PARAMETER ESTIMATION IN IMAGE PROCESSING

by

Achilles Venetoulia
Stanford University

TECHNICAL REPORT NO. 19
JUNE 1989

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT DMS 84-11300

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
PARAMETER ESTIMATION IN IMAGE PROCESSING

by
Achilles Venetoulia
Stanford University

TECHNICAL REPORT NO. 19
JUNE 1989

PREPARED UNDER THE AUSPICIES
OF
NATIONAL SCIENCE FOUNDATION
GRANT DMS 84-11300

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Abstract

This report concerns the problem of parameter estimation in the context of image processing applications. The report can be conceptually divided into two parts. The first part deals with the estimation of the parameters of a binary Markov model. A method is developed for obtaining moment estimates for the parameters, when the model is used to describe the spatial structure of a black and white image. Unlike previously suggested estimation procedures, this method is trivial to implement and is driven entirely by the data (that is, does not rely on some preliminary pixel classification).

The second part develops a method for estimating the variance of an estimate of an image parameter. This method creates replicate parameter estimates within the image and uses them to estimate the variance. The replicates are obtained by dividing the pixel lattice into strips of rows (or columns). The variance estimate is obtained from analyzing the estimates of the strips as a time series. This method can be extended into a valuable tool for the analysis of image data.
Chapter 1

Introduction

Markov random field models have recently become a popular tool in image processing. This report develops a method for estimating the parameters of a particular class of Markov random fields. Specifically, estimates are obtained for the parameters of a binary Markov random field in the context of image processing applications. This report also suggests a method for estimating the variance of the parameter estimates. The method is quite general and can be used to estimate the variance of the estimates of any image parameters.

Image processing is a broad, interdisciplinary area of research which concerns the analysis of images. In the past fifteen years scientists from many fields have been active in developing methodologies for analyzing images. The reason is that images have become a convenient and flexible form of data collection which is suitable for a number of disciplines. Doctors rely on tomographic images to specify the exact location of tumors. Physicists and astronomers use two-dimensional arrays of detectors to capture signals from the universe. Geologists interpret satellite images to identify mineral or other deposits of interest. Computer scientists build systems which recognize three-dimensional objects from two-dimensional images in order to endow robots with sophisticated vision. This wide range of applications has resulted in an overwhelming proliferation of literature in image processing [Vene88].

The term “image processing” encompasses today an array of problems which represent different aspects of the analysis of images. Some of the common examples are image enhancement, restoration of images, classification of imagery data, image compression, etc. Most of these problems are related to each other because they constitute manifestations of a basic statistical problem. This problem is how to deduce the unobservable attribute(s) of a spatial domain, given some observations (on the domain) which are somehow related to the actual attribute(s). This vague definition pertains to the majority of issues in the area of spatial statistics and
may, therefore, appear un informatively general. In the context of image processing, the spatial
domain is the domain of definition of the image, the attribute is some property of interest and
the observations are the measurements which comprise the data on the image. This definition
becomes more concrete in the next chapter.

Markov random fields are models which extend the concept of one-dimensional Markov
processes (e.g., Markov chains) to the two-dimensional plane. Such models have been applied
to problems in many different fields: physics, mathematical biology, spatial statistics, etc.
In image processing Markov random fields are used as models of the spatial structure of an
image. Their popularity stems from some attractive mathematical properties (e.g., consistent
spatial distributions on the pixel lattice [Besa74]) and some successful results which have
been obtained from the application of such models [GeGe84, Besa86a]. In most image
processing applications Markov random fields are parametric models and their use requires
some knowledge of the parameters involved. Unfortunately, parameter estimation is often
a difficult problem. The main object of this report is to describe a method for obtaining
moment estimates for the parameters of a binary Markov random field, when this field is used
to describe the spatial structure of black and white images (see section 2.1).

This report addresses the problem of parameter estimation in the classification of imagery
data. The distinction about the context of application is important. The goal is not to
undertake a theoretical investigation of the properties of a Markov random field in order to
obtain descriptions of its parameters without any regard to a specific application. Instead,
the goal is to obtain adequate estimates of the parameters of a Markov random field that will
make it a useful model for describing an image. This is a practical problem and the proposed
solution is to construct simple moment estimates. These estimates are easy to compute and
analyze and they seem to perform satisfactorily in practice.

Parameter estimation is a problem of significant practical value because it constitutes an
essential component of the process of applying an algorithm to an image. It is also a problem
of interest in its own right because any attempt to estimate the model parameters provides
insight into the data. For example, parameter estimation could aid the interpretation of the
model parameters in terms of tangible image properties and thus highlight some previously
unnoticed properties. Or, it could yield some indication as to how appropriate the used model is and, possibly, suggest an alternative.

This report is organized as follows. The next chapter explores the relationship between statistics and image processing. It defines the general problem, discusses some of the issues that arise and identifies the role of parameter estimation. Chapter 3 describes the application of Markov random fields to image processing and introduces the model for which a parameter estimation technique is developed in chapter 5. Chapter 4 presents a slight modification to a commonly used reconstruction algorithm, Besag's ICM [Bes86a]. The modification amounts to adapting this algorithm to the Markov model which is studied in chapter 5. Chapter 5 develops the moment estimator for the parameters of a binary Markov random field from univariate observations. It presents the properties of the estimator and suggests a possible implementation. Chapter 6 presents a method for estimating the variance of the moment estimates. This method exploits the analogy between the spatial index of a two-dimensional Markov process (e.g., a Markov field) and the time index of an one-dimensional Markov process (e.g., a time series). It is motivated by the estimator of chapter 5, but it is sufficiently general to be applicable to any other estimator as well. In fact, this idea provides a simple method for estimating the variance of the estimates of any quantities of interest in an image. Chapter 7 discusses two extensions of the moment estimator (images with multivariate observations and local applications of the estimator) and one extension of the variance estimation technique. Finally, chapter 8 presents some empirical results obtained from applying the modified ICM algorithm and the proposed estimators (i.e., the moment estimator and the variance estimator) to a number of images.
Chapter 2

Statistical Image Processing

This chapter outlines some of the statistical issues in image processing. The next two sections introduce the general and the specific problem respectively with which this report is concerned. Section 2.1 provides an informal model of the statistical problem underlying most of image processing. An overview and some examples are discussed from the point of view of applications. This overview sets the framework for the more specific problem of parameter estimation in image processing. Section 2.2 briefly introduces different aspects of this problem and puts in perspective the special case of parameter estimation in Markov random fields.

2.1. The General Problem.

2.1.1. Domains

An image usually describes a bounded region in Euclidean space of one, two or three dimensions. The bounded region, $D$, is the spatial domain of definition of the image. Each point in the domain is denoted by a spatial index, $t$. The domain $D$ possesses some attribute of interest which is a function $I(t)$. This attribute is appropriately defined on $D$ and cannot be directly observed. Instead, there is another, observable function, $O(t)$, which is also defined on $D$. $O(t)$ is related to $I(t)$ and is observed on a subdomain $D' \subseteq D$. The goal is to infer the image \{I(t) : \forall t \in D\} from the data \{O(t) : \forall t \in D'\}.

This model conceptually defines a number of image processing problems. However, its constituents are yet to be specified. In order to make the definition of each problem concrete, one must specify the spatial domains $D$, $D'$, the functions $I(t)$, $O(t)$, and the relationship between them. Many possibilities exist and different combinations of \{D, D', I(t), O(t)\} give rise to different applications.
The domain $D$ is most frequently a subset of the two-dimensional plane. In this case, the spatial index $t$ assumes the role of the two coordinates for each point in $D$. Typical examples of two-dimensional domains are: a field on the surface of the earth, the object of a photograph, a region in the sky, etc. Sometimes the spatial domain may be a subset of one-dimensional Euclidean space (e.g., a trajectory in a certain direction). One-dimensional domains are often used for the development of theoretical methods, which are then extended to real images in two or more dimensions. The reason is that there is a greater abundance of well-understood models in one dimension than in two. Domains in three or more dimensions are also possible, especially if time is incorporated as an additional coordinate into the domain of definition (e.g., in multitemporal images where the same image is observed at various points in time). For the sake of simplicity, the remainder of this section deals exclusively with two-dimensional spatial domains. These are the ones most frequently encountered in applications and they suffice for the purposes of this discussion. Furthermore, this report concerns only two-dimensional images.

It is often impossible, or highly impractical, for a recording device to sample every point in the domain of definition of an image. This fact leads to a distinction between the domain of definition, $D$, and the corresponding domain of observation, $D'$. $D$ is always an essential component of the image (as part of its definition), whereas $D'$ may exist purely for observational purposes. This difference becomes obvious for spatially continuous images where the observations can occur only at a few points in $D$. In this case, it is typical to divide $D$ into a grid of small areas ("pixels") and sample one point per pixel (e.g., its center). The pixels are usually rectangular but other shapes (triangular, hexagonal, circular, etc.) are sometimes useful as well. The point is that the size and the shape of the pixels do not have to represent any characteristics of the image itself and may be determined by the recording device. A common example is satellite images of agricultural scenes. Here, the size of the pixels depends only on the satellite (e.g., on its resolution capability) and different pixel grids are possible regardless of the actual scene. In other words, the image $\{I(t) : \forall t \in D\}$ exists independently from the choice of the sampling set $D'$.

On the other hand, the choice of the pixel grid depends not only on the actual image $I(t)$ but on the particular application as well. One reason is that the relative size of the pixels
depends on the nature of the application and on the size and the spatial characteristics of the image. However, size (of pixels and images alike) is a concept which is difficult to define and will not be discussed any further.

The distinction between $D$ and $D'$ raises another relevant issue. If the observations $O(t)$ are available only at points in $D'$, the question is how much information can be extracted from $\{O(t) : \forall t \in D'\}$ about the image $I(t)$ at the points in $D - D'$. That is, to what extent can information about $\{I(t) : \forall t \in D - D'\}$ be extrapolated from $\{O(t) : \forall t \in D'\}$ (this is the same problem as in kriging). A simple way of circumventing this difficulty is to identify each pixel with its center. Then, the spatial index $t$ denotes both the center of the pixel and the pixel itself, and $D$ reduces to $D'$. This approach is in agreement with the cases where $D$ naturally coincides with $D'$. For granular images in which the objects form a discrete lattice (e.g., for a DNA molecule) the domain of definition, $D$, and the domain of observation, $D'$, are identical. In this case, $D'$ is intrinsically associated with the image itself and is not an extraneous artifact of the observation process. The problem now becomes to infer the image $\{I(t) : \forall t \in D'\}$ from the data $\{O(t) : \forall t \in D'\}$. This is the model assumed in this report: for a spatially continuous image the observations are collected on a pixel grid, and the image values are being sought at the same points where the observations are available. It is, therefore, implied through the rest of this section, as well as through the rest of the report, that $D' \equiv D$.

### 2.1.2. Images and Data

The function $I(t)$ defines the image by describing its property of interest. The reasons for which the property $I(t)$ is not directly observable can be quite diverse: the observations are contaminated with noise, the image pertains to a geographically inaccessible area, the cost of direct observation is prohibitive, etc. Common examples for $I(t)$ are categorical and positive, continuous functions. The categorical functions correspond to discrete properties of the image: the existence or lack of a certain feature, the type of rock, the color of a surface, the type of agricultural crop, etc. The positive, continuous functions correspond to continuous, or almost continuous, properties of the image: the grey level of a surface, the intensity of a light source, the concentration of a compound, the textural characteristic of some material, etc.

Similarly, the nature of the observations $O(t)$ can also vary. Categorical observations
represent possibly erroneous recordings of a discrete measurement or discretized versions of a continuous measurement. Continuous observations represent, as before, intensities, concentrations, grey level measurements, etc. For continuous observations, and especially for concentrations and intensities, the measurement \( O(t) \) at pixel \( t \) is frequently the integral of the observation function over some area around the point \( t \). It should be mentioned that the area of integration is not necessarily the same as the pixel \( t \).

The image and the observations can be either univariate (with one measurement for every point) or multivariate (with a number of observations for every point). Multivariate images, or observations, may comprise a combination of the types mentioned. The relationship between \( I(t) \) and \( O(t) \) follows from their respective definitions. Some examples are:

**Example 1.** The image \( I(t) \) is categorical and assumes only two values. These values are for convenience taken to be 0 and 1 and they represent respectively the absence and the presence of a certain characteristic. This characteristic may be a rock type, a color, a type of crop, a boundary, etc. When the characteristic is the presence or the absence of a boundary, the boundary detection problem arises. The observations are multivariate intensities recorded at different bandwidths (e.g., infrared and ultraviolet). This problem is the main focus of this work: the classification problem for binary images (black/white images) with multispectral observations (see sections 4.1, 5.1, 7.2).

**Example 2.** The same as in Example 1, except that the image \( I(t) \) assumes more than two values. This is the classification (or the image segmentation) problem for multi-class images. When the image is some characteristic of a three-dimensional object (e.g., \( I(t) \) takes values in a set \{corner, edge, face, ...\}) and the observations are, say, curvature measurements, the problem becomes the reconstruction of three-dimensional objects from two-dimensional images.

**Example 3.** The image \( I(t) \) are the true grey levels of a picture (regarded as a continuous function on the integers from 0 to 255). The observations are the same grey levels which have been degraded by noise: \( O(t) = I(t) + e(t) \), where \( e(t) \) is the noise process. This is the image enhancement or the image restoration problem.

**Example 4.** The image \( I(t) \) are the true intensities of a field, and the observations are the same intensities which have been blurred and distorted by noise: \( O(t) = H * I(t) + e(t) \), where
\( H \) is a blurring operator. This is again the problem of image restoration. A typical application in which blurring occurs naturally are astronomical images.

2.1.3. Estimation of Images

In trying to estimate \( I(t) \) from \( O(t) \) there are two major obstacles to overcome. The first one, and usually simpler of the two, is to filter out the noise which is present in the observations. In the simplest case \( O(t) = I(t) + e(t) \), where \( e(t) \) is the noise process (as in Example 3). The noise is almost always spatially correlated, because the errors in one part of the image are often due to a common aberration. For instance, when a cloud interferes with a satellite recording, the cloud affects the measurements of all the pixels in some area of the image. In practice, however, for reasons of simplicity the noise is usually considered independent from one pixel to another. What is needed is a method for separating the noise \( e(t) \) from the signal \( I(t) \). Different assumptions about the noise lead to different reconstruction methods but in principle a smoothing algorithm could suffice. The smoothing algorithm can assume the form of Kalman filtering [ZhCV86], Bayesian smoothing [DeEC84], local spatial smoothing [Swit80], etc. Nonsmoothing techniques for extracting the signal from the observations have also been suggested; two examples are principal components and min/max autocorrelation factors ("MAF") [SwGr84].

The second obstacle is to model the spatial behavior of \( I(t) \) on \( D \). This is equivalent to describing those spatial properties of the image which are relevant for the particular application. Describing the spatial structure of the true image \( I(t) \) turns out to be a complicated task. Most useful images are not random juxtapositions of arbitrary pixels: they obey some connectivity principle which dictates that a number of connected regions (objects) form the overall pattern of the image. This is a concept which is easy to understand but difficult to quantify. For example, in a black and white image the fact that a pixel \( t \) is white may increase the chances that its neighboring pixel \( t' \) is also white. To quantify this statement one must devise a formula which accurately measures this increase in probability and which is consistent with the whole image.

The values of \( I(t) \) may be treated as independent over the domain \( D \), but more often than not this is an unrealistic assumption which yields unsatisfactory results. If the pixels are assumed to be independent, the classification problem for binary images reduces to the problem
Section 2.1: The General Problem

of classifying a bunch of independent observations into one of two populations. As such, it is amenable to maximum likelihood estimation and/or standard discriminant analysis techniques (linear or quadratic) which disregard the relative positions of the pixels. This approach used to be common practice, but it has been shown to be inferior to methods which exploit the spatial structure of the image. Numerous schemes, of varying degree of rigor, attempt to utilize the spatial context of an image. Their range includes (a) augmenting the observation at each pixel with the observations at its neighboring pixels and using the augmented observations as data, (b) locally smoothing the data before carrying out the estimation, (c) devising geometric models for the patterns of discrete images, (d) imposing formal probabilistic specifications on the spatial distribution of the image, etc. The latter two of these examples ((c) and (d)) are instances of a large class of Bayesian paradigms for image estimation.

The Bayesian paradigm naturally incorporates a model for the spatial distribution of $I(t)$ into the estimation of $I(t)$ from $O(t)$. A model, $P\{I(t)\}$, is constructed for the spatial distribution of $I(t)$ and a conditional distribution, $P\{O(t)|I(t)\}$, is specified for the observations $O(t)$ given the actual image values $I(t)$. These two components together determine the "posterior" distribution of the image given the data: $P\{I(t)|O(t)\} \propto P\{O(t)|I(t)\} \cdot P\{I(t)\}$. The image is estimated by minimizing the expectation of some loss function with respect to the posterior distribution $P\{I(t)|O(t)\}$.

In Bayesian terminology, the image model $P\{I(t)\}$ plays the role of the prior distribution, the conditional distribution $P\{O(t)|I(t)\}$ plays the role of the likelihood of the data and the distribution of the image given the data $P\{I(t)|O(t)\}$ plays the role of the posterior likelihood. However, nothing changes if non-Bayesian terms like conditional distribution and marginal distribution are adopted for $P\{O(t)|I(t)\}$ and $P\{I(t)\}$ respectively. It is important to note that the specifications for the spatial distribution $P\{I(t)\}$ and the conditional distribution $P\{O(t)|I(t)\}$ are ad hoc models which may not bear much relevance to the true image. In other words, the Bayesian paradigm provides just a convenient formulation, not a definitive framework, for the problem.

The appeal of this idea rests to a great extent with its flexibility. The model for $I(t)$, the conditional distribution $P\{O(t)|I(t)\}$ and the loss function introduce ways in which this
approach can be tailored to a particular application. The spatial distribution $P\{I(t)\}$ and the
conditional distribution $P\{O(t)|I(t)\}$ can be selected so that they reflect prior information
about the image and the observations respectively. An empirical Bayesian approach is often
used because it allows even greater freedom: the distribution $P\{I(t)\}$ is parametrized and its
parameters, which hopefully have a meaningful interpretation in terms of image properties,
are obtained from the data. Examples of image models are spatially stationary stochastic
processes, geometric models for discrete images, autoregressive schemes, Markov random fields
(see section 3.1), etc. The object of this report is to develop a method for estimating the
parameters of the distribution $P\{I(t)\}$, when the model for the true image is a binary Markov
random field (see sections 4.1, 4.3 and chapter 5).

The conditional distribution $P\{O(t)|I(t)\}$ is usually assumed to be normal, since this
appears to be a harmless and in many cases an almost realistic assumption. A typical example is
the classification problem for binary images where $I(t) \in \{0, 1\}$ and the observations $O(t)$ are
multivariate. Then, $P\{O(t)|I(t) = 0\} = N(\mu_0, \Sigma_0)$ and $P\{O(t)|I(t) = 1\} = N(\mu_1, \Sigma_1)$. Similarly, in the multi-class classification problem, where $I(t) \in \{I_1, I_2, \ldots, I_k\}$, the (conditional) class distributions are assumed to be normal: $P\{O(t)|I(t) = I_l\} = N(\mu_l, \Sigma_l)$, for $1 \leq l \leq k$; it is frequently convenient to assume further that the covariance matrices $\Sigma_l$ are all equal:
$\Sigma_l = \Sigma$, for all $l$. Another common assumption regarding the distribution $P\{O(t)|I(t)\}$ is the
conditional independence of the observations $O(t)$ and $O(t')$ given the actual image values at $t$ and $t'$ respectively: $P\{O(t), O(t')|I(t), I(t')\} = P\{O(t)|I(t)\} \cdot P\{O(t')|I(t')\}$. This condition of independence expresses the concept that the stochastic nature of any observation $O(t)$
depends on the image only through its value, $I(t)$, at the point $t$.

The loss function can also be chosen in accord with the application at hand. More generally,
many image estimation procedures aim at the optimization of some objective function;
the minimization of a loss function in the Bayesian procedures is a special case. This objective
function constitutes the performance criterion by which one judges the quality of the estimated
image. The evaluation of the performance of an image estimation technique is an issue which
is currently not well-understood. Two elusive aspects of this problem are (a) how to select the
most appropriate performance criterion and (b) how to construct performance criteria which
capture some desirable characteristic. This characteristic may be either some feature of the
actual image $I(t)$ or some desirable property of the estimated image $\hat{I}(t)$.

The most obvious criterion for estimation is visual fidelity. However, this criterion is subjective and, in most cases, infeasible due to the unavailability of the original image. Any measure of fidelity between the estimated image $\hat{I}(t)$ and the true image $I(t)$ is a potential objective function. Some common examples are:

Example 5. $\hat{I}(t)$ is obtained by maximizing the entropy of the image: $\hat{I}(t)$ maximizes $-\sum t q_t \log q_t$, where $q_t = I(t)/\sum t I(t)$, subject to some constraint $C(I(t), O(t))$. The constraint accounts for the existence of the data $O(t)$ by relating it to the image $I(t)$ [BuGS83].

Example 6. $\hat{I}(t)$ is obtained by maximizing the posterior probability at each pixel $t_0$: $\hat{I}(t_0)$ maximizes $P\{I(t_0)|O(t), t \in D\}$. This criterion corresponds to a loss function which is the sum of the misclassification probabilities for the individual pixels [Hjor85a].

Example 7. $\hat{I}(t)$ is obtained by maximizing the overall posterior probability of the scene: the estimated image $\{\hat{I}(t), t \in D\}$ maximizes $P\{I(t), t \in D|O(t), t \in D\}$. This is the maximum a posteriori ("MAP") estimate. The MAP estimate corresponds to a loss function which is 0 if there are no misclassification errors at all and 1 otherwise [GeGe84].

Example 8. $\hat{I}(t)$ is obtained by maximizing the posterior marginal probability of a particular class for the whole image. That is, if $I(t) \in \{I_1, I_2, \ldots, I_k\}$ then $\{\hat{I}(t), t \in D\}$ maximizes the probability $P\{I(t) = I_i, t \in D|O(t), t \in D\}$ for some $i$ in $\{1, 2, \ldots, k\}$. This objective function is pertinent when it is important to estimate accurately the frequency of occurrence of a certain class.

Example 9. $\hat{I}(t)$ is obtained by minimizing the expectation of a quadratic loss functional with respect to the posterior distribution of the image: $\hat{I}(t)$ minimizes $E\left(\sum t (I(t) - \hat{I}(t))^2\right)$, where the expectation is with respect to $P\{I(t), t \in D|O(t), t \in D\}$ [MaMP87].

Estimated images which are derived from some optimality criterion do not necessarily possess either the visual qualities of the original image or its statistical properties. In fact, it is not rare that visual inspection is more revealing about the poor performance of a reconstruction algorithm than the performance criterion itself. One possible remedy is to post-process the reconstructed image (by a smoothing filter) in order to rectify some of its visual characteristics.
Chapter 2: Statistical Image Processing

[Swit83, JiCh86]. Nevertheless, since the occurring discrepancies can be quite severe, it is sometimes preferable to use estimation procedures which do not satisfy any optimality criteria but exhibit a more agreeable behavior.

The presentation in the section has by necessity ignored a number of issues. Among others, the details of the spatial models (e.g., their probabilistic structure, the estimation of their parameters, their relevance to a specific image, etc.), the question of the size of an image and the evaluation of the object functions (e.g., how to estimate the probabilities of misclassification errors when the original image is unknown) and the properties of the estimated images (e.g., Examples 5–9) have not been addressed. More comprehensive discussions of these issues exist in many review articles. Two relatively recent papers which contain a more detailed and systematic overview of the whole problem are Ripl86 and Swit87. An early mathematical treatment of the classification problem appears in Swit69: the presentation there is sufficiently general and, at the same time, detailed enough to provide a good reference. A different source of similar information is the detailed annotated bibliography of Vene88.

2.2. Parameter Estimation.

Parameter estimation is a problem which warrants attention for reasons that have already been mentioned. First, the processing of an image is often based on parametric models, the application of which relies on the values of the parameters involved. Second, the process of parameter estimation may reveal useful information about an image. This information can be used to build and/or select a more suitable model for the particular image. Indeed, a model is sometimes chosen from a family of candidate models which are indexed by a set of parameters. In these cases choosing a model amounts to specifying the desirable values of the indexing parameters [KaCh83, ChKa85], and the model is actually identified by the parameter estimates.

Parameter estimation appears under many guises in image processing. The reason is that the parameters of interest are determined in each case by the specific application and the algorithm(s) to be used. The wide range of of applications, the abundance of spatial models for describing images and the arbitrarily large selection of image processing algorithms explain the multitude and the diverse nature of the existing methods for parameter estimation. For
example, in image restoration (Examples 1–4) different reconstruction methods may necessitate different parameter estimation techniques even for the same image and the same model (i.e., the exact same parameters) [Besa86b]. Or, different applications may require different degrees of accuracy in the obtained estimates because their objectives are less, or more, sensitive to the actual values of the parameters. Consequently, a particular parameter estimation technique may be sufficient for some applications but inadequate for others.

Most, if not all, of the models which are used for image processing are essentially ad hoc in nature. Therefore, it is difficult to systematically categorize, or address, all the aspects of parameter estimation. In general, there are two kinds of parameters: those related to the observations $O(t)$ and those related to the image $I(t)$. This distinction is somewhat artificial (see Example 12 below) but at the same time quite useful, because it facilitates the exposition of this section.

It is usually a simpler task to estimate the parameters related to the observations than those related to the image. The salient reason for this phenomenon is the very nature of the processes $I(t)$ and $O(t)$: $I(t)$ is not observable whereas $O(t)$ is observed directly. Furthermore, the models commonly used for $I(t)$ tend to be more arbitrary and, naturally, less verifiable than the models used for $O(t)$. As a consequence of this latter fact, one may conjecture that the parameters of the image models play a less significant role than the parameters of the observation models. However, the sensitivity of an image restoration algorithm (e.g., Examples 5–9) to the values of its parameters depends on the specific application and the nature of the algorithm.

2.2.1. Parameters

The parameters related to the observations $O(t)$ can be associated with: (a) the conditional distribution $P \{ O(t) | I(t) \}$, (b) some algorithmic procedure which operates on $O(t)$ and (c) some direct model for $O(t)$. The next three examples correspond to these cases.

**Example 10.** In the multi-class classification problem (see Example 2), the image $I(t)$ is categorical and the observations $O(t)$ are multivariate: for all $t \in D$, $I(t) \in \{I_1, I_2, \ldots, I_k\}$ and $O(t) \in \mathbb{R}^p$. Each class distribution is assumed to be normal with its own mean vector and
its own covariance matrix: \( P \{ O(t) | I(t) = I_l \} = N(\mu_l, \Sigma_l) \), for \( 1 \leq l \leq k \). The parameters \( \mu_l \) and \( \Sigma_l \) are related to the observations and are associated with the distribution \( P \{ O(t) | I(t) \} \).

This model for the observations sets the framework for the main problem of this report (see sections 5.1, 5.2 and 7.2).

**Example 11.** In the image restoration problem (see Example 3), the image \( I(t) \) assumes values on a continuous grey scale. The observations \( O(t) \) are the image values corrupted by additive noise: \( O(t) = I(t) + e(t) \), where \( e(t) \) is random noise. A common restoration method is local spatial smoothing [Swit83]. Local spatial smoothing smoothes the observation \( O(t) \) at pixel \( t \) by some statistic (e.g., mean or median) of the observations in a neighborhood \( N_t \) of \( t \). That is, the smoothed value of \( O(t) \) is \( \phi(\{O(t') : \forall t' \in N_t\}) \), where \( \phi \) is the smoothing statistic. Consider the case where \( \phi \) is a weighted average of all the observations \( O(t') \) in \( N_t \):

\[
\phi(t) = \sum_{t' \in N_t} w_{t'} O(t'),
\]

with the weights \( w_{t'} \) subject to a desirable specification [DeMo78, Dehn85, HaHo85a]. The weights can be thought of as observation parameters associated with the smoothing procedure.

**Example 12.** In the image restoration problem, the observations \( O(t) \) are assumed to follow a simultaneous autoregressive ("SAR") model: \( O(t) = \sum_{t' \in N_t} \theta_{tt'} O(t') + \sqrt{\rho} e(t) \), where \( N_t \) is a neighborhood of pixels around \( t \) and \( e(t) \) is a noise process with mean 0 and variance equal to 1 [ChKa82]. This kind of model has been extensively used [Kash81, ChCh85a, Chel85] as a spatial model for image description. The weights \( \theta_{tt'} \) and the error variance, \( \rho \), are the parameters of the model for \( O(t) \).

The last two examples illustrate two points. First, the apparent similarity in the equations of Examples 11 and 12 is subject to different interpretations. In Example 11, \( \phi(t) \) is viewed as a smoother without any underlying model for either \( I(t) \) or \( O(t) \). In Example 12, the implication is more restrictive: the observations constitute a realization of the SAR model. Second, since the relation between \( O(t) \) and \( I(t) \) is assumed to be known (for the problem to make sense), the observation parameters which are associated with a direct model for \( O(t) \) are also associated with the image \( I(t) \). For instance, it may be convenient to translate the observation parameters of a model similar to the one in Example 12 to image parameters of an analogous model. This is the reason for which the distinction between observation and image
parameters is to a certain extent artificial.

The parameters related to the image $I(t)$ depend on the model which is used to describe the spatial structure of $I(t)$. As a general rule it is desirable to use parameters which can be interpreted in terms of tangible image properties, but such a parametrization may not always be possible. Clearly, there are numerous possibilities for meaningful spatial parameters:

**Example 13.** For a polychotomous image (see Example 2), where $I(t) \in \{I_1, I_2, \ldots, I_k\}$, the frequency of occurrence of a particular class is often important. This parameter is the marginal class probability $P\{I(t) = I_i\}$, for any class $I_i$; in Bayesian terms, this is the prior probability of class $I_i$. Another commonly used parameter is the probability that a pixel $t$ belongs to a certain class $I_i$ given that a neighboring pixel $t'$ belongs to another class $I_m$: $P\{I(t) = I_i | I(t') = I_m\}$. This is the class transition probability between classes $I_i$ and $I_m$. The pixels $t$ and $t'$ are almost always adjacent to each other, but the transition probability can be defined for any pixels which are a fixed distance apart (i.e., for any $t$ and $t'$ for which the distance between them remains constant). The marginal class probability and the transition probability play an important role in the estimation problem considered in chapter 5 (see sections 5.2 – 5.4).

**Example 14.** For a black/white image (see Example 1), one measure of homogeneity of the pixels is the percent of heterogeneous pixels (i.e., those pixels which belong partially to one color and partially to the other). This parameter provides a crude lower bound for any classification algorithm which assumes that each pixel is homogeneous and allocates it as a whole to one of the two colors (for an alternative see KeMa87). The reason is that such an algorithm will, by definition, misclassify the heterogeneous pixels [Swit83, SwVe85].

**Example 15.** For a continuous image which is assumed to follow a conditional autoregressive ("CAR") model [Besa74, MoRi87, KiCa85] the parameters of the model are of interest. Specifically, let $E( I(t) | I(t'), \forall t' \in D - \{t\} ) = \sum_{r \in N_t} \alpha_{ir} I(r)$ and $\text{Var}( I(t) | I(t'), \forall t' \in D - \{t\} ) = \alpha_{00}$, where $N_t$ is a neighborhood of pixels around pixel $t$ and the $\alpha_{ij}$'s are appropriately defined constants (scalars when $I(t)$ is a univariate process and matrices when $I(t)$ is multivariate). The $\alpha$'s are the parameters which identify one such model from a whole class of similar models.
Example 16. For a binary image \( I(t) \in \{0, 1\} \) which is assumed to obey a Markov random field [Besa74, GeGe84, Besa86a] the parameter of spatial interaction is of interest. The Markov random field model is specified by the conditional distribution of the image value \( I(t) \) at any pixel \( t \), given the values \( I(t') \) at all the other pixels \( t' \):

\[
P \{ I(t) = 1 | I(t'), t' \in D - \{t\} \} = P \{ I(t) = 1 | I(t'), t' \in N_t \} = \frac{e^{\beta S_t(1)}}{e^{\beta S_t(0)} + e^{\beta S_t(1)}},
\]

where \( N_t \) is some neighborhood around \( t \), \( S_t(0) = \{ t' \in N_t : I(t') = 0 \} \) is the number of neighbors of \( t \) which are equal to 0 and \( S_t(1) = \{ t' \in N_t : I(t') = 1 \} \) is the number of neighbors of \( t \) which are equal to 1. Although it may not be apparent at first, this model is essentially the same as the one in Example 15 [Besa74]. The parameter \( \beta \) is a measure of the spatial interaction between neighboring pixels: the dependence of the value \( I(t) \) on the values \( \{ I(t'), t' \in N_t \} \) increases with \( \beta \). The estimation of \( \beta \) is the main focus of this report. The Markov random field model is presented in detail in the next chapter (see sections 3.1–3.3), and a method for obtaining moment estimates for \( \beta \) is given in chapter 5 (section 5.3).

Example 17. The success of many reconstruction algorithms depends on the degree of spatial variability of the actual image. That is, it depends on how slowly the image values change over the spatial domain. For a categorical image one way of quantifying this concept is to measure the number of existing boundaries among the regions which belong to the different classes. A Poisson field [Swit65, Owen84] has been used as a model for the process that generates the boundaries: the intensity of the field is the parameter which determines how many boundaries are generated. This is an example of a parameter which is related to the image \( I(t) \) in a less direct fashion than the parameters discussed in the previous examples.

2.2.2. Methods

The most common means for parameter estimation are training samples (or, training data). The idea is to somehow obtain complete knowledge of the image on a subregion \( \tilde{D} \) of the domain of definition \( D \). In other words, the image values \( \{ I(t) : \forall t \in \tilde{D} \} \) become part of the data along with all the observations \( \{ O(t) : \forall t \in \tilde{D} \} \). Then, the parameters of interest are estimated from the training data \( \{ O(t) \text{ and } I(t), \text{ for } t \in \tilde{D} \} \).

The use of training data is a very natural concept in the processing of images with remotely
sensed data. For instance, a satellite may record the reflected solar energy measurements (for selected wavelength bands) of an agricultural area; this data is to be used to estimate the crop composition of the area [EILF85]. In this case, training data can be obtained by inspection of a small region (e.g., one corner, the center, a few random sections) of the observed area. Another example arises in the identification of certain textures (e.g., cork, paper, sand, etc.) from their images [ChKa85, GeGe84]. Here, training data can consist of prototype images of each type of texture. As indicated in the examples that follow training data is a convenient and rather flexible tool.

Example 19. (Example 10, continued) The parameters of interest are the means $\mu_i$ and the covariance matrices $\Sigma_i$. Let $S_i$ be the training sample for class $I_i$ namely, $S_i = \{ t \in \hat{D} : I(t) = I_i \}$. Both $\mu_i$ and $\Sigma_i$ can be estimated by their sample estimates:

$$\hat{\mu}_i = \frac{1}{|S_i|} \sum_{t \in S_i} O(t)$$

and

$$\hat{\Sigma}_i = \frac{1}{|S_i| - 1} \sum_{t \in S_i} (O(t) - \hat{\mu}_i)(O(t) - \hat{\mu}_i)' .$$

These estimates are independent of any distributional assumptions regarding the class distributions $P \{ O(t) | I(t) = I_i \}$. Alternatively, if $P \{ O(t) | I(t) = I_i \}$ is assumed to be of a particular distributional form (say, normal as in Example 10), maximum likelihood or moment estimates can also be derived for the parameters of interest from the class training sample $S_i$. The data $\{ O(t) : \forall t \in S_i \}$ constitutes a sample from the class distribution $P \{ O(t) | I(t) = I_i \}$ and, therefore, any standard estimation technique can be applied to it.

Example 20. (Example 13, continued) The parameters of interest are $P \{ I(t) = I_i \}$ and $P \{ I(t) = I_i, I(t') = I_m \}$. For simplicity let $t$ and $t'$ be adjacent pixels in $\hat{D}$. An obvious estimate for $\pi_i = P \{ I(t) = I_i \}$ is $\hat{\pi}_i = |S_i| / |\hat{D}|$. Similarly, if $S_{im} = \{(t, t') : I(t) = I_i, I(t') = I_m \}$ an estimate for $\pi_{im} = P \{ I(t) = I_i, I(t') = I_m \}$ is $\hat{\pi}_{im} = |S_{im}| / |\{(t, t')\}|$ and an estimate for $P \{ I(t) = I_i, I(t') = I_m \}$ is $\hat{\pi}_{im} / \hat{\pi}_m$. More generally, an estimate for the probability of occurrence of any spatial pattern can be obtained from the training data by counting the number of its occurrences.

Example 21. (Example 14, continued) The percent of heterogeneous pixels can be estimated, as in the previous example, by $\left| \left\{ t \in \hat{D} : t \text{ is heterogeneous} \right\} \right| / |\hat{D}|$.

The extensive use of training data has resulted in a considerable amount of literature
on the subject. The relevant issues which have been discussed include the independence of the training observations, the possible bias in the sample estimates due to existing spatial correlations [TuCo78], the optimal choice of a training sample, the implementation of certain estimation techniques for specific models [Owen84] and the pertinence of the obtained estimates for the whole image. This last issue, however, suggests one reason for which estimation through training samples may not always be desirable.

The training data is usually constrained by feasibility considerations (e.g., cost) to a small fraction of the whole image which is situated in an easily accessible part of the image. Thus, the estimates obtained from the training data may not represent accurately the values of the parameters from the whole image. Or, it may be desirable to allow the parameter (and their estimates) to vary with location. This idea is taken up in chapter 7 (see also Swit83, MoRi87 and Clif86). Clearly, in such a situation the estimates from one piece of training data may be deemed insufficient. Furthermore, training data may not be available (because, say, inspection of an area in Antarctica is impossible) or it may be considered inefficient (because it does not use the data from outside the training area).

One alternative to training data estimation is to combine the data from the rest of the image \( \{ O(t) : \forall t \in D - \hat{D} \} \) with the training data \( \{ I(t) \) and \( O(t) \), for \( t \in \hat{D} \} \). This idea attempts to exploit all the available observations and thus avoid the loss of information which stems from restricting the estimation procedure to the training area. With the appropriate formulation for the problem, the EM algorithm can be used to produce maximum likelihood estimates for the parameters [Hjor85b]. The following example illustrates this idea.

**Example 22.** Consider a binary image \( I(t) \in \{0, 1\} \), for \( t \in D \). The class probabilities are \( \pi_0 = P \{I(t) = 0\} \), \( \pi_1 = P \{I(t) = 1\} \) and the class (conditional) densities are \( f_0(t) = f_0(O(t)|I(t) = 0) \), \( f_1(t) = f_1(O(t)|I(t) = 1) \). Let \( \hat{D}_{\text{ind}} \) be a subset of the training set \( \hat{D} \) which consists of pixels that are sufficiently far apart from each other to be considered independent. Let also \( D_{\text{ind}} \) be a subset of \( D - \hat{D} \) which satisfies the same property. The density of each observation in \( \hat{D}_{\text{ind}} \) can be written as a mixture of \( f_0 \) and \( f_1 \):

\[
f(t) = \pi_0 f_0(t) + \pi_1 f_1(t) = f_1(t) + \pi_0 (f_0(t) - f_1(t)), \quad \forall t \in D_{\text{ind}}.
\]
Assume further that the pixels in $D_{\text{ind}}$ are independent from the pixels in $\hat{D}_{\text{ind}}$ (this can always be achieved by an appropriate selection of the pixels that comprise the two sets). Then the joint likelihood of $D_{\text{ind}}$ and $\hat{D}_{\text{ind}}$ is

$$L = L(D_{\text{ind}}, \hat{D}_{\text{ind}}) = L(D_{\text{ind}}) \cdot L(\hat{D}_{\text{ind}}) = \prod_{t \in D_{\text{ind}}} f(t) \cdot \left( \prod_{t_0 \in \hat{D}_{\text{ind}} \setminus I(t_0)=0} f_0(t_0) \prod_{t_1 \in \hat{D}_{\text{ind}} \setminus I(t_1)=1} f_1(t_1) \right).$$

This likelihood is a function of the class probabilities $\pi_0, \pi_1$ and the parameters of the conditional densities $f_0, f_1$. Thus, maximum likelihood estimates can be derived by maximizing $L$ with respect to these parameters. In practice, an iterative algorithm (like the EM algorithm) is needed in order to obtain the required estimates. Clearly, this idea can also be extended to the case of a categorical image with more than two possible values. In fact, the use of subimages which satisfy the independence property of $D_{\text{ind}}$ and $\hat{D}_{\text{ind}}$ has been a popular device for parameter estimation in a number of different contexts (see Example 24 in this chapter and section 3.4 in chapter 3).

Another way of combining the training data with the all the observations is to obtain moment estimates for the marginal probabilities of the classes in a categorical image [SwKL82].

**Example 23.** Let $I(t) \in \{I_1, \ldots, I_k\}$ be a polychotomous image and $\hat{I}(t)$ be an initial image estimate which results from some classification scheme. The class probabilities are $\pi_l = P\{I(t) = I_l\}$, for all $1 \leq l \leq k$. The estimate $\hat{I}(t)$ yields the confusion matrix, $C_D$, for the training data $\hat{D}$: each element $c_{ij}$ is the proportion of class $I_j$ pixels which are estimated by $\hat{I}(t)$ as class $I_i$. This matrix is assumed to be constant over the spatial domain. Let $p_l(t)$ be the proportion of pixels in a neighborhood $N_t$ of $t$ which are estimated to be of class $I_l$, and write $p(t) = (p_1(t), \ldots, p_k(t))^\prime$. The probability vector $\pi(t) = (\pi_1(t), \ldots, \pi_k(t))^\prime$ is estimated for each pixel $t$ from the expectation equation

$$p(t) = C_D \cdot \pi(t).$$

This idea is attractive because it allows estimates of the probabilities $\pi(t)$ which vary with the location $t$ in the spatial domain.

In the complete absence of training data other methods must be devised for parameter
estimation which use only the observations \( \{O(t) : \forall t \in D\} \). Maximum likelihood [Owen84, Sclo83], approximations to maximum likelihood [KaCh83, ChCh85b] and least squares estimation [KaCh83, Hara84, JiCh86] have all been used for this purpose. In addition, other more specialized techniques have been suggested as well. These are designed for use either with a specific model [Besa74, Ripl86] or with a particular image reconstruction method [MoHJ85, Besa86a]. For example, a very interesting approach is adopted in Besa77a, where an errors-in-variables formulation is developed for the problem (see also KaWD83, ChJi85 and JiCh86).

There is an abundance of parameter estimation techniques and comparisons between them. Nevertheless, the general problem of estimating the statistics of the original image \( I(t) \) from the “degraded image” \( O(t) \) does not appear to have been solved. The following examples are a sample of some of these methods.

**Example 24.** (Example 22, continued) [Hjor85b] If \( \hat{D} \) and \( \hat{D}_{\text{ind}} \) are not available, in principle maximum likelihood estimates can still be obtained by maximizing \( L(D_{\text{ind}}) \) with respect to the parameters of interest. However, the function \( L(D_{\text{ind}}) \) may be unbounded and/or it may be difficult to optimize with respect to many parameters. A viable alternative is to adopt this approach when the class descriptions (i.e., the class densities \( f_0 \) and \( f_1 \)) are known. Then, \( L(D_{\text{ind}}) \) is maximized only with respect to the class probabilities \( \pi_0 \) and \( \pi_1 \). Again an iterative algorithm is required in order to arrive at the estimates \( \hat{\pi} \) which maximize \( L(D_{\text{ind}}) \). Possibilities include a Newton-Raphson numerical optimization or the EM algorithm [DeLR77]. The idea of using a set of independent pixels \( D_{\text{ind}} \) for parameter estimation is basically the coding scheme of Besa74; it has been used repeatedly [CrJa83, CoCo84a, HjMo84, HoYu82]. Section 3.4 describes one way for constructing the set \( D_{\text{ind}} \) when the image obeys a Markov random field model (see Example 16).

**Example 25.** A different approach of applying the maximum likelihood method is to treat the observation parameters and the image values together [Sclo83]. Specifically, let \( f_i(t) = f_i(O(t); \theta_i|I(t) = I_i) \) be the class density for class \( I_i \) in a categorical image \( I(t) \in \{I_1, I_2, \ldots, I_l, \ldots, I_k\} \). The observation parameters for class \( I_i \) are denoted by \( \theta_i \). Writing \( f_i(t) = f_i(O(t); \theta_i) \) and replacing the \( I(t) \)'s by a set of indicators \( \{y_i(t), l = 1, 2, \ldots, k\} \) leads
to consideration of the likelihood

\[ L(O(t); \theta, y) = \prod_{t \in D} \prod_{l=1}^{k} [f_l(O(t); \theta_l)]^{y_l(t)}. \]

Both the indicators \( y \) and the parameters \( \theta \) can be estimated by maximizing the likelihood with respect to \( \{y, \theta\} \). However, this approach results in serious biases [Titt84] in the parameter estimates \( \hat{\theta} \). Such biases can be avoided by resorting to the mixture model of the previous example.

**Example 26.** A general scheme for estimation of spatial parameters is suggested in Hall88b. This method first estimates the minimum distance between pixels so that they are independent [Mora48, Mora50] and then divides the image into congruent, disjoint regions ("tiles") which are apportioned into a number of groups. All the members of a group are separated from each other by at least the estimated independence distance. Unbiased estimates for the parameters of interest are obtained from each tile, and the estimates from all the tiles in the same group are bootstrapped in order to derive a confidence interval (with equal tail probabilities) of arbitrary precision. Controlling the significance level for each such interval (there are as many intervals as the number of groups) and using the Bonferroni inequality allows construction of a confidence interval for the parameters. An interesting analogy exists between this method and the variance estimation technique which is presented in chapter 6 (see section 6.4).

**Example 27.** (Example 16, continued) Resorting to simple geometry leads to estimates, or a region of desirable values, for the parameter \( \beta \) [Rip88]. For the sake of concreteness, let \( N_t \) be the 8 pixels surrounding \( t \). If all the pixels in \( N_t \) are 0, it is natural that one wants \( P \{I(t) = 1|I(t') = 0, t' \in N_t\} \) to be small. Imposing an arbitrary bound (say, \( P_{\text{small}} \)), this restriction becomes a bound for \( \beta \):

\[
P \{I(t) = 1|I(t') = 0, t' \in N_t\} = \frac{1}{1 + e^{8\beta}} \leq P_{\text{small}} \Rightarrow \beta \geq \frac{1}{8} \log \left( \frac{1 - P_{\text{small}}}{P_{\text{small}}} \right).
\]

At the same time if all the pixels in \( N_t \) are 1, the probability \( P \{I(t) = 1|I(t') = 1, t' \in N_t\} \) should be large. Thus, this probability should exceed some other bound, \( P_{\text{big}} \), and one gets an
upper bound for $\beta$:

$$P \{ I(t) = 1 | I(t') = 1, t' \in N_t \} = \frac{e^{8\beta}}{1 + e^{8\beta}} \geq P_{\text{big}} \Rightarrow \beta \leq \frac{1}{8} \log \left( \frac{P_{\text{big}}}{1 - P_{\text{big}}} \right).$$

These cases are somewhat extreme but they illustrate the method. Similarly, a single estimate of $\beta$ can be obtained by assigning an arbitrary probability to any configuration of the pixels \( \{ t, N_t \} \) (see also section 3.4).

The last example is one method for estimating the spatial interaction parameter in a binary Markov random field. Other available methods (e.g., pseudolikelihood, coding, etc.) are discussed in greater detail in the next chapter (see section 3.4). This report presents another method for estimating the parameters in a binary Markov random field (see chapter 5). This estimation problem is a special case of the more general problem reviewed in this section namely, the estimation of the spatial parameters in image models.
Chapter 3

Markov random fields

This chapter presents the Markov random field models which are used in the rest of the report. Section 3.1 defines the concept of a Markov random field, as it applies to the pixel lattice of an image. Sections 3.2 and 3.3 describe respectively two broad classes of Markov random field models: symmetric (noncausal) models and mesh (causal) models. In particular, section 3.3 introduces the binary Markov field models for which a parameter estimation technique is developed in chapter 5. Section 3.4 reviews some of the previously suggested methods for parameter estimation.

3.1. Definitions.

The concept of a Markov random field was introduced in Examples 15 and 16 of Chapter 2. Let \( D \) be the pixel lattice (see section 2.1.1) and \( t \) be a pixel (or a site) in \( D \). For an image \( I(t) \), let \( p(I) \) be the probability distribution which assigns values to \( I(t) \). The conditional probability of the image value at site \( t \) given the image values at all the other sites \( D - \{t\} \) is \( p_t = p \{ I(t)|I(t'), t' \in D - \{t\} \} \). Viewed through these conditional probabilities \( p_t \), the probability distribution \( p(I) \) is called a Markov random field.

A Markov random field is local, or locally dependent, if the conditional distribution \( p_t \) depends only on the values of the pixels in the vicinity of \( t \). That is, if \( N_t \) is some neighborhood of pixels in the vicinity of \( t \), the Markov random field \( p(I) \) is locally dependent iff

\[
p_t = p \{ I(t)|I(t'), t' \in D - \{t\} \} = p \{ I(t)|I(t'), t' \in N_t \}, \quad \forall t \in D. \tag{3.1.1}
\]

In this condition the probability \( p_t \) and the neighborhood \( N_t \) are specific to pixel \( t \).

The neighborhood \( N_t \) is a subset of \( D - \{t\} \) and the pixels which belong to it are the neighbors of \( t \). The construction of a Markov random field on \( D \) is usually carried out by first
imposing a neighborhood structure on $D$ and then choosing a probability distribution which satisfies condition (3.1.1). This “a priori” specification of a Markov random field may at first seem unrealistic but it turns out to be rather convenient and less restrictive than it appears. The model $p(I)$ can be, at least partially, adapted to the image at hand by allowing $p(I)$ to contain some parameters. The alternative approach of trying to fit a model to a particular image (by choosing, for example, the most “appropriate” neighborhood structure) has been used occasionally [KaCh83] but is not very common.

There are many possibilities for a neighborhood structure. Two examples of commonly used neighborhoods are shown below:

<table>
<thead>
<tr>
<th></th>
<th>$t'$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$t'$</td>
<td>$t$</td>
<td>$t'$</td>
</tr>
<tr>
<td>$t'$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$t'$</th>
<th>$t'$</th>
<th>$t'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t'$</td>
<td>$t$</td>
<td>$t'$</td>
<td></td>
</tr>
<tr>
<td>$t'$</td>
<td></td>
<td>$t'$</td>
<td></td>
</tr>
<tr>
<td>$t'$</td>
<td>$t'$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first neighborhood is the cross which consists of the four pixels adjacent to $t$. The second neighborhood is the square of the eight pixels surrounding $t$. These neighborhoods are the first two members of a whole hierarchy of symmetric neighborhoods: the cross neighborhood is the first-order neighborhood, the square neighborhood is the second-order neighborhood, etc. On a square pixel lattice this hierarchical neighborhood system is translation invariant.

The almost arbitrary nature of the neighborhood structure gives rise to a number of Markov random fields. For example, each of the two previous neighborhoods leads to a different field on $D$. However, it is important to mention that the probability distribution (3.1.1) is generally subject to restrictive consistency conditions which are not at all obvious. These conditions are identified by the Hammersley–Clifford theorem and are thoroughly discussed in Besa74. For instance, this theorem dictates that “neighbor” is a symmetric relationship: if pixel $t'$ is a neighbor of $t$, then $t$ must be a neighbor of $t'$ as well.

Another issue which is related to the choice of a neighborhood structure is what happens at the boundaries of the lattice. The problem is that pixels situated on the boundaries (i.e., on the last row or column of pixels on $D$) may not have as many neighboring pixels as the neighborhood structure requires. Two major approaches exist for dealing with this end effect. The first is the free boundary assumption which essentially truncates the neighborhoods of
all the boundary pixels according to their relative position on the lattice. If a first-order neighborhood is used, according to the free boundary assumption all the pixels in the interior of the lattice have 4 neighbors, all the pixels on the boundaries have 3 neighbors and the four corner pixels have only 2 neighbors. The second approach is to regard the pixel lattice as a torus wrapped around its edges. Then, a pixel on the boundary becomes a neighbor of the corresponding pixel on the opposite boundary. Thus all the pixels have the same number of neighbors. The free boundary and the torus approach are illustrated below for the first-order neighborhood of a corner pixel $t$:

<table>
<thead>
<tr>
<th>$t$</th>
<th>$t'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t'$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$t$</th>
<th>$t'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t'$</td>
<td></td>
</tr>
</tbody>
</table>

The torus approach is intuitively less natural than the free boundary approach and, for that reason, has been criticized. However, both approaches have been used repeatedly (see ChKa82, KaCh83, Hall88b for a torus approach and CoCo84a, GeGe84 for a free boundary approach).

Associated with the neighborhood structure is the notion of a clique. A clique is a set of pixels which either consists of a single pixel or, otherwise, all the pixels in it are neighbors of one another. As an example, consider the first-order neighborhood and let $t_x$, $t_y$ denote respectively the row and the column of pixel $t$ in $D$. Then the cliques are of the form \{$(t_x, t_y)$\}, \{$(t_x, t_y)$, $(t_x - 1, t_y)$\} and \{$(t_x, t_y)$, $(t_x, t_y - 1)$\}. Cliques are essential to the construction of a valid Markov field. In particular, any Markov field $p(I)$ can be written as

$$p(I) \propto \exp \left[ \sum_{t \in D} G_t(I(t)) + \sum_{t \in D} \sum_{t' \in D} G_{tt'}(I(t), I(t')) + \sum_{t \in D} \sum_{t' \in D} \sum_{t'' \in D} G_{tt't''}(I(t), I(t'), I(t'')) + \ldots \right],$$

(3.1.2)

where the sum in the square brackets is finite and the last term involves as many indices as there are pixels in $D$. Here the functions $G$ are arbitrary, subject to the restriction that $G_{tt''}$ must be 0 if the indexing pixels \{$t, t', \ldots$\} do not form a clique [Besa74]. The $G$'s represent
the interaction terms between the pixel values: the $G_t$'s are the first-order interactions (of a pixel with itself), the $G_{tt'}$'s are the second-order interactions (between pairs of pixels), the $G_{tt't''}$'s are the third-order interactions (between triplets of pixels), etc.

Equation (3.1.2) gives an expansion of $p(I)$ in terms of arbitrary functions of all the existing cliques on the lattice $D$. The neighborhood structure enters in (3.1.2) implicitly through the definition of the cliques. The relation between the order of the neighborhood and the order of the interactions is discussed in Kitt86. In practice, for reasons of simplicity higher order interaction terms are often neglected. If, for instance, all the cliques of size three or more do not contribute anything to the expansion (i.e., all the $G$ functions with three or more arguments are taken to be 0), (3.1.2) simplifies to

$$p(I) \propto \exp \left( \sum_{t \in D} G_t(I(t)) + \sum_{t \in D} G_{tt'}(I(t), I(t')) \right).$$

(3.1.3)

The fields which satisfy this condition are called pairwise interactions Markov random fields.

As an example, consider a categorical image $I(t) \in \{I_1, I_2, \ldots, I_k\}$, with $k$ unordered classes. Let $G_t(I(t)) = \alpha_t$, iff $I(t) = I_t$ and $G_{tt'}(I(t), I(t')) = -\beta_{tm}$, iff $I(t) = I_t$ and $I(t') = I_m$. Then, (3.1.3) becomes

$$p(I) \propto \exp \left( \sum_{1 \leq t \leq k} \alpha_t n_t - \sum_{1 \leq t \leq m \leq k} \beta_{tm} n_{tm} \right),$$

(3.1.4)

where $n_t$ is the number of pixels which belong to class $I_t$ and $n_{tm}$ is the number of pairs of neighboring pixels which belong respectively to classes $I_t$ and $I_m$. The constants $\{\alpha_t, \beta_{tm}\}$ are arbitrary parameters (see also Example 16 in Chapter 2).

Locally Markov random fields have been widely used as image models for a variety of reasons. First, they are mathematically appealing because they, as a subclass of Markov random fields, describe the spatial structure of an image in a probabilistically formal way. Unlike most other image models, they specify an actual probability distribution which is consistent over the whole pixel lattice. Second, they enjoy substantial computational advantages because of the local dependence of $p_t$ on the values $\{I(t') : \forall t' \in D - \{t\}\}$. They are amenable to parallel implementations [GeGe84, CoCo84b] and often allow for a computationally simple optimiza-
tion of the posterior probability of the image (see section 2.1.3). Third, they have constituted the basis for a number of relatively successful image restoration algorithms.

Additionally, locally dependent Markov random fields are used to describe the local properties of an image. They attempt to capture the intuitive notion that the image value at \( t \) depends mostly on the values of the pixels around it. Indeed, condition (3.1.1) is a Markov-like property: in the absence of an obvious time ordering (as in a Markov process), the neighborhood \( N_t \) plays the role of the conditioning set. However, even simple Markov fields can exhibit positive correlations over arbitrary distances. In this sense, the name "Markov" is misleading. More importantly, this observation contradicts the fact that spatial correlations of image values decrease with distance. Consequently, Markov random fields can possess large-scale characteristics which may be inconsistent with the properties of an image. Such characteristics can be undesirable and may render a particular model inappropriate.

The last reservation can be partially remedied by an image restoration algorithm which relies on consistency between image and model only at a local level. Such an algorithm is Besag's ICM [Bes86a] which is discussed in chapter 4. A Markov random field model is not, and should not be viewed as, a definitive representation of an image. Rather, it is a convenient, and often coarse, tool for describing the spatial structure of an image. The model remains useful as long as it functions adequately in the context of a restoration algorithm. This is why specifying a model in advance is not that severe a limitation. On the other hand, any attempt to interpret the model in a stronger sense (e.g., as the true spatial distribution of the image) is far fetched. For example, artificial images which are created by simulating Markov random fields are often completely uncharacteristic of real images. As an additional word of caution, Markov random fields are not appropriate for all types of images. In particular, the consensus opinion indicates that they are less satisfactory in satellite images than, say, in tomographic images (see the discussion on Bes86a). Future research will show whether these models can be adapted to even more applications than the ones they currently enjoy.

The conditional probability formulation of (3.1.1) is not the only possible definition of a Markov random field. There exists an equivalent definition in terms of energy functions (or Gibbs potentials). This second definition will not be needed and is not presented here. The
interested reader can find a comprehensive mathematical treatment in KiSn80 and an image restoration technique based on the alternative definition in GeGe84.

3.2. Spatially Symmetric Models.

Consider a categorical image \( I(t) \in \{ I_1, I_2, \ldots, I_l, \ldots, I_k \} \) and a pairwise-interactions Markov random field, \( p(I) \), of the form (3.1.4):

\[
p(I) \propto \exp \left[ \sum_{1 \leq i \leq k} \alpha_i n_i - \sum_{1 \leq i < m \leq k} \beta_{lm} n_{im} \right].
\]

(3.2.1)

The \( \alpha \)'s and the \( \beta \)'s are the model parameters. A standard argument is to compare the probabilities of two realizations of the process \( I(t) \) which differ only at the value of pixel \( t \) \[Besa74]. This shows that for any class \( I_l \):

\[
p_t = p \{ I(t) = I_l | I(t'), t' \in D - \{ t \} \} \propto \exp \left[ \alpha_l - \sum_{m \neq l} \beta_{lm} S_t(m) \right],
\]

(3.2.2)

where \( S_t(m) \) is the number of neighbors of pixel \( t \) which belong to class \( I_m \). The last equation defines the field \( p(I) \) through the conditional probabilities \( p_t \); this is the Markovian definition of (3.1.1).

The parameters \( \{ \alpha_l, \beta_{lm} \} \) allow for a certain degree of flexibility and can be used to adapt the model (3.2.1) to a particular image. The \( \alpha \)'s are associated with the marginal probabilities of the classes \( p \{ I(t) = I_l \} \); they are related to each other by virtue of the fact that \( \sum_{1 \leq i \leq k} n_i = |D| \). The \( \beta \)'s are the interaction parameters and are required to satisfy a natural condition of symmetry: \( \beta_{lm} = \beta_{ml} \), for all \( l \) and \( m \). Apart from these minor restrictions, the \( \alpha \)'s and the \( \beta \)'s are essentially arbitrary.

Model (3.2.1) implies that all the pairs of neighboring pixels of the form \( (I_l, I_m) \) carry the same weight \( \beta_{lm} \) in the likelihood of the image \( p(I) \). This situation can be changed by adopting either a more complicated or a simpler version of (3.2.1). A more complicated model would allow for different \( \beta \)'s among horizontal than among vertical neighbors (this, for example, may necessarily be the case if the pixels are rectangles but not squares). On the other hand, a simpler model could have all the \( \beta \)'s equal to each other. Then, (3.2.2) becomes

\[
p_t = p \{ I(t) = I_l | I(t'), t' \in N_t \} \propto \exp [\alpha_l + \beta S_t(l)],
\]
which gives the conditional probabilities

\[ p_t = p \{ I(t) = I| I(t'), t' \in N_t \} = \frac{e^{(\alpha_I + \beta S_I(l))}}{\sum_{m=1}^{k} e^{(\alpha_m + \beta S_I(m))}}. \]  

(3.2.3)

This model can be further simplified to

\[ p_t = p \{ I(t) = I| I(t'), t' \in N_t \} = \frac{e^{\beta S_I(l)}}{\sum_{m=1}^{k} e^{\beta S_I(m)}}, \]  

(3.2.4)

if all the classes \(I_1, I_2, \ldots, I_k\) are equally likely: \( p\{ I(t) = I_l \} = \frac{1}{k}, \) for all \( 1 \leq l \leq k. \)

An interesting property of (3.2.3) is that the geometry of the neighborhood does not affect the conditional probability \( p\{ I(t) = I| I(t'), t' \in N_t \}. \) This probability depends only on the number of neighbors of pixel \( t \) which belong to class \( I_l \) and not on the positions of these neighbors in \( N_t. \) For instance, in a binary image \( I(t) \in \{0, 1\} \) the two second-order neighborhoods

\[
\begin{array}{ccc}
1 & 0 & 1 \\
0 & t & 0 \\
0 & 0 & 0 \\
\end{array}
\quad
\begin{array}{ccc}
0 & 1 & 0 \\
0 & t & 0 \\
0 & 1 & 0 \\
\end{array}
\]

assign to pixel \( t \) the same probability of being 1: \( p\{ I(t) = 1| I(t'), t' \in N_t \} = \frac{e^{(\alpha_1 + 2\beta)}}{e^{(\alpha_0 + 6\beta) + e^{(\alpha_1 + 2\beta)}}}. \)

In reality, of course, the alternating triangular pattern of the first neighborhood may be more unlikely than the linear pattern of the second neighborhood.

### 3.3. Markov Mesh Models.

A special family of Markov random field models are the Markov mesh models [AbHK65, Kana80]. A Markov mesh model is defined by the conditional distribution of \( I(t), \) given the values \( \{ I(t') \} \) of all the pixels which precede \( t. \) To define the concept of precedence, let \( t = (t_x, t_y) \) and \( t_x \) and \( t_y \) denote respectively the row and the column of pixel \( t \) on \( D. \) Then, pixel \( t' \) precedes pixel \( t \) iff either \( t'_x < t_x \) or \( t'_x = t_x \) and \( t'_y < t_y. \) The set of predecessors of \( t \) is graphically depicted as

\[
\begin{array}{cccc}
\quad & t' & t' & t' \\
\quad & t' & t' & t' \\
\quad & t' & t' & t \\
\quad & t & & \\
\end{array}
\]
Markov mesh models are also called “causal” Markov random fields because their unilateral nature resembles the time causality present in time series.

In accordance with the Markovian condition (3.1.1), the conditional distribution of $I(t)$ given $\{I(t'), t' \text{ precedes } t\}$ is taken to be a function of only a few of the predecessors of $t$. That is, $p\{I(t)|I(t'), t' \text{ precedes } t\} = p\{I(t)|I(t'), t' \in N_t\}$, where $N_t$ is a neighborhood of pixels preceding $t$. The simplest interesting case arises from the neighborhood $N_t = \{(t_x - 1, t_y), (t_x, t_y - 1)\}$:

<table>
<thead>
<tr>
<th>$t'$</th>
<th>$t = (t_x, t_y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For notational simplicity write $I(t_x, t_y)$ for $I(t)$. Then, the corresponding field is defined by the probabilities $p\{I(t)|I(t_x - 1, t_y), I(t_x, t_y - 1)\}$, for all $t \in D$. As an example, consider the field (3.2.5) with this neighborhood:

\[
\begin{align*}
p\{I(t) = 1|I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 0\} &= \frac{e^{\alpha_1}}{e^{\alpha_0 + 2\beta} + e^{\alpha_1}}, \\
p\{I(t) = 1|I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0\} &= \frac{e^{\alpha_1 + \beta}}{e^{\alpha_0 + \beta} + e^{\alpha_1 + \beta}}, \\
p\{I(t) = 1|I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 1\} &= \frac{e^{\alpha_1 + \beta}}{e^{\alpha_0 + \beta} + e^{\alpha_1 + \beta}}, \\
p\{I(t) = 1|I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1\} &= \frac{e^{\alpha_0 + 2\beta}}{e^{\alpha_0} + e^{\alpha_1 + 2\beta}}.
\end{align*}
\] (3.3.1)

These equations clearly illustrate that the information contained in $\alpha_0$ and $\alpha_1$ is redundant (see section 3.2). The two $\alpha$'s can be replaced by a single $\alpha = \alpha_1 - \alpha_0$, and the model becomes a two parameter model.

Another example of interest is the field

\[
p\{I(t) = j|I(t'), t' \in N_t\} = \frac{e^{(\alpha + \beta(S_t(j) - 1))I(t)}}{1 + e^{(\alpha + \beta(S_t(1) - 1))}}, \quad \text{for } j \in \{0, 1\},
\] (3.3.2)

where again $S_t(j)$ are the neighbors of $t$ that belong to class $j$. Fields of this form are special cases of (3.1.3) and are called auto-binomial [Besa74]. When the neighborhood of pixel $t$ is
\( N_t = \{(t_x - 1, t_y), (t_x, t_y - 1)\} \), this field takes a form similar to (3.3.1):

\[
\begin{align*}
    p \{ I(t) = 1 | I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 0 \} &= \frac{e^{(\alpha - \beta)}}{1 + e^{(\alpha - \beta)}}, \\
    p \{ I(t) = 1 | I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0 \} &= \frac{e^{\alpha}}{1 + e^{\alpha}}, \\
    p \{ I(t) = 1 | I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 1 \} &= \frac{e^{\alpha}}{1 + e^{\alpha}}, \\
    p \{ I(t) = 1 | I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1 \} &= \frac{e^{(\alpha + \beta)}}{1 + e^{(\alpha + \beta)}}.
\end{align*}
\] (3.3.3)

The parameters \( \alpha, \beta \) are again arbitrary and need to be specified for the model to become useful. Chapter 5 describes a method of parameter estimation which applies equally well to either field (3.3.1) or (3.3.3).

The Markov fields (3.3.1) and (3.3.3) are similar to the field (3.2.3) in that the geometric pattern of the neighborhood \( N_t \) does not influence the probability \( p \{ I(t) = 1 | I(t'), t' \in N_t \} \). All these fields simply count the neighbors of pixel \( t \) which belong to class 1 in order to determine the conditional probability that pixel \( t \) also belongs to class 1.

It is worth noticing that the neighborhood \( N_t = \{(t_x - 1, t_y), (t_x, t_y - 1)\} \) is the smallest possible neighborhood which gives rise to a two-dimensional Markov model. A process \( I(t) \) like the one in (3.3.1) makes explicit the simultaneous dependence of \( I(t) \) on both coordinate axes of the two-dimensional domain \( D \). This feature distinguishes models like (3.3.1) or (3.3.3) from one-dimensional Markov chains which have been used as image models by having been applied to either the rows or the columns of the pixel lattice.

It should also be emphasized that conditioning only on the predecessors of pixel \( t \) is quite different from conditioning on all the other pixels \( D - \{t\} \). In a model like (3.3.3) (or (3.3.1)) the conditional distribution of the value \( I(t) \) given all the predecessors of pixel \( t \) depends only on the values \( I(t_x - 1, t_y) \) and \( I(t_x, t_y - 1) \). Nevertheless, for the same model the conditional distribution of the value \( I(t) \) given all the other pixels \( D - \{t\} \) depends on the values \( I(t_x, t_y - 1), I(t_x + 1, t_y - 1), I(t_x - 1, t_y), I(t_x + 1, t_y), I(t_x - 1, t_y + 1) \) and \( I(t_x, t_y + 1) \) [Besa74]. That is, the probability \( p \{ I(t) | I(t'), t' \in D - \{t\} \} \) is a function of the 6 neighbors of \( t = (t_x, t_y) \) shown...
This can be seen by the same argument that yields equation (3.2.2), i.e., by comparing the probabilities of two realizations of the process $I(t)$. Let $R_t = \{t' : t' \in D - \{t\}, t \text{ precedes } t'\}$ be the set of successors of pixel $t$ and $P_t = \{t' : t' \in D - \{t\}, t' \text{ precedes } t\}$ be the set of predecessors of pixel $t$. Then the Markov field $p(I)$ is

$$p(I) = p(I(I(t), \{I(t'), t' \in R_t\}), \{I(t''), t'' \in P_t\})$$

$$= p(I(t''), t'' \in R_t|I(t), \{I(t'), t' \in P_t\}) p(I(t), \{I(t'), t' \in P_t\})$$

$$= p(I(t''), t'' \in R_t|I(t), \{I(t'), t' \in P_t\}) p(I(t)|I(t'), t' \in P_t) p(I(t'), t' \in P_t).$$

The first term $p(I(t''), t'' \in R_t|I(t), \{I(t'), t' \in P_t\})$ depends on $I(t)$ through the probabilities $p\{I(t_x, t_y + 1)|I(t_x - 1, t_y + 1), I(t)\}$ and $p\{I(t_x + 1, t_y)|I(t), I(t_x + 1, t_y - 1)\}$. The second term $p(I(t)|I(t'), t' \in P_t)$ depends on $I(t)$ through $p\{I(t)|I(t_x - 1, t_y), I(t_x, t_y - 1)\}$, and the third term $p(I(t'), t' \in P_t)$ does not depend on $I(t)$ at all. Thus $p(I)$ depends on the value of $I(t)$ through the factor

$$g_t(I(t)) = p\{I(t_x + 1, t_y)|I(t), I(t_x + 1, t_y - 1)\} p\{I(t_x, t_y + 1)|I(t_x - 1, t_y + 1), I(t)\}$$

$$\cdot p\{I(t)|I(t_x - 1, t_y), I(t_x, t_y - 1)\}. \tag{3.3.4}$$

Considering two realizations of the process $I$ which differ only at the value of $I(t)$ gives

$$p\{I(t) = 1|I(t'), t' \in D - \{t\}\} = \frac{q_t(1)}{q_t(1) + q_t(0)}, \tag{3.3.5}$$

$$p\{I(t) = 0|I(t'), t' \in D - \{t\}\} = \frac{q_t(0)}{q_t(1) + q_t(0)}.$$ 

The exact form of $q_t(I(t))$ is determined by the nature of the conditional probabilities of $I(t)$ given the predecessors of $t$, but equation (3.3.5) holds for any mesh model which is based on the neighborhood $N_t = \{(t_x - 1, t_y), (t_x, t_y - 1)\}$.

The unilateral nature of Markov mesh models differentiates such model from the noncausal fields of the previous section. The consequences of the existing causality can be both negative and positive. On the negative side, models which obey the unilateral structure of (3.3.1) have been criticized as restrictive [Bes86] and not natural for a spatial context because (a) the
unilateral nature of Markov mesh models sometimes makes it difficult to accommodate special properties of the image and (b) they suffer from a spatial asymmetry which can manifest itself with directionality side-effects in the reconstructed images. Such effects have been observed for reconstruction algorithms which are based on the conditional probability of \( I(t) \) given the values of the predecessors of \( t \). This asymmetry can sometimes be partially mitigated \([\text{Besa74}]\) or altogether avoided in some special models \([\text{Pick80, Hasl85}]\), but is generally an undesirable characteristic. Two ways for dealing with this asymmetry are discussed in section 4.3.

On the positive side, their unilateral structure frees Markov mesh models from the stringent consistency conditions of the Hammersley–Clifford theorem (see section 3.1). It also permits the direct simulation of such models in a raster-scan fashion along the rows of the pixel lattice (whereas the simulation of noncausal, discrete Markov fields is an impossible task \([\text{Besa86}]\)). More importantly, it enables the analytical computation of the posterior probability at each pixel. This has practical ramifications because it reduces greatly the computational complexity of the restoration algorithms (i.e., of the maximization of the posterior probabilities).

The last advantage was the reason for which Markov mesh models were among the first Markov fields to be used in image processing applications \([\text{DeEC84, KiFo84}]\). This is also the reason for which Markov mesh models are presented here. The estimation procedure of chapter 5 exploits the mathematical tractability of these models to obtain moment estimates for the parameters of the model. Once the parameters are estimated, the model is used for image reconstruction with the algorithm of section 4.3.

### 3.4. Parameter Estimation.

In practice, to use the field (3.2.3) as part of a restoration algorithm one must know the values of the parameters \( \{\alpha, \beta\} \). This report addresses the problem of estimating these parameters in the special case of a binary Markov random field. That is, the field of interest is (3.2.3) for binary images:

\[
p \{ I(t) = j | I(t'), t' \in N_t \} = \frac{e^{(\alpha_j + \beta S_t(j))}}{e^{(\alpha_0 + \beta S_t(0))} + e^{(\alpha_1 + \beta S_t(1))}}, \quad \text{for} \quad j \in \{0, 1\}. \tag{3.4.1}
\]

Here, as before, \( S_t(j) \) stands for the number of class \( j \) neighbors of pixel \( t \). The parameter \( \beta \)
can be positive or negative. When $\beta$ is positive the model discourages pairs of neighbors of the form $(0,1)$. This is the more interesting case because it corresponds to the intuitive notion of spatial continuity.

The estimation of $\{a_i, \beta\}$ for the field (3.2.3) is a necessary and nontrivial task. The main difficulty is that when the neighborhood structure is a symmetric one (e.g., a second-order neighborhood), the normalizing constant and the marginal probabilities of the distribution $p(I)$ defined by (3.2.3) are generally intractable. Thus, for example, maximum likelihood estimation is often not possible, since the majority of used models are symmetric. A second difficulty stems from the fact that in most image processing applications the hypothesized realization of the Markov field (i.e., the actual image $\{I(t)\}$) is not directly observable. Thus, when training data are not available, the parameters cannot be estimated from the process $I(t)$ to which they are actually related (see section 2.2.2).

One way of estimating the parameters of a Markov field is the geometric method of Example 27 in Chapter 2. The idea is to assign some probability to a particular configuration of a set of pixels and solve the resulting equation for the parameters of interest. As an example, consider a binary field of the form (3.2.4) with a second-order neighborhood:

$$p \{I(t) = 1 | I(t'), t' \in D - \{t\}\} = p \{I(t) = 1 | I(t'), t' \in N_t\} = \frac{e^{\beta S_t(1)}}{e^{\beta S_t(0)} + e^{\beta S_t(1)}}. \quad (3.4.2)$$

The set $\{t, N_t\}$ is a convenient choice for a set of pixels, because different configurations of this set can be turned immediately into equations for $\beta$ via (3.4.2). Specifically, it may be desirable that the pattern

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>t : $I(t) = 1$</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

gets assigned probability equal to some number, say $3/4$. It follows from equation (3.4.2) that $p \{I(t) = 1 | I(t'), t' \in N_t, N_t \text{ as above}\} = \frac{e^{\beta}}{e^{\beta} + e^{\beta}} = \frac{3}{4}$, which gives $\beta = \frac{1}{4} \ln 3$.

This method is conceptually simple and effectively cost free. Also, the probability assignment for any set of pixels can be made so that it reflects the frequency of occurrence of a particular pattern. Of course, the frequency of any pattern is a property of the image and thus, some knowledge of the image is required in order to make a realistic probability assign-
Section 3.4: Parameter Estimation

ment. On the other hand, it is obvious that this estimation method is very ad hoc and suffers from consistency problems. Different probability assignments to different sets of pixels may, and will in general, lead to distinct estimates of $\beta$. It is not clear how to achieve a compromise between such estimates. This is the reason for which this method may prove more useful in providing a range of values for $\beta$, rather than a specific estimate. A range of desirable values can be obtained as in Example 27 of Chapter 2 namely, by assigning a lower and an upper bound (instead of a specific value) to the probability (3.4.2). Nevertheless, even this modification does not satisfactorily resolve the consistency issue, because the bounds are arbitrary and different configurations can still produce different ranges for $\beta$.

Another way of estimating the parameters of a Markov field is the method of "coding". Coding was first introduced in Besa72, further described in Besa74 and subsequently adopted by several researchers [CrJa83, CoCo84a, HjMo84, HoYu82]. The idea has already been presented in Example 24 of Chapter 2. First, a subset of pixels, $D_{\text{ind}}$, is selected from the interior of the pixel lattice so that all the pixels in this subset can be regarded as independent. Pixels on the boundaries of the lattice are excluded to avoid possible end effects related to the artificial nature of the model around the edges. Then, the conditional likelihood of \{$I(t) : \forall t \in D_{\text{ind}}$\} given the values \{$I(t') : \forall t' \in D - D_{\text{ind}}$\} becomes the product of its components:

$$L(D_{\text{ind}}|D - D_{\text{ind}}) = \prod_{t \in D_{\text{ind}}} p \{ I(t)|I(t'), t' \in N_t \}.$$  \hspace{1cm} (3.4.3)

Thus, conditional maximum likelihood estimates of the unknown parameters can be obtained by maximizing this product with respect to the parameters.

To illustrate the method consider a first-order Markov field. The lattice can be divided into ♠'s and ✷'s as follows:

Since the field is assumed to be of order one, the ♠ pixels are mutually independent given
the values of the ♢ pixels. Thus, one possibility for $D_{\text{ind}}$ are all the ♣ pixels. Conditional maximum likelihood estimates can be obtained for the field parameters by maximizing $L(I(t), t = ♣ | I(t'), t' = ♢)$. Naturally, another possibility for $D_{\text{ind}}$ are the ♢ pixels and estimates can also be obtained from $L(I(t'), t' = ♢ | I(t), t = ♣)$. The estimates obtained from the ♢'s and the ♣'s are not independent. However, they can be combined appropriately to produce a final estimate (say, by averaging the two estimates).

The advantages of coding are simplicity and flexibility. The disadvantages are its relative inefficiency [BeMo75, KaCh83] and that it requires a realization of the process $I(t)$ in order to be carried out. In practice, an estimate of $I(t)$ is used in equation (3.4.3) to yield the desired estimates. For instance, in the implementation of an iterative algorithm like ICM, an initial estimate of $I(t)$ gives initial coding estimates for the parameters. These are used to obtain an improved estimate of the image, which in turn can be used to reestimate the image, etc. The problem is that the intermediate estimates of $I(t)$ (i.e., not the final estimate to which ICM converges) may be unreliable and misleading. They do not necessarily correspond to the nature of the true image (or even to the nature of the final estimate) and thus cannot be trusted to yield good parameters estimates.

An alternative to coding is the “pseudolikelihood” method [Besa75, Besa77b, Besa78]. Pseudolikelihood estimates the parameters of the Markov field $p(I)$ by maximizing the product $\prod_{t \in D} p \{ I(t) | I(t'), t' \in N_t \}$. This product is obviously not a true likelihood since $p(I)$ cannot be decomposed into such a product form except in the case of independence. Moreover, the product does not include the normalizing constant of $p(I)$. As with coding, boundary pixels may be excluded from the pseudolikelihood product. Pseudolikelihood is more efficient than coding and its estimates have been shown to be asymptotically consistent [GuYo86] and normally distributed [GeGr86] as the pixel lattice grows in size. It is perhaps for these reasons that pseudolikelihood is the most widely used estimation technique. However, just like coding, pseudolikelihood depends on some intermediate estimate of $I(t)$ which may not bear much relevance to the actual image. In fact, using the pseudolikelihood in the ICM algorithm to estimate iteratively the parameter $\beta$ for a model like (3.4.2) tends to increase the value of $\beta$ on every iteration. When $\beta$ is small this makes sense, but when $\beta$ is large raising its value even more is not as desirable because it results in reconstructed images which are far too smooth.
[Owen86].

The methods described here are not the only possibilities for estimating the parameters of a Markov random field. Other techniques have also been suggested [GeGe86, Kuns86] but are not commonly used. In fact, in practice it is often the case that the parameter $\beta$ in model (3.4.2) is selected by trial and error. The criterion is usually the visual quality of the resulting reconstructions. Traditionally, the values of $\beta$ that are adopted in the ICM algorithm range from 0.5 to 1.5, but higher values can be useful as well. To a certain extent, this fact reflects the ad hoc nature of the image model and of the parameter itself. It also provides an indication to the relationship between the values of $\beta$ and the reconstruction algorithms. Experience shows that the algorithms are not terribly sensitive to parameter values, provided the latter fall within some reasonable limits (e.g., they do not specify implausible probabilities). For example, the ICM algorithm is quite robust to the values of $\beta$ as long as smaller values are used in the first few iterations [Besa86a]. This relative robustness, however, does not negate the importance of parameter estimation techniques.
Chapter 4

The ICM Algorithm

This chapter presents a modification to a widely used image reconstruction algorithm. The algorithm is the ICM (iterated conditional modes) algorithm, and is commonly attributed to Besag [Bes86a]. Section 4.2 briefly describes the original version of ICM, and section 4.3 presents the modified version. The difference between the two versions lies in the Markov random field they use to model the unobserved image.

4.1. The Problem.

Consider a two-dimensional image $I(t)$ which is defined on the spatial domain $D$. The domain $D$ consists of a grid of pixels and each point $t$ in $D$ represents the center of a pixel in the grid (recall from section 2.1.1 that the domain of definition $D$ and the domain of observation $D'$ coincide). The shape of the grid is not important: for the sake of concreteness, let it be a rectangular grid of square pixels. The image $I(t)$ is an unobservable, binary function: $I(t) \in \{0, 1\}$. The values 0 and 1 are simply labels for a dichotomous characteristic (e.g., black/white, oil/no oil, water/ground, crop 1/crop 2, etc.); there is no (implied) ordering among them. At each point $t$, in $D$, there exists an observation $O(t)$ which is somehow related to the actual value $I(t)$ of the image at that point. The objective is to use the information contained in the data $\{O(t)\}$ in order to classify each pixel $t$ into one of the two classes: \{t : I(t) = 0\} (class 0) and \{t : I(t) = 1\} (class 1).

4.2. The Original Version.

The classification of the image can be accomplished by any one of the existing reconstruction algorithms (see section 2.1.3). It is assumed in this report that this task is to be carried out by the ICM algorithm of Besa86a. The ICM algorithm is an iterative procedure which constructs successive image estimates $\hat{I}^{(0)}(t), \hat{I}^{(1)}(t), \hat{I}^{(2)}(t), \ldots$ until convergence is attained.
The initial estimate \( \hat{I}^{(0)}(t) \) can be constructed by any classification procedure and is usually obtained by maximum likelihood. On the \( n \)th iteration, the new estimate \( \hat{I}^{(n)}(t) \) is obtained by maximizing the conditional probability \( P \left\{ I(t)\mid O(t), O(t'), \hat{I}^{(n-1)}(t'), t' \in D - \{t\} \right\} \). The maximization of this probability is performed independently for each pixel \( t \), while the labels of all the other pixels \( D - \{t\} \) are treated as known from the previous iteration.

As explained in chapter 2 (section 2.1.3), specification of the conditional probabilities \( P \left\{ I(t)\mid O(t), O(t'), \hat{I}^{(n-1)}(t'), t' \in D - \{t\} \right\} \) requires two components. The first component is a spatial model for the image. The second component is a model for the conditional distributions, \( P \{O(t)\mid I(t)\} \), of the observations given the corresponding image values \( I(t) \).

The ICM algorithm employs a nondegenerate, local Markov random field to represent the local characteristics of the image (see section 3.2). The field is taken to be symmetric because spatial symmetry is considered to be a desirable feature (see chapter 3). Also, the field is almost always restricted to be a pairwise interactions Markov random field (i.e., one of the form (3.1.3)) for reasons of computational simplicity. One possibility is the model (3.2.3)

\[
p \{ I(t) = j \mid I(t'), t' \in N_t \} = \frac{e^{(\alpha_j + \beta S_t(j))}}{e^{(\alpha_0 + \beta S_t(0))} + e^{(\alpha_1 + \beta S_t(1))}}, \quad \text{for} \quad j \in \{0, 1\}, \tag{4.2.1}
\]

with a second-order neighborhood \( N_t \) which consists of the eight pixels surrounding \( t \):

<table>
<thead>
<tr>
<th>t'</th>
<th>t'</th>
<th>t'</th>
</tr>
</thead>
<tbody>
<tr>
<td>t'</td>
<td>t</td>
<td>t'</td>
</tr>
<tr>
<td>t'</td>
<td>t'</td>
<td>t'</td>
</tr>
</tbody>
</table>

As before, \( S_t(j) \) denotes the number of pixels in the neighborhood \( N_t \) which belong to class \( j \) (\( j = 0, 1 \)). This is a spatially symmetric model which becomes operational when specific values are assigned to the parameters \( \alpha_0, \alpha_1 \) and \( \beta \).

With regard to the observations \( O(t) \), the conditional distributions \( P \{O(t)\mid I(t)\} \) are often known or assumed to be known. Moreover, different pixels are assumed to be conditionally independent given the corresponding image values: \( P \{O(t), O(t')\mid I(t), I(t')\} = P \{O(t)\mid I(t)\} P \{O(t')\mid I(t')\} \). As mentioned in chapter 2, these assumptions are both realistic and harmless.

Under these assumptions, the probability \( p \left\{ I(t)\mid O(t), O(t'), \hat{I}^{(n-1)}(t'), t' \in D - \{t\} \right\} \) can
be written as
\[
P \left\{ I(t) | O(t), O(t'), \hat{I}^{(n-1)}(t'), t' \in D - \{t\} \right\} \propto P \{ O(t) | I(t) \} P \left\{ I(t) | \hat{I}^{(n-1)}(t'), t' \in D - \{t\} \right\} = P \{ O(t) | I(t) \} p \left\{ I(t) | \hat{I}^{(n-1)}(t'), t' \in N_t \right\}.
\] (4.2.2)

This product is a function of the unknown value \( I(t) \). The first factor is specified by the conditional model for the observations and represents the contribution of the data. The second factor is specified by the Markov random field and represents the contribution of the spatial model. On the nth iteration, ICM derives the new estimate \( \hat{I}^{(n)}(t) \) for pixel \( t \) by maximizing the product
\[
P \{ O(t) | I(t) \} p \left\{ I(t) | \hat{I}^{(n-1)}(t'), t' \in N_t \right\}
\]
with respect to \( I(t) \). This is essentially a trivial task because for most of the commonly used local Markov random fields the probabilities \( p \left\{ I(t) | \hat{I}^{(n-1)}(t'), t' \in N_t \right\} \) are given by simple expressions (see chapter 3). On each iteration all the pixels are visited once, but the order in which they are visited is not important: any sequential arrangement is adequate.

The ICM algorithm (in the form described here) is guaranteed to converge. In fact, one of the most attractive features of the algorithm is its rapid convergence. This feature ensures that the final estimate, \( \hat{I}(t) \), for the image values is not affected by the long range properties of the Markov field \( p(I) \) (see section 3.1). The final estimate to which the algorithm converges is a local maximum of the posterior probability \( P \{ I(t), \forall t \in D | O(t), \forall t \in D \} \). A number of rather successful applications have established it as one of the standard reconstruction algorithms [HaHo86, Owen86, Ripl86, JeJu87, KeMa87].

4.3. A Modified Version.

As described in the previous section, the ICM algorithm always uses a symmetric Markov field to model the spatial structure of an image. Symmetry is desirable for reasons which have been outlined in chapter 3. On the other hand, symmetry sometimes renders certain parts of the implementation (e.g., parameter estimation) harder. This section proposes a modification to the ICM algorithm which attempts to attain a compromise between symmetry and tractability of the Markov model: the modified algorithm can use the parameter estimation technique of chapter 5 (or any other method) without sacrificing much spatial symmetry.
The modification to the ICM algorithm consists simply of using a Markov mesh model (like (3.3.1) or (3.3.3)) instead of a fully symmetric model (like (4.2.1)). The trick, though, is to use the mesh model not in the formulation of its definition (e.g., (3.3.1) or (3.3.3)) but in the alternative formulation which is derived from conditioning on all the other pixels (e.g., (3.3.5)). That is, on each iteration conditioning at pixel \( t \) takes place with respect to all the other pixel \( D - \{ t \} \) rather than just with respect to the predecessors of \( t \). Then, the implementation of the modified ICM is identical to the implementation described in the previous section.

The arguments in favor of local Markov fields (computational simplicity and consistency between the algorithm and the image only at a local level) apply to this case as well. Therefore, the models (3.3.1) and (3.3.3) become two obvious candidates for image reconstruction with the modified ICM. By analogy to (4.2.2), for either one of these models, the new estimate \( \hat{I}^{(n)}(t) \) for pixel \( t \) (on the \( n \)th iteration) is obtained by maximizing the product

\[
P \{ O(t) | I(t) \} p \left( I(t) | \hat{I}^{(n-1)}(t'), t' \in D - \{ t \} \right) = P \{ O(t) | I(t) \} \frac{q_t(I(t))}{q_t(1) + q_t(0)}, \tag{4.3.1}
\]

where \( q_t(I(t)) \) is defined in equation (3.3.4) and is determined by the specific model which is being used:

\[
q_t(I(t)) = p \{ I(t_x + 1, t_y) | I(t), I(t_x + 1, t_y - 1) \} p \{ I(t_x, t_y + 1) | I(t_x - 1, t_y + 1), I(t) \} \cdot p \{ I(t) | I(t_x - 1, t_y), I(t_x, t_y - 1) \}. \tag{4.3.2}
\]

Again, this maximization is performed separately for each pixel \( t \), while the labels of all the other pixels \( D - \{ t \} \) are regarded as known. The maximization of (4.3.1) is a little more complicated than the maximization of (4.2.2) because the computation of the quantity \( q_t(I(t)) / (q_t(1) + q_t(0)) \) is not as simple as the computation of \( p \{ I(t) | \hat{I}^{(n-1)}(t'), t' \in N_t \} \).

The idea behind conditioning on all the pixels \( D - \{ t \} \) is to avoid, or moderate, any directionality effects that may result from the causality of the mesh model. This feature distinguishes this algorithm from other reconstruction algorithms which are based on Markov mesh models, because most of them use one-sided conditioning. For models (3.3.1) and (3.3.3), as well as for any mesh model which is defined in terms of the predecessor neighborhood \( \{(t_x, t_y - 1), (t_x - 1, t_y)\} \), conditioning on all the other pixel \( D - \{ t \} \) is equivalent to conditioning on the six neighbors \( (t_x, t_y - 1), (t_x + 1, t_y - 1), (t_x - 1, t_y), (t_x + 1, t_y), (t_x - 1, t_y + 1), (t_x, t_y + 1) \).
1) (see section 3.3). Clearly, this can be thought of as an approximation to conditioning on the eight neighbors of the second-order neighborhood (e.g., in model (4.2.1)). Carrying this analogy to an extreme, one may forego the more complicated optimization of (4.3.1) in favor of the simpler optimization of (4.2.2) with only six neighbors entering in the probability \( p \left\{ I(t) | I^{(n-1)}(t'), t' \in N_t \right\} \). Some results on the performance of the modified algorithm are given in chapter 8.

The modified ICM algorithm tries to extract more spatial context from the Markov mesh model than an algorithm which would only use the two preceding neighbors, \((t_x - 1, t_y)\) and \((t_x, t_y - 1)\), of pixel \( t \). Nevertheless, the algorithm still does not use a fully symmetric neighborhood. First, the two diagonal neighbors \((t_x - 1, t_y - 1)\) and \((t_x + 1, t_y + 1)\) are omitted. Second, the six neighbors which enter in the conditional probability \( p \{ I(t) | I(t'), t' \in D - \{ t \} \} \) do not all enter in the same way (this is the case in (4.2.2)); see equation (4.3.2). This fact implies that the reconstructed images may exhibit some features which are associated with the direction of the underlying Markov model. A simple way of overcoming such features is to use different direction on successive iterations: the employed mesh model is based on the probabilities \( p\{ I(t) | I(t_x - 1, t_y), I(t_x, t_y - 1) \} \) for the first iteration, on the probabilities \( p\{ I(t) | I(t_x - 1, t_y), I(t_x, t_y + 1) \} \) for the second iteration, on the probabilities \( p\{ I(t) | I(t_x, t_y + 1), I(t_x + 1, t_y) \} \) for the third iteration, on the probabilities \( p\{ I(t) | I(t_x + 1, t_y), I(t_x, t_y - 1) \} \) for the fourth iteration, etc. The quantities \( q_t(I(t)) \) must then be adjusted accordingly but this is the only difference from (4.3.1).

The modified ICM algorithm shares the attractive properties of the original ICM algorithm (i.e., guaranteed convergence, trivial computation, rapid convergence and independence from the long-range effects of the Markov field). At the same time, parameter estimation for the modified ICM algorithm is a much easier task than for the original ICM algorithm. By virtue of the fact that it uses a more primitive Markov model, the modified ICM algorithm is amenable to the parameter estimation technique of the next chapter. As a result, the overall implementation of the algorithm is rather simple and, as indicated by chapter 8, the obtained reconstructions are comparable to the ones obtained from the original ICM.
Chapter 5

The Moment Estimator

This chapter develops a moment estimator for the parameters of a Markov random field in the context of an image reconstruction algorithm. The next section describes the framework of the problem and section 5.2 states the assumptions under which the problem is solved. Sections 5.3, 5.4, 5.5 and 5.6 develop the moment estimator. Section 5.7 presents an interesting special case. Section 5.8 discusses the features, both negative and positive, of the moment estimator.

5.1. The Problem.

Consider the classification of a two-dimensional, binary image $I(t)$ from the observed data $O(t)$ (see section 4.1). The spatial domain is a rectangular array of square pixels and the image values $I(t)$ are 0 or 1 (with no ordering between them). Suppose that the classification of the image is carried out by the modified ICM algorithm of section 4.3, and that the spatial structure of the image is described by a Markov mesh model (e.g., a model like (3.3.1) or (3.3.3)). Such a model becomes operational, and, therefore, applicable with the ICM algorithm, when its parameters are specified. This chapter develops a method for estimating the parameters of the field (3.3.3) from the observed data $O(t)$. As a result, the modified ICM algorithm becomes a useful reconstruction tool which is entirely driven by the available data.

5.2. Assumptions.

For the sake of concreteness, it is assumed that the modified ICM algorithm uses the Markov model (3.3.3) with the neighborhood $N_t = \{(t_x - 1, t_y), (t_x, t_y - 1)\}$ to describe the
\[ p \{ I(t) = 1 | I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 0 \} = \frac{e^{\alpha - \beta}}{1 + e^{\alpha - \beta}}, \]
\[ p \{ I(t) = 1 | I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0 \} = \frac{e^{\alpha}}{1 + e^{\alpha}}, \]
\[ p \{ I(t) = 1 | I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 1 \} = \frac{e^{\alpha}}{1 + e^{\alpha}}, \]
\[ p \{ I(t) = 1 | I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1 \} = \frac{e^{\alpha + \beta}}{1 + e^{\alpha + \beta}}. \]

(5.2.1)

There is no loss of generality from this assumption, because the estimation technique of the next section can be used for any Markov mesh model (and not only for (5.2.1)). This Markov random field involves two parameters: \( \alpha \) and \( \beta \). These are arbitrary and can be used to finetune the model to a particular image.

The \( \beta \) is the most important of the two parameters; intuitively, \( \beta \) quantifies the degree of spatial interaction between neighboring pixels (or, equivalently, measures the degree of spatial homogeneity among neighboring pixels). The case \( \beta = 0 \) corresponds to complete independence between neighbors (i.e., \( I(t) \) and \( I(t') \) are independent for all \( t \neq t' \)). The case \( \beta > 0 \) corresponds to positive interactions between neighbors: the field (5.2.1) encourages neighboring pixels to belong to the same class (hence, the name “attractive” field). The case \( \beta < 0 \) corresponds to negative interactions between neighbors: the field (5.2.1) discourages neighboring pixels from belonging to the same class (hence, the name “repulsive” field).

The parameter \( \alpha \) is associated with the marginal (or, “prior”) probabilities of the classes: \( \pi \triangleq p \{ I(t) = 1 \} \) and \( 1 - \pi \triangleq p \{ I(t) = 0 \} \). To see this, consider the trivial case \( \beta = 0 \) which reduces equation (5.2.1) to

\[ p \{ I(t) = 1 | I(t'), t' \in N_t \} = \frac{e^{\alpha}}{1 + e^{\alpha}}. \]

This verifies the independence between \( I(t) \) and \( \{ I(t'), t' \in N_t \} \), since the right hand side of the equation does not depend on the \( I(t') \)’s at all. Thus,

\[ \frac{e^{\alpha}}{1 + e^{\alpha}} = p \{ I(t) = 1 | I(t'), t' \in N_t \} = p \{ I(t) = 1 \} = \pi. \]

It follows that \( \pi = \frac{1}{2} \iff \alpha = 0 \). In other words, if the two classes are equally likely (i.e., \( \pi = p \{ I(t) = 1 \} = p \{ I(t) = 0 \} = \frac{1}{2} \)) then the parameter \( \alpha \) vanishes. This is equivalent to
Section 5.2: Assumptions

saying that $\alpha_1 = \alpha_2$ in (3.2.3). Indeed, when the classes are equally likely the $\alpha$'s must all be equal and, in light of their redundancy, they can be taken to be 0 (see also equation (3.3.1)).

Even though it is true that the $\alpha$ is associated with the marginal probability $\pi$, it is usually hard, or even impossible, to derive an explicit formula which expresses their relationship. The reason is that the probability $\pi$ is often an intractable quantity (see chapter 3). Equation (5.2.1) implicitly gives rise to a probabilistic constraint between $\pi$, $\alpha$ and $\beta$. This constraint, however, cannot be obtained analytically except in few special cases (e.g., in the uninteresting case $\beta = 0$ above). In particular, the relationship between $\pi$ and $\alpha$ cannot be isolated from $\beta$.

The goal of this chapter is to develop a procedure for estimating the parameters $\alpha$ and $\beta$ from the data $\{O(t)\}$. For that purpose it is convenient to assume at first that the data consists of univariate observations: $O(t) \in \mathbb{R}$. This assumption is made throughout this chapter. The observations can generally be multivariate, but the general case is not any harder than the simpler case of univariate observations. The moment estimator is extended to images with multivariate observations in chapter 7.

A model for the observations is usually specified through the conditional distribution $p\{O(t)|I(t)\}$ of the observations given their image values (see section 2.1.3). Since $I(t)$ takes values in the set $\{0,1\}$ there are two such distributions: $p\{O(t)|I(t) = 0\}$ and $p\{O(t)|I(t) = 1\}$. It is assumed here that the observations $O(t)$ obey the linear model

$$O(t) = (1 - I(t))[\mu_0 + \epsilon(t)\sigma_0] + I(t)[\mu_1 + \epsilon(t)\sigma_1] ,$$  

(5.2.2)

where $\epsilon(t)$ is a noise process with mean 0 and variance 1. The $\epsilon$'s are independent from each other and from all the image values $I(t)$. In other words, $\epsilon(t)$ and $\epsilon(t')$ are independent for all $t, t' \in D$, and $\epsilon(t)$ is independent of $\{I(t') : \forall t' \in D\}$ for all $t$. The class means $\mu_0$, $\mu_1$ and the class variances $\sigma_0^2$, $\sigma_1^2$ are assumed to be known. If it is further assumed that the noise process $\epsilon(t)$ is normal, then the conditional distributions $p\{O(t)|I(t) = j\}$ (for $j = 0,1$) are completely specified: they are normal distributions with the given means and variances. Normality is a harmless and popular assumption (see section 2.1.3). In fact, some of the computational simplicity of the ICM algorithm is due to the normality of the distributions $p\{O(t)|I(t) = j\}$. However, for the problem of estimating the parameters $\alpha$ and $\beta$ and for the method of moments developed in this chapter it is not necessary to require a particular distributional form for
p \{O(t)|I(t)\}. As long as \(\mu_0, \mu_1, \sigma_0^2\) and \(\sigma_1^2\) are known, the conditional distribution of \(O(t)\) given \(I(t)\) can be arbitrary.

It should be noted that the spatial independence of the noise process implies that, given the image values \(I(t)\) and \(I(t')\), the observations \(O(t)\) and \(O(t')\) are conditionally independent:

\[
p \{O(t), O(t')|I(t), I(t')\} = p \{O(t)|I(t)\} \cdot p \{O(t')|I(t')\}, \quad \forall t, t' \in D.
\]

This is a common assumption which formalizes the notion that the stochastic behavior of the observation \(O(t)\) depends on the image \(\{I(t') : \forall t' \in D\}\) only through the corresponding image value \(I(t)\). In other words, the distribution of \(O(t)\) is determined solely by the value of \(I(t)\) and the spatial correlations in the data are exclusively due to the correlations of the image values. This assumption plays an important role in the development of the moment estimator, because it allows one to express the moments of \(O(t)\) in terms of the unknown parameters \(\alpha\) and \(\beta\) without an additional model for the behavior of the \(\epsilon(t)\)'s.

This section describes the framework of the parameter estimation problem by introducing a number of assumptions. It is useful to recapitulate and distinguish among them those that are crucial to the solution of the problem. The shape of the pixel grid, the use of the (modified) ICM algorithm and the (possible) normality of the distributions \(p \{O(t)|I(t)\}\) are of secondary importance. The actual estimation procedure does not depend on these elements at all. They serve only to compose a meaningful context in which the parameters \(\alpha\) and \(\beta\) are to be estimated. For example, the parameters are still needed – and can be estimated according to section 5.3 – even if the field (5.2.1) is used with a reconstruction algorithm other than ICM. The assumption of univariate data is just a simplification which is eventually relaxed in chapter 7. Lastly, the following assumptions are critical:

- The image \(I(t)\) is binary.
- A Markov mesh model, like (5.2.1), is used to describe the image \(I(t)\). The exact form of the Markov random field is not important.
- The noise in the observations, \(\epsilon(t)\), is spatially independent, and the means and the variances of the conditional distributions \(p \{O(t)|I(t)\}\) are known.
The estimation of $\alpha$ and $\beta$ rests on these assumptions.

5.3. The Moment Estimator.

The field (5.2.1) can be written as

\[
\begin{align*}
p \{ I(t) = 1 | I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 0 \} &= \frac{A}{A + B}, \\
p \{ I(t) = 1 | I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0 \} &= \frac{A}{1 + A'}, \\
p \{ I(t) = 1 | I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 1 \} &= \frac{A}{1 + A'}, \\
p \{ I(t) = 1 | I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1 \} &= \frac{AB}{1 + AB},
\end{align*}
\]

(5.3.1)

by putting $A = e^\alpha$ and $B = e^\beta$. This parametrization is equivalent to the original parametrization, and the estimation of $(\alpha, \beta)$ amounts to the estimation of $A$ and $B$ under the feasibility constraint $A > 0, B > 0$. The remainder of this chapter adopts the $(A, B)$ parametrization because it simplifies notation.

For this Markov field consider the neighborhood $N_t = \{(t_x - 1, t_y), (t_x, t_y - 1)\}$ of pixel $t$

<table>
<thead>
<tr>
<th></th>
<th>(t_x - 1, t_y)</th>
<th>(t_x, t_y - 1)</th>
<th>(t_x, t_y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t_x - 1, t_y)</td>
<td>(t_x - 1, t_y)</td>
<td>(t_x, t_y - 1)</td>
<td>(t_x, t_y)</td>
</tr>
</tbody>
</table>

and let

\[\pi_{11} = p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1\} = p\{I(t_x, t_y) = 1, I(t_x, t_y - 1) = 1\} \]

denote the probability that the image values of two adjacent pixels are both 1, and

\[\tilde{\pi}_{11} = p\{I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1\} \]

denote the probability that the image values of two diagonally adjacent pixels (and in the orientation of $(t_x - 1, t_y), (t_x, t_y - 1)$) are both 1. It is convenient to tabulate the conditional probabilities of (5.3.1):
Table 5.1 Conditional probabilities for the Markov mesh model.

<table>
<thead>
<tr>
<th>$I(t_x - 1, t_y) + I(t_x, t_y - 1)$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p{I(t_x, t_y) = 1</td>
<td>I(t'_x), t'_y \in N_t }$</td>
<td>$\frac{A}{A+B}$</td>
<td>$\frac{A}{1+A}$</td>
</tr>
<tr>
<td>$p{I(t_x, t_y) = 0</td>
<td>I(t'_x), t'_y \in N_t }$</td>
<td>$\frac{B}{A+B}$</td>
<td>$\frac{1}{1+A}$</td>
</tr>
</tbody>
</table>

It follows, from this table, that

\[ p\{I(t_x, t_y) = 1 | I(t_x - 1, t_y) = 1 \} = \]

\[= p\{I(t_x, t_y) = 1 | I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0 \} \cdot p\{I(t_x, t_y - 1) = 1 | I(t_x - 1, t_y) = 1 \} + p\{I(t_x, t_y) = 1 | I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0 \} \cdot p\{I(t_x, t_y - 1) = 0 | I(t_x - 1, t_y) = 1 \}, \]

and, therefore,

\[ p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1 \} = \]

\[= \frac{AB}{1+AB} \cdot p\{I(t_x, t_y - 1) = 1 | I(t_x - 1, t_y) = 1 \} + \frac{A}{1+A} \cdot p\{I(t_x, t_y - 1) = 0, I(t_x - 1, t_y) = 1 \}, \]

or

\[
\pi_{11} = \frac{AB}{1+AB} \hat{\pi}_{11} + \frac{A}{1+A} (\pi - \hat{\pi}_{11}). \tag{5.3.2}
\]

Similarly, it follows from Table 5.1 that

\[ p\{I(t_x, t_y) = 1 | I(t_x - 1, t_y) = 0 \} = \]

\[= p\{I(t_x, t_y) = 1 | I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 1 \} \cdot p\{I(t_x, t_y - 1) = 1 | I(t_x - 1, t_y) = 0 \} + p\{I(t_x, t_y) = 1 | I(t_x - 1, t_y) = 0, I(t_x, t_y - 1) = 0 \} \cdot p\{I(t_x, t_y - 1) = 0 | I(t_x - 1, t_y) = 0 \}, \]

and, therefore,

\[ p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 0 \} = \]

\[= \frac{A}{1+A} \cdot p\{I(t_x, t_y - 1) = 1, I(t_x - 1, t_y) = 0 \} + \frac{A}{A+B} \cdot p\{I(t_x, t_y - 1) = 0, I(t_x - 1, t_y) = 0 \}, \]

or

\[
\pi - \pi_{11} = \frac{A}{1+A} (\pi - \hat{\pi}_{11}) + \frac{A}{A+B} (1 - 2\pi + \hat{\pi}_{11}). \tag{5.3.3}
\]

The two equations (5.3.2) and (5.3.3) involve the parameters $A, B$ and the probabilities $\pi, \pi_{11}, \hat{\pi}_{11}$. If these probabilities are known, the values of the parameters $A, B$ can be obtained.
by solving (5.3.2) and (5.3.3). This is the idea of the moment estimator: to estimate the probabilities \( \pi, \pi_{11}, \hat{\pi}_{11} \) and treat (5.3.2) and (5.3.3) as a system of two equations with two unknowns. The unknowns are the parameters of interest \( A, B \) and their moment estimates are the solutions to the system.

In other words, a three-dimensional statistic is used to estimate the two unknown parameters. The statistic is the triplet of estimates \((\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11})\) and the two unknown parameters are \( A, B \) (or, \( \alpha \) and \( \beta \)). This is an appealing approach because it neatly separates image from data parameters (see section 2.2.1). All the relevant image parameters \( \pi, \pi_{11}, \hat{\pi}_{11}, \alpha, \beta \) are estimated from the data parameters \( \mu_0, \mu_1, \sigma_0 \) and \( \sigma_1 \); only the latter need be known. This is also a realistic situation because in practice it is quite common to have some exogenous information on the nature of the data (e.g., the data parameters) independently of any information (or lack thereof) on the image itself.

It is worth mentioning that the system (5.3.2), (5.3.3) can be replaced by any other pair of probabilistic constraints which involve the parameters \( A \) and \( B \). For instance, the two parameters can also be estimated from the joint probabilities of the image values at pixels which are not necessarily adjacent or diagonally adjacent. This is possible provided expressions can be derived for these probabilities. Of course, the problem of computing the joint probabilities becomes more involved as the distance between the pixels grows larger (see BaBe69). Equations (5.3.2) and (5.3.3) are the most obvious candidates, because they are the two equations which are most readily related to the definition of the Markov random field.

5.4. Estimation of the marginal probability.

One way of estimating the marginal probability \( \pi \) is from the moments of the observed data \( \{ O(t) : \forall t \in D \} \). This can be done in a number of ways because constant functions \( G(\{ O(t) \}) \) of the data can produce a moment estimate for \( \pi \). For example, the expectation of any polynomial in the \( O(t) \)'s (e.g., any moment \( E(O^m(t)) \) of order \( m \)) is a linear function of \( \pi \) and, therefore, can be used to obtain an estimate \( \hat{\pi} \). The simplest possible moment estimate is derived from the mean of \( O(t) \)

\[
E(O(t)) = (1 - \pi) \mu_0 + \pi \mu_1,
\]

5.4.1
which gives

\[ \hat{\pi} \left( \{ O(t) + \eta \} \right) = \frac{\frac{1}{|D|} \sum_{t \in D} O(t + \eta) - (\mu_0 + \eta)}{(\mu_1 + \eta) - (\mu_0 + \eta)} = \hat{\pi} (\{ O(t) \}). \]

Here, \(|D|\) stands for the total number of pixels in the lattice \( D \). That is, if \( D \) is an \( n \times m \) rectangular grid then \(|D| = nm\). The estimate \( \hat{\pi} \) is, by design, unbiased. It is also translation invariant. That is, if the observations are perturbed by a constant amount, say \( \eta \), the estimate \( \hat{\pi} \) remains unaltered:

\[ \hat{\pi} (\{ O(t) + \eta \}) = \frac{\frac{1}{|D|} \sum_{t \in D} O(t + \eta) - (\mu_0 + \eta)}{(\mu_1 + \eta) - (\mu_0 + \eta)} = \hat{\pi} (\{ O(t) \}). \]

The only moment estimate which enjoys this property of translation invariance is the estimate \( \hat{\pi} \) obtained from the mean \( E(O(t)) \). Higher order moments of \( O(t) \) do not yield translation invariant estimates. This can be seen easily from the estimate \( \hat{\pi}_{(2)} \) which is obtained from the second moment \( E(O^2(t))\):

\[ \hat{\pi}_{(2)} (\{ O(t) \}) = \frac{\frac{1}{|D|} \sum_{t \in D} O^2(t) - \mu_0^2 - \sigma_0^2}{\mu_1^2 + \sigma_1^2 - \mu_0^2 - \sigma_0^2}, \]

\[ \hat{\pi}_{(2)} (\{ O(t) + \eta \}) = \frac{\frac{1}{|D|} \sum_{t \in D} O^2(t + \eta) - \mu_0^2 - \sigma_0^2 + 2\eta \left[ \frac{1}{|D|} \sum_{t \in D} O(t) - \mu_0 \right]}{\mu_1^2 + \sigma_1^2 - \mu_0^2 - \sigma_0^2 + 2\eta [\mu_1 - \mu_0]}. \]

Some numerical results concerning \( \hat{\pi} \) are presented in chapter 8.

Moment estimates may not always be asymptotically efficient but they are usually consistent because of the laws of large numbers. In particular, the estimate \( \hat{\pi} \) would be consistent if the observations \( O(t) \) were independent. However, as the observations depend on the image values and those are not independent, the consistency of \( \hat{\pi} \) is determined by the spatial structure of the image. To get a handle on the consistency of \( \hat{\pi} \) as the lattice \( D \) grows to infinity, the variance \( \text{var}(\hat{\pi}) \) must be computed. This is done in the next proposition. Let \( \rho_{tt'} = \text{corr}(I(t), I(t')) \) be the spatial correlation between two image values \( I(t) \) and \( I(t') \).

**Proposition 5.4.1** The variance of \( \hat{\pi} \) is given by the following formula:

\[ \text{var}(\hat{\pi}) = \frac{\pi(1 - \pi)}{|D|} + \frac{(1 - \pi)\sigma_0^2 + \pi\sigma_1^2}{|D|\mu^2} + 2 \frac{\pi(1 - \pi)}{|D|^2} \sum_{t \neq t'} \rho_{tt'}, \quad (5.4.3) \]

where \( \mu = \mu_1 - \mu_0 \).
Section 5.4: Estimation of the marginal probability

Proof. To simplify notation let $S$ be the sum $\sum_{t \in D} O(t)$, $k$ be the number $|D|$ of pixels in $D$ and $\mu = \mu_1 - \mu_0$. Then,

$$\hat{\pi} = \frac{1}{k} \frac{S - \mu_0}{\mu}$$

and the variance of $\hat{\pi}$ is given by

$$\text{var}(\hat{\pi}) = \frac{\text{var}(S)}{k^2 \mu^2}.$$  \hspace{1cm} (5.4.4)

Since

$$E(S) = k [(1 - \pi) \mu_0 + \pi \mu_1],$$

the variance of $S$ is

$$\text{var}(S) = E \left( S^2 \right) - k^2 [(1 - \pi) \mu_0 + \pi \mu_1]^2$$

$$= \sum_{t \in D} E \left( O^2(t) \right) + 2 \sum_{t \neq t'} E \left( O(t)O(t') \right) - k^2 [(1 - \pi) \mu_0 + \pi \mu_1]^2. \hspace{1cm} (5.4.5)$$

Since

$$E(O(t)O(t')) = E_{I(t),I(t')} \left( O(t)O(t') | I(t), I(t') \right)$$

$$= (1 - \pi) \mu_0^2 + \pi \mu_1^2 - \pi (1 - \pi) \mu^2 (1 - \rho_{tt'})$$

$$= [(1 - \pi) \mu_0 + \pi \mu_1]^2 + \pi (1 - \pi) \mu^2 \rho_{tt'} \hspace{1cm} (5.4.6)$$

and similarly,

$$E \left( O^2(t) \right) = (1 - \pi) (\mu_0^2 + \sigma_0^2) + \pi (\mu_1^2 + \sigma_1^2) \hspace{1cm} (5.4.7)$$

equations (5.4.4) and (5.4.5) obtain (5.4.3). \hfill \blacksquare

The formula for the variance of $\hat{\pi}$ makes explicit the dependence of the consistency of $\hat{\pi}$ on the spatial behavior of the image $I(t)$. This dependence can be formulated as a necessary and sufficient condition.

Proposition 5.4.2 A necessary and sufficient condition for $\hat{\pi}$ to be consistent is

$$\sum_{t \neq t'} \rho_{tt'} = o \left( |D|^2 \right). \hspace{1cm} (5.4.8)$$

Proof. Since $\hat{\pi}$ is unbiased, it is consistent iff $\text{var}(\hat{\pi}) \to 0$ as the lattice $D$ grows large. From
Proposition 5.4.1,
\[ \text{var}(\hat{\pi}) = \frac{o(|D|)}{|D|} + 2 \frac{\pi(1 - \pi)}{|D|^2} \sum_{t \neq t', t,t' \in D} \rho_{tt'} \]
and the result follows. □

This proposition gives an indication as to how the spatial structure of the image affects the properties of \( \hat{\pi} \). Condition (5.4.8) resembles some of the regularity conditions that appear in the analysis of time series. For instance, a similar condition stipulates that the cumulants of the process \( I(t) \) are summable (see section 2.6 in Bril81). The second-order cumulants are just the covariances and, therefore, summability of the cumulants implies summability of the correlations \( \rho_{tt'} \). However, summability of the \( \rho_{tt'} \)'s is a more stringent requirement than (5.4.8) which only gives a \(|D|^2\) rate for the sum \( \sum \rho_{tt'} \). Also, this condition is exactly the same for all the estimates of \( \pi \) which are obtained from the moments \( E(O^m(t)) \) of \( O(t) \) (namely, \( \hat{\pi}, \hat{\pi}(2), \hat{\pi}(3), \) etc.).

For some image models (e.g., the Markov random field (3.2.2)) it may be hard, or even impossible, to obtain the spatial correlations or evaluate their sum. Also, (5.4.8) does not always hold (take \( \rho_{tt'} = 1 \), for all \( t, t' \)). Nevertheless, (5.4.8) is a pretty weak condition. The worst case asymptotic behavior of \( \sum_{t,t' \in D} \rho_{tt'} \) is \( O \left( |D|^2 \right) \): this happens only in the unlikely event where all the \( \rho_{tt'} \) are equal to 1 and the sum \( \sum \rho_{tt'} \) becomes \( \frac{|D|(|D|-1)}{2} \). Furthermore, (5.4.8) does hold for the frequently used isotropic correlation function
\[ \rho_{tt'} = \rho^{\text{dist}(t,t')}, \]
where \( \rho \in [0,1) \) is a persistence parameter and \( \text{dist}(t,t') \) is some measure of distance between pixels \( t \) and \( t' \). The following propositions show that condition (5.4.8) prevails when \( \text{dist}(t,t') \) is the Euclidean distance or the city block distance. The proofs are deferred to the Appendix. As before, let \( t = (t_x, t_y) \) and \( t_x, t_y \) denote the coordinates of pixel \( t \) on \( D \).

**Proposition 5.4.3** Let \( \delta(t,t') = \text{vert} t_x - t'_x \) + \( |t_y - t'_y| \) be the city block distance between pixels \( t \) and \( t' \) on the lattice \( D \). If \( \rho_{tt'} = \rho^{\delta(t,t')} \), then \( \sum_{t,t' \in D} \rho_{tt'} = o \left( |D|^2 \right) \).
Section 5.4: Estimation of the marginal probability

Proposition 5.4.4 Let \( d(t, t') = \sqrt{|t_x - t'_x|^2 + |t_y - t'_y|^2} \) be the Euclidean distance between pixels \( t \) and \( t' \) on the lattice \( D \). If \( \rho_{tt'} = \rho_{d(t, t')} \), then \( \sum_{t, t' \in D} \rho_{tt'} = o(|D|^2) \).

As it turns out, for some special cases of Markov mesh models the correlations between the image values at different pixels can be obtained analytically. These correlations are of the form \( \rho_{tt'} = \rho_{\text{dist}(t, t')} \) [BaBe69] and, therefore, satisfy condition (5.4.8).

The estimate \( \hat{\pi} \) does not depend on any model for the image \( I(t) \). This independence is an advantage and a liability. It is an advantage in that it allows \( \hat{\pi} \) to be applicable under a wide range of circumstances. At the same time, it is a liability in that it does not take advantage of the properties of a particular model. For instance, \( \hat{\pi} \) does not exploit the Markovian structure of model (5.2.1) and thus may not be a fully efficient estimate of the probability \( \pi \) for the situation of section 5.1. If the spatial structure of the image is ignored, the estimation of \( \pi \) becomes equivalent to estimating the mixing probability in a mixture of two known densities. This is a problem which has been studied extensively and for which a number of estimators are available (a comprehensive reference is TiSM85). In this context, however, some of the existing estimators depend heavily on the distributional form of \( P \{ O(t)|I(t) \} \) for which nothing is assumed here.

Apart from possibly being inefficient, the estimator \( \hat{\pi} \) may also result in unacceptable estimates (e.g., values below 0 or greater than 1). This annoying feature is common to almost all moment estimators (see, for example, SwKl82 and sections 5.6, 5.8) and is dealt with on an individual basis from case to case. If the potential difficulties with \( \hat{\pi} \) are too serious to be rectified in an easy manner, one can always resort to more elaborate estimation procedures for \( \pi \). Examples 22 and 24 in chapter 2 describe one approach for computing approximate maximum likelihood estimates for \( \pi \). Another possibility is to obtain an initial estimate \( \hat{I}(t) \) of the image and use it to estimate \( \pi \) as in Example 20 of chapter 2 (see also section 3.4). However, in many cases the moment estimate \( \hat{\pi} \) (from (5.4.2)) is a simple and fairly reliable estimate.

Lastly, it should be mentioned that simple moment estimates \( \hat{\pi}_I \) can be obtained for the marginal probabilities \( \pi_I = P \{ I(t) = I_l \} \) even when the \( I(t) \) assumes more than two values: \( I(t) \in \{ I_1, I_2, \ldots, I_k \} \). The estimates \( \hat{\pi}_I \), for \( 1 \leq l \leq k - 1 \), constitute the solution to a linear
system of equations which is constructed from the first \( k - 1 \) moments of \( O(t) \).

5.5. Estimation of the joint probabilities.

The joint probabilities \( \pi_{11}, \bar{\pi}_{11} \) can also be estimated from the moments of the observed data \( \{O(t) : \forall t \in D\} \). For the estimation of \( \pi_{11} \) consider two adjacent pixels \( t, t' \) (e.g., \( (x, y) \) and \( (x - 1, y) \) or \( (x, y) \) and \( (x, y - 1) \)). Let

\[
\rho_{tt'} = \text{corr}(I(t), I(t'))
\]

be the correlation between their image values and notice that

\[
\pi_{11} = p\{I(t) = 1, I(t') = 1\} = E(I(t)I(t')) = \pi(1 - \pi)\rho_{tt'} + \pi^2.
\]

Therefore,

\[
\pi_{11} = \pi(1 - \pi)\rho_{tt'} + \pi^2 = \pi - \pi(1 - \pi)(1 - \rho_{tt'}).
\]

and the estimation of \( \pi_{11} \) reduces to the estimation of \( \rho_{tt'} \). A moment estimate for the latter can be obtained from the expectation of the product \( O(t)O(t') \) which is easy to calculate (see (5.4.6)):

\[
E (O(t)O(t')) = (1 - \pi)\mu_0^2 + \pi \mu_1^2 - \pi(1 - \pi)\mu^2 + \pi(1 - \pi)\mu^2 \rho_{tt'}.
\]

Thus \( \rho_{tt'} \) can be estimated from

\[
\hat{\rho}_{tt'} = \frac{\hat{E}(O(t)O(t')) - [(1 - \pi)\mu_0 + \pi \mu_1] ^2}{\pi(1 - \pi)\mu^2},
\]

and the corresponding estimate for \( \pi_{11} \) is

\[
\hat{\pi}_{11} = \pi^2 + \pi(1 - \pi)\hat{\rho}_{tt'}
= \pi^2 + \frac{\hat{E}(O(t)O(t')) - [(1 - \pi)\mu_0 + \pi \mu_1] ^2}{\mu^2}.
\]

The estimation of \( \bar{\pi}_{11} \) is analogous. Consider two diagonally adjacent pixels \( t, t'' \) which are in the orientation of \( (x - 1, y), (x, y - 1) \) and let

\[
\hat{\rho}_{tt''} = \text{corr}(I(t), I(t'')).
\]

The same argument as before gives

\[
\hat{\rho}_{tt''} = \frac{\hat{E}(O(t)O(t'')) - [(1 - \pi)\mu_0 + \pi \mu_1] ^2}{\pi(1 - \pi)\mu^2}
\]
and
\[
\hat{\pi}_{11} = \pi^2 + \pi(1 - \pi)\hat{\rho}_{11}
\]
\[
= \pi^2 + \frac{\bar{E}(O(t)O(t'')) - [(1 - \pi)\mu_0 + \pi\mu_1]^2}{\mu^2}
\] (5.5.6)

The only difference between equations (5.5.4) and (5.5.6) is the relative location of the pixels \(t, t'\) and \(t, t''\).

The estimates \(\hat{\pi}_{11}, \hat{\pi}_{11}\) are unbiased and consistent if the estimates of the expectations \(E(O(t)O(t'))\), \(E(O(t)O(t''))\) are unbiased and consistent. These expectations can be estimated by their sample averages:
\[
\bar{E}(O(t)O(t')) = \frac{1}{|\{(t, t') : t, t' \in D\}|} \sum_{t, t'} O(t)O(t'),
\] (5.5.7)
\[
\bar{E}(O(t)O(t'')) = \frac{1}{|\{(t, t'') : t, t'' \in D\}|} \sum_{t, t''} O(t)O(t'').
\] (5.5.8)

Equations (5.5.7) and (5.5.8) differ in an important detail. The sum in equation (5.5.7) is over all pixels \(t, t'\) which are adjacent to each other in either direction. This is the case because the Markov field (5.3.1) is symmetric with respect to the pairs \((t_x - 1, t_y), (t_x, t_y)\) and \((t_x, t_y), (t_x, t_y - 1)\):

\[
\text{corr}(I(t_x - 1, t_y), I(t_x, t_y)) = \text{corr}(I(t_x, t_y), I(t_x, t_y - 1)).
\]

On the other hand, the sum in equation (5.5.8) is over all the pixels \(t, t''\) which are diagonally adjacent but also conform to the configuration of pixels \((t_x - 1, t_y), (t_x, t_y - 1)\). That is, since the Markov field (5.3.1) is a mesh model, direction is important in its specification and

\[
\text{corr}(I(t_x - 1, t_y), I(t_x, t_y - 1)) \neq \text{corr}(I(t_x - 1, t_y - 1), I(t_x, t_y)).
\]

In practice one may choose to ignore the directionality of the mesh model and adopt the more symmetric estimator which extends the sum in (5.5.8) to all diagonally adjacent pixels. Strictly speaking this is not mathematically consistent with the model (5.3.1), because it is equivalent to imposing two such models on the pixel lattice: one is the model determined by the conditional probabilities \(p\{I(t_x, t_y)|I(t_x - 1, t_y), I(t_x, t_y - 1)\}\) and the other is the model determined by the conditional probabilities \(p\{I(t_x, t_y)|I(t_x - 1, t_y), I(t_x, t_y + 1)\}\) (see section 4.3 for a similar implementation of the modified ICM algorithm). On the other hand, this approximate estimator enjoys the advantage of greater spatial symmetry.
Chapter 5: The Moment Estimator

The estimates \( \hat{E}(O(t)O(t')) \), \( \hat{E}(O(t)O(t'')) \) are trivially unbiased. Therefore, establishing their consistency is equivalent to showing that their variance goes to 0 as the size of the pixel lattice grows. This is a tedious task which will not be pursued here. The reason is that the variance of \( \hat{E}(O(t)O(t')) \) (or \( \hat{E}(O(t)O(t'')) \)) depends not only on the variances of the individual terms \( O(t)O(t') \) (or \( O(t)O(t'') \) respectively) but also on the covariances between such terms, and it is not clear how the computation of these covariances can be accomplished.

It is rather obvious that the data functions \( E(O(t)O(t')) \) and \( E(O(t)O(t'')) \) are not the only recourse to estimating the correlations \( \rho_{tt'} \) and \( \rho_{tt''} \). Any other function of the data which involves these correlations can be used as well. For example, estimating \( \rho_{tt'} \) from \( E(O(t)O(t')) \) is equivalent to estimating it from

\[
E(O(t) - O(t'))^2 = 2E(O^2(t)) - 2E(O(t)O(t')) \\
= 2[(1 - \pi)(\mu_0^2 + \sigma_0^2) + \pi(\mu_1^2 + \sigma_1^2)] - 2[(1 - \pi)\mu_0 + \pi\mu_1]^2 \\
- 2\pi(1 - \pi)\mu^2\rho_{tt'}.
\]

However, since \( E(O(t)O(t')) \) is – in some sense – the simplest possible function of the data which involves the correlation \( \rho_{tt'} \) it appears to be the most natural choice.

5.6. Estimation of the parameters.

The system of equations (5.3.2) and (5.3.3) is linear in \( B \) and quadratic in \( A \):

\[
A^2B(\pi_{11} - \pi) + AB(\pi_{11} - \bar{\pi}_{11}) + A(\pi_{11} - \pi + \bar{\pi}_{11}) + \pi_{11} = 0, \\
AB(\bar{\pi}_{11} - \pi_{11}) + A^2(2\pi - 1 - \pi_{11}) + A(3\pi - 1 - \pi_{11} - \bar{\pi}_{11}) + B(\pi - \pi_{11}) = 0.
\]

If the values of \( \pi, \pi_{11}, \bar{\pi}_{11} \) are known, the solution for \( A, B \) is straightforward. Eliminating \( B \) yields a fourth degree equation in \( A \):

\[
-(\pi_{11} - \pi)(\pi_{11} - 2\pi + 1)A^4 + [\pi - 3\pi^2 - 2\pi_{11}^2 - 2(1 - 3\pi)\pi_{11} + (1 - \pi)\bar{\pi}_{11}]A^3 \\
- (1 - 2\pi)(\pi_{11} - \bar{\pi}_{11})A^2 - (\pi\bar{\pi}_{11} - 2\pi_{11}^2 + 2\pi\pi_{11} - \pi^2)A - \pi_{11}(\pi - \pi_{11}) = 0.
\]

Inspection shows that \( A = -1 \) satisfies this equation. Of course, \( A = -1 \) is not an acceptable solution because \( A \) must be positive. However, this observation reduces the previous equation to a third degree equation for \( A \):

\[
-(\pi_{11} - \pi)(\pi_{11} - 2\pi + 1)A^3 + [(1 - \pi)\bar{\pi}_{11} - \pi_{11}^2 - \pi^2 - (1 - 3\pi)\pi_{11}]A^2 \\
+ [\pi(\pi - \bar{\pi}_{11}) - \pi_{11}(\pi - \pi_{11})]A - \pi_{11}(\pi - \pi_{11}) = 0.
\]
This equation always has a real root. More importantly, it always has a positive real root. To see that (5.6.2) has a positive root recall that the product of the three roots of any equation
\[ \gamma_3 z^3 + \gamma_2 z^2 + \gamma_1 z + \gamma_0 = 0 \]
is given by
\[ z_1 z_2 z_3 = -\frac{\gamma_0}{\gamma_3}. \]
In this case, the product of the three roots $A_1, A_2, A_3$ of (5.6.2) is
\[ A_1 A_2 A_3 = \frac{\pi_{11}(\pi - \pi_{11})}{(\pi_{11} - \pi)(\pi_{11} - 2\pi + 1)} = \frac{\pi_{11}}{\pi_{11} - 2\pi + 1}. \] (5.6.3)
This quantity is positive since $\pi_{11} \geq 2\pi - 1$ (by the Bonferoni inequality). Thus,
\[ A_1 A_2 A_3 > 0 \]
and (5.6.2) must have one positive root. If the three roots $A_1, A_2, A_3$ are real they cannot all be negative. If, on the other hand, two of them are complex they must comprise a conjugate pair; this makes their product positive and forces the third root to be also positive.

The estimate $\hat{A}$ of $A$ is a positive root of the equation (5.6.2):
\[ \hat{A} = \text{a positive root of (5.6.2)}. \] (5.6.4)

The corresponding estimate $\hat{B}$ for $B$ is obtained from either one of the two equations of (5.6.1):
\[ \hat{B} = \frac{1}{A} \frac{\pi_{11} - \frac{\hat{A}}{1 + \hat{A}}(\pi - \hat{\pi}_{11})}{\hat{\pi}_{11} - \pi_{11} + \frac{\hat{A}}{1 + \hat{A}}(\pi - \hat{\pi}_{11})} = \frac{\hat{A} \pi_{11} + (\hat{A} + 1)\pi_{11} - \hat{A} \pi}{\hat{A}^2 \pi - \hat{A}(\hat{A} + 1)\pi_{11} + \hat{A} \hat{\pi}_{11}}. \] (5.6.5)
Again, it is easy to see that this equation defines a positive quantity. Observe that
\[ \pi_{11} = p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1\} \]
\[ \geq p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0\} \]
\[ = p\{I(t_x, t_y) = 1|I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0\}p\{I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0\} \]
\[ = \frac{A}{1 + A} [p\{I(t_x - 1, t_y) = 1\} - p\{I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1\}] \]
\[ = \frac{A}{1 + A} (\pi - \hat{\pi}_{11}), \]
and

\[ \hat{\pi}_{11} = p\{I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1\} \]
\[ \geq p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1\} \]
\[ = p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 1\} \]
\[ + p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0\} \]
\[ - p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0\} \]
\[ = p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1\} - p\{I(t_x, t_y) = 1, I(t_x - 1, t_y) = 1, I(t_x, t_y - 1) = 0\} \]
\[ = \pi_{11} - \frac{A}{1 + A}(\pi - \hat{\pi}_{11}). \]

Hence, both the numerator and the denominator in (5.6.5) are positive and so is \( \hat{B} \).

It is certainly reassuring to know that the model is self-consistent, in that the system (5.6.1) always has a solution \((\hat{A}, \hat{B})\) for which both \( \hat{A} \) and \( \hat{B} \) are positive. In practice, however, the probabilities \( \pi, \pi_{11}, \hat{\pi}_{11} \) are not known and the estimates \( \hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11} \) are used in their place. These estimates may be such that none of the solutions to (5.6.1) are feasible. For example, it may be the case that \( \hat{\pi}_{11} < 2\hat{\pi} - 1 \) and equation (5.6.2) does not have a positive, real root. Such discrepancies may be due either to random fluctuations in the observed data or due to possible inadequacies of the Markov model.

One way of dealing with this problem is to enforce the necessary conditions on the estimates \( \hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11} \), so that \( \hat{A} > 0 \) and \( \hat{B} > 0 \). A set of sufficient conditions for \( \hat{A} > 0 \) is \( \{0 \leq \hat{\pi} \leq 1, 2\hat{\pi} - 1 \leq \hat{\pi}_{11} \leq \hat{\pi}\} \) (another possibility is the set \( \{0 \leq \hat{\pi}_{11} \leq \hat{\pi} \leq \frac{1}{2}\} \) but this is unnecessarily restrictive). Unfortunately, it has not been possible to obtain in general a set of sufficient conditions for \( \hat{B} > 0 \) solely in terms of \( \hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11} \). This has been possible only in the special case where \( \pi = \frac{1}{2} \) (i.e., \( A = 1 \)); see section 5.7 below. Nevertheless, an arbitrary choice can be enforced for \( \hat{B} \) if the estimated value turns out to be negative. The obvious choice is to set \( \hat{B} \) equal to 0 and make \( \hat{B} \) be equal to some large, negative number.

An interesting issue arises when equation (5.6.2) has more than one positive root. Then a decision must be made as to which positive root is the most appropriate estimate for \( A \). This decision is easily made if only one solution \((\hat{A}, \hat{B})\) to the system (5.6.1) satisfies the feasibility constraint \( \hat{A} > 0, \hat{B} > 0 \). If two or three of the solutions to (5.6.1) satisfy this constraint, an
arbitrary choice must be made for \((\hat{A}, \hat{B})\) on the basis of other considerations.

## 5.7. Some special cases.

A special case of interest is when the two classes \(\{t: I(t) = 0\}\) and \(\{t: I(t) = 1\}\) are equally likely. Then, \(\alpha = 0, A = 1\) and \(\pi = \frac{1}{2}\) (it is easy to check that when \(A = 1\) equations (5.3.1) imply \(\pi = \frac{1}{2}\)). Table 5.1 becomes

<table>
<thead>
<tr>
<th>(I(t_x - 1, t_y) + I(t_x, t_y - 1))</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p{I(t_x, t_y) = 1</td>
<td>I(t'), t' \in N_t})</td>
<td>(\frac{1}{1+B})</td>
<td>(\frac{1}{2})</td>
</tr>
<tr>
<td>(p{I(t_x, t_y) = 0</td>
<td>I(t'), t' \in N_t})</td>
<td>(\frac{B}{1+B})</td>
<td>(\frac{1}{2})</td>
</tr>
</tbody>
</table>

and the Markov field (5.2.1) degenerates to a one-parameter model [BaBe69]. In this case, there is only \(B\) to estimate and the system (5.6.1) consists of a single equation (e.g., equation (5.3.2)):

\[
\pi_{11} = \frac{B}{1+B} \bar{\pi}_{11} + \frac{1}{2}(\frac{1}{2} - \bar{\pi}_{11}).
\]

It follows that the estimate of \(B\) is

\[
\hat{B} = \frac{\frac{1}{2} - \bar{\pi}_{11} - 2\pi_{11}}{2\pi_{11} - \frac{1}{2} - \bar{\pi}_{11}}.
\] (5.7.1)

It is easy to see that this value is positive, provided \(\pi_{11} \leq \pi = \frac{1}{2}\). Observe that

\[
\text{sign}(\hat{B}) = \text{sign}(\frac{1}{2} - \bar{\pi}_{11} - 2\pi_{11})\text{sign}(2\pi_{11} - \frac{1}{2} - \bar{\pi}_{11})
\]

\[
= \text{sign}(\bar{\pi}_{11}^2 - \frac{1}{2} - 2\pi_{11})^2
\]

\[
= \text{sign}(\bar{\pi}_{11}^2 + 2 - \frac{1}{4} - 4\pi_{11}^2).
\]

and that

\[
\pi_{11}^2 \leq \pi^2 = \frac{1}{4} \Rightarrow -4\pi_{11}^2 \geq -1 \Rightarrow \bar{\pi}_{11}^2 + 2 - \frac{1}{4} - 4\pi_{11}^2 \geq \bar{\pi}_{11}^2 + \frac{3}{4} > 0.
\]

Therefore, \(\hat{B}\) is an acceptable estimate of \(B\) if \(\pi_{11} \leq \frac{1}{2}\). This inequality is obviously satisfied by the true probabilities \(\pi, \pi_{11}\) but it gives a sufficient condition for the feasibility of \(\hat{B}\) in terms of the estimates \(\hat{\pi}, \hat{\pi}_{11}\). That is, if the estimates \(\hat{\pi}, \hat{\pi}_{11}\) are used, they must satisfy this inequality for the estimate \(\hat{B}\) to be acceptable.

Other special cases of interest are the three cases of limiting behavior for \(B\). When \(B \to 0\) (i.e., \(\beta \to -\infty\)) the Markov field becomes more and more repulsive, and when \(B \to \infty\) (i.e.,
$\beta \to \infty$) the Markov field becomes more and more attractive (see section 5.2). When $B = 1$
or approaches 1 (i.e., $\beta = 0$ or $\beta \to 0$) the Markov field reduces to a trivial specification
of independence between the image values at different pixels. These cases are conceptually
interesting but do not present any difficulties, as they follow from the discussion of the previous
sections.


The moment estimator of the previous sections is simple and almost trivial to implement.
The data is used only once, at the beginning of the reconstruction process, to estimate the
parameters of the Markov field. There is no need for interpreting the intermediate classifica-
tions that arise during the iterations of the modified ICM algorithm (see sections 3.4 and 4.3).
Nor is there need for resorting to the degree of arbitrariness that is present in the geometric
method of Example 27 in chapter 2 (see also section 3.4). Even if the parameters are eventu-
ally estimated in an iterative fashion (by alternating one phase of image reconstruction with
one phase of parameter estimation), the moment estimator provides a good starting point.

The simplicity of the estimation process makes possible the repetition of its application
with relatively little effort. Thus the image can be divided into subimages and the parameters
can be estimated locally, by applying the method of section 5.3 to each subimage separately.
Local parameter estimation is sometimes desirable because the values of the parameters may
be changing spatially with the location of the image (see section 7.2). Also, local parameter
estimation could potentially be used as a diagnostic tool for the adequacy of the adopted model.

Another advantage of the moment estimator is that the obtained estimates are given by
explicit formulas. Unlike most estimates obtained by numerical optimization of some objec-
tive function, the estimates obtained from (5.6.1) can be written analytically as functions of
$\hat{\tau}, \hat{\tau}_{11}, \hat{\tau}_{11}$. This makes easier the manipulation and the examination of the properties of the
derived estimates. For instance, the system (5.6.1) can be viewed as a mapping from the
$(\hat{\tau}, \hat{\tau}_{11}, \hat{\tau}_{11})$ space to the $(\hat{A}, \hat{B})$ space, and an approximate estimate for the variance of $\hat{A}$ and
$\hat{B}$ can be derived with the delta method (once the variances of $\hat{\tau}, \hat{\tau}_{11}$ and $\hat{\tau}_{11}$ are computed; see the next chapter and section 7.3).

The moment estimator has been presented in the previous sections for the Markov model
(3.3.3) but exactly the same approach can be used for any Markov mesh model with the neighborhood \( N_t = \{(t_x - 1, t_y), (t_x, t_y - 1)\} \). The details of the implementation (i.e., the exact form of the equations which comprise a system for the parameters) will vary with the particular form of the mesh model, but the technique remains the same. Furthermore, the approach can be used with any Markov model (symmetric or not): the parameters can be estimated from a number of moment equations, by first estimating the probabilities involved and then solving the system. These probabilities, however, become more complex (i.e., correspond to higher-order statistics) as the size of the neighborhood grows larger. For instance, in a Markov model with a first-order symmetric neighborhood (see chapter 3) the required joint probabilities would be of the form \( p\{I(t_x - 1, t_y), I(t_x, t_y - 1), I(t_x + 1, t_y), I(t_x, t_y + 1)\} \) instead of \( p\{I(t_x - 1, t_y), I(t_x, t_y - 1)\} \).

The most important liability of the moment estimator is its limitation to binary images. At the moment it is not at all clear how to extend the method to images which can assume more than two values. Another potential problem may be that the obtained estimates are inappropriate (e.g., \( \hat{B} < 0 \)). This problem is common to all moment estimates, and its solution usually varies according to the application at hand. Finally, the method developed in sections 5.3–5.7 works nicely for Markov mesh models (like (5.2.1)) but gets significantly more complicated for really symmetric models (like (3.2.3)). A similar idea has been applied to the Ising model [FrPi88], but the results are somewhat different and the estimates cannot be obtained in closed form.
Chapter 6

Estimation of Variance

This chapter suggests a method for estimating the variance of the estimator of the previous chapter. The method is sufficiently simple to be applicable to the estimator of any quantity of interest in an image. The next section motivates the estimation of the variance of a parameter estimate. Section 6.2 describes the variance estimation technique as it applies to the estimator of chapter 5. Section 6.3 extends the idea to an arbitrary estimator and presents a heuristic algorithm for its implementation. Since the variance estimation technique appears to be new, section 6.4 identifies some possibilities for future research in this direction.

This chapter is a digression from the previous two chapters and from the Markov random field framework in general. Especially sections 6.3 and 6.4 are not related to Markov random fields at all and can be read almost independently of chapters 3, 4, 5 and 7. Section 6.2 pertains to the variance of the estimator of chapter 5 but even in this case the relationship (between section 6.2 and chapter 5) is tangential: the details of chapter 5 are not necessary for understanding section 6.2.

6.1. Motivation.

In practice it is the estimate of a parameter that is important and not the estimate of its variance. In most cases the objective of an image processing effort is the reconstruction of an image, and it is the estimate of a parameter that is used to drive a reconstruction algorithm. As a result, the variance of the estimates of image parameters has not attracted much attention in the past. For example, the ICM algorithm [Besa86a] only needs a reasonable value for $\beta$ and not a measure of accuracy for this value; notice that it could be argued that a "right" $\beta$ value does not even exist (see section 3.1).

It is legitimate then to wonder why one should bother at all with the variance of a param-
Section 6.2: The Variance of the Moment Estimator

The answer is twofold. First, on a purely statistical level, an estimate carries little, if any, information without a corresponding measure of accuracy. Second, on a practical level, an estimate of the variance can provide valuable information about the parameter itself. For instance, an extremely high variance for \( \hat{\beta} \) may suggest that it is more prudent to adopt different values for \( \beta \) in different parts of an image. This information is, and should be regarded as, secondary when compared to the actual estimate \( \hat{\beta} \), but it can be used as a diagnostic tool or as an aid to improve model selection. Such an application is rather unlikely for Markovian models (because they are only convenient vehicles for reconstruction algorithms, not realistic models for the "true" image; see section 3.1), but it can be valuable for non-Markovian models (e.g., for local averaging with neighborhoods of varying size). Therefore, there are valid reasons for which the estimation of the variance is a worthwhile task.

6.2. The Variance of the Moment Estimator.

For the sake of simplicity, the discussion in this chapter focuses only on one image parameter. This simplification makes the presentation easier and, as it will become obvious, does not result in any loss of generality. Thus the remainder of this section concentrates on the most important of the two parameters of the model (5.2.1) namely, the spatial interaction parameter \( \beta \).

As mentioned in section 5.8, one of the advantages of the estimator of chapter 5 is that it can be applied locally to parts of an image. That is, the parameter \( \beta \) of the model (5.2.1) can be estimated from the data of any subimage \( \{O(t) : \forall t \in D_{\text{sub}}\} \), where \( D_{\text{sub}} \subseteq D \). The only requirement is that \( D_{\text{sub}} \) defines a contiguous subset of \( D \) (e.g., strips of rows or columns, one corner of the image, some block of pixels, etc.). This feature can be exploited in order to estimate the variance of the estimate \( \hat{\beta} \).

Consider a division of the pixel lattice \( D \) into nonoverlapping strips of rows of pixels which cover the whole lattice. If the lattice is an \( n \times m \) rectangular array of pixels, such a division consists of \( s \) consecutive strips of \( \frac{n}{s} \times m \) pixels, where \( s \) is a divisor of \( n \). All the strips contain the same number of rows, \( r = n/s \). This number is a critical parameter of the variance estimation procedure and its choice is discussed below. It is obvious that the choice of \( s \) determines \( r \) and vice-versa. Therefore, \( s \) and \( r \) are equivalent parameters. For the time
being, \( r \) is considered to be a constant specified by the fixed number of strips \( s \). The division of the lattice can schematically be illustrated as follows:

\[
D_1 = \left\{ \begin{array}{l}
\text{Row 1} \\
\text{Row 2} \\
\vdots \\
\text{Row } r \\
\text{Row } r + 1 \\
\text{Row } r + 2 \\
\vdots \\
\text{Row } 2r \\
\vdots \\
\text{Row } (s - 1)r + 1 \\
\text{Row } (s - 1)r + 2 \\
\vdots \\
\text{Row } n \\
\end{array} \right.
\]

\[
D_2 = \left\{ \begin{array}{l}
\text{Row 1} \\
\text{Row 2} \\
\vdots \\
\text{Row } r \\
\text{Row } r + 1 \\
\text{Row } r + 2 \\
\vdots \\
\text{Row } 2r \\
\vdots \\
\text{Row } (s - 1)r + 1 \\
\text{Row } (s - 1)r + 2 \\
\vdots \\
\text{Row } n \\
\end{array} \right.
\]

Looking at the pixel lattice from top to bottom, the first strip, \( D_1 \), contains the first \( r \) rows of pixels, the second strip, \( D_2 \), contains the second \( r \) rows of pixels, etc.

Once the lattice is divided into the strips \( D_1, \ldots, D_s \), the estimation procedure of chapter 5 can be applied to each strip separately to yield estimates of the parameter \( \beta \). That is, for each strip \( D_i \) the data \( \{O(t) : \forall t \in D_i\} \) is used to obtain an estimate \( \hat{\beta}_i \), according to equation (5.6.5) (\( \hat{\beta} = \log \hat{B} \)). There are exactly \( s \) \( \hat{\beta}_i \)'s (i.e., as many as there are strips in the division of \( D \)):

\[
\begin{array}{c}
D_1 \xrightarrow{(5.6.5)} \hat{\beta}_1 \\
D_2 \xrightarrow{(5.6.5)} \hat{\beta}_2 \\
\vdots \\
D_s \xrightarrow{(5.6.5)} \hat{\beta}_s \\
\end{array}
\]

The idea behind estimating the variance of \( \hat{\beta} \) is to approximate it by the variance of \( \hat{\beta}_1, \ldots, \hat{\beta}_s \).

The estimates \( \hat{\beta}_i \) are not independent and, therefore, the sample variance

\[
\frac{1}{s - 1} \sum_{i=1}^{s} (\hat{\beta}_i - \bar{\beta})
\]

(6.2.1)
is not an appropriate estimate of the variance of $\hat{\beta}_1, \ldots, \hat{\beta}_s$. Since the strips $D_i$ and $D_{i+1}$ are adjacent to each other, the corresponding estimates $\hat{\beta}_i$ and $\hat{\beta}_{i+1}$ are correlated. The degree to which these estimates are correlated depends on the image $I(t)$ and the nature of the strips $D_i$ and $D_{i+1}$ (e.g., on the number of rows $r$). In particular, the correlation between $\hat{\beta}_i$ and $\hat{\beta}_{i+1}$ grows smaller as the size of the strips $D_i$ and $D_{i+1}$ grows larger and vice-versa. This is analogous to how the spatial correlation between image values depends on the size of the pixels: the smaller the pixels the larger their spatial correlation is. In addition, the correlation between the $\hat{\beta}_i$'s decreases with the distance of the corresponding strips $D_i$. In other words, the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$ grows stronger as the distance between $D_i$ and $D_j$ becomes smaller. This is again analogous to how the spatial correlation between image values decreases with the distance of the corresponding pixels: the farther apart two pixels are the smaller the correlation between their image values is.

The last property resembles the behavior of a time series and suggests a method for estimating the variance of $\hat{\beta}$. The $\hat{\beta}_i$'s can be modeled as a stationary time series, where the index $i$ plays the role of the time index. Then, if the overall estimate $\hat{\beta}$ is viewed roughly as an average of the strip estimates $\hat{\beta}_i$, its variance $\text{var}(\hat{\beta})$ can be approximated by the variance of the average of the fake time series $\{\hat{\beta}_1, \ldots, \hat{\beta}_s\}$; the latter is obtained from the variance of the $\hat{\beta}_i$'s which is computed by standard time series techniques. This approximation is obviously going to be less accurate when $\hat{\beta}$ is a non-linear function of the $\hat{\beta}_i$'s. Nevertheless, even then it can prove useful because it can be combined with a simple propagation of errors argument to yield variance estimates (see section 7.3).

Naturally, the computation of the variance of $\{\hat{\beta}_1, \ldots, \hat{\beta}_s\}$ requires that some model is imposed on this series (e.g., an autoregressive model, a moving average model, etc.). Different models may result in different estimates of variance and a question arises as to how to select the appropriate model (or, equivalently, the appropriate variance estimate). Unfortunately, it is well known that the selection of a time series model is generally not an easy problem. Moreover, the selection of a model for $\{\hat{\beta}_1, \ldots, \hat{\beta}_s\}$ can be considered as an even harder case than the general problem because of the artificial nature of the model. On the other hand, this same artificial nature eases the need for a strictly formal model selection process and lends itself to convenient, though ad hoc, choices.
One criterion for selecting a time series model is the size of the resulting variance. A small variance is almost always preferable to a large one and, therefore, one may be inclined to select a model that yields a small estimate of variance. However, such a model selection process may be criticized as unreasonable, because the goal is to estimate accurately the variance. Adjusting the artificial model for \( \{\hat{\beta}_1, \ldots, \hat{\beta}_s\} \) so that it produces a small estimate of variance is defeating the purpose.

The criterion adopted in this report is simplicity. That is, a very simple autoregressive process of order 1 is used to capture the variability in the \( \hat{\beta}'s \):

\[
\hat{\beta}_i - \mu_{\hat{\beta}} = \omega (\hat{\beta}_{i-1} - \mu_{\hat{\beta}}) + e_i, \quad \text{with} \ E(e_i) = 0 \ \text{and} \ \text{var}(e_i) = \sigma_e^2. \tag{6.2.2}
\]

Here \( \mu_{\hat{\beta}} \) is the mean of the process for the \( \hat{\beta} \)'s, \( \omega \) is the autoregression parameter and the \( e \)'s are the pure noise process. The \( e \)'s are assumed to be pointwise independent: \( e_i \) is independent from \( e_j \), for all \( i \) and \( j \). This autoregression is probably not a realistic model for describing all of the spatial structure contained in the \( \hat{\beta} \)'s. Nevertheless, from a practical standpoint it is deemed adequate for most applications for reasons which are explained below. Some numerical results regarding model (6.2.2) are presented in chapter 8.

The main advantage of the autoregressive model (6.2.2) is that its properties are well known [Chat84]. Provided \( |\omega| < 1 \), the variance of the \( \hat{\beta}_i \)'s is given by the standard formula

\[
\sigma_{\hat{\beta}_i}^2 = \text{var}(\hat{\beta}_i) = \frac{\sigma_e^2}{1 - \omega^2}, \tag{6.2.3}
\]

and the process \( \hat{\beta}_i \) is second-order stationary. Its autocovariance function \( \gamma \) of lag \( k \) is given by

\[
\gamma(k) = E(\hat{\beta}_i - \mu_{\hat{\beta}})(\hat{\beta}_{i+k} - \mu_{\hat{\beta}}) = \omega^k \sigma_{\hat{\beta}_i}^2, \quad \text{for} \ k > 0. \tag{6.2.4}
\]

The autocorrelation function \( \rho \) of lag \( k \) is given by

\[
\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\gamma(k)}{\sigma_{\hat{\beta}_i}^2} = \omega^k.
\]

When \( k < 0 \), \( \gamma(k) = \gamma(-k) \) and \( \rho(k) = \omega^{-k} \). The condition \( |\omega| < 1 \) is necessary to ensure that the \( \hat{\beta}_i \)'s form a proper stationary process.

To use the variance \( \text{var}(\hat{\beta}_i) \) as an estimate for the variance of \( \hat{\beta} \), one must first estimate
the variance of the $e$'s, $\sigma_e^2$, and the autoregression parameter $\omega$. The parameters $\mu_\hat{\beta}$, $\sigma_e^2$ and $\omega$ can be estimated either from the moments of the process $\hat{\beta}_t$ (via its autocovariance function (6.2.4)) or from least squares. The usual estimates are the least squares estimates which are obtained from the equations

$$
\hat{\mu}_\hat{\beta} = \frac{1}{1 - \hat{\omega}} \left[ \frac{1}{s - 1} \sum_{i=2}^{s} \hat{\beta}_i - \hat{\omega} \frac{1}{s - 1} \sum_{i=1}^{s-1} \hat{\beta}_i \right],
$$

$$
\hat{\omega} = \frac{\sum_{i=1}^{s-1} (\hat{\beta}_i - \hat{\mu}_\hat{\beta}) (\hat{\beta}_{i+1} - \hat{\mu}_\hat{\beta})}{\sum_{i=1}^{s-1} (\hat{\beta}_i - \hat{\mu}_\hat{\beta})^2},
$$

$$
\hat{\sigma}_e^2 = \frac{1}{s - 1} \sum_{i=1}^{s-1} (\hat{\beta}_i - \hat{\omega} \hat{\beta}_{i-1})^2.
$$

(6.2.5)

The approximation $\frac{1}{s-1} \sum_{i=1}^{s-1} \hat{\beta}_i \approx \frac{1}{s-1} \sum_{i=2}^{s} \hat{\beta}_i \approx \frac{1}{s} \sum_{i=1}^{s} \hat{\beta}_i \equiv \beta$ reduces the least squares equations (6.2.5) into the approximate versions

$$
\hat{\mu}_\hat{\beta} \approx \frac{1}{s} \sum_{i=1}^{s} \hat{\beta}_i = \beta,
$$

$$
\hat{\omega} \approx \frac{\sum_{i=1}^{s-1} (\hat{\beta}_i - \beta) (\hat{\beta}_{i+1} - \beta)}{\sum_{i=1}^{s-1} (\hat{\beta}_i - \beta)^2},
$$

$$
\hat{\sigma}_e^2 = \frac{1}{s - 1} \sum_{i=1}^{s-1} (\hat{\beta}_i - \hat{\omega} \hat{\beta}_{i-1})^2.
$$

(6.2.6)

It is common practice to use the estimates $\hat{\mu}_\hat{\beta}$, $\hat{\omega}$ and $\hat{\sigma}_e^2$ which are obtained from the approximate equations (6.2.6) rather than the ones obtained from the exact equations (6.2.5). Substituting these estimates into (6.2.3) yields the estimate for the variance of $\hat{\beta}_t$:

$$
\text{var}(\hat{\beta}_t) = \hat{\sigma}_e^2 \hat{\beta}_t = \frac{\hat{\sigma}_e^2}{1 - \hat{\omega}^2}.
$$

(6.2.7)

The variance of the estimate $\hat{\beta}$ is the variance $\beta$ of the average of the $\hat{\beta}_t$'s and can be computed explicitly as a function of the autoregression parameter $\omega$:

$$
\text{var}(\hat{\beta}) = \text{var}(\beta) = \frac{1}{s^2} \left[ n \cdot \text{var}(\beta) + 2(s - 1)\hat{\omega} + 2(s - 2)\hat{\omega}^2 + 2(s - 3)\hat{\omega}^3 + \ldots + 2\hat{\omega}^{s-1} \right].
$$

A more useful formula for the variance of $\hat{\beta}$ is the approximation (see page 568 in BoJe76)

$$
\text{var}(\hat{\beta}) = \text{var}(\hat{\beta}_t) \approx \frac{\hat{\sigma}_e^2}{s(1 - \hat{\omega})^2}.
$$

(6.2.8)
In light of equation (6.2.6), the cost associated with the computation of this variance estimate is essentially the same as the cost of running the regression of \( \{ \hat{\beta}_i - \overline{\beta}, 2 \leq i \leq s \} \) on \( \{ \hat{\beta}_{i-1} - \overline{\beta}, 1 \leq i \leq s - 1 \} \).

The autoregressive model (6.2.2) is clearly an arbitrary choice, but it can be defended on the grounds of convenience. It is rather helpful to have a model whose properties are well understood and readily available, especially if the estimate of the variance of \( \hat{\beta} \) is to be used as a diagnostic tool. Other factors which contribute to the defense of a simple model are (a) the heuristic nature of the variance estimation procedure (i.e., the fake time series model for \( \{ \hat{\beta}_1, \ldots, \hat{\beta}_s \} \)), (b) the inherent artificiality of the parameter \( \beta \) and, more generally, of the Markov random field model (5.2.1), and (c) the fact that the variance of \( \hat{\beta} \) is indeed auxiliary information as far as the main objective (i.e., image reconstruction) is concerned. The underlying principle is to avoid expending too much energy in an effort occulted by many levels of indirection.

A problem which is related to the problem of choosing a model for \( \{ \hat{\beta}_1, \ldots, \hat{\beta}_s \} \) is how to select appropriately the size of the strips \( r \). Since the division of \( D \) into the strips \( D_1, \ldots, D_s \) is arbitrary, the strips can vary in size. The two extreme cases are \( r = 1 \) and \( r = n \). When \( r = 1 \), each row of pixels comprises one strip and there exist \( n \) strips in total (i.e., as many as there are rows in the lattice \( D \)). When \( r = n \), all the rows together comprise one strip and there exists only 1 strip; this strip is identical to the whole image.

These extreme cases are impractical for rather apparent reasons, but they help to illustrate a trade-off associated with the choice of \( r \). When \( r = 1 \) (and \( s = n \)) there are \( n \) estimates \( \hat{\beta}_i \). That is, the sequence \( \{ \hat{\beta}_1, \ldots, \hat{\beta}_n \} \) contains the maximum number of possible points. This is the ideal situation for estimating the variance of \( \hat{\beta} \), precisely because there are as much data (i.e., \( \hat{\beta}_i \)'s) as possible. However, the individual estimates \( \hat{\beta}_i \) are not reliable because each of them is based on a single row of pixels. Indeed, on a single row the estimation of \( \beta \) from (5.6.5) becomes one-dimensional. The model (5.2.1) does not even have a chance to show its spatial behavior and the estimates \( \hat{\beta}_i \) are almost meaningless. Furthermore, a single row of pixels is likely to contain too few pixels and thus result in unstable estimates \( \hat{\beta}_i \). Consequently, the variance of \( \{ \hat{\beta}_1, \ldots, \hat{\beta}_s \} \) is estimated with greater precision but the obtained estimate of
variance bears less relevance to a meaningful estimate $\hat{\beta}$. This compensation effect between the number of available strips and the goodness of the corresponding estimates is evident in the approximate variance formula (6.2.8).

When $r = n$ (and $s = 1$) there is only one estimate $\hat{\beta}_1$. The sequence $\{\hat{\beta}_1, \ldots, \hat{\beta}_s\}$ contains one point and estimation of the variance becomes impossible. The single estimate $\hat{\beta}_1$ is the estimate $\hat{\beta}$ obtained from the whole image and, therefore, it may be the most pertinent estimate possible. On the other hand, a single estimate for $\beta$ may not be ideal if, in reality, different $\beta$'s are better suited for different parts of the image.

One difficulty which is involved in the choice of $r$ is the trade-off between "better estimates $\hat{\beta}_i$" and "better estimate of the variance of $\hat{\beta}$". Another difficulty is a second trade-off between independent $\hat{\beta}_i$'s and spatially correlated $\hat{\beta}_i$'s. The estimates $\hat{\beta}_i$ become more and more independent as $r$ grows larger. The reason is that the coarser the scale on which quantities of an image are measured the less spatially correlated these quantities are. As a result, if the strips $D_i$ contain a sufficiently large number of rows the estimates $\hat{\beta}_i$ will be effectively independent. In this case, there is no need for a model like (6.2.2) or any other model which attempts to account for the spatial correlation (due to proximity and scale of measurement) between the $\hat{\beta}_i$'s. The sample variance (6.2.1) gives an estimate for the variance of the $\hat{\beta}_i$'s and, in turn, for the variance of $\hat{\beta}$.

The degree to which these trade-off's ("better estimates $\hat{\beta}_i$" versus "better estimate $\sigma^2_\beta$" and "correlated $\hat{\beta}_i$'s" versus "independent $\hat{\beta}_i$'s") affect an optimal choice for $r$ depends on the image at hand. For example, the size of $r$ which yields effectively uncorrelated $\hat{\beta}_i$'s is determined by (a) the spatial structure of the image, (b) how faithfully a Markovian model like (5.2.1) describes this structure, and (c) how sensitive the estimation procedure is to the spatial changes in the image.

The dependence of the selection of $r$ on the particular image makes it difficult to develop an algorithm which will select the most appropriate $r$. This problem will be the subject of future research (see section 6.4). A heuristic guideline is to take $r$ close to $\sqrt{n}$, because this choice appears to strike an acceptable compromise in the trade-off between meaningful estimates of $\beta$ and a reasonable estimate of the variance of $\hat{\beta}$ (the other trade-off between "independence"
and "correlation" is less important. The \( \sqrt{n} \) rule is adopted in the general algorithm of the next section; limited numerical experience indicates that it does not lead to disastrous results (see chapter 8).

### 6.3. A Heuristic Algorithm.

The variance estimation technique of the previous section does not rely on any special properties of the estimation procedure of chapter 5. The basic idea is to artificially introduce replication in a situation where there isn't any. The same idea can readily be applied to any estimation procedure: simply replace the estimation procedure of chapter 5 by the desired estimation procedure in the discussion of the previous section. All the details remain the same, except that a different estimator is used on each strip \( D_i \). Furthermore, there is nothing magical about the parameter \( \beta \) or its Markovian model either. The parameter of interest can be any parameter \( \theta \) of an arbitrary image model. In other words, the method of section 6.2 is independent from both the parameter \( \beta \) and the moment estimator of chapter 5. This independence makes the variance estimation technique a general tool which can be used in a number of applications (see also section 7.3).

Consider a spatial model \( m(I) \) for the image \( I(t) \). Assume that the model \( m(I) \) is a parametric model and that \( \theta \) is its critical parameter. The model \( m(I) \) could be, but is not necessarily, a Markovian model like \( p(I) \) in (3.2.3) or (5.2.1); in fact, it is denoted by \( m(I) \) to emphasize this distinction. For example, if local averaging is going to be used as a smoothing operation, \( \theta \) may be the size of the averaging neighborhoods. Suppose that there exists some estimation procedure \( E(\theta) \) which produces an estimate \( \hat{\theta} \) for \( \theta \) from the data of the image \( \{O(t)\} \). The estimation procedure \( E(\theta) \) can be thought of as a black-box mechanism and its details are not important. The only requirement is that \( E(\theta) \) can be applied to parts of an image in order to produce local estimates of \( \theta \). This is not a serious constraint because in practice any subimage can be regarded as a separate image.

Let \( \hat{\theta} \) be the estimate of interest for the parameter \( \theta \); this estimate is obtained through \( E(\theta) \). The variance of \( \hat{\theta} \) can be estimated according to the method described in the previous section. It is useful to formulate the variance estimation technique in a procedural fashion. The resulting algorithm is called VETIPE (Variance Estimation Technique for Image Parameter
Estimates). For an $n \times m$ rectangular lattice, VETIPE applies to $\hat{\theta}$ as follows:

- Let $r$ be the integer divisor of $n$ which is closest to $\sqrt{n}$ and set $s = n/r$.

- Divide the lattice $D$ into $s$ nonoverlapping strips $D_1, D_2, \ldots, D_i, \ldots, D_s$ of rows of pixels. The dimensions of each strip are $r \times m$.

- Obtain $s$ estimates $\hat{\theta}_i$ by applying $E(\theta)$ separately to $\{O(t), t \in D_i\}$, for $1 \leq i \leq s$.

- Select some simple time series model for the sequence $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_i, \ldots, \hat{\theta}_s\}$.

- Estimate the parameters of the time series model and fit the model to the $\hat{\theta}_i$'s.

- Compute the desired variance estimate $\hat{\sigma}_\theta^2$ from the parameters of the time series model.

This description of VETIPE is more of a schema for an algorithm than an algorithm in itself. The algorithm is outlined at a high-level of abstraction without laying out the specifics of the various steps. The advantage, of course, is that abstraction offers a higher degree of generality.

In VETIPE step 3 is the only step which is affected by the nature of the parameter $\theta$ and the estimation procedure $E(\theta)$. The implementation of steps 5 and 6 is determined entirely by step 4 (i.e., which model is chosen for $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s\}$). One possibility for step 4 is to adopt the simple model (6.2.2), in which case steps 5 and 6 are described in the previous section. In general, however, the last three steps of the algorithm are application specific.

It should also be noted that the first step follows the $\sqrt{n}$ heuristic in choosing the number of rows $r$ for the strips $D_i$. The trade-off's associated with the choice of $r$ are identified in the previous section for the parameter $\beta$ of the Markov random field (5.2.1). They pertain to the general case of an arbitrary parameter $\theta$ in exactly the same way. That is, the estimates $\hat{\theta}_i$ are more reliable and less correlated when $r$ is larger. At the same time, the variance of $\hat{\theta}$ is estimated more accurately when $r$ is smaller. It is difficult to say what the optimal value of $r$ is. How appropriate a particular value is depends on the image $I(t)$, the model $m(I)$, the parameter $\theta$ and the procedure $E(\theta)$. More significantly, it depends on the purpose for which the estimates $\hat{\theta}$ and $\hat{\sigma}_\theta^2$ are going to be used.

The variance estimation technique generates replicate estimates of the parameter and uses them to measure the variability of the overall estimate. To measure variability, VETIPE
creates and exploits a correspondence between the time index of a time series and the spatial index of a spatial data series. Creating such a correspondence is a device that has been used before in image processing and in the analysis of spatial data in general. The reason is that some properties of spatial data are often analogous to certain properties of time dependent data. This fact is evidenced in image processing by the early attempts to use Markov chains to describe the spatial structure of images; these attempts eventually culminated into the formalization of Markov random fields as image models.

Setting up a correspondence between a time and a space index presents an intrinsic difficulty in matching the dimensions of the two indices. A time index is one-dimensional and, therefore, provides a natural ordering for the observations. A space index is at least two-dimensional (for most images) and, therefore, provides only the notion of proximity not an ordering. The method presented in this chapter resolves this difficulty by dividing the lattice into strips of rows and obtaining one parameter estimate per strip. Thus, it effectively hides the columns of the lattice and reduces the dimensionality of the space index. The same trick of dividing the image into strips of rows appears in a different context in DeEC84.

VETIPE can also be used when there are two parameters of interest (say, $\theta$ and $\phi$) and it is the covariance of their estimates, $\text{cov}(\hat{\theta}, \hat{\phi})$, that is being sought. For instance, $\theta$ may be $\pi$ and $\phi$ may be $\pi_{11}$. The use of VETIPE remains the same: (a) the image is divided into strips of rows, (b) estimates, $\hat{\theta}_i, \hat{\phi}_i$, are obtained for the parameters from each strip $i$, (c) a time series model is fitted to the obtained estimates and (d) a covariance estimate $\text{cov}(\hat{\theta}, \hat{\phi})$ is computed from the time series model. This is illustrated in chapter 8 for the covariance of the estimates $\hat{\pi}, \hat{\pi}_{11}$.

6.4. Future Research.

The previous two section raise a number of interesting questions which suggest some directions for future research. For instance, an issue of practical importance is the development of a procedure which will allow the selection of the size of the strips, $r$, in a systematic way.

One of the potential uses of the variance estimate $\hat{\sigma}_\theta^2$ is as a diagnostic tool namely, to detect an inordinate amount of spatial variability in the estimate $\hat{\theta}$. In this capacity the estimate $\hat{\sigma}_\theta^2$ wastes considerable information because it condenses all the $\hat{\theta}_i$'s into a single
number. An alternative approach is to use the bootstrap in an effort to utilize the data \( \{ \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s \} \) more efficiently. This is an appealing approach because both the bootstrap and the variance estimation method rely on the same idea of creating artificial replication.

The bootstrap cannot be applied directly to the \( \hat{\theta}_i \)'s because they are not independent. However, it can be applied to the residuals \( \{ e_1, e_2, \ldots, e_s \} \) which result from fitting a model to \( \{ \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s \} \) (see section 5.7 in Efro82). Suppose that a model is selected for the \( \hat{\theta}_i \)'s in step 4 of VETIPE. Fitting this model produces residuals \( e_1, e_2, \ldots, e_s \) which are, by assumption, independent. The residuals are bootstrapped to yield \( B \) bootstrap samples \( \{ e_1^{(1)}, e_2^{(1)}, \ldots, e_s^{(1)} \}, \{ e_1^{(2)}, e_2^{(2)}, \ldots, e_s^{(2)} \}, \ldots, \{ e_1^{(B)}, e_2^{(B)}, \ldots, e_s^{(B)} \} \). Each bootstrap sample of residuals \( \{ e_1^{(b)}, e_2^{(b)}, \ldots, e_s^{(b)} \} \) gives rise to a bootstrap sample of \( \hat{\theta} \)'s through the selected model:

\[
\begin{align*}
D_1 & \xrightarrow{\mathcal{E}(\hat{\theta})} \hat{\theta}_1 & e_1 & e_1^{(1)}, e_1^{(2)}, \ldots, e_1^{(B)} & \hat{\theta}_1^{(1)}, \hat{\theta}_1^{(2)}, \ldots, \hat{\theta}_1^{(B)} \\
\vdots & & \vdots & \vdots & \vdots \\
D_i & \xrightarrow{\mathcal{E}(\hat{\theta})} \hat{\theta}_i & e_i & e_i^{(1)}, e_i^{(2)}, \ldots, e_i^{(B)} & \hat{\theta}_i^{(1)}, \hat{\theta}_i^{(2)}, \ldots, \hat{\theta}_i^{(B)} \\
\vdots & & \vdots & \vdots & \vdots \\
D_s & \xrightarrow{\mathcal{E}(\hat{\theta})} \hat{\theta}_s & e_s & e_s^{(1)}, e_s^{(2)}, \ldots, e_s^{(B)} & \hat{\theta}_s^{(1)}, \hat{\theta}_s^{(2)}, \ldots, \hat{\theta}_s^{(B)}
\end{align*}
\]

This operation results in \( B \) bootstrap samples of estimates for the parameter \( \theta \):

\[
\{ \theta_1^{(b)}, \theta_2^{(b)}, \ldots, \theta_s^{(b)} \}, \quad \text{for } \ 1 \leq b \leq B.
\]

Each sample contains \( s \) estimates which correspond to the \( s \) strips \( D_i \).

Once a bootstrap sample \( \{ \theta_1^{(b)}, \theta_2^{(b)}, \ldots, \theta_s^{(b)} \} \) is constructed, it is easy to obtain an estimate \( \hat{\mu}_\theta^{(b)} \) of its mean. This is done according to the model which has been chosen for the \( \hat{\theta}_i \)'s (see, for example, equations (6.2.6)). Regarding the \( \hat{\mu}_\theta^{(b)} \)'s as estimates of \( \hat{\theta} \), a bootstrap confidence interval for \( \hat{\theta} \) can be constructed from \( \{ \hat{\mu}_\theta^{(1)}, \hat{\mu}_\theta^{(2)}, \ldots, \hat{\mu}_\theta^{(B)} \} \). A number of methods exist for constructing bootstrap confidence intervals; see Efro82, Efro87, EfTi86, Hjor85c and Hall88a.

The confidence interval for \( \hat{\theta} \) can serve as a substitute measure of variability for \( \hat{\theta}^2 \). For instance, if \( \hat{\theta} \) falls outside this interval there may be reason for concern regarding the value \( \hat{\theta} \). Unfortunately, such an interval can be interpreted only in one direction. The fact that the interval contains the value \( \hat{\theta} \) does not necessarily imply that this value is an appropriate one.
Chapter 6: Estimation of Variance

This application of the bootstrap is conceptually analogous to the one suggested by Hall88b (see Example 26 in chapter 2). Hall prefers to avoid dealing with the correlations between the $\hat{\theta}_i$'s, rather than resort to a time series model in order to account for them. The idea is again to obtain estimates $\hat{\theta}_i$ for $\theta$ from different parts of the image and use these estimates to assess the variability of $\theta$ (i.e., construct bootstrap confidence intervals). The difference is that the obtained $\hat{\theta}_i$'s are independent of each other (unlike the $\hat{\theta}_i$'s from the strips $D_i$) and, therefore, can be bootstrapped directly. To achieve independence between the $\hat{\theta}_i$'s, the image is divided into groups of subimages in such a way that all the members of a group are sufficiently far apart to be considered independent. This approach is conceptually similar to VETIPE in that both methods rely on artificial replication. On the other hand, the mechanics of Hall's approach (namely, the method for creating independent replicates) resemble more closely the coding scheme of Besa74 (see section 3.4 and Example 24 in chapter 2) than VETIPE's division of the lattice into strips of rows.

Another possibility for future research stems from the observation that in VETIPE the strips of rows can just as easily be replaced by strips of columns. The lattice can be divided into strips of columns (instead of rows) of pixels, these strips can produce estimates $\hat{\theta}_i$ for the parameter $\theta$, and the variance of $\theta$ can be estimated – as before – from the $\hat{\theta}_i$'s. In other words, there is a duality between rows and columns. It is, therefore, conceivable that one may be able to take advantage of this duality in order to extract more information out of the estimate $\hat{\theta}_i$.

Consider two divisions of the lattice $D$, the first into strips of rows and the second into strips of columns. The strips of rows $D_1^{\text{Row}}, \ldots, D_i^{\text{Row}}, \ldots, D_s^{\text{Row}}$ yield parameter estimates $\hat{\theta}_1^{\text{Row}}, \ldots, \hat{\theta}_i^{\text{Row}}, \ldots, \hat{\theta}_s^{\text{Row}}$, and the strips of columns $D_1^{\text{Col}}, \ldots, D_j^{\text{Col}}, \ldots, D_s^{\text{Col}}$ yield estimates $\hat{\theta}_1^{\text{Col}}, \ldots, \hat{\theta}_j^{\text{Col}}, \ldots, \hat{\theta}_s^{\text{Col}}$. Clearly, there is some redundancy in the information contained in the $\hat{\theta}_1^{\text{Row}}$'s and the $\hat{\theta}_j^{\text{Col}}$'s. This redundancy could provide some additional information on the variability of $\hat{\theta}$. For example, a "comparison" between $\{\hat{\theta}_1^{\text{Row}}, \ldots, \hat{\theta}_s^{\text{Row}}\}$ and $\{\hat{\theta}_1^{\text{Col}}, \ldots, \hat{\theta}_s^{\text{Col}}\}$ might reveal that $\hat{\theta}$ is changing more rapidly in one direction than in the other. Or, it might indicate whether the spatial trends in the $\hat{\theta}_i^{\text{Row}}$'s and the $\hat{\theta}_j^{\text{Col}}$'s are similar. This possibility can be investigated only if (a) the comparison of $\{\hat{\theta}_1^{\text{Row}}, \ldots, \hat{\theta}_s^{\text{Row}}\}$ to $\{\hat{\theta}_1^{\text{Col}}, \ldots, \hat{\theta}_s^{\text{Col}}\}$ is a precisely defined notion and (b) a method is developed for realizing it. An example of such an analysis is presented in chapter 8 (section 8.3).
Chapter 7

Extensions

This chapter presents two extensions to the moment estimator of chapter 5 and one extension to the variance estimation technique of chapter 6. The next section extends the estimation method of section 5.3 to images with multivariate observations and section 7.2 outlines some possibilities for local applications of the moment estimator. Section 7.3 describes a variation to VETIPE. This method actually uses VETIPE and is a useful alternative in situations where the overall parameter estimate is not a straight average of the individual strip estimates.

7.1. Multivariate Images.

Chapter 5 discusses exclusively the case of univariate observations $O(t)$, but the moment estimator of $\alpha$ and $\beta$ (section 5.3) can readily be extended to images with multivariate observations. Let $O(t) \in \mathbb{R}^p$ and write $O(t) = (O_1(t), \ldots, O_p(t))^T$, with $O_i(t) \in \mathbb{R}$ for $1 \leq i \leq p$. The $I(t)$'s are again 0 or 1, and the conditional distribution of $O(t)$ is described – as before – by:

$$O(t) = (1 - I(t))[\mu_0 + \epsilon(t)\Sigma_0] + I(t)[\mu_1 + \epsilon(t)\Sigma_1],$$

where $\epsilon(t)$ is the multivariate noise process. The process $\epsilon(t)$ has mean 0 and covariance equal to the $p \times p$ identity matrix $I$. The means $\mu_i$ and the covariance matrices $\Sigma_i$ of $O(t)$ are known.

The parameters $\alpha$ and $\beta$ of the Markov random field (5.2.1) can still be estimated from the system (5.6.1). The only difference with chapter 5 is that the estimates $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ of $\pi, \pi_{11}, \pi_{11}$ must now incorporate the $p$ components of the observation vectors.

An estimate $\hat{\pi}$ for the marginal probability $\pi$ can be obtained with the method of moments of section 5.4. When the observations are multivariate the moment equation (5.4.1) becomes
Chapter 7: Extensions

For the estimation of $\pi_{11}$, $\hat{\pi}_{11}$, let $-\text{as in chapter 5} - t$, $t', t''$ be two adjacent pixels, $t, t''$ two diagonally adjacent pixels in the orientation of $(t_x - 1, t_y), (t_x, t_y - 1)$ and $\mu = \mu_1 - \mu_0$. The probabilities $\pi_{11}, \hat{\pi}_{11}$ can be estimated from the equivalent equations to (5.5.4) and (5.5.6). Since

$$E(O^T(t)O(t')) = (1 - \pi) \mu_0^T \mu_0 + \pi \mu_1^T \mu_1 - \pi (1 - \pi) (1 - \rho_{tt'}) \mu^T \mu$$

$$= [(1 - \pi) \mu_0 + \pi \mu_1]^T [(1 - \pi) \mu_0 + \pi \mu_1] + \pi (1 - \pi) \rho_{tt'} \mu^T \mu,$$

and the same equation holds with $t''$ replacing $t'$, the estimates $\hat{\pi}_{11}, \hat{\pi}_{11}$ are

$$\hat{\pi}_{11} = \pi^2 + \frac{\hat{E}(O^T(t)O(t')) - [(1 - \pi) \mu_0 + \pi \mu_1]^T [(1 - \pi) \mu_0 + \pi \mu_1]}{\mu^T \mu},$$

$$\hat{\pi}_{11} = \pi^2 + \frac{\hat{E}(O^T(t)O(t'')) - [(1 - \pi) \mu_0 + \pi \mu_1]^T [(1 - \pi) \mu_0 + \pi \mu_1]}{\mu^T \mu}.$$

Alternatively, equivalent estimates can also be obtained from

$$E \left( (O(t) - O(t'))^T (O(t) - O(t')) \right) =$$

$$= 2E (O^T(t)O(t)) - 2E (O^T(t)O(t'))$$

$$= 2 [(1 - \pi) [\text{trace}(\Sigma_0) + \mu_0^T \mu_0] + \pi [\text{trace}(\Sigma_1) + \mu_1^T \mu_1]] - 2 [(1 - \pi) \mu_0^T \mu_0 + \pi \mu_1^T \mu_1]$$

$$+ 2\pi (1 - \pi) (1 - \rho_{tt'}) \mu^T \mu$$

$$= 2 [(1 - \pi) \cdot \text{trace}(\Sigma_0) + \pi \cdot \text{trace}(\Sigma_1)] + 2\pi (1 - \pi) (1 - \rho_{tt'}) \mu^T \mu$$

instead of $E(O^T(t)O(t'))$ and $E \left( (O(t) - O(t''))^T (O(t) - O(t'')) \right)$ instead of $E(O^T(t)O(t''))$. These expectations, and the corresponding estimates, are directly analogous to the univariate case.
7.2. Local Parameter Estimation.

A common concern about a Markovian model like (5.2.1) is the uniformity of the parameter $\beta$ over the whole image [Clif86]. The question is whether one value of $\beta$ is adequate for the whole image or whether the value of $\beta$ should change with location. This is an important issue and local adjustments to the value of $\beta$ have been suggested repeatedly in the literature. In contrast, the number of implementations which allow the parameters of the Markov random field to vary is rather small. This is the case partly because of the substantial difficulty of parameter estimation (see sections 2.2 and 3.4). The complexity of some of the existing parameter estimation methods prohibits their local application on small segments of an image.

On the contrary, the simplicity of the moment estimator makes local estimation of the parameters $\alpha$ and $\beta$ a realistic possibility. The parameters can be estimated locally from any part of the image, $D_{\text{sub}}$, by applying the estimation procedure of section 5.3 to the data $\{O(t) : \forall t \in D_{\text{sub}}\}$ of the particular subimage. For instance, the moment estimator can be applied locally either to strips of rows (or columns) of pixels or to blocks of pixels (see chapter 6). Therefore, the estimator of chapter 5 makes possible the implementation of a reconstruction algorithm with spatially varying parameters. At the same time, local estimation of the parameters on different parts of the image facilitates the investigation of the spatial behavior of the parameters. Chapter 6 describes how to combine a local application of the moment estimator with a variance estimation technique in order to estimate the variance of the parameter estimates. Either the variance estimation technique of chapter 6 (see section 6.2) or the confidence interval method of Hall88b (see section 6.4) can be used for this purpose.

7.3. Propagation of Errors.

The method of chapter 6 prescribes that the variance of an overall estimate $\hat{\theta}$ be approximated by the variance of the average of the strip estimates $\hat{\theta}_1, \ldots, \hat{\theta}_s$. This approximation is exact when the overall estimate $\hat{\theta}$ happens to be the average of the $s$ strip estimates $\hat{\theta}_1, \ldots, \hat{\theta}_s$. For example, if $\theta = \pi$ is the parameter and $\hat{\theta} = \bar{\pi}$ is the moment estimate of section 5.4, then $\hat{\theta}$ is just the average of the strip estimates $\hat{\theta}_i = \bar{\pi}_i$.

The approximation can also be made exact, without much difficulty, when the overall
estimate is a general linear combination of the strip estimates. For instance, equation (6.2.8) will no longer hold for the autoregressive model (6.2.2) but an analogous formula can easily be derived. There may be situations, however, where the overall estimate $\hat{\theta}$ is not a linear combination of the individual strip estimates. An example is the estimate $\hat{\beta}$ of section 5.3. Each strip estimate $\hat{\beta}_i$ is obtained by solving the system (5.3.2), (5.3.3) for that particular strip, whereas the overall estimate $\hat{\beta}$ is obtained by solving the same system for the whole image.

When the overall estimate is not a linear combination of the strip estimates, one way of applying the method of chapter 6 is to use the average of the strip estimates as the overall estimate. For example, this would imply that the average of the $\hat{\beta}_i$'s be used as an estimate for $\beta$ (rather than the estimate obtained from solving the system (5.3.2), (5.3.3) for the whole image). This approach has the advantage that the variance of the estimate can be calculated directly with the method of chapter 6. If this approach is inadequate, an alternative is to derive estimates of variance with the delta method.

Let $\hat{\theta}$ be the estimate of the parameter under investigation and suppose that there exists some other parameter $\zeta$ whose overall estimate $\hat{\zeta}$ is a linear combination of the strip estimates $\hat{\zeta}_1, \ldots, \hat{\zeta}_s$. Suppose further that $\hat{\theta}$ can be written as a known, well-behaved function of the estimate $\hat{\zeta}$. Then an estimate of the variance $\text{var}(\hat{\theta})$ can be computed from an estimate of the variance $\text{var}(\hat{\zeta})$; the latter can be obtained by applying VETIPE to $\hat{\zeta}$.

To make things concrete, consider again the model (5.2.1) and let $(\hat{\alpha}, \hat{\beta})$ be the estimates of interest. According to section 5.3, the estimates $\hat{\alpha}$ and $\hat{\beta}$ are obtained from the estimates $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ by solving the two equations (5.3.2), (5.3.3). Thus $\hat{\alpha}$ and $\hat{\beta}$ can be expressed as known, although implicit, functions of $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$. Furthermore, these moment estimates do satisfy the property that their overall value is a linear combination of their strip values. Therefore, the variances $\text{var}(\hat{\pi}), \text{var}(\hat{\pi}_{11}), \text{var}(\hat{\pi}_{11})$ and the pairwise covariances $\text{cov}(\hat{\pi}, \hat{\pi}_{11}), \text{cov}(\hat{\pi}, \hat{\pi}_{11}), \text{cov}(\hat{\pi}_{11}, \hat{\pi}_{11})$ can be found with the method of chapter 6 (see section 8.3 for details).

The variances of $\hat{\alpha}, \hat{\beta}$ can be derived as follows. Since the model (5.2.1) is a two parameter model, only two of the three probabilities $\pi, \pi_{11}, \pi_{11}$ are truly independent parameters.
Without loss of generality let \( \pi, \pi_{11} \) be the independent parameters and consider \( \hat{\pi}_{11} \) to be a function of \( \pi, \pi_{11} \). Differentiating equations (5.3.2), (5.3.3) with respect to \( \hat{\pi}, \hat{\pi}_{11} \) yields

\[
\begin{align*}
D_1 \frac{\partial \hat{\alpha}}{\partial \hat{\pi}} + D_2 \frac{\partial \hat{\beta}}{\partial \hat{\pi}} &= \Delta_1, \\
D_1 \frac{\partial \hat{\alpha}}{\partial \hat{\pi}_{11}} + D_2 \frac{\partial \hat{\beta}}{\partial \hat{\pi}_{11}} &= \Delta_2, \\
D_3 \frac{\partial \hat{\alpha}}{\partial \hat{\pi}} + D_4 \frac{\partial \hat{\beta}}{\partial \hat{\pi}} &= \Delta_3, \\
D_3 \frac{\partial \hat{\alpha}}{\partial \hat{\pi}_{11}} + D_4 \frac{\partial \hat{\beta}}{\partial \hat{\pi}_{11}} &= \Delta_4,
\end{align*}
\]

(7.3.1)

where the quantities \( D_1, D_2, D_3, D_4, \Delta_1, \Delta_2, \Delta_3, \Delta_4 \) are defined as

\[
\begin{align*}
D_1 &= 2e^{2\hat{\pi}} e^{\hat{\beta}} (\hat{\pi}_{11} - \hat{\pi}) + e^{\hat{\alpha}} e^{\hat{\beta}} (\hat{\pi}_{11} - \hat{\pi}_{11}) + e^{\hat{\alpha}} (\hat{\pi}_{11} - \hat{\pi} + \hat{\pi}_{11}), \\
D_2 &= e^{2\hat{\pi}} e^{\hat{\beta}} (\hat{\pi}_{11} - \hat{\pi}) + e^{\hat{\alpha}} e^{\hat{\beta}} (\hat{\pi}_{11} - \hat{\pi}_{11}), \\
D_3 &= e^{\hat{\alpha}} e^{\hat{\beta}} (\hat{\pi}_{11} - \hat{\pi}_{11}) + 2e^{2\hat{\pi}} (2\hat{\pi} - 1 - \hat{\pi}_{11}) + e^{\hat{\alpha}} (3\hat{\pi} - 1 - \hat{\pi}_{11} - \hat{\pi}_{11}), \\
D_4 &= e^{\hat{\alpha}} e^{\hat{\beta}} (\hat{\pi} - \hat{\pi}_{11}) + e^{\hat{\beta}} (\hat{\pi} - \hat{\pi}_{11}), \\
\Delta_1 &= e^{2\hat{\pi}} e^{\hat{\beta}} + e^{\hat{\alpha}} (e^{\hat{\beta}} - 1) \frac{\partial \hat{\pi}_{11}}{\partial \hat{\pi}} + e^{\hat{\alpha}}, \\
\Delta_2 &= -e^{2\hat{\pi}} e^{\hat{\beta}} + e^{\hat{\alpha}} e^{\hat{\beta}} \left( \frac{\partial \hat{\pi}_{11}}{\partial \hat{\pi}_{11}} - 1 \right) - e^{\hat{\alpha}} \left( 1 + \frac{\partial \hat{\pi}_{11}}{\partial \hat{\pi}_{11}} \right) - 1, \\
\Delta_3 &= e^{\hat{\alpha}} \left( \frac{\partial \hat{\pi}_{11}}{\partial \hat{\pi}} - 3 \right) - e^{\hat{\alpha}} e^{\hat{\beta}} \frac{\partial \hat{\pi}_{11}}{\partial \hat{\pi}_{11}} - 2e^{\hat{\alpha}} - e^{\hat{\beta}}, \\
\Delta_4 &= e^{2\hat{\pi}} + e^{\hat{\alpha}} \left( 1 + \frac{\partial \hat{\pi}_{11}}{\partial \hat{\pi}_{11}} \right) + e^{\hat{\alpha}} e^{\hat{\beta}} \left( 1 - \frac{\partial \hat{\pi}_{11}}{\partial \hat{\pi}_{11}} \right).
\end{align*}
\]

Assuming for a moment that the derivatives \( \partial \hat{\pi}_{11} / \partial \hat{\pi}, \partial \hat{\pi}_{11} / \partial \hat{\pi}_{11} \) are known, the equations (7.3.1) form a linear system in the unknown derivatives \( \partial \hat{\alpha} / \partial \hat{\pi}, \partial \hat{\beta} / \partial \hat{\pi}, \partial \hat{\alpha} / \partial \hat{\pi}_{11}, \partial \hat{\beta} / \partial \hat{\pi}_{11} \).

Thus these partial derivatives can be found by solving the system (7.3.1):

\[
\begin{align*}
\frac{\partial \hat{\alpha}}{\partial \hat{\pi}} &= \frac{D_4 \Delta_1 - D_2 \Delta_3}{D_1 D_4 - D_2 D_3}, \\
\frac{\partial \hat{\beta}}{\partial \hat{\pi}} &= \frac{D_1 \Delta_3 - D_3 \Delta_1}{D_1 D_4 - D_2 D_3}, \\
\frac{\partial \hat{\alpha}}{\partial \hat{\pi}_{11}} &= \frac{D_4 \Delta_2 - D_2 \Delta_4}{D_1 D_4 - D_2 D_3}, \\
\frac{\partial \hat{\beta}}{\partial \hat{\pi}_{11}} &= \frac{D_1 \Delta_4 - D_3 \Delta_2}{D_1 D_4 - D_2 D_3}.
\end{align*}
\]
Now the variance of the estimate $\hat{\beta}$ can be computed from the first-order Taylor expansion:

$$\hat{\beta} = \hat{\beta}(\hat{\pi}, \hat{\pi}_{11})$$

$$\approx \beta(E(\hat{\pi}), E(\hat{\pi}_{11})) + \left(\frac{\partial \hat{\beta}}{\partial \hat{\pi}}\right)_{E(\hat{\pi})} (\hat{\pi} - E(\hat{\pi})) + \left(\frac{\partial \hat{\beta}}{\partial \hat{\pi}_{11}}\right)_{E(\hat{\pi}_{11})} (\hat{\pi}_{11} - E(\hat{\pi}_{11})) + \ldots$$

Ignoring the higher order terms and taking variances leads to

$$\text{var}(\hat{\beta}) \approx \left(\frac{\partial \hat{\beta}}{\partial \hat{\pi}}\right)^2_{E(\hat{\pi})} \text{var}(\hat{\pi}) + \left(\frac{\partial \hat{\beta}}{\partial \hat{\pi}_{11}}\right)^2_{E(\hat{\pi}_{11})} \text{var}(\hat{\pi}_{11}) + 2 \left(\frac{\partial \hat{\beta}}{\partial \hat{\pi}}\right)_{E(\hat{\pi})} \left(\frac{\partial \hat{\beta}}{\partial \hat{\pi}_{11}}\right)_{E(\hat{\pi}_{11})} \text{cov}(\hat{\pi}, \hat{\pi}_{11}),$$

(7.3.2)

and, similarly, for $\hat{\alpha}$:

$$\text{var}(\hat{\alpha}) \approx \left(\frac{\partial \hat{\alpha}}{\partial \hat{\pi}}\right)^2_{E(\hat{\pi})} \text{var}(\hat{\pi}) + \left(\frac{\partial \hat{\alpha}}{\partial \hat{\pi}_{11}}\right)^2_{E(\hat{\pi}_{11})} \text{var}(\hat{\pi}_{11}) + 2 \left(\frac{\partial \hat{\alpha}}{\partial \hat{\pi}}\right)_{E(\hat{\pi})} \left(\frac{\partial \hat{\alpha}}{\partial \hat{\pi}_{11}}\right)_{E(\hat{\pi}_{11})} \text{cov}(\hat{\pi}, \hat{\pi}_{11}).$$

(7.3.3)

These equations provide a first-order approximation to the variances $\text{var}(\hat{\alpha}), \text{var}(\hat{\beta})$, as all the terms on the right hand side are known.

The remaining detail is the computation of the partial derivatives $\partial \hat{\pi}_{11}/\partial \hat{\pi}, \partial \hat{\pi}_{11}/\partial \hat{\pi}_{11}$. These are not known because the explicit form of the constraint between $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ is not known (see sections 5.2 and 5.3). In practice, however, there is a simple way of estimating these derivatives. Since $s$ pairs of estimates $(\hat{\pi}_i, \hat{\pi}_{11, i})$ are available from the $s$ strips, the function $\hat{\pi}_{11} = \hat{\pi}_{11}(\hat{\pi})$ can be approximated by fitting a smoother through the points $(\hat{\pi}_i, \hat{\pi}_{11, i})$. Then the derivative $\partial \hat{\pi}_{11}/\partial \hat{\pi}$ can be estimated by the derivative of the smoother. The selection of an appropriate smoother is not a major problem: basically any reasonable choice will do (e.g., a cubic spline). The derivative $\partial \hat{\pi}_{11}/\partial \hat{\pi}_{11}$ can be estimated similarly from the $s$ pairs $(\hat{\pi}_{11, i}, \hat{\pi}_{11, i})$. 
Chapter 8

Numerical Results

This chapter contains the results of some numerical experiments with images. These results illustrate the application of the methods discussed in chapters 4, 5, 6 and 7. Section 8.1 describes the two images which are used for the experiments. Section 8.2 presents an application of the moment estimator of chapter 5. This section shows the obtained estimates and the corresponding reconstructions with the modified ICM algorithm. Section 8.3 presents an application of the variance estimation technique of chapter 6: VETIPE is used to estimate the variance of the probability estimates obtained in section 8.2.

8.1. The Images.

Two binary images are used for the numerical experiments. These images are shown in Figure 8.1; they are squares of 256 \times 256 pixels. The first image, provided by Art Owen, is a real image of a map of northern Scotland. The black pixels represent land and the white pixels represent water. Each pixel is a square with sides approximately equal to one kilometer. This image is interesting because it exhibits a number of features: (a) the image contains both large homogeneous pieces and isolated islands, and (b) the right part of the image is considerably smoother than the left part. These properties are useful in assessing the behavior of the suggested methods (i.e., moment estimation of the parameters and estimation of the variance) under changing conditions.

The second image is an artificial image of a checkerboard with large squares. Each square in the checkerboard consists of 32 \times 32 pixels. This image is interesting because of its high degree of regularity. This regularity facilitates visual comparisons between reconstructed images. Moreover, it facilitates the interpretation of the obtained results because their expected properties can be figured out easily. For the purposes of the numerical experiments, this image compliments well the more irregular map of Scotland.
Figure 8.1 The original images.
Section 8.2: Application of a Markov Mesh Model

The "observed" images are degraded versions of the original images. Following equation (5.2.2), the value of $I(t)$ is 0 if the pixel is white in the original image and 1 if the pixel is black. That is, the mean of the conditional distribution $p\{O(t)|I(t) = 0\}$ is taken to be 0 and the mean of the conditional distribution $p\{O(t)|I(t) = 1\}$ is taken to be 1. The noise process $\epsilon(t)$ is normal with mean 0 and standard deviation 1. Different levels of degradation are obtained by setting the standard deviations, $\sigma_0$ and $\sigma_1$, of the distributions $p\{O(t)|I(t)\}$ to different values. The values of $\sigma_0$ and $\sigma_1$ are taken to be equal to a common standard deviation $\sigma$. Figure 8.2 shows four degraded versions of the map of Scotland; these images are obtained by setting $\sigma$ equal to 1, 2, 4 and 8 respectively. Figure 8.3 shows the corresponding degraded versions of the checkerboard. Since the difference in the means of the conditional distributions $p\{O(t)|I(t)\}$ is 1, each unit of $\sigma$ corresponds to one unit of separation between the two populations $\{O(t)|I(t) = 0\}$ and $\{O(t)|I(t) = 1\}$.

The displayed images of Figures 8.2, 8.3 are thresholded versions of the observations. Since the observations are continuous (according to model (5.2.2)), those that are greater than 0.5 are shown as black and those that are less or equal to 0.5 are shown as white. Therefore, Figures 8.2 and 8.3 effectively show the reconstructions of the simplest possible, non-contextual classification algorithm: a pixel is classified as white if its observation is smaller or equal to 0.5. The error rate (i.e., the percent of misclassified pixels) of this algorithm is the same for either the Scotland or the checkerboard image: it is 31% for $\sigma = 1$ (and grows to 40% for $\sigma = 2$, 45% for $\sigma = 4$ and 47% for $\sigma = 8$). Error rates are an objective criterion for comparisons but do not capture at all the spatial continuity of an image. More important than error rates is the visual quality of the reconstructed images. The thresholding algorithm results in reconstructions in which it is difficult to recognize the original images. This is evident from Figures 8.2 and 8.3, especially for high levels of noise (i.e., $\sigma = 2, 4, 8$).

8.2. Application of a Markov Mesh Model.

8.2.1. Moment Estimates

The moment estimation technique of chapter 5 has been applied to the Scotland and the checkerboard image for different levels of noise in the observed images. The tables below
Figure 8.2 Noisy realizations of Scotland.
Figure 8.3 Noisy realizations of the checkerboard.
present the obtained estimates for $\pi, \pi_{11}, \tilde{\pi}_{11}, \alpha$ and $\beta$, under the assumption that the model (5.2.1) is going to be used for reconstruction.

**Table 8.1** Parameter estimates from Scotland.

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>$\hat{\pi}$</th>
<th>$\hat{\pi}_{11}$</th>
<th>$\hat{\tilde{\pi}}_{11}$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 0$</td>
<td>0.657</td>
<td>0.644</td>
<td>0.641</td>
<td>0.188</td>
<td>2.259</td>
</tr>
<tr>
<td>$\sigma = 1$</td>
<td>0.654</td>
<td>0.643</td>
<td>0.644</td>
<td>0.232</td>
<td>2.171</td>
</tr>
<tr>
<td>$\sigma = 2$</td>
<td>0.651</td>
<td>0.644</td>
<td>0.654</td>
<td>0.346</td>
<td>2.007</td>
</tr>
<tr>
<td>$\sigma = 4$</td>
<td>0.645</td>
<td>0.647</td>
<td>0.695</td>
<td>0.660</td>
<td>1.576</td>
</tr>
<tr>
<td>$\sigma = 8$</td>
<td>0.633</td>
<td>0.666</td>
<td>0.861</td>
<td>-1.598</td>
<td>1.551</td>
</tr>
</tbody>
</table>

**Table 8.2** Parameter estimates from the checkerboard.

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>$\hat{\pi}$</th>
<th>$\hat{\pi}_{11}$</th>
<th>$\hat{\tilde{\pi}}_{11}$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 0$</td>
<td>0.500</td>
<td>0.486</td>
<td>0.473</td>
<td>0.000</td>
<td>3.426</td>
</tr>
<tr>
<td>$\sigma = 1$</td>
<td>0.497</td>
<td>0.485</td>
<td>0.476</td>
<td>-0.001</td>
<td>2.882</td>
</tr>
<tr>
<td>$\sigma = 2$</td>
<td>0.494</td>
<td>0.485</td>
<td>0.485</td>
<td>-0.008</td>
<td>2.342</td>
</tr>
<tr>
<td>$\sigma = 4$</td>
<td>0.488</td>
<td>0.488</td>
<td>0.525</td>
<td>-0.048</td>
<td>1.667</td>
</tr>
<tr>
<td>$\sigma = 8$</td>
<td>0.476</td>
<td>0.506</td>
<td>0.689</td>
<td>-0.173</td>
<td>1.062</td>
</tr>
<tr>
<td></td>
<td>0.476</td>
<td>0.506</td>
<td>0.689</td>
<td>-1.453</td>
<td>1.398</td>
</tr>
</tbody>
</table>

The first row (in each of the two tables) contains the parameter estimates which are obtained from applying the moment estimator to the uncorrupted images. These estimates represent the "true" values of the parameters and are useful for comparative purposes. They must, however, be interpreted with caution. A true value exists only for the probabilities $\pi, \pi_{11}, \tilde{\pi}_{11}$ (because these are well defined quantities regardless of any image model). On the other hand, the "true" values of $\alpha$ and $\beta$ are the values that these parameters would have assumed if the model (5.2.1) were the "correct" model for the underlying image. But, of course, a model like (5.2.1) can never be the exact mechanism which generates the true image.
As seen from Tables 8.1 and 8.2, the moment estimator produces acceptable estimates \( \hat{\alpha}, \hat{\beta} \) in all the tried cases. This is encouraging empirical evidence for the feasibility issue which is raised in sections 5.6 and 5.8 (namely, whether acceptable moment estimates exist). Experience with a number of other images also supports the belief that in practice feasibility is not a serious concern.

In two occasions there are more than one acceptable estimates: when \( \sigma = 4 \) in the Scotland image and when \( \sigma = 8 \) in the checkerboard image. In both cases there are three pairs of estimates \((\hat{\alpha}, \hat{\beta})\). It is interesting to note that these estimates differ significantly in the values of \( \hat{\alpha} \) and less so in the values of \( \hat{\beta} \). The selection of one of the multiple estimates can be fairly arbitrary. For example, a pair of estimates can be chosen on the basis of some desirable constraint (e.g., that both \( \alpha \) and \( \beta \) are positive). In lack of a clear guideline, the most reasonable approach is to select the estimates \((\hat{\alpha}, \hat{\beta})\) on the basis of the resulting reconstructions: simply choose those estimates which yield the best reconstruction.

Another point of interest is that the obtained estimates vary fairly smoothly with the level of noise present in the degraded image. As expected, this is more evident for the probability estimates \( \hat{\tau}, \hat{\tau}_{11}, \hat{\tilde{\tau}}_{11} \) than for the estimates \( \hat{\alpha}, \hat{\beta} \). The reason is that the probabilities \( \tau, \tau_{11}, \tilde{\tau}_{11} \) are directly related to tangible image properties, whereas the parameters \( \alpha, \beta \) pertain to a hypothetical, and unlikely, model for the underlying image. As a result, the estimation of \( \tau, \tau_{11}, \tilde{\tau}_{11} \) is simpler, more direct and more accurate than the estimation of the parameters of the Markov model.

### 8.2.2. Reconstructions with the Modified ICM

As an example of the modified ICM algorithm, the estimates \( \hat{\alpha}, \hat{\beta} \) of the previous section have been used to reconstruct the two test images. Figure 8.4 shows the reconstructed images of Scotland and the checkerboard from their noisy counterparts with \( \sigma = 1 \). Along with each reconstruction, Figure 8.4 shows a plot which highlights the difference between the original and the reconstructed image. In this plot the pixels which are misclassified are shown as black. The actual reconstructions are shown on the left and the differences from the original images are shown on the right.

The reconstructions have been carried out with a version of the modified ICM algorithm
Figure 8.4 Reconstructions with the modified ICM for $\sigma = 1$. 
which in fact does not even use the conditioning on all the other pixels in the lattice (see section 4.3). Instead, the conditioning is done only with respect to the predecessors of each pixel. This choice is made because it results in a significant speed-up of the computation. Also, the algorithm is implemented so that the Markov mesh model changes direction on successive iterations (see section 4.3).

The error rates of the modified ICM algorithm for the two images are 8% for the Scotland image and 7% for the checkerboard. These error rates compare favorably to the 31% of the thresholding algorithm and are comparable to the error rates achieved by the original ICM algorithm (4% with $\beta = 1$ and slightly larger with $\beta = 2$). However, at higher levels of noise (e.g., $\sigma = 4$ or $\sigma = 8$) the modified ICM performs very poorly, as the resulting reconstructions do not resemble at all the true images. In fairness, it should be mentioned that the original ICM algorithm exhibits similar behavior. For high levels of noise it has been found that ICE (another variant of the ICM algorithm suggested by Owen86) performs significantly better than either the ICM or the modified ICM algorithm. The behavior of the two algorithms will be investigated further in future work.

8.3. Application of the Variance Estimation.

The algorithm of chapter 6 suggests estimating the variance of an estimate $\hat{\theta}$ with the variance of the average of the fake time series $\hat{\theta}_1, \ldots, \hat{\theta}_s$, where $\hat{\theta}_i$ is the parameter estimate obtained from the $i$th strip in the image. As pointed out in section 7.3, this approach is perfectly suited for situations in which the overall estimate $\hat{\theta}$ is an average of the individual strip estimates $\hat{\theta}_i$. This is the case for the probability estimates $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ of sections 5.4 and 5.5. Thus the method of chapter 6 (VETIPE) is directly applicable to the estimates $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ (actually the overall estimates $\hat{\pi}_{11}, \hat{\pi}_{11}$ are not exact averages of the corresponding strip estimates, but the minor edge effect between the strips can, for all practical purposes, be ignored). This section illustrates the application of VETIPE to these estimates.

There are two reasons for which one may be interested in the variances (and the covariances) of $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$. First, the parameters $\pi, \pi_{11}, \pi_{11}$ are useful image parameters and can be of interest in their own right (that is, without any reference to a Markov model). Second, the estimates of $\text{var}(\hat{\pi}), \text{var}(\hat{\pi}_{11}), \text{var}(\hat{\pi}_{11})$ (and of the pairwise covariances) can be used to obtain
estimates of variance for $\hat{\alpha}, \hat{\beta}$ with the method of section 7.3.

To apply VETIPE, the images are divided into strips of 32, 16 and 8 rows of pixels. Three different strip sizes are considered in order to investigate the effect of the strip size on the accuracy of the results. Since the images are squares of $256 \times 256$ pixels, the division into strips of 32 rows contains 8 strips, the division into strips of 16 rows contains 16 strips and the division into strips of 8 rows contains 32 strips. These strip sizes are chosen so that there is some balance between the number of available strips and the goodness of the strip estimates: the division into strips of 16 rows represents the $\sqrt{n}$ rule (see section 6.2).

The variances of the estimates $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ are calculated with the help of an autoregressive model of order 1 (equation (6.2.2)). This model is

$$\hat{x}_i - \mu_{\hat{x}} = \omega (\hat{x}_{i-1} - \mu_{\hat{x}}) + \varepsilon_i,$$

with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2_{\varepsilon}$,

$$\hat{x}_{11,i} - \mu_{\hat{x}_{11}} = \psi (\hat{x}_{11,i-1} - \mu_{\hat{x}_{11}}) + \varepsilon_i,$$

with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2_{\varepsilon}$,

$$\hat{x}_{11,i} - \mu_{\hat{x}_{11}} = \phi (\hat{x}_{11,i-1} - \mu_{\hat{x}_{11}}) + r_i,$$

with $E(r_i) = 0$ and $\text{var}(r_i) = \sigma^2_r$,

and is determined by the number of available points $s$ (that is, the parameters $\omega, \psi, \phi, \sigma^2_{\varepsilon}, \sigma^2_{r}$ depend on the value of $s$). This model results in the following variance estimates for $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ (see equation (6.2.8)):

$$\text{var}(\hat{\pi}) = \text{var}(\hat{\pi}_i) \approx \frac{\sigma^2_{\varepsilon}}{s(1 - \omega)^2},$$

$$\text{var}(\hat{\pi}_{11}) = \text{var}(\hat{\pi}_{11,i}) \approx \frac{\sigma^2_{\varepsilon}}{s(1 - \psi)^2},$$

$$\text{var}(\hat{x}_{11}) = \text{var}(\hat{x}_{11,i}) \approx \frac{\sigma^2_r}{s(1 - \phi)^2}.$$

Here $s$ is 8, 16 or 32 depending on the strip size.

The covariances of the estimates $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ can also be computed from this model, provided some correlation structure is introduced in the autoregressive model. One way of doing so is to assume that the noise processes $\varepsilon, \mu, \nu$ and $r$ are correlated. That is, the error in the $\pi$ estimate from any strip is correlated with the error in the $\pi_{11}$ estimate from the same strip:

$$\text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \gamma_{\varepsilon\varepsilon}, & \text{if } i = j; \\ 0, & \text{if } i \neq j; \end{cases}$$

$$\text{cov}(\varepsilon_i, \tau_j) = \begin{cases} \gamma_{\varepsilon\tau}, & \text{if } i = j; \\ 0, & \text{if } i \neq j; \end{cases}.$$
Section 8.3: Application of the Variance Estimation

\[ \text{cov}(\epsilon_i, r_j) = \begin{cases} \gamma_{er}, & \text{if } i = j; \\ 0, & \text{if } i \neq j. \end{cases} \]

Under this assumption, the pairwise covariances between the estimates \( \hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11} \) are given by:

\[
\begin{align*}
\text{cov}(\hat{\pi}, \hat{\pi}_{11}) &= \frac{\gamma_{ee}}{s^2(1 - \omega \psi)} [s + (s - 1)(\omega + \psi)], \\
\text{cov}(\hat{\pi}_{11}, \hat{\pi}_{11}) &= \frac{\gamma_{er}}{s^2(1 - \psi \phi)} [s + (s - 1)(\psi + \phi)], \\
\text{cov}(\hat{\pi}, \hat{\pi}_{11}) &= \frac{\gamma_{er}}{s^2(1 - \omega \phi)} [s + (s - 1)(\omega + \phi)].
\end{align*}
\tag{8.3.1}
\]

To see this, treat \( \hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11} \) as the averages of the corresponding strip estimates and observe that

\[
\text{cov}(\hat{\pi}, \hat{\pi}_{11}) = \text{cov} \left( \frac{1}{s} \sum_{i=1}^{s} \hat{\pi}_{11,i}, \frac{1}{s} \sum_{i=1}^{s} \hat{\pi}_{11,i} \right) = \frac{1}{s^2} \sum_{i=1}^{s} \sum_{j=1}^{s} \text{cov}(\hat{\pi}_{11,i}, \hat{\pi}_{11,j}).
\]

In this sum all the terms are 0 except for the \( s \) terms \( \text{cov}(\hat{\pi}_{11,i}, \hat{\pi}_{11,i}) \), the \( s - 1 \) terms \( \text{cov}(\hat{\pi}_{11,i}, \hat{\pi}_{11,i-1}) \) and the \( s - 1 \) terms \( \text{cov}(\hat{\pi}_{11,i}, \hat{\pi}_{11,i+1}) \). Computing these terms

\[
\begin{align*}
\text{cov}(\hat{\pi}_{11,i}, \hat{\pi}_{11,i}) &= \frac{\gamma_{ee}}{1 - \omega \psi}, \\
\text{cov}(\hat{\pi}_{11,i}, \hat{\pi}_{11,i-1}) &= \frac{\omega \gamma_{ee}}{1 - \omega \psi}, \\
\text{cov}(\hat{\pi}_{11,i}, \hat{\pi}_{11,i+1}) &= \frac{\psi \gamma_{ee}}{1 - \omega \psi}
\end{align*}
\]

yields equation (8.3.1).

The covariances \( \gamma \) are estimated by the sample covariances

\[
\begin{align*}
\hat{\gamma}_{ee} &= \frac{1}{s} \sum_{i=1}^{s} (\hat{\pi}_i - \bar{\pi})(\hat{\pi}_{11,i} - \bar{\pi}_{11}) \approx \frac{1}{s} \sum_{i=1}^{s} (\hat{\pi}_i - \bar{\pi})(\hat{\pi}_{11,i} - \hat{\pi}_{11}), \\
\hat{\gamma}_{er} &= \frac{1}{s} \sum_{i=1}^{s} (\hat{\pi}_{11,i} - \bar{\pi}_{11})(\hat{\pi}_{11,i} - \bar{\pi}_{11}) \approx \frac{1}{s} \sum_{i=1}^{s} (\hat{\pi}_{11,i} - \hat{\pi}_{11})(\hat{\pi}_{11,i} - \hat{\pi}_{11}), \\
\hat{\gamma}_{er} &= \frac{1}{s} \sum_{i=1}^{s} (\hat{\pi}_i - \bar{\pi})(\hat{\pi}_{11,i} - \bar{\pi}_{11}) \approx \frac{1}{s} \sum_{i=1}^{s} (\hat{\pi}_i - \hat{\pi})(\hat{\pi}_{11,i} - \hat{\pi}_{11}),
\end{align*}
\]

and the corresponding estimates for the covariances \( \text{cov}(\hat{\pi}, \hat{\pi}_{11}), \text{cov}(\hat{\pi}_{11}, \hat{\pi}_{11}), \text{cov}(\hat{\pi}, \hat{\pi}_{11}) \) are obtained by replacing the covariances \( \gamma \) and the autoregressive parameters \( \omega, \psi, \phi \) by their estimates in equation (8.3.1).

The next two tables summarize the standard deviations and the covariances obtained from the map of Scotland and the checkerboard image. The results are presented for three
levels of noise in the observed images: $\sigma = 1, 2, 4$. The true values of $\pi, \pi_{11}, \hat{\pi}_{11}$ are 0.657, 0.644 and 0.641 respectively for the image of Scotland, and 0.500, 0.486, 0.473 for the checkerboard image (see Tables 8.1, 8.2).

Table 8.3 Variances of probability estimates from the rows of Scotland.

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>Strip Size</th>
<th>$\sqrt{\text{var}(\hat{\pi})}$</th>
<th>$\sqrt{\text{var}(\hat{\pi}_{11})}$</th>
<th>$\sqrt{\text{var}(\hat{\pi}_{11})}$</th>
<th>$\text{cov}(\hat{\pi}, \hat{\pi}_{11})$</th>
<th>$\text{cov}(\hat{\pi}, \hat{\pi}_{11})$</th>
<th>$\text{cov}(\hat{\pi}<em>{11}, \hat{\pi}</em>{11})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 1$</td>
<td>32</td>
<td>0.1395</td>
<td>0.1384</td>
<td>0.1325</td>
<td>0.0110</td>
<td>0.0109</td>
<td>0.0107</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.1022</td>
<td>0.1008</td>
<td>0.1001</td>
<td>0.0071</td>
<td>0.0071</td>
<td>0.0069</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.0906</td>
<td>0.0851</td>
<td>0.0880</td>
<td>0.0042</td>
<td>0.0042</td>
<td>0.0041</td>
</tr>
<tr>
<td>$\sigma = 2$</td>
<td>32</td>
<td>0.1308</td>
<td>0.1314</td>
<td>0.1149</td>
<td>0.0103</td>
<td>0.0101</td>
<td>0.0098</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.0912</td>
<td>0.0868</td>
<td>0.0857</td>
<td>0.0066</td>
<td>0.0065</td>
<td>0.0062</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.0876</td>
<td>0.0768</td>
<td>0.0834</td>
<td>0.0039</td>
<td>0.0039</td>
<td>0.0037</td>
</tr>
<tr>
<td>$\sigma = 4$</td>
<td>32</td>
<td>0.1137</td>
<td>0.1151</td>
<td>0.0853</td>
<td>0.0090</td>
<td>0.0086</td>
<td>0.0081</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.0743</td>
<td>0.0641</td>
<td>0.0599</td>
<td>0.0053</td>
<td>0.0053</td>
<td>0.0045</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.0780</td>
<td>0.0695</td>
<td>0.0796</td>
<td>0.0036</td>
<td>0.0039</td>
<td>0.0035</td>
</tr>
</tbody>
</table>

Table 8.4 Variances of probability estimates from the rows of the checkerboard.

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>Strip Size</th>
<th>$\sqrt{\text{var}(\hat{\pi})}$</th>
<th>$\sqrt{\text{var}(\hat{\pi}_{11})}$</th>
<th>$\sqrt{\text{var}(\hat{\pi}_{11})}$</th>
<th>$\text{cov}(\hat{\pi}, \hat{\pi}_{11})$</th>
<th>$\text{cov}(\hat{\pi}, \hat{\pi}_{11})$</th>
<th>$\text{cov}(\hat{\pi}<em>{11}, \hat{\pi}</em>{11})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 1$</td>
<td>32</td>
<td>0.0042</td>
<td>0.0045</td>
<td>0.0043</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.0034</td>
<td>0.0048</td>
<td>0.0034</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.0037</td>
<td>0.0046</td>
<td>0.0052</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>$\sigma = 2$</td>
<td>32</td>
<td>0.0086</td>
<td>0.0083</td>
<td>0.0108</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.0070</td>
<td>0.0117</td>
<td>0.0068</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.0073</td>
<td>0.0095</td>
<td>0.0118</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>$\sigma = 4$</td>
<td>32</td>
<td>0.0173</td>
<td>0.0159</td>
<td>0.0247</td>
<td>0.0002</td>
<td>0.0003</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.0139</td>
<td>0.0271</td>
<td>0.0201</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.0146</td>
<td>0.0234</td>
<td>0.0323</td>
<td>0.0002</td>
<td>0.0003</td>
<td>0.0003</td>
</tr>
</tbody>
</table>
The parameters $\pi, \pi_{11}, \hat{\pi}_{11}$ are well-estimated by the estimates $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ and this is reflected in Tables 8.3 and 8.4. Especially for the checkerboard image, the variability in the obtained estimates is negligible. This is to be expected because the horizontal strips are aligned with the pattern in the image and become essentially "copies" of each other. Perhaps the most striking conclusion from Tables 8.3, 8.4 is that for the image of Scotland (which is the more interesting of the two images) the amount of noise present in the observations does not dramatically affect the estimates of variance.

As suggested by section 6.4, the division of the image into strips is not only a vehicle for estimating the variance of an estimate but, more generally, a useful tool for data analysis (of image data). The map of Scotland is much smoother in the right part of the image than in the left. It is, therefore, interesting to apply VETIPE to strips of columns and compare the obtained variances to those obtained from the strips of rows. Such a comparison is an example of the type of analysis which is suggested in section 6.4. Table 8.5 shows the estimates of variance for $\hat{\pi}, \hat{\pi}_{11}, \hat{\pi}_{11}$ which are computed by dividing the image into strips of columns (the true values of the probabilities are 0.657, 0.644 and 0.641):

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>Strip Size</th>
<th>$\sqrt{\text{Var}(\hat{\pi})}$</th>
<th>$\sqrt{\text{Var}(\hat{\pi}_{11})}$</th>
<th>$\sqrt{\text{Var}(\hat{\pi}_{11})}$</th>
<th>$\text{cov}(\hat{\pi}, \hat{\pi}_{11})$</th>
<th>$\text{cov}(\hat{\pi}, \hat{\pi}_{11})$</th>
<th>$\text{cov}(\hat{\pi}<em>{11}, \hat{\pi}</em>{11})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 1$</td>
<td>32</td>
<td>0.1240</td>
<td>0.1241</td>
<td>0.1264</td>
<td>0.0105</td>
<td>0.0107</td>
<td>0.0108</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.1481</td>
<td>0.1392</td>
<td>0.1529</td>
<td>0.0074</td>
<td>0.0076</td>
<td>0.0076</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.1321</td>
<td>0.1291</td>
<td>0.1291</td>
<td>0.0041</td>
<td>0.0042</td>
<td>0.0042</td>
</tr>
<tr>
<td>$\sigma = 2$</td>
<td>32</td>
<td>0.1213</td>
<td>0.1149</td>
<td>0.1197</td>
<td>0.0098</td>
<td>0.0098</td>
<td>0.0096</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.1416</td>
<td>0.1150</td>
<td>0.1399</td>
<td>0.0071</td>
<td>0.0072</td>
<td>0.0071</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.1244</td>
<td>0.0971</td>
<td>0.1186</td>
<td>0.0038</td>
<td>0.0040</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\sigma = 4$</td>
<td>32</td>
<td>0.1158</td>
<td>0.0894</td>
<td>0.1202</td>
<td>0.0079</td>
<td>0.0093</td>
<td>0.0077</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.1289</td>
<td>0.0829</td>
<td>0.1207</td>
<td>0.0065</td>
<td>0.0067</td>
<td>0.0063</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.1007</td>
<td>0.0724</td>
<td>0.0859</td>
<td>0.0035</td>
<td>0.0036</td>
<td>0.0037</td>
</tr>
</tbody>
</table>

The numbers in Tables 8.3 and 8.5 are not that different. This is not surprising since the shown
standard deviations are supposed to represent the standard deviation of the same overall estimate (i.e., \( \hat{\pi} \) or \( \hat{\pi}_{11} \) or \( \hat{\pi}_{11} \)). What is more interesting is to compare visually the "time series" from the division into strips of rows to the "time series" from the division into strips of columns.

Figure 8.5 shows plots of the estimates \( \hat{\pi} \) which are obtained from the image of Scotland by dividing it into strips of size 2. The division is done both by rows and by columns; in each case there is a total of 128 strips. The obtained estimates are shown for three different values of \( \sigma \) (1, 2 and 4). It is clear that the estimates from the strips of rows behave differently than the estimates from the strips of columns.

Plots like those of Figure 8.5 can be used as diagnostic tools for the analysis of the image. For example, according to Figure 8.5 there is a sharp change in the values of \( \hat{\pi} \) when the image is divided into strips of rows of pixels. This change occurs around the 60th strip and suggests that the value of \( \pi \) is much higher in the lower part of the image than in the upper part. This observation can serve as a guideline for dividing the image into two parts and analyzing them separately. In the Scotland example, a division into a lower and an upper subimage may be appropriate, especially if \( \pi \) is the parameter of prime interest. Incidentally, a glance at Figure 8.1 verifies the validity of the heuristic conclusion that the lower part of the image contains more land.

This analysis helps illustrate the point that the idea behind the VETIPE algorithm is not restricted to the computation of an estimate of variance. The general idea of creating replicates for the estimates of interest is a valuable tool for the analysis of image data (see chapter 6 for some interesting connections to the bootstrap and the jackknife). Potential applications of this idea will be identified in a forthcoming report.
Figure 8.5 Estimates of $\hat{\pi}$ from strips of Scotland.
Bibliography


Clif86 Clifford, P. Comment on Besa86, *Journal of the Royal Statistical Society*, B, **48**,
3, 284, 1986.


HaHo86  Haslett, J. and Horgan, G., Linear models in spatial discriminant analysis, Department of Statistics, Trinity College, Dublin, Ireland, 1986.


HoYu82  Holmgaard, S. and Yu Song, Y., Gaussian Markov random fields applied to image segmentation, Preprint No. 1, Institute of Mathematical Statistics, University of Copenhagen, Copenhagen, 1982.


JiCh86  Jinchi, H. and Chellappa, R., Restoration of blurred and noisy images using Gaussian Markov random field models, *Proceedings of the Nineteenth Annual Conference on Information Sciences and Systems*, 34–39, John Hopkins University, Baltimore,
Maryland, 1986.


Kittler, J., Comment on *Besa86*, *Journal of the Royal Statistical Society*, B, 48,


SwGr84 Switzer, P. and Green, A. A., Min/max autocorrelation factors for multivariate spatial imagery, Technical Report No. 6, Department of Statistics, Stanford University, Stanford, California, 1984.


Appendix

This appendix contains the proofs of Propositions 5.4.3 and 5.4.4. As before, the lattice $D$ is considered to be an $n \times m$ rectangular array of square pixels and $|D| = nm$ denotes the cardinality of the set of pixels in $D$. Each pixel $t$ is identified with its coordinates on $D$ namely, $t = (t_x, t_y)$. The next lemma provides a necessary technical detail.

**Lemma 1** Let $\delta(t, t') = |t_x - t'_x| + |t_y - t'_y|$ be the city block distance between points $t$ and $t'$ on the rectangle $[0, n] \times [0, m]$. If $\rho_{tt'} = \rho^{\delta(t, t')}$, for some $\rho \in [0, 1)$, then

$$\int_{t,t' \in [0, n] \times [0, m]} \rho_{tt'} d\nu(t) d\nu(t') = o\left(n^2 m^2\right),$$

where $\nu$ stands for the Lebesgue measure on the plane.

**Proof.** The region of integration $\{t = (t_x, t_y), t' = (t'_x, t'_y) : t, t' \in [0, n] \times [0, m], t \neq t'\}$ can be decomposed into nine regions which are defined by whether $t_x = t'_x, t_x > t'_x, t_x < t'_x$ and $t_y = t'_y, t_y > t'_y, t_y < t'_y$. On each of these regions the integral can be obtained by straightforward integration: all the integrals are $o\left(n^2 m^2\right)$ and the result follows. \hfill \blacksquare

With the help of this lemma it is easy to prove Propositions 5.4.3 and 5.4.4.

**Proposition 5.4.3** Let $\delta(t, t') = |t_x - t'_x| + |t_y - t'_y|$ be the city block distance between pixels $t$ and $t'$ on the lattice $D$. If $\rho_{tt'} = \rho^{\delta(t, t')}$, then $\sum_{t,t' \in D} \rho_{tt'} = o\left(|D|^2\right)$.

**Proof.** The sum $\sum_{t,t'} \rho_{tt'}$ can be evaluated exactly. The distance $\delta(t, t')$ is an integer between 1 and $n + m - 2$ and it is possible, though messy, to count for each pixel $t$ the number of pixels $t'$ which are respectively 1, 2, \ldots, $n + m - 2$ units away from $t$. Then the sum $\sum_{t,t'} \rho_{tt'}$ follows from an appropriate summation over all pixels $t$. However, this approach is cumbersome because
the summation involved is tedious. Instead, it is easier to derive an $o\left(|D|^2\right)$ upper bound for $\sum_{t \neq t'} \rho_{tt'}$:

$$\sum_{\substack{t \neq t' \\
(\neq t', (t', 0, n) \times [0, m]}} \rho_{tt'} \, d\nu(t) \, d\nu(t') = \int_{\substack{t \neq t' \\
(\neq t', (t', 0, n) \times [0, m]}} \rho_{tt'} \, d\nu(t) \, d\nu(t'),$$

where $\nu$ is the Lebesgue measure on the plane. Lemma 1 shows that the last integral is $o\left(n^2 m^2\right) = o\left(|D|^2\right)$ and the result follows. □

**Proposition 5.4.4** Let $d(t, t') = \sqrt{|t_x - t'_x|^2 + |t_y - t'_y|^2}$ be the Euclidean distance between pixels $t$ and $t'$ on the lattice $D$. If $\rho_{tt'} = \rho^{d(t, t')}$, then $\sum_{\substack{t \neq t' \\
(\neq t', (t', 0, n) \times [0, m]}} \rho_{tt'} = o\left(|D|^2\right)$.

**Proof.** Since

$$\delta(t, t') \leq \sqrt{2} d(t, t') \leq \sqrt{2} \delta(t, t'),$$

it follows that

$$\rho^{\delta(t, t')} \leq \rho^{d(t, t')} \leq \rho^{\sqrt{2} \delta(t, t')}.$$ 

Applying Lemma 1 to $\rho^{\sqrt{2}}$ and $\rho$ shows that

$$\int_{\substack{t \neq t' \\
(\neq t', (t', 0, n) \times [0, m]}} \rho^{d(t, t')} \, d\nu(t) \, d\nu(t') = o\left(n^2 m^2\right) = o\left(|D|^2\right),$$

from which

$$\sum_{\substack{t \neq t' \\
(\neq t', (t', 0, n) \times [0, m]}} \rho_{tt'} = \sum_{\substack{t \neq t' \\
(\neq t', (t', 0, n) \times [0, m]}} \rho^{d(t, t')} \leq \int_{\substack{t \neq t' \\
(\neq t', (t', 0, n) \times [0, m]}} \rho^{\delta(t, t')} \, d\nu(t) \, d\nu(t') = o\left(|D|^2\right)$$

and the proof is complete. □