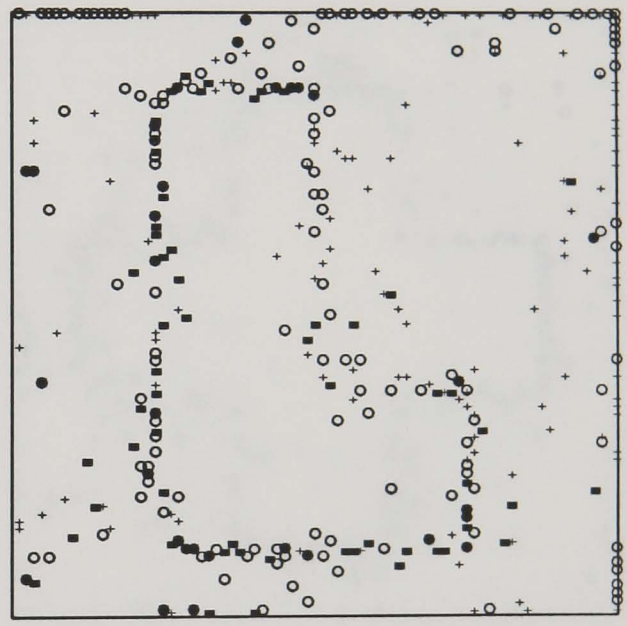


Fig A2-4(d) : 3.0 seconds

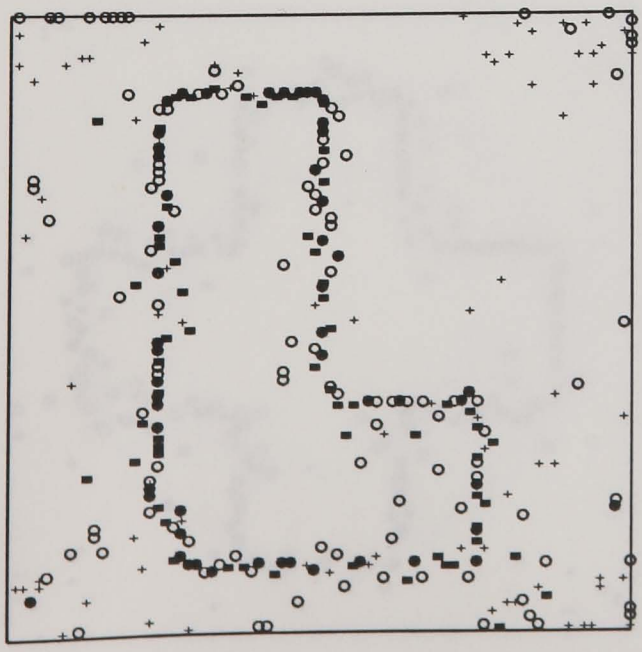


Fig A2-4(e) : 112.6 seconds

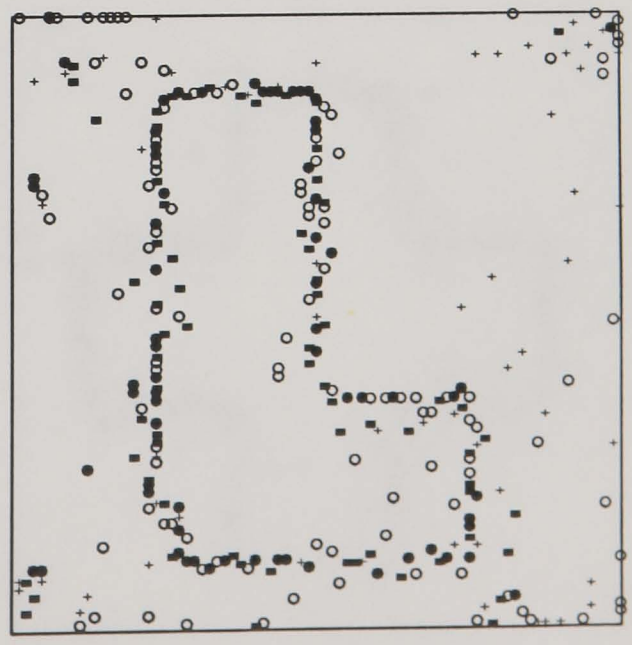


Fig A2-4(f) : 21.1 seconds

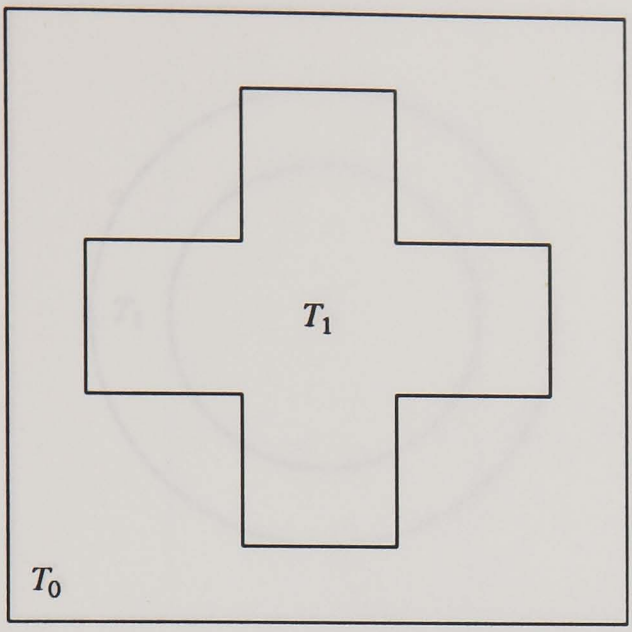


Fig A2-5(a) : true scene

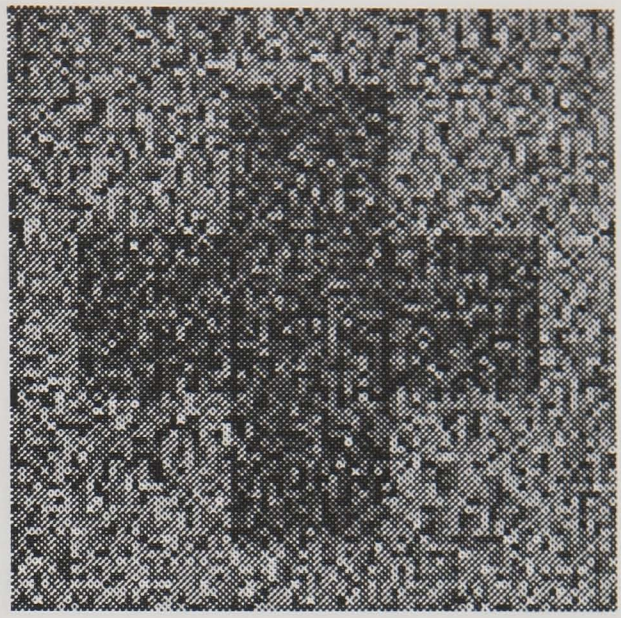


Fig A2-5(b) : image

Fig A2-5(c) : 1.8 seconds

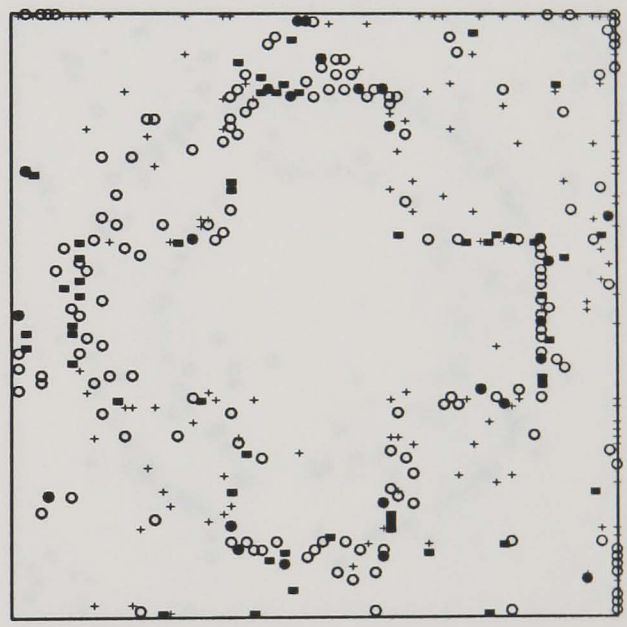


Fig A2-5(d) : 3.2 seconds

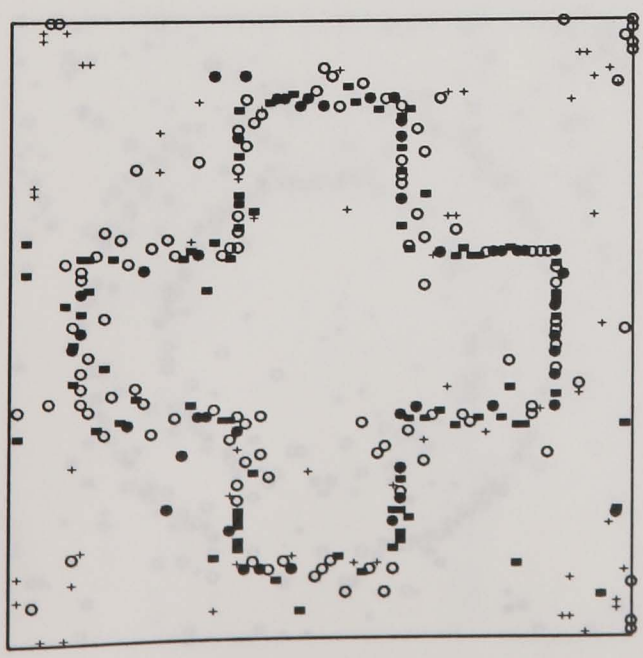


Fig A2-5(e) : 121.8 seconds

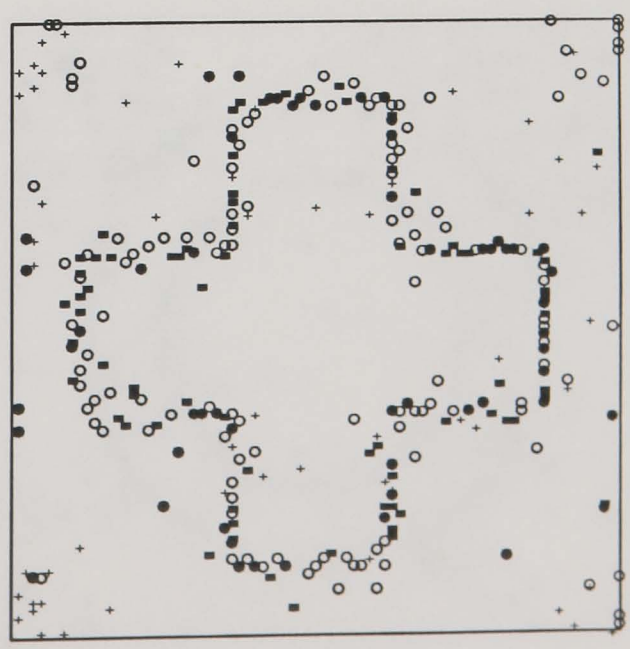


Fig A2-5(f) : 17.3 seconds

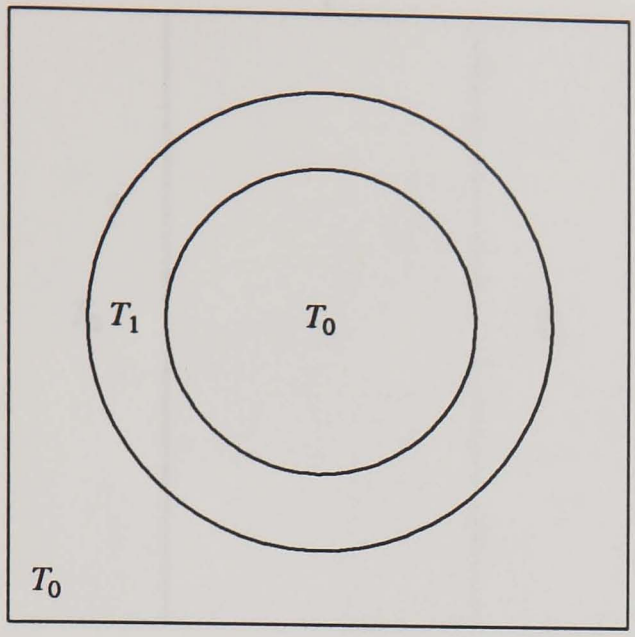


Fig A2-6(a) : true scene

Fig A2-6(b) : image

Fig A2-6(c) : 2.0 seconds

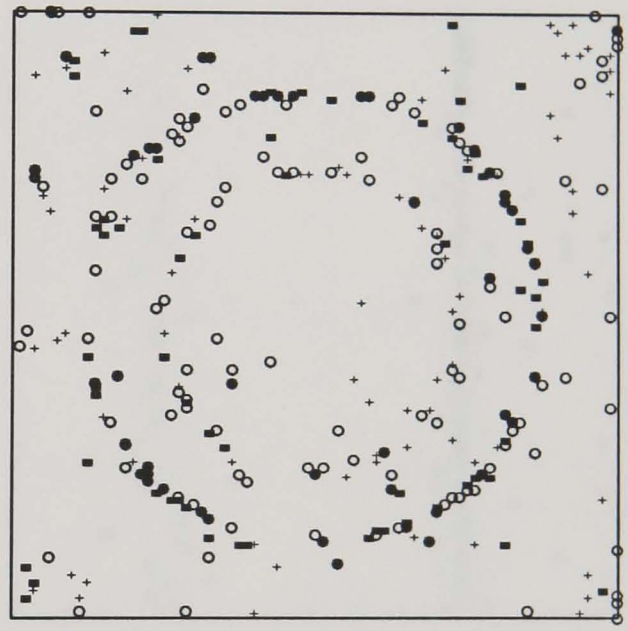


Fig A2-6(d) : 22.9 seconds

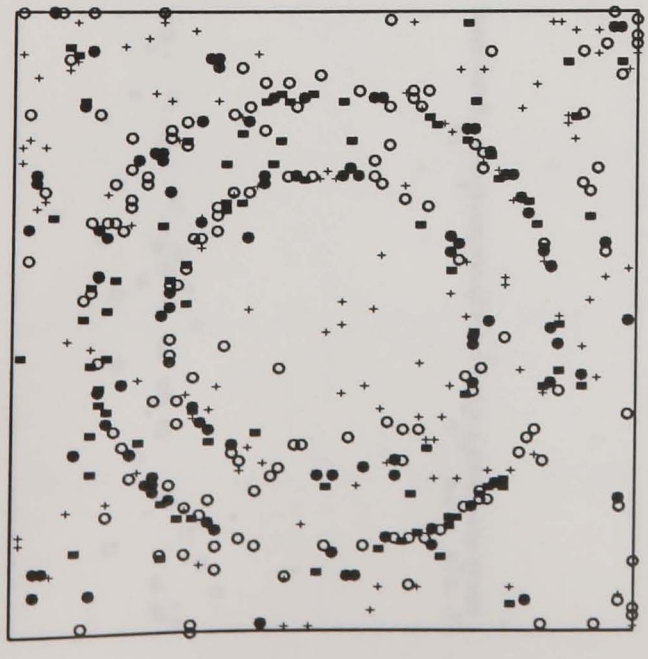


Fig A2-6(e) : 30.1 seconds

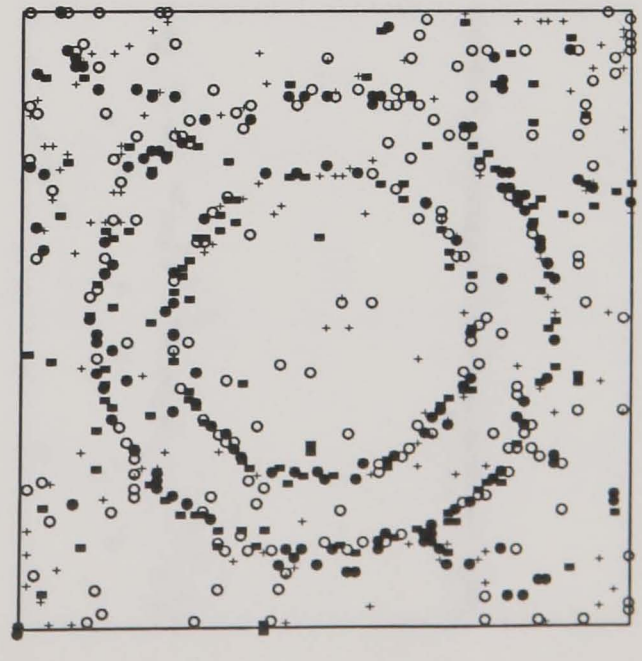


Fig A2-6(f) : 51.5 seconds

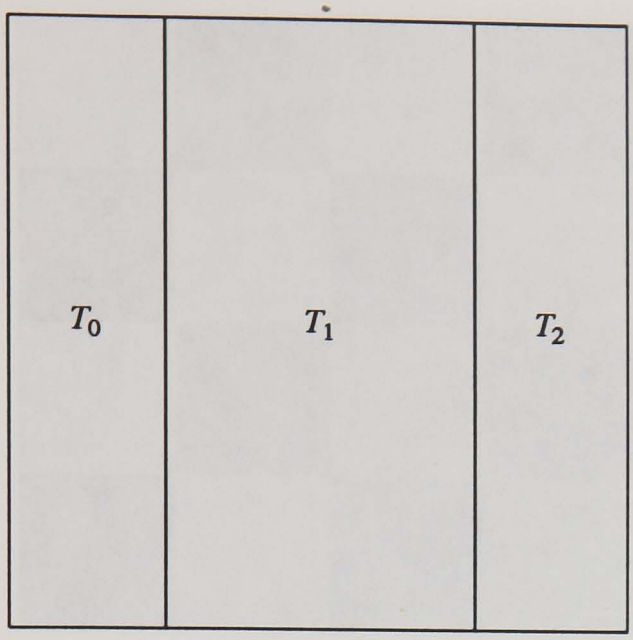


Fig A2-7(a) : true scene

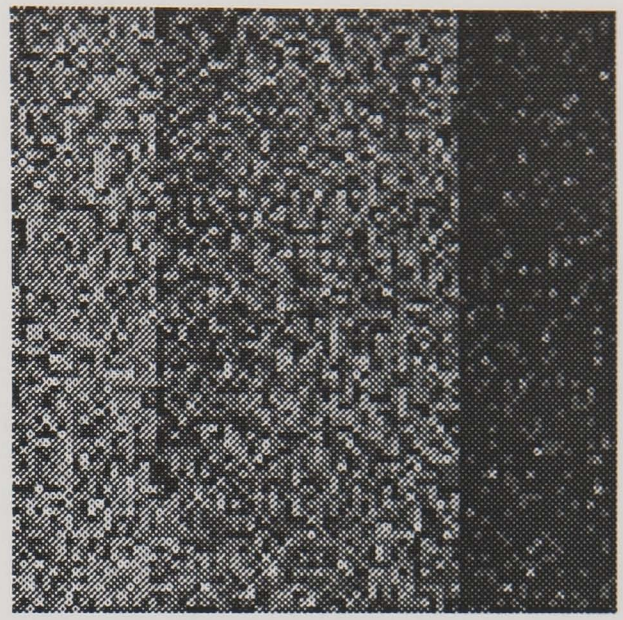


Fig A2-7(b) : image

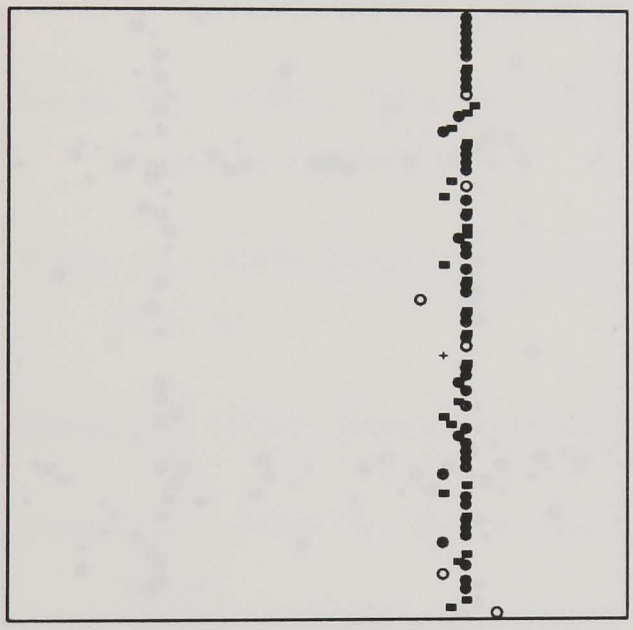


Fig A2-7(c) : 0.92 seconds

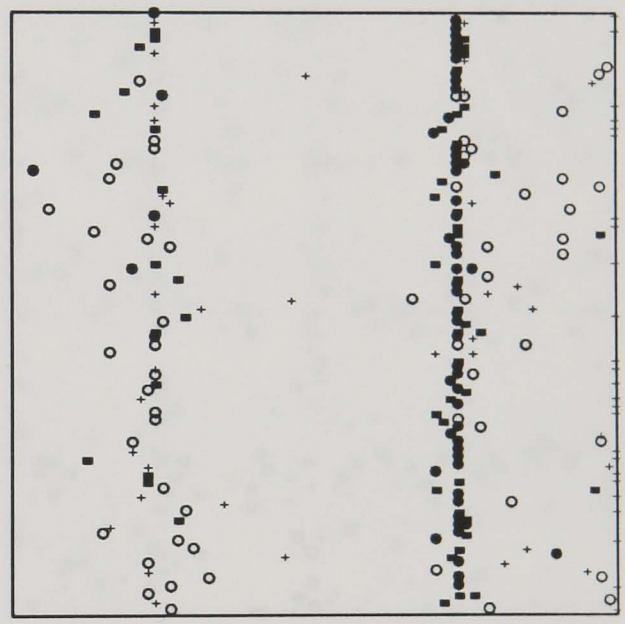


Fig A2-7(d) : 1.84 seconds

Fig A2-7(e) : 62.54 seconds

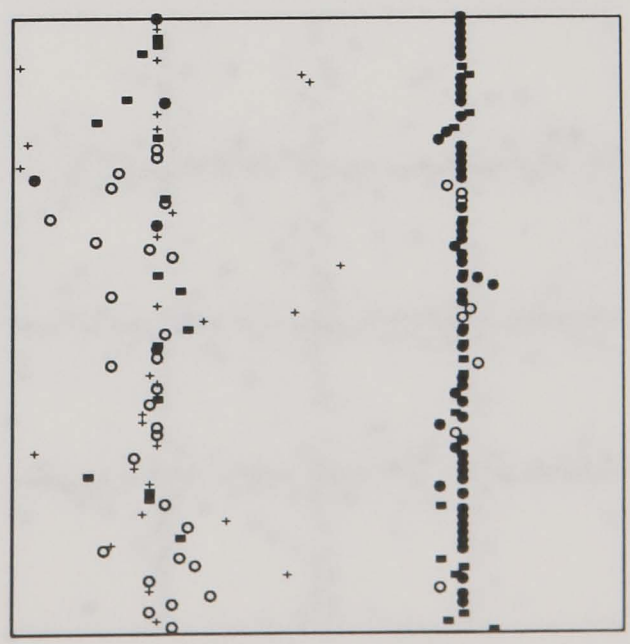


Fig A2-7(f) : 6.50 seconds

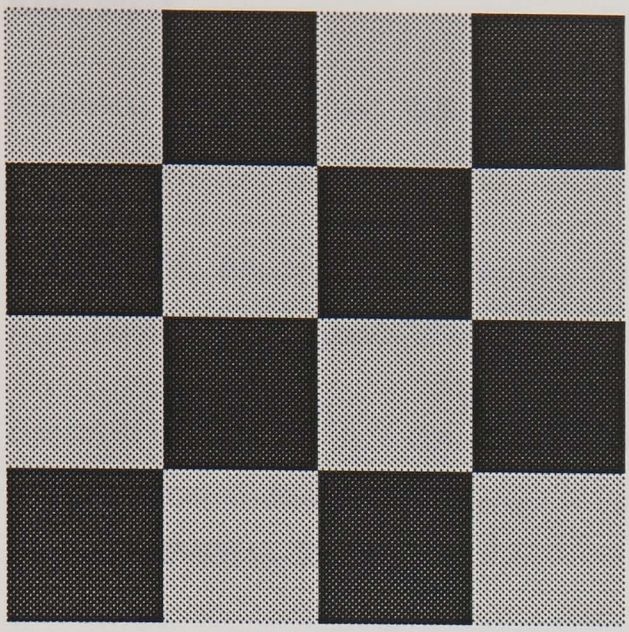


Fig A2-8(a) : true scene

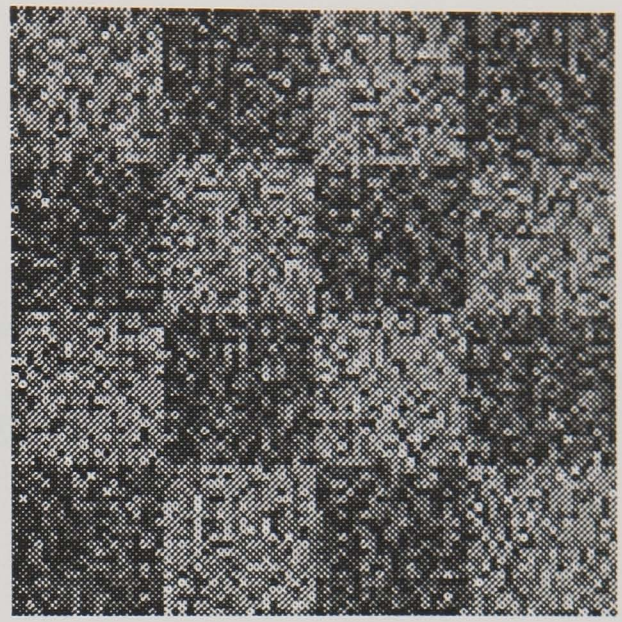


Fig A2-8(b) : image

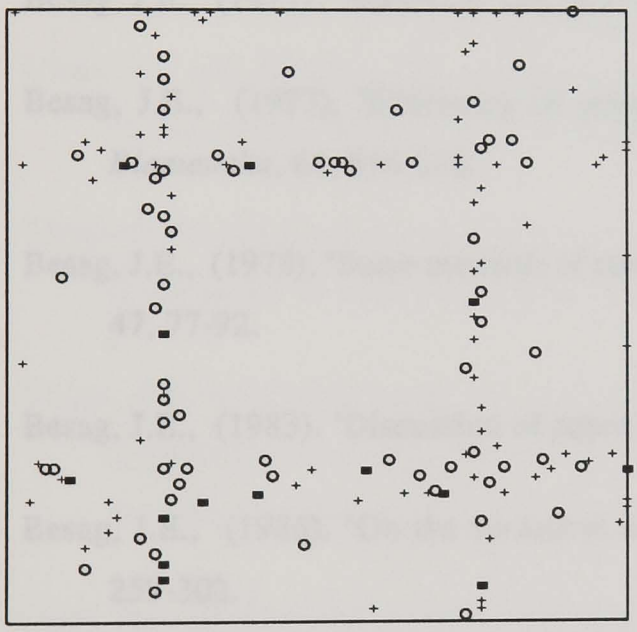


Fig A2-8(c) : 1.86 seconds

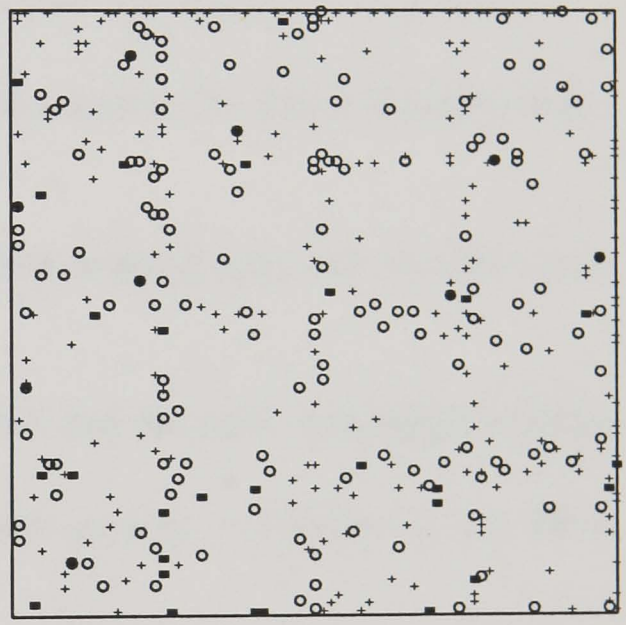


Fig A2-8(d) : 3.60 seconds

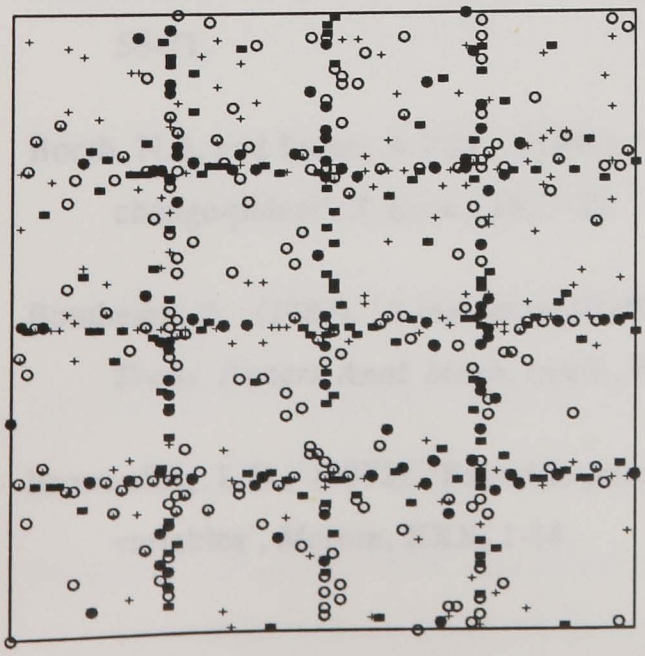


Fig A2-8(e) : 38.84 seconds

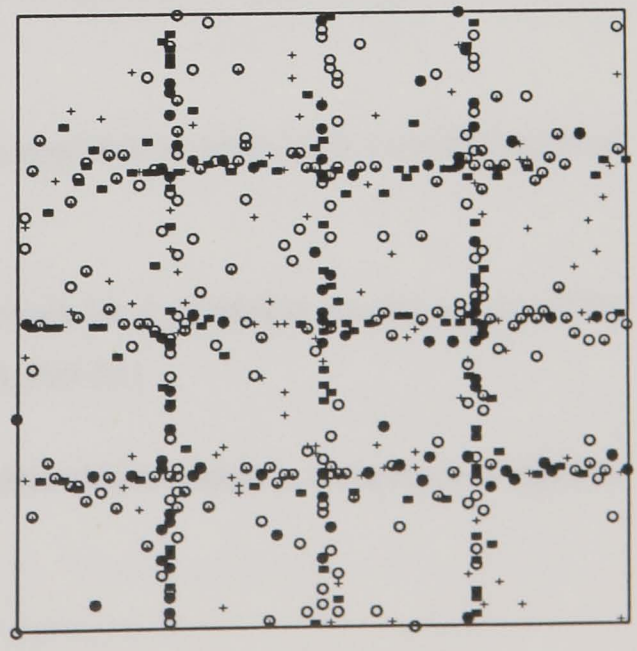


Fig A2-8(f) : 128.72 seconds

References

Andrews, H.C. and Hunt, B.R., (1977). *Digital Image Restoration*, Prentice-Hall, Englewood Cliffs.

Bartels, R.H., Beatty, J.C., and Barsky, B.A., (1987). *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*, Morgan Kaufman, California.

Basseville, M., (1981). 'Edge detection using sequential methods for change in level - Part II: Sequential detection of change in mean', *IEEE Trans. Acoust. Sound, Signal Processing*, **ASSP-29**(1), 32-50.

Besag, J.E., (1974). 'Spatial interaction and the statistical analysis of lattice systems', *J. R. Statist. Soc. B*, **36**(2), 192-236.

Besag, J.E., (1975). 'Statistical analysis of non-lattice data', *The Statistician*, **24**, 179-195.

Besag, J.E., (1977). 'Efficiency of pseudolikelihood estimation for simple Gaussian fields', *Biometrika*, **64**, 616-618.

Besag, J.E., (1978). 'Some methods of statistical analysis for spatial data', *Bull. Int. Statist. Inst.*, **47**, 77-92.

Besag, J.E., (1983). 'Discussion of paper by P. Switzer', *Bull. Int. Statist. Inst.*, **50**(3), 422-425.

Besag, J.E., (1986). 'On the statistical analysis of dirty pictures', *J. R. Statist. Soc. B*, **48**(3), 259-302.

Blake, A. and Zisserman, A., (1987). *Visual Reconstruction*, MIT press, Massachusetts.

Bookstein, F.L., (1979). 'Fitting conic sections to scattered data', *Comp. Graph. Im. Proc.*, **9**, 56-71.

Booth, N.B. and Smith, A.F.M., (1982). 'A Bayesian approach to retrospective identification of change-points', *J. Econ.*, **19**, 7-22.

Bouthemy, P., (1989). 'A maximum likelihood framework for determining moving edges', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(5), 499-511.

Broemeling, L.D., (1972). 'Bayesian procedures for detecting a change in a sequence of random variables', *Metron*, **XXX**, 1-14.

Broemeling, L.D., (1974). 'Bayesian inferences about a changing sequence of random variable variables', *Commun. Statist.*, **3**, 243-255.

Buck, C.E., Cavanagh, W.G., and Litton, C.D., (1988). 'The spatial analysis of site phosphate data', *Computer Applications in Archaeology, B.A.R. International series*, **446**(i).

Cannon, R.L., Dave, J.V., and Bezdek, J.C., (1986). 'Efficient implementation of the fuzzy c-means clustering algorithms', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-8**(2), 248-255.

Canny, J., (1986). 'A computational approach to edge detection', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-8**(6), 679-698.

Carlin, B.P., Gelfand, A.E., and Smith, A.F.M., (1989). Hierarchical Bayesian analysis of change point problems. Technical report, Carnegie-Mellon University

Chen, J.S. and Medioni, G., (1989). 'Detection, localization, and estimation of edges', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(2), 191-198.

Chernoff, H. and Zacks, S., (1964). 'Estimating the correct mean of a Normal distribution which is subjected to a change in time', *Ann. Math. Statist.*, **35**, 999-1018.

Cohen, F.S. and Cooper, D.B., (1987). 'Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-9**(2), 195-315.

Cook, R.D. and Weisberg, S., (1982). *Residuals and Influence in Regression*, Chapman and Hall, New York.

Cooper, D.B. and Sung, F.P., (1983). 'Multiple-window parallel adaptive boundary finding in computer vision', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-5**(3), 299-314.

Cross, G.R. and Jain, A.K., (1983). 'Markov random field texture models', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-5**(1), 25-39.

De Boor, C., (1978). *A Practical Guide to Splines*, Springer-Verlag, New York.

De Micheli, E., Caprile, B., Ottonello, P., and Torre, V., (1989). 'Localisation and noise in edge detection', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(10), 1106-1116.

Dempster, A.P., Laird, N.M., and Rubin, D.B., (1977). 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Statist. Soc. B*, **39**(1), 1-38.

Derin, H., Elliot, H., Cristi, R., and Geman, D., (1984). 'Bayesian smoothing algorithms for segmentation of binary images modeled by Markov random fields', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-6**(6), 707-720.

Derin, H. and Cole, W.S., (1986). 'Segmentation of textured images using Gibbs random fields', *Comp. Vis. Graph. Im. Proc.*, **35**, 72-98.

Derin, H. and Elliot, H., (1987). 'Modeling and segmentation of noisy and textured images using Gibbs random fields', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-9**(1), 39-55.

Dubes, R.C. and Jain, A.K., (1989). 'Random field models in image analysis', *J. Appl. Statist.*, **16**(2), 131-164.

Duda, R.D. and Hart, P.E., (1973). *Pattern Classification and Scene Analysis*, Wiley, New York.

Forbes, A.B., (1987). Fitting an ellipse to data. NPL report DITC 95/87.

Fu, K.S. and Yu, T.S., (1980). *Statistical Pattern Classification using Contextual Information*, Research Studies Press, Chichester.

Fukunaga, K., (1972). *Introduction to Statistical Pattern Recognition*, Academic Press, New York.

Gath, I. and Geva, A.B., (1989). 'Unsupervised optimal fuzzy clustering', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(7), 773-781.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M., (1989). Illustration of Bayesian inference in normal data models using Gibbs Sampling. Technical report, Department of Statistics, University of Nottingham

Gelfand, A.E. and Smith, A.F.M., (1990). Sampling based approaches to calculating marginal densities. To appear.

Geman, S. and Geman, D., (1984). 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-6**, 721-741.

Geman, S. and Graffigne, C., (1987). 'Markov random field image models and their applications to computer vision', *Proc. Int. Congr. Math.*, A.M. Gleason (ed.). Amer. Math. Soc., Providence, RI.

Green, P.J. and Silverman, B.W., (1979). 'Constructing the convex hull of a set of points in the plane', *The Computer Journal*, **22**(3), 262-266.

Greig, D.M., Porteous, B.T., and Seheult, A.H., (1989). 'Exact maximum a posteriori estimation for binary images', *J. R. Statist. Soc. B*, **51**(2), 271-279.

Gull, S.F. and Skilling, J., (1985). 'The entropy of an image', in *Maximum-Entropy and Bayesian Methods in Inverse Problems*, W.T. Gandy (ed.), Reidel, Dordrecht, 287-301.

Hansen, F.R. and Elliot, H., (1982). 'Image segmentation using simple Markov field models', *Comp. Graph. Im. Proc.*, **20**, 101-132.

Haralick, R.M., (1984). 'Digital step edges from zero crossing of second directional derivatives', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-6**(1), 58-68.

Hartigan, J.A., (1975). *Clustering Algorithms*, Wiley, New York.

Haslett, J., (1985). 'Maximum likelihood discriminant analysis on the plane using a Markovian model of Spatial Context', *Pattern Recognition*, **18**(3-4), 287-296.

Hassner, M. and Sklansky, J., (1980). 'The use of Markov random fields as models of texture', *Comp. Graph. Im. Proc.*, **12**, 357-370.

Hills, S.E., (1989). 'The Parametrisation of Statistical Models', Ph.D. thesis, University of Nottingham.

Hinkley, D.V., (1970). 'Inference about the changepoint in a sequence of random variables', *Biometrika*, **57**(1), 1-17.

Hinkley, D.V., (1971). 'Inference about the change-point from cumulative sum tests', *Biometrika*, **58**(3), 509-523.

Johnson, W. and Geisser, S., (1982). 'Assessing the predictive influence of observations', in *Statistics and Probability: Essays in Honor of C.R. Rao*, J.K. Ghosh (ed.), North-Holland, 343-358.

Kanal, L.N., (1980). 'Markov mesh models', *Comp. Graph. Image Proc.*, **12**, 371-375.

Kashyap, R.L., Chellappa, R., and Ahuja, N., (1981). 'Decision rules for choice of neighbors in random field models of images', *Comp. Graph. Image Proc.*, **15**, 301-318.

Kashyap, R.L. and Chellappa, R., (1983). 'Estimation and choice of neighbours in spatial-interaction models of Images', *I.E.E.E. Trans. Inf. Th.*, **IT-29**(1), 60-72.

Kashyap, R.L. and Eom, K.-B., (1989). 'Texture boundary detection based on the long correlation model', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(1), 58-67.

Kent, J.T. and Mardia, K.V., (1988). 'Spatial classification using fuzzy membership models', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-10**(5), 659-671.

Khotanzad, A. and Kashyap, R.L., (1987). 'Feature selection for texture recognition using image synthesis', *IEEE Trans. Syst., Man., Cybern.*, **SMC-17**, 1087-1095.

Kimeldorf, G. and Wabha, G., (1970). 'A correspondence between Bayesian estimation on stochastic processes and smoothing by splines', *Ann. Math. Statist.*, **41**, 495-502.

Klein, R. and Press, S.J., (1987). 'Spatial structure in Bayesian classification', Tech. Rep. No. 157, Dept. of Statistics, University of California Riverside.

Kunsch, H.R., (1987). 'Intrinsic autoregressions and related models of the two-dimensional lattice', *Biometrika*, **74**(3), 517-524.

Lakshmanan, S. and Derin, H., (1989). 'Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(8), 799-813.

Lu, Y. and Jain, R.C., (1989). 'Behavior of edges in scale space', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(4), 337-356.

Mardia, K.V., (1984). 'Spatial discrimination and classification maps', *Comm. Stat. B*, **13**(18), 2181-2197.

Mardia, K.V. and Hainsworth, T.J., (1988). 'A spatial thresholding method for image segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-10**(6), 919-927.

Marr, D. and Hildreth, E., (1980). 'Theory of edge detection', *Proc. R. Soc. Lond. B*, **207**, 187-217.

Mascarenhas, N.D.A. and Prado, L.O.C., (1980). 'A Bayesian approach to edge detection in images', *I.E.E.E. Trans. Aut. Cont.*, **AC-25**(1), 36-43.

Molina, R. and Ripley, B.D., (1989). 'Using spatial models as priors in astronomical image analysis', *J. Appl. Statist.*, **16**(2), 193-206.

Moore, M., (1984). 'On the estimation of a convex set', *Ann. Statist.*, **12**(3), 1090-1099.

Nalwa, V.S. and Binford, T.O., (1986). 'On detecting edges', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-8**(6), 699-714.

Nevatia, R. and Babu, K.R., (1980). 'Linear feature extraction and description', *Comp. Graphics and Image Processing*, **13**, 257-269.

Peli, T. and Malah, D., (1982). 'A study of edge detection algorithms', *Comp. Graph. Im. Proc.*, **20**, 1-21.

Perez, A. and Gonzalez, R.C., (1987). 'An iterative thresholding algorithm for image segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-9**(6), 742-751.

Pettit, L.I. and Smith, A.F.M., (1983). 'Outliers and influential observations in linear models', in *Bayesian Statistics 2*, J.M. Bernardo et al. (ed.), Elsevier Science Publishers B.V. (North-Holland), 473-494.

Pettitt, A.N., (1980). 'A simple cumulative sum type statistic for the changepoint problem with zero-one observations', *Biometrika*, **67**(1), 79-84.

Pratt, W.K., (1978). *Digital Image Processing*, Wiley, New York.

Rao, C.R, (1973). *Linear statistical inference and its applications*, Wiley, New York.

Ridler, T.W. and Calvard, S., (1978). 'Picture thresholding using an iterative selection method', *IEEE Trans. Syst., Man., Cybern.*, **SMC-8**, 630-632.

Ripley, B. D. and Rasson, J.-P., (1977). 'Finding the edge of a Poisson forest', *J. Appl. Prob.*, **14**, 483-491.

Ripley, B.D., (1981). *Spatial Statistics*, Wiley, New York.

Ripley, B.D., (1987). *Stochastic Simulation*, Wiley, New York.

Ripley, B.D., (1988). *Statistical Inference for Spatial Processes*, CUP, Cambridge.

Roberts, G.O. and Polson, N.G., (1990). A survey of geometric ergodicity of Markov chains. Technical report, Department of Statistics, University of Nottingham.

Rosenfeld, A. and Kak, A.C., (1982). *Digital Picture Processing*, Academic Press, New York.

Schowengerdt, R.A., (1983). *Techniques for Image Processing and Classification in Remote Sensing*, Academic Press, Orlando.

Shaban, S.A., (1980). 'Change point problem and two-phase regression: an annotated bibliography', *Inst. Statist. Review*, **48**, 83-93.

Shahraray, B. and Anderson, D.J., (1989). 'Optimal estimation of contour properties by cross-validated regularization', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(6), 600-610.

Silverman, B.W. and Titterington, D.M., (1981). 'Minimum covering ellipses', *SIAM J. Sci. Stat. Comput.*, **1**(4), 401-409.

Silverman, B.W., (1985). 'Some aspects of the spline smoothing approach to non-parametric regression curve fitting', *J. R. Statist. Soc. B*, **47**(1), 1-52.

Silverman, B.W., Jones, M.C., Wilson, J.D., and Nychka, D.W., (1990). A smoothed EM approach to indirect estimation problems with particular reference to stereology and emission tomography. To appear.

Smith, A.F.M., (1975). 'A Bayesian approach to inference about a change-point in a sequence of random variables', *Biometrika*, **62**(2), 407-416.

Smith, A.F.M. and Spiegelhalter, D.J., (1980). 'Bayes factors and choice criteria for linear models', *J. R. Statist. Soc. B*, **42**(2), 213-220.

Smith, A.F.M., (1983). 'Bayesian approaches to outliers and robustness', in *Specifying Statistical Models*, J. Florens et al. (ed.), 13-35.

Smith, A.F.M., (1986). 'Some Bayesian thoughts on modelling and model choice', *The Statistician*, **35**, 97-102.

Spiegelhalter, D.J. and Smith, A.F.M., (1982). 'Bayes factors for linear and log-linear models with vague prior information', *J. R. Statist. Soc. B*, **44**(3), 377-387.

Switzer, P., (1980). 'Extensions of linear discriminant analysis for statistical classification of Remotely Sensed satellite imagery', *Mathematical Geology*, **12**(4), 367-376.

Switzer, P., (1983). 'Some spatial statistics for the interpretation of satellite data', *Bull. Int. Statist. Inst.*, **50**(2), 962-972.

Tanner, M. and Wong, W., (1987). 'The calculation of posterior distributions by data augmentation', *J. Amer. Statist. Ass.*, **82**, 528-550.

Torre, V. and Poggio, T.A., (1986). 'On edge detection', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-8**(2), 147-163.

Woods, J.W., (1981). 'Two-dimensional Kalman filtering', in *Two-Dimensional Transforms and Filters*, T.S. Huang (ed.), Springer-Verlag, Berlin, 155-205.

Woods, J.W., Dravida, S., and Mediavilla, R., (1987). 'Image estimation using doubly stochastic Gaussian random field models', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-9**(2), 245-253.

Zacks, S., (1982). 'Classical and Bayesian approaches to the changepoint problem : fixed sample and sequential procedures', *Statistique et Analyse des Donnees*, **1**, 48-81.

Zhou, Y.T., Venkateswar, V., and Chellappa, R., (1989). 'Edge detection and linear feature extraction using a 2-D random field model', *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-11**(1), 84-94.

Additional reference - Chapter 6.

Mardia, K.V., and Holmes, D., (1980). 'A statistical analysis of megalithic data under elliptic pattern', *J. R. Statist. Soc. A*, **143**, 293-302.