LINEAR DISCRIMINANT ANALYSIS IN IMAGE RESTORATION AND THE PREDICTION OF ERROR RATE

John Haslett
Department of Statistics
Trinity College Dublin
Dublin 2, Ireland

and

Graham Horgan[1]
Scottish Agricultural Statistics Service
University of Edinburgh
Edinburgh EH9 3JZ, Scotland

ABSTRACT

Much attention has recently been focussed on the use of statistical and probabilistic models in the restoration of digital images corrupted by noise. A possibly multivariate "signal" (data) x_i is observed everywhere on a regular finite lattice, and carries information on a small number of unobserved "labels", or colors' c_i at each pixel i, where $i = 0 \ldots N$ labels the sites of a lattice. The objective is to restore the labels (that is, to classify the pixels) $\underline{c} \equiv \{c_i\}$, given the data $\underline{x} \equiv \{x_i\}$. The key idea is that many (or even all) of the signals are used to classify *every* pixel. This is known variously as image segmentation, image restoration or contextual classification. A number of algorithms are now available for this problem. Many of these, though *ad hoc*, are quite satisfactory in practice. But as such it is impossible to conduct any detailed theoretical analysis of their performance, even in terms of something as apparently fundamental as error rate, and restorations must be regarded as equivalent to point-estimates, unqualified in any way. In this paper we show that a simple proposal of Switzer (1980) can be extended to yield a method which is not only perfectly adequate in at least some important cases but admits of fairly detailed analysis of its properties.

---

[1] Supported by the National Board for Science and Technology and ERA Limited.

Keywords: Kriging; Contextual Classification; Correlated Noise; Plug-in Estimates of Error Rate.

1980 Mathematics Subject Classification (1985 Revision). 22-02, 60G35.

1. Introduction

Much attention has been focussed recently on the use of statistical and probabilistic models in the restoration of digital images which have been corrupted in some way; a particular case is where the corruption is by "noise". In this context a signal $\underline{x}_i$ is observed everywhere on a finite regular lattice, and carries information at each pixel on one of a small number of unobserved labels or colors c_i. The objective is to restore the labels $\underline{c} = \{c_i\}$ (that is, to classify each pixel according to the data $\underline{x} = \{x_i\}$). Recent work is concerned with using many (or even, in principle, all) of the signals to classify every pixel; this has become known as "contextual" classification. While the problem has been addressed for many years, initially on the (non-contextual) basis of using only x_i to classify pixel i (Duda and Hart 1973), and subsequently in terms of *ad hoc* smoothers (Switzer 1983), recent work has pursued the idea that the image, and the noise, may be modelled via spatial stochastic processes.

The paper by Geman and Geman (1984) is regarded as seminal; they have pioneered the use of non-causal Markov Random Fields (Besag, 1974) as models of the underlying "true" or uncorrupted image. See also Yu and Fu (1983), Hansen and Elliot (1982), Besag (1986), and Derin and Elliot (1987). Others have used causal Markovian models, which while sometimes more tractable, are essentially asymmetric; see Devijver (1985), Derin et al. (1984), and Haslett (1985). The relaxation methods of Rosenfeld (1978) are motivated by similar ideas, but fall short of being models. Alternative non-Markovian ideas have been proposed by Switzer (1980), Owen (1984), Campbell and Kiiveri (1985), Kent and Mardia (1986) and Haslett and Horgan (1985, 1987). It is from the latter that this paper is derived; its closest parallel is in the work of Switzer (1980).

A number of aspects have emerged from the debate; see the discussion following Besag (1986). Firstly, there is agreement that most "contextual" methods can achieve considerable improvements over *naive* or non-contextual methods. This improvement is clearly most marked when the signal-to-noise ratio is low, and when the corruption exhibits a very simple spatial structure (such as added white noise). There have however been relatively few systematic evaluations on real images; an exception is that of Saebo et al. (1985), who worked with satellite imagery. Switzer (1986) is critical of the extensive use in the methodological literature of artificial examples based on uncorrelated noise.

Secondly, however, there has emerged some disagreement as to how to *measure* the performance of any given restoration method. In particular the natural measure, the proportion of pixels misclassified, while widely criticized as inadequate, is often used in presentation of results. For a discussion on this, see Derin et al. (1984), Geman and Geman (1984), Ripley (1986), Besag (1986), Marroquin (1985, 1987), Titterington (1986) and Switzer (1986). It has been pointed out (Besag (1986), Marroquin (1985)) that algorithms, which have as their objective the minimization of this quantity, are in effect adopting, in a Bayesian context, the criterion of *marginal maximum a posteriori probability* (*marginal MAP*). Thus, one seeks to evaluate, for each pixel i, the posterior probability, given *all* the data $\underline{x}$, $\Pr(C_i = c|\underline{x})$, for each possible c, and to allocate, sequentially for each i, pixels to that c which maximizes this. Others (in particular Geman and Geman (1984) and subsequently others) have suggested a *global MAP* approach, based on simultaneously allocating *all* pixels to that $\underline{c}$ which maximizes $\Pr(\underline{C} = \underline{c}|\underline{x})$. However, Besag (1986) and Greig, Porteus and Seheult (1986) (see also this volume) have pointed out that this goal, while well-intentioned, can yield degenerate solutions in the context of the model most commonly adopted, the discrete Markov Random Field, (MRF). Marroquin (1985) provides graphic examples to support this. In particular,

because of the very long range properties of the discrete MRF (the Ising model is a well known example; see Besag (1973), Bartlett (1975)), the most *globally* probable coloring, being too model dependent, can involve an image comprising only one color! Besag's (1986) Iterated Conditional Modes algorithm, ICM, is partially justified by the fact that it *avoids* the above criterion. He thus makes the point that the MRF model, and the *global* criterion, are to be used with caution. In many ways ICM, and many other similar approaches, are model based only to the extent that the model is used to motivate an algorithm; see Titterington (1986). Thus ICM is not only simple to implement, but is remarkably insensitive to departures from the nominal requirements of the method. It is however, completely impenetrable to any analysis of its properties, in common with most other methods.

It will be observed that whatever the basis adopted, all such algorithms in fact work by computing, iteratively or recursively, a function of the data surrounding each pixel i, and using it to classify pixels, sequentially or in parallel (it matters not, for this simply redefines the function). This classification function we denote as $g_i(\underline{x})$. There are many difficulties in such computations, and that proposed by Geman and Geman (1984), namely *stochastic annealing*, in fact has a random component in the very definition of this function.

Finally, although much attention has been given to exploring these different routes, and some to their comparison in terms of misclassification rates on simulated and occasionally real images, little attention seems yet to have been given to a theoretical investigation of these error rates, or indeed any other performance measure. In what theoretical circumstances are these rates higher or lower? For a given restoration, is there any estimate, from the fitted model, of the error rate? This latter question can be explored by simulation, given the MRF model; see Haslett (1986) and work by Anderson reported elsewhere is this volume. These second-order questions are of course normal, and dominate the conventional classification literature, along with related questions, such as the certainty with which any particular case (pixel) is classified, and whether indeed it should be classified at all. There are many good reasons why such questions have not yet been approached.

Firstly, there remains doubt as to whether the error rate itself is the most useful performance measure; yet it is widely used, and no clear alternative has yet emerged. Secondly, it requires that we discuss the distribution of $g_i(\underline{x})$, conditional on knowing c_i. Given the form of $g_i(\underline{x})$ associated with most algorithms, this is effectively impossible. An exception is the rule proposed by Switzer (1980), wherein $g_i(\underline{x})$ is a linear function of the data in a small neighborhood of i. This is the approach which we shall generalize in the following, and relate to a spatial stochastic model of the observed image. Finally, such discussions are necessarily very reliant on the model proposed; if it is inappropriate, then the results are meaningless. In particular, it is not clear whether the MRF model is appropriate, because of its global properties (Besag 1986). We will deliberately avoid relying on any strongly specified model of the underlying image, relying rather on a non-parametric description.

In the following, therefore, we approach this cautiously, considering only the simple case of an underlying binary image (each pixel has one of only two possible colors). We confine ourselves to additive Gaussian noise. We use bounds and a simple approximation for the error rate for theoretical studies, and we propose a simple "plug-in" rule for the evaluation of a given restoration. We explicitly confine ourselves to an approximation to a *marginal MAP* solution. We rely only on the covariance properties of the underlying image, and make no statements about joint probabilities about sets of pixels, other

than pairs. We do not confine ourselves necessarily to uncorrelated noise, however. Some generalization to images possessing more than two colors is possible, but with difficulty. The method proposed, while claiming no optimality, will be seen to compete satisfactorily with others in these simple cases, as far as the restoration is concerned. It does not generalize in any simple way to the very rich range of hierarchical models currently being tackled by iterative methods based on the MRF. But it does allow some access to the second-order properties mentioned above, which we hope to illustrate, and these properties can serve as a useful baseline against which claims of optimality can be compared. The approach may further be used to identify "outliers", and indeed generalize this concept, as well as to identify pixels where there is insufficient evidence to make a firm classification.

2. A linear solution for a simple case

We consider a binary underlying image, corrupted by noise. We denote the two possible colors by $c = 0$ and $c = 1$; in this underlying image all pixels have intensities that are either μ_0 or μ_1. (Our notation will be such as to indicate which of our arguments can be extended to more than two colors; see also Haslett and Horgan (1987), but for simplicity we confine ourselves here to the binary case). We label the pixels in arbitrary order as $0, 1, \ldots, N$, using a single subscript for notational simplicity. We shall take the true underlying image $\underline{c}$, the observed image $\underline{x}$ and the noise $\underline{\varepsilon}$ as finite portions of realizations of random variables $(\underline{C}, \underline{X}, \underline{E})$, themselves generated by appropriate spatial stochastic processes, such that

$$X_i = \mu_{c_i} + \varepsilon_i \quad i = 0, 1, \ldots, N; \ \underline{E} \text{ and } \underline{C} \text{ are uncorrelated, and } \varepsilon_i \sim N(0, \sigma^2). \quad (2.1)$$

Without loss of generality in the binary case we have taken pixel intensities as scalar; further we take $\mu_0 = 0$ and $\mu_1 = 1$ which allows (2.1) to be written more simply as

$$X_i = C_i + \varepsilon_i \quad (2.2)$$

Suppose further that $\underline{C}$ and $\underline{E}$ are second order stationary processes for which

$$\Pr\{C_i = 1\} = p, \quad \text{and} \quad \mathrm{Cov}\{C_i, C_j\} = \Pr\{C_i = 1, C_j = 1\} - p^2 = k_c(\underline{d}_{ij}) \quad (2.3)$$

where $\underline{d}_{ij}$ denotes the separation of pixels i and j; and

$$\mathrm{Cov}\{\varepsilon_i, \varepsilon_j\} = k_\varepsilon(\underline{d}_{ij}) \quad (2.4)$$

It follows that

$$\mathrm{Cov}(X_i, X_j) = k_x(\underline{d}_{ij}) = k_c(\underline{d}_{ij}) + k_\varepsilon(\underline{d}_{ij}) \quad (2.5)$$

Thus any two of k_x, k_c and k_ε specify the second order properties. Our objective is then to classify pixel i on the basis of some near optimal classification functions $g_{i,c}(\underline{x})$ which are sufficiently simple to allow access to their distributional properties. We will propose below a linear combination of the data $\underline{x}$

$$g_{i,c}(\underline{x}) = \underline{\lambda}_{i;c}^T \underline{x}$$

for some weights $\underline{\lambda}_{i;c}^T$. In principle these weights are defined for *every* pixel; in practice we will in the subsequent take the weights to be zero unless the corresponding pixel is close to pixel i. We will see that such functions can be motivated in a number of

ways including Linear Discriminant Analysis (LDA) and, for binary underlying images, kriging and Wiener Filters. In combination with the Gaussian assumption (ii) above, reasonable approximations can thus be made for probability statements concerning the function $g_{i;c}(\underline{x})$. Note that in the binary case it will be sufficient to confine attention (for the purpose of the allocation rule) to

$$g_i(\underline{x}) = g_{i;0}(\underline{x}) - g_{i;1}(\underline{x}) = \underline{\lambda}_i^T \underline{x} \tag{2.6}$$

2.1 Spatial linear discriminant analysis

Recall that in general LDA may be motivated in a number of closely interrelated ways: viz. (i) the Mahalanobis distance of a given observation to the mean $\underline{\mu}_k$ of each of k possible populations, assuming a common (or, in practice, pooled) covariance; (ii) maximum likelihood for multivariate normal populations, and the closely related Bayesian minimum error rate; and (iii) in the case of two populations only, Fisher's Discriminant Function and prediction via multiple linear regression on an appropriately defined indicator variable. See, for example, Seber (1984) Section 6.9, Mardia et al. (1979), Chap. 11, and Flury and Riedwyl (1985). We will show how our spatial problem may conveniently be formulated as (i) and (iii) which are non-parametric. We will show also that analogies with (ii) may be made as a good approximation to yield the second-order properties we desire. Haslett and Horgan (1985) have sketched approach (iii) for binary data; an outline of the more general arguments has been presented in Haslett and Horgan (1987).

The neighborhood. Let us confine our attention to a single pixel, which we label 0; we define an n-neighborhood, V_n, of this pixel as all pixels within $\pm n$ (vertically and horizontally) of pixel 0, and let $m = (2n+1)^2$. We shall ignore for the moment the question of edge-effects, although they pose no more than notational and computational inconveniences, and we leave open the choice of n, which can, in theory, be arbitrarily large. Let the labelling be such that pixels with labels exceeding m are not in V_n. Clearly the ideas to follow can be translated, by second order stationarity, to any other pixel.

Let $\underline{X}^{V_n} = (X_0, X_1, X_2, \ldots, X_m)^T$ denote the random variable of which $\underline{x}^{V_n}$, the totality of the signal information in V_n, is a realization. We shall be interested in the second-order properties of $\underline{X}^{V_n}$, particularly when we condition on $C_o = c$. In particular we note that:

(i) $E[\underline{X}^{V_n}] = p\underline{1}_m$ where $\underline{1}_m$ is an m- dimensional vector of 1's;

(ii) $\Sigma_T = \mathrm{Cov}\{\underline{X}^{V_n}\}$ has as its (i,j)th elements $\sigma_{ij;T} = k_x(\underline{d}_{ij})$;

(iii) $E[\underline{X}^{V_n}|C_0 = c] = \underline{\mu}_c$ has as its ith elements $\mu_{i;c}$ where

$$\begin{aligned} \mu_{i;c} &= p - k_c(\underline{d}_{i0})/(1-p) \quad \text{if} \quad c = 0 \\ &= p - k_c(\underline{d}_{i0})/p \quad \text{if} \quad c = 1 \end{aligned} \tag{2.7}$$

and

(iv) while $\Sigma_{w,c} = \mathrm{Cov}[\underline{X}^{V_n}|C_0 = c]$ is not available, the pooled within group variance covariance matrix $\Sigma_w = (1-p)\Sigma_{w,0} + p\Sigma_{w,1}$ has as its (i,j)th element

$$\sigma_{ij;w} = k_x(\underline{d}_{i,j}) - k_c(\underline{d}_{i,0})k_c(\underline{d}_{j,0})\{p(1-p)\}^{-1} \tag{2.8}$$

In classical terminology, Σ_w is the (pooled) within-group variance-covariance matrix, and Σ_T is the total variance-covariance matrix.

We note that unless all of the pixels within V_n have the same colour, $\underline{X}^{V_n}$ does *not* follow a multivariate Normal distribution. Non-parametric linear discriminant functions of themselves make no such requirement; we shall invoke the multivariate Normal distribution, however, as a means of obtaining approximate second order information.

Linear discriminant function. Classical results show that a rule, based on comparing the Mahalanobis distance of a given $\underline{x}^{V_n}$ to the two means $\underline{\mu}_c^{V_n}$, with respect to Σ_w, leads to comparing the two functions

$$D_c^2 = (\underline{x}^{V_n} - \underline{\mu}_c^{V_n})^T \Sigma_w^{-1} (\underline{x}^{V_n} - \underline{\mu}_c^{V_n}) \tag{2.9}$$

for $c = 0$ and 1. This is equivalent to a decision rule based on

$$g_0(\underline{x}^{V_n}) = \underline{\lambda}_n^T \underline{x}^{V_n} \tag{2.10}$$

where

$$\underline{\lambda}_n = \Sigma_w^{-1}\{\underline{\mu}_1^{V_n} - \underline{\mu}_0^{V_n}\} = \Sigma_w^{-1}\underline{k}_n\{p(1-p)\} \tag{2.11}$$

where $\underline{k}_n$ has $k_c(\underline{d}_{i,0})$ as its ith element.

This is our desired linear combination. The above arguments can be generalized straightforwardly for the case of more than 2 colors (Haslett and Horgan 1987). It is insightful however to view (2.10) from a prediction point of view.

Prediction. If we seek that linear combination $\widehat{C}_0 = \alpha_n + \underline{\beta}_n^T \underline{x}^{V_n}$ of the data in V_n which minimizes $E\{(\widehat{C}_0 - C_0)^2\}$ we find, after routine algebra which closely follows indicator kriging, (Haslett and Horgan 1987, Ripley 1981, Journel and Huigbregts 1978) that:

$$\underline{\beta}_n = \Sigma_w^{-1}\underline{k}_n \quad \text{and} \quad \alpha_n = p(1 - \underline{\beta}_n^T \underline{1}_m) \tag{2.12}$$

This approach is akin to performing 2-group LDA by regression on an appropriately defined indicator variable (Flury and Riedwyl 1985). That $\underline{\beta}_n$ and $\underline{\lambda}_n$ are identical to within a multiplicative constant follows from classical results. It is also clear that (2.12) is simply a finite version of the Wiener filter.

Operationally (2.10) amounts to smoothing the image by a linear filter. In principle adjustments must be made near the edges, and the formulation of (2.12) in particular shows that there are no difficulties in predicting from other than a symmetric neighborhood. We ignore such details here. Following the smoothing, allocation is performed by contrasting $g_i(\underline{x}^{V_n})$ with some threshold to be determined. We pursue this below in the context of minimizing the mis-classification rate. In the subsequent we shall omit explicit reference to the neighborhood size n.

2.2 Misclassification rate.

It follows from the above that:

$$\min E[(\widehat{C}_0 - C_0)^2] \equiv \widehat{\sigma}^2 = p(1-p) - S \tag{2.13}$$

where $S = \underline{\beta}^T \underline{k}$,

$$E[\widehat{C}_0 | C_0 = 1] \equiv m_1 = p + S/p$$

and

$$E[\widehat{C}_0|C_0 = 0] \equiv m_0 = p - S/(1-p) \tag{2.14}$$

and finally, though $\widehat{\sigma}_c^2 = \text{Var}[\widehat{C}_0|C_0 = c]$ is not available, the pooled variance $\widetilde{\sigma}^2 = (1-p)\widehat{\sigma}_0^2 + \widehat{\sigma}_1^2$ is given by

$$\widetilde{\sigma}^2 = S[1 - S\{p(1-p)\}^{-1}] \tag{2.15}$$

If we further assume that, to a reasonable first approximation, the conditional distributions of $\widehat{C}_0$ are approximately Normal, with common variance $\widetilde{\sigma}^2$, we have that the allocation rule *"allocate to colour 1 if $\widehat{C}_0 > h$; else allocate to colour 0"* has misclassification rate of:

$$\begin{aligned} \text{PIC}_n &= \Pr\{\text{Incorrect classification; } h, n\} \\ &= (1-p)\Phi\{(h - m_0)/\widetilde{\sigma}\} + p\Phi\{(m_1 - h)/\widetilde{\sigma}\} \end{aligned} \tag{2.16}$$

which is minimized, wrt h, when

$$h = \gamma + S\{p(1-p)\}^{-1}(1/2 - \gamma) \tag{2.17}$$

where $\gamma = p + p(1-p)\ln\{(1-p)/p\}$

At this threshold, the rate is

$$\text{PIC}_n = (1-p)[1/2D_{01;n}^{-1} + D_{01,n}\ln\{(1-p)/p\}] + p[1/2D_{01;n}^{-1} + D_{01;n}\ln\{(1-p)/p\}] \tag{2.18}$$

where $D_{01;n}^2 = (m_1 - m_0)^2/\widetilde{\sigma}_n^2$ can be shown to be the Mahalanobis distance between $\underline{\mu}_0^{V_n}$ and $\underline{\mu}_1^{V_n}$ *wrt* Σ_w. We discuss in Section 3 the approximation implicit here, and possibilities for improving it.

Thus, subject to the Normal approximation above, we have from (2.14, 2.15) that for a specific pixel

$$\text{Odds}(C_0 = 1|\widehat{C}_0 = \widehat{c}_0) \propto \exp\{\widehat{c}_0/\widetilde{\sigma}^2\} \tag{2.19}$$

the familiar logistic function. Equivalently, from (2.10, 2.11) we have that

$$\text{Odds}(C_0 = 1|\underline{x}^{V_n}) \propto \exp\{g_0(\underline{x}^{V_n})\}$$

which emphasizes the interrelationships between the alternative approaches behind (2.10) and (2.12).

It is seen then that the proposed allocation rule is approximately *marginally MAP* and is based on the classification function $g_0(\underline{x}^{V_n}) = \widehat{c}_0/\widetilde{\sigma}^2$. This may be contrasted with that from Besag's ICM which yields, at convergence, a classification function which can be expressed as $g_0^{\text{ICM}}(\underline{x})$, where

$$\text{"Odds"}(C_0 = 1|\underline{x}) \propto \exp(g_0^{\text{ICM}}(\underline{x})) \tag{2.20}$$

with $g_0^{\text{ICM}}(\underline{x}) = x_0/\sigma^2 + \beta\widetilde{v}$, the quotes indicating that the ultimate classification criterion is not in fact an odds ratio conditional only on $\underline{x}$. Here β is the parameter of the MRF (see Besag, 1986), and $\widetilde{v}$ is the number of nearest neighbours of pixel 0 of colour 1 at the *previous* iteration. That iteration of course drew on the data at their neighbours, including of course pixel 0, and so on back through the iteration. If s iterations are

used, then that MRF model which is based on 8 nearest neighbours, effectively uses $\underline{x}^{V_s}$ rather than $\underline{x}$ in defining the classification function $g_0^{\mathrm{ICM}}(\ )$.

In the subsequent sections of this paper we evaluate these suggestions, and discuss other operational aspects.

2.3 Implementation issues. We outline here some of the practical issues to be faced in implementing the foregoing. Some of these, although interesting, will be given only passing consideration due to considerations of space.

Firstly, from where might one obtain the key information needed - namely m_0 and m_1 (taken to be 0 and 1 respectively above), p, and any two of $\{k_x(\ ), k_c(\ ), k_e(\ )\}$?

In theoretical studies we may take these as known; $k_c(\ )$ in this case is the empirical covariance function for the given *clean* image. Here the interest lies in investigating, as a theoretical issue, the extent to which increasing the noise auto-correlation, for example, affects PIC. There is currently a complete lack of theoretical tools with which to study even such basic issues. We briefly outline some results in Section 4.

But the issue of how to estimate such parameters, and of the resulting impact on accuracy given, *in extremis*, only the image, is a critical question which has been taken up by many authors in image segmentation. Of course the issue has been extensively researched in a non-spatial context, and the fact that the method proposed above is so close to classical discriminant analyses suggests that there may be considerable room for some further theoretical analyses here; see for example the not unrelated study of Lawoko and McLachan (1985). The key requirements are: (i) estimates of the class means μ_0 and μ_1, for from these, and a knowledge that the noise process is zero mean, follow good estimates of p, given that N is typically very large; and (ii) an estimate $\widehat{k}_\varepsilon(\ )$ of the covariance structure in the noise process (taken to be uncorrelated with the underlying image) for from this, and an empirically obtained estimate $\widehat{k}_x(\ )$ from the given image, follows $\widehat{k}_c(\ )$. If the noise process is known to be spatially uncorrelated, then this is straightforward. If not, then the autocorrelations of $\underline{E}$ and $\underline{C}$ will be impossibly confounded. This issue is discussed in geostatistics as "structural analysis" (Journel and Huigbregts 1978), and is not a simple issue. The reader will note that this problem remains unresolved in the general image segmentation literature, for the estimation procedures that have been discussed normally assume that the signals $\underline{X}$, conditional on the underlying image, are independent. Switzer (1986) has pointed out that this is unrealistic in certain cases; examples include satellite imagery, where the "noise" is in fact often "unimportant detail" rather than atmospheric and other interference.

It is interesting to note however, that the natural emphasis is on the means, μ_0 and μ_1, and on the noise process. The need to model the underlying image reduces, given stationarity, to a non-parametric description of its second order properties. If it is known to have a Markovian representation, then *in principle* a parametric description follows. (In practice of course this is an unsolved mathematical problem). In the subsequent therefore we confine ourselves to the illustration of some of the consequences of estimating all parameters, bar μ_0 and μ_1; we confine this illustration to the case where the noise is iid.

The second major issue is that of choice of neighborhood Vn. Recall that an iterative method such as Besag's ICM, with s iterations, effectively uses a neighborhood of size $(2s + 1)^2$. For low signal-to-noise ratios - in this case measured by $\{p(1 - p)\}/\sigma^2$-convergence with ICM is often achieved in 6 or so iterations; with a high ratio, convergence can be almost immediate. In our case we shall *presume* that a limited

neighborhood is sufficient. This can be measured via PICn, as in (2.18) or equivalently via $\tilde{\sigma}^2$ in (2.15), and it is easy to show empirically that these converge rapidly with n, providing guidance in choice of n. A more interesting approach is to select a reasonably large n^* at the outset, and to use subset methods on equation (2.12) to select an appropriate subset of the $m^* = (2n^* + 1)^2$ potential discriminators. In combination with an exploitation of the symmetry involved in this spatial application, (pixels i and j must have *equal* discriminatory power, and coefficients, if $\underline{d}_{i0} = -\underline{d}_{j0}$ (and if isotropy is assumed, if $|\underline{d}_{i0}| = |\underline{d}_{j0}|$)) this can reduce the numbers of pixels (or combinations thereof with a *priori* identical coefficients) to a very small number. However space does not permit us to develop this argument here; see Haslett and Horgan (1985) and Haslett (1989). Similarly, the handling of edge effects is entirely straightforward, but for simplicity here we ignore this entirely, and treat our image as though surrounded by an uninformative strip, in which all the pixels have intensity h, the threshold.

3. Bounds for the misclassification rate

In the above, the approximate expressions for PIC (2.16, 2.18) are in fact remarkably satisfactory, as we shall see in Section 4. In this section we explain why this is so, and indicate how formal bounds may be obtained for PIC. Given that the classification method of Section 2 is sub-optimal (restricted as it is to linear combinations of the observed data), the main interest for practical purposes lies in the *upper* bound, for this bound is then generally relevant to all restoration algorithms pertaining to this problem. We adopt the notation of (2.12) and drop explicit reference for the moment to the neighborhood size n, for simplicity. We note that:

$$\hat{C}_0 = \alpha + \underline{\beta}^T\{\underline{C}^V + \underline{\varepsilon}^V\} = Z + Y \tag{3.1}$$

where $Y = \underline{\beta}^T\underline{\varepsilon}^V \sim N(0, \sigma_y^2)$, with $\sigma_y^2 = \underline{\beta}^T\Sigma_\varepsilon^V\underline{\beta}$, the elements of Σ_ε^V being available from $k_\varepsilon(\underline{d}_{ij})$, adn $Z = \alpha + \underline{\beta}^T\underline{C}^V$. Letting $Z_i \stackrel{D}{=} (Z|C_0 = i)$ we have, for a given threshold h, that

$$\text{PIC} = 1 - p - (1-p)\Pr(Z_0 + Y \le h) + p\Pr(Z_i + Y \le h) \tag{3.2}$$

Our task below is to bound these probabilities, given the Normal distribution of Y, and the available information on the distributions of the Z_i. This is that

(i) $EZ_i = m_i$ from (2.14)

(ii) $(1-p)\operatorname{Var}(Z_0) + p\operatorname{Var}(Z_1) = \tilde{\sigma}_y^2 - \sigma_y^2 \equiv \sigma_z^2$ from (2.15)

(iii) $Z_0 \in [a_0, b_0]$, where $a_0 = \alpha + \sum_{i \neq 0} \min(0, \beta_i)$ and $b_0 = \alpha + \sum_{i \neq 0} \max(0, \beta_i)$ and

(iv) $Z_1 \in [a_1, b_1]$, where $a_1 = \beta_0 + a_0, b_1 = \beta_0 + b_0$.

Our approximations and bounds are therefore essentially concerned with:

$$\Pr(Z_i + Y \le h) \equiv \underset{Z_i}{E}[\Phi\{(h - Z_i)/\sigma_y\}] \tag{3.3}$$

and with the approximation or bounding in (a_i, b_i) of $F(z) \equiv \Phi\{(h-z)/\sigma_y\}$.

Many approximations in this range are excellent for, since $\sigma_y^2 > \sigma_z^2$ in practice, the exact form used for the distribution of Z_i is not of critical importance. In particular the approximation implicit in the statement that $Z_i \sim$ Normal, as used in (2.16), is very adequate for this reason. Quadratic approximations for $F(\cdot)$ yield predictions almost

indistinguishable from those derived with this Normal approximation. We may however go further and formally bound $F(\cdot)$ above and below by suitably simple functions, and obtain formal bounds for PIC. We illustrate this with the linear bounds, and concentrate on the upper bound.

3.1 A linear upper bound.

Lemma.

$$\text{PIC} \leq \text{PIC}^+ \equiv 1 - p - (1-p)\pi_0^- + p\pi_1^+ \tag{3.4}$$

where

$$\begin{aligned} \pi_0^- &= F(m_0) \quad \textit{if} \quad b_0 > a_0 \geq h \\ &= (a_0 - c_0)^{-1}[(a_0 - m_0)F(c_0) + (m_0 - c_0)F(a_0)] \qquad \textit{otherwise} \end{aligned}$$

where

$$\begin{aligned} c_0 &= \min(h - \sigma_y \eta(a_0), b_0) \qquad \textit{and} \\ \pi_1^- &= F(m_1) \quad \textit{if} \quad b_0 > a_0 \geq h \\ &= (b_1 - c_1)^{-1}[(b_1 - m_1)F(c_1) + (m_1 - c_1)F(b_1)] \qquad \text{otherwise} \end{aligned}$$

where $c_1 = \max(h - \sigma_y \eta(b_1), a_1)$.

Here $\eta \equiv \eta(\tau)$ satisfies $\phi(\eta) = \{\Phi(\tau) - \Phi(\eta)\}/(\tau - \eta)$, where $\Phi(\cdot)$ is the Normal Distribution Function, and ϕ its derivative; see Figure 1.

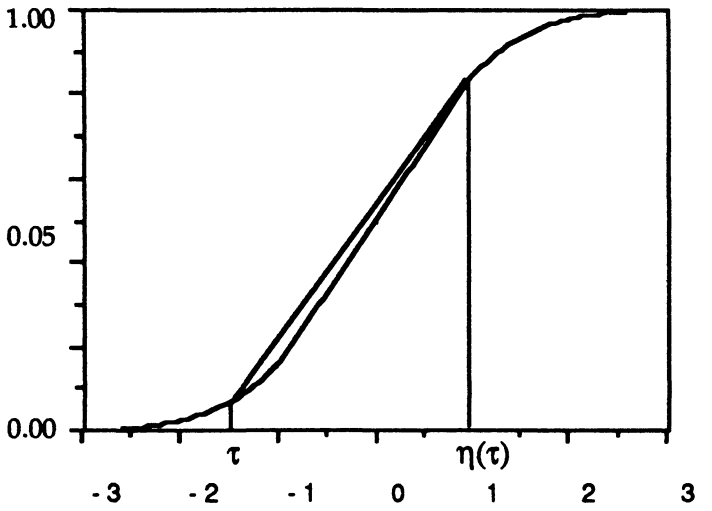

Figure 1. Definition of $\eta(\tau)$.

Proof.

$$\Pr(Z_0 + Y \leq h) = \underset{z_0}{E} F(Z_0)$$

If $a_0 \leq h$, $F(z)$ is concave in (a_0, b_0), and by Jensen's inequality

$$\underset{z_0}{E} F(Z_0) \geq F\{E(Z_0)\} = F(m_0)$$

If $b_0 \leq h, F(z)$ is convex in (a_0, b_0) and

$$F(z) \geq (a_0 - b_0)^{-1}[(a_0 - z)F(b_0) + (z - b_0)F(a_0)] \text{ in } (a_0, b_0)$$

If $a_0 \leq h < b_0$ however, then

$$F(z) \geq (a_0 - d_0)^{-1}[(a_0 - z)F(d_0) + (z - d_0)F(a_0)] \text{ in } (a_0, b_0)$$

where $d_0 = h - \sigma_y \eta(\alpha_0)$.

If $b_0 \leq h$ however, then $b_0 \leq d_0$; we therefore write $c_0 = \min(b_0, d_0)$; thus if $h > a_0$, these two expressions lead, on taking expectations *wrt* Z_0, to the result,

$$\underset{z_0}{E} F(Z_0) \geq (a_0 - c_0)^{-1}[(a_0 - m_0)F(c_0) + (m_0 - c_0)F(a_0)]$$

The proof for the upper bound π_0^+ follows similarly; equally we could obtain a lower bound for PIC, were that of interest.

Finally, we note that (3.4) defines a *family* of upper bounds, for neither the threshold h nor the neighborhood size n have been specified. In practice however, neither (2.16) nor (3.4) exhibit any sensitivity to variations in h near the optimal value; but for large neighborhoods PIC^+ can be very pessimistic, for the events corresponding to the upper and lower extremes $Z_0 = b_0$ and $Z_1 = a_1$, respectively, correspond to very unlikely events, such as a black pixel being completely surrounded in a large neighborhood by white pixels. If we accept that PIC is a decreasing function of the separation $D^2_{01;n}$, then if $N = \{n_0, n_1, n_2, \ldots\}$ defines a set of neighborhoods with increasing $D^2_{01;n}$, we may with advantage define $\text{PIC}^+ = \min_{n \in N} \text{PIC}_n^+$. It is rather easy in practice to define such a set of neighborhoods, in the context of the subset selection methods referred to in Section 2, and we use this below to tighten our upper bounds.

4. Prediction of misclassification rate-Results

In this section we evaluate the predictions of PIC made in Sections 2 and 3. For this purpose we have used a simulation approach on two test images CAT and CHESS; see Figures 2 and 3. The CAT image is similar to other test images used in this field; the CHESS image is quite artificial, and is designed to allow the exploration of an extreme case.

In each case the image is a 50×50 array, rather smaller than is likely to be encountered in practice, and consequently somewhat more prone to edge-effect problems. To these arrays was added an array of simulated Gaussian "noise", with zero mean, specified variance σ^2 and correlation structure as below. The correlated noise was generated by the *turning bands* method (Ripley 1987); in this method 1-dimensional realizations of correlated noise are generated along each of a number of "spokes", and combined to yield noise with a known 2-dimensional correlation structure.

For our testing purposes we have generated AR-1 noise on each spoke, with 1-dimensional correlation structure at lag s given by $\rho_1(s) = \exp(-\rho s)$ and corresponding 2-dimensional structure

$$\rho_2(s) = (2/\pi) \int_0^1 \exp(-\rho v)(1 - v^2)^{-1/2} dv$$

Clearly at $\rho = 0$ we have iid Gaussian noise. This function decays quite slowly with distance; for example, at $\rho = 0.2$, $\rho_2(1) = 0.41$ and $\rho_2(5) = 0.08$.

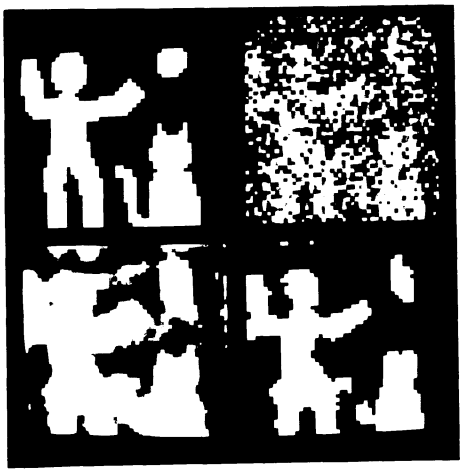

Figure 2. (a) CAT image; (b) *plus* noise ($\sigma = 1$, $\rho = 0$); (c) $\hat{C}$; (d) reconstruction by linear method (error rate = 7.9%)

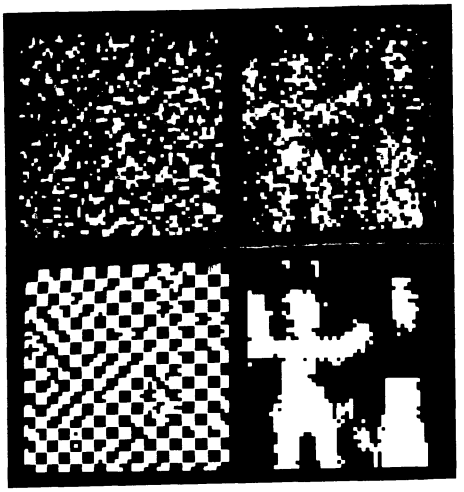

Figure 3. (a) CHESS image *plus* noise ($\sigma = 1$, $\rho = 0$); (b) CAT image *plus* noise ($\sigma = 1$, $\rho = 0$); (c) Error rate 10.2%; (d) Error rate 9.5%

A number of experiments are reported below for different combinations of image and (ρ, σ^2). In particular we consider two basic experiments. In the first, in which we justify the model, we shall presume that the all parameters are known; this presumes knowledge of the second order structure of the underlying *true* image; see Figure 2 for an example. We shall see that in this case the method, as a classification method, works well in comparison with Besag's ICM recipe. But further, the predictions of accuracy are almost exactly as given in (2.16), and the bounds from Section 3 are tight. Our main thesis will thus be vindicated, and it becomes possible to make some general statements about precision. In the second, oriented at rather more practical applications, we shall presume the existence of a single noisy image and knowledge only of the correlation

structure of the *noise*, and its independence from the underlying image, together with the mean signal for each colour pixel; the correlation structure of the image, and the parameters p and σ^2 can then be estimated, and used, with (2.16) to yield *plug-in* estimates of PIC. We illustrate this only with iid noise; see Figure 3. In principle the method can be used in general, but in practice it is impossible, from a single image, to separate the correlation structure of the noise and the underlying image. We will see that, in common with such estimates in classical LDA, the resulting predictions are somewhat biased.

Table 1 sets out the results of 10 replications for each combination. Edge effects were handled for simplicity by assuming the existence of a border with uniform uninformative signal equal to the threshold h. A 9×9 neighborhood was used in each case, and experiments indicated no measurable advantage in using a larger neighborhood. Besag's ICM was implemented for comparison, for the CAT image; it is not naturally adapted to the CHESS image (although this can be done, by appropriate definition of *nearest neighbor*). In this implementation ICM's β parameter sequentially assumes the values .2 and 1.7 in steps of .3; as this involves 6 iterations ICM is thus using in practice a neighborhood of 13×13. With this β-schedule ICM is suboptimal, for the β parameter could in principle have been fitted to the *true* image; our experience is that this makes remarkably little difference, and in any case does not effect our conclusions. Additionally, no attempt was made to adapt ICM to the case of correlated noise, and the conventional recipe was followed even in such cases, with remarkably good results as will be seen. The standard errors derived from the 10 replications are reported also.

The conclusions from the first set of experiments (see Table (1) are that (a) the linear method, though crude, achieves accuracies (as measured by PIC) as good as ICM; (b) the predictions of PIC are excellent in all cases, and the upper bound is quite tight; and (c) ICM performs quite remarkably well, in comparison with this bound, given our implementation, and given that it is not even intended to minimize PIC. Additionally the method predicts a decline in accuracy with increasing s2 in both cases, which is confirmed, but only in the CAT image with increasing ρ, which is also confirmed.

The second set of experiments are reported in Table 2; here 5 replications were used; the predictions and the upper bound now vary with replication, as does accuracy. As anticipated, the predictions are optimistically biased. ICM, implemented as above, generally yields better accuracy, but of course no prediction of accuracy. Had the β parameter been fitted "on the fly" to the noisy image as per Besag's recipe, the result for ICM would be similar.

The actual reconstructions generated by the two methods differ considerably in "texture", with the linear method (particularly with high s and with covariance structure estimated from noisy images) generating rather more "speckled" images. This is a natural consequence of (i) adopting a *marginally* rather than *globally MAP* approach and (ii) basing the algorithm only on the *pairwise* joint or conditional probabilities of pixel colour, rather than, as in ICM (or any MRF based procedure) on conditional probabilities, given all 8 neighbours. Indeed if, a *priori*, isolated pixels of a given colour are known to be very unlikely, *post- smoothing* (Switzer 1983) a restoration, generated by this linear method, can be advantageous. Many such smoothers are available; indeed ICM is one such. Significant improvements are available in such cases, though not for ICM generated reconstructions, which are already smooth. For example the linear method followed by a simple *opening* (Serra 1982) yields an image very little different in appearance or error rate, though somewhat in detail from that generated by ICM. However, the blind application of such smoothers has effects which cannot be

predicted, and would clearly be quite inappropriate for images such as CHESS, and even in advantageous cases no prediction of decrease in error rate is possible.

5. Conclusions

In this paper we have investigated linear methods for restoring binary images degraded by additive Gaussian noise. Our objectives have not been to obtain restorations that are "better" - in terms of accuracy of reconstruction, an ill-defined term in any case - but to produce methods which, while adequate in this respect, are simple enough to admit of some analysis of their properties. For we share the belief of Titterington (1986) that there is "little to choose visually" between many different methods of such reconstruction, and his concern that we seek specifically *statistical* contributions to this rapidly growing field.

In this spirit we have employed the classical statistical tool of Linear Discriminant Analysis to a simple version of this problem, and have shown that some of its properties are attractive. In particular we have shown that it is possible not

Table 1
Achieved and Predicted PIC% Parameters Known

CAT image		$\sigma = 0.5$		$\sigma = 1.0$	
	ρ	PIC	(SE)	PIC	(SE)
Pred	0.0	1.74		7.43	
Upper		3.43		7.46	
Linear		2.94	(0.13)	7.93	(0.24)
ICM		3.11	(0.17)	9.27	(0.33)
Pred	0.2	8.90		22.35	
Upper		10.15		22.94	
Linear		8.62	(0.36)	22.26	(0.93)
ICM		8.13	(0.37)	22.98	(0.95)
Pred	0.4	11.68		25.30	
Upper		12.83		25.99	
Linear		12.00	(0.54)	26.75	(1.45)
ICM		11.43	(0.83)	28.13	(1.46)
CHESS image					
	ρ				
Pred	0.0	0.77		8.10	
Upper		2.16		10.14	
Linear		1.42	(0.089)	7.77	(0.24)

Pred	0.2	0.46		7.75	
Upper		1.75		8.01	
Linear		1.83	(0.089)	8.23	(0.43)
Pred	0.4	0.15		5.76	
Upper		1.46		6.14	
Linear		1.93	(0.089)*	6.21	(0.14)

Notes: (i) SEs based on 10 replications.
(ii) Edge effect just measurable at case *; PIC internally to image is $\sim$ 1.50%; on the edge it is 10.50%.

only to achieve reconstructions which compare well with other methods, in particular with ICM, but also to predict pixel misclassification rates with surprising accuracy, even in circumstances normally avoided in such methods, namely that of *correlated* noise. The method has been shown to be capable of handling images of widely differing texture, and correctly predicting that which was initially surprising, that increasing the correlation in the noise does not always lead to an increase in error rate. Additionally, since formal upper bounds for error rates can now be determined, for a method which makes no claim to optimality, standards are now available for methods claiming optimality. By these standards ICM is seen to perform remarkably well even in circumstances departing substantially from its nominal requirements.

However, as Critchley (1986) points out, our main practical interest is not performance in some theoretical situation, but *for the given image*. Simple *plug-in* rules for the error rate have been proposed, and have been shown as elsewhere to be useful, though somewhat optimistically biased. These may in principle be contrasted with those of Anderson reported elsewhere in this volume. Additionally pixels whose classification cannot, in honesty, be anything other than guesswork (typically those on the edges of regions) can be identified as such since we have posterior probabilities which may reasonably be interpreted as such (equation (2.19)). But also, we have some possibilities for checking whether the model is indeed valid for every part of the given image, in that it is possible to refuse to classify a pixel whose immediate neighborhood seems unreasonable in the context of the image as a whole. For we note that D_c^2 as defined in (2.9) has, under the same normal approximation, a distribution which is approximately χ_m^2, if pixel 0 is of colour c. We may then

Table 2
Achieved and Predicted PIC Parameters Estimated Independent Noise

	$\sigma = 0.5$			$\sigma = 1.0$		
	Predicted	Upper	Achieved	Predicted	Upper	Achieved
CAT						
Linear	1.33	3.28	3.29	5.51	6.27	8.76
SE	0.28	0.14	0.14	1.14	1.03	1.13
ICM			3.11			9.27
SE			0.17			0.33

CHESS						
Linear	0.84	2.53	1.90	5.35	10.24	11.46
SE	0.37	0.42	0.13	1.24	0.57	0.29

Note: SEs based on 5 replications.

determine, by comparing $\min\{D_0^2, D_1^2\}$ with this distribution, whether a pixel is conceivably of either colour. If it is we may allocate it by comparison of D_0^2 and D_1^2, or equivalently by comparing $\widehat{C}_0$ with the threshold h, and return, via (2.19) a posterior probability of its being colour 1, together with an estimate of the overall misclassification rate from (2.18), if it is not, we simply return it as an "outlier". These steps are routine in discriminant analyses studies elsewhere. The reader is referred to Haslett and Horgan (1987) for examples of the application of these ideas to identifying *Doubters* and *Textual Outliers*. In other words it is possible to produce reconstructions which are, as is normal in any other statistical exercise, appropriately qualified by reference to a simple but adequate model.

In principle the method can be extended to images with more than two colors; the extension from Section 2 is straightforward, and is outlined in Haslett and Horgan (1987); even the misclassification rates can be predicted, though now from higher dimensional Normal distributions. In practice however the non-parametric nature of the description of the underlying image becomes too complicated to sustain, particularly when this must be estimated from a given noisy image. But given an appropriate parameterization of the *covariance* structure of such an image (which is regrettably unavailable from models such as the Markov Random Field, but which can be achieved within the general class of *coverage models* (Narendra and Schachter 1983) progress may be possible also.

The model, being linear, can readily be extended, at least in principle, to many other areas of interest. A particular case is that in which the noise is correlated with the underlying image. For given the cross-correlation, (2.5) may easily be generalized. In practice however it is difficult to envisage how, from a single noisy image, one might isolate the separate components of the spatial variability. Given the knowledge that the noise is independent of the given underlying image, it is possible to envisage studies aimed at the analysis of the noise *per se*. It should be noted that many of the examples discussed, for example in the discussion of Besag (1986) pertain to remote sensing of the Earth. Here there is often very little noise as such (except as in cloud, an extreme example of correlated noise), and the problem is one of 'unimportant detail'.

The main remaining, and as yet insurmountable, difficulty is that error rate is but one measure of performance, and sometimes a poor one at that.

Acknowledgements

This paper has benefitted greatly from the advice of many people since the first version was written. In particular, Norm Campbell, Kanti Mardia, John Kent, Adrian Raftery, Julian Besag, Don Geman and Paul Switzer deserve mention.

References

Bartlett, M. S. (1975). *The Statistical Analysis of Spatial Pattern.* Chapman and Hall.

Besag, J. (1986). On the Statistical Analysis of Dirty Pictures (with Discussion). *Journal of the Royal Statistical Society* B **48**, 259-302

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion). *Journal of the Royal Statistical Society* B **36**, 192- 236.

Campbell, N. A., & Kiiveri, H. T. (1985). *Allocation of Remotely-Sensed Data*, CSIRO Technical Report.

Critchley, F. (1986) *Journal of the Royal Statistical Society* B **48**, 286 -287, in discussion of Besag.

Derin, H., Elliot, H., Christi, R., & Geman, D. (1984). Bayes Smoothing Algorithms for Segmentation of Binary Images Modelled by Markov Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*-**6**, 707-720.

Derin, H., & Elliot, H. (1987). Modelling and Segmentation of Noisy and Textured Images using Gibbs Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*-**9**, 39-55.

Devijver, P. A. (1985). Probabilistic Labelling in a Hidden Second-Order Markov Mesh. Gelsema, E. S., & Kanal, L. N. (1985, eds.), *Pattern Recognition in Practice II*, North Holland, 13-124.

Duda, R. O., & Hart, P. E. (1983). *Pattern Classification and Scene Analysis*, Wiley.

Flury, B. N., & Riedwyl, H. (1985). T^2 - tests, the Linear Two-Group Discriminant Function and their Computation by Linear Regression. *The American Statistician* **39**, 20-25.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*-**6**, 721-741.

Greig, D. M., Porteus, B. T., & Seheult, A. H. (1986). *Journal of Royal Statistical Society* B **48**, 282-84, in discussion of Besag; see also this volume.

Hand, D. J. (1981). *Discrimination and Classification.* Wiley.

Hansen, F. R., & Elliot, H. (1982). Image Segmentation using Simple Markov Field Models. *Computer Graphics and Image Processing* **20**, 101-132.

Haslett, J. (1986). *Journal of the Royal Statistical Society* B **48**, 292 in discussion of Besag.

Haslett, J. (1988). *Geostatistical Neighborhood and Subset Selection.* Amstrong, Y. (1988, ed.), *Geostatistics*, Volume 2, Kluwer Academic Publishers, 569–577.

Haslett, J. (1985). Maximum Likelihood Discriminant Analysis on the Plane, Using a Markovian Model of Spatial Context. *Pattern Recognition* **18**, 287-296.

Haslett, J., & Horgan, G. (1985). *Spatial Discriminant Analysis - A Linear Discriminant Function for the Black/White Case.* Technical Report, Department of Statistics, Trinity College, Dublin.

Haslett, J., & Horgan, G. (1987). Linear Models in Spatial Discriminant Analysis. Devijver, P. A., & Kittler, J. (1987, eds.), *Pattern Recognition: Theory and Practice*, Springer-Verlag, 47-56.

Journel, A. G., & Huigbregts, C. J. (1987). *Mining Geostatistics.* Academic Press.

Kent, J. T., & Mardia, K. V. (1986). Spatial Classification using Fuzzy Membership Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*-**10**, 659–671.

Lawoko, C. R., & McLachan, G. J. (1985). Discrimination with Autocorrelated Observations. *Pattern Recognition* **18**, 145-149.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press.

Marroquin, J. L. (1985). *Probabilistic Solution of Inverse Problems.* MIT Ph.D. Thesis, LIDS-TH-1500.

Marroquin, J. L., Mitter, S., & Poggio, T. (1987). Probabilistic Solution of Ill-posed Problems in Computational Vision. *Journal of American Statistical Association* **82**, 76-89.

Narendra, A., & Schachter, B. J. (1983). *Pattern Models.* Wiley.

Ripley, B. D. (1981). *Spatial Statistics.* Wiley.

Ripley, B. D. (1986). Statistics, Images and Pattern Recognition. *Canadian Journal of Statistics* **14**, 83-111.

Ripley, B. D. (1987). *Stochastic Simulation.* Wiley.

Rosenfeld, A. (1978). Iterative Methods in Image Analysis. *Pattern Recognition* **10**, 181-187.

Saebo, H. V., Braten, K., Hjort, N. L., Llewellwn, B., & Mohn, E. (1985). *Contextual Classification of Remotely Sensed Data: Statistical Methods and Development of a System*, Rept. 768, Norwegian Computing Centre.

Seber, G. A. F. (1984). *Multivariate Observations.* Wiley.

Serra, J. (1982). *Image Analysis and Mathematical Morphology.* Academic Press.

Switzer, P. (1980). Extensions of Linear Discriminant Analysis for Statistical Classification of Remotely Sensed Imagery. *Journal of Int. Ass. Math. Geol.* **12**, 367-376.

Switzer, P. (1983). Some Spatial Statistics for the Interpretation of Satellite Data. *Bull, ISI,* **50**, 962-972.

Switzer, P. (1986). *Journal of the Royal Statistical Society* B, 295 in discussion of Besag.

Titterington, D. M. (1986). *Journal of the Royal Statistical Society* B, 280-281, in discussion of Besag 1986.

Yu, T. S., & Fu, K. S. (1983). Recursive Contextual Classification using a Spatial Stochastic Model. *Pattern Recognition* **16**, 89-108.