STATISTICAL MODELLING AND INFERENCE IN IMAGE ANALYSIS

Wei QIAN, B.Sc., M.Sc.

A Dissertation Submitted To The

UNIVERSITY OF GLASGOW

For The Degree Of

Doctor of Philosophy

TO MY PARENTS

ACKNOWLEDGEMENTS

The author would like to express his sincere gratitude to his supervisor, Professor D. M. Titterington, for introducing him to this interesting research area, and for his guidance and encouragement during the period from 1988 - 1990, and in particular, for his help in writing this thesis. In addition, the author would like to thank Dr. Jim Kay for useful discussions and all the members in the Department of Statistics for their assistance, as well as friendship.

The author is indebted to many friends in Glasgow, especially he would like to express his appreciation to C. Hu, Z.Li, Z. Wang, T. Tang, X. Chen, and F. Yang who have made the author's life easier.

This work was supported by a Postgraduate Scholarship from Glasgow University and an Overseas Research Students award.

Finally, The author would like to thank his family, his parents and his wife for their understanding and support.

The work included in Chapter 2 was published in Statist. Prob. Letters, 1990, 10, 49-58. A short version of Section 3.3 was published in Adv. Appl. Prob., 1990, 22, 755-757. Parts of Sections 3.4, 3.5 and that of Chapter 5 were published in J. Appl. Statist., 1989, 16, 267-281, other parts of them have been submitted to J. Statist. Comput. Simul.. The work included in Chapter 4 has been submitted to Stochastic Processes & Their Applications, and that included in Chapter 6 will appear in Signal Processing. Finally, the work in Chapter 7 will appear in J. R. Statist. Soc., B, 1991.

CONTENTS

Chapter 1

Introduction

Chapter 2

Parameter Estimation For Gibbs Chains And Hidden Gibbs Chains

Chapter 3

Inference For Markov Random Field Models

Chapter 4

Normal Approximations For Lattice Systems

Chapter 5

Parameter Estimation For Hidden Markov Random Fields

Chapter 6

Three-Dimensional Markov Mesh Models

Chapter 7

Multi-Dimensional Markov Chain Models For Image Textures

Chapter 8

SUMMARY

The aim of the thesis is to investigate classes of model-based approaches to statistical image analysis. We explored the properties of models and examined the problem of parameter estimation from the original image data and, in particular, from noisy versions of the the scene. We concentrated on Markov random field (MRF) models, Markov mesh random field (MMRF) models and Multi-dimensional Markov chain (MDMC) models.

In Chapter 2, for the one-dimensional version of Markov random fields, we developed a recursive technique which enables us to achieve maximum likelihood estimation for the underlying parameter and to carry out the EM algorithm for parameter estimation when only noisy data are available. This technique also enables us, in just a single pass, to generate a sample from a one-dimensional Markov random field. Although, unfortunately, this technique cannot be extended to two- or multi-dimensional models, it was applied to many cases in this thesis. Since, for two-dimensional Markov random fields, the density of each row (column), conditionally on all other rows (columns) is of the form of a one-dimensional Markov random field, and since the distribution of the original image, conditionally on the noisy version of data, is still a Markov random field, the technique can be used on different forms of conditional density of one row (column). In Chapter 3, therefore, we developed the line-relaxation method for simulating MRFs and maximum line pseudo-likelihood estimation of parameter(s), and in Chapter 5, we developed a simultaneous procedure of parameter estimation and restoration, in which line pseudo-likelihood and a modified EM algorithm were used.

The first part of Chapter 3 and Chapter 4 concentrate on inference for two-dimensional MRFs. We obtained a matrix expression for partition functins for general models, and a more explicit form for a multi-colour Ising model, and thus located the positions of critical points of this multi-colour model. We examined the asymptotic properties of an asymmetric, two-colour Ising model. For general models, in Chapter 4, we explored asymptotic properties under an "independence" or a "near independence" condition, and then developed the approach of maximum approximate-likelihood estimation.

For three-dimensional MMRF models, in chapter 6, a generalization of Devijver's F-G-H algorithm is developed for restoration.

In Chapter 7, the recursive technique was again used to introduce MDMC models, which form a natural extension of a Markov chain. By suitable choice of model parameters, textures can be generated that are similar to those simulated from MRFs, but the simulation procedure is computationally much more economical. The recursive technique also enables us to maximize the likelihood function of the model.

For all three sorts of prior random field models considered in this thesis, we developed a simultaneous procedure for parameter estimation and image restoration, when only noisy data are available. The currently restored image was used, together with noisy data, in modified versions of the EM algorithm. In simulation studies, quite good results were obtained, in terms of estimation of parameters in both the original model and, particularly, in the noise model, and in terms of restoration.

Chapter 1

Introduction

1.1 Image analysis and problems

A digital image is a multi-dimensional vector with positive elements, which represent a pattern of radiant energy emitted by objects in space. There is a very wide range of practical problems requiring image processing. Examples include: various types of satellite data; ultra-sound; thermal images; nuclear medicine; computer vision; electron micrography and astronomy.

There are usually two kinds of image; grey-level images and texture images. In the former each element takes a real value, while in the latter each element takes a value, or state, or colour, from a finite state space, for example {1,2,....,S}. The present thesis is mainly concentrated on texture images.

Due to the nature of image blurring and data acquisition, a general problem in image processing is to remove the effect of blur and noise. For this inverse problem, models play a very important role. Models in image analysis are required to serve a dual role, both as descriptions of images that are observed in practice and also as means to generate synthetic images from image parameter(s).

Therefore, there are three main problems in image analysis. The first is to find and to investigate the properties of, suitable models which may include the original models for the underlying images and the 'noise-models' for observed data; the second is to recover the original image from a noisy image(**Restoration**); the third is to estimate the parameter(s) in the models.

1.2 Stochastic constraints and Bayesian inference

The stochastic model-based approach to image analysis is currently very active, involving the use of random fields to provide constraints on the original images. Some authors regard the constraints on the original images as penalization. In the real scene, the image value or colour at one site typically has a relationship with the colours of its neighbouring sites. The objective of stochastic constraints is mainly to capture mostly these neighbour relationships.

Let $x=\{x_t;\ t\in D\}$ be the original image, where D is usually a regular lattice in multi-dimensional space. The lattice elements(**pixels**) correspond to small patches in the scene. Some other problem-specific image attributes, such as classification or boundary labels, can also be regarded as part of x (Geman and Geman, 1984, Silverman, et. al. 1990). Let Ω denote the set of all possible x. It is assumed in the **Bayesian approach** to image analysis that X=x is a sample realization from a distribution $p(x|\beta)$ over Ω. $p(x|\beta)$ is then referred to as the '**Prior distribution**'. It is also usually assumed that for all $t\in D$, x_t takes values from a specific set, either infinite or finite, corresponding to grey-level images and texture images respectively.

Now let Y denote the observed process, and Y=y be the observed data. (In practice, there may be repeated observations of Y, or additional observations obtained through different mechanisms from that for Y, or even more complicated situations, but we shall imagine that there is only a single observation process.) Denote by $f(y|x,\theta)$ the conditional distribution of Y given X. $f(y|x,\theta)$ could represent non-random transformations from X to Y, such as linear transformations, or random transformations which involve, for instance, optical blurring or obscurations(i.e. missing observations). The sample space of Y is often different from that of X. For example, X may be a texture image and Y a set of observed intensities.

The two distributions, $p(x|\beta)$ and $f(y|x,\theta)$, obviously determine the joint distribution of X and Y as well as the **Posterior** distribution of X given Y, denoted by $P(x|y,\beta,\theta)$. The modern methods of model-based image restoration, such as **Maximum A Posteriori(MAP)** restoration(Geman and Geman, 1984) and the **Iterated Conditional Modes(ICM)** method(Besag, 1986), all depend upon the posterior $P(x|y,\beta,\theta)$, under the assumption that the parameters, β and φ, are known.

The estimation problem for parameters β and θ, which may well not be known in practice, from the observed process Y=y, can be regarded as an incomplete-data or missing-data problem, or as a mixture-distribution problem. It is well known that the EM algorithm (Dempster, et al., 1977) is often used to treat this kind of problem. However, due to the complicated structure of the distributions, especially the prior distribution $p(x|\beta)$, the EM algorithm is impossible to implement, except in some very special cases (Geman and McClure, 1987, Titterington, et al 1985. Titterington 1989) or by Monto Carlo methods (Geman and McClure, 1985, Chalmond, 1988, Younes 1988b).

1.3 Markov random field models and Markov mesh models

There has been considerable recent interest in Bayesian inference, particularly involving prior models based on Markov random fields. It has been mentioned that in real scenes, one pixel is typically related to its neighbouring pixels. The neighbouring pixels of any pixel are determined by a neighbourhood system, and the '**order**' of the model reflects the size of the considered neighbourhood. A Markov random field model is classified as either causal or noncausal depending on the structure of the neighbourhood. In causal models, the concept of the '**past**' of a pixel is introduced and only the past neighbouring pixels influences the current pixel. Causal Markov random fields are generally called Markov mesh random fields or simply Markov meshes, while noncausal Markov random fields are called Markov random fields. We now review some results for these two sorts of model, together with their application in image analysis.

1.3.1 **Markov random fields(MRF)**

Following Cross and Jain(1983), we give the following definition of a Markov random field.

Definition: A Markov random field is a joint probability density $p(x)$ on Ω subject to the following conditions:

1). Positivity: $p(x) > 0$ for any $x \in \Omega$;

2). Markov property: $p(x_i|x\backslash x_i) = p(x_i|x \text{ at neighbours of } i)$.

In practice, we shall define a Markov random field by first choosing a neighbourhood system, by naming the neighbours of each pixel, and then selecting $p(x)$ from within the corresponding class of probability distributions. Denote by ∂i the neighbours of pixel i, so that $p(x_i|x\backslash x_i) = p(x_i|x_{\partial i})$. Due to the not-immediately-obvious consistency conditions, identified by the Hammersley-Clifford theorem (see for example, Besag(1974, 1986)), it is necessary to preserve symmetry in neighbourhood system: that is, if j is a neighbour of i then i must be a neighbour of j. The general form of MRF distributions is given by Besag(1974). It is well known that the MRF model is equivalent to the **Gibbs random field model**, and we therefore only give the form of the **Gibbs distribution** as follows:

$$p(x|\beta) = \frac{1}{C(\beta)}\exp\{-U(x,\beta)\} \qquad (1.3.1)$$

where $C(\beta)$ is a normalizing constant, called the **partition function** in the discrete-model case, and the energy function $U(x,\beta)$ is of the form

$$U(x,\beta) = \sum_{c} V_c(x,\beta) \qquad (1.3.2)$$

in which c ranges over **cliques** associated with the specified neighbourhood system on D, and the potentials $V_c(x,\beta)$ are functions supported on them. (A clique c is either a single pixel or a set of pixels such that every pair of distinct pixels in c are neighbours.)

Although they have been studied extensively during the last 50 years in statistical physics,——especially the Ising model (Kaufman, 1949, Kramers and Wannier, 1944, Newell and Montroll, 1953), Markov random fields are relative newcomers on the mathematical and statistical scenes, remarkably little progress on **inference methodology** has been made, especially in the discrete-model case. Tjostheim(1978, 1983) developed classes of spatial series models which are extensions of one dimensional time-series models to multi-dimensional space, and the asymptotic properties of estimation are investigated. Pickard(1976, 1977, 1982, 1987) developed the asymptotic inference for the Ising model. Possolo(1986) discussed methods of parameter estimation for binary MRFs. For the application of MRF to image processing, we refer to two seminal papers by Geman and Geman(1984) and Besag(1986). Geman and Geman introduced an ingenious technique for maximizing the posterior distribution, through simulated annealing and Gibbs samplers. Although the procedure enables escape from any local maxima to occur, it is, computationally very demanding. Besag's(1986) **iterated conditional modes(ICM)** method concentrates on the local dependence structure of the MRFs, and produces restored images very cheaply and quickly, but it usually only exhibits local convergence. For further work on the use and estimation of MRF models with applications in image analysis, see Chapter 4 and Chapter 5 of the present thesis.

We shall concentrate on pairwise interaction MRFs(Besag, 1974, 1986), in which, for any x in the sample space Ω,

$$p(x|\beta) = \frac{1}{C(\beta)}\exp\{ \sum_i G_i(x_i,\beta) + \sum_i\sum_j G_{ij}(x_i,x_j,\beta) \} \qquad (1.3.3)$$

where, to ensure the Markov property, $G_{ij}\equiv 0$ unless i, j are neighbours.

Among this class of MRF models, Auto-normal models are often used in practice, especially in plant ecology. Following Besag(1974, 1986), for the simplest description of a Gaussian MRF with zero mean, $X_i=x_i$ has conditional density:

$$p(x_i|x_{\partial i}) \propto \exp\{-\tfrac{1}{2}(x_i - \sum_{j\neq i} \beta_{ij}x_j)^2/\lambda_i\} \qquad (1.3.4)$$

where (1) $\beta_{ij}\equiv 0$ unless pixels i and j are neighbours, (2) $\beta_{ij}\lambda_j=\beta_{ji}\lambda_i$. There may sometimes be further constraints on parameters. Then it follows that

$$p(x) = (2\pi)^{-\frac{1}{2}n}|Q|^{\frac{1}{2}}\exp\{ -\tfrac{1}{2}x'Qx\} \qquad (1.3.5)$$

where $Q=\Lambda^{-1}B$, Λ is the nxn diagonal matrix with diagonal entries λ_i, B is an nxn matrix with unit diagonal entries and with $-\beta_{ij}$ as (i,j)-th element, and n is the dimension of x. We shall use formula (1.3.5) in our later discussion about the ICM method and the EM algorithm in Chapter 5.

1.3.2 Markov meshes

The Markov mesh model is a sort of multi-dimensional generalization of the Markov chain. It was introduced by Abend et al(1965), and was further developed by Kanal(1980), Devijver(1988) and Lacroix(1987) for the two-dimensional case. In the model, the pixels are ordered, for example, in the two-dimensional case, through a diagonal direction. To be precise, the **past** of pixel (i,j) is defined as {(m,n): m<i or n<j}. Then, for the second-order model, the conditional density of pixel (i,j), given states at all past pixels, is given by

$$P(x_{ij}|x_{mn}, m<i \text{ or } n<j) = P(x_{ij}|x_{i-1,j},x_{i,j-1}) \qquad (1.3.6)$$

All these conditional densities determine the distribution over the entire lattice. Although the models exhibit causal dependence in that samples from them are generated directionally, they can find application in image segmentation, texture analysis and synthesis. For the pixel labelling problem, Devijver(1988) and Lacroix(1987) develoed the F-G-H algorithm, in which, a "local decomposition relationship" and a "lattice recurrence relationship" are used to calculate the density of each pixel, conditionally upon all the observed data on the past rectangular pixels. In Chapter 6, we shall extend this algorithm to the three-dimensional case.

1.4 Missing data and the EM algorithm

It often happens in practice that some data are not observed directly, but only indirectly, through other data. More precisely, we assume there are two sample spaces $\mathfrak{X}$ and $\mathfrak{Y}$. X from $\mathfrak{X}$ is not observed; only a realization from $\mathfrak{Y}$, namely y, is observed. The EM algorithm is then an iterative computational approach for maximizing the marginal distribution of y to find maximum likelihood estimates. Since the relationship between X and Y can vary widely in nature, the EM algorithm is useful for many practical problems, as Dempster et al described:

The EM algorithm is remarkable in part because of the simplicity and generality of the associated theory, and in part because of the wide range of examples which fall under its umbrella.

Suppose that the joint density over $\mathfrak{X}\times\mathfrak{Y}$ is $f(x,y|\Phi)$ with unknown parameter Φ, and that Y=y is observed. x is then called **missing data.** The EM algorithm is therefore designed to find the maximizer of

$$g(y|\Phi) = \int_{\mathfrak{X}} f(x,y|\Phi)dx. \qquad (1.4.1)$$

We give a brief illustration of the EM algorithm for the following exponential family cases. Suppose we know the distributional form of X, denoted by $p(x|\beta)$, and the conditional distribution of Y, given X, namely, $f(y|x,\theta)$, and that they take the form

$$p(x|\beta) = B_1(x)\exp\{\mu(\beta)'T_1(x)\}/A_1(\beta) \qquad (1.4.2)$$

$$f(y|x,\theta) = B_2(x,y)\exp\{\eta(\theta)'T_2(x,y)\}/A_2(\theta), \qquad (1.4.3)$$

where $\mu(\beta)$, $\eta(\theta)$, $T_1(x)$ and $T_2(x,y)$ are all vectors with compatible dimensions. It is known that Y is a sort of mixture distribution if X only takes values from a finite state space. The EM algorithm consists of two steps, called **E-step** and **M-step**, as follows:

E-step: Assume that $\beta^{(k)}$ and $\theta^{(k)}$ are current values, compute

$$E(T_1(X)|y,\beta^{(k)},\theta^{(k)}) = T_1{}^{(k)} \qquad (1.4.4)$$

and

$$E(T_2(X,y)|y,\beta^{(k)},\theta^{(k)}) = T_2{}^{(k)}, \qquad (1.4.5)$$

M-step: Find $\beta^{(k+1)}$ to maximize

$$\mu(\beta)'T_1{}^{(k)} - \log[A_1(\beta)] \qquad (1.4.6)$$

and $\theta^{(k+1)}$ to maximize

$$\eta(\theta)'T_2{}^{(k)} - \log[A_2(\theta)]. \qquad (1.4.7)$$

Both (1.4.6) and (1.4.7) are of the familiar form of the log-likelihood for maximum-likelihood estimation given data from a regular exponential family, in particular, for this problem, data from distribution (1.4.2) and from the conditional distribution (1.4.3) respectively. Furthermore, $T_1{}^{(k)}$ and $T_2{}^{(k)}$ are the expected values of the sufficient statistics computed from observed data, based on the (conditional) distributions. In many cases, (1.4.6) and (1.4.7) can be maximized either directly or by an iterative computational approach, such as the Newton-Raphson method. In the following chapters of this thesis, we will sometimes use the EM algorithm to maximize marginal or conditional marginal distributions of some random variables. Those random variables together with some other missing variables will satisfy the (conditional) distributional forms given in (1.4.2) and (1.4.3). Therefore, detailed descriptions of the EM algorithm are omitted at those corresponding places of the present thesis.

However, in some cases, especially in image analysis, the conditional distribution of X given Y=y is very complicated, so that the E-step is infeasible. Although the Monte Carlo method may be a useful approach to the computation of T_1 and T_2, it may still not be satisfied because of either the difficulty of generating samples from the corresponding conditional distribution or the heavy computational burden. Another difficulty involved in the EM algorithm is that it may not always be possible to maximize (1.4.6) and (1.4.7) due to the complexity of the log-likelihood functions of the corresponding processes.

In statistical image processing, x represents the original image, while y is the observed process. Little is known about the probabilistic or statistical properties of those prior distributions which are commonly used in the current literature, while the conditional distributions or the posteriors are equally or even more complicated. Both E-step and M-step are infeasible, except for some special cases. For instance, in chapter 2, we develop a recursive technique for a one-dimensional version of Markov random fields in order to carry out an iterative procedure for maximizing the log-likelihoood and to compute the corresponding conditional expectations. For continuous-valued images, the EM algorithm was used in the case of a prior Markov random field model with only one parameter by Geman and McClure(1987). Chalmond(1988) used the EM algorithm together with a Monte Carlo technique, known as the "Gibbs sampler " to maximize a pseudo-likelihood, his iterative approach was in fact made up of simultaneous parameter estimation and reconstruction. Besag(1986) also proposed an iterative procedure for simultaneous parameter estimation and image restoration, by using his coding technique(Besag, 1974, 1976) or its modification for parameter estimation based on some x obtained by a restoration technique. The EM algorithm was not used, however, so the estimated parameters may be unreliable when compared with the true values.

The EM algorithm increases the likelihood at each cycle, and the well-known concavity property of the log-likelihood for regular exponential families guarantees convergence. However, in some cases, the maximizing value lies on the boundary of the parameter space, or there is more than one maximizing point. Therefore, the EM algorithm is usually of local convergence, and the converged value depends upon the starting point. For more detail about the EM algorithm and its properties, see Dempster et al(1977) and Wu(1983).

1.5 Arrangement of the thesis

The aim of this project is to investigate classes of model based approaches to statistical image analysis, and to examine the three problems mentioncd in the first section of this chapter. We present brief summaries for each chapter as follows.

In Chapter 2, for the one-dimensional version of Markov random fields, (we shall refer to them as Gibbs chains), we develop a recursive technique which enables us to simulate them directly and to carry out maximum likelihood estimation of parameters. Usually, this iterative procedure is linearly convergent. The same technique also enables us to compute the conditional expectations associated with the EM algorithm, when only a partially observed process is available, with the result that both the E-step and the M-step can be carried out. Some asymptotic properties are also discussed.

In Chapter 3, we first examine a simple example of a Markov random field, namely, the natural extension of the Ising model to the multi-colour case. A matrix representation of the partition function is derived and, from it, the critical points of the model are found. For a more special case, the asymmetric two-colour Ising model with periodic boundary conditions, limiting distributional results are obtained for the sufficient statistics. Applications of these asymptotic properties are discussed. Then, for general Markov random fields, we develop line stochastic relaxation, using the recursion technique for the one-dimensional version of Markov random fields, in order to generate samples from the fields. Finally in this Chapter, we discuss maximum pseudo-likelihood parameter estimation for general Markov random fields. The principal tools are the conditional distributions of a block (usually a line) of pixels, given the values at their neighbouring pixels.

In Chapter 4, we still consider Markov random fields on two-dimensional rectangular lattices. It is shown that if the interaction parameter(s) between rows is zero or nearly equal to zero, certain statistics are asymptotically normal when the number of rows is large and that of columns is fixed. These results can be used to obtain an approximate likelihood function, for Markov random fields, from which one can estimate the underlying parameter(s). Other statistical applications of the results are also discussed.

In Chapter 5, we discuss the problem of parameter estimation for Markov random fields from noisy data. Since this is usually done together with image restoration, we first discuss existing restoration methods. A detailed discussion of Besag's ICM mothod(1986) in the auto-normal model case is provided. Then, we point out the difficulties of the EM algorithm. Again in the auto-normal case, we examine the difference between the EM algorithm and the iterative procedure proposed by Besag(1986) for simultaneous parameter estimation and image restoration. Finally, an iterative procedure of simultaneous parameter estimation and restoration is proposed. It is almost the same as that proposed by Besag(1986), except that a modified EM algorithm is used at each cycle of the iteration.

In Chapter 6, we discuss problems associated with Markov mesh models. For the three-dimensional case, we consider a third-order model, which is a natural generalization from two to three dimensions, and a generalization of Devijver's F-G-H algorithm for image restoration is developed. We then discuss the parameter estimation problem, in particular, for the two-dimensional case, from either the original image or a noisy version of the image. A modified EM algorithm, similar to that in Chapter 5, is examined.

In Chapter 7, we introduce a causal dependence model, namely, a multi-dimensional Markov chain model. It is a natural extension of the Markov chain, in which each row represents one point of a special Markov chain. By suitable choice of model parameters, textures can be simulated that are similar to those generated by Markov random fields, (see Cross and Jain(1983) for various textures simulated from Markov random fields), but the simulation procedure is computationally much more economical. The problem of parameter estimation is examined for the cases of non-noisy and noisy data. In the latter case, again using an idea similar to that in Chapter 5, procedures are developed for simultaneous parameter estimation and image restoration.

Finally, in Chapter 8, we present further general discussion about problems in statistical image analysis. Concluding remarks about the present thesis are given there.

In each of Chapters 2 to 7, illustrative examples or numerical results are provided at appropriate places.

Chapter 2

Parameter Estimation For Gibbs Chains And Hidden Gibbs Chains

2.1. Introduction

A stationary Markov chain $X=(X_1,X_2,...,X_N)'$, $X_i \in \{1,2,...,S\}$, is generally represented by its probabilistic functions,

$$P_i = \Pr(X_1=i) \qquad i=1,2,...S \tag{2.1.1}$$

$$p_{ij} = \Pr(X_{t+1}=j|X_t=i) \qquad i,j=1,2,..S. \tag{2.1.2}$$

Although this gives a very simple description of a Markov chain, it is not always convenient as a basis for statistical inference. In this chapter, we will discuss Gibbs chains, which follows a more general model than the Markov chain model above, but which satisfy Markov properties as follows:

$$\Pr(X_t|\{X_j:\ j \neq t\}) = \Pr(X_t|X_{t\pm1},X_{t\pm2},..,X_{t\pm r}) \tag{2.1.3}$$

$$\Pr(X_t|X_1,X_2,..X_{t-1}) = \Pr(X_t|X_{t-1},X_{t-2},..,X_{t-r}), \tag{2.1.4}$$

where r is called the order of the model. We will provide the definition of the Gibbs chain, which is in fact the one-dimensional version of the Markov random field or Gibbs field. It will also be shown that the Markov chain represented by (2.1.1) and (2.1.2) can be represented as a first-order Gibbs chain. It will be helpful to refer to the subscript of X_t as the time-point.

It often happens that the stochastic process itself cannot be observed directly. Instead, another process, either with continuous or discrete state space, is observed. This is known as a Hidden Markov model(HMM) and, in particular, as a partially observed Gibbs chain(POGC) model. We also refer to the observed data as the noisy process. The model finds application in various areas, including signal processing and medical statistics. Both the original chain model and the noise model are parameterized. In this chapter, some recurrence techniques are developed for carrying out both steps, especially the E-Step, of the EM algorithm.

Let the observed process be $Y = (Y_1,Y_2,...,Y_N)'$. We make a simple assumption for the conditional distribution $\Pr(Y|X,\theta)$, namely, that of conditional independence:

$$\Pr(Y|X,\theta) = \prod_{i=1}^{N} f_i(Y_i|X_i,\theta).$$

For the Markov chain represented by (2.1.1) and (2.1.2) with observed parallel discrete process Y, Baum and Petrie(1966) proved the consistency of maximum likelihood estimation(MLE) based on the likelihood function Pr(Y), and an iterative procedure for finding the MLE was derived(Baum and Eagon, 1967, Baum et al., 1970). The procedure is mainly based on forward and backward recursions. The algorithm was later generalized to the case with multi-dimensional noisy data at each time-point by Liporace(1982) and Rabiner et al (1985), and was used for the recognition of isolated word vocabularies(Rabiner et al, 1984). Baum's method can be shown to be an example of the EM algorithm. In this chapter, we develop corresponding forward and backward recursions in the context of Gibbs chains. The structure of the Gibbs chain treats the forward and backward directions symmetrically and this symmetry is repeated in the techniques of this chapter. Baum's technique, on the other hand, treated the two directions asymmetrically, in parallel with the familiar way of defining the Markov chain as in (2.1.1) and (2.1.2).

This chapter is arranged as follows. In Section 2.2, we give the definition of pairwise interaction Gibbs chains, then discuss the problem of parameter estimation and its asymptotic properties, and we show that the Markov chain, described by (2.1.1) and (2.1.2), is a first-order Gibbs chain. In Section 2.3, we will discuss the procedure for carrying out the EM algorithm in the case where only the noisy process is observed. In Section 2.4, some simulation results are presented. The main results of this chapter appear in Qian and Titterington(1990a).

2.2. Inference for Gibbs chains

Definition: A stochastic process $X=(X_1,X_2,...,X_N)'$ is called a pairwise interaction Gibbs chain(PIGC) or simply Gibbs chain if its probability function can be written

$$p(X=x|\beta) = \exp\{ \sum_{\mu=0}^{r} \sum_{i=1}^{N-\mu} G_{\mu i}(x_i,x_{i+\mu},\beta)\}/C(\beta) \qquad (2.2.1)$$

where $C(\beta)$ is a normalizing factor, β is the parameter, and r is called the order of the model. We rewrite $G_{0i}(x_i,x_i,\beta)$ as $g_i(x_i,\beta)$.

Property: If X is a PIGC, then

$$P(x_i|\{x_j:\ 1\leqslant j\leqslant N,\ j\neq i\}) = P(x_i|x_{i\pm1},x_{i\pm2},\dots x_{i\pm r}) \qquad (2.2.2)$$

Example: $$p(X=x|\beta) = \exp\{\beta \sum_{i=1}^{N-1} \delta(x_i,x_{i+1})\}/C(\beta) \qquad (2.2.3)$$

where $$\delta(s,t) = \begin{cases} 1 & s=t \\ 0 & s\neq t, \end{cases} \qquad (2.2.4)$$

If $\beta>0$, model (2.2.3) implies that neighbouring time-points tend to have the same value, and that all the states are treated equally.

Assume that X=x is observed. In order to obtain the maximum likelihood estimator of β from (2.2.1), we are required to deal with the normalizing constant $C(\beta)$. When β is one-dimensional, we can use a search method to maximize (2.2.1) if we know how to compute $C(\beta)$. On the other hand, if we assume that the exponential part of probability function (2.2.1) is linear in the parameter β, ie. $P(X|\beta)\propto\exp(\beta'Z(X))$, where $Z(X)$ is a vector with the same dimension as β, maximizing $P(X|\beta)$ is usually equivalent to solving the following equation,

$$lp(\beta) = Z(X) - \frac{1}{C(\beta)}\frac{d}{d\beta}C(\beta) = 0 \qquad (2.2.5)$$

where $d/d\beta$ denotes the gradient vector. This equation is also equivalent to

$$Z(X) - E_\beta Z(X) = 0. \qquad (2.2.6)$$

It is therefore very important to be able to compute $C(\beta)$ or to compute the expectation of the exponential part of $P(X|\beta)$. The following double-theorem solves this problem for the first-order and the second-order cases respectively. The results can be naturally extended to higher-order cases, although the computational burden increases with the order r. For clarity, explicit mention of β is largely omitted in the notation of the theorem.

Theorem 2.2.1(a) Let $X=(X_1,X_2,\dots,X_N)'$ be a first order Gibbs chain (r=1), with state space $\{1,2,\dots,S\}$, and probability function (2.2.1). Also let $a_i = [a_i(1),\dots a_i(S)]'$, and $b_i = [b_i(1),\dots b_i(S)]'$ be S-dimensional vectors, obtained by the following forward and backward recursions,

$$a_1(s) = \exp(g_1(s)) \qquad s=1,2,..,S$$

$$a_i(s) = (\sum_{t=1}^{S} a_{i-1}(t)\exp\{G_{1,i-1}(t,s)\})\exp\{g_i(s)\}$$

$$i=2,3,\ldots N, \quad s=1,2,\ldots,S$$

and

$$b_N(s) = \exp\{g_N(s)\} \qquad s=1,2,\ldots,S$$

$$b_i(s) = (\sum_{t=1}^{S} b_{i+1}(t)\exp\{G_{1,i}(s,t)\})\exp\{g_i(s)\}$$

$$i=N-1,N-2,\ldots,1, \quad s=1,2,\ldots,S.$$

Then,

1). $\Pr(X_{i+1}=t|X_i=s,X_{i-1},..X_1) = \Pr(X_{i+1}=t|X_i=s)$

$\propto b_{i+1}(t)\exp\{G_{1,i}(s,t)\}.$

2). $\Pr(X_i=s) = a_i(s)b_i(s)\exp\{-g_i(s)\}/C.$

3). $\Pr(X_i=s,X_{i+1}=t) = a_i(s)b_{i+1}(t)\exp\{G_{1,i}(s,t)\}/C.$

4). $C(\theta) = \sum_{s=1}^{S} a_i(s)b_i(s)\exp\{-g_i(s)\}$ for any i.

#

Theorem 2.2.1(b) For the second-order Gibbs chain, let $a_i=(a_i(s,t))$ and $b_i= (b_i(s,t))$, $s,t=1,2,\ldots S$, $i=1,2,\ldots N$, be S×S matrices defined by the following forward and backward recursions, respectively:

$$a_1 = \mathrm{diag}(\exp\{g_1(1)\},\ldots,\exp\{g_1(S)\})$$

$$a_i(s,t)=\exp\{g_i(t)+G_{1,i-1}(s,t)\} \sum_{\nu=1}^{S} a_{i-1}(\nu,s)\exp\{G_{2,i-2}(\nu,t)\},$$

$$s,t=1,2,\ldots S, \quad i=2,3,\ldots N,$$

and

$$b_N = \mathrm{diag}(\exp\{g_N(1)\},\ldots,\exp\{g_N(S)\})$$

$$b_i(s,t)=\exp\{(g_i(s)+G_{1,i}(s,t)\} \sum_{\nu=1}^{S} b_{i+1}(t,\nu)\exp\{G_{2,i}(s,\nu)\}$$

$$s,t=1.2.\ldots S, i=N-1,N-2,\ldots 1.$$

Then,

1) $\Pr(x_i|x_{i-1},x_{i-2},\ldots x_1) = \Pr(x_i|x_{i-1},x_{i-2})$

$\propto \exp\{G_{2,i-2}(x_{i-2},x_i)\}\cdot b_{i-1}(x_{i-1},x_i).$

2) $Pr(x_i,x_{i+1}) = a_{i+1}(x_i,x_{i+1})b_i(x_i,x_{i+1})\times$

$$\exp\{-g_i(x_i)-g_{i+1}(x_{i+1})-G_{1,i}(x_i,x_{i+1})\}/C.$$

3) $Pr(x_i,x_{i+1},x_{i+2}) = a_{i+1}(x_i,x_{i+1})b_{i+1}(x_{i+1},x_{i+2})\times$

$$\exp\{G_{2,i}(x_i,x_{i+2}) - g_{i+1}(x_{i+1})\}/C.$$

#

The proofs are straightforward. For instance, for Theorem 2.2.1, note that for any i and j with i+1<j,

$$Pr(x_{i+1},x_{i+2},\ldots,x_j) = \sum_{s=1}^{S} Pr(x_i=s,x_{i+1},\ldots x_j) \qquad (2.2.7)$$

we only need to show that

$$Pr(x_i,x_{i+1},\ldots,x_j) = a_i(x_i)b_j(x_j)\times$$

$$\exp\{\sum_{\nu=i-1}^{j-1} g_\nu(x_\nu) + \sum_{\nu=i}^{j-1} G_{1,\nu}(x_\nu,x_{\nu+1})\} \qquad (2.2.8)$$

Therefore, the detailed proofs are omitted. For numerical stability in practical calculation, we normalize a_i and b_i at each time point to prevent overflow. The theorems enable us to compute the expectation of $g_i(X_i)$, $G_{1,i}(X_i,X_{i+1})$, etc, and thereby, that of Z(X), so that we may solve equation (2.2.6) with the help of the following iteration,

$$\beta_{n+1} = \beta_n + M^{-1}[Z(X) - E(Z(X)|\beta_n)], \qquad (2.2.9)$$

where M is a positive definite matrix. This is not the exact Newton-Raphson method, but note that the derivative matrix of $lp(\beta)$ is equal to $-Var[Z(X)|\beta]$. Since this is usually a negative definite matrix, the iteration is therefore usually linearly convergent to a local maximum if M is large enough, in the sense of exceeding $Var[Z(X)|\beta]$ in terms of the Loewner ordering. In some very special cases, for the example provided above, $Var[Z(X)|\beta)$ can be calculated exactly, (see Chapter 4 for details), and we can then use the Newton-Raphson method, which is of second order of convergence.

Note that from 1) of the theorem, $Pr(X_i|X_\nu=x_\nu,\ \nu<i,\ \beta)$ can be calculated, so that the Gibbs chain can be generated from X_1 to X_N in one pass, if we first compute all the $b_i(s)$. Similarly, if we first compute all the $a_i(s)$, we can generate the Gibbs chain from X_N to X_1. For multi-dimensional Gibbs fields, the same results are

unlikely to obtain, but in the following chapters the recursion technique is used in several cases associated with multi-dimensional lattice systems.

We now concentrate on asymptotic properties. Write $Z(X)$ as $Z_N(X)$, and $\log C(\beta)$ as $lc_N(\beta)$. Assume that the domain of β is Ω, a connected open subset of a finite dimensional space, and that $\beta_0 \in \Omega$ is the true value of the parameter. Denote by $\hat{\beta}_N$ the maximum likelihood estimate of β_0. For the first-order case, note that the recurrence relationships for a_i and b_i in Theorem 2.2.1(a) can be written in the matrix form,

$$a_i = A_i B_{i-1}{}' a_{i-1} \qquad (2.2.10)$$

$$b_i = A_i B_i b_{i+1} \qquad (2.2.11)$$

where $A_i = \mathrm{diag}(\exp\{g_i(1)\},\ldots,\exp\{g_i(S)\})$ and $B_i = [\exp\{G_{1,i}(s,t)\}]_{S\times S}$.

We thus obtain the following formula for $C(\beta)$:

$$C(\beta) = L_S{}' A_1 B_1 A_2 \ldots B_{N-1} A_N L_S, \qquad (2.2.12)$$

where $L_S = (1,1,\ldots 1)'$.

<u>Corollary</u>: When $r=1$, if $g_i \equiv g$, $i=1,2,\ldots N$, and $G_{1,i} \equiv G$, $i=1,2,..N-1$, let α be the maximum eigenvalue of matrix $A^{½}BA^{½}$. (α is positive since all the elements of matrix $A^{½}BA^{½}$ are positive(Varga, 1962).) Then

$$N^{-1} lc_N(\beta) \longrightarrow \log\alpha, \quad \text{as } N \longrightarrow \infty.$$

#

We will not specify any particular form for the functions g_i and G_{ij}, nor any particular condition on them, in order to discuss the properties of $lc_N(\Theta)$ or its asymptotic behaviour. Instead, we provide the following lemma about the asymptotic properties of $Z_N(X)$, obtained under some assumptions about the asymptotic properties of $lc_N(\beta)$.

<u>Lemma</u> 2.2.1: Suppose that, for any $\beta \in \Omega$, as $N \longrightarrow \infty$,

$$N^{-1} lc_N(\beta) \longrightarrow \alpha_0(\beta) \qquad (2.2.13)$$

$$N^{½}[N^{-1}\frac{d}{d\beta} lc_N(\beta) - \alpha_1(\beta)] \longrightarrow 0 \qquad (2.2.14)$$

$$N^{-1}\frac{d^2}{d\beta^2} lc_N(\beta) \longrightarrow \alpha_2(\beta) \qquad (2.2.15)$$

where $\alpha_0(\beta)$, $\alpha_1(\beta)$ and $\alpha_2(\beta)$ are defined on Ω. Of these, $\alpha_1(\beta)$ is a vector with the same dimension as β and $\alpha_2(\beta)$ is a matrix. The convergence is assumed to be uniform on any compact subset of Ω. Then, as $N \longrightarrow \infty$,

$$N^{-1}Z_N(X) \xrightarrow{Pr} \alpha_1(\beta_0) \qquad (2.2.16)$$

$$N^{½}[N^{-1}Z_N(X) - \alpha_1(\beta_0)] \xrightarrow{D} N(0,\alpha_2(\beta_0)). \qquad (2.2.17)$$

Proof: Choose a positive number δ such that $D=\{\beta: |\beta-\beta_0|\leqslant\delta\}\subset\Omega$. For any vector λ with the same dimension as β, and for any scalar t, $\beta_0+N^{-1}t\lambda\in D$ and $\beta_0+N^{-½}t\lambda\in D$ when N is sufficiently large. Consider the random variables,

$$V = N^{-1}\lambda' Z_N(X); \qquad W = N^{-½}\lambda'[Z_N(X) - N\alpha_1(\beta_0)].$$

The moment-generating functions of V and W are

$$M_V(t|\beta_0,\lambda) = C_N(\beta_0+N^{-1}t\lambda)/C_N(\beta_0)$$

and

$$M_W(t|\beta_0,\lambda) = \exp\{-N^{½}t\lambda'\alpha_1(\beta_0)\}C_N(\beta_0+N^{-½}t\lambda)/C_N(\beta_0).$$

Thus, $\log M_V(t|\beta_0,\lambda) = N^{-1}\lambda'\frac{d}{d\beta}lc_N(\beta_0 + N^{-1}\tilde{t}_1\lambda)$,

and

$$\log M_W(t|\beta_0,\lambda)=N^{½}\lambda'[N^{-1}\frac{d}{d\beta}lc_N(\beta_0)-\alpha_1(\beta_0)]+(2N)^{-1}t^2\lambda'\frac{d^2}{d\beta^2}lc_N(\beta_0+N^{-½}t\tilde{\lambda}_2)\lambda$$

where $|\tilde{t}_1|, |\tilde{t}_2|\leqslant|t|$.

Then the uniform convergence assumed in (2.2.13)—(2.2.15) on the compact set D ensures that for any $t\in(-\infty,+\infty)$.

$$M_V(t|\beta_0,\lambda) \longrightarrow \exp\{t\lambda'\alpha_1(\beta_0)\}$$

and

$$M_W(t|\beta_0,\lambda) \longrightarrow \exp\{t^2\lambda'\alpha_2(\beta_0)\lambda/2\}$$

Since $\exp\{t^2\lambda'\alpha_2(\beta_0)\lambda/2\}$ is the generating function of the normal distribution with zero mean and variance $\lambda'\alpha_2(\beta_0)\lambda$, and since $\exp\{t\lambda'\alpha_1(\beta_0)\}$ is that of the degenerate distribution at $\lambda'\alpha_1(\beta_0)$. it follows that (Moran(1968))

$$V \xrightarrow{Pr} \lambda'\alpha_1(\beta_0)$$

$$W \xrightarrow{D} N(0, \lambda'\alpha_2(\beta_0)\lambda). \qquad \#$$

Theorem 2.2.2. Suppose that the conditions in the Lemma hold, and that

$$\alpha_1(\beta) = \frac{d}{d\beta}\alpha_0(\beta); \quad \alpha_2(\beta) = \frac{d}{d\beta}\alpha_1(\beta); \quad \alpha_2(\beta)>0,$$

Then, as $N \longrightarrow \infty$

1) $\hat{\beta}_N \xrightarrow{Pr} \beta_0$

2) $N^{1/2}(\hat{\beta}_N - \beta_0) \xrightarrow{D} N(0, \alpha_2(\beta_0)^{-1})$.

Proof: 1). Note that $h(\beta)=\beta'\alpha_1(\beta_0)-\alpha_0(\beta)$ is a concave function, and that β_0 is the maximum point. For any $\varepsilon>0$, define

$$D_1 = \{\beta;\ |\beta-\beta_0|=\varepsilon,\ \beta\in\Omega\}$$

$$\delta_1 = \inf\{h(\beta_0)-h(\beta),\ \beta\in D_1\}.$$

Then δ_1 is positive. Consider another concave function $h_N(\beta)=N^{-1}[\beta' Z_N(X)-lc_N(\beta)]$. $\hat{\beta}_N$ is the maximum point of h_N. The uniform convergence of $N^{-1}lc_N(\beta)$ assumed in the lemma on the compact subset D_1 of Ω therefore ensures that, for N sufficiently large, if

$$|N^{-1}Z_N(X) - \alpha_1(\beta_0)|<\delta,$$

then, $h_N(\beta) < h_N(\beta_0)$ for any $\beta\in D_1$.

Since h_N is concave with maximum point $\hat{\beta}_N$, $\hat{\beta}_N$ must lie in the region $\{\beta;\ |\beta-\beta_0|<\varepsilon\}$. That means that for any $\varepsilon>0$, there exists a $\delta>0$ such that, for N sufficiently large,

$$Pr(|\hat{\beta}_N-\beta_0|<\varepsilon) \geqslant Pr(|N^{-1}Z_N(X)-\alpha_1(\beta_0)|<\delta).$$

The result then follows from the Lemma.

2). By using the result in 1), the proof for 2) is standard.

#

The assumption in the lemma and the theorem is very natural'. For the particular case in the Corollary to Theorem 2.2.1. we know that $\alpha_0(\beta)$, $\alpha_1(\beta)$ and $\alpha_2(\beta)$ are usually the logarithm of the maximum eigenvalue of the matrix $A^{1/2}BA^{1/2}$, and its first-order and second-order derivatives, repectively. For the example represented by (2.2.3) and (2.2.4),

$$A_i = I_S, \quad B_i = (e^{\beta}-1)I_S + L_S L_S',$$

$$\alpha = e^{\beta} + S - 1, \quad \text{and} \quad C_N(\beta) = S(e^{\beta} + S - 1)^{N-1}.$$

Then

$$\alpha_0(\beta)=\log(e^{\beta}+S-1); \; \alpha_1(\beta)=e^{\beta}/(e^{\beta}+S-1); \; \alpha_2(\beta)=e^{\beta}(S-1)/(e^{\beta}+S-1)^2.$$

For the first-order stationary Markov chain represented by (2.1.1) and (2.1.2),

$$\Pr(X|P,p) = p_{x(1)} \prod_{s=1}^{S} \prod_{t=1}^{S} p_{st}{}^{V(X,s,t)}$$

where

$$V(X,s,t) = \sum_{i=2}^{N} \delta(x_{i-1},s)\delta(x_i,t)$$

and $\delta(s,t)=\delta_{st}$, the Kronecker delta function.

Its probability function can therefore be written in Gibbs-chain form. However, because of the constraints among the probabilistic parameters, it is difficult to make the exponential part linear in the parameter even if the model is re-parameterized, for example by $\gamma_{ij}=\log p_{ij}$, etc.

Finally, in this section, we mention two applications of this Gibbs chain model and the recursion technique described, that appear in the later parts of this thesis. One is in Chapter 5, for the two-dimensional Markov random field(MRF) (also see Qian and Titterington, 1989, 1990a), where the technique is used to obtain the relaxation method for simulating MRF and to develop the 'coding' method for estimating the parameters of MRF. Another is in Chapter 7, which describes a new texture model, based on a multi-dimensional Gibbs chain which was introduced and shown to be very useful in Qian and Titterington(1990c).

2.3 Parameter estimation for hidden Gibbs chains

In this section we discuss the problem of parameter estimation for the partially observed Gibbs chain(POGC). It was mentioned earlier that the Gibbs chain may itself be unobservable, but that, instead, an observed parallel noisy process Y=y may be available. Assume that the density of X is as in (2.2.1), and that, given the original chain, X, the noisy data y_i at different time-points are conditionally independent, each noise variable depending only on the original state at the same time-point. It is not necessary for each noise variable to have the same conditional distribution, but we assume that all such distributions can be written in exponential-family form. Thus

$$P(Y=y|X=x,\theta) = \exp\{\sum_{i=1}^{N} d_i(x_i,y_i,\theta)\} \tag{2.3.1}$$

where θ is an unknown parameter.

Lemma 2.3.1 Given Y=y, the conditional probability function of X is still of Gibbs-chain form with the same order as that of X:

$$\Pr(X=x|y,\theta,\beta)=\exp\{\sum_{i=1}^{N}\bar{g}(x_i|y_i,\theta,\beta)+\sum_{\mu=1}^{r}\sum_{i=1}^{N-\mu}G_{\mu i}(x_i,x_{i+1},\beta)\}/C(Y,\theta,\beta), \tag{2.3.2}$$

$$\Pr(X=x|y,\theta,\beta)=\exp\{\sum_{i=1} g(x_i|y_i,\theta,\beta)+\sum_{\mu=1}\sum_{i=1}G_{\mu i}(x_i,x_{i+1},\beta)\}/C(Y,\theta,\beta), \tag{2.3.2}$$

where

factor.

Note that, if $G_{\mu i}\equiv 0$, for $\mu \geqslant 1$, the y_i are independent mixture variables, in which case the EM algorithm can be used to obtain parameter estimates (Titterington et al, 1985, Titterington, 1989). Note that the E-step of the EM algorithm can be carried out with the help of the above Lemma and the recursion technique in the previous section. The EM algorithm for the POGC model can therefore be described as follows.

E-step: Assume θ_k and β_k are the current values for θ and β. Compute

$$E[\log\Pr(X,y|\theta,\beta)|y,\theta_k,\beta_k],$$

which is a function of θ and β. Since $\Pr(X,Y|\theta,\beta)=P(X|\beta)\Pr(Y|X,\theta)$, the conditional expectation can be separated into two parts, one of which is a function of β, namely, $Q_1(\beta)=E[\log P(X|\beta)|y,\theta_k,\beta_k]$, while the other is a function of θ, $Q_2(\theta)=E[\log P(y|X,\theta)|y,\theta_k,\beta_k]$.

M-step: Maximize the above two functions $Q_1(\beta)$ and $Q_2(\theta)$ to obtain the new values, θ_{k+1} and β_{k+1}, respectively. For the particular case where $P(X|\theta)\propto\exp\{\theta' Z(X)\}$, θ_{k+1} is obtained by maximizing $\theta' E[Z(X)|Y,\theta_k,\beta_k] - \log C(\theta)$ or, equivalently, by solving the equation

$$E[Z(X)|Y,\theta_k,\beta_k] - E[Z(X)|\beta] = 0$$

Therefore, as we mentioned in Chapter 1, the M-step is a procedure equivalent to maximizing a likelihood function. it is necessary to use the recursion technique described in the last section.

The ease or otherwise of computation in practice depends on the exact forms of $P(X|\beta)$ and $P(Y|X,\theta)$. In the M-step, the maximization procedure for $Q_1(\beta)$ is similar to the procedure described in Section 2.2 for estimating θ from X, which might itself require an iterative procedure. At each cycle of the EM algorithm, a reasonable, approximate procedure might be constructed for carrying out this iteration in the M-step.

Another problem is whether or not $Q_2(\theta)$ can be maximized. For regular exponential family, it is usually not difficult to maximize $Q_2(\theta)$. We have assumed conditional independence for $P(Y|X,\theta)$. For some special cases, for example, $Y=AX+\varepsilon$, where A is a band matrix, and ε is a multi-dimensional normal varible, it may be still possible to carry out the EM algorithm, except that the order of the conditional distribution $P(X|Y)$ is higher than that of the original chain.

2.4. Numerical results

We obtained simulation results for several cases. The original model for X was that of the Example in Section 2, involving only one parameter, β. For all the iterative procedures, we used

$$\omega = \|(\theta^{(k)},\beta^{(k)})'-(\theta^{(k+1)},\beta^{(k+1)})'\|_2^2$$

as the basis for a stopping rule: if $\omega<\omega_0$, the iteration stopped. $\|*\|_2$ denotes the Euclidean Norm, and ω_0 was pre-specified.

(1). In the first case, we took S=2 and S=3, and imposed Gaussian additive noise with variance σ^2. We took various values for N and σ^2, and, for each situation, we generated 50 samples. We obtained parameter estimates for the complete-data case and for the case where X is missing, and the results were summarized by the sample means and the sample variances. The results are presented in Table 2.1 for S=2 and Table 2.2 for S=3. The true β_0 is 1.5, and ω_0=0.000001. The starting values of β and σ^2 were both taken to be 1.0. COM denotes the complete data case, INC the case where X is missing, β_V denotes the asymptotic variance of the MLE of $\hat{\beta}_N$, namely, $\alpha_2(\beta_0)^{-1}/N$, and $\hat{\beta}_M$, $\hat{\sigma}^2{}_M$, $\hat{\beta}_V$ and $\hat{\sigma}^2{}_V$ denote the sample means and sample variances of β and σ^2, respectively.

Our experiment showed that convergence of the EM algorithm in this case was quite fast. For most of the situations described above, the iteration stopped within 20 cycles. The estimates for the INC cases were quite close to those for the COM cases, and the sample variances of β and σ^2 decreased as either $N \to \infty$ or $\sigma^2 \to 0$. Note that $\alpha_2(\beta_0)^{-1}$, which is the asymptotic variance of $N^{½}(\hat{\beta}_N - \beta_0)$, is now $(e^{1.5}+S-1)^2/[(S-1)e^{1.5}]$. Divided by the sample number N, they, ie, β_V in the tables, provide comparisons with the sample variances $\hat{\beta}_V$ at the COM case. They turn out to be quite similar.

2). Table 2.3 provides simulated results for the same original model with S=3, but where the imposed noisy chains were assumed to be discrete-valued with the same state space {1,2,3} as X, and

$$P(y_i=t|x_i=s) = \exp\{\theta\delta(s,t)\}/(2 + e^{\theta}),$$

involving only a single parameter, θ. The larger θ is, the higher is the probability with which y_i takes the same value as x_i. Note that the trend in the sample variances from Table 2.3 is similar, as θ increases, to that observed as σ^2 decreases in Table 2.1 and Table 2.2.

In the iterations for this model, we took ω_0 to be 0.00005, the true values β_0 and θ_0 were 1.0, and the starting values of θ and β were both taken to be 0.5. For each situation, we simulated 30 samples. We noticed that, for many situations, the estimates converged slowly. It is also possible that, when N is small and θ large, Y is almost the same as X, in which case the estimate of θ might be very large.

N	COM					INC			
	β_V	$\hat{\beta}_M$	$\hat{\beta}_V$	$\hat{\sigma}^2_M$	$\hat{\sigma}^2_V$	$\hat{\beta}_M$	$\hat{\beta}_V$	$\hat{\sigma}^2_M$	$\hat{\sigma}^2_V$
	(1). σ^2 = 0.36								
200	0.0335	1.5172	0.0392	0.3586	0.0015	1.5056	0.1731	0.3483	0.0015
400	0.0168	1.5286	0.0179	0.3604	0.0005	1.5693	0.0749	0.3610	0.0011
600	0.0112	1.5214	0.0081	0.3608	0.0003	1.5229	0.0389	0.3598	0.0006
	(2). σ^2 = 0.16								
200				0.1594	0.0003	1.5378	0.0750	0.1557	0.0003
400				0.1602	0.0001	1.5590	0.0750	0.1621	0.0002
600				0.1603	0.0001	1.5380	0.0177	0.1607	0.0001
	(3). σ^2 = 0.04								
200				0.0398	0.0000	1.5212	0.0420	0.0397	0.0000
400				0.0400	0.0000	1.5249	0.0186	0.0401	0.0000
600				0.0401	0.0000	1.5233	0.0095	0.0400	0.0000

Table 2.1. The results for the symmetric binary chain

N	COM					INC			
	β_V	$\hat{\beta}_M$	$\hat{\beta}_V$	$\hat{\sigma}^2{}_M$	$\hat{\sigma}^2{}_V$	$\hat{\beta}_M$	$\hat{\beta}_V$	$\hat{\sigma}^2{}_M$	$\hat{\sigma}^2{}_V$
	(1). $\sigma^2 = 0.36$								
200	0.0234	1.5118	0.0295	0.3586	0.0015	1.5362	0.0835	0.3498	0.0027
400	0.0117	1.5181	0.0083	0.3604	0.0005	1.5235	0.0382	0.3665	0.0014
600	0.0077	1.5113	0.0061	0.3608	0.0003	1.5276	0.0174	0.3622	0.0009
	(2). $\sigma^2 = 0.16$								
200				0.1594	0.0003	1.5362	0.0501	0.1595	0.0005
400				0.1602	0.0001	1.5256	0.0179	0.1638	0.0002
600				0.1603	0.0001	1.5218	0.0101	0.1614	0.0001
	(3). $\sigma^2 = 0.04$								
200				0.0398	0.0000	1.5062	0.0289	0.0395	0.0000
400				0.0400	0.0000	1.5110	0.0088	0.0396	0.0000
600				0.0401	0.0000	1.5105	0.0062	0.0401	0.0000

Table 2.2. The results for S=3.

N	COM					INC			
	β_V	$\hat{\beta}_M$	$\hat{\beta}_V$	$\hat{\theta}_M$	$\hat{\theta}_V$	$\hat{\beta}_M$	$\hat{\beta}_V$	$\hat{\theta}_M$	$\hat{\theta}_V$
	(1). $\theta = 0.9$								
300	0.0137	0.9963	0.0174	0.9021	0.0129	0.7435	0.1679	0.7923	0.1956
600	0.0068	1.0196	0.0052	0.9115	0.0074	0.6610	0.0422	0.7628	0.1147
900	0.0046	0.9859	0.0041	0.9022	0.0023	0.8372	0.1163	0.8824	0.0838
	(2). $\theta = 1.2$								
300				1.2138	0.0139	0.9004	0.1403	1.1453	0.2041
600				1.2033	0.0078	0.9991	0.0894	1.1512	0.0597
900				1.2024	0.0021	1.0083	0.0265	1.2249	0.0367
	(3). $\theta = 1.5$								
300				1.5185	0.0170	0.9945	0.1099	1.5091	0.2329
600				1.5118	0.0068	1.1313	0.0566	1.4690	0.0566
900				1.5021	0.0037	1.1453	0.0439	1.4265	0.0367

Table 2.3. The results with discrete noise

Chapter 3

Inference For Markov Random Field Models

3.1 Introduction

Statistical inference for spatial models was originally motivated by geographical and ecological data, but was severely hampered by the dearth of realistic parametric alternatives to spatial independence. Markov random fields have clearly filled this gap. Markov random fields, and the Ising models, in particular, have been studied extensively during the last 50 years in statistical physics, where they are known as Gibbs ensembles(Pickard, 1987). In the discrete case, Gibbs ensembles are known as Ising models, which are important in theoretical physics because they exhibit phase transitions, that is, as parameter values increase past critical values, abrupt changes occur in qualitative behaviour. In particular, pixel variables lose their asymptotic long-range-independence, and the simulated samples are then very likely to be almost one colour. Another major difficulty in statistical inference for Markov random fields is that their apparently simple likelihood functions are in fact surprisingly intractable, even for very simple cases.

The difficulties involved with discrete Markov random fields are due to the intractablity of the partition functions. The properties of a MRF are determined by the behaviour of its partition function, which is analytic in its parameters. However, it has proved surprisingly difficult to determine the partition function asymptotically in order to explain the asymptotic properties of the model, even in simple cases. Kramers and Wannier(1941) introduced a matrix method for determining partition functions (see also Newell and Mantroll, 1953). The method is quite similar to what we used for the one-dimensional case in the last chapter. Onsager(1944) used this method, obtaining a matrix representation of the partition function for the classic Ising model with only two colours, with a first-order neighbourhood system and with periodic boundary conditions. Kaufmann(1949) obtained a direct product decomposition for this matrix expression. Kaufmann and Onsager(1949) determined, approximately, the correlation structure for the same model. Baxter(1972) obtained the partition function for the eight-vertex lattice model. Recently, Pickard(1976, 1987) used Kaufmann(1949)'s decomposition results, obtaining limit theorems for the sample correlation between nearest neighbours for the symmetric two-colour Ising model with periodic boundary conditions. These results provide a basis for asymptotic testing and estimation.

The partition function has also been expressed as a determinant or as a Pfaffian; see Pickard(1977), where he used this expression to examine the asymptotic properties of a two-colour Ising model in which two parameters were involved.

For the two-colour Ising model with only one parameter, Kramers and Wannier(1941) deduced an inversion transformation under which the partition function is "invariant" when the parameter is transformed from a low to a high value. The important property of this transformation is that its fixed point determines the transition point of the lattice. It was also later found by Kaufmann(1949), by examining the exact matrix representation of the partition function, that this fixed point is just the critical point. He also located the critical points for the two-parameter case. Potts(1952) used a method similar to that of Kramers and Wannier(1941) to develop the inversion transformation for multi-colour models and obtained the fixed points, which should also be the critical points. His work was for the symmetric case where only one parameter is involved. In this chapter, we shall obtain the critical points for multi-colour model with two parameters and first-order neighbourhood system.

For parameter estimation for MRFs, there are four sorts of approach. The first is the method of **moment estimation**. However, the behaviours of moments are very difficult to determine, even asymptotically. This method can only be applied to very special cases such as the two-colour Ising model for which some asymptotic properties have been derived. The second is **maximum asymptotic likelihood estimation**. Similarly to the first method, it depends on the asymptotic behaviour of the likelihoood function or the partition function. For the two-colour Ising model, results in Pickard(1976, 1977, 1986) can be used. We shall also examine the two-colour Ising model with periodic boundary conditions in this chapter. The third method is **maximum pseudo-likelihood estimation**. This was first introduced by Besag(1974) in the form of the coding technique, using the local dependence of the fields and a set of conditional distributions for single pixels. Geman and Graffigne(1987) proved the consistency of the method. The fourth is **maximum approximate-likelihood estimation**. We shall concentrate on this approach in the next chapter.

This chapter is arranged as follows. In section 3.2, we provide matrix representation of partition function for general pairwise interaction Markov random fields with a first-order neighbourhood system, and we then discuss the partition function of a special multi-colour Ising model which is a generalization of the two-colour Ising model. Critical points are found. In section 3.3, we consider the asymmetric two-colour Ising model with periodic boundary conditions, in otherwords, the field is wrapped around a torus. Limiting distributions for sample correlations between nearest neighbours are obtained. Applications of these results to parameter estimation and the problem of testing for symmetry are discussed. In section 3.4, we discuss the problem of simulating Markov random fields. Following on from Geman and Geman's(1984) stochastic relaxation method, we examine a line-relaxation method. In section 3.5, we develop pseudo-likelihood parameter estimation, by using the distribution of a line of pixels, conditionally upon its neighbour lines. Some simulation results are presented in section 3.2, section 3.4 and section 3.5. More discussion is provided in section 3.6.

3.2 Partition functions and critical points

3.2.1 Some matrix notation

We first introduce some notation and provide some related properties. Let $A=(a_{ij})$ be an $M\times p$ complex matrix, and $B=(b_{ij})$ an $N\times q$ complex matrix. Then the Kronecker product of A and B is defined by

$$A\otimes B = (a_{ij}B)$$

which is an MN×pq matrix. The following properties are standard results related to this product-operation.

Proposition 3.2.1

(1). $(A\otimes B)\otimes C = A\otimes(B\otimes C)$

(2). $(A + B)\otimes C = A\otimes C + B\otimes C$

(3). $A\otimes(B + C) = A\otimes B + A\otimes C$

(4). If A, B, C and D are appropriately compatible, then

$$(AB)\otimes(CD) = (A\otimes C)(B\otimes D)$$

(5). $e^{A}\otimes I = e^{A\otimes I}$, $I\otimes e^{A} = e^{I\otimes A}$, where I is an identity matrix.

(6). $(A\otimes B)^{*} = A^{*}\otimes B^{*}$, where * means the corresponding conjugate transpose matrix. □

If A_i, i=1,2,..T are T matrices, let

$$\bigotimes_{t=1}^{T} A_t = A_1\otimes A_2\otimes\ldots\otimes A_T .$$

Let $I_{(S,i)}$ denote the identity matrix of dimension S^i, $L_{(S,i)}$ denote S^i dimensional vector with all elements being unity, $I_S=I_{(S,1)}$ and $L_S=L_{(S,1)}$. Then

$$I_{(S,i+j)} = I_{(S,i)}\otimes I_{(S,j)} = \bigotimes_{t=1}^{i+j} I_S$$

$$L_{(S,i+j)} = L_{(S,i)}\otimes L_{(S,j)} = \bigotimes_{t=1}^{i+j} L_S.$$

3.2.2. Partition functions in general pairwise-interaction cases

For simplicity of notation, we only consider homogenous MRFs; that is, the interaction is assumed invariant over the entire lattice.

Consider a Markov random field $X=(X_{ij})$ over an M×N rectangular lattice, where each $X_{ij}\in\{1,2,\ldots S\}$, with distribution

$$P(X=x|\beta) = \frac{1}{C(\beta)}\exp\{\sum_{i=1}^{M}\sum_{j=1}^{N} g(x_{ij},\beta) + Z_1(x,\beta) + Z_2(x,\beta)\} \qquad (3.2.1)$$

where

$$Z_1(x,\beta) = \sum_{i=1}^{M}\sum_{j=1}^{N-1} G_1(x_{ij},x_{i,j+1},\beta)$$

and

$$Z_2(x,\beta) = \sum_{i=1}^{M-1}\sum_{j=1}^{N} G_2(x_{ij},x_{i+1,j},\beta).$$

The two terms Z_1 and Z_2 are associated with interactions along row directions and column directions respectively. The normalizing constant $C(\beta)$, known as the **partition function**, is given by

$$C(\beta) = \sum_{x} \exp(\sum_{i=1}^{M} \sum_{j=1}^{N} g(x_{ij},\beta) + Z_1(x,\beta) + Z_2(x,\beta)), \qquad (3.2.2)$$

where the summation is over all possible x, so that there are altogether S^{MN} terms. It is therefore impractical to compute $C(\beta)$ by using (3.2.2) except when both M and N are very small. The following matrix method is almost the same as that used earlier for the one-dimensional version of MRF, but the detailed procedure was omitted in the previous chapter. In fact, if we consider one row as a "point", (3.2.1) can be regarded as a first-order stationary Markov chain with S^N states in the corresponding state space. Therefore, from (2.2.12), $C(\beta)$ can be written as

$$C(\beta) = L_{(S,N)}{}'ABA....BAL_{(S,N)} \qquad (3.2.3)$$

where A and B are $S^N \times S^N$ matrices and A is diagonal. However, it is not easy to find A and B exactly. Before examining this problem, we introduce some more notation. Let

$$X_{(ij)} =\{x_{st},\ s>i,\ \text{or}\ s=i,t>j\}$$

$$a_{\nu} = \exp\{g(\nu,\beta)\}$$

$$b_{\nu\mu}{}^{(1)} = \exp\{G_1(\nu,\mu)\}$$

$$b_{\nu\mu}{}^{(2)} = \exp\{G_2(\nu,\mu)\}$$

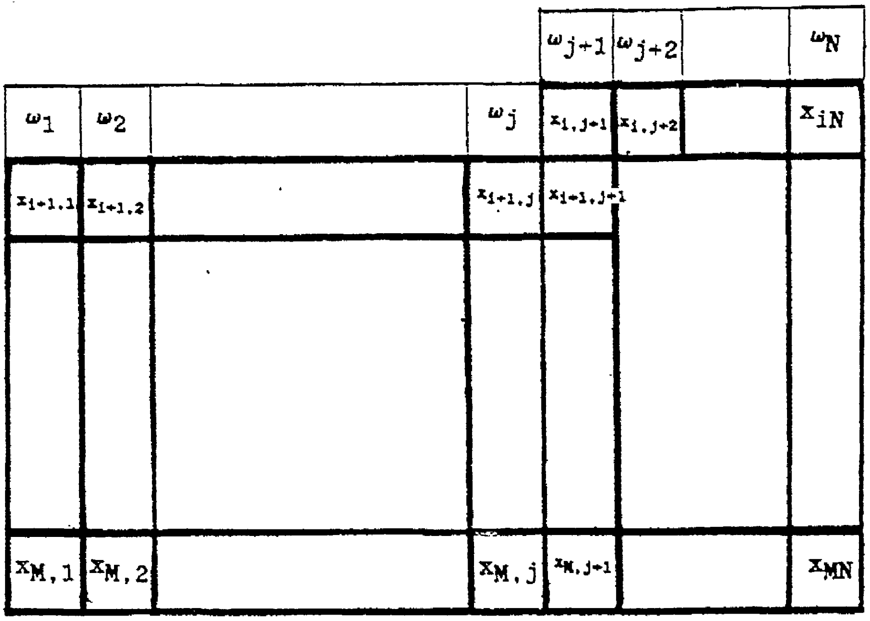


Fig 3.1 $X_{(ij)}$ and notation relevant for inference about $C(\beta)$

$X_{(ij)}$ is shown in Fig 3.1. It is a vector of variables over all pixels coming after pixel (ij), if we order the pixels one by one and row by row. Also, reference to Fig 3.1 makes it easy to imagine that, when $j<N$, $\Pr(X_{(ij)}|\beta)$ can be written in the following form:

$$\Pr(x_{(ij)}|\beta) = \sum_{\omega_1=1}^{S} \sum_{\omega_2=1}^{S} \ldots \sum_{\omega_N=1}^{S} d_{ij}(\omega_1,\omega_2,\ldots\omega_N)\exp\{W_{ij}^{(1)}+W_{ij}^{(2)}\}/C(\beta),$$

where

$$W_{ij}^{(1)} = g(x_{i,j+1}) + G_1(\omega_j,x_{i,j+1}) + G_1(x_{i,j+1},x_{i,j+2})$$
$$+ G_2(\omega_{j+1},x_{i,j+1}) + G_2(x_{i,j+1},x_{i\div 1,j+1})$$

and

$$W_{ij}^{(2)} = \sum_{\nu=j+2}^{N} g(x_{i\nu}) + \sum_{\iota=i+1}^{M} \sum_{\nu=1}^{N} g(x_{\iota\nu})$$
$$+ \sum_{\nu=j+2}^{N} G_2(\omega_\nu,x_{i\nu}) + \sum_{\nu=j+2}^{N-1} G_1(x_{i\nu},x_{i,\nu+1})$$
$$+ \sum_{\nu=1}^{j} G_2(\omega_\nu,x_{i+1,\nu}) + \sum_{\nu=j+2}^{N} G_2(x_{i\nu},x_{i+1,\nu})$$
$$+ \sum_{\iota=i+1}^{M} \sum_{\nu=1}^{N-1} G_1(x_{\iota\nu},x_{\iota,\nu+1}) + \sum_{\iota=i+1}^{M-1} \sum_{\nu=1}^{N} G_2(x_{\iota\nu},x_{\iota+1,\nu}).$$

Note that $W_{ij}^{(1)}$ is a function associated with $x_{i,j+1}$ and its four neighbour states, ω_j, ω_{j+1}, $x_{i,j+2}$ and $x_{i+1,j+1}$, while $W_{ij}^{(2)}$ is a summation of those potential functions, namely, g, G_1 and G_2, that are associated with all ω_ν and $x_{(ij)}$ except those associated with $x_{i,j+1}$.

When j=N, $W_{iN}^{(1)}$ is a sum of those potential functions associated with $x_{i+1,1}$, so it contains only four terms, while $W_{ij}^{(2)}$ shows little change as well. Note that, since

$$\Pr(x_{(i,j+1)}|\beta) = \sum_{x_{i,j+1}=1}^{S} \Pr(x_{(ij)}|\beta),$$

we can then have

$$d_{i,j+1}(\omega_1,\ldots\omega_j,t,\omega_{j+2},\ldots\omega_N) =$$

$$a_t b_{\omega_j t}{}^{(1)} \sum_{s=1}^{S} b_{st}{}^{(2)} d_{ij}(\omega_1,..\omega_j,s,\omega_{j+2},\dots,\omega_N)$$

for $1 \leqslant j \leqslant N-1$,

and

$$d_{i+1,1}(t,\omega_2,\dots\dots,\omega_N) = a_t \sum_{s=1}^{S} b_{st}{}^{(2)} d_{iN}(s,\omega_2,\dots,\omega_N).$$

Denote $d_{ij}{}'$ by

$$\begin{aligned} d_{ij}{}' = (&d_{ij}(1,1,..,1),d_{ij}(1,1,\dots,2),\dots\dots,d_{ij}(1,1,\dots.S),\\ &d_{ij}(1,..,2,1),d_{ij}(1,\dots,2,2),\dots\dots,d_{ij}(1,\dots,2,S),\\ &\dots\dots\dots\dots\dots\\ &d_{ij}(S,S,..,1),d_{ij}(S,S,\dots,2),\dots\dots d_{ij}(S,S,\dots.S)) \end{aligned}$$

which is an S^N-dimensional vector. We therefore have

$$d_{i,j+1} = I_{(j-1)}\otimes B_1 \otimes I_{(N-j-1)} \cdot I_{(j)}\otimes(AB_2)\otimes I_{(N-j-1)} d_{ij}$$

and

$$d_{i+1,1} = (AB_2)\otimes I_{(N-1)} d_{iN} ,$$

where, for simplicity, $I_{(\nu)}$ denotes $I_{(S,\nu)}$, A is an SxS diagonal matrix with a_ν as diagonal entries, B_1 is an $S^2 \times S^2$ diagonal matrix with $[(\nu-1)S+\iota]$-th diagonal element $b_{\nu\iota}{}^{(1)}$, and B_2 is an SxS matrix with (ν,ι) element $b_{\nu\iota}{}^{(2)}$. Note that

$$I_{(j)}\otimes(AB_2)\otimes I_{(N-j-1)} = I_{(j)}\otimes A \otimes I_{(N-j-1)} \cdot I_{(j)}\otimes B_2 \otimes I_{(N-j-1)}.$$

There are therefore three sorts of matrices, namely, $I_{(j)}\otimes A\otimes I_{(N-j-1)}$, $I_{(j-1)}\otimes B_1 \otimes I_{(N-j-1)}$ and $I_{(j)}\otimes B_2\otimes I_{(N-j-1)}$. The first two are diagonal, and thereby commutative. Furthermore, when $j>i$,

$$\begin{aligned} &I_{(j)}\otimes B_2\otimes I_{(N-j-1)} \cdot I_{(i)}\otimes A \otimes I_{(N-i-1)} \\ &\quad = I_{(i)}\otimes A\otimes I_{(N-i-1)} \cdot I_{(j)}\otimes B_2 \otimes I_{(N-j-1)} \end{aligned}$$

and

$$\begin{aligned} &I_{(j)}\otimes B_2\otimes I_{(N-j-1)} \cdot I_{(i-1)}\otimes B_1 \otimes I_{(N-i-1)} \\ &\quad = I_{(i-1)}\otimes B_1\otimes I_{(N-i-1)} \cdot I_{(j)}\otimes B_2 \otimes I_{(N-j-1)}, \end{aligned}$$

which means that when $j>i$, the last sort of matrix is commutative with the first two. We thus obtain

$$d_{iN} = D_1 D_2 D_3 d_{i-1,N}$$

where D_1, D_2 and D_3 are $S^N \times S^N$ matrices which can be written in following forms:

$$D_1 = \prod_{j=1}^{N} I_{(j-1)} \otimes A \otimes I_{(N-j)} = \bigotimes_{j=1}^{N} A \qquad (3.2.4)$$

$$D_2 = \prod_{j=1}^{N-1} I_{(j-1)} \otimes B_1 \otimes I_{(N-j-1)} \qquad (3.2.5)$$

$$D_3 = \prod_{j=1}^{N} I_{(j-1)} \otimes B_2 \otimes I_{(N-j)} = \bigotimes_{j=1}^{N} B_2 \qquad (3.2.6)$$

Since there is no effect associated with B_2 in the first row, we find that

$$d_{1,N} = D_1 D_2 L_{(N)},$$

where $L_{(N)} = L_{(S,N)}$.

Note that since $C(\beta) = L_{(N)}{}' d_{MN}$, we obtain the following formula for $C(\beta)$:

$$C(\beta) = L_{(N)}{}' D_1 D_2 D_3 D_1 D_2 \ldots\ldots D_1 D_2 D_3 D_1 D_2 L_{(N)}. , \qquad (3.2.7)$$

3.2.3 Partition function in a special case

We consider now a special multi-colour Ising model, which is a direct generalization of the two-colour Ising model. The model involves two parameters, and the function g in the model is zero, so that D_1 is an identity matrix. The distribution of the model is given by

$$\Pr(X=x|\alpha,\beta) = \exp\{\alpha Z_1(x) + \beta Z_2(x)\}/C(\alpha,\beta) \qquad (3.2.8)$$

where $Z_1(x) = \sum_{i=1}^{M} \sum_{j=1}^{N-1} \delta(x_{ij}, x_{i,j+1})$

and $Z_2(x) = \sum_{i=1}^{M-1} \sum_{j=1}^{N} \delta(x_{ij}, x_{i+1,j})$.

For the two-colour Ising model, Onsager(1944) expressed D_2 and D_3 in exponential form. The function δ in his model is slightly different from δ here, in that $\delta(s,t)=-1$, when $s \neq t$. D_2 is a diagonal matrix, and therefore, not difficult to express in exponential form, while D_3 is slightly more complicated. Our aim is to express D_2 and D_3 in a form which is similar to that for the two-colour case and which is associated with some basic matrices in $S^N \times S^N$ matrix space, in such a way that D_2 and D_3 are related in a very specific way. This relationship can be very important so far as the properties of the model are concerned. This more explicit representation could also be important for examining the asymptotic behaviour of the model, as has been done for the two-colour case. Note that D_2 and D_3 are only related to B_1 and B_2 respectively, of which, B_1 is diagonal. We then re-write B_1 and B_2 as follows.

$$\begin{aligned} B_1 &= \text{diag}(\underbrace{e^{\alpha},1,\ldots,1}_{S},\underbrace{1,e^{\alpha},\ldots,1}_{S},\underbrace{\ldots\ldots}_{\ldots\ldots},\underbrace{1,1,\ldots,1,e^{\alpha}}_{S}) \\ &= \exp\{\alpha\cdot\text{diag}(\underbrace{1,0,\ldots0}_{S},\underbrace{0,1,\ldots.0}_{S},\underbrace{\ldots\ldots}_{\ldots\ldots},\underbrace{0,0,\ldots0,1}_{S})\} \end{aligned} \tag{3.2.9}$$

$$B_2 = (e^{\beta}-1)I_{(1)} + L_{(1)}L_{(1)}'. \tag{3.2.10}$$

Now we introduce some basic matrices in $S \times S$ matrix space and present some properties of them. Let ψ be a non-trivial S-th root of unity. where non-trivial means $\psi^i \neq 1$, when $i<S$, and $\psi^S=1$. ψ is a complex number. One choice for ψ is $\exp\{2\pi i/S\}$. Define

$$u = \text{diag}\{1,\psi,\psi^2,\ldots,\psi^{S-1}\}$$

$$v = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

and $u_i = u^i$ $\qquad 1 \leqslant i \leqslant S-1$

$v_i = v^i$ $\qquad 1 \leqslant i \leqslant S-1.$

Proposition 3.2.2:

(1). For any $1 \leqslant i \leqslant S-1$, $v_i{}^S = u_i{}^S = I$

(2). If $V = v_1 + v_2 + \ldots + v_{S-1}$, then

$$V^2 = (S-2)V + (S-1)I.$$

The proof is trivial. Property (2) also implies that

$$[\frac{2}{S}V - \frac{S-2}{S} I]^2 = I.$$

Therefore, we have

$$\exp\{\gamma[2V-(S-2)I]/S\} = ½(e^{\gamma} + e^{-\gamma})I + ½(e^{\gamma} - e^{-\gamma})[2V-(S-2)I]/S$$

$$= [\frac{1}{S}e^{\gamma} + \frac{S-1}{S}e^{-\gamma}]I + \frac{1}{S}(e^{\gamma} - e^{-\gamma})V$$

Note that

$$V_2 = \begin{bmatrix} 0 & 0 & 0 \ldots\ldots 0 & 1 & 0 \\ 0 & 0 & 0 \ldots\ldots 0 & 0 & 1 \\ 1 & 0 & 0 \ldots\ldots 0 & 0 & 0 \\ \ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots \\ 0 & \ldots 1 & 0 \quad 0 & 0 & 0 \\ 0 & \ldots 0 & 1 \quad 0 & 0 & 0 \\ 0 \ldots & \ldots 0 & 0 \quad 1 & 0 & 0 \end{bmatrix}$$

....

$$V_{S-1} = \begin{bmatrix} 0 & 1 & 0 & \ldots\ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

We therefore have that $V=L\cdot L' - I$, implying that all the non-diagonal elements of V are unity, while all diagonal elements are zero. Now define β^*, which is a function of β, by the following equation:

$$\frac{e^{\beta^*/2} + (S-1)e^{-\beta^*/2}}{e^{\beta^*/2} - e^{\beta^*/2}} = e^{\beta}, \qquad (3.2.11)$$

otherwise written as

$$\beta^* = \log[(S-1 + e^{\beta})/(e^{\beta} - 1)]. \qquad (3.2.12)$$

We can then show that

$$B_2 = \frac{S}{2\cosh(½\beta^*)}\exp\{½\beta^* W\} = \frac{\exp\{-½(S-2)\beta^*/S\}}{2\cosh(½\beta^*)}\exp\{\frac{1}{S}\beta^* V\}$$

$$= d(\beta)\exp\{\beta^* V/S\}, \qquad (3.2.13)$$

where $d(\beta) = [e^{\beta}-1]^{1-1/S}[S-1+e^{\beta}]^{1/S}/S$.

There is a symmetry with the transformation: $\beta \to \beta^*$, in that $\beta = (\beta^*)^*$. We now examine B_1. Consider the following $S^2 \times S^2$ matrix:

$$U = u_1 \otimes u_1^* + u_2 \otimes u_2^* + \ldots + u_{S-1} \otimes u_{S-1}^*,$$

where u_i^* is the conjugate transpose matrix of u_i. The $[(i-1)S+j]$-th diagonal element of U is

$$\sum_{k=1}^{S-1} (\psi^k)^{i-1} (\bar{\psi}^k)^{j-1} = \sum_{k=1}^{S-1} (\psi^{i-j})^k$$

$$= \begin{cases} S-1 & i=j \\ -1 & i \neq j \end{cases}$$

We therefore obtain the following formula for B_1:

$$B_1 = \exp\{\alpha(U + I)/S\} = \exp\{\alpha/S\}\exp\{\alpha U/S\} \qquad (3.2.14)$$

We have now obtained a quite explicit expression for B_1 and B_2, and we can then write D_2 and D_3 as follows:

$$D_2 = \exp\{\frac{(N-1)\alpha}{S}\}\exp\{\frac{\alpha}{S} \sum_{j=1}^{N-1} \sum_{\nu=1}^{S-1} U_{j\nu} U_{j+1,\nu}^*\} \text{,} \qquad (3.2.15)$$

$$D_3 = [d(\beta)]^N \exp\{\frac{\beta^*}{S} \sum_{j=1}^{N} \sum_{\nu=1}^{S-1} V_{j\nu}\}, \qquad (3.2.16)$$

where

$$U_{j\nu} = I_{(j-1)} \otimes u_\nu \otimes I_{(N-j)} \qquad (3.2.17)$$

$$V_{j\nu} = I_{(j-1)} \otimes v_\nu \otimes I_{(N-j)}. \qquad (3.2.18)$$

Consider now the case with periodic boundary conditions, which is equivalent to wrapping the lattice around a torus. To be precise, consider the following distribution function:

$$P(X=x|\alpha,\beta) = \frac{1}{C(\alpha,\beta)}\exp\{ \sum_{i=1}^{M} \sum_{j=1}^{N} \alpha\delta(x_{ij}, x_{i,j+1}) + \beta\delta(x_{ij}, x_{i+1,j})\},$$

$$(3.2.19)$$

where $x_{i,N+1} = x_{i,1}$, $1 \leqslant i \leqslant M$ and $x_{M+1,j} = x_{1,j}$, $1 \leqslant j \leqslant N$. Then there are interactions between the first row and the last row and between the first column and the last column. In fact, the interaction matrix between rows is still the same as D_3, but the intra-interaction within

within one row changes slightly, due to the interaction between the first pixel and the last pixel. Noting the equivalence of each pixel in one row, we have

$$D_2 = \exp\{\frac{(N-1)\alpha}{S}\}\exp\{\frac{\alpha}{S}\sum_{j=1}^{N}\sum_{\nu=1}^{S-1} U_{j\nu}U_{j+1,\nu}{}^{*}\} \qquad (3.2.20)$$

where $U_{N+1,\nu}=U_{1,\nu}$. Again, because of the interaction between the first row and the last row, expression (3.2.7) for the partition function changes to the following form:

$$C(\alpha,\beta) = \text{trace}[D_2D_3D_2\ldots D_3] = \text{trace}[D_2D_3]^M \qquad (3.2.21)$$

$$= \sum_{\lambda} \lambda^M, \qquad (3.2.22)$$

where λ are the eigenvalues of matrix D_2D_3. Therefore, when $M\to\infty$, $C(\alpha,\beta)$ approximately depends only on the largest eigenvalues of D_2D_3.

3.2.4 Critical points

For the multi-colour Ising model discussed above, we have obtained two matrices, D_2 and D_3, and the representation (3.2.21) and (3.2.22) for the partition function $C(\alpha,\beta)$. There is a similarity between D_2 and D_3. For the two-colour case, both Onsager(1944) and Kaufmann(1949) found somewhat different transformations which interchange D_2 and D_3, and therefore enable us to locate the critical points. We now consider the similarity relationship between D_2 and D_3 in the multi-colour case.

We have defined two basic matrices, u and v, which satisfy the following commutative law:

$$uv = \gamma vu \qquad (3.2.23)$$

Note that the eigenvalues of u and v consist of 1, γ, γ^2, $\ldots,\gamma^{S-1}$, so u and v are **unitary equivalent** to each other. Potts(1952) considered the 2×∞ lattice, but in fact he only examined the interchange between $\exp\{\beta^*(v_1 + v_2\ldots+ v_{S-1}\}$ and $\exp\{\beta(u_1 + u_2 +\ldots u_{S-1}\}$, then located the fixed transtion point defined by $\beta^*=\beta$. For the M×N lattice, the similarity between D_2 and D_3 is more complicated.

Since a matrix is equivalent to a linear operator, we consider S^N dimensional linear complex space with the operator:

$$D = V_{1,1}V_{2,1}\cdots V_{N,1}. \qquad (3.2.24)$$

Similarly to $V_{i,\upsilon}$, the eigenvalues of D consist of 1, ψ, ψ^2 $\ldots,\psi^{S-1}$, each of them repeated $S^{(N-1)}$ times. Therefore there are S subspaces of S^N-dimensional space, denoted by Ω_i respectively, where Ω_i is an S^{N-1}-dimensional linear space, satisfying

$$D\xi = \psi^{i-1}\xi \qquad \xi\in\Omega_i. \qquad (3.2.25)$$

Note that $V_{j,1}$ is commutative with D. From the commutative rule in (3.2.23), we can also show that $U_{j,1}U_{j+1,1}{}^*$ commutes with D. Therefore, all $V_{j,1}$ and $U_{j,1}U_{j+1,1}{}^*$ can be regarded as linear operators in Ω_i.

Consider a group of linear operators in Ω_i,

$$V_{1,1}.\ \ U_{1,1}U_{2,1}{}^*,\ \ V_{2,1},\ \ U_{2,1}U_{3,1}{}^*,\ldots,V_{N-1,1},\ \ U_{N-1,1}U_{N,1}{}^* \qquad (3.2.26)$$

re-written in short notation as

$$W_1,\ W_2,\ W_3,\ \ldots\ldots,\ W_{2(N-1)}, \qquad (3.2.27)$$

which satisfies the condition that each element commutes with its neighbours by the appropriate rule associated with the commutative rule in (3.2.23) and commutes with all other elements in the sequence in the ordinary sense. We can show that all of them are linearly independent, and furthermore, that all following $S^{2(N-1)}$ operators in Ω_i are linearly independent:

$$W_1{}^{i_1}W_2{}^{i_2}\ldots\ldots W_{2(N-1)}{}^{i_{2(N-1)}}, \qquad (3.2.28)$$

where i_1, $i_2\ldots i_{2(N-1)} = 0, 1,\ldots S-1$.

It is known that for any $S^{(N-1)}$-dimensional space, all linear operators form an algebra which is equivalent to the matrix algebra of $S^{(N-1)}$ dimensions. So the algebra generated by those basic elements in (3.2.28) is equivalent to the matrix algebra of $S^{(N-1)}$ dimensions, although those matrices in (3.2.28) are S^N dimensional, we can regard them as $S^{(N-1)}\times S^{(N-1)}$ matrices. The previously independent operator $V_{N,1}$ is now expressible in Ω_i by others; that is,

$$V_{N,1} = \psi^{i-1}V_{1,1}{}^{S-1}V_{2,1}{}^{S-1}\ldots V_{N-1,1}{}^{S-1} \qquad (3.2.29)$$

$$= \psi^{i-1} W_1{}^{S-1} W_3{}^{S-1} \ldots\ldots\ldots W_{2N-3}{}^{S-1}. \qquad (3.2.30)$$

We can also notice that

$$U_{N,1} U_{1,1}{}^{*} = W_2{}^{S-1} W_4{}^{S-1} \ldots\ldots\ldots W_{2(N-1)}{}^{S-1}. \qquad (3.2.31)$$

Now consider another group of basis elements of the algebra

$$U_{1,1}U_{2,1}{}^{*},\quad V_{2,1},\quad U_{2,1}U_{3,1}{}^{*},\ V_{3,1},\ldots U_{N-1,1}U_{N,1}{}^{*},\quad V_{N,1} \qquad (3.2.32)$$

which has the same properties as (3.2.26) or (3.2.27), including the commutative rule. Therefore the transformation

$$\begin{aligned} V_{j,1} &\longrightarrow U_{j,1}U_{j+1,1}{}^{*} \\ U_{j,1}U_{j+1,1}{}^{*} &\longrightarrow V_{j+1,1} \qquad j=1,2,\ldots N-1 \end{aligned}$$

describes an automorphism of the whole algebra. Consider two matrices

$$A_1 = \sum_{j=1}^{N} \sum_{\nu=1}^{S-1} U_{j\nu}U_{j+1,\nu}{}^{*} \qquad A_2 = \sum_{j=1}^{N} \sum_{\nu=1}^{S-1} V_{j\nu}.$$

Under the automorphism we have

$$A_1 \longrightarrow \sum_{j=2}^{N} \sum_{\nu=1}^{S-1} V_{j\nu} + \sum_{\nu=1}^{S-1} \psi^{-\nu(i-1)} V_{1\nu} \qquad (3.2.33)$$

$$A_2 \longrightarrow \sum_{j=1}^{N-1} \sum_{\nu=1}^{S-1} U_{j,\nu}U_{j+1,\nu}{}^{*} + \sum_{\nu=1}^{S-1} \psi^{\nu(i-1)} U_{N,\nu}U_{1,\nu}{}^{*}. \qquad (3.2.34)$$

Note that under the automorphism, both transformed matrices are only slightly different from A_2 and A_1, respectively. In particular, corresponding to the subspace Ω_1,

$$A_1 \longrightarrow A_2; \qquad\qquad A_2 \longrightarrow A_1,$$

or more compactly,

$$f(A_1,A_2) \longrightarrow f(A_2,A_1) \qquad (3.2.35)$$

in the algebra associated with space Ω_1. Application to the operators D_2 and D_3 given by (3.2.20) and (3.2.16) shows that the following two operators

$$\exp\{-N\alpha/S\}[d(\beta)]^{-N}D_2(\alpha)D_3(\beta) = \exp\{\alpha A_2/S\}\exp\{\beta^* A_1/S\}$$

and

$$\exp\{-N\beta^*/S\}[d(\alpha^*)]^{-N}D_2(\beta^*)D_3(\alpha^*) = \exp\{\beta^* A_2/S\}\exp\{\alpha A_1/S\},$$

can be regarded as the same under the automorphism, so that the set of $(S^{(N-1)})$ eigenvalues of the matrix $D_2(\alpha)D_3(\beta)$, which are associated with the space Ω_1, can be obtained from the corresponding parts of the eigenvalues of $D_2(\beta^*)D_3(\alpha^*)$ by the relation

$$\lambda_{(1)}(\alpha,\beta) = [(e^{\alpha}-1)(e^{\beta}-1)/S]^N\lambda_{(1)}(\beta^*,\alpha^*) \qquad (3.2.36)$$

This result applies in particular to the largest eigenvalues of the operators. The reason is that the operators are matrices with all positive entries, so each of them has one and only one positive eigenvector, ξ, say, which corresponds to the largest eigenvalue (Varga, 1962, see also Bushell(1973) and Istratescu(1981)), then because of the commutability between them and D, $D\xi$ is also a positive eigenvector associated with the largest eigenvalue, and therefore $D\xi=\xi$. That means ξ must be in Ω_1. Therefore,

$$\lambda_{\max}(\alpha,\beta) = [(e^{\alpha}-1)(e^{\beta}-1)/S]^N\lambda_{\max}(\beta^*,\alpha^*)' \qquad (3.2.37)$$

On the other hand, if we disregard the effect of ψ in (3.2.33) and (3.2.34), we know that the eigenvalues of $\exp\{\alpha A_2/S\}\exp\{\beta^* A_1/S\}$ are the same as those of $\exp\{\beta^* A_2/S\}\exp\{\alpha A_1/S\}$. Therefore, approximately or asymptotically (as $M \longrightarrow \infty$),

$$C(\alpha,\beta) = [(e^{\alpha}-1)(e^{\beta}-1)/S]^{MN}C(\beta^*,\alpha^*). \qquad (3.2.38)$$

Although we do not have a theoretical guarantee of the existence of the critical points of the multi-colour Ising model, a simulation study has shown their existence. We know, for each pair of parameters (α,β), that there exists a corresponding pair (β^*,α^*), which exchanges with (α,β). Thus when $C(\alpha,\beta)$ is analytic at (α,β), it must also be at (β^*,α^*), and if (α,β) is a critical point, so is (β^*,α^*). Thus if there are not many critical points, they must occur at the fixed points of the transformation: $(\alpha,\beta) \longrightarrow (\beta^*,\alpha^*)$. To be precise, the critical points satisfy the equation

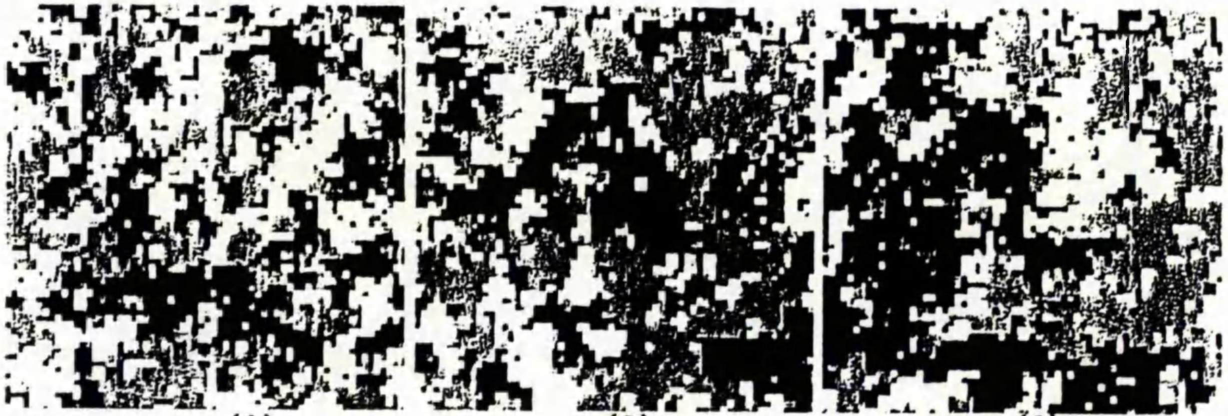
$$(e^{\alpha} - 1)(e^{\beta} - 1) = S. \qquad (3.2.39)$$

In the case $\alpha=\beta$, the critical point is

$$\log(1 + \sqrt{S}) \tag{3.2.40}$$

which is the same as given by Potts(1952).

Some simulated images with parameter slightly bigger and smaller than the critical point, in the case of three-colour model with $\alpha=\beta$ and with free boundary conditions, are shown in Fig.3.2a and Fig.3.2b respectively. Note that now $\alpha=\beta=1.00505$ gives the critical point. We use the point-relaxation method (Geman and Geman, 1984) to generate these images. In Section 3.4, we will discuss further the simulation of Markov random fields.

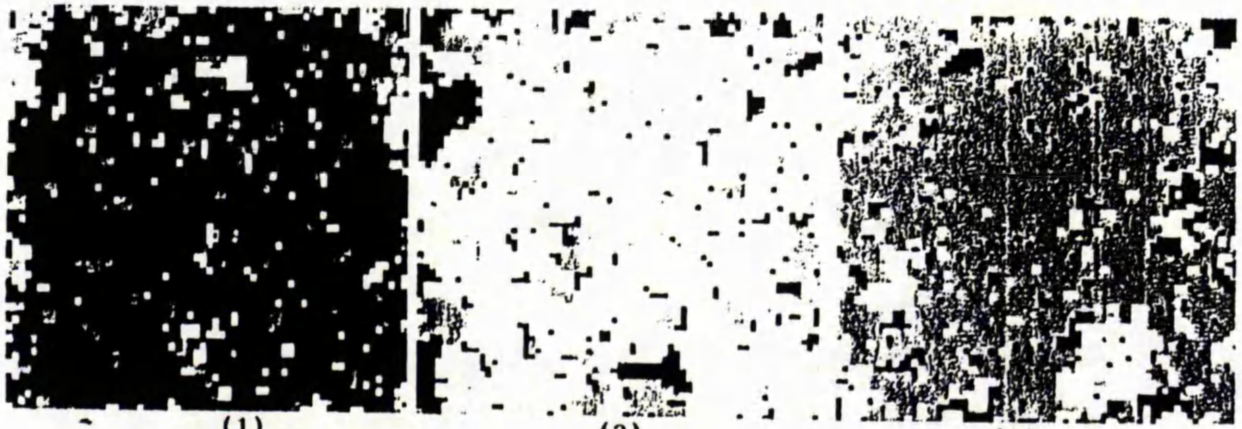


(1) $\alpha=0.95$ 7000 iterations

(2) $\alpha=0.98505$ 10000 iterations

(3) $\alpha=0.98505$ 10000 iterations

Fig 3.2a Simulated images with α smaller than the critical point



(1) $\alpha=1.05$ 1000 iterations

(2) $\alpha=1.02505$ 1000 iterations

(3) $\alpha=1.02505$ 10000 iterations

Fig 3.2b Simulated images with α bigger than the critical point

3.3 Asymptotic inference for an asymmetric Ising model around a torus

For the two colour Ising model, Pickard(1976) obtained a weak law of large numbers and a central limit theorem for the sample correlation coefficient between nearest neighbours for the isotropic (symmetric) case which involves only one parameter. In his later paper(Pickard, 1977), the results were extended to the asymmetric case for the sample correlations along rows and columns. His work in Pickard(1976) was for a lattice around a torus, ie, with perodic boundary counditions, while in Pickard (1977), he did not assume this condition. However, as indicated in Pickard(1982, 1986), although the maximum likelihood estimator of parameters is consistent and asymptotically normal, it cannot be computed. An alternative solution is therefore taken, namely, to solve an asymptotic-normal equation. The resulting asymptotic-MLE is consistent but the central limit result cannot be derived, except for the case with periodic boundary conditions(Pickard, 1976). In this section, we extend the results of Pickard(1976) for the symmetric case to the asymmetric case with boundary conditions, so that the asymptotic-MLE is asymptotically normal, thereby enabling us to discuss the asymptotic behaviour of the asymptotic-likelihood-ratio test for symmetry.

3.3.1 The model

As mentioned in the last section, the imposition of periodic boundary conditions is equivalent to wrapping the lattice around a torus. The model we consider here is for the two-colour case; the potential functions are slightly different from those of the multi-colour model in the last section. We specify the joint distribution function, which involves two parameters, α and β, by

$$P(X=x|\alpha,\beta)=\frac{1}{C(\alpha,\beta)}\exp\{\sum_{i=1}^{M}\sum_{j=1}^{N}[\alpha x_{ij}x_{i,j+1}+\beta x_{ij}x_{i+1,j}]\}. \qquad (3.3.1)$$

By the symmetry of all pixels in (3.3.1), the random variables, x_{ij} are identically distributed with equal probabilities at -1 and +1, but they are not independent unless $\alpha=\beta=0$.

Consider the vector random variable

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \sum_{i=1}^{M} \sum_{j=1}^{N} \begin{bmatrix} x_{ij}x_{i,j+1} \\ x_{ij}x_{i+1,j} \end{bmatrix}, \tag{3.3.2}$$

whose distribution is determined by the parameter values. Clearly, Q/MN gives the sample correlations between nearest neighbours along rows and columns. Our first aim is to determine the asymptotic properties of Q. For the symmetric case, where $\alpha=\beta$, Pickard(1976) obtained a central limit theorem and a weak law of large numbers for Q_1+Q_2. We will achieve this in the case $\alpha\neq\beta$ by the same method of analysing the partition function.

The matrix method(Kramers and Wannier, 1941, Newell and Montroll, 1953) was used to obtain a representation of $C(\alpha,\beta)$ as follows; see also Kaufmann(1949) and Pickard(1976). We use the same notation as in these papers.

$$C(\alpha,\beta) = \mathrm{trace}(V_2V_1)^M = \sum_{\lambda} \lambda^M \tag{3.3.3}$$

with

$$V_1 = \prod_{r=1}^{N} (e^{\beta}I + e^{-\beta}c_r); \qquad V_2 = \exp(\alpha \sum_{r=1}^{N} s_r s_{r+1}),$$

where all matrices are of dimension 2^N, I is the identity, and $s_{N+1}=s_1$. $\{\lambda\}$ are the eigenvalues of V_2V_1, and s_r and c_r are Kronecker products of N quaternion matrices, ie,

$$c_r = I_2 \otimes I_2 \otimes \ldots \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \ldots \otimes I_2$$

$$s_r = I_2 \otimes I_2 \otimes \ldots \otimes \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes \ldots \otimes I_2 .$$

Note that c_r and s_r are special cases of $V_{r.1}$ and $U_{r,1}$ respectively, but because of the difference of interaction functions, V_1 and V_2 are slightly different from D_2 and D_3 in the last section, respectively.

3.3.2 The eigenvalues of V_1V_2

Kaufmann(1949) obtained a product decomposition of V_2V_1 in terms of matrices V^+ and V^- which are representative of rotations in 2N-space. Half of the eigenvalues of V^+ and half of those of V^- are the eigenvalues of V_2V_1. Pickard(1976) changed Kaufmann's notation slightly, and provided the eigenvalues of V_2V_1 for the case $\alpha=\beta$. As mentioned in the last section, it is well known that, asymptotically, the partition function has critical points. For the asymmetric case, a careful analysis of the way of choosing eigenvalues from V^+ and V^- to be those of V_2V_1 enables us to obtain a critical curve, CL_1, which is defined (for $0<\alpha$, $\beta<\infty$) by

$$\cosh 2\alpha \cosh 2\beta = \sinh 2\alpha + \sinh 2\beta. \tag{3.3.4}$$

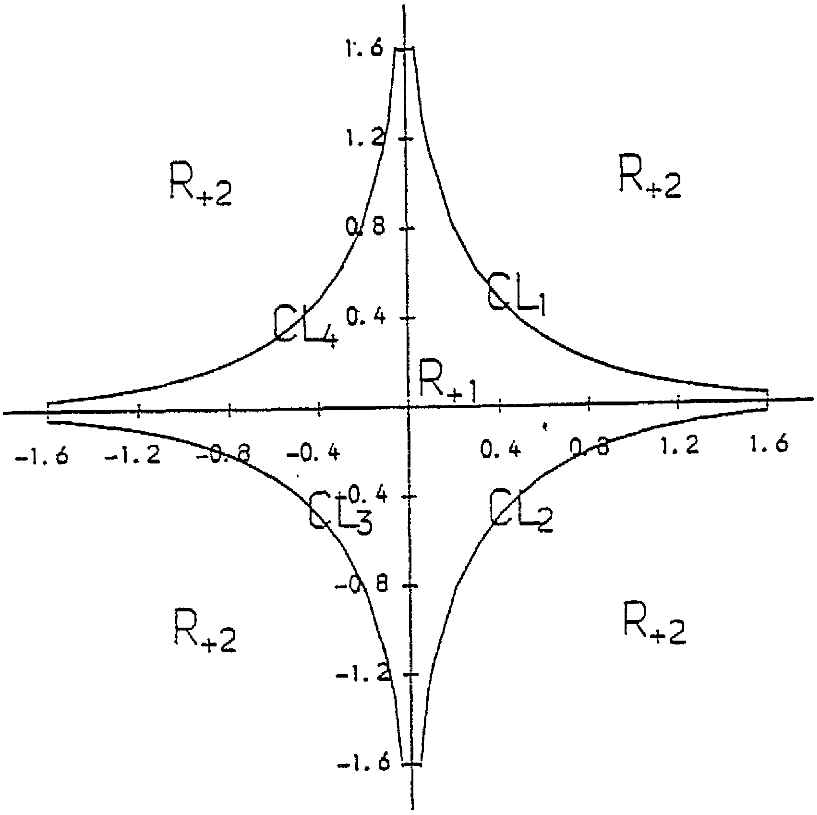


Fig 3.3 The critical lines and the corresponding two areas

Note that the critical points are in fact those which make γ_0 equal to zero, (γ_0 is associated with the eigenvalues of V and is defined later.) As shown in Fig 3.3, the area $\{\alpha>0, \beta>0\}$ is separated by CL_1 into two parts, denoted by R_{+1} and R_{+2}. The eigenvalues of V_2V_1, denoted by λ^+ and λ^- corresponding to V^+ and V^- respectively, are then given (for $0<\alpha$, $\beta<\infty$) by

$$\lambda^{+} = \lambda^{+}(\alpha,\beta) = (2\sinh 2\beta)^{N/2}\exp\{\tfrac{1}{2}\sum_{i=1}^{N} \delta_{2i-1}\gamma_{2i-1}\} \qquad (3.3.5)$$

$$\lambda^{-} = \lambda^{-}(\alpha,\beta) = \nu^{\delta_0}(2\sinh 2\beta)^{N/2}\exp\{\tfrac{1}{2}\sum_{i=1}^{N} \delta_{2i-2}\gamma_{2i-2}\} \qquad (3.3.6)$$

where, in each exponent, an even number of δ's are -1 while the other are $+1$, giving a total of 2^N eigenvalues of V_2V_1, and

$$\gamma_i = \gamma(\alpha,\beta,i\pi/N),$$

$$\gamma(\alpha,\beta,\omega) = \cosh^{-1}\left(\frac{\cosh 2\alpha\cosh 2\beta - \sinh 2\alpha\cos\omega}{\sinh 2\beta}\right),$$

$$\nu = \begin{cases} \exp(-\gamma_0) & \text{if } (\alpha,\beta)'\epsilon R_{+1} \\ 1 & \text{if } (\alpha,\beta)'\epsilon R_{+2}+CL_1. \end{cases}$$

Since $\cosh^{-1}x=\log[x + (x-1)^{\frac{1}{2}}]$ for $x\geqslant 1$, and noting that for $0<\alpha$, $\beta<\infty$,

$$\frac{\cosh 2\alpha\cosh 2\beta-\sinh 2\alpha\cos\omega}{\sinh 2\beta} \geqslant \frac{\cosh 2\alpha\cosh 2\beta-\sinh 2\alpha}{\sinh 2\beta} \geqslant 1,$$

we can write

$$\gamma(\alpha,\beta,\omega) = \psi(\alpha,\beta,\omega) - \log\sinh 2\beta,$$

where

$$\psi(\alpha,\beta,\omega) = \log\{\cosh 2\alpha\cosh 2\beta - \sinh 2\alpha\cos\omega + [(\cosh 2\alpha\cosh 2\beta-\sinh 2\alpha\cos\omega)^2 - (\sinh 2\beta)^2]^{\frac{1}{2}}\}.$$

If we write $\psi_i=\psi(\alpha,\beta,i\pi/N)$, the eigenvalues of V_2V_1 can be re-written as

$$\lambda^{+}=\lambda^{+}(\alpha,\beta)=2^{N/2}(\sinh 2\beta)^{\iota+}\exp\{\tfrac{1}{2}\sum_{i=1}^{N} \delta_{2i-1}\psi_{2i-1}\} \qquad (3.3.7)$$

and

$$\lambda^{-}=\lambda^{-}(\alpha,\beta)=\mu^{\delta_0}2^{N/2}(\sinh 2\beta)^{\iota-}\exp\{\tfrac{1}{2}\sum_{i=1}^{N} \delta_{2i-2}\psi_{2i-2}\}, \qquad (3.3.8)$$

where $\iota+$ and $\iota-$ are the numbers of δ's which are equal to -1 in the corresponding exponents and

$$\mu = \begin{cases} \exp\{-\psi_0\}\sinh 2\beta & (\alpha,\beta)'\in R_{+1} \\ 1 & (\alpha,\beta)'\in R_{+2}+CL_1, \end{cases}$$

It is clear that the largest eigenvalues of λ^+ and λ^- occur when all δ's are +1, ie,

$$\lambda^+_{max}=2^{N/2}\exp\{\tfrac{1}{2}\sum_{i=1}^{N}\psi_{2i-1}\}; \quad \lambda^-_{max}=2^{N/2}\mu\exp\{\tfrac{1}{2}\sum_{i=1}^{N}\psi_{2i-2}\}.$$

The representation of the partition in terms of λ^+ and λ^- is now only restricted to $0<\alpha$, $\beta<\infty$. It is desirable to extend it to $\mathbb{R}^2$. As in Pickard(1976) for the symmetric case, we require M and N to be even, so that $C(\alpha,\beta)$ satisfies

$$C(\alpha,\beta) = C(-\alpha,-\beta) = C(-\alpha,\beta). \qquad (3.3.9)$$

Note that, when M and N are not both even, the above property does not hold, so that the following results have to be regarded as proved only for the case of even M and N.

Also note that $\psi(\alpha,\beta)$ and therefore λ^+ and λ^- are defined for $\alpha=0$ or $\beta=0$ in (3.3.7) and (3.3.8). Appendix 1 shows that (3.3.3) provides the representation of the partition for $\{\alpha\geqslant 0,\ \beta\geqslant 0\}$, with the help of (3.3.7) and (3.3.8). Hence the partition can be extended to $\mathbb{R}^2$ by extending λ^+, λ^- and ψ to $\mathbb{R}^2$ by replacing α, β by their absolute values. There are then all four critical curves, which partition $\mathbb{R}^2$ into five parts. We still denote the area around the origin by R_{+1} and the others by R_{+2}. Finally, in this section, we indicate that the critical curves (Fig.3.3) are equivalently defined by

$$\sinh 2\alpha \sinh 2\beta = \pm 1. \qquad (3.3.10)$$

3.3.3 Asymptotic behaviour of $C(\alpha,\beta)$

It is convenient to deal with

$$lc(\alpha,\beta) = \log C(\alpha,\beta) = \log \sum_{\lambda} \lambda^M. \qquad (3.3.11)$$

Pickard(1977) provided $lc(\alpha,\beta)$ in terms of a two-dimensional Riemann sum plus an error term, with the help of a Pfaffian. Although this was done in the context of a free boundary condition, all that is known about the error term and its derivatives is that they are no larger than O(M), so the central limit theorem obtained is not completely satisfatory.

Write $\boldsymbol{lc}(\alpha,\beta)$ as

$$\boldsymbol{lc}(\alpha,\beta) = J(\alpha,\beta) + K(\alpha,\beta), \tag{3.3.12}$$

where

$$J(\alpha,\beta) = \log\lambda_{\max}^{+\,M} = ½MN\log2 + ½M\sum_{i=1}^{N}\psi_{2i-1}$$

and

$$K(\alpha,\beta) = \log\sum_{\lambda\neq\lambda_{\max}^{+}}(\lambda/\lambda_{\max}^{+})^{M}.$$

For the moment, we shall restrict our attention to $[0,+\infty)\times[0,+\infty)$. Consider the function defined by the Riemann integral

$$\boldsymbol{B}(\alpha,\beta) = (4\pi)^{-1}\int_{0}^{2\pi}\psi(\alpha,\beta,\omega)d\omega \tag{3.3.13}$$

Then the properties of ψ ensure that: the first partial derivatives of $\boldsymbol{B}$ are continuous on $[0,+\infty)\times[0,+\infty)$; the higher-order partial derivatives of $\boldsymbol{B}$ are continuous on $[0,+\infty)\times[0,+\infty)\setminus CL_1$ and undefined on CL_1; and these derivatives , are obtained by differentiation under the integral sign. For s=1,2,...,

$$\frac{1}{2N}\sum_{i=1}^{N}\frac{\partial^{s}\psi_{2i-1}}{\partial\alpha^{j}\partial\beta^{s-j}} \quad \text{and} \quad \frac{1}{2N}\sum_{i=1}^{N}\frac{\partial^{s}\psi_{2i-2}}{\partial\alpha^{j}\partial\beta^{s-j}}$$

converge to $\dfrac{\partial^{s}\boldsymbol{B}}{\partial\alpha^{j}\partial\beta^{s-j}}$, wherever it exists. Moreover, convergence is uniform on any compact subset of the domain of $\dfrac{\partial^{s}\boldsymbol{B}}{\partial\alpha^{j}\partial\beta^{s-j}}$.

Note that, as $\psi(\alpha,\beta,\omega)$, λ^{+}, λ^{-} and $B(\alpha,\beta)$ can also be extended to $\mathbb{R}^2$ to satisfy $\boldsymbol{B}(\alpha,\beta)=\boldsymbol{B}(-\alpha,-\beta)=\boldsymbol{B}(-\alpha,\beta)$,the extended function $\boldsymbol{B}(\alpha,\beta)$ is such that, for a fixed α (β), $\boldsymbol{B}(\alpha,*)$ ($\boldsymbol{B}(*,\beta)$) have right-handed derivatives of all orders at the origin, and the odd-ordered derivatives vanish there. These results are clear by noticing the following representation of $\boldsymbol{B}$:

$$\boldsymbol{B}(\alpha,\beta) = \frac{1}{2}\log\{2\cosh2\alpha\cosh2\beta\} - \frac{1}{2}\sum_{i=1}^{\infty}\frac{1}{2i}\binom{2i}{i}\left\{\left[\frac{d_1}{2}\right]^{2i} + \left[\frac{d_2}{2}\right]^{2i}\right\}$$

$$- \frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{(2i+2j-1)!}{(i!)^2(j!)^2}\left[\frac{d_1}{2}\right]^{2i}\left[\frac{d_2}{2}\right]^{2j},$$

where $d_1=\tanh 2\alpha/\cosh 2\beta$ and $d_2=\tanh 2\beta/\cosh 2\alpha$.

We have therefore proved that, as $N \longrightarrow \infty$,

$$(MN)^{-1}J(\alpha,\beta) \longrightarrow \tfrac{1}{2}\log 2 + \mathcal{B}(\alpha,\beta) \quad \text{on } \mathbb{R}^2,$$

and

$$(MN)^{-1}\frac{\partial^s J(\alpha,\beta)}{\partial\alpha^j\partial\beta^{s-j}} \longrightarrow \frac{\partial^s \mathcal{B}(\alpha,\beta)}{\partial\alpha^j\partial\beta^{s-j}} \tag{3.3.14}$$

on $\mathbb{R}^2$ for $s=1$ and on $R_{+1}+R_{+2}$ for $s\geqslant 2$, and that the convergence is uniform in $(\alpha,\beta)'$ on any compact subset of the appropriate region.

In Pickard(1976) analytic complex functions were used to determine the approximate speed of convergence of a Riemann sum to its Riemann integral. In our case, we take G to be any compact subset of $R_{+1}+R_{+2}$. Then there is a positive number ρ such that, for any fixed $(\alpha,\beta)'\in G$, the complex function

$$g(\alpha,\beta,z) = \log\{A(\alpha,\beta,z) + [A(\alpha,\beta,z)^2 - \sinh^2 2\beta]^{\frac{1}{2}}\},$$

is analytic in the annulus $H=\{z: e^{-\rho}\leqslant|z|\leqslant e^{\rho}\}$, where $A(\alpha,\beta,z) = \cosh 2\alpha\cosh 2\beta - \tfrac{1}{2}\sinh 2\alpha(z + 1/z)$. Thus,

$$(MN)^{\frac{1}{2}}|(MN)^{-1}\nabla J(\alpha,\beta) - \nabla\mathcal{B}(\alpha,\beta)| = O((MN)^{\frac{1}{2}}e^{-\rho N}),$$

where ∇ denotes the first-order derivative vector. (We will use ∇^2 to denote the second-order derivative matrix.) Therefore, as M, N $\longrightarrow\infty$ (with $M\leqslant N^{\theta}$ for any fixed $\theta>0$),

$$(MN)^{\frac{1}{2}}[(MN)^{-1}\nabla J(\alpha,\beta) - \nabla\mathcal{B}(\alpha,\beta)] \longrightarrow 0 \tag{3.3.15}$$

uniformly for $(\alpha,\beta)'\in G$.

In order to deal with $K(\alpha,\beta)$, by almost the same procedure as that of Pickard(1976) for the case $\alpha=\beta$, we can prove that, as M, $N\longrightarrow\infty$ (provided $N^{\theta_1}\leqslant M\leqslant N^{\theta}$ for any fixed θ_1 and θ with $0<\theta_1<\theta<\infty$),

$$K(\alpha,\beta) \longrightarrow \begin{cases} 0 & \text{if } (\alpha,\beta)'\in R_{+1} \\ \log 2 & \text{if } (\alpha,\beta)'\in R_{+2} \end{cases} \tag{3.3.16}$$

and

$$\frac{\partial^s K(\alpha,\beta)}{\partial\alpha^j\partial\beta^{s-j}} \longrightarrow 0 \quad \text{for } s=1,2,\ldots,\ (\alpha,\beta)'\in R_{+1}+R_{+2}. \tag{3.3.17}$$

Again, the convergence is uniform on any compact subset of $R_{+1}+R_{+2}$. Combining (3.3.14) and (3.3.17) we obtain

$$(MN)^{-1}\frac{\partial^s \boldsymbol{lc}(\alpha,\beta)}{\partial\alpha^j\partial\beta^{s-j}} \longrightarrow \frac{\partial^s \boldsymbol{B}(\alpha,\beta)}{\partial\alpha^j\partial\beta^{s-j}}, \qquad (3.3.18)$$

$$(\alpha,\beta)'\in R_{+1}+R_{+2},\ s=1,2,\dots.$$

as M, N⟶∞ (provided $N^{\Theta_1}\leqslant M\leqslant N^{\Theta}$ for any fixed Θ_1 and Θ with $0<\Theta_1<\Theta<\infty$). Furthermore, convergence is uniform in $(\alpha,\beta)'$ on any compact subset of $R_{+1}+R_{+2}$.

3.3.4 The limiting theorems

Let $(\alpha,\beta)'\in R_{+1}+R_{+2}$ and choose any compact subset $D_\delta=\{(x,y)':(x-\alpha)^2+(y-\beta)^2\leqslant\delta\}$ such that $D_\delta\subset R_{+1}+R_{+2}$. For a fixed two-dimensional vector $\eta=(\eta_1,\eta_2)'$, consider the random variables

$$S = (MN)^{-1}\eta' Q \qquad \text{and} \qquad T = (MN)^{1/2}\eta'[(MN)^{-1}Q - \nabla\boldsymbol{B}(\alpha,\beta)].$$

Denote by $M_S(t:\alpha,\beta)$ and $M_T(t:\alpha,\beta)$ the moment-generating functions of S and T. Then

$$\log M_S(t:\alpha,\beta) = \boldsymbol{lc}(\alpha+(MN)^{-1}t\eta_1,\beta+(MN)^{-1}t\eta_2) - \boldsymbol{lc}(\alpha,\beta)$$

and

$$\log M_T(t:\alpha,\beta) =$$
$$\boldsymbol{lc}(\alpha+(MN)^{-1/2}t\eta_1,\beta+(MN)^{-1/2}t\eta_2)-\boldsymbol{lc}(\alpha,\beta)-t(MN)^{1/2}\eta'\nabla\boldsymbol{B}(\alpha,\beta).$$

When M and N are large enough, $(\alpha+(MN)^{-1}t\eta_1,\ \beta+(MN)^{-1}t\eta_2)'\in D_\delta$ and $(\alpha+(MN)^{-1/2}t\eta_1,\beta+(MN)^{-1/2}t\eta_2)'\in D_\delta$. Noting (3.3.11) and applying Taylor's theorem to K and J, we obtain

$$\log M_S(t:\alpha,\beta) = (MN)^{-1}t\eta'\nabla\boldsymbol{lc}(\alpha+(MN)^{-1}\bar{t}_1\eta_1,\beta+(MN)^{-1}\bar{t}_1\eta_2)$$

and

$$\log M_T(t:\alpha,\beta) = t^2(MN)^{-1}\eta'\nabla^2 J(\alpha+(MN)^{-1/2}\bar{t}_2\eta_1,\beta+(MN)^{-1/2}\bar{t}_2\eta_2)\eta$$

$$+(MN)^{1/2}\eta'[(MN)^{-1}\nabla J(\alpha,\beta)-\nabla B(\alpha,\beta)]+K(\alpha+(MN)^{-1/2}t\eta_1,\beta+(MN)^{-1/2}t\eta_2)-K(\alpha,\beta)$$

where $|\bar{t}_1|\leqslant|t|$, $|\bar{t}_2|\leqslant|t|$. The uniform convergence of (3.3.14)), (3.3.16) and (3.3.18) in D_δ, and the convergence of (3.3.15), imply that, as M, N⟶∞ (provided $N^{\Theta_1}\leqslant M\leqslant N^{\Theta}$ for any fixed Θ_1 and Θ with

$0<\theta_1<\theta<\infty$),

$$M_S(t:\alpha,\beta) \longrightarrow \exp\{t\eta'\nabla\boldsymbol{B}(\alpha,\beta)\}$$

$$M_T(t:\alpha,\beta) \longrightarrow \exp\{t^2\eta'\nabla^2\boldsymbol{B}(\alpha,\beta)\eta\}$$

Since $\exp\{t^2\eta'\nabla^2\boldsymbol{B}(\alpha,\beta)\eta\}$ is the generating function of the normal distribution with zero mean and variance $\eta'\nabla^2\boldsymbol{B}(\alpha,\beta)\eta$, and since $\exp\{t\eta'\nabla\boldsymbol{B}(\alpha,\beta)\}$ is that of the degenerate distribution at $\eta'\nabla\boldsymbol{B}(\alpha,\beta)$, it follows that

$$(MN)^{-1}Q \xrightarrow{\text{pr}} \nabla\boldsymbol{B}(\alpha,\beta) \qquad (3.3.19)$$

$$(MN)^{-\frac{1}{2}}((MN)^{-1}Q - \nabla\boldsymbol{B}(\alpha,\beta)) \xrightarrow{D} N(0,\nabla^2\boldsymbol{B}(\alpha,\beta)). \qquad (3.3.20)$$

Pickard's(1977) numerical results of the asymptotic correlation between Q_1 and Q_2 showed that $\nabla^2\boldsymbol{B}(\alpha,\beta)$ could be a positive definite matrix when $(\alpha,\beta)'$ is not a critical point. It is very difficult to prove this, because of the complexity of the function $\boldsymbol{B}(\alpha,\beta)$. We will assume this result in the remainder of present discussion.

The likelihood function (3.3.1) cannot be maximized, since $C(\alpha,\beta)$ and $\nabla C(\alpha,\beta)$ are almost impossible to compute. An obvious alternative solution is to maximize an asymptotic-likelihood, $\alpha Q_1+\beta Q_2-MN\boldsymbol{B}(\alpha,\beta)$, or to solve the asymptotic-normal equation

$$(MN)^{-1}Q = \nabla\boldsymbol{B}(\alpha,\beta). \qquad (3.3.21)$$

Suppose α_0, β_0 are the true values of the parameters and $(\alpha_0,\beta_0)'$ is not a critical point. Denote by $(\hat{\alpha},\hat{\beta})$ the solution of (3.3.21). The standard method therefore yields that, as M, N$\longrightarrow\infty$ (provided $N^{\theta_1}\leqslant M\leqslant N^{\theta}$ for any fixed θ_1 and θ with $0<\theta_1<\theta<\infty$),

$$(MN)^{\frac{1}{2}}\begin{bmatrix}\hat{\alpha} - \alpha_0\\ \hat{\beta} - \beta_0\end{bmatrix} \xrightarrow{D} N(0,[\nabla^2\boldsymbol{B}(\alpha_0,\beta_0)]^{-1}) \qquad (3.3.22)$$

and

$$\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix} \xrightarrow{\text{pr}} \begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}. \qquad (3.3.23)$$

Note that (3.3.23) holds even if $(\alpha_0,\beta_0)'\in CL$. For the symmetric case, with $\alpha_0=\beta_0$, suppose the asymptotic-MLE $\bar{\alpha}$ maximises $\alpha\mathit{1}'Q-MN\boldsymbol{B}(\alpha,\alpha)$, where $\mathit{1}=(1,1)'$. Then

$$(MN)^{-1}(\bar{\alpha} - \alpha_0) \xrightarrow{D} N(0,\ [\mathit{1}'\nabla^2\boldsymbol{B}(\alpha_0,\beta_0)\mathit{1}]^{-1})$$

and

$$\bar{\alpha} \xrightarrow{\text{pr}} \alpha_0 .$$

3.3.5 Testing against asymmetry

In this sub-section, we are interested in testing the null hypothesis

$$H_0: \quad \alpha = \beta, \qquad \alpha \in \mathbb{R}-\{\pm\tfrac{1}{2}\sinh^{-1}1\}$$

against the general alternative hypothesis of 'not H_0'. We shall consider the use of the likelihood-ratio test. However, as indicated in the previous sub-section, it is almost impossible to maximize the likelihood function. The asymptotic-likelihood-ratio test is therefore used and is defined as follows. Define the test statistic

$$\begin{aligned}\Lambda(Q) &= \operatorname*{Sup}_{\alpha,\beta} \boldsymbol{lal}(Q|\alpha,\beta) - \operatorname*{Sup}_{\alpha} \boldsymbol{lal}(Q|\alpha,\alpha) \\ &= \hat{\alpha}Q_1 + \hat{\beta}Q_2 - MN\boldsymbol{B}(\hat{\alpha},\hat{\beta}) - \bar{\alpha}(Q_1 + Q_2) + MN\boldsymbol{B}(\bar{\alpha},\bar{\alpha}),\end{aligned}$$

where $(\hat{\alpha},\hat{\beta})'$ is the maximizing point of $\boldsymbol{lal}(Q|\alpha,\beta)$, while $\bar{\alpha}$ is that of $\boldsymbol{lal}(Q|\alpha,\alpha)$; $\boldsymbol{lal}$ is the log-asymptotic-likelihood, namely, $\alpha Q_1+\beta Q_2-MN\boldsymbol{B}(\alpha,\beta)$. We do not know the distribution of $\Lambda(Q)$, but we may use its asymptotic properties. For a size δ test, we shall define the upper δ-point q_δ, by

$$\max_{\alpha\in\mathbb{R}-\{\pm\frac{1}{2}\sinh^{-1}1\}} \lim_{M,N\to\infty} \Pr(\Lambda(Q)\geqslant q_\delta|\alpha=\beta) = \delta.$$

If $\alpha=\beta=\alpha_0\in\mathbb{R}-\{\pm\frac{1}{2}\sinh^{-1}1\}$ provides the true value of the parameters, clearly, $\hat{\alpha}$, $\hat{\beta}$ and $\bar{\alpha}$ converge to α_0 in probability. Let $\varepsilon=(MN)^{\frac{1}{2}}\Delta^{-\frac{1}{2}}[Q-\nabla\boldsymbol{B}(\alpha_0,\alpha_0)]$, where $\Delta=\nabla^2\boldsymbol{B}(\alpha_0,\alpha_0)$. It can be derived that

$$\Lambda(Q) = \tfrac{1}{2}\varepsilon'[I-\frac{(\Delta^{\frac{1}{2}}\boldsymbol{1})(\Delta^{\frac{1}{2}}\boldsymbol{1})'}{\boldsymbol{1}'\Delta\boldsymbol{1}} + o(1)]\varepsilon,$$

where $o(1)$ means that it converges to zero if $\hat{\alpha}$, $\hat{\beta}$ and $\bar{\alpha}$ converge to α_0, so it converges to zero in probability. Let $d=\Delta^{\frac{1}{2}}\boldsymbol{1}/\|\Delta^{\frac{1}{2}}\boldsymbol{1}\|=(d_1,d_2)'$, $U=\begin{bmatrix} d_1 & -d_2 \\ d_2 & d_1 \end{bmatrix}$, and $\varepsilon_1=U\varepsilon=(\varepsilon_{11}, \varepsilon_{12})'$. Then U is an orthogonal matrix, and ε_1 is asymptotically normal with zero mean and covariance matrix I. Therefore, as M, N$\longrightarrow\infty$ (provided $N^{\theta_1} \leq M \leq N^{\theta}$ for any fixed θ_1 and θ with $0<\theta_1<\theta<\infty$),

$$\lim_{M,N\longrightarrow\infty} \Pr[\Lambda(Q)\geq q_\delta] = \lim_{M,N\longrightarrow\infty} \Pr(\tfrac{1}{2}\varepsilon_{11}{}^2\geq q_\delta);$$

ε_{11} is asymptotically normal with zero mean and variance 1, and the upper δ-point, q_δ, is therefore easily obtained.

Now suppose $\alpha_0 \neq \beta_0$ are the true values, $(\alpha_0,\beta_0)\in R_{+1}+R_{+2}$. Define $\bar{\alpha}_0$ by

$$\mathit{1}'\nabla B(\bar{\alpha}_0,\bar{\alpha}_0) = \mathit{1}'\nabla \mathbf{\mathit{B}}(\alpha_0,\beta_0).$$

Note that the above equation has one and only one solution. Clearly, $(\hat{\alpha},\hat{\beta})'$ converges to $(\alpha_0,\beta_0)'$ in probability and $\bar{\alpha}$ converges to $\bar{\alpha}_0$ in probability. Therefore

$$(MN)^{-1}\Lambda(Q) \longrightarrow \mathbf{\mathit{B}}(\bar{\alpha}_0,\bar{\alpha}_0) - (\bar{\alpha}_0-\alpha_0,\ \bar{\alpha}_0-\beta_0)\nabla\mathbf{\mathit{B}}(\alpha_0,\beta_0) - \mathbf{\mathit{B}}(\alpha_0,\beta_0)$$

Since $\mathbf{\mathit{B}}(\alpha,\beta)$ is a strictly concave function, the right part of the above formula is positive, implying that, when the alternative hypothesis holds, the test statistic $\Lambda(Q)$ has, in probability, the same order as MN. While H_0 holds, $\Lambda(Q)$ satisfies a $x^2(1)$ distribution.

3.4 Simulation of MRFs; Stochastic relaxation

Methods of Monte Carlo simulation of Markov random fields are now not completely satisfatory. The theoretically valid speeds of convergence of these iterative methods are slow, and the computational demands are very substantial. The stochastic relaxation method of Geman and Geman(1984) treats each pixel individually at each stage, as far as updating is concerned. As we mentioned before, if we imagine a block of pixels, for example, one line, as one point, fields are still Markov random fields, or Markov chains, and their distributions are still of Gibbs-distribution form. Somewhat in the spirit of Clifford's discussion of Besag(1974), we may update Markov random fields a line (row or column) of pixels at a time. Thanks to the techniques we developed for one-dimensional Gibbs fields in Chapter 2 and because the density of one line, conditional upon its neighbour lines, is a Gibbs distribution, this sort of updating for one line can be carried out.

The stochastic relaxation method is not the best method for simulating Markov random fields(Ripley and Kirkland, 1990). In this section, We discuss theoretical convergence properties of line relaxation, and rely mainly on a simulation study to compare the practical convergence rates of point relaxation and block relaxation, in particular, line relaxation.

3.4.1 Line relaxation

For a Gibbs field with distribution p(X=x), Geman and Geman(1984) proved the following. Suppose $\{n_{t,}\ t\geqslant 1\}$ is a sequence of pixels such that it contains each pixel infinitely often. Whatever the starting configuration, x(0), we visit the pixel n_t at time t, and decide a new state for it according to the local probability properties and the current states of the neighbouring pixels of n_t. Then

$$\lim_{t\to\infty} \Pr(X(t)=w|x(0)) = p(X=w). \qquad (3.4.1)$$

As described, the relaxation only changes the state of one pixel at a time, although synchronous updating is also possible. The rate of convergence depends on

$$\delta = \inf_{(ij),x} \ p(X_{ij}=x_{ij}|x_{\partial ij}) \qquad (3.4.2)$$

If we assume that the pixels are visited cyclicly, and let x(t) denote the result after the t-th cycle, then

$$\sup_{x(t),x(o)} |\Pr(x(t)|x(0)) - p(x(t))| \leqslant r^t, \qquad (3.4.3)$$

where $r=1-S^L\delta^L$, S is the number of possible labels for each pixel and L=M*N. For a particular case, say, the first order Markov random field discussed in the last two sections, with a single parameter $\beta>0$ and with only two states, we have $\delta=1/(1+\exp(4\beta))$, so that r might be very near unity. Although we can obtain a better rate, it can only be improved a very little, as in the case, for example, for the above first-order model. There are, therefore, often practical problems in using relaxation to simulate Gibbs samples.

Stochastic relaxation is based on the local properties of Gibbs random fields. We concentrate on the first- or second-order pairwise interaction MRFs with appropriate distribution forms such as (1.3.3). Denote by X_i the i-th row. It follows that the density of X_i, conditional upon all other rows, depends only on its two nearest rows. To be precise,

$$p(X_i=x_i|x_{\partial i},\beta) = p(x_i|x_{i-1},x_{i+1},\beta)$$

$$= \exp\{ \sum_{j=1}^{N} \bar{G}_{ij}(x_{ij}) + \sum_{j=1}^{N-1} G_{(ij),(i,j+1)}(x_{ij},x_{i,j+1})\}/C(x_{i-1},x_{i+1}), \qquad (3.4.4)$$

where $\bar{G}_{ij}(x_{ij})$ depends on $x_{i-1,j}$ and $x_{i+1,j}$ (for the first-order case) together with $x_{i-1,j-1}$, $x_{i-1,j+1}$, $x_{i+1,j-1}$ and $x_{i+1,j+1}$ (for the second-order case). All these functions $\bar{G}$ or G depend on the parameter β, although in (3.4.4), we omit explicit mention of the parameter. Note that the above distribution, (3.4.4), is of first-order Gibbs chain form. Thus, in the same way that we change the state for one pixel, we can obtain the new states of the i-th row, given $x_1,..x_{i-1}$, $x_{i+1},...,$ x_M. This is a relaxation by replacement of one row instead of one point. We can visit the rows one by one, or carry out synchronous updating. In this case, the rate of convergence depends on

$$\delta_1 = \inf_{i,x} p(x_i|x_{i-1},x_{i+1}) \qquad (3.4.5)$$

and

$$\sup_{x(t),x(0)} |p(x(t)|x(0)) - B(x(t))| \leqslant r_1{}^t, \qquad (3.4.6)$$

where $r_1 = 1-S^L\delta_1{}^M$. Generally, $r_1<r$(equivalently, $\delta_1>\delta^N$). Since the nearby pixels of a Gibbs field tend to have the same states, δ_1 should be $p(X_i=(1)'|X_{i-1}=X_{i+1}=(0)')$ for the first-order model with two states, for which we have discussed the corresponding δ above, where (1)' denotes a vector of 1's and (0)' a vector of 0's. This δ_1 is $1/R_1$, where

$$R_1 = \frac{1+2e^{\beta}+e^{4\beta}-(1+e^{2\beta})\lambda_2}{\lambda_1 - \lambda_2}\cdot\lambda_1{}^{N-1} + \frac{(1+e^{2\beta})\lambda_1-1-2e^{\beta}-e^{4\beta}}{\lambda_1 - \lambda_2}\cdot\lambda_2{}^{N-1}$$

and

$$\lambda_1=\{1+e^{2\beta}+[(e^{2\beta}-1)^2+4)]^{\frac{1}{2}}\}/2, \quad \lambda_2=\{1+e^{2\beta}-[(e^{2\beta}-1)^2+4]^{\frac{1}{2}}\}/2.$$

It is easy to see that $R_1<[1+e^{4\beta}]^N$.

3.4.2. Simulation study

In the next section we will discuss in detail several kinds of pseudo-likelihoods. Here we mention only point pseudo-likelihood (Besag, 1974, 1986), which is based on the local conditional distribution of a pixel, and line pseudo-likelihood, which is based on the local conditional distribution of a line of pixels. In the practical computation involved in stochastic relaxation, pixels or lines are visited periodically. We also note that, for line relaxation, rows and columns can be visited alternately, in that we can first visit all rows, then all columns, then rows again, and so on.

Before we describe our simulation results, we specify the particular form of Markov random field, which we discussed in the previous two sections. However, we only consider the symmetric case, so that only one parameter $\alpha=\beta$ is involved. This is the first-order Markov random field which treats all colours equally. The simulations are based on this particular case, although the methods can be used straightforwardly for general Markov random fields.

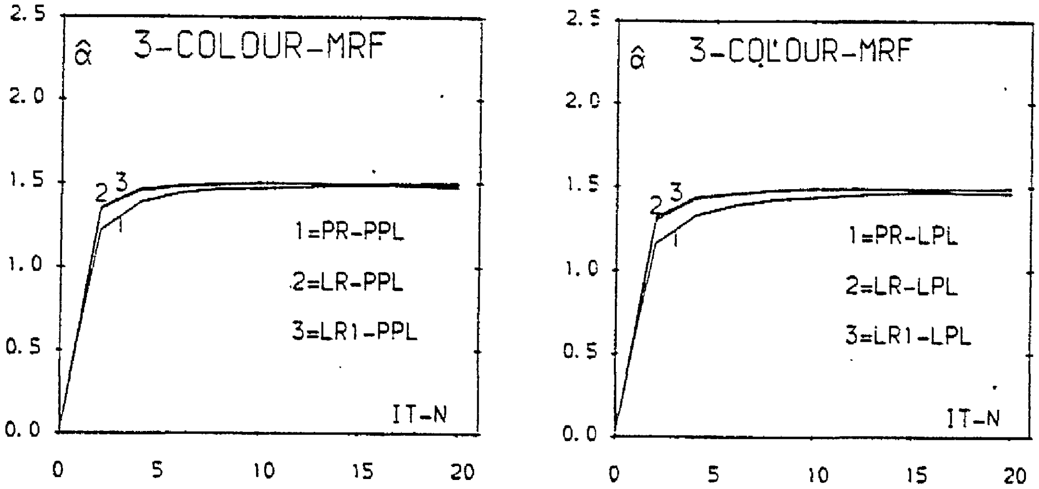


Fig 3.4 Results of first 20 cycles of relaxations for 3-colour MRFs

Fig 3.4 provides results from the first 20 cycles of iteration; where one cycle means that every pixel is visited once. It is based on a three-colour MRF with distribution (3.2.8), on a 128×128 lattice. IT-N denotes the iteration number. We initialised the simulation with a white noise image. The true value of α is 1.5 and the results for $\hat{\alpha}$ are in the form of sample means from 10 replicates. PR means point relaxation, LR means line(row) relaxation, while LR1 denotes alternate row-column relaxation. PPL denotes maximum point pseudo-likelihood estimation and LPL denotes maximum (one) line pseudo-likelihood estimation. From Fig 3.4, we see that line relaxation or alternate row-column relaxation is only a little better than point relaxation, and for all three methods, the maximum pseudo-likelihood estimates are, on average, almost the same as the true value after 15 cycles of iteration.

Although the average maximum pseudo-likelihood estimate is almost equal to the true value, we still cannot say that the iteration has converged. Fig.3.5 shows results for binary Markov random fields with many more cycles of iteration. There is only one sample for each relaxation. AL denotes maximum asymptotic-likelihood estimation (Pickard, 1976, 1987). Theoretically, because of its asymptotic normality, AL gives a better idea about whether or not the iteration has converged. We still find that LR or LR1 converges faster than PR.

Unfortunately, it is known that, when α is big (bigger, for example, than the critical point for binary Markov random fields), the simulated scenes are usually close to being one-colour images. Ripley and Kirkland(1990) reported this phenomenon. We found that the images are close to being one-colour once the AL estimates are almost equal to the true values, which may be a consequence of the clustering property of Markov random fields. We also note from Fig 3.5 that, at a certain stage in AL, there is what appears to be almost a jump to the true value for each realization. Fig 3.6 and Fig 3.7 provide some simulated patterns with small and large values for the parameter, respectively, but they are for second-order case, for which the distribution is almost the same as that in the last two sections, but each pixel has 8 neighbour pixels. In both figures, the upper three are those simulated by PR, while the lower three are simulated by RR. Monochromatic images are not usually useful in practical contexts and, in practice, Markov random fields are usually simulated under the condition that there be a prescribed number of pixels of each colour or that the boundary pixels be of predetermined colours, but the resulting patterns may be unrepresentative of the corresponding theoretical distributions.

Other block relaxations, such as two-line or more-line relaxation, could also be used, but our simulations showed that two-line relaxation is almost equivalent in performance to one-line relaxation. Even for one-line relaxation, although the maximum pseudo-likelihood

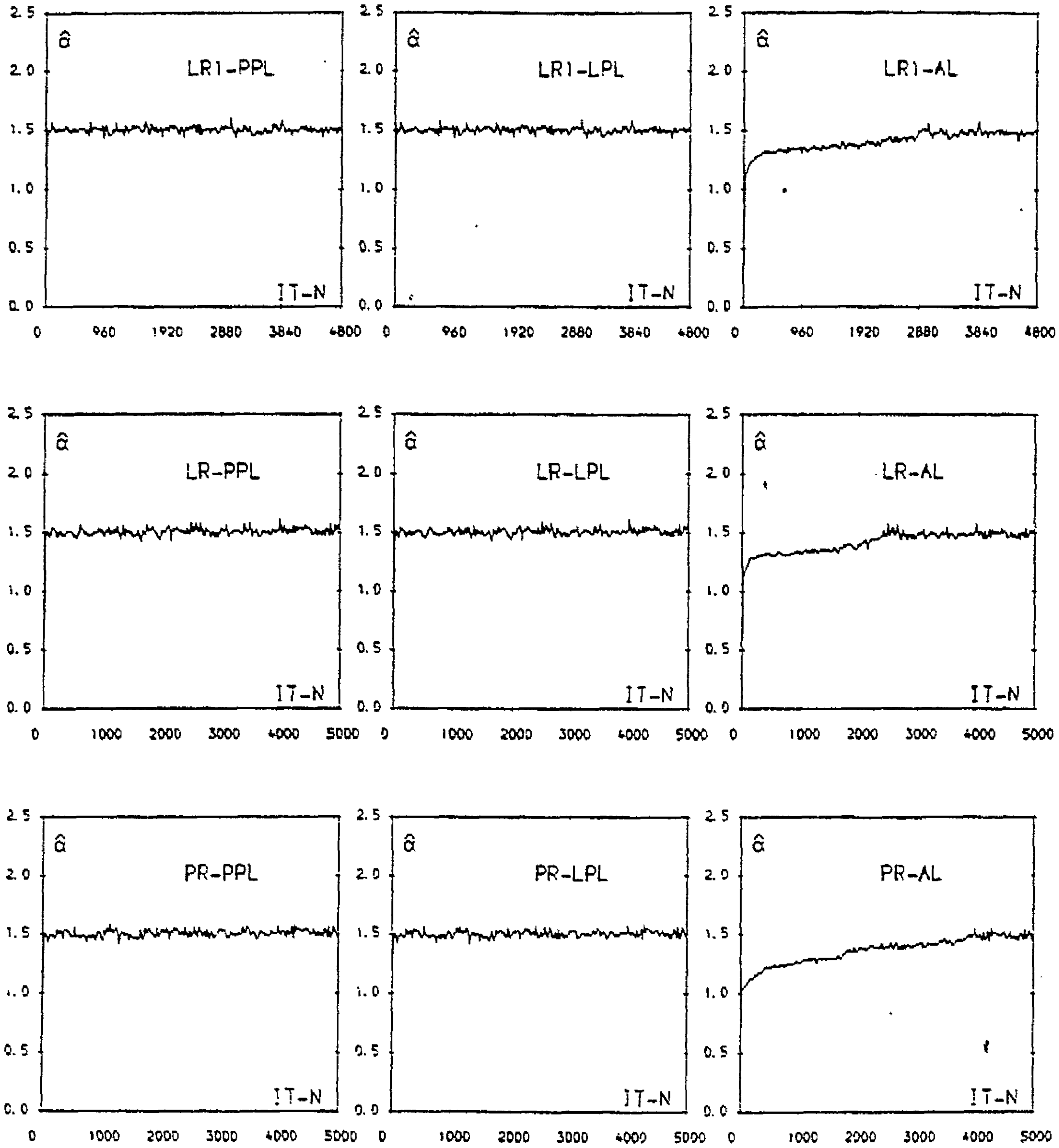


Fig 3.5 Results of relaxations for binary MRFs

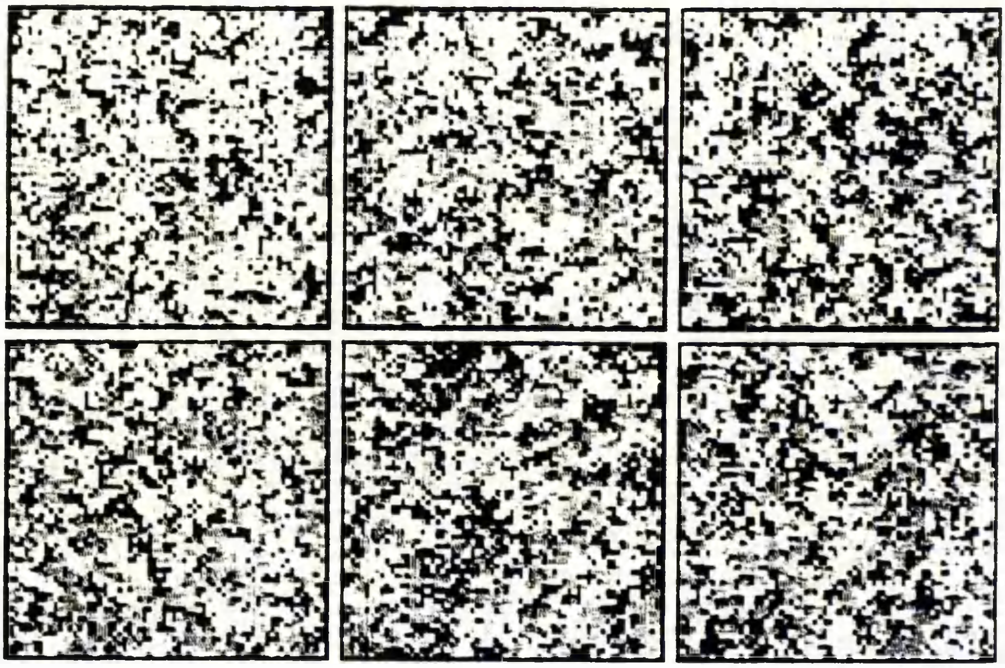


Fig 3.6 Simulated patterns with α=0.3 after 500 cycles

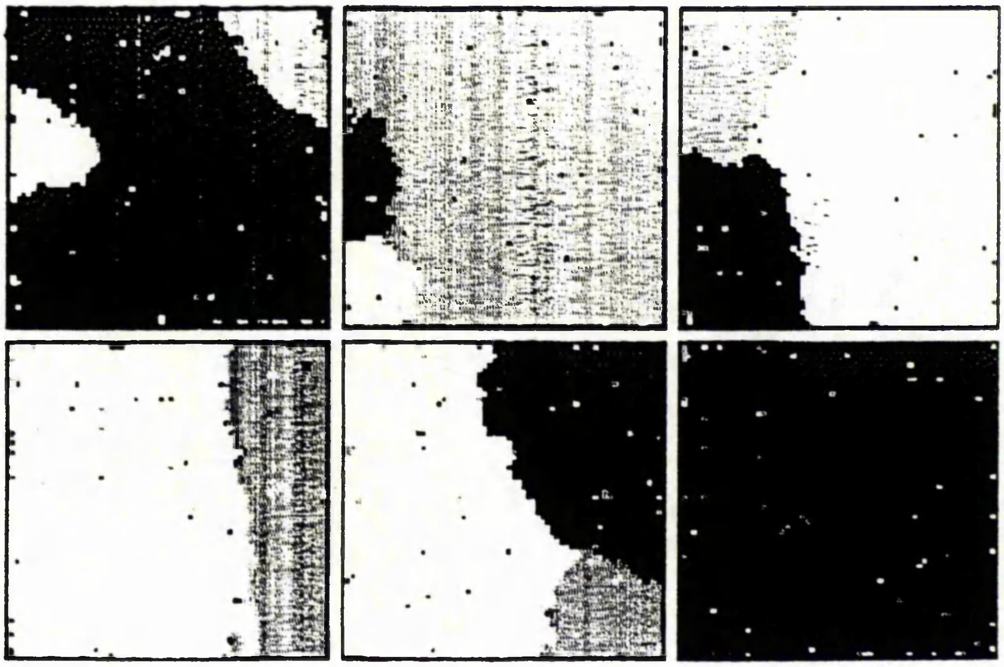


Fig 3.7 Simulated patterns with α=0.7 after 500 cycles

estimates converge to the true values faster than is the case with point relaxation, it is possible that the asymptotic-likelihood estimates converge more slowly.

3.5 Pseudo-likelihood parameter estimation

For general Markov random fields, it is difficult to write down useful expressions for their partition functions, or to compute them, so it is difficult to maximize the likelihood $p(x|\beta)$ directly(Besag, 1976, Possolo, 1986). Younes(1988a) used a Monte Carlo technique to maximize the log likelihood function. However, as mentioned in the last section, the Monte Carlo technique requires heavy computation. Alternatively, Besag(1974, 1976) introduced the coding method, in which we maximize

$$\prod_{ij\in\Delta} p(x_{ij}|x_{\partial ij},\beta) \qquad (3.5.1)$$

where Δ is a subset of the whole pixel set such that, for any ij, $mn\in\Delta, ij\neq mn$, ij and mn are not neighbours of each other. The Markov properties of Markov random fields ensure that $p(x_{ij}|x_{\partial ij},\beta)$ depends only on several neighbour pixels of ij, $\prod p(x_{ij}|x_{\partial ij},\beta)$ is therefore easily maximized. One could extend the definition of (3.5.1) and take Δ to be the whole pixel set, which results in the so-called point pseudo-likelihood form. Since the density of one line, conditional upon all other lines, is a Gibbs chain, and depends only on several neighbouring lines, according to the neighbourhood system, we (Qian and Titterington, 1989, 1990e) introduced line(row) pseudo-likelihood functions, among which the one-line case is

$$\prod_{i} p(x_i|x_{\partial i},\beta), \qquad (3.5.2)$$

where ∂i means the neighbouring rows of the i-th row. Another way of writing a pseudo-likelihood function is

$$\prod_{\Delta\in\Omega} p(x_\Delta|x_{\partial\Delta},\beta), \qquad (3.5.3)$$

where Ω is a class of some subsets of the pixel set, and Δ denotes elements of Ω. Geman and Graffigne(1987) proved the consistency of pseudo-likelihood parameter estimation. Expression (3.5.3) can then be considered as a block pseudo-likelihood function, and (3.5.1) and

(3.5.2) are special forms of (3.5.3). Note that each conditional density $p(x_i|x_{\partial i},\beta)$ in (3.5.2) corresponds to a Gibbs chain, and we can therefore compute its normalizing factor, ie, the partition function, and thereby to compute $p(x_i|x_{\partial i},\beta)$. If then the parameter β is one dimensional, we may use the Golden-Section Search method to obtain the maximizing point.

The recursive technique developed in Chapter 2 also enables us to compute the expectation of the exponential part of a Gibbs chain, and therefore, in the case where the exponential part is linear in the parameter, we may maximize the likelihood by the iterative procedure (2.2.9), which converges linearly. Denote by $\boldsymbol{B}(\beta)$ the above pseudo-likelihood form. Then the corresponding iterative procedure for maximizing $\boldsymbol{B}(\beta)$ is

$$\beta_{n+1} = \beta_n - P^{-1}(\partial/\partial\beta)\log[\boldsymbol{B}(\beta_n)], \tag{3.5.4}$$

where P is either a positive constant or a positive definite matrix. As in Chapter 2, the difficulty is the choice of P which influences the convergence rate. For (3.5.2), we can usually compute the second-order differential of $\log[\boldsymbol{B}(\beta)]$, so P can be chosen so that the iteration is the Newton-Raphson method, while, for (3.5.3), it is difficult to compute the second-order differential, so that the Newton-Raphson method is not available. However, when the exponential part of $p(x_i|x_{\partial i},\beta)$ is linear in β, the second differential of $-\mathrm{Log}[p(x_i|x_{\partial i},\beta)]$ is the conditional covariance matrix of the exponential part, given the neighbouring lines. Thus the second-order differential of $\mathrm{Log}[\boldsymbol{B}(\beta)]$ is usually a negative definite matrix. If it has a negative upper bound with respect to β, choosing P to be non-small and positive can ensure convergence.

Note that, if we combine two points such as $(X_{ij},X_{i+1,j})$ together as one point, Z_{ij}, say, then Z_{ij} has S^2 states. The conditional distribution of (X_i,X_{i+1}) can therefore be written as

$$p(X_i,X_{i+1}|X_{\partial(i,i+1)},\beta) = P(Z_{i1},Z_{i2},\ldots Z_{iN}|X_{\partial(i,i+1)},\beta). \tag{3.5.5}$$

It is clear that the above distribution is still a one-dimensional Markov random field, but each point has S^2 states. We can therefore consider another form of block pseudo-likelihood,

$$\prod_{1\leqslant i\leqslant M-1} p(X_i,X_{i+1}|X_{\partial(i,i+1)},\beta) \tag{3.5.6}$$

We refer to this as two-line pseudo-likelihood. We can also consider three-or-more-line pseudo-likelihoods, but the computational burden increases quickly as the number of lines increases.

Some simulation results are provided in Chapter 5, together with parameter estimation from noisy data. There, we generate images, estimate parameters, then add noise, and estimate parameters again, from noisy data.

3.6 Discussions

We have presented some theoretical and simulation results in this chapter. The explicit matrix expressions for partition functions are only obtained for simple cases, ie, the Ising models. For the two-colour case, Kaufmann(1949) obtained an eigenvalue expression, which enabled Pickard(1976) to examine the asymptotic properties of the model. For the multi-colour case, although we obtained similar results, as far as transition or the critical points is concerned, it may require theoretical results about finite matrix algebra and the representation of finite Lie groups, in order to get further properties associated with the partitions. So far as the Monte Carlo method for generating Markov random fields are concerned. we emphasize that the stochastic relaxation method is not the best method and that practical simulation are usually carried out under some restrictions. Maximum pseudo-likelihood estimation can be used for general MRFs, while some other methods are available only for special cases. Pseudo-likelihood functions are based on local conditional properties. However, for first-order MRF, for example, one pixel has four neighbouring pixels, while one line only has two neighbouring lines, therefore line (block) pseudo-likelihood uses fewer pixels as "condition" in the corresponding conditional distributions than does point pseudo-likelihood. It may therefore be more "close" to the true likelihood and thereby may provide more efficient estimates. Since pseudo-likelihood estimators are consistent, the variance(covariance) of block pseudo-likelihood estimators may be "smaller" than that of point pseudo-likelihood estimators.

Chapter 4

Normal Approximations For Lattice Systems

4.1 Introduction

In this chapter, we still consider Markov random fields on a rectangular lattice. Let $X=\{X_{ij}: 1\leqslant i\leqslant M, 1\leqslant j\leqslant N\}$ be an array of random variables on the M×N rectangular lattice, where $X_{ij}\in\{1,2,...S\}$. Throughout this chapter, we shall fix N. For i=1,2,...M, define $X_i=(X_{i1},X_{i2},...X_{iN})'$. Assume that $X=(X_1,X_2,...X_M)$ is distributed as a particular Markov random field with

$$P(X|\alpha,\beta) = \frac{1}{C(\alpha.\beta)}\exp\{\alpha'\sum_{i=1}^{M} f(X_i) + \beta'\sum_{i=1}^{M-1} g(X_i,X_{i+1})\} \qquad (4.1.1)$$

where the parameters α, β might be vectors, and f, g are vectors with the same dimension as α and β respectively. Model (4.1.1) is a general stationary Markov random field, stationary in the sense that the interaction terms are independent of location. Particular forms of f and g can represent first-order and second-order interaction MRFs. If we consider X_i to be a combination of two rows or more, (M would then be even or a multiple of some fixed integer), model (4.1.1) could represent a high-order MRF. β is the interaction parameter along the column direction (i.e., between rows), while α, in whole or in part, is that along the row direction. Strauss(1975), Saunders et al.(1979) and Kryscio et al.(1980) considered the asymptotic properties of the sample correlations of neighbouring pixels for binary lattice processes under the condition that all interaction parameters are zero or almost zero. Asymptotic normality properties were then used to obtain an approximate likelihood for parameter estimation (Possolo. 1986) and to compare the power of some tests for randomness(i.e., whether or not the pixels are independent)(Kryscio, et al., 1980).

Define

$$Y_M = \sum_{i=1}^{M} f(X_i); \qquad Z_M = \sum_{i=1}^{M-1} g(X_i.X_{i+1}).$$

In what follows, we use the central limit properties of finite-step-dependent stationary processes to show that the random variables

$(Y_M, Z_M)'$ have a normal limiting distribution when M becomes large, subject to conditions on the parameter β. We first handle the case where the rows are independent ($\beta=0$), and then extend to the case where β is proportional to the square root of the variance of Z_M. As in Possolo(1986), limit results are also used in relation to **maximum approximate- likelihood estimation(MALE)** for some particular cases. We discuss the problem of testing for the randomness of rows, and provide simulation results and comparisons with MALE based on the asymptotic normality results obtained under the condition that all interaction parameters are zero(Kryscio et al., 1980, Possolo, 1986).

4.2 Limiting result when β=0

Consider an independent chain $\{y_i\}$, i=1,2,..., where $y_i \in \{1,2,\ldots,T\}$, with $p_j=\Pr(y_i=j)$. Let $A(y_i,y_{i+1})$ be a finite function defined on $\{1.2,..T\} \times \{1,2,..T\}$, where A might be a vector. Define

$$W_M = \sum_{i=1}^{M-1} A(y_i,y_{i+1}).$$

We then have the following lemma.

<u>**Lemma**</u>: Suppose $0<p_j<1$, j=1,2,..T. Then $M^{-\frac{1}{2}}(W_M - EW_M)$ has a normal limiting distribution with zero mean and variance (or covariance matrix) V_W as $M \longrightarrow \infty$, where

$$V_W = EA_1A_1' + E(A_1A_2' + A_2A_1') - 3(EA_1)(EA_1)'$$

where $A_1 = A(y_1,y_2)$, $A_2 = A(y_2,y_3)$. #

Note that the process $\{A(y_i,y_{i+1})\}$ is a one-step-dependent stationary process. Furthermore, since A(*,*) has a finite sample space, $\{A(y_i,y_{i+1})\}$ is a first-order strictly stationary Markov chain. Standard methods can be used to prove that $(M-1)^{-\frac{1}{2}}(W_M - EW_M)$ is asymptotically normal, as $M \rightarrow \infty$. In fact, it can be shown by some long and tedious calculations that the moments of W_M converge to the moments of the limiting normal distribution, thus establishing the lemma (Moran, 1968). (Details of the proof of the lemma are omitted.)

When β=0, the rows in model (4.1.1) are independent and identically distributed. Each row can be considered as one point with S^N states. Therefore, two straightforward results for model (4.1.1) follow from the lemma and are presented in the following theorem.

Theorem 1: When $\beta=0$, then, as $M \longrightarrow \infty$,

1. $M^{-\frac{1}{2}}(Z_M - EZ_M)$ has a normal limiting distribution with zero mean and variance (or covariance matrix, if β is a vector).

$$V_Z = Egg' + E(g(X_1,X_2)g(X_2,X_3)' + g(X_2,X_3)g(X_1,X_2)') - 3(Eg)(Eg)' \tag{4.2.1}$$

2. $M^{-\frac{1}{2}}\begin{bmatrix} Y_M - EY_M \\ Z_M - EZ_M \end{bmatrix}$ has a multivariate normal limiting distribution with zero mean and covariance matrix $V = (v_{ij})$,

where $v_{11} = Eff' - (Ef)(Ef)'$;

$v_{22} = V_Z$;

$v_{12} = v_{21}' = E[f(X_1) + f(X_2) - 2Ef][g(X_1,X_2) - Eg]'$. #

4.3 Limiting results under close alternatives

In this section we consider the limiting distribution of Z_M under another condition, namely that $\beta=M^{-\frac{1}{2}}\beta_1$, where β_1 is a constant. Thus, under this alternative hypothesis, the interactions along columns are sufficiently small that the assumption of independence among rows is 'nearly' true.

Let $\mathcal{M}_M(t|\alpha)$ denote the moment generating function of $M^{-\frac{1}{2}}(Z_M - EZ_M)$ when $\beta=0$, and let $\mathcal{M}_M(t|\alpha,\beta)$ denote that in the case where $\beta\neq 0$. Then

$$\mathcal{M}_M(t|\alpha,\beta) = \sum_X \frac{1}{C(\alpha,\beta)} \exp\{\alpha Y_M + \beta Z_M\} \exp\{tM^{-\frac{1}{2}}(Z_M - EZ_M)\} \tag{4.3.1}$$

and

$$\mathcal{M}_M(t|\alpha) = \mathcal{M}_M(t|\alpha,0). \tag{4.3.2}$$

It is easy to find that:

$$\mathcal{M}_M(t|\alpha, M^{-\frac{1}{2}}\beta_1) = \mathcal{M}_M(t + \beta_1|\alpha)/\mathcal{M}_M(\beta_1|\alpha). \tag{4.3.3}$$

If the sequence of moment generating functions $\mathcal{M}_M(t|\alpha)$ converges to that of a normal distribution for all t, it follows from (4.3.3) that $M^{-\frac{1}{2}}(Z_M - EZ_M)$ has a normal limiting distribution under close alternatives, since the ratio of the limiting moment generating functions in (4.3.3) is also the generating function of a normal distribution. We proved in the previous section that $M^{-\frac{1}{2}}(Z_M - EZ_M)$

converges in distribution to a normal random variable. In general, however, convergence in distribution alone does not guarantee that the corresponding sequence of moment generating functions will converge to the moment generating function of the limiting random variable. Although some tedious calculations for the moments of $M^{-½}(Z_M-EZ_M)$ could be used to establish convergence of the sequence of moment generating functions, we use an argument similar to that of Kryscio et al.(1980). Saunders et al.(1979) showed that convergence in distribution implies convergence of moment generating functions provided the sequence of moment generating functions is uniformly bounded in an interval containing zero. Thus, to establish the asymptotically normal properties of Z_M under close alternatives it is sufficient to show that, if $\beta=0$, then there is a constant c>0, such that

$$E\exp\{tM^{-½}(Z_M - EZ_M)\} \leqslant c \qquad (4.3.4)$$

for $-\Delta \leqslant t \leqslant \Delta$ with $\Delta>0$.

Note that

$$Z_M-EZ_M = \sum_{i=1}^{[M/2]} [g(X_{2i-1},X_{2i})-Eg] + \sum_{i=1}^{[(M-1)/2]} [g(X_{2i},X_{2i+1})-Eg] \qquad (4.3.5)$$

where [a] denotes the integer part of a, and that, for each sum in (4.3.5), denoted by S_1 and S_2 respectively, elements are independent and identically distributed. Thus,

$$E\exp\{tM^{-½}(Z_M-EZ_M)\} \leqslant [E\exp\{2tM^{-½}S_1\}E\exp\{2tM^{-½}S_2\}]^{1/2}$$
$$= [E\exp\{2tM^{-½}(g-Eg)\}]^{M/2}.$$

Consider the joint density of two rows of variables, X_1 and X_2, with partition function $C_2(\alpha,\beta)$, ie

$$P(X_1.X_2|\alpha,\beta) = \frac{1}{C_2(\alpha,\beta)}\exp\{\alpha f(X_1) + \alpha f(X_2) + \beta g(X_1,X_2)\}.$$

Clearly, when α is fixed, $C_2(\alpha,*)$ is an analytic function. Note that $E\exp\{2tM^{-½}g(X_1,X_2)\} = C_2(\alpha,2tM^{-½})/C_2(\alpha,0)$. Thus

$$\log E\exp\{tM^{-½}(Z_M-EZ_M)\} \leqslant \frac{M}{2}[\log C_2(\alpha,2tM^{-½})-\log C_2(\alpha,0)-2tM^{-½}Eg].$$

Since $Eg=\frac{\partial}{\partial\beta}\log C_2(\alpha,\beta)\Big|_{\beta=0}$, application of simple analysis to the right-hand side of the above inequality shows that it is uniformly bounded, for t in any finite interval containing zero. We have therefore proved (4.3.4). Since

$$M_M(t+\beta_1|\alpha) \longrightarrow \exp\{½V_Z(t+\beta_1)^2\}, \quad \text{as } M\longrightarrow\infty$$

and

$$M_M(\beta_1|\alpha) \longrightarrow \exp\{½V_Z\beta_1{}^2\}, \quad \text{as } M\longrightarrow\infty,$$

we can summarize the above discussion as follows.

Theorem 2: When $\beta=M^{-½}\beta_1$, $M^{-½}(Z_M-EZ_M)$ has a normal limiting distribution with mean $V_Z\beta_1$ and variance V_Z, as $M\longrightarrow\infty$. #

Remark: In this and the previous section, the operator 'E' denotes expectation under the condition that $\beta=0$. If β is a multi-dimensional parameter, similar results hold, but with vector-matrix notation as appropriate.

Theorem 1 and Theorem 2 provide a basis for a statistical test of the null hypothesis H_0: $\beta=0$, against the alternative hypothesis H_a: $\beta=M^{-½}\beta_1$, $\beta_1\neq0$. If α is known, then the appropriate test of H_0 versus H_a is based on the statistic

$$T = M^{-½}V_Z{}^{-1}(Z_M - EZ_M).$$

By Theorem 1, this statistic has, under H_0, a standard normal limiting distribution, while, by Theorem 2, under H_a, it has a normal limiting distribution with variance 1 but with non-zero mean.

4.4 EZ_M and V_Z in particular cases

We now consider how to obtain EZ_M and V_Z. For convenience, assume β be a scalar parameter. When $\beta=0$, each row is independent of the others. Therefore,

$$P(X|\alpha,0) = \prod_{i=1}^{M} P_1(X_i|\alpha), \qquad (4.4.1)$$

where

$$P_1(X_i|\alpha) = \frac{1}{C_1(\alpha)}\exp\{\alpha f(X_i)\} \tag{4.4.2}$$

and $C_1(\alpha)$ is the partition function for one row. In Chapter 2, we developed a recursive technique for one-dimensional Markov random fields; see also Qian and Titterington(1989, 1990a) . This technique enables us to compute $C_1(\alpha)$ and $Eg(X_i,X_{i+1})$ for a variety of forms of f and g. In particular, suppose

$$f(X_i) = \sum_{j=1}^{N-1} \delta(X_{ij},X_{i,j+1}) \tag{4.4.3}$$

and

$$g(X_i,X_{i+1}) = \sum_{j=1}^{N} \delta(X_{ij},X_{i+1,j}), \tag{4.4.4}$$

where δ is the Kronecker delta function. This particular case corresponds to a first-order Markov random field which treats all colours equally. From Chapter 2, we know that, when $\beta=0$, each row is simply a strictly stationary Markov chain, with

$$\Pr(x_{ij},x_{i,j+1},..x_{i,j+t})=\frac{1}{S(e^{\alpha}+S-1)^t}\exp\{\alpha \sum_{\nu=j}^{j+t-1} \delta(x_{i\nu},x_{i,\nu+1})\}. \tag{4.4.5}$$

Clearly,

$$C_1(\alpha) = S(e^{\alpha} + S - 1)^{N-1} \tag{4.4.6}$$

and $$E\delta(x_{ij},x_{i+1,j}) = 1/S. \tag{4.4.7}$$

Thus, $EZ_M = N(M-1)/S$.

From Theorem 1, we know that

$$V_Z = a + 2b,$$

where $$a = E[g(X_1,X_2) - Eg]^2;$$

$$b= E[g(X_1,X_2) - Eg][g(X_2,X_3) - Eg].$$

Note that, for the particular case,

$$b = E\{E[g(X_1,X_2)|X_2]E[g(X_2,X_3)|X_2]\} - N^2/S^2$$

$$= E\{\frac{N}{S}\cdot\frac{N}{S}\} - \left[\frac{N}{S}\right]^2 = 0$$

Thus, although $g(X_1,X_2)$ and $g(X_2,X_3)$ are not independent, the correlation between them is zero. We therefore only need to know a.

Consider the random chain $\{\delta(x_{1i},x_{2i})\}$. It is clearly a strictly stationary first-order Markov chain with only two states {0,1} and

$$a = Nr_0 + 2 \sum_{i=1}^{N-1} (N-i)r_i, \qquad (4.4.8)$$

where $r_i = E[\delta(x_{11},x_{21})-E\delta][\delta(x_{1,1+i},x_{2,1+i})-E\delta]$, the i-th order autocorrelation of $\{\delta(x_{1i},x_{2i})\}$. In view of the properties of strictly stationary binary Markov chains, r_i takes the form (Appendix 2)

$$r_i = r_0\lambda^i, \qquad i=1,2,... \qquad (4.4.9)$$

Simple calculations from the density functions of (x_{11},x_{21}) and $(x_{11},x_{12},x_{21},x_{22})$ give

$$r_0 = (S-1)/S^2 \qquad (4.4.10)$$

and $$r_1 = \frac{e^{2a} + S - 1}{S(e^{\alpha} + S - 1)^2} - \frac{1}{S^2}. \qquad (4.4.11)$$

Thus, $$\lambda = \frac{S(e^{2\alpha}+S-1) - (e^{\alpha}+S-1)^2}{(S-1)(e^{\alpha}+S-1)^2}. \qquad (4.4.12)$$

We have now obtained EZ_M and V_Z for this particular case. Note that (4.4.9) is associated with the first-order strictly stationary Markov chain with only two states, and λ is the eigenvalue of its transition matrix that is different from unity. In general, if g has the form

$$g(X_i,X_{i+1}) = \sum_{j=1}^{N} d(x_{ij},x_{i-1.j}), \qquad (4.4.13)$$

where d is not now the Kronecker delta function (cf (4.4.4)), both a and b have a form similar to (4.4.8) (see Appendix) with

$$r_i = \sum_j \mu_j\lambda_j^{\,i}, \qquad (4.4.14)$$

where the $\{\lambda_j\}$ are the eigenvalues of the transition matrix of the corresponding first-order strictly stationary Markov chain $\{d(x_{1j},x_{2j})\}$, excluding the eigenvalue unity. The number of these eigenvalues is associated with the number of values that can be assumed by d. In some cases, the order of the corresponding Markov chain may be higher than 1. However, similar results could be also obtained by using formulae similar to (4.4.8) for higher order Markov chains. For example, if two points are combined as one, a second-order Markov chain with S states can be regarded as a first-order Markov chain with S^2 states. Thus, we only need to calculate several low-order autocorrelations in order to obtain V_Z. It should be pointed out that (4.4.3) is a special form which renders each row strictly stationary. For other forms of f, with the condition that each row is a finite-order, strictly stationary Markov chain, similar results for EZ_M and, especially, for V_Z, might also be obtained. For the calculation of EZ_M, it is not necessary to let each row be stationary.

4.5 MALE for Markov random fields

In this section. we consider the use of the approximate form of the partition function $C(\alpha,\beta)$ in parameter estimation. From (4.1.1),

$$C(\alpha,\beta) = \sum_X \exp\{\alpha \sum_{i=1}^{M} f(X_i) + \beta \sum_{i=1}^{M-1} g(X_i,X_{i+1})\}$$

$$= [C_1(\alpha)]^M \exp\{\beta(M-1)Eg\} \sum_X \exp\{\beta(Z_M-EZ_M)\} \prod_{i=1}^{M} P_1(X_i|\alpha)$$

$$= [C_1(\alpha)]^M \exp\{\beta(M-1)Eg\} \mathcal{M}_M(M^{\frac{1}{2}}\beta|\alpha) \qquad (4.5.1)$$

where $\mathcal{M}_M(*|\alpha)$, defined by (4.3.2), is the moment generating function of $M^{-\frac{1}{2}}(Z_M-EZ_M)$. Since we have proved the asymptotic normality property and the convergence of the sequence of the moment generating functions for $M^{-\frac{1}{2}}(Z_M-EZ_M)$, and although $M^{\frac{1}{2}}\beta$ is related to M (not fixed), we may still use $\exp\{\frac{1}{2}M\beta^2V_Z\}$ as an approximation to $\mathcal{M}_M(M^{\frac{1}{2}}\beta|\alpha)$. We therefore obtain the following approximate form of $C(\alpha,\beta)$:

$$\bar{C}(\alpha,\beta) = [C_1(\alpha)]^M \exp\{\beta(M-1)Eg\}\exp\{\tfrac{1}{2}M\beta^2V_Z\}. \qquad (4.5.2)$$

For the particular case described in (4.4.3) and (4.4.4), suppose both M and N are large enough for us to omit the difference between N and N-1. M and M-1. By noting that

$$V_Z = N\frac{S-1}{S^2}[1 + 2\sum_{i=1}^{N-1}(1 - \frac{i}{N})\lambda^i] \approx N\frac{(S-1)(1+\lambda)}{S^2(1-\lambda)} ,$$

we therefore have the approximate log-likelihood function

$$\frac{1}{MN}\{-\log\bar{C}(\alpha,\beta) + \alpha\sum_{i=1}^{M}\sum_{j=1}^{N-1}\delta(x_{ij},x_{i,j+1}) + \beta\sum_{i=1}^{M-1}\sum_{j=1}^{N}\delta(x_{ij},x_{i+1,j})\}$$

$$\approx \frac{1}{MN}\{\alpha\sum_{i=1}^{M}\sum_{j=1}^{N-1}\delta(x_{ij},x_{i,j+1}) + \beta\sum_{i=1}^{M-1}\sum_{j=1}^{N}\delta(x_{ij},x_{i+1,j})\}$$

$$- \log(e^{\alpha}+S-1) - \frac{\beta}{S} - \frac{(S-1)(1+\lambda)}{2S^2(1-\lambda)}\beta^2, \qquad (4.5.3)$$

where λ is defined by (4.4.12).

For the binary case (S=2) when $\alpha=\beta$, note that

$$P(X|\alpha) \approx \frac{1}{C(\alpha)}\exp\{-4\alpha\Sigma x_{ij} + 2\alpha\,\Sigma' x_{ij}x_{\mu\nu}\},$$

where Σ' denotes the sum over nearest neighbour pixels. and $x_{ij}\in\{0,1\}$, for all i,j. This is the first-order form of the density for binary MRF considered by Strauss(1975), Saunders et al.(1979), Kryscio et al.(1980) and Possolo(1986). Consider another density involving two parameters, namely,

$$P(X|u,v) = C(u,v)^{-1}\exp\{u\Sigma x_{ij} + vy\}, \qquad (4.5.4)$$

where $y=\Sigma' x_{ij}x_{\mu\nu}$. When v=0, all points are independent with the same distribution function. Bloemena(1964) and Kryscio et al.(1980) proved the asymptotic normality properties of y when v=0. The technique, which uses the moment generating function of the normal distribution to approximate the moment generating function of another random variable, was also used to obtain an approximate form of C(u,v)(Possolo, 1986). Bloemena(1964) gave explicit formulae for Ey and Var(y) under v=0. Omitting the lower-order terms, we have that, when v=0,

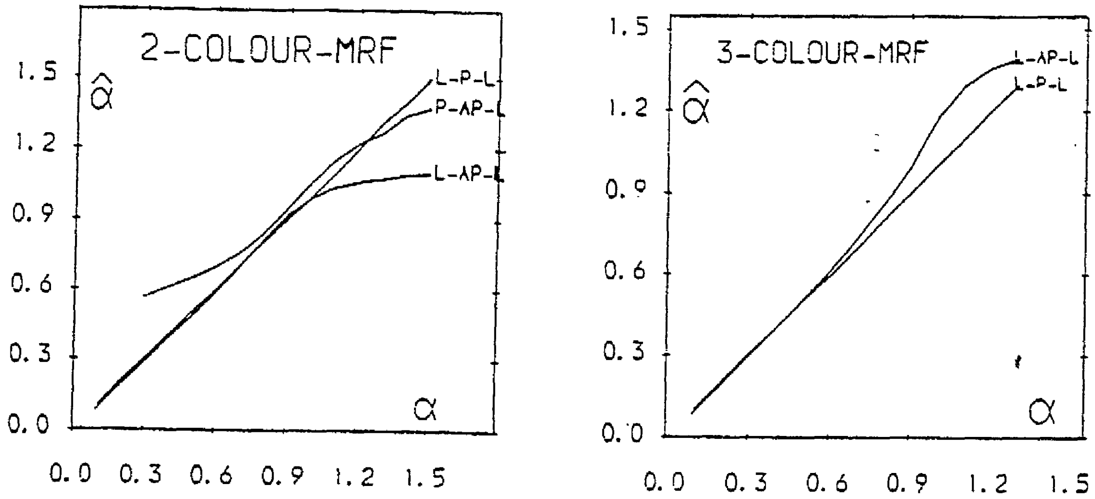
$$Ey = 2MN\theta^2$$

and $$Var(y) = 2mn\theta^2(1-\theta)(1+7\theta),$$

where $\theta=e^u/(1+e^u)$. We can hence obtain an approximate form of the log-likelihood of (4.5.4) as follows(see Possolo, 1986):

$$u\frac{1}{MN}\Sigma x_{ij}+v\frac{1}{MN}y-\log(1+e^{u})-2v\theta^{2}-v^{2}\theta^{2}(1-\theta)(1+7\theta). \qquad (4.5.5)$$

The results of a simulation study of parameter estimation based on the maximization of both approximate likelihoods, (4.5.3) and (4.5.5), for binary MRF involving only one parameter, namely, $\alpha=\beta$, is presented in Fig 4.1a. There are 128×128 pixels. The generating MRFs are simulated by the method of stochastic relaxation(Geman and Geman, 1984). Although there is no guarantee for convergence of the simulation procedure for MRFs, by noting the simulation results in the previous chapter, we know that, after 20 sweeps of relaxation, the generated patterns are very 'close' to the MRFs, especially in the cases where the parameter is less than the critical point. Denote by L-P-L the line pseudo-likelihood estimation method which we described in the last chapter (see Qian and Titterington(1989, 1990e)), while P-AP-L and L-AP-L denote the maximum approximate likelihood estimates from (4.5.5) and (4.5.3) respectively. ((4.5.5) and (4.5.3) are based on the assumptions of point-independence and line-independence, respectively.) Fig 4.1b represents some results for three colour MRFs. In both Fig 4.1a and Fig 4.1b, the values are means of 10 samples for each value of the parameter.



(a) (b)

Fig 4.1 Simulation results of MALEs for two and three colour MRFs

From the figure, we know that the maximum approximate likelihood estimates are near the true values only in a small region. Although we find that L-AP-L is better than P-AP-L, the quality of each approximation depends upon β and M. Since $\mathcal{M}_M(t|\alpha)$ does not converge to $\exp\{\frac{1}{2}t^2V_Z\}$ uniformly for $-\infty<t<\infty$, whereas $M^{\frac{1}{2}}\beta$ is infinite when $M\to\infty$, the practice of using $\exp\{\frac{1}{2}M\beta^2V_Z\}$ to approximate $\mathcal{M}_M(M^{\frac{1}{2}}\beta|\alpha)$ could cause totally different behaviours.

In conclusion, we may remark that the MALE can only be used for a small region of parameters and that similar results might also be obtained for the conditional distribution of Z_M given Y_M under the condition that $Y_M=\Sigma f(X_i)$ is known. Since (Y_M,Z_M) are jointly sufficient for (α,β), the asymptotic results about the conditional distribution of Z_M under Y_M can provide an approach to the statistical testing problem in the case where α is unknown. The idea of the "line" normal approximation described in this chapter is valid for a large range of models; the difficulty is the calculation of limiting means and variances or covariances. There is the same difficulty in the "point" normal approximation. It is known that the lemma in Section 2 holds under much weaker conditions, so the results may be obtained for a variety of models.

Chapter 5

Parameter Estimation For Hidden Markov Random Fields

5.1 Introduction

We have examined some problems for Markov random fields in Chapter 3. In practice, however, a random field itself is usually not observed. Instead, a blurred and/or noisy version of it is observed. This is the case that we consider in this chapter. We mainly concentrate on the problem of parameter estimation.

To use Markov random fields as priors in image analysis is a new and active subject in recent years. Cross and Jain(1983) provided various patterns of texture images which were simulated from MRF models. These simulated images were very realistic. Geman and Geman(1984) proposed simulated annealing methods to obtain **Maximum a posteriori(MAP)** estimation for image restoration from noisy data, but the computational problem is enoumous. Alternatively, Besag(1986) proposed the **Iterated conditional modes(ICM)** method, which concentrates on the local dependence structure of MRFs, and produces restored images very cheaply and quickly; see also Glendinning(1989). Jubb and Jennison(1988) suggested a modification of ICM, which extends the range of ICM to very noisy images and greatly reduces computational costs. Another modification, called Iterated Conditional expectation(ICE), was proposed by Owen(1986, 1989). Chellappa(1985) and Derin and Elliott (1987) also examined the use of Gibbs distributions for texture images. For parameter estimation, Besag (1986) proposed an iterative procedure for simultaneous parameter estimation and image restoration, based on his Coding technique. Kay and Titterington(1986) pointed out the difficulties involved in the EM algorithm in the context of MRFs. Geman and McClure(1985) presented a method of parameter estimation, based on the EM algorithm and the Monte Carlo technique of generating MRFs, where only one parameter is involved in the prior distribution function. Chalmond (1988) used a so-called Gibbsian EM algorithm, computing some posterior mathematical expectations by the Monte Carlo technique. Younes(1988b) generalized his own iterative technique(Younes, 1989), which also use the Monte Carlo technique, to the case of imperfectly observed Gibbs fields.

which includes the situation with noisy data. Frigessi and Piccioni(1988a, 1988b) examined consistent parameter estimation using the moment method, but only for the two-colour Ising model corrupted by noise. Simulation study was used by Thompson et al(1990) to examine methods of choosing smoothing parameter in the two-dimensional smoothing problem. Further relevant work appears in Geman and Graffigne(1988).

For the ICM method, we can carry out both asynchronous updating and synchronous updating. For the latter case, practical computing environment or special computer languages can be used, so that the computation is fast. However, as mentioned in Besag(1986), updating is most conveniently implemented as a raster scan, and in that case, it converges faster in term of one cycle. Besag's(1986) experiments for the case with one parameter, β, involved in the MRF model, where β represents the interaction between neighbouring pixels, showed that when β increases, the convergence rate decreases. He therefore took β increasing during the ICM procedure. In Section 5.2, we consider the case of continuous intensities, and try to describe these phenomena from a mathematical viewpoint. We will also show the different behaviours of asynchronous updating and synchronous updating, by giving a counter example.

In section 5.3, we discuss the difficulties of the EM algorithm for multi-dimensional Markov random fields. We have shown that, for one-dimensional version of MRFs, both E-step and M-step can be carried out. Due to extremely large computational demands, it is infeasible to do the same thing for multi-dimensional cases. We will also, for the auto-normal case, show the different bahaviour between Besag's(1986) iterative procedure of simultaneous parameter estimation and restoration, and the iterative procedure of the EM algorithm.

In section 5.4, we develop Besag's(1986) iterative procedure for simultaneous parameter estimation and restoration. The procedure is based on a restoration method, such as ICM, and a modified EM algorithm which, in each cycle, uses partially the observed data and partially the image restored in the last cycle. In a similar way to pseudo-likelihood estimation, the modified EM algorithm is also based on local conditional densities; for instance, conditional distributions of one line or two-lines of pixels. Numerical results and some further discussion are presented in section 5.5.

5.2 Discussions on the ICM method

In this section, we use a single subscript to represent a pixel, whether the image is one or multi-dimensional. Suppose there are L pixels which are labelled in some manner by the integer i=1,2,...L. Thus, the pixels are ordered. Let $x^{(k)}=(x_1{}^{(k)}, x_2{}^{(k)},\ldots x_L{}^{(k)})'$ be the estimate of the true scene, x^*, at the k-th cycle of iteration, and $x^{(k+1)}$ at the (k+1)-th cycle estimation. The ICM is based on the local posterior probability $p[x_i|y, x_{\partial i}]$, where Y=y are observed image.

Using asynchronous updating means choosing $x_i{}^{(k+1)}$ by maximizing

$$p[x_i|y,x_1{}^{(k+1)},\ldots x_{i-1}{}^{(k+1)},x_{i+1}{}^{(k)},\ldots x_L{}^{(k)}] \qquad (5.2.1)$$

for each i at each cycle of iteration. Hence, it can be ensured that

$$p[x^{(k+1)}|y] \geqslant p[x^{(k)}|y]. \qquad (5.2.2)$$

The above inequality ensures convergence to a local maximum point. For synchronous updating, $x_i{}^{(k+1)}$ is chosen by maximizing

$$p[x_i|y,x_1{}^{(k)},\ldots x_{i-1}{}^{(k)}, x_{i+1}{}^{(k)},\ldots,x_L{}^{(k)}]. \qquad (5.2.3)$$

(5.2.3) can only ensure that

$$p[x_1{}^{(k)},\ldots x_{i-1}{}^{(k)},x_i{}^{(k+1)},x_{i+1}{}^{(k)}\ldots x_L{}^{(k)}|y]$$
$$\geqslant p[x_1{}^{(k)},\ldots x_{i-1}{}^{(k)},x_i{}^{(k)},x_{i+1}{}^{(k)},\ldots x_L{}^{(k)}|y]$$

for each i, but (5.2.2) cannot necessarily be obtained. That means that convergence is not guaranteed.

Now consider the continuous-intensity case, where

$$p(x) \propto \exp\{-\tfrac{1}{2}x'Qx\} \qquad (5.2.4)$$

where $Q=\{b_{ij}\}$ is a positive definite matrix. This is slightly different from the form given in Chapter 1; see also Besag(1986). Re-write Q as

$$Q = \Lambda - B_1 - B_2 \qquad (5.2.5)$$

where B_1 is a lower-triangular matrix, B_2 is upper-triangular, and both B_1 and B_2 have zero diagonal entries, with

$$B_1 = B_2'$$

For simplicity, we assume $\Lambda= \lambda I$ and that the observations $y=\{y_i\}$

are independent Gaussian records with mean x^* and variance $1/\kappa$. The MAP estimate of x^* is

$$\hat{x} = \kappa(\kappa I + Q)^{-1}y. \tag{5.2.6}$$

Synchronous updating can be expressed in matrix form as follows:

$$x^{(k+1)} = \frac{1}{\kappa + \lambda}[\kappa y + (B_1 + B_2)x^{(k)}] \tag{5.2.7}$$

while asynchronous updating can be written as

$$x^{(k+1)} = \frac{1}{\kappa + \lambda}[\kappa y + B_1 x^{(k+1)} + B_2 x^{(k)}] \tag{5.2.8}$$

or, in another form, as

$$x^{(k+1)} = [(\kappa+\lambda)I - B_1]^{-1}[\kappa y + B_2 x^{(k)}]. \tag{5.2.9}$$

For the former case, convergence depends on the norm of matrix $A_1=(\kappa+\lambda)^{-1}[B_1+B_2]$, while, for the latter, it depends on that of $A_2=[(\kappa+\lambda)I-B_1]^{-1}B_2$. Note that parameter λ represents, in some respect, the interaction between neighbour pixels. The 'bigger λ, the less tendency there is that neighbouring pixels have same or similar colours, that is the less is the interaction. We would like to remark that the notation here is slightly different from that in Besag(1986), where λ and κ could be regarded as the inverse values of λ and κ here, respectively. We can also note that, in both cases, when λ increases, the convergence speed of the iterations increases as well. This phenomenon corresponds to that the parameter has controversy relation with converging speed in discrete MRF models(Besag, 1986). We shall first show that $\tilde{\rho}(A_2) = \max|\alpha_i| < 1$, where α_i are eigenvalues of A_2. That means that, for any λ and κ, asynchronous updating converges. If η_i is the eigenvector corresponding to α_i, then,

$$A_2\eta_i = \alpha_i\eta_i$$

thus,

$$B_2\eta_i = \alpha_i[(\kappa+\lambda)I - B_1]\eta_i.$$

Since $\eta_i{}'B_2\eta_i = \eta_i{}'B_1\eta_i = ½\eta_i{}'(B_1 + B_2)\eta_i$,

$$\alpha_i = \frac{\lambda\eta_i{}'\eta_i - \eta_i{}'Q\eta_i}{(2\kappa+\lambda)\eta_i{}'\eta_i + \eta_i{}'Q\eta_i} \tag{5.2.10}$$

and since Q is a positive definite matrix, it follows that $|\alpha_i|<1$. Now consider matrix A_1. Note that the eigenvalues of A_1 can be written as

$$\alpha=(\lambda-\beta)/(\kappa+\lambda), \qquad (5.2.11)$$

where β are eigenvalues of the matrix Q. Although Q is positive definite, $\beta>0$, for all β, which can only ensure $\alpha<1$. It is possible that, for some β, $\alpha<-1$. In that case, updating is not convergent for some starting points $x^{(1)}$.

Example: $L=3, \lambda=1$,

$$Q = \begin{bmatrix} 1 & -0.75 & 0.25 \\ -0.75 & 1 & -0.75 \\ 0.25 & -0.75 & 1 \end{bmatrix}$$

The eigenvalues of Q are 0.75, $\overline{(9}-\sqrt{73})/8$ and $\overline{(9}+\sqrt{73})/8$, Q is therefore positive definite. The smallest eigenvalue of A_1 is

$$-(\sqrt{73}+1)/[8(\kappa+1)].$$

When κ is small, the above value is smaller than -1. Θ

Therefore it cannot be guaranteed that synchronous updating always converges. However, (5.2.7) and (5.2.9) are only two iterative methods for solving the linear equation (5.2.6). There are certainly some other methods. we can consider, for instance, the following iterative method:

$$x^{(k+1)} = \frac{1}{\kappa + c}[\kappa y + (cI - Q)x^{(k)}], \qquad (5.2.12)$$

where c is a positive number. Clearly, it is equivalent to

$$x_i^{(k+1)} = \frac{1}{\kappa+c}[\kappa y_i + (c-\lambda)x_i^{(k)} + \sum_{j\neq i} b_{ij}x_j^{(k)}], \qquad (5.2.13)$$

for each i.

Note that (5.2.7) is a particular case of (5.2.12) with $c=\lambda$. The convergence of (5.2.12) depends on $(c-\beta)/(\kappa+c)$, where β are still eigenvalues of the matrix Q. Obviously, when c is big enough, $(c-\beta)/(\kappa+c)$ lies between -1 and 1, and the iteration therefore converges.

Even if (5.2.6) itself converges, in circumstances when we know the eigenvalues of Q we can still choose a number c to get faster convergence. The best choice of c is

$$c = (\beta_{min} + \beta_{max})/2$$

then $$\rho\{(cI-Q)/(\kappa+\lambda)\} = (\beta_{max}-\beta_{min})/(2\kappa+\beta_{max}+\beta_{min}).$$

Note that we only use $x_j^{(k)}$, $j\neq i$, in (5.2.6), but use $x_i^{(k)}$ as well in (5.2.12). Note that the original image x^* is represented by its noisy version, ie, the observed data, y. I am quite sure that as the iteration proceeds, the restored image contains more and more information about the original image x^* from the observation y. At each cycle of iteration, the newly obtained image gets information not only from y but also from the image estimated in the last cycle. (5.2.12) uses the information contained in $x_i^{(k)}$, so it may get more information than (5.2.6), with the result that it converges faster. For asynchronous updating, we can also modify the algorithm in a similar way to get faster convergence.

Although the synchronous updating formula (5.2.6) is not very practical, we can still conclude that, for the continuous intensity case, some modification involving the use of $x_i^{(k)}$ in the replacement for x_i at the (k+1)-th cycle can increase the converging speed.

For discrete Markov random fields, although it is not certain that synchronous updating converges, it is difficult to provide similar discussion. We would like to pose a question to finish this section.

Can we choose $x_i^{(k+1)}$, by using $x_i^{(k)}$ together with states at all other pixels or at its neighbouring pixels, to ensure faster convergence or get better restoration?

5.3 Difficulties in the EM algorithm

The recursive techniques in Chapter 2 enable us to carry out the EM algorithm for general one-dimensional versions of Markov random fields. Two groups of vectors are used there. If the number of states at each point is not large, it is practical and not difficult to implement these recursive computations. For two-dimensional Markov random fields, we have mentioned in Chapter 3 and Chapter 4 that they can be regarded as one-dimensional versions if we regard one row or a set of neighbouring rows as one point. However, the number of possible states at such a point is now S^N, or even higher, where S is the number of possible states or colours at each pixel, and N is the number of columns. Even if S is equal to 2, it is impractical to compute two S^N-dimensional vectors for each row. It is therefore infeasible to carry out similar recursive techniques for multi-dimensional Markov random fields themselves or such fields corrupted by noisy data.

The prior distribution considered in this chapter is of Gibbs form. It is known that the conditional distribution, given the observed image, is still a Gibbs distribution, although the corresponding neighbourhood system may change. It is therefore still of exponential distribution form. If the exponential part is linear in the parameters, the conditional mean and variance of the exponent, given the observed image, depend only on the normalizing factor, ie, the partition function, which is now related to the observed data. From Section 1.4, we know that the computation of the conditional mean of the exponential part is just the E-step of the EM algorithm. Chapter 3 has pointed out difficulties in this sort of computation. Although the Monte Carlo technique of generating samples from Gibbs distributions can be used, the resulting computational demands are also very substantial, so the Monte Carlo technique for the E-step is not an completely satisfatory approach.

As mentioned in Chapter 3, there is no explicit method for maximizing the likelihood function associated with a realisation of a Markov random field. We also know that the M-step of the EM algorithm is in fact equivalent to maximizing the joint distribution of the original image and the observed image, as if the original image were observed. This can usually be regarded as consisting of two parts. The first is to maximize the prior distribution, and the second is to maximize the noise model. Since the noise models are currently supposed to be relatively simple, it is not very difficult to achieve the second maximization. Chapter 3 pointed out the difficulties involving in parameter estimation for Markov random fields. Although Monte Carlo methods can be used here too, they are computationally time consuming. Therefore, even we can compute the conditional mean of the exponential part, ie, carry out the E-step, it is still not easy to do the M-step.

It is therefore impractical and very difficult to implement the EM algorithm directly from the original distribution and the noise model. Although the Monte Carlo method can be used for both steps, we have to examine alternative or modified methods. Chalmond(1988) applied the EM procedure to a point pseudo-likelihood together with the Monte Carlo technique. Note that, if the original image is partially observed, for instance(see the next section, and also Qian and Titterington, 1989). if all even rows are known, we can, by using the recursive technique for Gibbs chains, carry out the EM algorithm for the conditional distribution of all odd rows, with all image data at odd rows replaced by noisy data. We shall therefore examine the application of the EM algorithm to pseudo-likelihood functions in the next section.

For the case of auto-normal continuous intensities, the situation is different. By similar notation to that in the previous section, let x_i denote value at a single pixel, and the density of x be

$$p(x|\beta) = |Q|^{\frac{1}{2}} \cdot \exp\{-x'Qx/2\}/(2\pi)^{L/2} \qquad (5.3.1)$$

where Q, an L×L matrix, depends on the parameter β. Suppose the noisy is additive white noisy with conditional density

$$f(y|x,\sigma^2) = \prod_{i=1}^{L} f_1(y_i|x_i,\sigma^2) \qquad (5.3.2)$$

with

$$f_1(y_i|x_i,\sigma^2) = \exp[-(y_i-x_i)^2/(2\sigma^2)]/(2\pi\sigma^2)^{\frac{1}{2}} \qquad (5.3.3)$$

where σ^2 is an unknown parameter. The corresponding E-Step and M-Step are then as follows.

E-Step: We must compute,

$$x^{(k)} = E(x|y,\beta^{(k)},\sigma^{2(k)}) \qquad (5.3.4)$$

and

$$V^{(k)} = E(xx'|y,\beta^{(k)},\sigma^{2(k)}) \qquad (5.3.5)$$

It is easy to show that

$$x^{(k)} = (I+\sigma^{2(k)}Q^{(k)})^{-1} \cdot y \qquad (5.3.6)$$

and

$$V^{(k)} = \sigma^{2(k)}(I+\sigma^{2(k)}Q^{(k)})^{-1}. \qquad (5.3.7)$$

M-Step: $\beta^{(k+1)}$ is obtained by maximizing

$$-(x^{(k)\prime}Qx^{(k)} + tr(Q \cdot V^{(k)})) - \log[|Q|] \qquad (5.3.8)$$

and

$$\sigma^{2(k+1)} = [(y-x^{(k)})'(y-x^{(k)}) + tr(V^{(k)})]/L \qquad (5.3.9)$$

where tr(*) denotes the trace of the matrix.

The maximization of (5.3.8) is dictated by the dependence of Q on β. For simple models, it is not difficult in theory. Suppose $\beta^{(k)}$ and $\sigma^{2(k)}$ converge to β^* and σ^{2*}, respectively, they then satisfy

$$x^* = (I + \sigma^{2*}Q^*)^{-1}y \tag{5.3.10}$$

and $$V^* = \sigma^{2*}(I+ \sigma^{2*}Q^*)^{-1} \tag{5.3.11}$$

with the corresponding β^* maximizing

$$-(x^{*\prime}Qx^* + tr(Q\cdot V^*)) - \log[|Q|], \tag{5.3.12}$$

also $$\sigma^{2*} =[(y-x^*)'(y-x^*)+Tr(V^*)]/L \tag{5.3.13}$$

The method of estimation within ICM in Besag(1986) goes as follows. A single cycle of ICM is performed with the current parameters $\beta^{(k)}$ and $\sigma^{2(k)}$, giving a new $x^{(k)}$. This $x^{(k)}$ is treated as "known" and provides updated estimates $\beta^{(k+1)}$ and $\sigma^{2(k+1)}$. However, the ICM method just gives (5.3.6). This is equivalent, at every step, to taking $V^{(k)}$ as zero in (5.3.8) and (5.3.9)) to obtain the new values of β and σ^2. We therefore know that, if convergence obtains for the values of β and σ^2, the limiting values $\beta^{\#}$ and $\sigma^{2\#}$ satisfy

$$x\# = (I+\sigma^{2\#}Q^{\#})^{-1}y \tag{5.3.14}$$

$$\sigma^{2\#} = (y-x^{\#})'(y-x^{\#})/L, \tag{5.3.15}$$

with $\beta^{\#}$ maximizing

$$-x^{\#\prime}Qx^{\#} - \log|Q|. \tag{5.3.16}$$

It is obvious that the limiting values are different from those associated with EM algorithm, since generally $V^{(k)}$, the conditional covariance matrix of x, given y, is not zero. Therefore, the estimators in the ICM procedure without the EM algorithm may be biased. From our simulated results in the next section, we will also find that for discrete MRFs, the estimators in the ICM procedure are also biased.

5.4 Simultaneous parameter estimation and restoration

5.4.1 Basic idea

It is known that maximum pseudo-likelihood estimation is currently a useful method for Markov random fields. Consider the first-order case and pixel (ij) with its four neighbours. The conditional density of X_{ij}, given all other pixels, depending only on these neighbours, can be written as

$$p(X_{ij}|X_{i,j-1},X_{i-1,j},X_{i,j+1},X_{i+1,j},\beta). \qquad (5.4.1)$$

The point pseudo-likelihood is just a product of a set of such conditional densities. We only consider a single one here. Imagine $X_{i,j-1}$, $X_{i-1,j}$, $X_{i,j+1}$ and $X_{i+1,j}$ are four known constants and that β, θ are unknown parameters. Maximum pseudo-likelihood estimation involves maximizing (5.4.1). If X_{ij} is missing, but alternatively, y_{ij} is observed with conditional density $f(Y_{ij}|X_{ij},\theta)$, our aim is to maximize

$$Pr(Y_{ij}=y_{ij}|X_{i,j-1},X_{i-1,j},X_{i,j+1},X_{i+1,j},\beta,\theta) \qquad (5.4.2)$$

The EM algorithm can be used for this purpose and is not difficult to carry out. However, it is impractical to assume that $X_{i,j-1}$, $X_{i-1,j}$, $X_{i,j+1}$ ans $X_{i+1,j}$ are observed; they are usually missing together with X_{ij}. One solution to this problem is to use estimated values to replace them. It is natural to use restored image data. Since restoration depends on the parameters, we propose an iterative procedure, ie, restoration, then estimation, then restoration again, and so on. In the rest of this section we shall develop this idea in detail.

5.4.2 The case of $\{X_{2i}\}$ known

In this subsection, for simplicity, we concentrate on the first-order pairwise interaction model with distribution in (1.3.3). We suppose the exponential part is linear in the parameter β.

We consider the case where $\{X_{2i-1}, i=1,2,...M/2\}$ are missing, but $\{X_{2i}, i=1,2,...M/2\}$ are observed. Although this case is not realistic in practice, our aim is to illustrate the application of the EM algorithm to conditional distributions. We assume M is even; if M is odd, the results are very similar. We also assume the conditional density $f(Y_{ij}|X_{ij},\theta)$ be of linear-exponential form, proportional to $\exp\{\theta h(Y_{ij},X_{ij})\}$, say. Consider the following two conditional likelihoods

$$f(Y=y,\{X_{2i-1}\}=\{x_{2i-1}\}|\{X_{2i}\}=\{x_{2i}\},\beta,\theta)$$

$$= \prod_{i=1}^{M/2} p(x_{2i-1}|x_{2i-2},x_{2i},\beta) \prod_{i=1}^{M} f(y_i|x_i,\theta) \qquad (5.4.3)$$

and

$$f(Y=y|\{X_{2i}\}=\{x_{2i}\},\beta,\theta) = \sum_{\{x_{2i-1}\}} f(y,\{x_{2i-1}\}|\{x_{2i}\},\beta.\theta), \qquad (5.4.4)$$

where $p(x_i|x_{i-1},x_{i+1},\beta)$

$$= \frac{1}{C(x_{i-1},x_{i+1},\beta)} \exp\{\beta \sum_{j=1}^{N} \bar{g}_{ij}(x_{ij}) + \beta \sum_{j=1}^{N-1} G_{[ij][i,j+1]}(x_{ij},x_{i,j+1})\},$$

$$(5.4.5)$$

with $\bar{g}_{ij}(x_{ij}) = g_{ij}(x_{ij}) + G_{[ij][i-1,j]}(x_{ij},x_{i-1,j})$

$$+ G_{[ij][i+1,j]}(x_{ij},x_{i+1,j}). \qquad (5.4.6)$$

and

$$f(y_i|x_i,\theta) \propto \exp\{\theta \sum_{j=1}^{N} h(y_{ij},x_{ij})\}.$$

If we ignore the conditioning $\{X_{2i}\}$, maximization of (5.4.3) is just maximization of two exponential distributions for β and θ respectively, while maximization of (5.4.4) is that with $\{X_{2i-1}\}$ as incomplete data. We therefore discuss how to use the EM algorithm to maximize $f(y|\{X_{2i}\},\beta,\theta)$ in this subsection. Geman and Geman(1984) proved that the conditional distribution of x given y is still of Gibbs form. For the case we discuss here, the conditional distribution is also still of pairwise interaction form. Note that

$$f(\{X_{2i-1}\}|\{x_{2i-1}\},y,\beta,\theta) = \prod_{i=1}^{M/2} f(X_{2i-1}|x_{2i-2},x_{2i},y_{2i-1},\beta,\theta),$$

where f for each row corresponds to a first-order Gibbs chain. To be precise,

$$f(X_i=x_i|x_{i-1},x_{i+1},y_i,\beta,\theta) =$$

$$\frac{1}{C(x_{i-1},x_{i+1}.y_i,\beta,\theta)} \exp(\sum_{j=1}^{N} \breve{g}_{ij}(x_{ij}) + \beta \sum_{j=1}^{N-1} G_{[ij][i,j+1]}(x_{ij},x_{i,j+1}))$$

where $\breve{g}_{ij}(x_{ij}) = \beta\bar{g}_{ij}(x_{ij}) + \theta h(y_{ij}.x_{ij}).$

Thus, the corresponding E-step and M-step for maximizing

$f(y|\{x_{2i}\},\beta,\theta)$ are as follows.

E-step: Assume that $\beta^{(k)}$, $\theta^{(k)}$ are the current values. Note that both $f(X_{2i-1}|x_{2i-2},x_{2i},y_{2i-1},\beta,\theta)$ and $f(X_{2i-1}|x_{2i-2},x_{2i},\beta)$ are of Gibbs chain form, and that $f(Y_{ij}=y_{ij}|X_{ij},\theta)$ is exponential, so that we only need to compute

$$E(\bar{g}_{2i-1,j}(X_{2i-1,j})|x_{2i-2},x_{2i},y_{2i-1},\beta^{(k)},\theta^{(k)}),$$

$$E(G_{[2i-1,j][2i-1,j+1]}(X_{2i-1,j},X_{2i-1,j+1})|x_{2i-2},x_{2i},y_{2i-1},\beta^{(k)},\theta^{(k)}),$$

and

$$E(h(y_{2i-1,j},X_{2i-1,j})|x_{2i-2},x_{2i},y_{2i-1},\beta^{(k)},\theta^{(k)})$$

for all possible i and j.

M-Step: It is easy to see that the maximization can be regarded in two parts, namely, for β and θ respectively. For β, it is almost the same as line pseudo-likelihood, as illustrated in Chapter 3. We have to use the recursive technique developed in Chapter 2, the only difference being that we deal simultaneously with a number of chains with the same parameter. For θ, it depends on the precise structure of $h(Y_{ij},X_{ij})$. When Y_{ij} is normal with mean X_{ij} and veriance σ^2, it is easy to obtain new value for σ^2.

5.4.3 The case where $\{x_i\}$ are all missing

In practice, $\{x_{2i}\}$ are missing as well as $\{x_{2i-1}\}$. As pointed out in Section 5.3, if we try to maximize $f(y|\theta,\sigma^2)$, there will be many difficulties. It is natural to replace x_{2i} by their estimates, and we therefore propose the following iterative procedure.

(1). Obtain initial estimates of both parameters β and θ, and an initial estimate $\{\hat{x}_{2i}\}$ of $\{X_{2i}\}$. (The Maximum likelihood classification or another better restoration method such as ICM can be used to obtain an estimate of X; see Besag, 1986).

(2). Concentrate on $f(y|\{\hat{x}_{2i}\},\theta,\sigma^2)$, and use one or several cycles of the E-Step and M-Step described in the last subsection, to obtain new values of β and θ.

(3). Re-obtain $\{\hat{x}_{2i}\}$. A cheap way is to apply one or several cycles of ICM(Besag,1986) to all the pixels, with the parameters estimated in step (2).

(4). Return to (2) for a fixed number of cycles or until convergence seems to have occurred.

We know little about the theoretical properties of the above algorithm, such as whether or not it is convergent. We use the estimates $\{\hat{x}_{2i}\}$ as if they were true, and one problem may arise, namely, that the parameter estimaters may be biased. In (3), to re-obtain new estimates for all pixels means to reconstruct the image simultaneously with parameter estimation.

5.4.4 Other pseudo-likelihood approaches with the EM algorithm

Besag's(1986) iterative procedure of simultaneous parameter estimation and image restoration is as follows: restore the image, $\hat{x}$, from noisy data, y, then maximize a pseudo-likelihood to obtain new estimates, $\hat{\beta}$, from the restored image, $\hat{x}$. ($f(y|\hat{x},\theta)$ is maximized to obtain a new value $\hat{\theta}$.) $\hat{\beta}$ and $\hat{\theta}$ are then used for restoration in the next cycle. Note the difference between our procedure and Besag's. At each cycle of procedure, we use the EM algorithm to maximize a single conditional likelihood, say, $\Pi f(Y|\{\hat{x}_{2i}\},\beta,\theta)$, while in Besag's, two (conditional) likelihood functions, namely, $\Pi p(\hat{x}_{ij}|\hat{x}_{\partial ij},\beta)$ and $f(y|\hat{x},\theta)$, are maximized. Although the estimated image is used in both procedures, our simulation showed that their behaviours are quite different.

The EM algorithm in the above iterative procedure is only applied to the product of conditional distributions of all odd rows. It is easy to note that it can be extended to the product over all rows, if all pixels are restored at each cycle. We refer to it as LPL-EM, in correspondence with line pseudo-likelihood. Also note that the EM algorithm can be used to maximize $f(y_i,y_{i+1}|x_{\partial(i,i+1)},\beta,\theta)$, provided that $X_{\partial(i,i+1)}=x_{\partial(i,i+1)}$ is known. Denote by 2LPL-EM the corresponding iterative procedure in which two-line pseudo-likelihood function is used. (See Chapter 3.) Finally, $\Pi p(y_{ij}|\hat{x}_{\partial ij},\beta,\theta)$ is more easy to maximize by the EM algorithm, where the product is over all pixels. We refer to the associated iterative procedure as PPL-EM. Obviously, the procedure can usually be used with all sorts of pseudo-likelihoods. Since the restored image is determined by the observed image, the noisy image contains more information about the original image and the unknown parameters, so using the EM algorithm may result in better parameter estimation.

5.5 Simulation studies and discussion

The results from the procedure outlined in Section 5.4.3 are given in Table 5.1 and Table 5.2, but they correspond to second-order MRFs with only one parameter β, over a 64×64 lattice. We used 30 cycles of row-by-row relaxation to generate random fields, and imposed normal noise with different variances. The choice of X_{2i}, i=1,...M/2, is by ICM updating for all the pixels. The results, with complete data, with $\{X_{2i}\}$ known and with $\{X_i\}$ all missing are compared in the tables. The numbers of simulated random fields are also gven in the tables. Of course, the method leads to a restoration of the original image. From it and the estimation method for complete data, we obtain another set of parameter estimates, given in the tables as the fourth set of results.

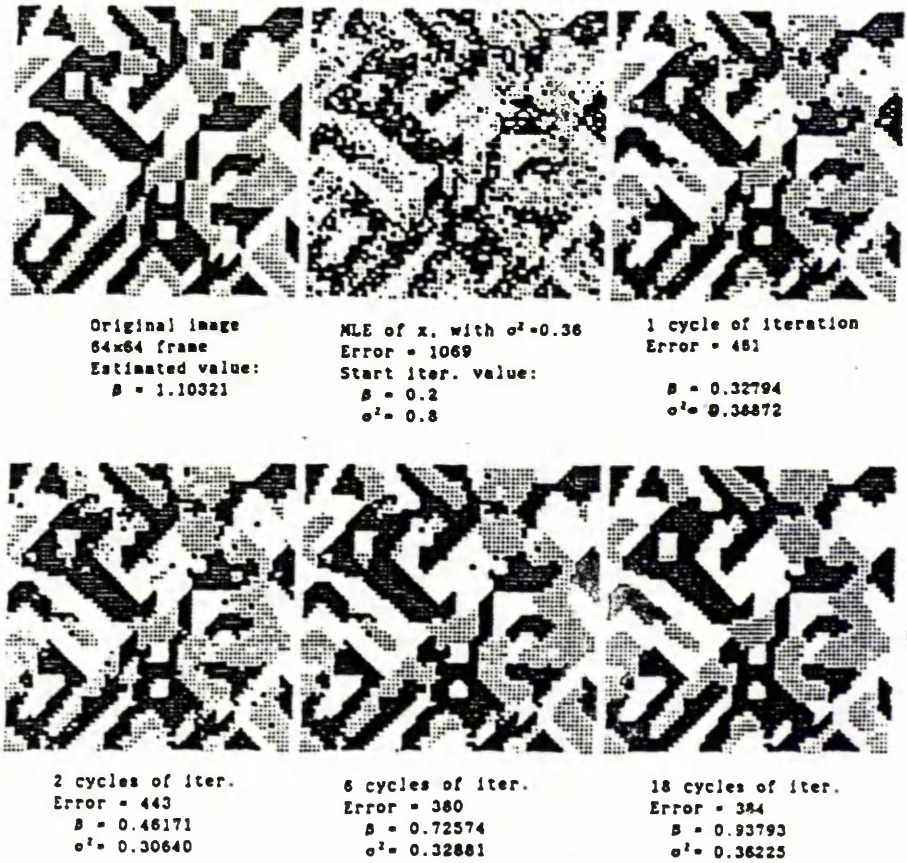
1. the variance of noise is 0.36, number of fields: 50				
	β		σ^2	
	N-Mean	N-Vari.	N-Mean	N-vari.
Complete data	0.49766	0.00048	0.35938	0.00008
$\{X_{2i}\}$ true	0.50366	0.00113	0.35921	0.00009
$\{X_i\}$ missing	0.49664	0.00210	0.37134	0.00012
	0.75911	0.01101	0.37742	0.00012
2. The variance of noise is 0.16, number of field: 30				
complete data	0.49933	0.00039	0.15971	0.00002
$\{X_{2i}\}$ true	0.50354	0.00088	0.15980	0.00002
$\{X_i\}$ missing	0.49151	0.00091	0.16159	0.00005
	0.57735	0.00279	0.16250	0.00006
3. The variance of noise is 0.04, number of fields: 30				
complete data	0.49933	0.00039	0.03996	0.00000
$\{X_{2i}\}$ true	0.49842	0.00043	0.03998	0.00000
$\{X_i\}$ missing	0.49842	0.00043	0.03991	0.00000
	0.50131	0.00042	0.03985	0.00000

Table 5.1 Parameter estimation for two-colour MRFs

Simulation results show that with different parameters, the differences between images restored by ICM, are relatively small, so Besag's(1986) iteratve procedure, if it converges, should be very similar in performance to the fourth set of results. Note that those results show positive biases. When the variance of the noise is small, the restored image is almost the same as the original one, so all the sets of results are almost equally good. The reason for this may be that the maximum probability restoration has the tendency to make the probability, with which the neighbouring pixels have different colours, become smaller than the true probability. In other words, a pixel with colour different from that of its neighbouring pixels, might be smoothed to the same colour as its neighbours after restoration, so that the estimated parameter within the MRF model becomes bigger.

1. Variance of Y_{ij}: 0.36, number of fields: 10				
	β		σ^2	
	N-Mean	N-Vari.	N-Mean	N-Vari.
complete data	0.49148	0.00044	0.35873	0.00001
$\{X_{2i}\}$ true	0.48671	0.00054	0.35912	0.00005
$\{X_i\}$ missing	0.46746	0.00039	0.34840	0.00006
	0.58595	0.00064	0.34345	0.00007
2. variance of Y_{ij}: 0.16 Number of fields: 10				
complete data	0.49148	0.00044	0.15932	0.00000
$\{X_{2i}\}$ known	0.48840	0.00040	0.15931	0.00001
$\{X_i\}$ missing	0.47810	0.00032	0.15484	0.00002
	0.52180	0.00037	0.15053	0.00002
3. Variance of y_{ij}: 0.04, number of fields: 10				
complete data	0.48148	0.00044	0.03969	0.00000
$\{X_{2i}\}$ true	0.49022	0.00039	0.03975	0.00000
$\{X_i\}$ missing	0.48987	0.00040	0.03953	0.00000
	0.49287	0.00044	0.03928	0.00000

Table 5.2 Parameter estimation for three-colour MRFs



Original image
64x64 frame
Estimated value:
β = 1.10321

MLE of x, with σ^2 =0.36
Error = 1069
Start iter. value:
β = 0.2
σ^2 = 0.8

1 cycle of iteration
Error = 481
β = 0.32794
σ^2 = 0.38872

2 cycles of iter.
Error = 443
β = 0.46171
σ^2 = 0.30640

6 cycles of iter.
Error = 380
β = 0.72574
σ^2 = 0.32881

18 cycles of iter.
Error = 384
β = 0.93793
σ^2 = 0.36225

Fig 5.1 Iterative procedure for an artifical images

Fig 5.1 provides images at several steps of the procedure for an artificial image, together with the estimated parameters and the error rates at those steps. We used a second-order model. Note, in particular, that, although the last two images are very similar in terms of error rate, the estimates of β are quite different. The final estimates of σ^2 are very close to the true value.

Fig 5.2 provides results of Besag's(1986) iterative procedure of simultaneous parameter estimation and restoration. The underlying binary images were simulated by the point relaxation method, under the Ising model with one parameter $\alpha=\beta$. We added white noise with zero mean and variance $\sigma=0.5$. For each value of α, 40 replicates were simulated. The straight lines in the figure indicate the true values. We note that the estimated parameters, $\hat{\alpha}$, are larger than the true values, with the implication that the restored images are oversmoothed. The reason for this may be that the restored images contain less information than the noisy images. The application of the EM algorithm to pseudo-likelihood functions may enable us to get more information from the observed images. Fig 5.3 presents results of PPL-EM and LPL-EM for the same model as in Fig 5.2. LPL denotes estimation from the originally simulated fields, using one-line pseudo-likelihood. The average LPL is the almost straight line in the figure, which is very close to the true value. In Fig 5.3(a), the middle line from $\alpha=0.3$ to 1.0 and the bottom one from $\alpha=1.0$ to 1.5 display the results of LPL-EM, we then see that LPL-EM is slightly better than PPL-EM. However, both of them represent considerable improvements on the results presented in Fig 5.2. The top two figures of Fig 5.4 give similar results for 2LPL-EM, where 2LPL also denotes parameter estimates from the data originally simulated but now by the two-line pseudo-likelihood method. Fig 5.4 is based on only 20 samples for each α. Note that the behaviours of 2LPL and 2LPL-EM are almost the same. The bottom two figures of Fig 5.4 show the sample variance of the above estimated parameters. $\hat{\sigma}^2$ shows similar behaviour, while the variance of $\hat{\alpha}$ increases quite quickly as α increases. However, even for α equal to 1.5, which is much bigger than the critical point of about 0.88, the sample variance is still not very large.

For all the above simulations, we used the method of Golden-Section-Search to find estimates of the single parameter $\alpha(\beta)$.

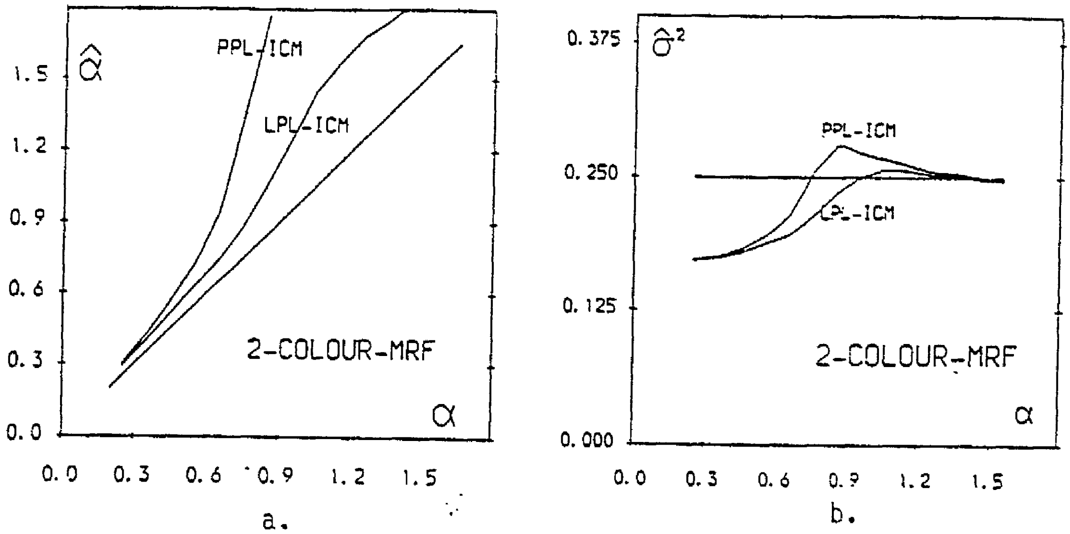


Fig 5.2 Parameter estimation without the Modified EM algorithm

We use ICM for restoration in our simulation. It is known that ICM only converges to a local maximum point. However, the pseudo-likelihood functions together with the EM algorithm, especially based on the two-line pseudo-likelihood, result in very good parameter estimation. Therefore we conclude in this chapter that, whatever method is used for restoration, the modified EM algorithm is very useful for the parameter estimation.

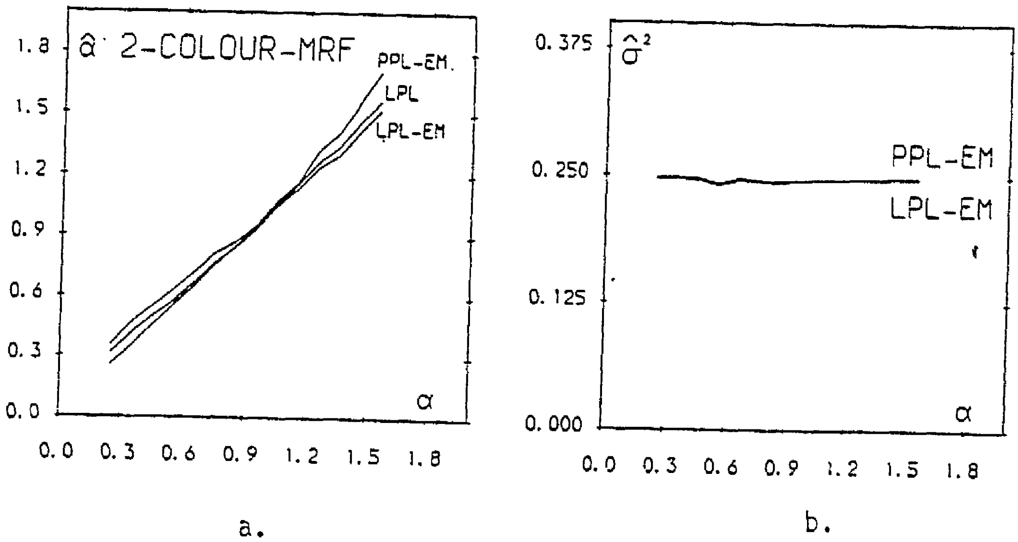


Fig 5.3 Parameter estimation with the modified EM algorithm

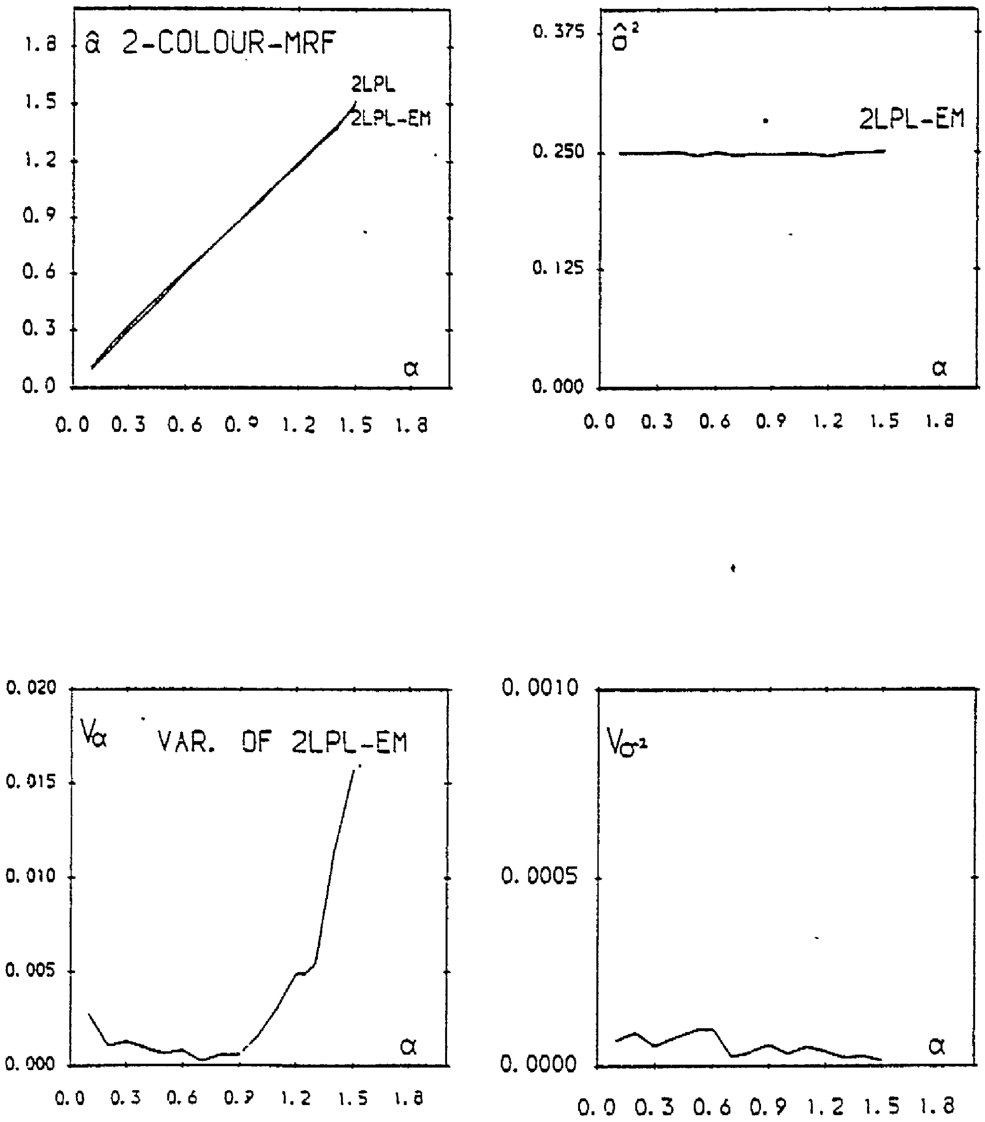


Fig 5.4 Results by two-line EM algorithm and sample variances

Chapter 6
Three-Dimensional Markov Mesh Models

6.1. Introduction

Although Markov random fields can be thought of as a two-dimensional version of the one-dimensional concept of a Markov chain, they, in general, differ from Markov chains in an important respect, in that they are not easily simulated. Realizations of Markov chains can be simulated by a single pass along the (one-dimensional) set of sites (Chapter 2), but this is not true of Markov random fields. However, for a **Markov Mesh Random Field (MMRF) model**, which, as mentioned in Chapter 1, is in fact a causally-dependent MRF and also a sort of generalization of a Markov chain, single pass simulation can be achieved. In the two-dimensional case this can be done by a raster scan.

In principle, there is no fundamental need for the set of sites in the random field models to correspond to a two-dimensional lattice of pixels. This chapter considers the three dimensional case of MMRF models. The model is a natural generalization of the two-dimensional version of MMRF. We only consider the case where the only available data is a noisy version of the true scene. We shall effectively present a direct analogue of the "two-dimensional" paper by Lacroix(1987). The closeness of the analogy will become clear as this chapter develops. Instead of pixels, the sites are a set of volume elements or **voxels**. Such images are also common in practice in the study of materials and in medical imaging.

Markov models for two-dimensional scenes are important because of the need to represent (spatial) contextual association among neighbouring pixels. Their extension to genuinely three-dimensional versions is crucial, particularly in the above contexts, in which interframe, as well as intraframe associations must be modelled. The MMRF models we discuss are comparatively general and are not formulated for special three-dimensional phenomena such as short-range motion across a sequence of two-dimensional frames.

For Markov random field models, the conditional distribution, given the observed noisy image, is still that of a Markov random field, with a slightly changed neighbourhood system, but it is not true of MMRF models. It is almost always computationally impossible to seek Maximum a Posteriori(MAP) restoration for MMRF models. Some sort of approximation is required, and a modified criterion is then used for the estimation of the true scene; see Devijver(1988) and Lacroix(1987) for two-dimensional cases.

In Section 6.2 the Markov Mesh models and assumptions are specified and in Section 6.3 the algorithm is derived. It takes the generalized form of the F-G-H algorithm of Devijver(1988); see also Lacroix(1987). It will be emphasised in Section 6.3 that simplifying assumptions are necessary, in order to creat a practicable algorithm. The treatment of boundary voxels is also specified there. Experimental results are presented in Section 6.4. In Section 6.5, we discuss the problem of parameter estimation from noisy data. An iterative procedure similar to that in Chapter 5, together with the modified EM algorithm, can be used. We present some simulation results but only for the two-dimensional case. Section 6.6 contains a brief discussion of conclusions and possible further research.

It should be mentioned here that notation in this chapter are different from those in the rest of this thesis.

6.2. The hidden Markov Mesh random field(MMRF) model

In this section we establish the basic assumptions underlying our Markov Mesh models for three-dimensional scenes and for the generation of noisy versions thereof. The true scenes are described by a lattice of voxel labels, denoted generically by λ, and the observed images by a corresponding set of feature vectors, for which the letter x is used. The development and detailed notations parallel very closely the two-dimensional version in Lacroix(1987).

6.2.1 The MMRF model

In notation very similar to that in Lacroix(1987), let

* V_{MNL} = { (m,n,ℓ), $0 \leqslant m \leqslant M$ rows, $0 \leqslant n \leqslant N$ columns, $0 \leqslant \ell \leqslant L$ layers} be a finite integer lattice;

* (a,b,c) be a voxel at the intersection of row a, column b and layer c;

* λ_{abc} and x_{abc} be the label and feature vector, respectively, at voxel (a,b,c), with each λ_{abc} taking a value from a finite set of

"colours", where x_{abc} could also be one-dimensional;

* V_{abc} denote the rectangular parallelopiped array of voxels depicted in Fig 6.1, with $V_{abc}^{(\lambda)}$ and $V_{abc}^{(x)}$ as the corresponding arrays of labels and feature vectors.

We shall also use the superscripts (λ) and (x) to denote arrays of labels and feature vectors associated with other sets of voxels. For the concept of the **Past** of voxel (a,b,c), we use the generalization of the definition introduced in Abend et al(1965) and used later in Kanal(1980) and Lacroix(1987) for the two-dimensional case. The **past** of voxel (a,b,c) is the set $\{(m,n,\ell)$: $m<a$ or $n<b$ or $\ell<c\}$.

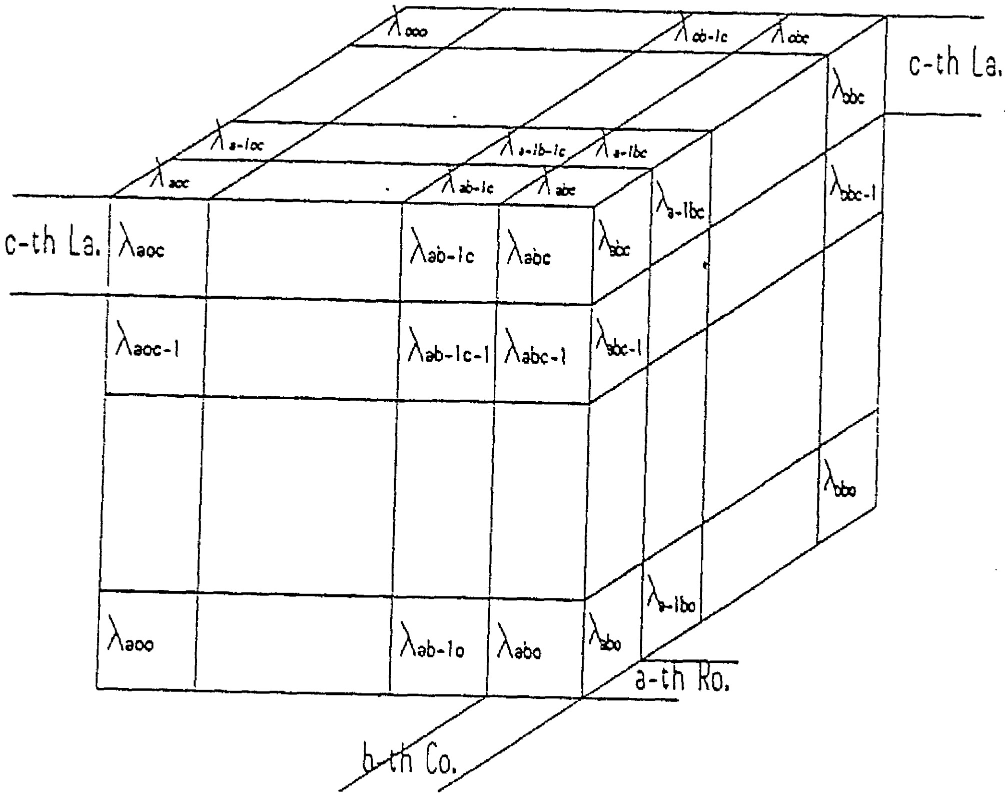


Fig 6.1 Definition of V_{abc} and $V_{abc}^{(\lambda)}$

The model chosen for $V_{MNL}^{(\lambda)}$ is that of a homogeneous, third-order Markov Mesh, thereby creating the natural analogue of the second-order process used by Lacroix(1987) for the two-dimensional case.

<u>Third-order Markov Mesh</u>: The Markovian assumption for this model is that

$$P[\lambda_{abc}|\{\lambda_{mn\ell} : m<a \text{ or } n<b \text{ or } \ell<c\}] = P[\lambda_{abc}|V_{abc}{}^{(\lambda)}\backslash\lambda_{abc}]$$

$$= P[\lambda_{abc}|\lambda_{a-1,bc}, \lambda_{a,b-1,c}, \lambda_{ab,c-1}]. \qquad (6.2.1)$$

The term "third-order" reflects the fact that the probabilities are conditional on three λ's.

Versions of this for the boundaries are obvious and will not be written out in detail.

<u>**Homogeneous**</u>: We now impose further simplifying assumption on (6.2.1). Let S denote the finite state space for the labels. Then for $0<a\leqslant M$, $0<b\leqslant N$, $0<c\leqslant L$ and for q, r, s, $t\in S$,

$$P[\lambda_{abc}=q|\lambda_{a-1,bc}=r, \lambda_{a,b-1,c}=s, \lambda_{ab,c-1}=t] = P_{q|rst},$$

independent of a, b and c. For the boundary conditions, which are also to be considered homogeneous, we make the natural assumptions that

$$P_q = P[\lambda_{ooo}=q] \qquad \text{(initial voxel)}$$

$$P_{q|r\square\square} = P[\lambda_{aoo}=q|\lambda_{a-1,oo}=r] \qquad \text{(initial Row axis, } a\geqslant1)$$

$$P_{q|\square s\square} = P[\lambda_{obo}=q|\lambda_{o,b-1,o}=s] \qquad \text{(initial Column axis, } b\geqslant1)$$

$$P_{q|\square\square t} = P[\lambda_{ooc}=q|\lambda_{oo,c-1}=t] \qquad \text{(initial Layer axis, } c\geqslant1)$$

$$P_{q|rs\square} = P[\lambda_{abo}=q|\lambda_{a-1,bo}=r, \lambda_{a,b-1,o}=s] \text{ (initial RC-plane, } a\geqslant1,b\geqslant1)$$

$$P_{q|\square st} = P[\lambda_{obc}=q|\lambda_{o,b-1,c}=s, \lambda_{ob,c-1}=t] \text{ (initial CL-plane, } b\geqslant1,c\geqslant1)$$

$$P_{q|r\square t} = P[\lambda_{aoc}=q|\lambda_{a-1,oc}=r, \lambda_{ao,c-1}=t] \text{ (initial LR-plane, } a\geqslant1,c\geqslant1)$$

These assumptions lead to the following properties for the process.

Property 1. For any $(a,b,c)\in V_{MNL}$,

$$P[V_{abc}{}^{(\lambda)}] = \prod_{m=0}^{a} \prod_{n=0}^{b} \prod_{\ell=0}^{c} P[\lambda_{mn\ell}|\lambda_{m-1,n\ell}, \lambda_{m,n-1,\ell}, \lambda_{mn,\ell-1}]. \quad (6.2.2)$$

(Thus, the joint probability function of the labels on V_{abc} factorises into a simple form reminiscent of "independence" models. It is easy to handle such a single (conditional) density, which looks like a local conditional distribution of a Markov random field. Therefore, similarly to the case of pseudo-likelihood for MRFs, it is usually not difficult to get likelihood estimates for the unknown parameter(s) involved in the model, from observed $V_{MNL}{}^{(\lambda)}$, directly or by an iterative procedure.)

Property 2. For any $(a,b,c)\in V_{MNL}$,

$$P[\lambda_{abc}|\{(\lambda_{mn\ell}: (m,n,\ell)\neq(a,b,c)\}] = P[\lambda_{abc}|\{\lambda_{mn\ell}: (m,n,\ell)\in\eta(abc)\}] \qquad (6.2.3)$$

$$\text{where} \quad \eta(abc) = \left\{\begin{array}{ccc} (a-1,b,c) & (a,b-1,c) & (a,b,c-1) \\ (a+1,b,c) & (a,b+1,c) & (a,b,c+1) \\ (a+1,b-1,c) & (a-1,b+1,c) & (a-1,b,c+1) \\ (a+1,b,c-1) & (a,b+1,c-1) & (a,b-1,c+1) \end{array}\right]$$

with appropriate modifications at the boundaries.

Property 3. The rows (resp. columns, layers) of a homogeneous MMRF form a stationary vector Markov chain of dimension (N+1)(L+1) (resp. (L+1)(M+1), (M+1)(N+1)).

6.2.2 The noise model

As in Devijver(1988) and Lacroix(1977) and previous chapters, we assume that the noise variables on different voxels are conditionally independent, given $V_{MNL}^{(\lambda)}$, and that each feature vector has the same conditional probability function $f(x_{abc}|\lambda_{abc})$, dependent only on λ_{abc}. Thus

$$P[V_{abc}^{(x)}|V_{abc}^{(\lambda)}] = \prod_{m=0}^{a}\prod_{n=0}^{b}\prod_{\ell=0}^{c} f(x_{mn\ell}|\lambda_{mn\ell}). \qquad (6.2.4)$$

6.2.3 Statement of the problem

We first assume that all the parameters in the MMRF model and in the noise model are known. In Section 6.5, we shall discuss the problem of parameter esimation from noisy data. Our problem is now to identify the original array $V_{MNL}^{(\lambda)}$ from $V_{MNL}^{(x)}$, the latter representing the available data, so that the original array is hidden by the noise. The approach we adopt for the labelling is different from that we used for Markov random fields, where MAP is used for the entire scene. In fact, we also use the Maximum a Posteriori(MAP) estimate, but for each pixel, conditionally on only part of the array of feature vectors: to be specific, for $(a,b,c)\in V_{MNL}$, we choose $\hat{\lambda}_{abc}=q$ where q comes from the rule:

$$P[\lambda_{abc}=q|V_{abc}^{(x)}] = \max_r P[\lambda_{abc}=r|V_{abc}^{(x)}]. \qquad (6.2.5)$$

This is in contrast with the usual practice with non-causal MRFs, where MAP estimation is undertaking conditionally on **all** feature data (Besag, 1986). One would attempt to choose $\hat{\lambda}_{abc}=q$ by

$$P[\lambda_{abc}=q|V_{MNL}^{(x)}] = \max_r P[\lambda_{abc}=r|V_{MNL}^{(x)}]. \qquad (6.2.6)$$

In the contexts of both MRF and MMRF models, implementation of (6.2.6) is very complicated. With MMRF models, the hope is that one can implement (6.2.5), or a plausible approximation thereof, using a single pass through the data. Clearly, the situation would be much simpler if the right-hand side of (6.2.6) were $P[\lambda_{abc}=r|V_{abc}^{(x)}]$, for then one can substitute from (6.2.1), using the labellings already established for voxels (a-1,b,c), (a,b-1,c) and (a,b,c-1), thereby easily finding q. Since, unfortunately, only x is available and since the true labels are spatially correlated, the exact expression even for $P[\lambda_{abc}=r|V_{abc}^{(x)}]$, when expanded in terms of voxel labels on V_{abc}, will turn out to be complicated.

6.3 Generalized F-G-H algorithm

For the two-dimensional case Devijver(1988) proposed the so-called F-G-H algorithm, and this procedure was generalized in Lacroix(1987). The general principle of the algorithm is to begin at one corner of the frame and to work diagonally downwards, labelling the pixels in the new "diagonals" as they enter into the pass. As remarked at the end of Section 6.2.3, the construction of a recursive procedure to effect a labelling following this type of pass will not feasible, computationally, without making approximations; these will be determined in terms of a "hypothesis", to be specified in Sectin 6.3.2.

In section 6.3.1 we introduce some notation, in Section 6.3.2 we derive the basic features of the algorithm itself, and, in Section 6.3.3, we discuss how to deal with the voxels at or close to the initial axes or planes, in other words, the boundary voxels.

6.3.1 Some notation

First we introduce the "**past of order h**", V_{abc}^{h}, and the "**diagonal of order h**", D_{abc}^{h}, for voxel (a,b,c).

We define

$$V_{abc}^{h} = \{(m,n,\ell): 0\le m\le a,\ 0\le n\le b,\ 0\le \ell\le c,\ m+n+\ell\le a+b+c-h\} \qquad (6.3.1)$$

and

$$D_{abc}^{h} = \{(a-i,b-j+i,c-h+j),\ 0\le i\le j\le h\} \qquad (6.3.2)$$

where we assume $h<\min(a,b,c)$. The boundary cases corresponding to $a\le h$ and/or $b\le h$ and/or $c\le h$, will be treated in Section 6.3.3.

The physical meanings of, and relationships among, these sets are depicted in Fig 6.2, which shows V_{abc}^{o}, D_{abc}^{o}, V_{abc}^{1}, D_{abc}^{1}, V_{abc}^{2} and D_{abc}^{2}. Note that $V_{abc}^{h}\backslash D_{abc}^{h}$ is a subset of all the past of all voxels of D_{abc}^{h} in V_{abc}. We also have $V_{abc}^{o}=V_{abc}$, and D_{abc}^{o} consists only of voxel (a,b,c).

Define

$$\Psi_i(D_{abc}^{i}(\lambda)) = P[D_{abc}^{i}(\lambda)\,|\,V_{abc}^{i}(x)]. \qquad (6.3.3)$$

Our problem is to maximize $\Psi_o(D_{abc}^{o}(\lambda))$, ie $\Psi_o(q)$. (Fig 6.2 may be helpful for us to understand the definition of functions Ψ_i.)

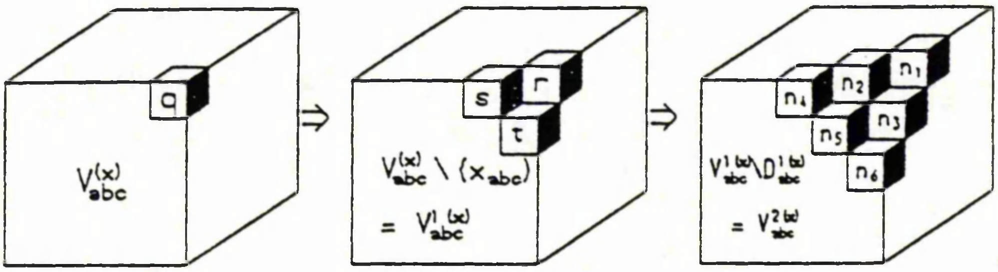


Fig 6.2 Definition of $V_{abc}^{h}(x)$ and $D_{abc}^{h}(\lambda)$

The cornerstones of the F-G-H algorithm are a "local decomposition relationship" and a lattice "recurrence relationship". The latter is associated with particular partitionings of V_{abc}^{i} and D_{abc}^{i}, and we discuss these next.

Fig 6.3 depicts the required partitioning of V_{abc}^{h} for $h\geq3$. The major element of the partition is $V_{a-1,b-1,c-1}^{h-3}$, while the other six components can be divided into two kinds, denoted by $A_{abc(*)}^{h}$ and $B_{abc(*)}^{h}$, where *=R,C,L refers to row, column or layer respectively. The A's are face and B's are edges, as Fig 6.3 demonstrates. There are various ways of relating and manipulating these components, as follows.

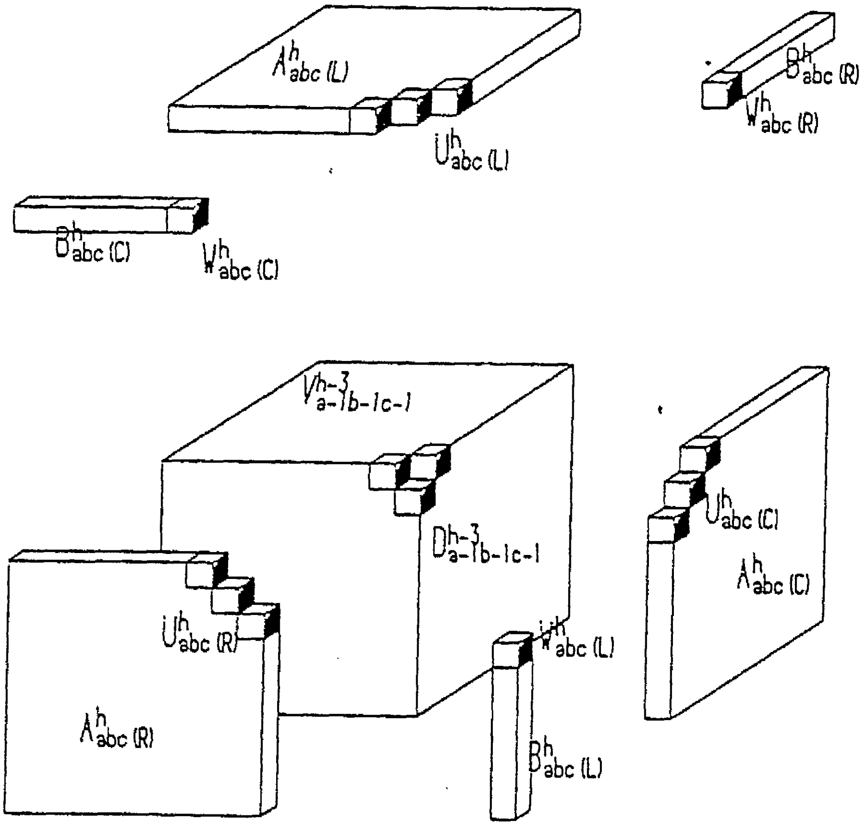


Fig 6.3 The decomposition of V_{abc}^{h} and D_{abc}^{h}

Decomposition relationships for V_{abc}^{h} ($h\geq3$)

Equations (6.3.4) and (6.3.5) describe two types of partitioning.

$$\left.\begin{aligned}
A_{abc(R)}^{h} + B_{abc(L)}^{h} + A_{abc(C)}^{h} + V_{a-1,b-1,c-1}^{h-3} &= V_{ab,c-1}^{h-1} \\
A_{abc(C)}^{h} + B_{abc(R)}^{h} + A_{abc(L)}^{h} + V_{a-1,b-1,c-1}^{h-3} &= V_{a-1,bc}^{h-1} \\
A_{abc(L)}^{h} + B_{abc(C)}^{h} + A_{abc(R)}^{h} + V_{a-1,b-1,c-1}^{h-3} &= V_{a,b-1,c}^{h-1}
\end{aligned}\right\} \qquad (6.3.4)$$

$$\left.\begin{aligned} A^{h}_{abc(R)} + V^{h-3}_{a-1,b-1,c-1} &= A^{h-2}_{a,b-1,c-1} \\ A^{h}_{abc(C)} + V^{h-3}_{a-1,b-1,c-1} &= A^{h-2}_{a-1,b,c-1} \\ A^{h}_{abc(L)} + V^{h-3}_{a-1,b-1,c-1} &= A^{h-2}_{a-1,b-1,c} \end{aligned}\right\} \qquad (6.3.5)$$

In these equations "+" denotes set-union operation.

Note that similar decomposition relationships hold for D^{h}_{abc}. These involve subsets of D^{h}_{abc}, to be called $U^{h}_{abc(*)}$ and $W^{h}_{abc(*)}$, that are direct analogues of $A^{h}_{abc(*)}$ and $B^{h}_{abc(*)}$, respectively. The decomposition relationships for D^{h}_{abc} (h≥3) are identical to those for V^{h}_{abc}, except that we replace V by D, A by U and B by W.

For example (c.f. (6.3.4)),

$$U^{h}_{abc(R)} + W^{h}_{abc(L)} + U^{h}_{abc(C)} + D^{h-3}_{a-1,b-1,c-1} = D^{h-1}_{ab,c-1}$$

and (c.f. (6.3.5)),

$$U^{h}_{abc(R)} + D^{h-3}_{a-1,b-1,c-1} = D^{h-2}_{a,b-1,c-1}.$$

$U^{h}_{abc(*)}$ and $W^{h}_{abc(*)}$ are also shown in Fig 6.3.

6.3.2 The algorithm

The algorithm is founded on two principle features, a <u>local decomposition relationship</u>(LDR) and a workable <u>recurrence relationship</u>(RR). In order to construct the (RR) from the (LDR) a simplifying hypothesis has to be imposed.

The Local Decomposition Relationship(LDR)

The (LDR) express $\Psi_i(D^{i}_{abc}(\lambda))$ in terms of corresponding functions $\Psi_{i+1}(D^{i+1}_{abc}(\lambda))$, related to D^{i+1}_{abc}, which is the diagonal of next highest order. Precisely, for i≤h-1, where, h<min(a,b,c):

(LDR) $\Psi_i(D_{abc}(\lambda)) =$

$$\frac{1}{N_i} P[D^{i}_{abc}(x) \,|\, D^{i}_{abc}(\lambda)] \sum_{D^{i+1}_{abc}(\lambda)} \Psi_{i+1}(D^{i+1}_{abc}(\lambda)) P[D^{i}_{abc}(\lambda) \,|\, D^{i+1}_{abc}(\lambda)], \qquad (6.3.6)$$

where $N_i = P[V^{i+1}_{abc}(x)]/P[V^{i}_{abc}(x)]$, is a normalizing number.

Proof of (LDR):

$$P[D^{i}_{abc}(\lambda),\ V^{i}_{abc}(x)] = \sum_{D^{i+1}_{abc}(\lambda)} P[D^{i}_{abc}(\lambda),\ D^{i+1}_{abc}(\lambda),\ V^{i+1}_{abc}(x),\ D^{i}_{abc}(x)]$$

$$= \sum_{D^{i+1}_{abc}(\lambda)} P[D^{i+1}_{abc}(\lambda),\ V^{i+1}_{abc}(x)]P[D^{i}_{abc}(\lambda)|D^{i+1}_{abc}(\lambda)]P[D^{i}_{abc}(x)|D^{i}_{abc}(\lambda)]$$

$$= P[D^{i+1}_{abc}(x)] \sum_{D^{i+1}_{abc}(\lambda)} P[D^{i+1}_{abc}(\lambda)|V^{i+1}_{abc}(x)]P[D^{i}_{abc}(\lambda)|D^{i+1}_{abc}(\lambda)]$$

$$\times P[D^{i}_{abc}(x)|D^{i}_{abc}(\lambda)]$$

$$= P[D^{i+1}_{abc}(x)] \sum_{D^{i+1}_{abc}(\lambda)} \Psi_{i+1}(D^{i+1}_{abc}(\lambda))P[D^{i}_{abc}(\lambda)|D^{i+1}_{abc}(\lambda)]P[D^{i}_{abc}(x)|D^{i}_{abc}(\lambda)].$$

Since $\Psi_i(D^{i}_{abc}(\lambda)) = P[D^{i}_{abc}(\lambda), V^{i}_{abc}(x)]/P[V^{i}_{abc}(x)]$, (LDR) follows.

Note that $P[D^{i}_{abc}(\lambda)|D^{i+1}_{abc}(\lambda)]$ are known, for the neighbouring past of any voxel of D^{i}_{abc} belongs to D^{i+1}_{abc}. Also since

$$P[D^{i}_{abc}(x)|D^{i}_{abc}(\lambda)] = \prod_{(m,n,l)\in D^{i}_{abc}} P[x_{mn\ell}|\lambda_{mn\ell}] \qquad (6.3.7)$$

these factors in (6.3.6) are also known. Thus if Ψ_h were known, we could recursively compute all of the Ψ_i, $0\leqslant i<h$, so that, in h steps, we can compute $\Psi_0(q)$, which is our basic goal.

The Simplifying Hypothesis

In order to exploit the (LDR), therefore, we require to compute Ψ_h. It is impractical to do this "exactly" and we impose a hypothesis which allows us to create a simplified expression for Ψ_h and thereby to determine the whole system. An equivalent procedure was carried out in Devijver(1988) and Lacroix(1987).

To motivate the need for simplifying hypotheses, let us try to compute Ψ_h. By its definition,

$$\Psi_h(D^{h}_{abc}(\lambda)) = P[D^{h}_{abc}(\lambda),\ V^{h}_{abc}(x)]/P[V^{h}_{abc}(x)] \qquad (6.3.8)$$

Next, we introduce some more notation. Define

$$\bar{A}_{(R)} = (A^{h\ (x)}_{abc(R)},\ U^{h\ (\lambda)}_{abc(R)})$$

and $$\bar{B}_{(R)} = (B^{h\ (x)}_{abc},\ W^{h\ (\lambda)}_{abc(R)})$$

along with similar definitions for $\bar{A}_{(C)}$, $\bar{A}_{(L)}$, $\bar{B}_{(C)}$ and $\bar{B}_{(L)}$.

Also define

$$\bar{V} = (V^{h-3\ (x)}_{a-1,b-1,c-1},\ D^{h-3\ (\lambda)}_{a-1,b-1,c-1})$$

for $h \geq 3$.

As a result of our earlier decomposition relationships(6.3.4) and (6.3.5), and with reference to Fig 6.3, it is clear that the denominator of (6.3.8) is

$$P[\bar{A}_{(R)},\ \bar{A}_{(C)},\ \bar{A}_{(L)},\ \bar{B}_{(R)},\ \bar{B}_{(C)},\ \bar{B}_{(L)},\ \bar{V}]. \qquad (6.3.9)$$

The crucial step in the computation of (RR) will be the replacement of (6.3.9) by the factorised form

$$\frac{P[\bar{B}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{V}]P[\bar{B}_{(C)},\bar{A}_{(L)},\bar{A}_{(R)},\bar{V}]P[\bar{B}_{(L)},\bar{A}_{(R)},\bar{A}_{(C)},\bar{V}]P[\bar{V}]}{P[\bar{A}_{(R)},\bar{V}]P[\bar{A}_{(C)},\bar{V}]P[\bar{A}_{(L)},\bar{V}]}. \qquad (6.3.10)$$

Clearly, (6.3.10) follows from (6.3.9) only under special independence assumptions involving the arguments of (6.3.9), and these assumptions constitute a suitable hypothesis. There are several approaches to the establishment of such a hypothesis. The most simple sufficient hypothesis is to assume $\bar{A}_{(*)}$ and $\bar{B}_{(*)}$ (altogether 6 terms) are independent of each other, conditionally on $\bar{V}$, ie,

$$P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{B}_{(R)},\bar{B}_{(C)},\bar{B}_{(L)}|\bar{V}]$$
$$= P[\bar{A}_{(R)}|\bar{V}]P[\bar{A}_{(C)}|\bar{V}]P[\bar{A}_{(L)}|\bar{V}]P[\bar{B}_{(R)}|\bar{V}]P[\bar{B}_{(C)}|\bar{V}]P[\bar{B}_{(L)}|\bar{V}], \qquad (6.3.11)$$

It is easy to use the above assumption to derive (6.3.10) from (6.3.9). However, note that there is a sort of "neighbour" relationship between $\bar{A}_{(*)}$ and $\bar{B}_{(*)}$. For example, some voxels corresponding to $\bar{A}_{(R)}$ are neighbours of some voxel corresponding to $\bar{B}_{(C)}$. Therefore, one may think of this hypothesis as not being realible. In fact, we shall state minimal assumptions under which

(6.3.10) follows from (6.3.9), and associated with such minimal assumptions, we can quote a more generous but symmetric set of assumptions.

A suitable minimal hypothesis (M) is represented by the following set of assumptions, expressed in four stages, (M1)-(M4).

Hypothesis (M):

M1 $(\bar{A}_{(L)}, \bar{B}_{(R)}, \bar{B}_{(C)})$ are assumed statistically independent of $\bar{B}_{(L)}$, conditionally on $(\bar{V}, \bar{A}_{(R)}, \bar{A}_{(C)})$.

M2 $(\bar{A}_{(C)}, \bar{B}_{(R)})$ are assumed independent of $\bar{B}_{(C)}$, conditionally on $(\bar{V}, \bar{A}_{(L)}, \bar{A}_{(R)})$.

M3 $\bar{A}_{(R)}$ is assumed independent of $\bar{B}_{(R)}$, conditionally on $(\bar{V}, \bar{A}_{(C)},\bar{A}_{(L)})$.

M4 $\bar{A}_{(R)}$, $\bar{A}_{(C)}$ and $\bar{A}_{(R)}$ are assumed to be mutually independent, conditionally on $\bar{V}$. ⊙

For this hypothesis to make sense, in terms of the Markov Mesh model, it is important that, in each of the above statements, the conditioning item separates the items assumed independent, in that it blocks any pathway between them that exists in the directed graph represented by the assumptions of the model. This was the case in the two-dimensional work in Devijver(1988) and Lacroix(1987), and its veracity can be easily be checked. (c.f. Fig 6.3), while it is not true of the assumption represented by (6.3.11)

Hypothesis (M) is not a unique set of minimal conditions: two others can be created by replacing "L" by "R" or "C" in (M1) and relabelling thereafter, as appropriate.

If this lack of uniqueness or the asymmetry among L, C and R is displeasing, the following symmetric hypothesis (S) suggested by a referee of our paper(Qian and Titterington, 1990b), is undoubtedly sufficient for our purposes. This hypothesis has two stages.

(S1). (i) Assume $(\bar{A}_{(L)}, \bar{B}_{(R)}, \bar{B}_{(C)})$ independent of $\bar{B}_{(L)}$, given $(\bar{V}, \bar{A}_{(R)}, \bar{A}_{(C)})$.

(ii) Assume $(\bar{A}_{(C)}, \bar{B}_{(L)}, \bar{B}_{(R)})$ independent of $\bar{B}_{(C)}$, given $(\bar{V}, \bar{A}_{(L)}, \bar{A}_{(C)})$.

(iii) Assume $(\bar{A}_{(R)}, \bar{B}_{(C)}, \bar{B}_{(L)})$ independent of $\bar{B}_{(R)}$, given $(\bar{V}, \bar{A}_{(C)}, \bar{A}_{(L)})$.

(S2). Same as (M4). ⊙

Note that (S1)(i) is the same as (M1), that (S1)(ii) implies (M2) and that (S1)(iii) implies (M3), thus verifying that, altogether, (S) is enough to imply (M), but not vice versa.

In practice, it is irrelevant which of (M) or (S) is assumed. They are both "respectable" in reflecting the type of "pathway blocking" mentioned above, and, as we will see, they both allow the deduction of (6.3.10) from (6.3.9).

(6.3.9) To (6.3.10) Under Hypothesis (M)

Here we show that (6.3.9) leads to (6.3.10), under the hypothesis summarised by (M1)-(M4).

By (M1),

$$P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{B}_{(R)},\bar{B}_{(C)},\bar{B}_{(L)},\bar{V}] = \frac{P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{B}_{(R)},\bar{B}_{(C)},\bar{V}]P[\bar{B}_{(L)},\bar{A}_{(R)},\bar{A}_{(C)},\bar{V}]}{P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{V}]}.$$

By (M2),

$$P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{B}_{(R)},\bar{B}_{(C)},\bar{V}] = \frac{P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{B}_{(R)},\bar{V}]P[\bar{B}_{(C)},\bar{A}_{(L)},\bar{A}_{(R)},\bar{V}]}{P[\bar{A}_{(L)},\bar{A}_{(R)},\bar{V}]}.$$

By (M3),

$$P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{B}_{(R)},\bar{V}] = \frac{P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{V}]P[\bar{B}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{V}]}{P[\bar{A}_{(C)},\bar{A}_{(L)},\bar{V}]}.$$

By (M4),

$$\frac{P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{V}]}{P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{V}]P[\bar{A}_{(L)},\bar{A}_{(R)},\bar{V}]P[\bar{A}_{(C)},\bar{A}_{(L)},\bar{V}]} = \frac{P[\bar{V}]}{P[\bar{A}_{(R)},\bar{V}]P[\bar{A}_{(C)},\bar{V}]P[\bar{A}_{(L)},\bar{V}]}.$$

Multiplication of the above four equations generates the required results.

The Recurrence Relationship(RR)

The step from (6.3.9) to (6.3.10) is the first stage in the derivation of the following Recurrence Relationship. It is helpful to introduce some further notation, representing certain subsets of $D^{h}_{abc}(\lambda)$.

We define

$$D^{1(\lambda)}_{h(R)} = (U^{h\ (\lambda)}_{abc(L)}, W^{h\ (\lambda)}_{abc(R)}, U^{h\ (\lambda)}_{abc(C)}, D^{h-3\ (\lambda)}_{a-1,b-1,c-1}), \qquad (6.3.12)$$

$$D^{2(\lambda)}_{h(R)} = (U^{h\ (\lambda)}_{abc(R)}, D^{h-3\ (\lambda)}_{a-1,b-1,c-1}) \qquad (6.3.13)$$

$$D^{3(\lambda)}_{h} = (D^{h-3}_{a-1,b-1,c-1}(\lambda)) \qquad (6.3.14)$$

and $D^{1(\lambda)}_{h(C)}$, $D^{1(\lambda)}_{h(L)}$, $D^{2(\lambda)}_{h(C)}$ and $D^{2(\lambda)}_{h(L)}$ by analogy with (6.3.12) and (6.3.13) respectively. Note that all these objects are composed of those labels of $D^{h}_{abc}(\lambda)$ located in $D^{h-1}_{a-1,bc}$, $D^{h-2}_{a,b-1,c-1}$, $D^{h-3}_{a-1,b-1.c-1}$, etc.

With the help of this notation, our recurrence relation (**RR**) is

$$\textbf{(RR)} \qquad \Psi_h(D^{h}_{abc}(\lambda))$$

$$= N_r \frac{\Psi^{1}_{h(R)}(D^{1(\lambda)}_{h(R)})\Psi^{1}_{h(C)}(D^{1(\lambda)}_{h(C)})\Psi^{1}_{h(L)}(D^{1(\lambda)}_{h(L)})\Psi^{3}_{h}(D^{3(\lambda)}_{h})}{\Psi^{2}_{h(R)}(D^{2(\lambda)}_{h(R)})\Psi^{2}_{h(C)}(D^{2(\lambda)}_{h(C)})\Psi^{2}_{h(L)}(D^{2(\lambda)}_{h(L)})} \qquad (6.3.15)$$

$$\text{where } N_r = \frac{P[V^{h-1}_{a-1,bc}(x)]P[V^{h-1}_{a,b-1,c}(x)]P[V^{h-1}_{ab,c-1}(x)]P[V^{h-3}_{a-1,b-1,c-1}(x)]}{P[V^{h-2}_{a,b-1,c-1}(x)]P[V^{h-2}_{a-1,b,c-1}(x)]P[V^{h-2}_{a-1,b-1,c}(x)]P[V^{h}_{abc}(x)]}$$

Note that, similarly to the functions Ψ_i, the functions $\Psi^{i}_{h(*)}$, (i=1,2,3, *=R,C,L) are associated directly with voxel (a,b,c). They are, in fact, directly expressible in terms of our previous notation in that $\Psi^{1}_{h(*)}$, $\Psi^{2}_{h(*)}$ and Ψ^{3}_{h} correspond respectively to Ψ_{h-1}, Ψ_{h-2} and Ψ_{h-3} associated with voxels (a-1,b,c) and (a-1,b-1,c) and (a-1,b-1,c-1), etc. For example, $\Psi^{1}_{h(R)}(D^{1(\lambda)}_{h(R)}) = P[D^{1(\lambda)}_{h(R)}|V^{h-1}_{a-1,bc}(x)]$, $\Psi^{2}_{h(R)}(D^{2(\lambda)}_{h(R)})=P[D^{2(\lambda)}_{h(R)}|V^{h-2\ (x)}_{a.b-1,c-1}]$ and $\Psi^{3}_{h}(D_h)=P[D^{3(\lambda)}_{h}|V^{h-3\ (x)}_{a-1,b-1,c-1}]$. As a result, from the Ψ_i associated with voxels belong to the past of voxel (a,b,c), (i<h), with the help of "local decomposition relationship", we can compute Ψ_h associated with voxel (a,b,c) and, thereby attain our goal.

*Proof of (RR):

$$\Psi_h(D_{abc}^{h}(\lambda)) = P[\bar{A}_{(R)},\bar{A}_{(C)},\bar{A}_{(L)},\bar{B}_{(R)},\bar{B}_{(C)},\bar{B}_{(L)},\bar{V}]/P[V_{abc}^{h}(x)]$$

$$= \frac{P[\bar{A}_{(C)},\bar{B}_{(L)},\bar{A}_{(R)},\bar{V}]P[\bar{A}_{(L)},\bar{B}_{(R)},\bar{A}_{(C)},\bar{V}]P[\bar{A}_{(R)},\bar{B}_{(C)},\bar{A}_{(L)},\bar{V}]P[\bar{V}]}{P[V_{abc}^{h}(x)]P[\bar{A}_{(R)},\bar{V}]P[\bar{A}_{(C)},\bar{V}]P[\bar{A}_{(L)},\bar{V}]}$$

$$= \frac{P[D_{h(R)}^{1(\lambda)}|V_{a-1,bc}^{h-1\ (x)}]P[D_{h(C)}^{1(\lambda)}|V_{a,b-1,c}^{h-1\ (x)}]P[D_{h(L)}^{1(\lambda)}|V_{ab,c-1}^{h-1\ (x)}]}{P[D_{h(R)}^{2(\lambda)}|V_{a,b-1,c-1}^{h-2\ (x)}]P[D_{h(C)}^{2(\lambda)}|V_{a-1,b,c-1}^{h-2\ (x)}]P[D_{h(L)}^{2(\lambda)}|V_{a-1,b-1,c}^{h-2\ (x)}]}$$

$$\times\frac{P[D_{h}^{3(\lambda)}|V_{a-1,b-1,c-1}^{h-3\ (x)}]P[V_{a-1,bc}^{h-1(x)}]P[V_{a,b-1,c}^{h-1\ (x)}]P[V_{ab,c-1}^{h-1(x)}]P[V_{a-1,b-1,c-1}^{h-3\ (x)}]}{P[V_{abc}^{h}(x)]P[V_{a,b-1,c-1}^{h-2}(x)]P[V_{a-1,b,c-1}^{h-2}(x)]P[V_{a-1,b-1,c}^{h-2}(x)]}$$

Θ

The recurrence relationship (6.3.15) holds if $h \geqslant 3$. For $h=1$ or 2, the decomposition relationships of V_{abc} and, in particular, of D_{abc} are different. To be precise, D_{abc} is only divided into six parts, none of which belongs to $V_{a-1,b-1,c-1}$, while, for D_{abc}, the situation is even more simple. However, if we formally define

$$V_{a-1,b-1,c-1}^{-1} = V_{a-1,b-1,c-1}^{-2} = V_{a-1,b-1,c-1}^{0} = V_{a-1,b-1,c-1},$$

define $D_{a-1,b-1,c-1}^{-1} = D_{a-1,b-1,c-1}^{-2}$ to be empty sets and define $U_{abc(*)}^{1}$, with $*=R,C,L$ also to be empty sets, then the decomposition relationships for $h=1,2$ can be written in terms of the same formulae as those for $h \geqslant 3$. We have the following recurrence relationships for $h=1$ and $h=2$. (c.f. Fig 6.3)

$$\Psi_1(r,s,t) \propto$$

$$P[\lambda_{a-1,bc}=r|V_{a-1,bc}^{(x)}]P[\lambda_{a,b-1,c}=s|V_{a,b-1,c}^{(x)}]P[\lambda_{ab,c-1}=t|V_{ab,c-1}^{(x)}] \tag{6.3.16}$$

$$\Psi_2(n_1,n_2,\ldots n_6) \propto$$

$$\frac{P[n_1,n_2,n_3|V_{a-1,bc}^{1\ (x)}]P[n_2,n_4,n_5|V_{a,b-1,c}^{1\ (x)}]P[n_3,n_5,n_6|V_{ab,c-1}^{1\ (x)}]}{P[n_2|V_{a-1,b-1,c}^{(x)}]P[n_3|V_{a-1,b,c-1}^{(x)}]P[n_5|V_{a-1,b-1,c}^{(x)}]}. \tag{6.3.17}$$

6.3.3 Boundary conditions

In this subsection, we assume h to denote the highest order adopted in both relationships. If voxel (a,b,c) belongs to the first h rows (or columns, or layers) ie., $\min(a,b,c) \leq h-1$, it is not possible to consider the h-th diagonal, and only functions Ψ_i such that $0 \leq i \leq \min(a,b,c)$ are defined as in the last two subsections. However, there is a difference between the three- and two-dimensional cases. In the two-dimensional case, if we consider pixel (a,b) and if $i=\min(a,b)-1$, $j=\max(a,b)-1$, we can define diagonals with order higher than i and lower than j, but where the numbers of pixels in these diagonals are the same as that in the i-th diagonal. For voxel (a,b,c) in the three-dimensional case, let $i=\min(a,b,c)-1$, and let $j=i+\min\{(a-i+1,b-i+1,c-i+1)\backslash(0)\}$, so that j+1 is the second smallest of a, b and c. If j>i, we can also define diagonals with order from i+1 to j. However, the numbers of voxels in these diagonals increase (see Fig 6.4). On the other hand, note that, for the initial planes, we can use the method for the two-dimensional case with high order. Therefore, if i<j, for a voxel with i and j defiped as above, we may either use local decomposition relationships only up to functions Ψ_i, or we may define functions Ψ_s (s>i) in such a way that we can use similar local decomposition relationships and recurrence relationships. If we adopt the former strategy we need the recurrence relationship for order 1 and order 2. These are presented in the preceding subsection, and we therefore do not discuss them here.

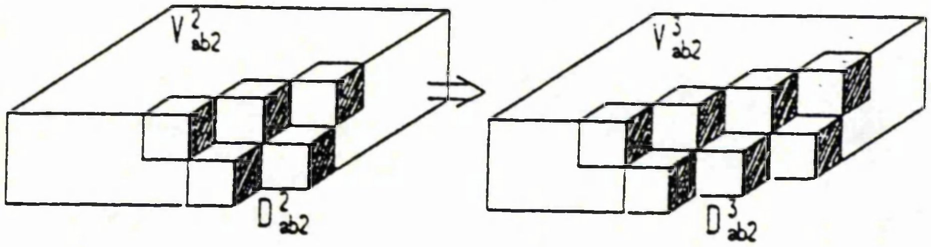


Fig 6.4 V^h_{abc} and D^h_{abc} at boundary case

From Figs 6.4 and 6.5, we can see that the functions Ψ with subscript higher than i have fewer arguments (which represent the states at those corresponding voxels), and, in terms of the local decomposition relationships, the numbers of voxels increase along only two directions. Thus both relationships, the recurrence relationship in particular, become simpler. To illustrate this, we only consider the case of i=c-1 in detail. For the other two cases, the argument is directly analogous.

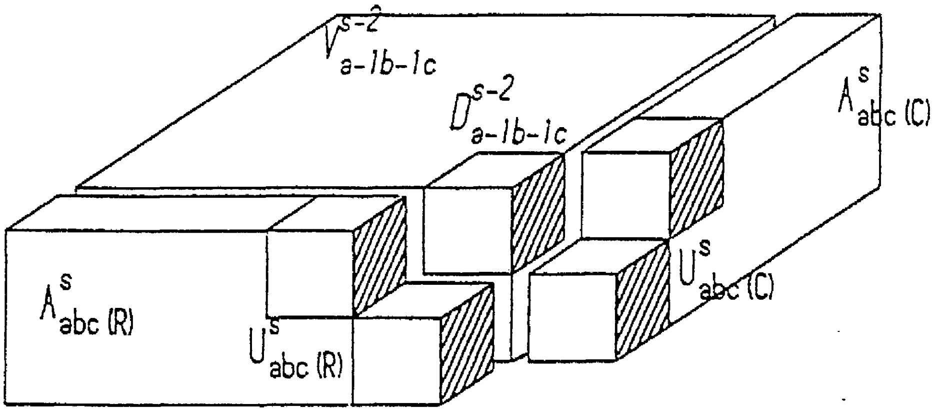


Fig 6.5 The decomposition of V^{h}_{abc} and D^{h}_{abc} at boundary case

Local decomposition relationship

When s>i, the LDR for function Ψ_s is almost the same as that in the last subsection, except that D^{s}_{abc}, V^{s}_{abc}, D^{s+1}_{abc} and V^{s+1}_{abc} are slightly different, in that fewer voxels are involved.

Recurrence relationship

When s>i we see that, as shown in Fig 6.5,

$$V^{s}_{abc} = A^{s}_{abc(R)} + V^{s-2}_{a-1,b-1,c} + A^{s}_{abc(C)} \qquad (6.3.18)$$

$$D^{s}_{abc} = U^{s}_{abc(R)} + D^{s-2}_{a-1,b-1,c} + U^{s}_{abc(C)}. \qquad (6.3.19)$$

We therefore impose the hypothesis that $(A^{s}_{abc(R)}{}^{(x)}, U^{s}_{abc(R)}{}^{(\lambda)})$ and $(A^{s}_{abc(C)}{}^{(x)}, U^{s}_{abc(C)}{}^{(\lambda)})$ are independent, conditionally upon $(V^{s-2}_{a-1,b-1,c}{}^{(x)}, D^{s-2}_{a-1,b-1,c}{}^{(\lambda)})$.

The RR is therefore obtained as follows:

$$\Psi_s(D^{s}_{abc}{}^{(\lambda)}) = [\Psi_{s-1(a-1,bc)}(U^{s}_{abc(R)}{}^{(\lambda)}, D^{s-2}_{a-1,b-1,c}{}^{(\lambda)})$$

$$\times\Psi_{s-1(a,b-1,c)}(U^{s}_{abc(C)}{}^{(\lambda)}, D^{s-2}_{a-1,b-1,c}{}^{(\lambda)})]/\Psi_{s-2(a-1,b-1,c)}(D^{s-2}_{a-1,b-1,c}{}^{(\lambda)})$$

where $\Psi_{s-1(a-1,bc)}$, $\Psi_{s-1(a,b-1,c)}$ and $\Psi_{s-2(a-1,b-1,c)}$ are those functions associated with voxels (a-1,b,c), (a,b-1,c) and (a-1,b-1,c) respectively. Note that the above recurrence relationship is almost the same as that in the two-dimensional case (Lacroix, 1987).

6.4 Experimental results

In order to analyse data from real, noisy images using Markov Mesh models, it is first necessary to guess, or estimate, values for unknown parameters from the data. This is usually a difficult problem, which we shall discuss in the next section. In our numerical work here, we used simulated binary images and an artifical ternary image corrupted by white noise, and we implemented our labelling algorithm under the assumption that the true parameters in the model were known.

In the following simulations the Markov meshes were defined by a single parameter, $\beta>0$, that represents a measure of the Markovian information contained in the true scenes. Suppose S_n denotes the number of members of the state space S. (In our examples, $S_n=2$ or 3.) Then we took

$$P_q = P[\lambda_{ooo} = q] = 1/S_n$$

$$P_{q|r\square\square} = P_{q|\square r\square} = P_{q|\square\square r} = \exp\{\beta\delta(q,r)\}/ \sum_{w\in S} \exp\{\beta\delta(w,r)\}$$

$$P_{q|rs\square} = P_{q|r\square s} = P_{q|\square rs}$$

$$= \exp\{\beta(\delta(q,r)+\delta(q,s))\}/ \sum_{w\in S} \exp\{\beta(\delta(w,r)+\delta(w,s))\}$$

$$P_{q|rst}=\exp\{\beta(\delta(q,r)+\delta(q,s)+\delta(q,t))\}/\sum_{w\in S} \exp\{\beta(\delta(w,r)+\delta(w,s)+\delta(w,t))\}$$

where $\delta(q,r)=1$ if $q=r$; $=0$ otherwise. The lower β is, the lower is the tendency for neighbouring voxels to have the same label, and the lower therefore is the Markovian information in the scene.

Since we have generated the images subsequently restored, we can easily count the number of voxels incorrectly restored and thereby report error rates.

Fig 6.6a shows an artifical ternary image on a 64×64×8 frame (ie. $M=N=63$, $L=7$). The first column displays layers 1-4 and the second column shows layers 5-8. Note that the picture changes slightly from one layer to the next, while there is a big difference between the first and last layer. The state space for the true labels was (0,1,2) and to each voxel independent white noise, with variance $\sigma^2=0.36$, was added.

First, the voxel-wise maximum likelihood(ML) classifier was applied. This simply assigned each voxel to the label "nearest" to its feature variable. The resulting labelling, shown in Fig 6.6b incurred 8410 errors(25.6%). Next, the image was labelled using our algorithm with h=1. The parameter β was chosen to be 1.5. For the boundaries we used the simplest possible approach, using only a one-step decomposition and recurrence iteration up to the nearest, past neighbouring voxels, of which there can be 1 or 2. The corresponding restored image is shown in Fig 6.6c and contains 1955 errors (6.0%).

We also implemented the "block constraint" method of Kanefsky and Strintzis(1978), as was done in Lacroix(1987) as a comparison with the two-dimensional algorithm. This method maximizes likelihoods corresponding to a small regions. The rule for assigning a state q to voxel (a,b,c) is:

$$K_{abc}(q) = \max_{w \in S} K_{abc}(w)$$

where

$$K_{abc}(w) = \sum_{r,s,t} P[r,s,t] f_r(x_{a-1,bc}) f_s(x_{a,b-1,c}) f_t(x_{ab,c-1}) f_w(x_{abc}) P_{w|rst}$$

in which $f_r(x_{abc}) = f(x_{abc}|\lambda_{abc}=r)$ and P[r,s,t] is the probability of the occurrence of the event $(\lambda_{a-1,bc}=r, \lambda_{a,b-1,c}=s, \lambda_{ab,c-1}=t)$ over the entire image. In practice, P[r,s,t] is unknown. In simulations, however, such quantities can be estimated either from the true scene or, less satisfactorily, from restorations obtained by some other method. Our experience was that, not surprisingly, it was better to use the true scene.

We compared the block-constraint method with our own algorithm for h=1 and h=2 in a study that parallels one reported by Lacroix(1987). We generated binary images on 20×20×20 frames with conditional means m_0=0.1 and m_1=0.9 for the white noise. For the block-constraint method, the P[r,s,t] were estimated from the true scenes.

First we fixed β=0.8 or 2.0 and varied σ, so that the error rate was a function of σ. Results shown in Fig 6.7 reflect a similar trend to that in Lacroix(1987), based on a measure of signal-to-noise ratio (SNR): the smaller σ is, the larger is SNR. In our example,

$$SNR = (m_1-m_0)^2/(4\sigma^2) = 0.16/\sigma^2.$$

In the second part of the study, we fixed σ at 0.4 or 0.7 and varied β. Note that the larger β is, the greater is the Markovian information, and the smaller therefore is the entropy. The results are shown in Fig 6.8. In both Figs 6.7 and 6.8, the error rates are averages of three replications.

Note from both Figs 6.7 and 6.8 that, when β and σ are both large, the results for h=1 are superior to those for h=2. This situation did not occur for the combinations of small β with large σ (e.g. β=0.8, σ=1.2) or large β with small σ(e.g. β=5.0, σ=0.5). The explanation may be that, for a particular voxel (a,b,c), some of the probabilities $\{P[D_{2(R)}^{(\lambda)}|V_{a,b-1,c-1}^{(x)}], P[D_{2(C)}^{(\lambda)}|V_{a-1,b,c-1}^{(x)}], P[D_{2(L)}^{(\lambda)}|V_{a-1,b-1,c}^{(x)}]\}$ (see (6.3.10)), are very small if β and σ are both large, and, in the recurrence relationship associated with h=2, where $D_{2(*)}$ is only one voxel, (for example, $D_{2(R)}$ is voxel (a,b-1,c-1)), the product of these three functions is used as a divisor of three other conditional probabilities (see (6.3.17)). Repetitions of such operations in numerical calculations may lead to the accumulation of round-off errors, with the result that unexpectedly many voxels are assigned wrong labels. To combat this we made the following modification: for each $\Psi_2(D_{abc}^{(\lambda)})$ obtained by (6.3.17), a small, positive regularising constant (specifically, 0.5) is added to each factor in the divisor. The results are presented in Figs 6.9 and 6.10, corresponding to Figs 6.7 and 6.8, respectively. We can note that the resulting figures were virtually identical to Figs 6.7 and 6.8 except that the anomalies in the results for h=2 disappeared. The performance of h=2 still shows little superiority over that of h=1.

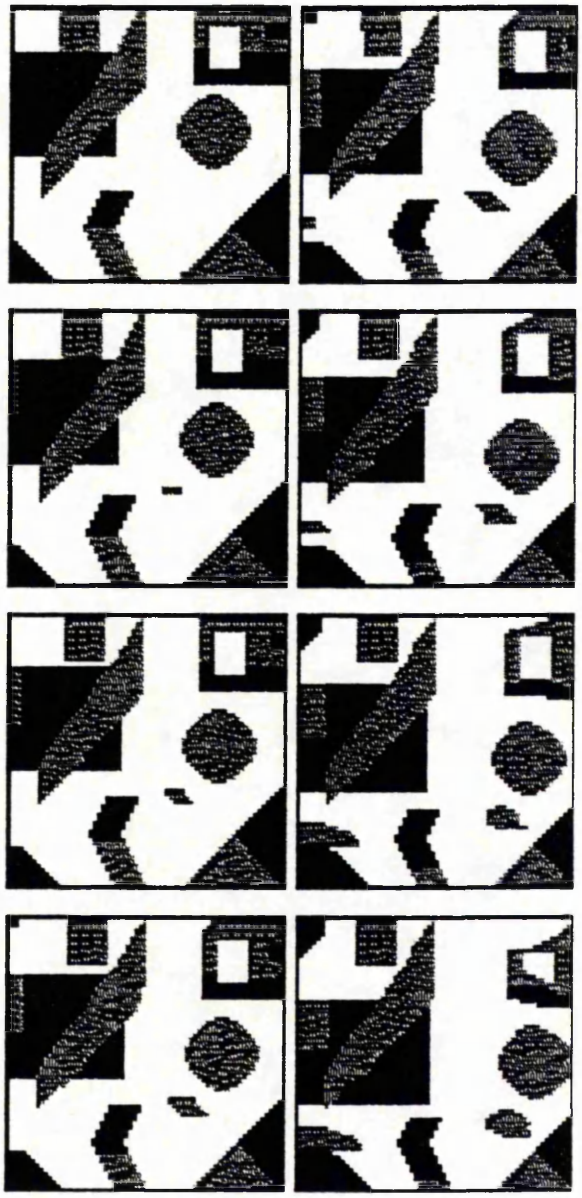


Fig 6.6a An artifical image: 64×64×8

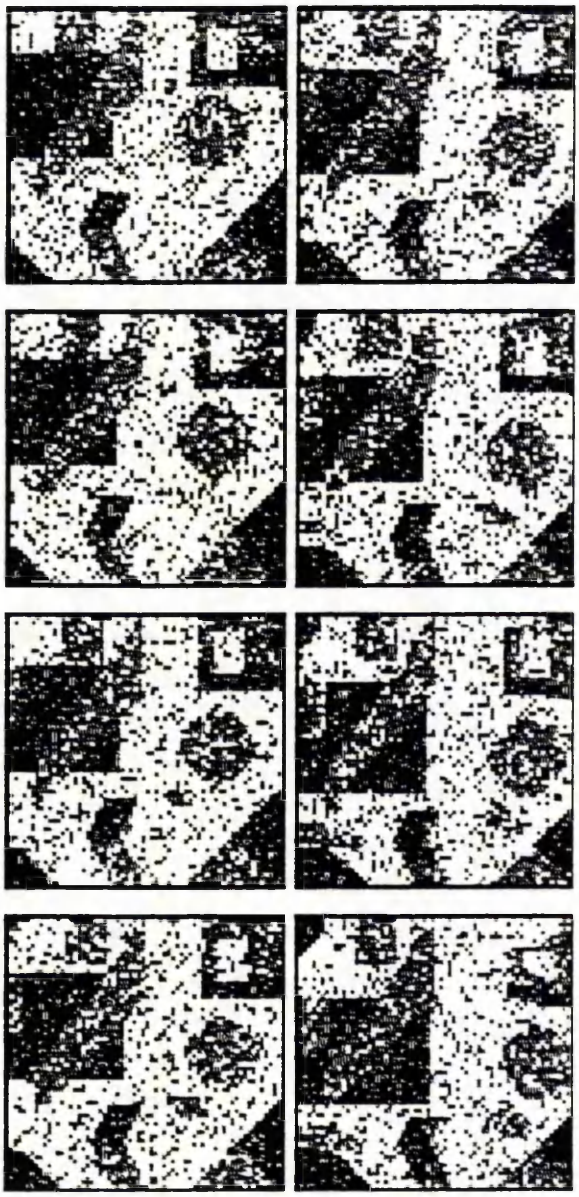


Fig 6.6b Maximum likelihood classifier: 25.6% error rate

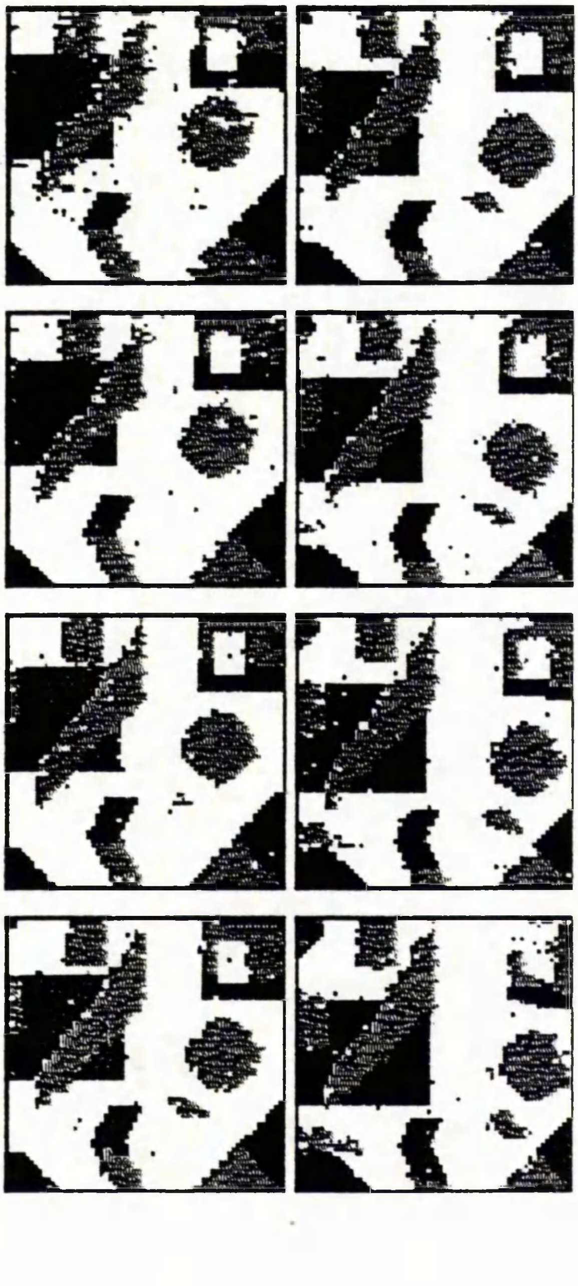


Fig 6.6c Restoration with h=1: 6.0% error rate

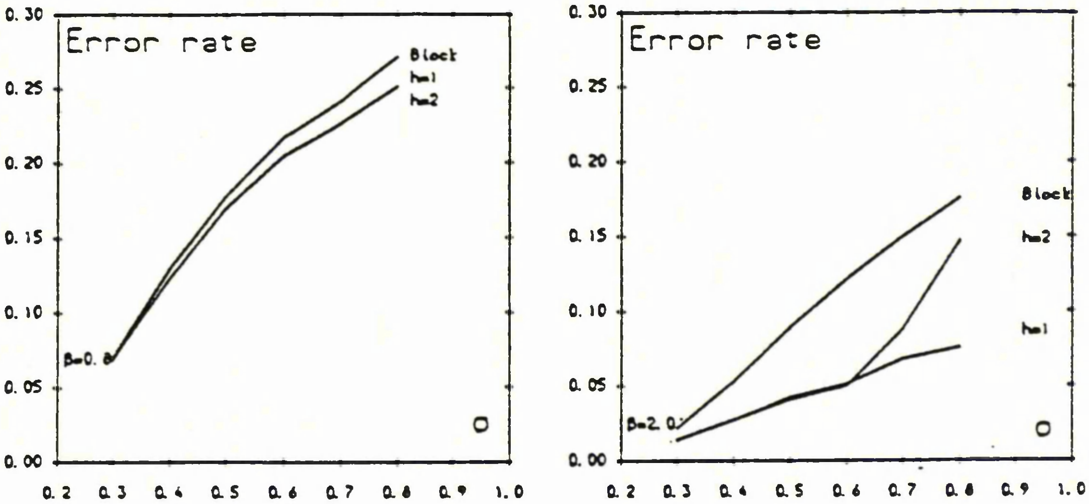


Fig 6.7 Some results of error rate against σ

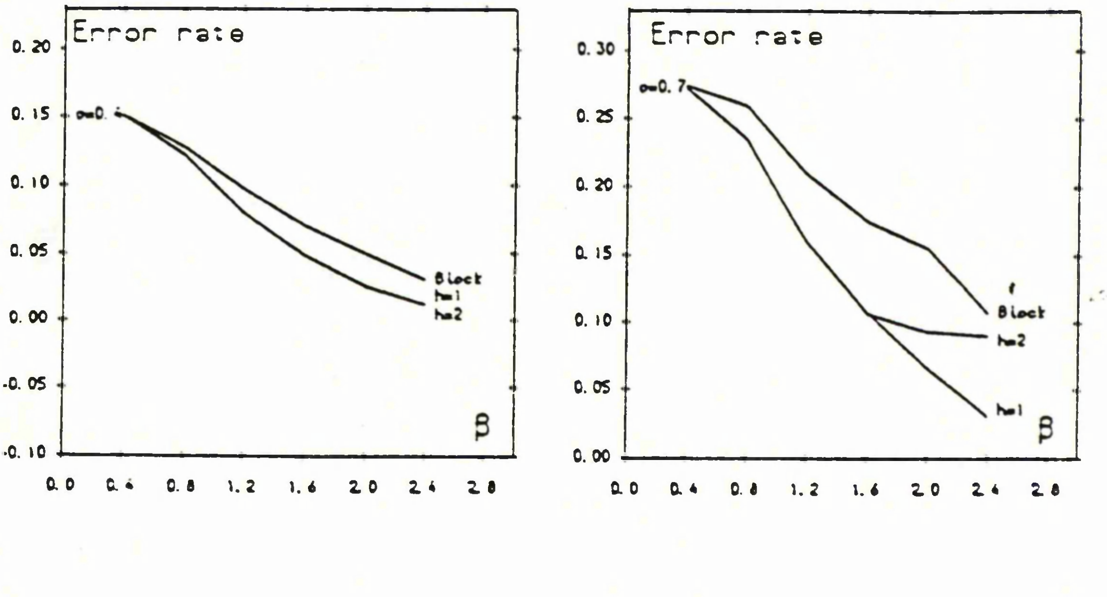


Fig 6.8 Some results of error rate against β

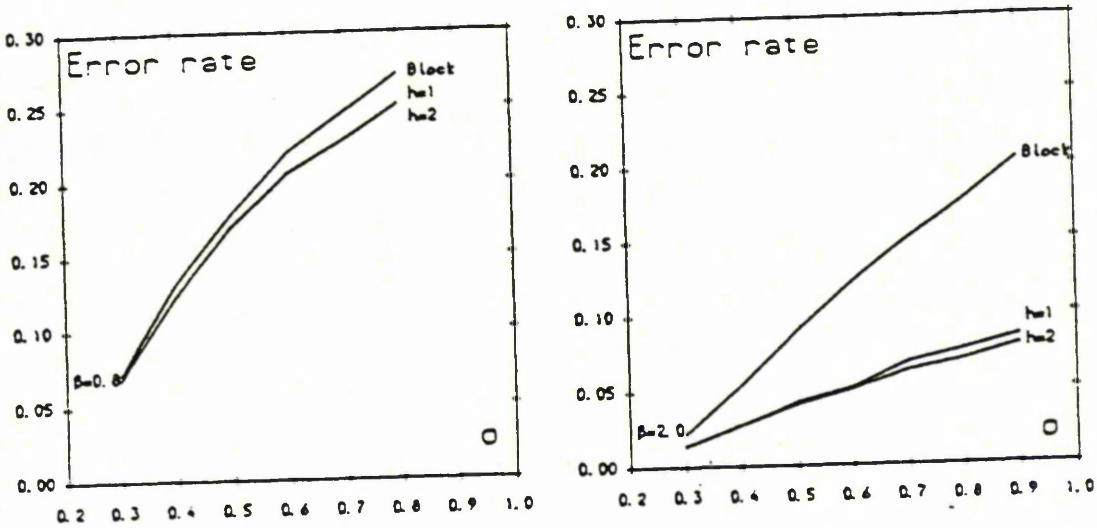


Fig 6.9 Error rates against σ after modification

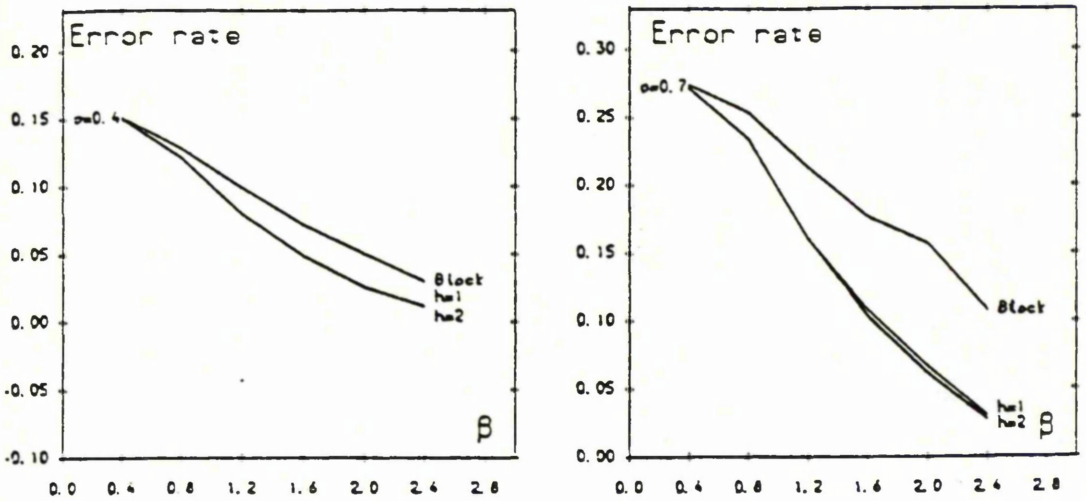


Fig 6.10 Error rates against β after modification

6.5 Parameter estimation from noisy data

Our aim in this section is to discuss the problem of estimating the model parameters. We concentrate only on the case of two-dimensional second-order models. The treatment is directly analogous for multi-dimensional cases. Devijver(1988) proposed a learning algorithm, which is based on a decision directed (DD) approach and is in fact a modification of the iterative procedure of the EM algorithm. In his procedure, the restored image was used at each cycle of iteration. Note from the last chapter that, in the iterative procedure of simultaneous parameter estimation and restoration for Markov random field models, a product of a set of local conditional distributions is maximized by using the EM algorithm. We can adopt the idea to hidden MMRF models as well. To be precise, at each cycle of iteration, after restoring the image, say, $\hat{x}$, we re-estimate the parameters by maximizing the function

$$\prod_{i=0}^{M} \prod_{j=0}^{N} \Pr[x_{ij}|\hat{\lambda}_{i-1,j},\hat{\lambda}_{i,j-1},\beta,\theta]. \qquad , \qquad (6.5.1)$$

(6.5.1) calls for some comments. First of all, it comes directly from the prior likelihood function

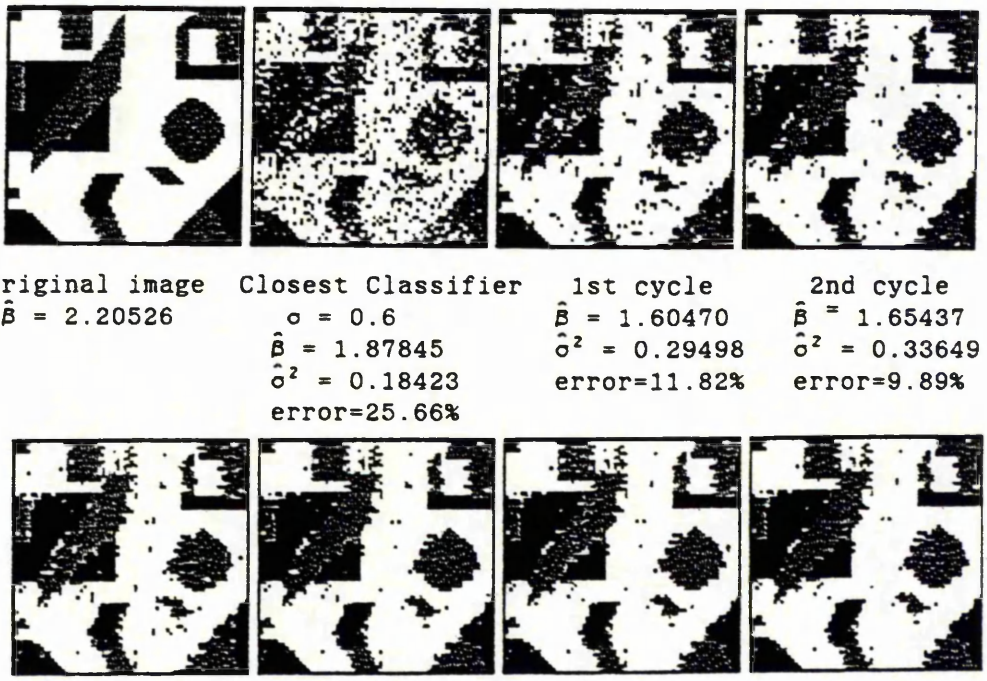
$$\prod_{i=0}^{M} \prod_{j=0}^{N} \Pr[\lambda_{ij}|\lambda_{i-1,j},\lambda_{i,j-1},\beta]. \qquad (6.5.2)$$

It is easy to deal with both sorts of conditional densities in (6.5.1) and (6.5.2), respectively, if $\Pr[x_{ij}|\lambda_{ij},\theta]$ is not complicated, and the EM algorithm can therefore be used to maximize (6.5.1). Secondly, for the case of MRF models, all the neighbouring pixels of one pixel are used in the corresponding local conditional density, while in (6.5.1) we only use the past neighbours. This may be a reason for the different behaviours of the procedures for the two models. Thirdly, we may, at each cycle, maximize (6.5.2), providing $\lambda=\hat{\lambda}$, and the other function involving the parameter θ, namely,

$$\prod_{i=0}^{M} \prod_{j=0}^{N} \Pr[x_{ij}|\hat{\lambda}_{ij},\theta], \qquad (6.5.3)$$

to obtain new values of the parameters. This is just the procedure described in Besag(1986).

Our experiments were performed under the additional assumptions that only one parameter, β, is involved in the prior model, (to be precise, the (conditional) probabilities are defined in a similar way to (6.4.1), (6.4.2) and (6.4.3)), and that the observed images are created with additive white noise with variance as the only unknown parameter.



Original image
$\hat{\beta} = 2.20526$

Closest Classifier
$\sigma = 0.6$
$\hat{\beta} = 1.87845$
$\hat{\sigma}^2 = 0.18423$
error=25.66%

1st cycle
$\hat{\beta} = 1.60470$
$\hat{\sigma}^2 = 0.29498$
error=11.82%

2nd cycle
$\hat{\beta} = 1.65437$
$\hat{\sigma}^2 = 0.33649$
error=9.89%

4th cycle
$\hat{\beta} = 1.77154$
$\hat{\sigma}^2 = 0.36763$
error=9.13%

8th cycle
$\hat{\beta} = 1.90790$
$\hat{\sigma}^2 = 0.38656$
error=8.91%

12th cycle
$\hat{\beta} = 1.95748$
$\hat{\sigma}^2 = 0.39385$
error=8.96%

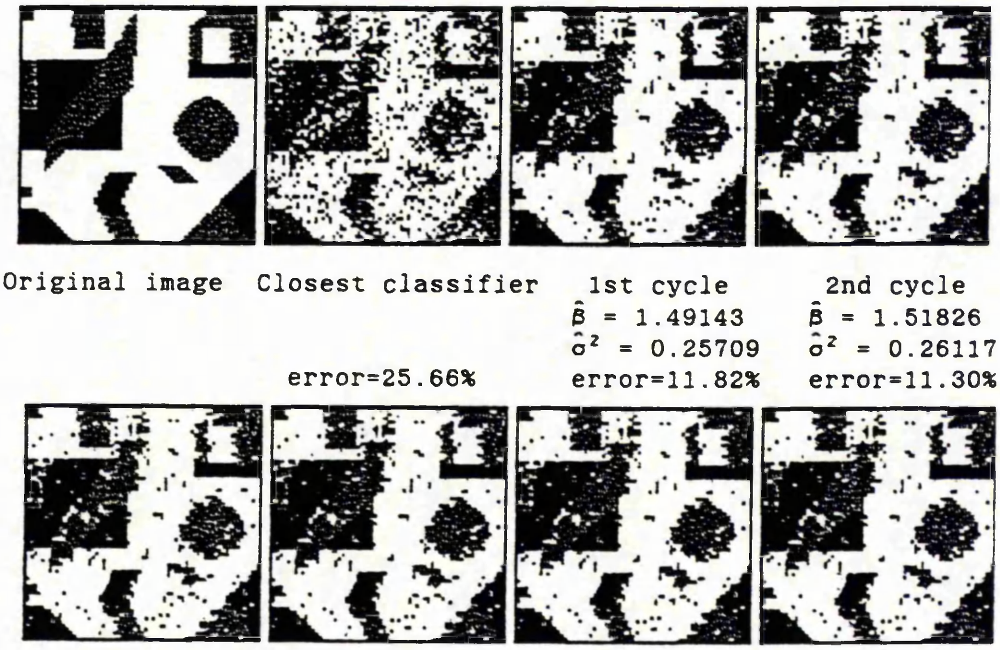
18th cycle
$\hat{\beta} = 1.96740$
$\hat{\sigma}^2 = 0.39578$
error=8.96%

Fig 6.11 Iterative procedure with EM algorithm

The illustrations concern an artifical 64×64 three-state image. White noise with variance σ=0.6 was added. We first restored the image by means of closest classification, and then estimated parameters by maximizing (6.5.2) and (6.5.3), assumimg that the restored image was the true scene. From these estimates, we started our iterative procedure which maximizes (6.5.1) at each cycle. Results of parameter estimation and restoration are shown in Fig 6.11. The procedure converges very quickly. Note that a similar phenomenon to the procedure for Markov random fields (see Fig 5.1) occured: the convergence of restoration is extremely rapid, and the restored image changes a little after one cycle, whereas the parameter estimates are quite different. Fig 6.12 illustrates the iterative procedure where

(6.5.2) and (6.5.3) were used in each cycle, where we also started the iterative procedure the same way as in Fig 6.11. This procedure converges faster, however, and we find that estimation using (6.5.1), especially for the variance of the noise, is much better than the procedure without the EM algorithm, although the difference between restorations is not so great.

Our experiments with different variances for the artifical image and for simulated images also showed that the estimated parameter $\hat{\beta}$ is usually smaller than the true value, whereas estimation of the variance of the noise is quite good. We can note from the last chapter a different phenomenon for the MRF model case, in that, for simulated images, the estimates of the parameter β are almost the same as the true values.



Original image

Closest classifier
error=25.66%

1st cycle
$\hat{\beta} = 1.49143$
$\hat{\sigma}^2 = 0.25709$
error=11.82%

2nd cycle
$\hat{\beta} = 1.51826$
$\hat{\sigma}^2 = 0.26117$
error=11.30%

4th cycle
$\hat{\beta} = 1.53522$
$\hat{\sigma}^2 = 0.26399$
error=11.06%

8th cycle
$\hat{\beta} = 1.54563$
$\hat{\sigma}^2 = 0.26590$
error=10.89%

12th cycle
$\hat{\beta} = 1.54769$
$\hat{\sigma}^2 = 0.26629$
error=10.86%

18th cycle
$\hat{\beta} = 1.54769$
$\hat{\sigma}^2 = 0.26629$
error=10.86%

Fig 6.12 Iterative procedure without EM algorithm

6.6 Discussion

As in the two-dimensional case(Lacroix, 1987), our algorithm consists of two parts, the second of which relies on a hypothesis whose validity cannot be guaranteed. If, in the three-dimensional case, the algorithm is applied to order greater than 2, then the number of voxels involved in the restoration of a single voxel becomes very great, with consequently high computational demands. There is also another computational problem. The diagonal D_{abc}^{h} contains many voxels, so that some of the associated conditional probabilities are small and prone to underflow, even after normalization. Finally, we are liable to suffer the problem of numerical instability discussed in Section 6.4. Introduction of the modification mentioned there transforms the procedure into a quite different algorithm.

These factors, along with the likelihood that the results with h>2 are unlikely to be dramatically superior to those with h=2, stimulate us to suggest that the algorithm be used only with h=1 or h=2.

Besag's(1986) ICM method is originally proposed for MRF models. It is an iterative procedure, while the method we discussed in this chapter is a single pass restoration. From Property 2 of our model (Section 6.2), we know that the conditional probabilities for one voxel, given labels at all other voxels, can easily be computed. For MMRF models, therefore, one can also use ICM for restoration.

The model discussed in this chapter is the simplest model for three-dimensional problems. One might use a more complex model in both two- and three-dimensional cases.

In the iterative procedure for parameter estimation and restoration, we can maximize a product of more complex conditional densities, such as,

$$\Pi P[x_{abc}, D_{abc}^{1}{}^{(x)} | D_{abc}^{2}{}^{(\hat{\lambda})}, \beta, \theta],$$

and this may lead to estimators with better properties.

Although our methods have produced reasonable results, further work can be carried out to improve the methods, particularly with a view to enhancing the numerical stability by deriving a better recurrence relationship.

Chapter 7

Multi-Dimensional Markov Chain Models For Image Textures

7.1 Introduction

In this chapter we explore the use of the multi-dimensional Markov chain as a model for texture. Consider a rectangular array of M rows and N columns, regarding each row as a random vector. Under the model, it is assumed that the distribution of the current row, conditionally upon all "previous rows", depends only on a small number of past, neighbouring rows. This conditional distribution will be assumed to reflect pairwise interaction between neighbouring pixels, in order that the properties of the model might be very similar to those of pairwise-interaction MRF models.

The MMRF model exhibits causal dependence in that samples from them can be generated directionally. As mentioned before, it can be regarded as a direct analogue of the (one-dimensional) Markov chain, or as a sort of two-dimensional time series. The asymmetry of its local dependence structure leads to the aesthetically unsatisfatory directional property. Causal dependence is not present in the MRF model, but realisations are usually generated by a iterative procedure. Some methods for simulating MRFs, such as that used in Cross and Jain(1983), are conditional on prescribed assumptions. The resulting patterns may therefore be wholly unrepresentative of the distribution of the corresponding MRF(Besag, 1986). Although the convergence of Geman and Geman's (1984) stochastic relaxation method can be proved, and parallel computational architectures can be exploited, the theoretically valid speeds of convergence are very slow.

Our model is a stationary multiple Markov chain. Although it is difficult to make it strictly stationary, and non-directionality cannot be guaranteed, local dependence on the past pixels is symmetric. As in the case of MRFs, values chosen for the underlying parameters reflect the nature of the interaction between neighbouring pixels.

In Chapter 2, (see also Qian and Titterington, 1990a), we investigated the one-dimensional, pairwise-interaction Gibbs chain, and developed a recursive technique for calculating its normalizing constant, thereby enabling us to simulate the Gibbs chain and to carry out maximum likelihood estimation of parameters. This "line technique" also enables us to handle the model which we shall introduce in this chapter.

Standard Markov chains have been used for generating textures. Connors and Harlow(1980) generated streaky line textures according to a simple Markov chain that ignores dependence among rows. Haralick and Yokoyama(1979) generated essentially one-dimensional textures. Although they injected some correlation between neighbouring rows by considering co-occurrence matrices, it is almost impossible to specify the spatial, probabilistic structure of the resulting pattern, and their simulation results were also limited in scope.

In Section 7.2, we define the multi-dimensional Markov chain model and simulate patterns corresponding to a variety of choices of neighbourhood systems and parameters. (The selection is made with a view to producing patterns similar to those of Cross and Jain(1983).) In Section 7.3, we discuss the problem of parameter estimation based on a realization of the model. Maximum likelihood estimation is achieved through an iterative algorithm which is similar to the procedure we illustrated in Chapter 2, and, since it is not exactly the Newton-Raphson procedure, it usually converges linearly. In Section 7.4 the ICM method is used to restore images from noisy data, and the problem of parameter estimation from noisy data is also addressed there. In Section 7.5 simulation results related to Section 7.3 and 7.4 are presented and compared. Concluding remarks are given in Section 7.6.

7.2 The multi-dimensional Markov chain (MDMC) model

We now consider an MxN frame. Let $X_i=(X_{i1},X_{i2},...X_{iN})$ denote the i-th row. where $X_{ij}\in\{1,2,...S\}$ for each i and j; $X=(X_1,X_2,...X_M)$ denotes the whole image. The model is based on the causal assumption that the distribution function of one row, conditionally on all past rows, depends only on a few immediately preceding rows. It is known that MRFs provide symmetric texture models, allowing us to consider neighbouring pixels in all directions. For MMRF models, although only a few past pixels are used in the definition of the model, the local dependence involves more pixels. (See Property 2 in Chapter 6.) Similarly, for this new model, we only consider neighbouring pixels in the past rows and in the current row in defining the conditional density of the current row.

Definition: X is called a first-order pairwise-interaction multi-dimensional Markov chain (MDMC) if, for certain functions (g_{ij}), (G_{ij}),

$$P(X_1=x_1) = p(x_1)= \frac{1}{C_1} \exp\{ \sum_{j=1}^{N} g_{1j}(x_{1j}) + \sum_{j=1}^{N-1} G_{1j}(x_{1j},x_{1,j+1})\}, \qquad (7.2.1)$$

and, for i=2,3,...,M,

$$P(X_i=x_i|X_{i-1}=x_{i-1},X_{i-2}=x_{i-2},\ldots,X_1=x_1) = P(x_i|x_{i-1}) \qquad (7.2.2)$$

$$= \frac{1}{C_i(x_{i-1})} \exp\{ \sum_{j=1}^{N} \bar{g}_{ij}(x_{ij}) + \sum_{i=1}^{N-1} G_{ij}(x_{ij},x_{i,j+1})\}, \qquad (7.2.3)$$

where

$$\bar{g}_{ij}(x_{ij}) = \begin{cases} g_{i1}(x_{i1}) + G_{i1}^{(0)}(x_{i-1,1},x_{i1}) + G_{i1}^{(1)}(x_{i-1,2},x_{i1}) & j=1 \\ g_{ij}(x_{ij}) + G_{ij}^{(-1)}(x_{i-1,j-1},x_{ij}) + G_{ij}^{(0)}(x_{i-1,j},x_{ij}) & \\ \qquad + G_{ij}^{(1)}(x_{i-1,j+1},x_{ij}) & 2\leqslant j\leqslant N-1 \\ g_{iN}(x_{iN}) + G_{iN}^{(-1)}(x_{i-1,N-1},x_{iN}) + G_{iN}^{(0)}(x_{i-1,N},x_{iN}) & j=N \end{cases}$$

$$(7.2.4)$$

i-1,j-1	i-1,j	i-1,j+1
i,j-1	i,j	i,j+1

Fig 7.1 Form of neighbourhood for the first-order case

Note that the exponential parts of these (conditional) probability functions are very similar to those of pairwise-interaction MRF's, and that the interaction relationship between neighbouring pixels is represented by the functions G_* and $G_*^{(*)}$. Fig 7.1 shows pixels ij and those neighbours which are considered to have an interaction relationship with it. It follows easily that, in an obvious notation,

$$P(x_{ij}|x_1,x_2,\ldots,x_{i-1},x_{i1},x_{i2},\ldots.x_{i,j-1},x_{i,j+1},\ldots,x_{iN})$$

$$= P(x_{ij}|x_{i-1,j-1},x_{i-1,j},x_{i-1,j+1},x_{i,j-1},x_{i,j+1})$$

$$\propto \exp\{g_{ij}(x_{ij}) + G_{i,j-1}(x_{i,j-1},x_{ij}) + G_{ij}(x_{ij},x_{i,j+1})$$

$$G_{ij}^{(-1)}(x_{i-1,j-1},x_{ij}) + G_{ij}^{(0)}(x_{i-1,j},x_{ij}) + G_{ij}^{(1)}(x_{i-1,j+1},x_{ij})\}. \tag{7.2.5}$$

Different forms of the functions $G*$ and $G*^{(*)}$ or different interaction relationships can be chosen to create different types of image.

If instead we use second-order chains, we can naturally extend the model to the second-order case with neighbourhoods of the form shown in Fig 7.2. For higher-order cases the theroy of high-order Gibbs chains is required.

	i-2,j-1	i-2,j	i-2,j+1	
i-1,j-2	i-1,j-1	i-1,j	i-1,j+1	i-1,j+2
i,j-2	i,j-1	i,j	i,j+1	i,j+2

Fig 7.2 Form of neighbourhood for the second-order case

Note also that the (conditional) probability functions (7.2.1) and (7.2.3) have the same form as that of the first-order Gibbs chain which we examined in Chapter 2. The model can therefore be simulated, from the first row to the last row, in just one scan, whereas existing methods for simulating MRF's have to use iterative procedures (Cross and Jain, 1983, Geman and Geman, 1984).

Cross and Jain(1983) generated some examples of MRF's, according to various settings of the parameters, in order to imitate a variety of real textures. In presenting some images generated from our model, we chose parameters with a view to creating patterns similar to those in Cross and Jain(1983). We can consider either ordered-colour or unordered-colour textures in the cases of both MRF's and MDMC's. We can also adopt different interaction relationships in different sections of the whole image in both classes of model. One of the differences between these two models, which might be to the disadvantage of MDMCs, is that, in MRFs, isotropy can be enforced, so that the interaction relationship between one pixel and its neighbouring pixels in different directions is the same, while, for the MDMCs, we cannot create exactly isotropic cases, since the effects of the interaction functions G_{ij} and $G_{ij}^{(*)}$ in the conditional density of the i-th row, $f(X_i|X_{i-1})$, are different. In other words, since

$$\Pr[X_{ij}|X_{\text{all others}}] \propto \frac{1}{C_{i+1}(X_i)}\exp\{g_{ij}+G_{i,j-1}+G_{i,j}+G_{ij}^{(-1)}+G_{ij}^{(0)} + G_{ij}^{(1)}+G_{i+1,j-1}^{(1)}+G_{i+1,j}^{(0)}+G_{i+1,j+1}^{(-1)}\}, \qquad (7.2.6)$$

we see that the density function for one pixel, conditionally upon all other pixels, depends not only on its 8 neighbours but also on all other pixels in the same row, while that of a second-order MRF depends only on its 8 neighbours. However, we might still be able to generate virtually isotropic images by choosing different parameters for different directions.

The descriptions of the images simulated are as follows. We take

$$\left.\begin{aligned} g_{ij}(x_{ij}) &\equiv 0 \\ G_{ij}(x_{ij},x_{i,j+1}) &= \beta\delta(x_{ij},x_{i,j+1}) \\ G_{ij}^{(-1)}(x_{i-1,j-1},x_{ij}) &= \beta_{-1}\delta(x_{i-1,j-1},x_{ij}) \\ G_{ij}^{(0)}(x_{i-1,j},x_{ij}) &= \beta_{0}\delta(x_{i-1,j},x_{ij}) \\ G_{ij}^{(1)}(x_{i-1,j+1},x_{ij}) &= \beta_{1}\delta(x_{i-1,j+1},x_{ij}) \end{aligned}\right\}, \qquad (7.2.7)$$

and

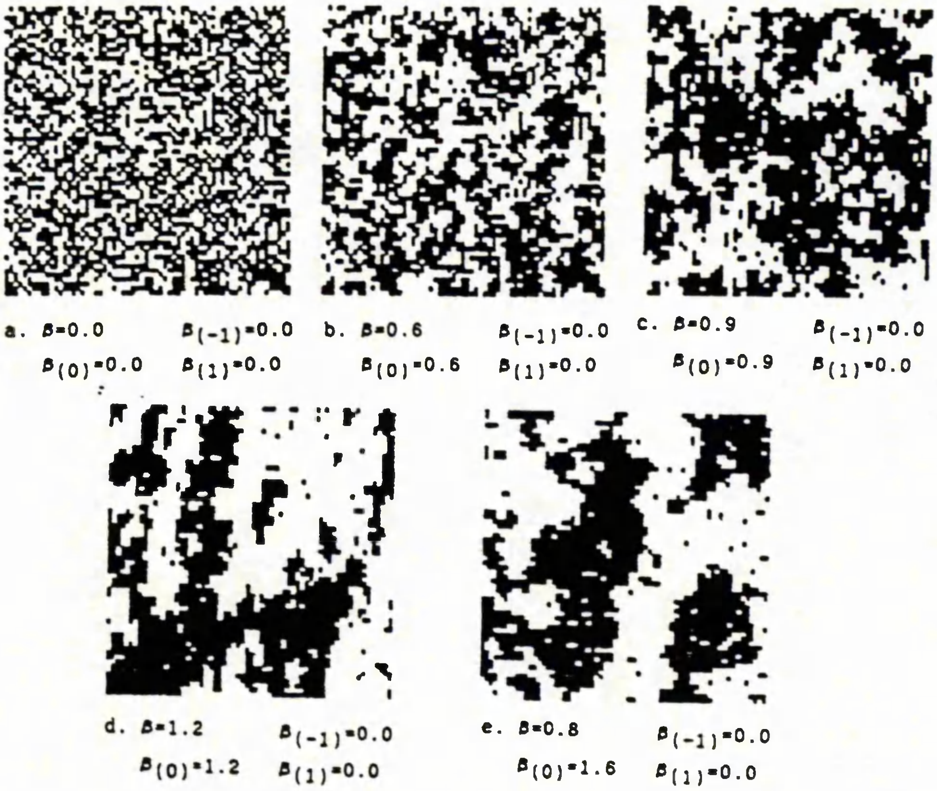
$$\delta(s,t)=\begin{cases} 1 & s=t \\ 0 & \text{otherwise} \end{cases} \qquad s,t \in \{1,2,\ldots S\} \qquad (7.2.8)$$

(1) Pseudo-Isotropic Effects: Fig 7.3 shows five simulated 64×64 binary textures, where the rows generated first are at the top of the patterns. Fig 7.3a represents the "noise", ie, with $\beta=\beta_{-1}=\beta_0=\beta_1=0$, and Fig 7.3b and Fig 7.3c are run-of-the-mill cases. Fig 7.3d is an image with equal parameters in the horizontal and vertical directions, whereas, in Fig 7.3e, the horizontal parameter β is bigger than β_0. We find therefore that there is more similarity among rows than among columns and that the model is vertically directional.

(2) Anisotropic Effects: Fig 7.4 shows extremely anisotropic 64×64 binary images. The parameters β, β_0, β_{-1} and β_1 control horizontal, vertical, NW-SE and NE-SW directional interactions respectively. For the three images, the parameter for one direction is large relative to those for the other three.

(3) Ordered Patterns: By enforcing negative interactions between one

pixel and its four nearest pixels, we can simulate the chessboard-like pattern shown in Fig 7.5. Note that a black pixel is very likely to be surrounded by four white pixels, and vice versa.



a. $\beta=0.0$ $\beta_{(-1)}=0.0$ $\beta_{(0)}=0.0$ $\beta_{(1)}=0.0$

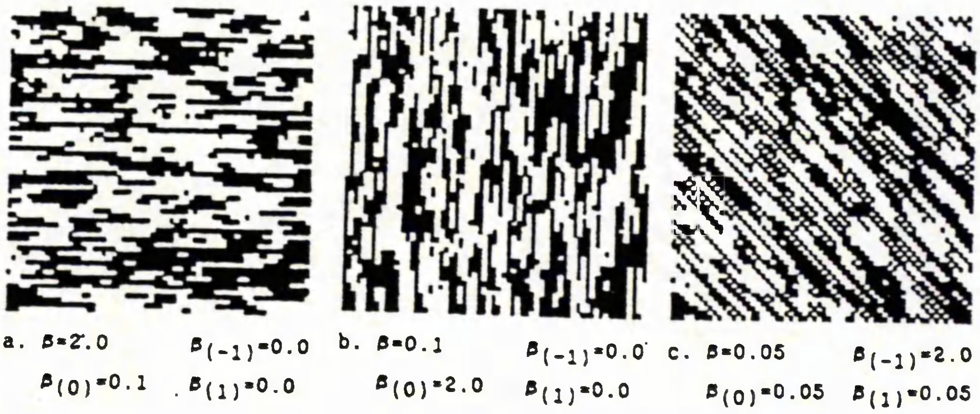
b. $\beta=0.6$ $\beta_{(-1)}=0.0$ $\beta_{(0)}=0.6$ $\beta_{(1)}=0.0$

c. $\beta=0.9$ $\beta_{(-1)}=0.0$ $\beta_{(0)}=0.9$ $\beta_{(1)}=0.0$

d. $\beta=1.2$ $\beta_{(-1)}=0.0$ $\beta_{(0)}=1.2$ $\beta_{(1)}=0.0$

e. $\beta=0.8$ $\beta_{(-1)}=0.0$ $\beta_{(0)}=1.6$ $\beta_{(1)}=0.0$

Fig 7.3 Some pseudo-isotropic examples

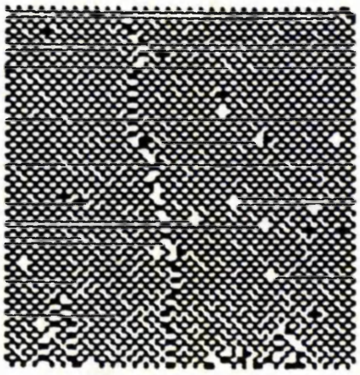


a. $\beta=2.0$ $\beta_{(-1)}=0.0$ $\beta_{(0)}=0.1$ $\beta_{(1)}=0.0$

b. $\beta=0.1$ $\beta_{(-1)}=0.0$ $\beta_{(0)}=2.0$ $\beta_{(1)}=0.0$

c. $\beta=0.05$ $\beta_{(-1)}=2.0$ $\beta_{(0)}=0.05$ $\beta_{(1)}=0.05$

Fig 7.4 Examples of anisotropic effects



Ordered pattern: $\beta=\beta_{(0)}=-1.8$

$\beta_{(-1)}=\beta_{(1)}=0.0$

Fig 7.5 Chessboard-like pattern

(4) Attraction-Repulsion Effects: As in Cross and Jain(1983), an attraction-repulsion process involves positive interaction between "near" neighbouring pixels, resulting in clustering but negative interaction between "far" neighbouring pixels, to inhibit the growth of clusters. If the interaction with far pixels is also positive, large areas with the same colour would be generated. Fig 7.6 illustrates inhibition in both diagonal directions: typical patterns have many horizontal and vertical lines. Fig 7.7 shows two images simulated from the second-order model. The twelve kinds of interaction functions associated with twelve neighbouring pixels (Fig 7.2) are taken to be of the same form as those of the first order case in (7.2.6). Twelve parameters are therefore listed there according to the positions of the corresponding pixels in the neighbourhood.

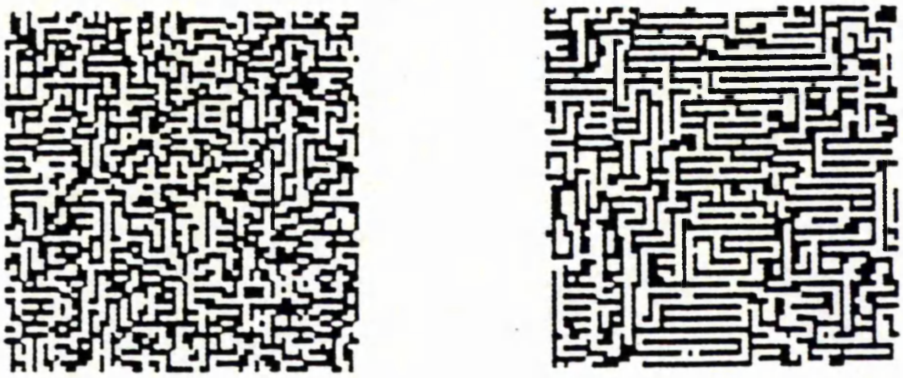
(5) Changing-Interaction Effects: For the above four cases, the interactions among neighbouring pixels are the same over the entire images. The texture model does allow us to consider different interaction relationship in different parts of the frame, and also to simulate it easily. Fig 7.8a shows an example in which interactions change gradually in the horizontal direction, ie

$$G_{ij}^{(-1)} \equiv G_{ij}^{(1)} \equiv 0; \quad G_{ij}(x_{ij}, x_{i,j+1}) = \frac{j}{N}\beta\delta(x_{ij}, x_{i,j+1})$$

and $G_{ij}^{(0)}(x_{i-1,j}, x_{ij}) = \beta_0\delta(x_{i-1,j}, x_{ij})$. In Fig 7.8b, on the other hand, which shows vertical change, we have $G_{ij}^{(-1)} \equiv G_{ij}^{(1)} \equiv 0$, and

$G_{ij}(x_{ij},x_{i,j+1})=\beta\delta(x_{ij},x_{i,j+1})$; $G_{ij}^{(0)}(x_{i-1,j},x_{ij})=\frac{i}{M}\beta_0\delta(x_{i-1,j},x_{ij})$.

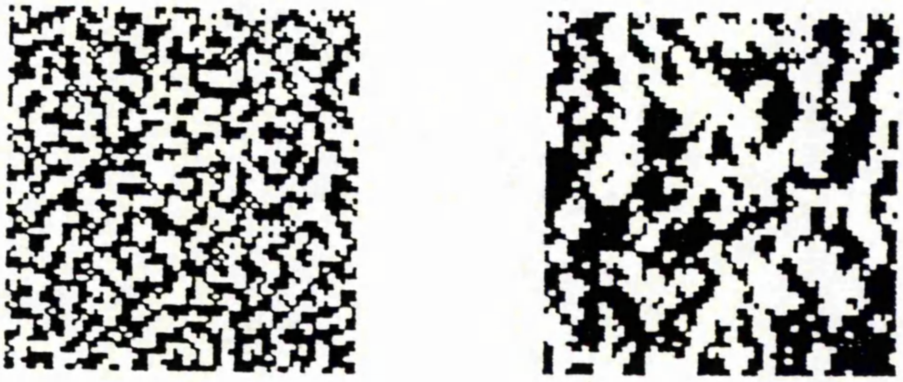
Fig 7.8c shows a pattern with changing parameters in both horizontal and vertical directions.



a. $\beta=0.2$ $\beta_{(-1)}=-0.8$
$\beta_{(0)}=0.2$ $\beta_{(1)}=-0.8$

b. $\beta=1.0$ $\beta_{(-1)}=-1.5$
$\beta_{(0)}=1.5$ $\beta_{(1)}=-1.5$

Fig 7.6 Examples of attraction-repulsion in 1st order case



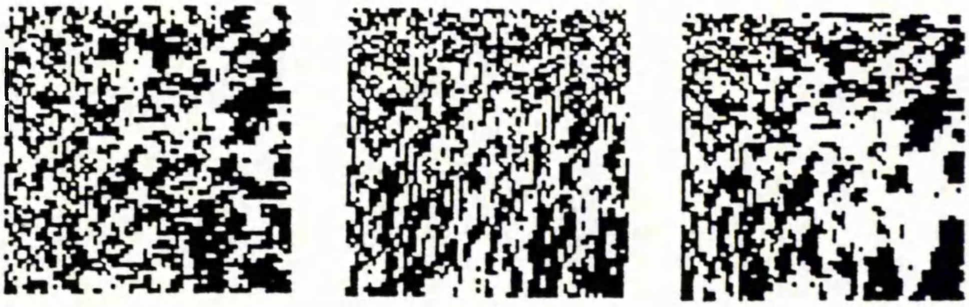
a.
-0.3 -0.3 -0.3
-0.3 0.35 0.35 0.35 -0.3
-0.3 0.45 0.45 -0.3

b.
-0.2 -0.2 -0.2
-0.2 0.7 0.7 0.7 -0.2
-0.2 0.9 0.9 -0.2

Fig 7.7 Examples of attraction-repulsion in 2nd order case

(6) Multi-Colour Patterns: Fig 7.9 shows one four-state picture and one five-state picture. They are simulated from the second-order model with the same parameters as listed there.

All these images look very realistic. The first four examples are very similar to those simulated by Cross and Jain(1983) from MRFs. This results from similar consideration of the interaction relationships. We can also note the re-appearance of the effect that these images appear out of focus, which is intrinsic to texture models based on stochastic assumptions.



a. Horizontal effect:

$\beta_{(-1)}=\beta_{(1)}=0.0$

$\beta=1.2 \quad \beta_{(0)}=0.5$

b. Vertical effect:

$\beta_{(-1)}=\beta_{(1)}=0.0$

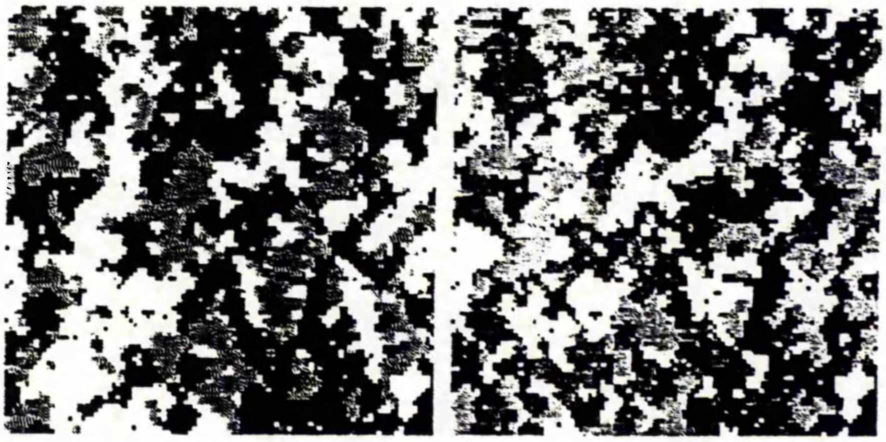
$\beta=0.7 \quad \beta_{(0)}=1.5$

c. Both direction effect:

$\beta_{(-1)}=\beta_{(1)}=0.0$

$\beta \ = \beta_{(0)} \ =1.5$

Fig 7.8 Examples of changing parameter effects



	-0.3	-0.2	-0.3	
-0.2	0.8	1.0	0.8	-0.2
-0.2	1.8		1.8	-0.2

Fig 7.9 Some multi-colour patterns

7.3 Parameter estimation

In this section, we express the (conditional) probability functions in the last section in terms of parameters, and assume that the the exponential parts of these densities are linear in the parameter, β. Thus,

$$p(x_1|\beta) = \frac{1}{C_1(\beta_1)}\exp\{\beta' Z_1(x_1)\} \tag{7.3.1}$$

$$p(x_i|x_{i-1},\beta) = \frac{1}{C_i(x_{i-1},\beta)}\exp\{\beta' Z_i(x_{i-1},x_i)\}, \tag{7.3.2}$$

where β is a multi-dimensional, unknown parameter and the Z_i are vector functions, which have still the interaction form shown in (7.2.1)-(7.2.4). Then

$$P(X=x,\beta) = p(x,\beta) = p(x_1|\beta) \prod_{i=1}^{M} p(x_i|x_{i-1},\beta) \tag{7.3.3}$$

and

$$\ell p(x,\beta) = \log p(x,\beta) = \beta' \sum_{i=1}^{M} Z_i - \sum_{i=1}^{M} \log C_i, \tag{7.3.4}$$

where $C_i=C_i(x_{i-1},\beta)$ is the normalizing factor of the i-th row. Thus,

$$\frac{\partial}{\partial\beta}\ell p(x,\beta) = \sum_{i=1}^{M} Z_i - \sum_{i=1}^{M} \frac{1}{C_i}\frac{\partial}{\partial\beta}C_i. \tag{7.3.5}$$

Note that the normalizing factors C_i satisfy

$$C_i(x_{i-1},\beta) = \sum_{X_i} \exp\{\beta' Z_i(x_{i-1},X_i)\}. \tag{7.3.6}$$

Thus

$$\frac{\partial}{\partial\beta}\ell p(x,\beta) = \sum_{i=1}^{M} Z_i(x_{i-1},x_i) - E[Z_1(X_1)|\beta] - \sum_{i=2}^{M} E[Z_i(x_{i-1},X_i)|x_{i-1},\beta] \tag{7.3.7}$$

and

$$\frac{\partial^2}{\partial\beta^2}\ell p(x,\beta) = -\sum_{i=1}^{M} \mathrm{Var}[Z_i(x_{i-1},X_i)|x_{i-1},\beta]. \tag{7.3.8}$$

In the above formulae, $\frac{\partial}{\partial\beta}$ denotes the partial derivative vector, $\frac{\partial^2}{\partial\beta^2}$ denotes the second-order partial derivative matrix and Var denotes a covariance matrix or a conditional covariance matrix. The results in Chapter 2 imply that $C_i(x_{i-1},\beta)$ and $E[Z_i(x_{i-1},X_i)|x_{i-1},\beta]$ can be computed easily. If therefore β is one-dimensional, we can use a search method to find the maximum point of the log-likelihood function (7.3.4). When β is multi-dimensional, as we pointed in Chapter 2 and Chapter 3, it is not easy to compute $\mathrm{Var}(Z_i|x_{i-1},\beta)$ exactly. As a consequence, we cannot use an exact Newton-Raphson iterative procedure to maximize $\ell p(x,\beta)$. If, however, we can find a positive-definite matrix A_0 such that

$$\frac{\partial^2}{\partial\beta^2} \ell p(x,\beta) \preccurlyeq -A_0 \prec 0 \quad \forall\ x,\ \beta, \tag{7.3.9}$$

where $\preccurlyeq$ ($\prec$) denote the (strict) Loewner ordering, we can still use the following iteration, which is similar to that in Chapter 2, to obtain maximum likelihood estimates. This iteration converges linearly.

$$\beta^{(k+1)} = \beta^{(k)} - A^{-1}[\sum_{i=1}^{M} Z_i - \sum_{i=1}^{M} E(Z_i|x_{i-1},\beta^{(k)})],\ k=0,1,\dots, \tag{7.3.10}$$

where $A \succ A_0$ is also a positive definite matrix.

Although it is difficult to check whether or not (7.3.9) holds, $\frac{\partial^2}{\partial\beta^2}\ell p(x,\beta)$ is usually a negative definite matrix. Thus, in practical computation, when A is "large" enough, the iteration will converge to the maximum likelihood estimates. Some simulation results together with results of reconstruction and parameter estimation in the case where only noisy data are available are provided in Section 7.5.

7.4 Image restoration and parameter estimation based on noisy data

As in previous chapters, we assume that, given the original image X=x, the noisy data on different pixels are conditionally independent, and that the noisy variable for pixel (i,j) depends only on x_{ij}. It is not necessary to have the same conditional distribution for each pixel. Thus

$$P(Y=y|X=x,\theta) = \prod_{i=M}^{M} \prod_{j=1}^{N} f_{ij}(y_{ij}|x_{ij},\theta), \tag{7.4.1}$$

where θ is an unknown parameter.

7.4.1 Image restoration

For the first-order model defined by (7.2.1)-(7.2.3), the marginal probability function for x_{ij}, the colour on pixel (i,j), conditionally upon all observed noisy data and on the true colours at all other pixels, satisfes

$$p[X_{ij}=x_{ij}|Y=y,x\backslash\{x_{ij}\}] \propto f_{ij}(y_{ij}|x_{ij})Pr[x_{ij}|x\backslash\{x_{ij}\}] \qquad (7.4.2)$$

The ICM method of Besag(1986) is a convergent method to maximize $P[X|Y=y,\beta,\theta)$, given the parameter β in the original image model and θ in the noise model. Noting the expression for $Pr[x_{ij}|x\backslash\{x_{ij}\}]$ in (7.2.6), we see that, apart from the term $(C_{i+1}(x_i))^{-1}$, (7.4.2) is the same as that of the MRF, However, $C_{i+1}(x_i)$ can be computed by the recursive technique in Chapter 2, and, as a consequence, we can use the ICM technique to obtain the Maximum Probability Reconstruction. Although the method incurs a heavier computational burden than in the case of MRFs, some techniques for computing the normalizing constant of the Gibbs chain can still be used to reduce the computational demands if we update the colours pixel-by-pixel and row-by-row. The techniques involve two groups of vectors; one is defined by forward recursion, and the i-th vector in the group depends only on the previous functions g and G; the other is defined by backward recursion, and the i-th one in the group depends only on the following functions g and G. Thus, in order to update colours from the first pixel to the last pixel in the i-th row, we can first compute all backward vectors, then compute forward recursion vectors element by element as the new colours are being chosen for each pixel at a time. Therefore compared with the MRF case, it is only the computational demands of these two groups of vectors that increase for each row: it is not necessary to compute $C_{i+1}(x_i)$ for each pixel in the i-th row. If S, the number of possible colours, is not large, the additional computational burden is not heavy.

Just as in the case of the MRF model, the ICM method is only of local convergence for our model. Instead of using the ICM method, we can adopt a rule which is similar to that for the MMRF model. This results in a non-iterative procedure, such that the image can be restored through a single pass. To be precise, we still start updating from the first row, and for the j-th pixel in the i-th row, we may either use

$$P[X_{ij}|\hat{x}_1,\hat{x}_2,\ldots,\hat{x}_{i-1},y_1,y_2,\ldots,y_i]=P[X_{ij}|\hat{x}_{i-1},y_i]$$

or

$$P[X_{ij}|\hat{x}_1,\ldots,\hat{x}_{i-1},\hat{x}_{i1},\hat{x}_{i2},\ldots\hat{x}_{i,j-1},y_1,y_2,\ldots y_i] =$$

$$P[X_{ij}|\hat{x}_{i-1},\hat{x}_{i1},\ldots,\hat{x}_{i,j-1},y_i]$$

to decide a colour for pixel (ij). Note that both the above conditional densities can be maximized by the recursive techniques.

7.4.2 Simultaneous parameter estimation and restoration

We now consider estimation of unknown parameters, β and θ. Since the original data are not available, we examine the use the EM algorithm. As in the case of the MRF model and MMRF model, it is infeasible to carry out the EM algorithm exactly. For the MRF model and MMRF model we have, in Chapter 5 and Chapter 6 respectively, developed the iterative procedure proposed by Besag(1986) for carrying out restoration and parameter estimation simultaneously. The approach used noisy data and the currently restored image to choose new parameters by using the EM algorithm to maximize a single pseudo-likelihood. For the MDMC model introduced in this chapter, based on the same idea as used in the previous chapter, we propose the following iterative procedure.

Suppose that $f_{ij}(y_{ij}|x_{ij},\theta) = \exp\{d_{ij}(x_{ij},y_{ij},\theta)\}$ and write y as $(y_1,y_2,\ldots,y_M)$ where y_i represents the noise or the observed data in the i-th row.

1. Obtain initial estimate $\hat{\beta}$ and $\hat{\theta}$.
2. Carry out ICM as in the last subsection, based on the current $\hat{\beta}$ and $\hat{\theta}$, thereby obtaining a new $\hat{x}$.
3. Obtain new values for $\hat{\beta}$ and $\hat{\theta}$ by maximizing the pseudo-likelihood

$$P(y_1|\beta,\theta)\prod_{i=2}^{M}\Pr(y_i|\hat{x}_{i-1},\beta,\theta). \qquad (7.4.3)$$

4. Return to 2 and continue for a fixed number of cycles or until an appropriate stopping rule is satisfied.

<u>Remarks</u>

(1) Note that both of the conditional densities $p(x_i|\hat{x}_{i-1},\beta)$ and $\Pr(x_i|\hat{x}_{i-1},y_i,\beta,\theta)$ are of Gibbs chain forms. The EM algorithm can therefore be used to maximize (7.4.3).

(2) Both Step 2 and Step 3 are in fact iterative procedures. At early stages in the whole procedure it is however not necessary to carry them out until (approximate) convergence. Experimental results suggest that a small number of iterations is sufficient,

for the ICM stage, in particular. As the procedure reaches convergence, of course, the component Steps 2 and 3 converge quickly in any case.

(3) Instead of the ICM method, we can use other restoration methods, such as the rule described at end of the last subsection.

(4) In Step 2, we might have tried, instead, to maximize the following two functions in order to obtain new estimates for the parameters, β and θ, respectively:

$$p(\hat{x}_1|\beta) \prod_{i=2}^{M} p(\hat{x}_i|\hat{x}_{i-1},\beta) \tag{7.4.4}$$

$$\prod_{ij} f_{ij}(y_{ij}|\hat{x}_{ij},\theta). \tag{7.4.5}$$

Suppose $X_{i-1} = x_{i-1}$ is not missing and consider the single factor $p(x_i|x_{i-1},\beta)$. When y_i is available, it is preferable to maximize $p(y_i|x_{i-1},\beta,\theta)$ rather than to maximize $p(\hat{x}_i|x_{i-1},\beta)$ and $p(y_i|\hat{x}_i,\theta)$ respectively, because $\hat{x}_i$ contains less information than y_i. Therefore, the results obtained by maximizing (7.4.3) should be better than those obtained by maximizing (7.4.4) and (7.4.5). We can observe this phenomenon for the MRF model and the MMRF model in Chapter 5 and Chapter 6, respectively.

(5) Similarly to the case with other models, little is known, theoretically, of the convergence properties of the above procedure. In our numerical experiments, in which we estimated four interaction parameters, corresponding to four directions, along with the variance of the Gaussian noise, by choosing positive initial estimates, the procedure always converged. Convergence did not always obtain if, instead, we maximized (7.4.4) and (7.4.5) in Step 3.

7.5 Numerical results

Figs 7.10-7.12 show some simulated results of the iterative procedure described in the last section, applied to 100×100 frames. The images at the upper-left corner were simulated from the model with parameters indicated as OP (original parameters). These parameters represent interaction relationships in four directions as expressed in (7.2.7). EP denotes the parameters estimated from the original image by the method described in Section 7.3. The upper-right images display the results of maximum likelihood classification after corruption with additive normal noise, with the stated variances. We started the iteration from initial estimates $\hat{\beta}=\hat{\beta}_{-1}=\hat{\beta}_0=\hat{\beta}_1=0.5$, and $\hat{\sigma}=0.3$ in all three cases. The lower-left images correspond to two cycles of iteration, while the lower-right indicates the converged state. IEP denotes the estimated parameters obtained by the iterative procedure in Section 7.4, and REP denotes parameter estimates obtained from the restored images, treated as true realizations of the model, by the method in Section 7.3.

Note that Fig 7.10 and Fig 7.11 are for the same original image but with noise of different variances. In the case of low level noise (Fig 7.10), both IEP and REP are quite close to the original parameters, while for higher noise variance IEP is better than REP.

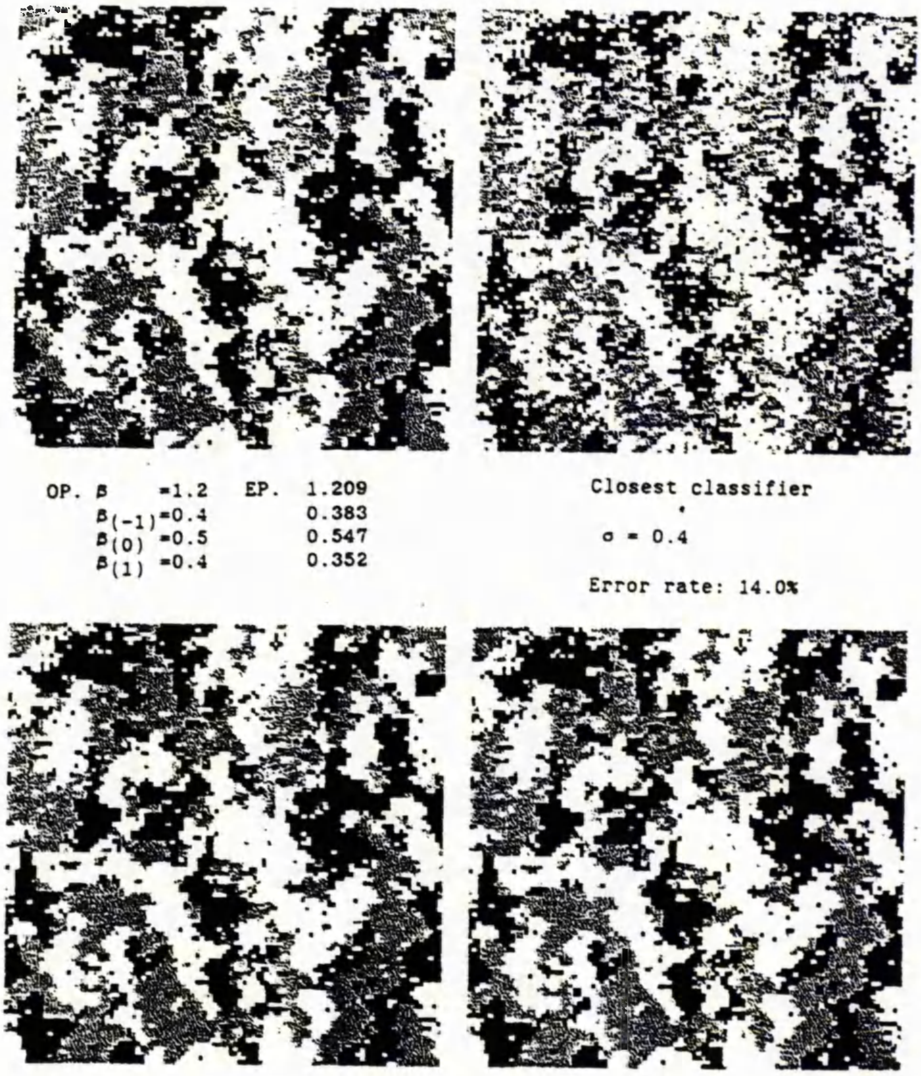
Fig 7.12 is an example to show the difference in performance if, in Step 3 of the algorithm, one maximizes (7.4.4) and (7.4.5), rather than (7.4.3). The iterative procedure based on (7.4.4) and (7.4.5) did not converge for this example since, after several cycles, parameters β_{-1} and β_1 became negative and the error rate for the currently restored image increased. This had an adverse effect on parameter estimation. From the figure we can also identify the effect by noticing that, although the reconstruction is quite good, one of REP values is negative, very unlike the original parameter and the IEP.

Although it can happen that REP is better than IEP, IEP performed better in most of our numerical experiments.

7.6 Concluding comments

In this chapter we have developed a multi-dimensional Gibbs chain model and used it as a model for image textures. We have demonstrated that, by choosing the underlying parameters appropriately, textures can be created that are very similar to those realised, after considerably more computational effort, from Markov random field models. The problems of image restoration from noisy data and of parameter estimation have been attacked using Besag's(1986) ICM algorithm coupled with an approximate version of the EM algorithm.

Important developments for the future include refinements of the algorithmic aspects of the methodology and theoretical investigation of the convergence properties of the iterative procedures.



OP. β $=1.2$ EP. 1.209

$\beta_{(-1)}=0.4$ 0.383

$\beta_{(0)}=0.5$ 0.547

$\beta_{(1)}=0.4$ 0.352

Closest classifier

$\sigma = 0.4$

Error rate: 14.0%

Error rate: 6.74%

$\hat{\beta}$ $=1.141$

$\hat{\beta}_{(-1)}=0.409$ $\hat{\sigma}=0.397$

$\hat{\beta}_{(0)}=0.512$

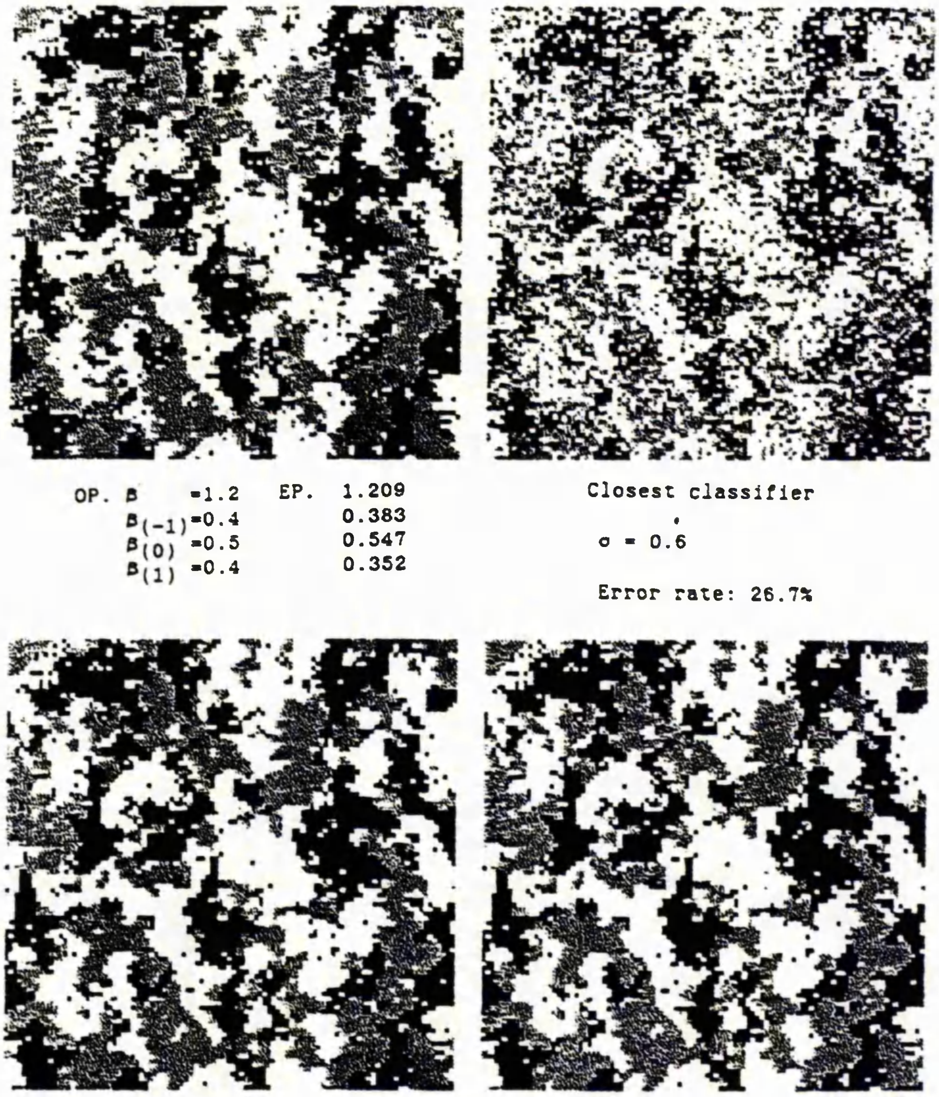
$\hat{\beta}_{(1)}=0.406$

Error rate: 6.69%

	1.143		1.328
IEP.	0.401	REP.	0.367
	0.527		0.508
	0.395		0.355

$\hat{\sigma}=0.397$

Fig 7.10 Example 1 of the iterative procedure



OP. β $=1.2$ EP. 1.209
$\beta_{(-1)}=0.4$ 0.383
$\beta_{(0)}=0.5$ 0.547
$\beta_{(1)}=0.4$ 0.352

Closest classifier

$\sigma = 0.6$

Error rate: 26.7%

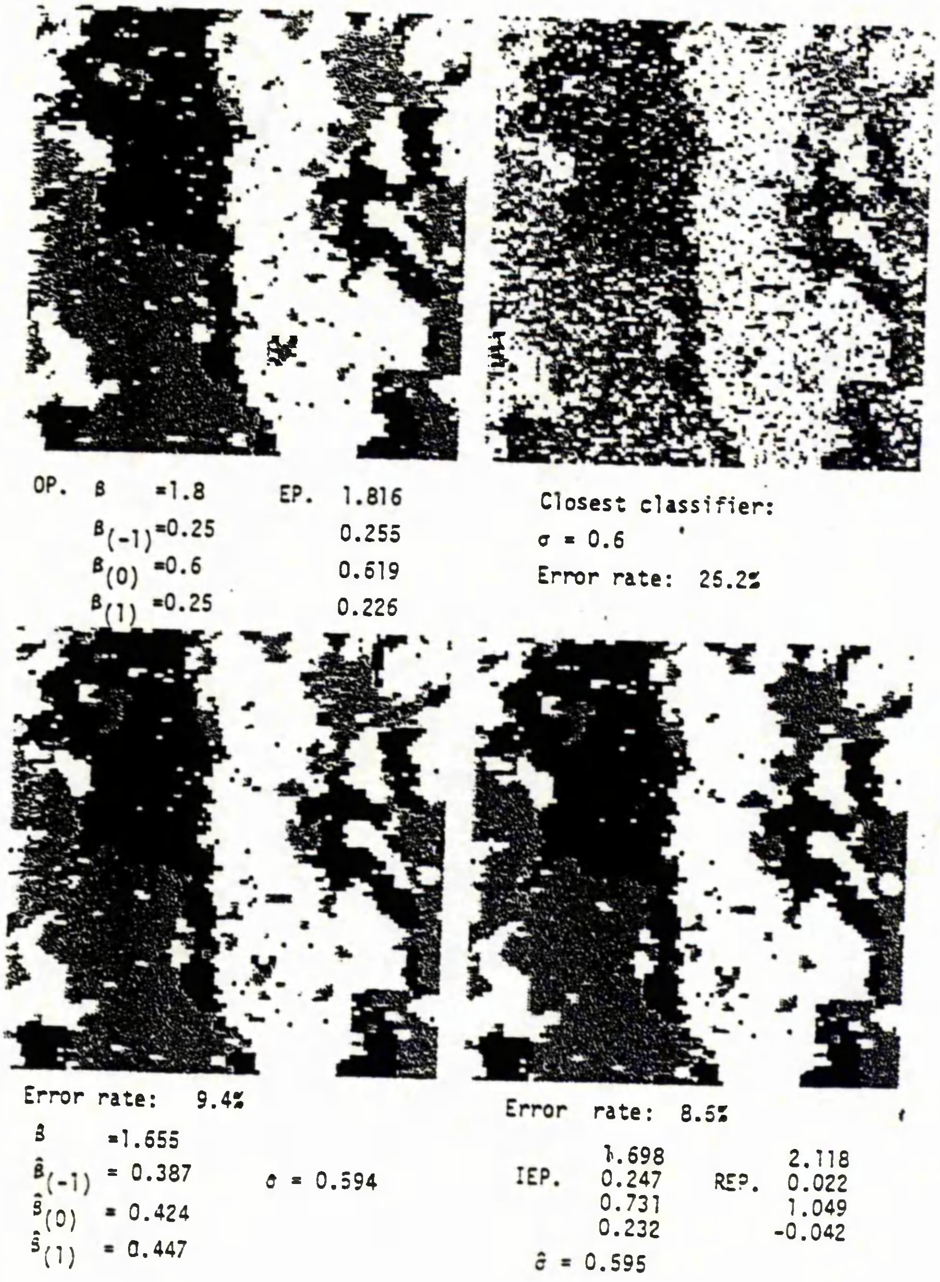
Error rate: 15.25%
$\hat{\beta}$ $=1.078$
$\hat{\beta}_{(-1)}=0.430$ $\hat{\sigma}=0.585$
$\hat{\beta}_{(0)}=0.468$
$\hat{\beta}_{(1)}=0.458$

Error rate: 14.94%

	1.073		1.419
IEP.	0.401	REP.	0.309
	0.563		0.710
	0.407		0.272

$\hat{\sigma}= 0.586$

Fig 7.11 Example 2 of the iterative procedure



OP. β =1.8 EP. 1.816

$\beta_{(-1)}$=0.25 0.255

$\beta_{(0)}$ =0.6 0.619

$\beta_{(1)}$ =0.25 0.226

Closest classifier:

σ = 0.6

Error rate: 25.2%

Error rate: 9.4%

$\hat{\beta}$ =1.655

$\hat{\beta}_{(-1)}$ = 0.387 $\hat{\sigma}$ = 0.594

$\hat{\beta}_{(0)}$ = 0.424

$\hat{\beta}_{(1)}$ = 0.447

Error rate: 8.5%

IEP. 1.698 REP. 2.118

0.247 0.022

0.731 1.049

0.232 -0.042

$\hat{\sigma}$ = 0.595

Fig 7.12 Example 3 of the iterative procedure

Chapter 8

Discussion And Concluding Remarks

This thesis mainly concentrates on spatial statistical models associated with applications to image analysis. Our aim is to explore the properties of the models and to examine the problem of parameter estimation from the original image data and, in particular, from the noisy versions of the scene.

In Chapter 2, we developed a recursion technique which enables us to deal with one-dimensional versions of Markov random fields, specifically, to achieve maximum likelihood estimation for the underlying parameters and to carry out the EM algorithm for parameter estimation when only noisy data are available. Although, unfortunately, this technique cannot be extended to two- or multi-dimensional models, it can still be used in many cases. We remark that it is used in Chapter 3, Chapter 4 and Chapter 5 for two-dimensional Markov field models, and that it enables us to introduce the new model, the MDMC model, in Chapter 7.

Chapter 3 and Chapter 4 concentrate on inference for two-dimensional Markov random fields. We obtained a matrix expression for partition functions for general models, and a more explicit matrix representation for multi-colour Ising models. This expression enables us to locate the positions of critical points. We also examined the asymptotic properties of asymmetric, two-colour Ising models. For general models, in Chapter 4, we explored the asymptotic properties of certain statistics under an "independence' or a "near independence" condition.

An important idea in this thesis is to regard a set of pixels as one "single point". It preserves the Markovian properties of MRF models. This idea was first adopted, in Chapter 3, in introducing the block pseudo-likelihood function, which is an extension of point pseudo-likelihood. Although the parameter estimates from both point pseudo-likelihood functions and line pseudo-likelihood functions are almost equally good, we can, from Chapter 5, note that line pseudo-likelihood performed better than the other one, when only noisy data were available and our modified EM algorithm was used. The idea is then used in Chapter 4. By assuming "independence" between rows, the limiting results of one-dimensional stochastic processes were used in two-dimensional cases, thereby enabling us to establish approximate normal results. This idea is then used in Chapter 7 together with the recursion technique.

For the Markov random field itself and its application to image analysis there are many aspects to be developed. They include asymptotic properties, special structures for practical phenomena, edge structures in the potential function, identification of the model, choice of the neighbourhood system, testing, influence analysis, etc.

Another main idea in this thesis is that the currently restored image might be used together with noisy data in iterative procedures for simultaneous parameter estimation and image reconstruction. The EM algorithm is then used at each cycle of the iteration, which is developed from Besag's(1986) procedure. For MRF models, Chapter 5 presented a simulation study of this procedure with different kinds of local conditional densities. The same procedure was also adopted in Chapter 6 and Chapter 7 for the MMRF model and the MDMC model, respectively. Quite good results have been obtained in terms of estimation of parameters in both the original model and, particularly, in the noise model, and in terms of image restoration, for all three sorts of prior random field model considered in this thesis.

In Chapter 6, we extended the MMRF model to the three-dimensional case. A generalized F-G-H algorithm for restoration was then proposed.

In Chapter 7, based on the results for Gibbs chains, we introduced the MDMC model. Although it is a causal-dependence random field model, textures can be simulated, by suitable choice of parameters, that are similar to those generated from MRF models, and, very importantly, the simulation procedure is computationally much more economical.

In this thesis, we have examined the three problems mentioned in Section 1.1. The author would like to remark finally that although, under a fixed type of model, restoration is not very sensitive to the values of parameters, the parameters should lie in a suitable region. Therefore, choice of the model and parameter estimation are very important for image analysis.

Appendix 1

——C(α,β) at boundaries

The following results are used in Section 3.3.

When $\alpha=0$ ($\beta=0$), X consists of N (M) independent Markov chains with the same distribution function. The matrix method for defining the normalizing constant can be simply used in those one-dimensional Markov chains, and yields

$$C(\alpha,0) = \left[\text{trace}\begin{bmatrix} e^{\alpha} & e^{-\alpha} \\ e^{-\alpha} & e^{\alpha} \end{bmatrix}^N\right]^M = [(e^{\alpha} + e^{-\alpha})^N + (e^{\alpha} - e^{-\alpha})^N]^M$$

and

$$C(0,\beta) = [(e^{\beta} + e^{-\beta})^M + (e^{\beta} - e^{-\beta})^M]^N.$$

As $\alpha=0$, $\psi_j \equiv \log(\cosh 2\beta+1)$. From (3.3.7) and (3.3.8)

$$\sum_{\lambda} \lambda^M = 2^{\frac{1}{2}MN} \sum_{i=1}^{N} \begin{bmatrix} N \\ i \end{bmatrix} (\sinh 2\beta)^{iM} (\cosh 2\beta+1)^{\frac{1}{2}MN-iM}$$

$$= (e^{\beta}+e^{-\beta})^{MN}\{1 + [(e^{\beta}-e^{-\beta})/(e^{\beta}+e^{-\beta})]^M\}^N = C(0,\beta).$$

As $\beta=0$, again, from (3.3.7) and (3.3.8), all λ are zero except $\lambda^{+}_{\max}$. In order to prove $(\lambda^{+}_{\max})^M = C(\alpha,0)$, it is required to show that

$$2^{2N} \prod_{j=1}^{N} (\cosh 2\alpha - \sinh 2\alpha \cos\theta_j) = [(e^{\alpha}+e^{-\alpha})^N + (e^{\alpha}-e^{-\alpha})^N]^2, \quad (*)$$

where $\theta_j=(2j-1)\pi/N$. In fact, the left-hand side of (*) is

$$2^{2N} \left| \prod_{j=1}^{N} (e^{-\alpha}\cos\frac{\theta_j}{2} + ie^{\alpha}\sin\frac{\theta_j}{2}) \right|^2 =$$

$$\left| \prod_{j=1}^{N} [(e^{\alpha}+e^{-\alpha})+(e^{\alpha}-e^{-\alpha})\exp(-i\theta_j)] \right|^2.$$

Since $(e^{\alpha}+e^{-\alpha})+(e^{\alpha}-e^{-\alpha})\exp(-i\theta_j)$, $1 \leqslant j \leqslant N$, are N roots of the equation

$$(x-e^{\alpha}-e^{-\alpha})^N + (e^{-\alpha}-e^{\alpha})^N = 0,$$

(*) is therefore proved. We have therefore proved that, with the help of (3.3.7) and (3.3.8), (3.3.3) provides the representation of the partition function for $\{\alpha \geqslant 0,\ \beta \geqslant 0\}$.

Appendix 2

——Autocorrelations of Markov chain

The following results are used in Chapter 4.

Consider a strictly stationary Markov chain with state space $\{i_1,i_2,\ldots i_S\}$, equilibrium probabilities $P=(P_1,P_2,\ldots P_S)'$ and transition matrix $Q=(p_{ij})_{S\times S}$. Suppose $\lambda_1<\lambda_2\ldots<\lambda_{S-1}<\lambda_S=1$ are the eigenvalues of Q, with eigenvectors η_1, η_2, ..., η_S, respectively, where $\eta_S=(1,1,\ldots 1)'$. Clearly,

$$P'\eta_i = P'Q\eta_i = \lambda_i P'\eta_i$$

Then $P'\eta_i = 0. \qquad 1\leqslant i\leqslant S-1$

Clearly, $P'\eta_S = 1$

Let $\xi=(i_1,i_2,\ldots i_S)'$; $\eta=(\eta_1, \eta_2,\ldots\eta_S)$; $\xi_1=\eta^{-1}\xi=(\xi_{11},\ldots\xi_{1S})'$;

and

$$d(P) = \boldsymbol{diag}\{P_1,P_2,\ldots P_S\}.$$

Then,

$$Ex_i = P'\xi = P'\eta\xi_1 = \xi_{1S}$$

and

$$\begin{aligned}Ex_ix_{i+j} &= \xi'd(P)Q^j\xi = \xi_1'\eta'd(P)Q^j\eta\xi_1\\ &= \xi_1'\eta'd(P)\eta\,\boldsymbol{diag}\{\lambda_1^j,\ldots\ldots,\lambda_S^j\}\xi_1\\ &= \sum_{i=1}^{S}\mu_i\lambda_i^j = \sum_{i=1}^{S-1}\mu_i\lambda^j + \mu_S,\end{aligned}$$

where $\mu_S = [\xi_{1S}]^2$. Therefore, the autocorrelations $\{AC(i)\}$ of the Markov chain have the form

$$AC(i) = \sum_{i=1}^{S-1}\mu_i\lambda_i^j.$$

References and Bibliography:

Abend K., Harley T.J. and Kanal L.N.(1965) Classification of binary random patterns. IEEE Trans. Inform. Theory, IT-11, 533-544

Baum L.E. and Eagon J.A.(1967) An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology, Bull Amer. Math. Soc. 73, 360-363.

Baum L.E. and Petrie T.(1966) Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Statist. 37, 1554-1563.

Baum L.E., Petrie T., Soules G. and Weiss N.(1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Statist., 41, 164-171.

Baxter R.J.(1972) Partition function of eight-vertex lattice model, Ann. Physics, 70, 193-228.

Besag J.E.(1974) Spatial interaction and the statistical analysis of lattice system (with discussion), J.R. Statist. Soc., B 36, 192-236.

———, (1975) Statistical analysis of non-lattice data, The Statistician, 24, 179-195.

———,(1976) Parameter estimation for Markov fields, Dept. of Statistics Tech. Report 108, Series 2, (Princeton, NJ, Princeton University).

———,(1978) Some methods of statistical analysis for spatial data, Bull. Int. Statist. Inst., 47, 77-92.

———,(1986) On the statistical analysis of dirty pictures (with discussion), J.R. Statist. Soc., B 48, 259-302

Bloemena A.R. (1964) Sampling From A Graph, Math. Center Tract 2, Amsterdam.

Bushell P.J.(1973) Hilbert's metric and positive contraction mappings in a Banach space, Arch. Rat. Mech. Math.,52, 330-338.

Chalmond B.(1988) An iterative Gibbsian technique for simultaneous structure estimation and reconstruction of M-ary images,

Chellappa R.(1985) Two dimensional discrete Gaussian Markov random fields for image processing, in: L.N. Kanal & A. Rosenfeld (eds)

Progress in Pattern Recognition, (Amsterdam, North-Holland).

Connors R.W. and Harlow C.A. (1980) A theoritical comparison of texture algorithms. IEEE Trans. Pattern Anal. Machine. Intell., PAMI-2, 204-222.

Cross G.R. and Jain A.K.(1983) Markov random field texture models, IEEE Trans. Pattern Anal. Machine. Intell., PAMI-5, 25-39.

Dempster A.P., Laird N.M., and Rubin D.B.(1977) Maximum likelihood from incomplete data via EM algorithm (with discussion), J.R. Statist. Soc., B 39, 1-38

Derin H. and Elliott H.(1987) Modeling and segmentation of noisy and textured image using Gibbs random fields, IEEE Trans. Pattern Anal. & Machine Intell., PAMI-9,39-55.

Devijver P.A.(1988) Image segmentation using causal Markov random fields, Lecture Notes in Computer Science, 301, 113-143. (Berlin, Springer.)

Frigessi A. and Piccioni M.(1988a) Parameter estimation for the two-dimensional Ising fields corrupted by noise. Quaderno, IAC-CNR, Roma.

————,(1988b) Parameter estimation for 2d Ising fields corrupted by noise: Numerical Experiments. Quaderno, IAC-CNR, Roma.

Geman S. and Geman D.(1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. & Machine Intell., PAMI-6, 721-741

Geman S. and Graffigne C. (1987) Markov random field image models and their application to computer version, in: A.M.Gleason(ed), Proceedings of International Congress of Mathematicians 1986, (Washington D.C., American Mathematical Society).

Geman S. and McClure D.E.(1987) Statistical methods for tomographic image reconstruction. Bull. Int. Statist. Inst., 5, 52-54.

Glendinning R.H. (1989) An evaluation of ICM algorithm for image reconstruction. J. Statist. Comput. Simul., 31, 169-185.

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, 271-279.

Istratescu V.I.(1981) Fixed Point Theory, Math. and Its Appl. 7, (D. Reidel Publ. Company).

Jubb M. and Jennison C.(1988) Aggregation and refinement in binary image restoration, School of Math. Science Tech. Report, Univ. of Bath, U.K..

Kanal L.N. (1980) Markov mesh models. Comput. Graphics Image Process., 12 371-375.

Kanefsky M. and Strintzis M.G.(1978) A decision theory approach to picture smoothing, IEEE Trans. Comput., C-27, 32-36

Kaufmann B. (1949) Crystal statistics II, partition function evaluated by spinor analysis, Phys. Rev. 76, 1232-1243.

Kay J.W.(1988) On the choice of regularisation parameter in image restoration, Springer Lecture Notes in Computer Science, 301, 587-596

Kay J.W. and Titterington D.M.(1986) Image labelling and the statistical analysis of incomplete data. Proc. 2nd Int Image Processing and Applications, IEE Conf. Publ. No. 256, 44-48.

Kirkland M.(1989) Simulation methods for Markov random fields, Ph.D Thesis, Dept. of Mathematics, Univ. of Strathclyde, Glasgow.

Kramers H A. and Wannier G.H. (1941) Statistics of the two-dimensional ferromagnet, Phys. Rev. 60, 252-262.

Kryscio R.J., Saunders R. and Funk G.M.(1980) Normal approximation for binary lattice systems, J. Appl. Prob., 17, 674-685

Lacroix V.(1987) Pixel labelling in a second-order Markov mesh, Signal Processing, 12, 59-82.

Liporace L.A.(1982) Maximum likelihood estimation for multivariate observations of Markov sources, IEEE Trans. on Inform. Theory, 28,729-734.

Moran P.A.P.(1968) An Introduction to Probability Theory, (Oxford Univ. Press, Oxford).

Newell G.F. and Montroll E. W. (1953) On the theory of the Ising model of ferromagnetism, Rev. Mod. Phys. 25, 353-389.

Onsager L.(1944) Crystal statistics I, Phys. Review, 65 117-149.

Owen A.B.(1986) Discussion of Ripley(1986), Can. J. Statist., 14, 106-110.

———,(1989) Image segmentation via iterated conditional expecta-

———,(1989) Image segmentation via iterated conditional expectations, Technical Report No. 254, Dept. of Statistics, Univ. of Chicago.

Pickard D.K.(1976) Asymptotic inference for an Ising lattice, J. Appl. Prob. 13, 486-497.

———,(1977) Asymptotic inference for an Ising lattice II, Adv. Appl. Prob. 9, 476-501.

———,(1982) Inference for general Ising Models, J. Appl. Prob., 19*, 345-357.

———,(1987) Inference for discrete Markov fields: The simplest nontrivial case, J. Amer. Statist. Ass. 82, 90-96.

Pickett E.E. and Whiting R.G. (1987) On the estimation of probabilistic functions of Markov chains, in: Lecture Notes in Econom. and Math. Systems, vol 297 (Springer, Berlin).

Possolo A.(1986) Estimation of binary Markov fields, Tech. Report, No 77, Dept. of Statistics, Univ. of Washington, Seattle.

Potts R.B.(1952) Some generalized order-disorder transformations, Proc. of the Cambridge Philosohpical Soc., 48, 106-109.

Qian W.(1990) A note on asymptotic inference for an asymmetric Ising model around a torus, Adv. Appl. Prob., 22, 755-757.

Qian W. and Titterington D.M.(1989) On the use of Gibbs Markov chain models in the analysis of images based on second-order pairwise interactive distributions, J. Appl. Statist., 16, 267-281.

———,(1990a) Parameter estimation for hidden Gibbs chains, Statist. Prob. Letters, 10, 49-58.

———,(1990b) Pixel labelling for three-dimensional scenes based on Markov mesh models, Signal Processing, to appear.

———,(1990c) Multidimensional Markov chain models for image texture, J.R. Statist. Soc., B, to appear(1991).

———,(1990d) Normal approximation for lattice systems. Submitted to " Stoch. Processes & Their Appl.".

———,(1990e) Stochastic relaxations and EM algorithm for Markov random fields. Submitted to "J. Statist. Comput. Simul.".

Rabiner L.R., Levinson S.E., and Sondhi M.M.(1984) On the use of hidden Markov models for speaker independent recognition of isolated word from a Medium-siz vocabulary. AT&T Bell Lab

Rabiner L.R., Juang B.H., Levinson S.E., and Sondhi M.M.(1985) Some properties of continuous hidden Markov model representations, AT&T Bell Lab. Tech. J. 65, 1251-1270

Ripley B.D.(1986) Statistics, images and pattern recognition, Can. J. Statist., 14, 83-110.

———,(1990) Thoughts on pseudorandom number generators, J. Comput. Appl. Math., 31, 153-163.

Ripley B.D. and Kirkland M.D.(1990) Iterative simulation methods, J. Comput. Appl. Math., 31, 165-172.

Saunders R., Kryscio R.J. and Funk G.M.(1979) Limiting results for arrays of binary random variables on rectangular lattices under sparseness conditions, J. Appl. Prob., 16, 554-566

Silverman B.W, Jennison C., Atander J.and Brown T.C.(1989) The specification of edge penalties for regular and irregelar pixel images, submitted for publication.

Strauss D.J.(1977) Analyzing binary lattice data with nearest neighbour property, J. Appl. Prob., 14, 135-143.

Thompson A.M., Brown J.C., Kay J.W. and Titterington D.M.(1990) A study of methods of choosing the smoothing parameter in image restoration by regularization, IEEE Trans. Pattern Anal. & Machine Intell., to appear.

Titterington(1989) Some recent research in the analysis of mixture distributions, Statistics, to appear.

Titterington D.M., Smith A.F.M. and Makov U.E.(1985) Statistical analysis of finite mixture distributions, (Wiley, Chichester UK).

Tjoitheim D.(1978) Statistical spatial series modelling. Adv. Appl. Prob., 10, 130-154

————,(1983) Statistical spatial series modelling II: some further results and unilateral lattice processes. Adv. Appl. Prob., 15, 562-584.

Varga R.S.(1962) Matrix Iterative Analysis, (Prentice-Hall, Englewood Cliffs, NJ).

Wu C.F.J.(1983) On the convergence properties of the EM algorithm, Ann. Statist., 11, 95-103.

Yokoyama R. and Haralick R.M.(1979) Texture pattern image generation by regular Markov chain. Pattern recognition, 11, 225-234.

Younes L.(1988) Estimation and annealing for Gibbsian fields. Ann. de. l'Inst. Henr. Poincare B, 24, 269-294.

————, (1989) Parametric inference for unperfectly observed Gibbsian fields. Prob. Theory Relat. Fields, 82, 625-645.