EXTENDED RULES FOR THE CLASSIFICATION OF DEPENDENT PARAMETERS

By

Kosgallana Liyana Durage Gunawardena

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

1988

ABSTRACT

EXTENDED RULES FOR THE CLASSIFICATION OF DEPENDENT PARAMETERS

By

Kosgallana Liyana Durage Gunawardena

In this dissertation we consider the classification problem in which the unknown states $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ are dependent.

We first review the compound approach to the construction of decision rules and then introduce the extended compound approach as a general approach to the construction of rules with favorable risk behavior. In most empirical Bayes literature, the $\{\theta_i\}$ are assumed to be i.i.d.. The construction of decision rules in the i.i.d. case is very simple. The problem becomes considerably more complex if the $\{\theta_i\}$ are dependent, since the Bayes risk may depend on n and the class of distributions on $\{\theta_i\}$ may be of such high dimension so as to make accurate estimation impossible. The extended compound decision theory developed by Gilliland and Hannan (1969) is utilised to obtain decision rules in the case when $\{\theta_i\}$ are dependent.

The one–dimensional case is generalized to two dimensions where the indices are positions in an integer lattice. The two–dimensional case has applications in image segmentation and pattern recognition where the image is a matrix of parameters $\underline{\theta} = \{\theta_{ij}\}$ taking values in a finite set Θ . The

observed image is a random matrix $\underline{X}$ having a distribution depending on $\underline{\theta}$. Here $\underline{\theta}$ may reasonably be assumed to be the realization of a Markov random field in some applications. The extended compound approach to the image case is described with the choice of a neighborhood system on the lattice.

Many simulations were run to compare the risk behavior of the extended compound rules with that of other rules. The simulations performed show that the extended rules perform better than the other empirical Bayes rules that have been proposed.

To my daughter Kalpanee

ACKNOWLEDGMENTS

I am truly grateful to my advisor Professor Dennis C. Gilliland for his continuous encouragement and guidance in the preparation of this dissertation.

I would also like to thank Professor James Hannan for the introduction to compound decision theory and for teaching me much about mathematical statistics.

I would like to thank Professors R. Ericson and H. Salehi for serving on my guidance committee.

Also thanks are due to Ms. Loretta Ferguson for typing the manuscript.

Finally, I wish to thank the Department of Statistics and Probability for providing the financial support which made my graduate studies in Statistics possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In this thesis we consider the classification problem in which each of n observations is to be classified as belonging to one of two classes.

1.1 A Literature Review

In most of the work relating to the classification problem, the successive classes are assumed to be independent. In many situations, though, it is unreasonable to assume that the successive classes are independent. One of the earliest, if not the first, papers to consider positively correlated classes and that a more sensitive scheme of classification can be obtained by exploiting this correlation is due to Cox (1960).

In Cox (1960) a sequence of batches $\{B_i\}$ is to be classified based on the number of defectives in random samples of predetermined fixed size. The $\{X_i\}$ are random observations, where X_i is the number of defectives in the sample from batch B_i. The $\{X_i\}$ are assumed to be independent and have a Poisson distribution with mean m_i, given the sequence $\{m_i\}$, where m_i is the true but unknown batch quality. To complete the model, the sequence $\{m_i\}$ is assumed to be a realization of a two–state Markov chain with specified states and known transition matrix. Cox suggests using some other X's in addition to X_i in making the decision about the i^{th} batch. In particular he suggests a one–step back rule based on (X_{i-1},X_i) and a one–step forward, one–step back rule based on (X_{i-1},X_i,X_{i+1}).

Preston (1971) seems to be the first person to consider the empirical Bayes approach to a classification problem in which the set of parameter values is a realization of a stationary Markov chain. The model considered

by Preston is very specialized. The random observation X is assumed to have the following conditional distribution:

$$P(X = 0|\theta = 0) = 1,\ P(X = 1|\theta = 1) = p,\ P(X = 0|\theta = 1) = 1-p$$

where p is known. It is assumed that the unknown classifications $\{\theta_i\}$ are a realization of a stationary Markov chain with unknown transition matrix and that the observations X_i are independent conditional on $\{\theta_i\}$. At stage n, the transition probabilities are estimated based on the partial sequence $\underline{X} = (X_1, X_2, \ldots, X_n)$ and substituted into the Bayes rule for the decision about θ_n.

Devore (1975) considers the Robbins' (1951) component with Markovian $\{\theta_i\}$. The Robbins' component has normal conditional distributions $N(2\theta - 1, 1)$ and independent observations conditional on $\{\theta_i\}$. In Devore (1975), the $\{\theta_i\}$ sequence is assumed to be a realization of a stationary two-state Markov chain with

$$P(\theta=0) = P(\theta=1) = .5$$

and transition matrix

$$\begin{bmatrix} .5(1 + \pi) & .5(1 - \pi) \\ .5(1 - \pi) & .5(1 + \pi) \end{bmatrix} \quad \text{for a } \pi \in [-1,1].$$

The above model can be considered as a generalization of the model with repetitions of the Robbins' component problem with independent θ_i and probabilities $P(\theta=0) = P(\theta=1) = .5$. For this model, the simple compound rule $\varphi^0 = (\varphi_1^0, \varphi_2^0, \ldots, \varphi_n^0)$ defined by

$$\varphi_i^0\text{: decide } \theta_i = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if } X_i \begin{matrix} \geq 0 \\ < 0 \end{matrix} \qquad i = 1,2,\ldots,n$$

is the minimax classification rule for $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$.

Devore studies the properties of classification rules for θ_i based on linear functions of X_j's within the context of his problem. The two rules considered are those resulting from optimization over classes of linear rules:

Rule 1

Among all the rules in the class given by:

$$\text{decide} \quad \theta_i = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if} \quad X_i + aX_{i-1} \begin{matrix} \geq k \\ < k \end{matrix} \qquad i = 2, \ldots, n$$

choose that rule, i.e., a and k, which minimizes the probability of misclassification of θ_i.

Rule 2

Among all the rules in the class given by:

$$\text{decide} \quad \theta_i = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if} \quad aX_{i-1} + X_i + aX_{i+1} \begin{matrix} \geq 0 \\ < 0 \end{matrix} \qquad i = 2, \ldots, n$$

choose that rule, i.e., a, which minimizes the probability of misclassification of θ_i.

In both cases the choice can be made independent of i for $i \geq 2$. The optimum rules have k=0 and a as a function of π. When π is unknown, a moment estimator $\hat{\pi}$ based on $\underline{X}$ is used for π and a "plug-in" rule based on $\hat{\pi}$ is considered.

Hill et al. (1984) consider the application of parametric empirical Bayes theory to the same classification problem, which they call the two–crop model on the transect. In Hill et al. (1984), $\{\theta_i\}$ represents the sequence of crop types (classifications) on a transect with each $\theta_i = 0$ or 1. Let X_i be the random observation at the i^{th} pixel. The X_i is assumed to have density $f_{\theta_i}(\cdot)$ where both f_0 and f_1 are known. The X_i's are assumed to be conditionally independent given $\{\theta_i\}$. The sequence of crop types is assumed to be a Markov chain with transition probability matrix

$$\underline{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} .$$

If the Markov chain is stationary, then the common marginal distribution for θ_i is

$$p = P(\theta=0) = \frac{p_{10}}{p_{10}+p_{01}} .$$

The induced prior density on $\underline{\theta}$ satisfies

(1.1) $$P(\underline{\theta}) \propto \exp(\phi_1 N - \phi_2 F)$$

where

$$N = \sum_{i=1}^{n} \theta_i$$

$$F = \sum_{i=1}^{n-1} (\theta_i - \theta_{i+1})^2$$

$$\phi_1 = \log \frac{p_{11}}{p_{00}}$$

and

$$\phi_2 = \frac{1}{2} \log \frac{p_{11}p_{00}}{p_{01}p_{10}}$$

if the end–points are ignored in the derivation of $P(\underline{\theta})$. Letting

$$Z_i = \log \frac{f_1(X_i)}{f_0(X_i)}, \tag{1.2}$$

we have

$$f(\underline{x}|\underline{\theta}) = \prod_{i=1}^{n} f_1(x_i)^{\theta_i} f_0(x_i)^{1-\theta_i}$$

$$= \prod_{i=1}^{n} f_0(x_i) \left[\frac{f_1(x_i)}{f_0(x_i)}\right]^{\theta_i} \tag{1.3}$$

$$\propto \exp\left(\sum_{i=1}^{n} \theta_i z_i\right).$$

Using (1.1) and (1.3), the posterior density of $\underline{\theta}$, given $\underline{x}$, is approximately

$$f(\underline{\theta}|\underline{x}) \propto \exp\left[\sum_{i=1}^{n} \theta_i(z_i + \phi_1) - \phi_2 \sum_{i=1}^{n-1} (\theta_i - \theta_{i+1})^2\right]. \tag{1.4}$$

Morris et al. (1985) reports that the exact Bayes approach gives a complicated joint posterior for $\underline{\theta}$ whose maximization is non–trivial and that efficient likelihood estimation of the parameters of the Markov chain is difficult.

Taking $f_{\theta_i}(\cdot)$ to be $N(\mu_{\theta_i}, \sigma)$ Hill et al. (1984) use empirical Bayes theory to classify and to estimate the posterior probability for each class. The parameters μ_0, μ_1 and σ are assumed to be known.

With $f_{\theta_i}(\cdot) = N(\mu_{\theta_i}, \sigma)$, (1.2) reduces to

$$Z_i = -\frac{1}{2\sigma^2}(\mu_0 - \mu_1)(2X_i - \mu_0 - \mu_1)$$

$$= \delta\left[\frac{X_i - \bar{\mu}}{\sigma}\right]$$

where $\delta = \frac{\mu_1 - \mu_0}{\sigma}$ and $\bar{\mu} = \frac{1}{2}(\mu_0 + \mu_1)$.

Let $\bar{Z} = \frac{1}{n}\sum_i Z_i$ and $r_j = \frac{1}{n}\sum_i (Z_{i+j} - \bar{Z})(Z_i - \bar{Z})$.

Then $$E\bar{Z} = \delta^2(\cdot 5 - p)$$

$$Er_j \simeq \delta^4 p(1-p)(p_{11} - p_{01})^j, \qquad j = 1,2,\ldots\ldots .$$

The unknown parameters p_{11} and p_{01} are estimated by solving the equations

$$\begin{aligned} \delta^2(\cdot 5 - p) &= \bar{Z} \\ \delta^4 p(1-p)(p_{11} - p_{01}) &= r_1 . \end{aligned} \tag{1.5}$$

Instead of maximizing (1.4) over all possible $\underline{\ell}$, Hill et al. (1984) develop a formula for the posterior log odds ratio for each component i. The posterior log odds ratio is approximated by a moving average

$$\log \frac{P(\theta_i = 1 \mid \underline{X})}{P(\theta_i = 0 \mid \underline{X})} \simeq \log \frac{1-p}{p} + Z_i + \sum_{j \geq 1} w_j (Z_{i-j} + Z_{i+j}) \tag{1.6}$$

where $w_j = (p_{11} - p_{01})^j$. Hence, the Bayes decision rule, where the loss is the average number of misclassification for $\theta_1, \theta_2, \ldots, \theta_n$, is approximated by the rule

$$\text{(1.7)} \qquad \text{decide} \quad \theta_i = \begin{cases} 1 & \text{if RHS (1.6)} > 0 \\ 0 & \text{otherwise .} \end{cases}$$

Hill et al. (1984) use (1.5) to estimate p and $p_{11} - p_{01}$, substitute the estimates into (1.6), and use (1.7) for classification. Thus, they have developed a set empirical Bayes decision procedure for the Robbins' component and Markov prior. We will refer to procedures with j restricted to {1} in (1.6) as HEB3 and to procedures with j restricted to {1,2} in (1.6) as HEB5.

In Chapter 2, Section 1 we derive an exact expression for the induced posterior density on $\underline{\theta}$, where the end–points are included. We also give an algorithm for the maximization of the joint posterior, and classification rules are derived based on this maximization.

The work reviewed thus far concerns models with a certain common underlying structure. The important features of this structure are:

1. There is a sequence of unobservable parameters $\{\theta_i\}$ where $\theta_i = 0$ or 1, $i = 1,2,\ldots$.

2. Conditional on $\{\theta_i\}$, independent $X_1, X_2, \ldots$ are observed where $X_i \sim P_{\theta_i}$, $i = 1,2,\ldots$. The two distributions P_0 and P_1 are assumed known.

3. The parameter sequence $\{\theta_i\}$ has a (joint) distribution G.

4. There is a decision $\hat{\theta}_i$ to be made about θ_i, specifically, a classification as 0 or 1, $i = 1,2,\ldots\ldots$.

For each n we let $\underline{\theta} = (\theta_1,\theta_2,\ldots\ldots,\theta_n)$ and $\underline{X} = (X_1,X_2,\ldots\ldots,X_n)$ denote partial sequences, let $\underline{G}$ denote the distribution of $\underline{\theta}$ and let $P_{\underline{\theta}}$ denote the conditional distribution of $\underline{X}$ given $\underline{\theta}$. We consider the set compound decision problem and allow the $\hat{\theta}_i$ to depend on $\underline{X}$, $i = 1,2,\ldots\ldots n$.

There are two loss functions that are commonly considered with this structure. They are

$$L_1(\underline{\theta},\hat{\underline{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} [\, \theta_i \neq \hat{\theta}_i \,] \tag{1.8}$$

and

$$L_2(\underline{\theta},\hat{\underline{\theta}}) = [\, \underline{\theta} \neq \hat{\underline{\theta}} \,] \tag{1.9}$$

where square brackets denote indicator functions. L_1 defines the loss as the average number of misclassifications across the n components. L_2 defines the loss as 0 or 1 depending upon whether all n components are correctly classified or not.

At stage n, a Bayes rule with respect to $\underline{G}$ and L_1 is determined by taking $\hat{\theta}_i$ as 0 or 1 depending upon which maximizes the posterior probability distribution of θ_i given $\underline{X}$, $i = 1, 2,\ldots\ldots,n$. Most of our work concerns comparing rules relative to loss L_1 and we refer to the Bayes rules in this case as simply "the Bayes rules." A Bayes rule with respect to $\underline{G}$ and L_2 selects $\hat{\underline{\theta}}$ to maximize the posterior probability of $\underline{\theta}$ given $\underline{X}$. We refer to such a rule as a MAP rule following the common usage of "Maximum a–Posteriori" to describe such a rule. Of course, a Bayes rule has components that are based on the marginal distributions of the posterior of $\underline{\theta}$ given $\underline{X}$.

There may be an ordering or spatial structure to the index set $\{1,2,.....\}$ that suggests certain classes of priors G for certain problems. The one–dimensional case with order of indices representing spatial or time order is called the transect case. We consider the transect case in Chapter 2. Here G might be taken to be a simple product measure $G = G_1 \times G_1 \times G_1 \times .....$ or a Markov chain, depending upon the degree of dependence that is being modeled. In the literature cited previously, i.e. Cox (1960), Preston (1971), Devore (1975) and Hill et al. (1984), the θ_i are assumed to be the realization of a stationary Markov chain. In most empirical Bayes literature, the θ_i are assumed to be i.i.d. G_1 where G_1 is unknown.

In the i.i.d. G_1 case, the Bayes rule and the MAP rule are very simple since a conditional distribution for $\underline{\theta}$ given $\underline{X}$ is the product of the conditional distributions of θ_i given X_i; i = 1,2,...,n. In the Markov chain case, the Bayes and MAP rules are sufficiently complicated so as to have motivated researchers to impose simplifying assumptions. Recall, for example, that Cox (1960) takes P_0, P_1 to be Poisson, G to be a known Markov chain distribution and derives restricted Bayes rules, specifically, $\hat{\theta}_i$ restricted to be Bayes with respect to the class of rules that are (X_{i-1}, X_i) (or (X_{i-1}, X_i, X_{i+1})) measurable. Preston (1971) takes the very special case P_0 is degenerate on 0 and P_1 is Bernoulli B(1, p) and assumes G to be from the family of stationary Markov distributions. Preston considers Bayes, MAP and restricted Bayes rules. Devore (1975) takes $P_0 = N(-1, 1)$, $P_1 = N(1, 1)$ and assumes G to be from a special one–parameter family of symmetric stationary Markov distributions. Hill et al. (1984) take $P_0 = N(\mu_0, \sigma)$, $P_1 = N(\mu_1, \sigma)$ (without loss of generality N(–1, 1) and N(1, 1)) and assumes G is from the family of stationary

Markov distributions. They derive an approximation to the Bayes rule. Preston (1971), Devore (1975) and Hill et al. (1984) give estimators for the transition probabilities defining G and, therefore, consider the empirical Bayes model in a case where the parameter sequence is not i.i.d..

1.2 A General Approach

Consider the decision problem involving $\underline{\theta}$ introduced in last section and suppose that the loss function is L_1. Then a decision rule $\hat{\underline{\theta}}$ is evaluated in terms of average number of errors (misclassifications) across the n components. We will first review the compound approach to the construction of $\hat{\underline{\theta}}$ and will then introduce the extended compound approach as a general approach to the construction of rules with favorable risk behavior. There is a long history of work on the two-state compound decision problem including Robbins (1951), Hannan and Robbins (1955), Hannan and Van Ryzin (1965), Van Ryzin (1966), Gilliland, Hannan and Huang (1976), and on the extended version including Gilliland and Hannan (1969), Ballard and Gilliland (1978), Ballard, Gilliland and Hannan (1975), Vardeman (1975), (1980), (1981). Our review will be relatively brief and the arguments will be heuristic.

For a (compound) decision rule $\hat{\underline{\theta}}$, the compound risk at $\underline{\theta}$ and n is

$$\underline{R}(\underline{\theta},\hat{\underline{\theta}}) = E_{\underline{\theta}}\, L_1(\underline{\theta},\hat{\underline{\theta}}), \tag{1.10}$$

where L_1 is defined in (1.8). If G_n^1 is the empirical distribution of $\theta_1,\theta_2,\ldots,\theta_n$ and R^1 is the Bayes envelope risk function in the component problem, then the modified regret of $\hat{\underline{\theta}}$ at $\underline{\theta}$ and n is

(1.11) $$\underline{D}(\underline{\theta},\hat{\underline{\theta}}) = \underline{R}(\underline{\theta},\hat{\underline{\theta}}) - R^1(G_n^1).$$

Let $\mathcal{G}$ be a given class of distributions G on $\{\theta_i\}$. For each n let $\mathcal{D}$ denote the class of all (set) compound decision rules $\hat{\underline{\theta}}$ with $\hat{\underline{\theta}}_i$ a function of $\underline{X}$, i=1,2,.....,n. Let $\underline{R}(\underline{G},\hat{\underline{\theta}})$ denote the $\underline{G}$ expectation of $\underline{R}(\underline{\theta},\hat{\underline{\theta}})$. Suppose that $\hat{\underline{\theta}}(\underline{G})$ minimizes $\underline{R}(\underline{G},\hat{\underline{\theta}})$ among all rules $\hat{\underline{\theta}}$ in $\mathcal{D}$. We call $\hat{\underline{\theta}}(\underline{G})$ a Bayes response and denote the minimum Bayes risk as $\underline{R}(\underline{G})$. Then any rule $\hat{\underline{\theta}}$ can be judged by its Bayes risk relative to that of $\hat{\underline{\theta}}(\underline{G})$, i.e., by the excess

(1.12) $$\underline{D}(\underline{G},\hat{\underline{\theta}}) = \underline{R}(\underline{G},\hat{\underline{\theta}}) - \underline{R}(\underline{G}).$$

Note that this is not simply the $\underline{G}$ expectation of (1.11). Any sequence of rules $\hat{\underline{\theta}}$ such that

(1.13) $$\lim_n \underline{D}(\underline{G},\hat{\underline{\theta}}) = 0 \qquad \text{for all } G \in \mathcal{G}$$

is said to be asymptotically optimal.

If $\mathcal{G}$ is a singleton set, i.e., G is assumed known, then the only difficulty that arises in the Bayes decision problem is in calculating a Bayes response $\hat{\underline{\theta}}(\underline{G})$, since it may be a complicated function of $\underline{G}$ and $\underline{X}$. In the literature reviewed in the last section, Cox (1960) took G to be a known Markov chain.

Robbins (1956), (1961) introduced the empirical Bayes decision problem by assuming that G is unknown and using $\underline{X}$ to estimate $\underline{G}$ or $\hat{\underline{\theta}}(\underline{G})$. (As introduced by Robbins, only the component observations $X_1,X_2,.....,X_i$ are available for making the decision about θ_i, i=1,2,..... so that his version

is sequence in nature.) Robbins proposed (bootstrap) empirical Bayes rules of the type $\hat{\underline{\theta}}(\hat{\underline{G}})$ where $\hat{\underline{G}}$ is an estimate of $\underline{G}$ and rules of the type where the Bayes response is estimated directly without estimating $\underline{G}$.

It is obvious that the class $\mathcal{G}$ of priors to which G is assumed to belong must be restricted for the empirical Bayes approach to produce reasonable methods. Robbins chose $\mathcal{G}$ to be the class of all symmetric products

$$\mathcal{G} = \{ G_1 \times G_1 \times G_1 \times \ldots\ldots | G_1 \in \mathcal{G}_1 \}$$

in which case the Bayes risk envelope is simply $\underline{R}(\underline{G}) = R^1(G_1)$, independent of n. This is the assumed structure for $\mathcal{G}$ in most empirical Bayes literature where much of the effort concerns finding sequence rules $\hat{\underline{\theta}}$ that are component–wise asymptotically optimal, such that

$$\lim_i E [\hat{\theta}_i \neq \theta_i] = R^1(G_1) \qquad \text{for all } G_1 \in \mathcal{G}_1, \tag{1.14}$$

where $\hat{\theta}_i$ depends on $\underline{X}$ through $X_1, X_2, \ldots\ldots, X_i$. The feasibility of constructing sequence empirical Bayes rules that are asymptotically optimal for this $\mathcal{G}$ is quite apparent. Since the observed random variables $X_1, X_2, \ldots\ldots$ are i.i.d. according to the G_1 mixture P_{G_1} of P_0 and P_1, if G_1 is identified by P_{G_1}, then consistent estimation of G_1 is possible.

Of course, the restriction to product measures is unreasonable in applications for which the parameters $\theta_1, \theta_2, \ldots\ldots$ are correlated. In the last section, we reviewed examples of empirical Bayes problems where $\mathcal{G}$ went beyond the product case.

The empirical Bayes problem becomes considerably more complex if the class of distributions contains distributions that are not symmetric product measures. The excess risk (1.12) is the appropriate measure of efficiency for a decision rule $\hat{\underline{\theta}}$, but problems arise in the construction of bootstrap empirical Bayes rules. For one thing, $\mathcal{G}$ may be of such high dimension so as to make accurate estimation of $\underline{G}$ or $\hat{\underline{\theta}}(\underline{G})$ based on $\underline{X}$ impossible except for very large n. Moreover, $\hat{\underline{\theta}}(\underline{G})$ may be such a complicated function of $\underline{G}$ and $\underline{X}$ so as to make estimation difficult and $\underline{R}(\underline{G})$ may depend on n.

The dimensionality problem is not severe when $\mathcal{G}$ is a family of Markov distributions. For the two-state component we are considering, the stationary Markov distributions are indexed by two real parameters. The literature reviewed in the last section concerned this family.

The second problem has led investigators to approximate the functional form of a Bayes response $\hat{\underline{\theta}}(\underline{G})$ and to estimate the approximation. For example, (1.7) is an approximation to the decision rule $\hat{\theta}_i(\underline{G})$ in the transect case with stationary Markov prior. Truncating the range of j in (1.6) to various neighborhoods of i produces additional approximations.

The extended compound decision theory developed by Gilliland and Hannan (1969) suggests a general approach for finding empirical Bayes decision rules in our applications. It provides a logical and systematic method for finding approximations to Bayes rules. It solves the dimensionality problem and the approximation problem by defining envelope risk in terms of that resulting by restricting decision rules to a class $\mathcal{D}^k \subset \mathcal{D}$. Basically, the restriction is limited so that the envelope risk is close to Bayes risk $\underline{R}(\underline{G})$ and, at the same time, the restricted Bayes decision rules are simple enough to be easily estimated. Finally it should be

noted that this approach leads to compound decision rules which have controlled risk behavior conditional on $\underline{\theta}$ whereas the competitors that we test have not been shown to enjoy this more robust risk behavior.

We will now give some details to illustrate this approach. This will require introduction of the extended envelope theory and Γ^k construct of Gilliland and Hannan (1969). The parameter vector $\underline{\theta}$ belongs to the parameter set $\underline{\Theta} = \Theta^n$ where $\Theta = \{0,1\}$, in our case. Suppose that associated with each index $i \in I = \{1,2,.....,n\}$ is an (ordered) k–tuple of indices, $\mathcal{N}_i^* \in I^k$, where i is the last index in $\mathcal{N}_i^*$. We let $\underline{\theta}_i^k$ denote the restriction of $\underline{\theta}$ to $\mathcal{N}_i^*$, $i \in I$. An example with $k = 3$ is provided by $\mathcal{N}_i^* = (i-1, i+1, i)$ and $\underline{\theta}_i^3 = (\theta_{i-1}, \theta_{i+1}, \theta_i)$. (One can avoid endpoint problems by using the mod–n arithmetic on I to define the $\underline{\theta}_i^k$ for all $i \in I$ or simply by restricting the indices, say to $i = 2,3,.....,n-1$ in this example.)

The decision problem concerning the last component of a k–tuple with the decision based on independent obsevations based on all k parameters is called a Γ^k construct by Gilliland and Hannan (1969). Thus, we see that associated with n, the compound decision problem $\underline{\theta}$, and a neighborhood system $\mathcal{N}^*$ are n Γ^k–decision problems with parameters $\underline{\theta}_1^k, \underline{\theta}_2^k, ..., \underline{\theta}_n^k$. With G_n^k denoting the empirical distribution of this sequence and R^k denoting the Bayes envelope risk for the Γ^k decision problem, the k–extended modified regret for a decision rule $\hat{\underline{\theta}}$ is

$$\underline{D}^k(\underline{\theta},\hat{\underline{\theta}}) = \underline{R}(\underline{\theta},\hat{\underline{\theta}}) - R^k(G_n^k), \tag{1.15}$$

which subsumes (1.11) as a special $k = 1$ case. Note that we are suppressing the display of dependence on $\mathcal{N}^*$ by displaying only the size

k of the neighborhoods. As shown by Gilliland and Hannan (1969), for $k > 1$, the R^k are more stringent envelopes than the (unextended) component envelope $R^1 = R$, and, since G_n^1 is a marginal of G_n^k,

(1.16) $$R^k(G_n^k) \leq R^1(G_n^1) \qquad \text{for all } \underline{\theta}, \text{ n.}$$

The envelope risk $R^k(G_n^k)$ is the envelope to compound risk obtained by minimizing over the class $\mathscr{D}^k$ of compound decision rules $\hat{\underline{\theta}}$ where $\hat{\underline{\theta}}_i$ is a fixed function of $\underline{X}_i^k$, i = 1,2,.....,n. (The function does not depend on i.)

We are interested in the Bayes risk obtained by taking expectation of compound risk with respect to the distribution $\underline{G}$ on $\underline{\theta}$. Suppose that G is a strictly stationary distribution on $\theta_1, \theta_2, \ldots\ldots$ with common marginal G^k on $\underline{\theta}_i^k$, i= 1,2,..... . Theorem 3 of Gilliland and Hannan (1969) shows, in our application, that any compound decision rule $\hat{\underline{\theta}}$ satisfying

(1.17) $$\limsup_n \underline{D}^k(\underline{\theta},\hat{\underline{\theta}}) \leq 0 \qquad \text{for all } \underline{\theta}$$

also satisfies

(1.18) $$\limsup_n \underline{D}^k(\underline{G},\hat{\underline{\theta}}) \leq 0 \qquad \text{for all } \underline{G}$$

where

(1.19) $$\underline{D}^k(\underline{G},\hat{\underline{\theta}}) = \underline{R}(\underline{G},\hat{\underline{\theta}}) - R^k(G^k).$$

Suppose that the neighborhood structure is such that $R^k(G^k)$ is close to Bayes risk $\underline{R}(\underline{G})$ for all $G \in \mathscr{G}$, specifically, suppose that

(1.20) $$R^k(G^k) \leq \underline{R}(\underline{G}) + \epsilon \qquad \text{for all G, n.}$$

If $\hat{\underline{\theta}}$ is an asymptotic solution to the extended compound problem, i.e., satisfies (1.17), then, by (1.12), (1.18) and (1.20),

$$\text{(1.21)} \qquad \limsup_n \underline{D}(\underline{G},\hat{\underline{\theta}}) \leq \epsilon \qquad \text{for all } G \in \mathcal{G}.$$

Thus, the compound decision rule will be asymptotically ϵ–Bayes. The extended rules we propose can be shown to satisfy (1.17) by suitable adaptation of the sequence extended work of Ballard, Gilliland and Hannan (1975).

Recall from the last section that Preston (1971), Devore (1975) and Hill et al. (1984) consider approximations to the Bayes response $\hat{\underline{\theta}}(\underline{G})$. They examine the component $\hat{\theta}_i(\underline{G})$ as a function of $\underline{X}$ and approximate it by a function of X_j for j in a neighborhood of i, $i \in I$. In the empirical Bayes situation where G is an unknown stationary Markov distribution, they propose empirical Bayes procedures that estimate G using $\underline{X}$ and plug these estimates into the component approximations. They give no rationale for the approximations. In contrast, the extended compound approach provides a rationale for determining the neighborhood structure. It should be large enough to control the excess $R^k(G^k) - \underline{R}(\underline{G})$ and small enough to allow for efficient estimation of the optimal decision procedure in the corresponding restricted class of compound decision rules $\mathcal{D}^k$. We find in our risk calculations in latter chapters that the extended compound rules outperform the competitors selected for comparison.

1.3 The Image Problem

We have called the case where the index set I = {1,2,..... } the transect case. Besides the transect case, we will consider the two–dimensional case where the indices index positions in an integer lattice, for example, the pixels in an image. We call this the image case. Here $\{\theta_{ij}\}$ may reasonably be assumed to be the realization of a Markov random field in some applications.

A brute force calculation of the MAP estimate for the image $\underline{\theta}$ with $n = N_1 \times N_2$ requires a search for a maximum across 2^n calculated posterior probabilities. There has emerged a considerable literature dealing with the computational issues that arise in the MAP approach to image reconstruction since an image may easily have $n = 256^2 = 65{,}536$ and, therefore, a posterior that is supported by $2^{65{,}536}$ points. (LANDSAT 1–3 scanned the earth every 18–19 days recording images on a 3240×2340 lattice. Ripley (1986).)

In the typical empirical Bayes approach to the two–dimensional case, the unknown Markov random field distribution has parameters that define the conditional distribution for each θ_{ij} given values of θ_{kl} in a specified neighborhood. It follows that the posterior probability for θ_{ij} given $\underline{X}$ depends most heavily upon the X_{kl} for (k,l) in a neighborhood of (i,j). The result is that approximations to Bayes rules and MAP rules are possible and that estimates of the parameters based on $\underline{X}$ provide for the construction of empirical Bayes procedures for classifying the components of $\underline{\theta}$.

The extended compound approach has the same advantages in the image case as in the transect case. In fact, Swain et al. (1981), Tilton et al. (1982) have used the compound decision theory approach to develop rules to classify image data. A difficulty encountered in the implementation of the

rules is the estimation of G_n^k. A "classify–and–count" method is proposed to estimate G_n^k, with the use of a training set of data (a set of data with known classifications or a representative sample from the image data). First, the training set of data is classified by using uniform prior probabilities and then from this classification estimates of G_n^k are obtained by counting the occurences of $\underline{\theta}_i^k$.

We will use the extended compound approach in the image case and compare the rules with the empirical Bayes classification rules suggested by Morris et al. (1985), Morris (1986) on selected images.

1.4 Summary

In this Chapter we have reviewed the most relevant literature and have set the stage for later work where the risk behavior of extended compound decision rules is compared with that of other rules.

In Chapter 2 we develop the details of the transect case. We derive an algorithm for the calculation of the MAP rule. We introduce an empirical MAP rule and describe further the empirical Bayes classification rules of Hill et al. (1984). The extended compound rules are developed and some results of simulations to determine risk behavior for the various rules are summarized.

In Chapter 3 we develop the details of the image case. We introduce the empirical Bayes classification rules of Morris et al. (1985), Morris (1986). The extended compound rules are developed and some results of simulations to determine risk behavior for the various rules are summarized.

In Chapter 4 we give the results of the many simulations that were run. In the transect case, only $\underline{\theta}$ generated as Markov chains are considered. In the image case both fixed pattern and randomly generated $\underline{\theta}$ are considered.

Risks conditional on ℓ are estimated in every case.

In image segmentation literature, the performance of a classification rule is often measured by the quantity "per cent classified correctly" (PCC). The measure PCC is related to the loss function L_1 by $PCC = 100(1 - L_1)$. However, a higher PCC does not necessarily mean that the image reproduced will be perceived to be better than one with a lower PCC. Swain et al. (1981) suggests measuring the performance of classification rules in two ways, one of which is PCC. The other measurement is "average by class accuracy" (ACA). The ACA is obtained by first computing PCC for each class and then taking their arithmetic average. ACA will show whether the classification rule favors one class more than the others. We will report risk behaviors in terms of PCC.

CHAPTER 2

THE TRANSECT CASE

2.1 Introduction

In this chapter we consider the classification on a transect. Let $\theta_i \in \Theta = \{0,1\}$, $i = 1,2,\ldots\ldots$. For the empirical Bayes application, it is assumed that parameter sequence $\{\theta_i\}$ has a stationary Markov chain distribution G as its prior. We derive an empirical MAP rule and extended compound rules for the classification of $\underline{\theta}$. The risk performances of these rules are evaluated in comparison with the empirical Bayes rules (1.7) suggested by Hill et al. (1984).

Let $\{X_i\}$ denote the random observations. For our calculations, we take the Robbins' component. Thus, the X_i are conditionally independent given $\{\theta_i\}$ with $X_i \sim P_{\theta_i}$, where P_{θ_i} is the normal distribution with mean $2\theta_i - 1$ and variance 1. We let f_θ denote the density of P_θ. The sequence of states $\{\theta_i\}$ is assumed to be a realization of a two state stationary Markov chain with initial probability distribution

$$(2.1) \qquad P(\theta = 0) = p = 1 - P(\theta = 1), \qquad 0 < p < 1$$

and transition matrix

$$(2.2) \qquad \underline{P} = \begin{pmatrix} \delta & 1-\delta \\ \frac{p(1-\delta)}{1-p} & 1 - \frac{p(1-\delta)}{1-p} \end{pmatrix}, \qquad 0 < \delta < 1.$$

If δ is close to 1 there are likely to be long sequences of 0's and 1's in $\{\theta_i\}$ and $\delta = p$ corresponds to independence.

It can be easily shown that the Markov chain defined by (2.1) and (2.2) is stationary. We will sometimes denote $\underline{P}$ by

$$\underline{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

The induced density on $\underline{\theta}$ is

$$p(\underline{\theta}) = P(\theta_1 = \theta_1) \prod_{i=2}^{n} P(\theta_i = \theta_i \mid \theta_{i-1} = \theta_{i-1})$$

$$= p^{1-\theta_1} (1-p)^{\theta_1} \prod_{i=2}^{n} \begin{bmatrix} 1-\theta_i & \theta_i \\ p_{00} & p_{01} \end{bmatrix}^{1-\theta_{i-1}} \begin{bmatrix} 1-\theta_i & \theta_i \\ p_{10} & p_{11} \end{bmatrix}^{\theta_{i-1}}$$

$$= p \left[\frac{1-p}{p}\right]^{\theta_1} \prod_{i=2}^{n} p_{00} \left[\frac{p_{10}}{p_{00}}\right]^{\theta_{i-1}} \left[\frac{p_{01}}{p_{00}}\right]^{\theta_i} \left[\frac{p_{00}\, p_{11}}{p_{01}\, p_{10}}\right]^{\theta_{i-1}\theta_i}$$

$$= p\, p_{00}^{n-1} \left[\frac{1-p}{p}\right]^{\theta_1} \prod_{i=2}^{n} \left[\frac{p_{10}}{p_{00}}\right]^{\theta_{i-1}} \left[\frac{p_{01}}{p_{00}}\right]^{\theta_i} \left[\frac{p_{00}\, p_{11}}{p_{01}\, p_{10}}\right]^{\theta_{i-1}\theta_i} . \tag{2.3}$$

In the above derivation, the end–points of $\underline{\theta}$ are included and all the entries in $\underline{P}$ are assumed to belong to (0, 1). When $p_{ij} \in \{0, 1\}$, the obvious modifications have to be done to RHS (2.3).

2.2 Classification by Maximum Posterior Probability (MAP)

In this section we derive the MAP rule. The MAP rule is Bayes with respect to the loss function

$$L_2(\underline{\theta},\hat{\underline{\theta}}) = [\ \underline{\theta} \neq \hat{\underline{\theta}}\ ]$$

and chooses a state sequence which maximizes $f(\underline{\theta}|\underline{x})$. Here

$$f(\underline{x}|\underline{\theta}) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} \exp\ \{-\tfrac{1}{2}(x_i - 2\theta_i + 1)^2\}\right] \tag{2.4}$$

and

$$f(\underline{x},\underline{\theta}) = \text{RHS(2.3)} \times \text{RHS(2.4)}.$$

Thus,

$$\log f(\underline{x},\underline{\theta}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - 2\theta_i + 1)^2$$

$$+\log\left[p\ p_{00}^{n-1}\right] + \theta_1 \log\left[\frac{1-p}{p}\right]$$

$$+\sum_{i=2}^{n}\left[\theta_{i-1}\log\frac{p_{10}}{p_{00}} + \theta_i \log\frac{p_{01}}{p_{00}} + \theta_{i-1}\theta_i \log\left[\frac{p_{00}p_{11}}{p_{01}p_{10}}\right]\right]$$

$$= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i + 1)^2 + \log(pp_{00}^{n-1})$$

$$+ 2\sum_{i=1}^{n} x_i\theta_i + \theta_1 \log\left[\frac{1-p}{p}\right]$$

$$+\sum_{i=2}^{n}\left[\theta_{i-1}\log\frac{p_{10}}{p_{00}} + \theta_i \log\frac{p_{01}}{p_{00}} + \theta_{i-1}\theta_i \log\left[\frac{p_{00}p_{11}}{p_{01}p_{10}}\right]\right]$$

$$= a + \sum_{i=1}^{n} b_i\theta_i + c\sum_{i=2}^{n}\theta_{i-1}\theta_i\ , \tag{2.5}$$

where

$$a = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i + 1)^2 + \log\left(pp_{00}^{n-1}\right),$$

$$b_1 = 2x_1 + \log\left[\frac{1-p}{p}\cdot\frac{p_{10}}{p_{00}}\right] = 2x_1 + \log\left[\frac{1-\delta}{\delta}\right],$$

$$b_i = 2x_i + \log\left[\frac{p_{10}p_{01}}{p_{00}^2}\right], \qquad 1 < i < n,$$

$$= 2x_i + \log\left[\left[\frac{1-\delta}{\delta}\right]^2\frac{p}{1-p}\right],$$

$$b_n = 2x_n + \log\left[\frac{p_{01}}{p_{00}}\right] = 2x_n + \log\left[\frac{1-\delta}{\delta}\right],$$

and

$$c = \log\left[\frac{p_{00}p_{11}}{p_{01}p_{10}}\right] = \log\left[\frac{\delta(1-2p+p\delta)}{p(1-\delta)^2}\right].$$

Maximizing $f(\underline{\theta}|\underline{x})$ is equivalent to maximizing (2.5). A brute force maximization of (2.5) is simply out of the question because there are 2^n possible values of $\underline{\theta}$ to be searched. Traditionally, dynamic programming [Bertsekas (1976), Cooper and Cooper (1981)] has been used to find "optimum" solutions to multistage decision problems. An optimum solution has generally been one that maximizes or minimizes a performance or cost functional.

Dynamic programming techniques have been used in signal processing problems [Scharf and Elliot (1981)] and image segmentation [Hansen and Elliot (1982)]. The multistage decision problem we consider is formulated in

a probabilistic framework and the criterion of optimality is MAP. The cost functional (2.5) is a multivariable likelihood function and (2.5) is maximized by the following dynamic programming recursion [see Appendix].

Let $A_1^0 = 0 \qquad A_1^1 = b_1$

For $1 \leq i \leq n-1$, let

$$A_{i+1}^0 = \max(A_i^0, A_i^1)$$

(2.6)

$$A_{i+1}^1 = \max(A_i^0 + b_{i+1}, \; A_i^1 + b_{i+1} + c).$$

Then,

$$\max_{\underline{\theta}} \log f(\underline{x},\underline{\theta}) = a + \max\,(A_n^0, A_n^1).$$

Note that, A_k^0 is the value of

$$(2.7) \qquad \max_{(\theta_1,\theta_2 \cdots,\theta_{k-1})} \left[\sum_{i=1}^{k} b_i \, \theta_i + c \sum_{i=2}^{k} \theta_{i-1}\theta_i \right],$$

with $\theta_k = 0$ and A_k^1 is the value of (2.7) with $\theta_k = 1$.

The MAP rule is given by

$$(2.8) \qquad \text{decide} \quad \theta_i = \begin{cases} 1 & \text{if } A_i^1 \geq A_i^0 \\ 0 & \text{otherwise} \end{cases} \qquad 1 \leq i \leq n.$$

As stated before, few modifications have to be done on (2.3) and (2.5) when $p_{ij} \in \{0, 1\}$; i, j = 0, 1. We shall consider the following three cases in addition to the one already considered. These cover all of the nontrivial possibilities. In all four cases the cost functional is maximized by using (2.7) and the state sequence is classified using (2.8).

Case 1. $0 < p < \cdot 5, \qquad \underline{P} = \begin{pmatrix} 0 & 1 \\ p_{10} & p_{11} \end{pmatrix}.$

In this case (2.3) reduces to

$$p(\underline{\theta}) = p^{1-\theta_1} (1-p)^{\theta_1} \prod_{i=2}^{n} \{(p_{10}^{1-\theta_i} p_{11}^{\theta_i})^{\theta_{i-1}} [(\theta_{i-1}, \theta_i) \neq (0, 0)]\} .$$

For all $\underline{\theta} \in \Theta^n$ with $(\theta_{i-1}, \theta_i) \in \Theta^2 - \{(0, 0)\}$; $\quad i = 2,3,.....,n$

$$f(\underline{x},\underline{\theta}) = \text{RHS (2.4)} \times p^{1-\theta_1} (1-p)^{\theta_1} \prod_{i=2}^{n} (p_{10}^{i-\theta_i} p_{11}^{\theta_i})^{\theta_{i-1}}$$

and (2.5) becomes

$$\log f(\underline{x},\underline{\theta}) = a + \sum_{i=1}^{n} b_i \theta_i + c \sum_{i=2}^{n} \theta_{i-1} \theta_i$$

where

$$a = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} (x_i + 1)^2 + \log p ,$$

$$b_1 = 2x_1 + \log \left[\frac{1-p}{p} \cdot p_{10}\right] = 2x_1,$$

$$b_i = 2x_i + \log \left[\frac{p}{1-p}\right], \qquad 1 < i < n,$$

$$b_n = 2x_n,$$

and

$$c = \log\left[\frac{p_{11}}{p_{10}}\right] = \log\left[\frac{1-2p}{p}\right].$$

Case 2. $\cdot 5 \leq p < 1$, $\quad \underline{P} = \begin{pmatrix} p_{00} & p_{01} \\ 1 & 0 \end{pmatrix}$.

In this case (2.3) reduces to

$$p(\underline{\theta}) = p^{1-\theta_1}(1-p)^{\theta_1} \prod_{i=2}^{n} \{(p_{00}^{1-\theta_i} p_{01}^{\theta_i})^{1-\theta_{i-1}} [(\theta_{i-1}, \theta_i) \neq (1, 1)]\} .$$

For all $\underline{\theta} \in \Theta^n$ with $(\theta_{i-1}, \theta_i) \in \Theta^2 - \{(1, 1)\}$; $\quad i = 2,3,.....,n$

$$f(\underline{x},\underline{\theta}) = \text{RHS (2.4)} \times p^{1-\theta_1}(1-p)^{\theta_1} \prod_{i=2}^{n} (p_{00}^{1-\theta_1} p_{01}^{\theta_i})^{1-\theta_{i-1}}$$

and (2.5) becomes

$$\log f(\underline{x},\underline{\theta}) = a + \sum_{i=1}^{n} b_i\theta_i$$

where

$$a = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i + 1)^2 + \log p + (n-1)\log p_{00} ,$$

$$b_1 = 2x_1 + \log\left[\frac{1-p}{p} \cdot \frac{1}{p_{00}}\right] = 2x_1 + \log\left[\frac{1-p}{p\delta}\right],$$

$$b_i = 2x_i + \log\left[\frac{p_{01}}{p_{00}^2}\right] = 2x_i + \log\left[\frac{1-\delta}{\delta^2}\right] \qquad 1 < i < n,$$

and

$$b_n = 2x_n + \log\left[\frac{p_{01}}{p_{00}}\right] = 2x_n + \log\left[\frac{1-\delta}{\delta}\right].$$

Case 3. $p = 1, \qquad \underline{P} = \begin{pmatrix} p_{00} & p_{01} \\ 1 & 0 \end{pmatrix}.$

In this case (2.3) reduces to

$$p(\underline{\theta}) = [\theta_1 = 0] \prod_{i=2}^{n} \{(p_{00}^{1-\theta_i} p_{01}^{\theta_i})^{1-\theta_{i-1}} [(\theta_{i-1}, \theta_i) \neq (1, 1)]\} .$$

For all $\underline{\theta} \in \Theta^n$ with $(\theta_{i-1}, \theta_i) \in \Theta^2 - \{(1, 1)\}$ and $\theta_1 = 0$; $i = 2,3,.....,n$

$$f(\underline{x},\underline{\theta}) = \text{RHS (2.4)} \times \prod_{i=2}^{n} (p_{00}^{1-\theta_i} p_{01}^{\theta_i})^{1-\theta_{i-1}}$$

and (2.5) becomes

$$\log f(\underline{x},\underline{\theta}) = a + \sum_{i=1}^{n} b_i\theta_i$$

where

$$a = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n} (x_i + 1)^2 + (n-1)\log p_{00} ,$$

$$b_1 = 2x_1 + \log\left[\frac{1}{p_{00}}\right] = 2x_1 + \log\left[\frac{1}{\delta}\right],$$

$$b_i = 2x_i + \log\left[\frac{p_{01}}{p_{00}^2}\right] = 2x_i + \log\left[\frac{1-\delta}{\delta^2}\right] \qquad 1 < i < n,$$

and

$$b_n = 2x_n + \log\left[\frac{p_{01}}{p_{00}}\right] = 2x_n + \log\left[\frac{1-\delta}{\delta}\right].$$

2.3 Empirical MAP

In order to implement the MAP rule (2.8) we need to obtain estimates for the unknown parameters of G. The proposed estimates are based on $\underline{X}$. When plugged into (2.8) the result is an empirical MAP rule.

Proposition 2.1.

(i) $U_n = \frac{1}{2}(1 - \overline{X}_n)$, where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, is an unbiased estimate for p.

(ii) Let $Z_i = [\text{sgn } X_i \neq \text{sgn } X_{i+1}]$, $i = 1,2,\ldots,n-1$ and

$$\overline{Z}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} Z_i.$$

Then $V_{n-1} = a\overline{Z}_{n-1} - b$ is an unbiased estimate for $p(1-\delta)$, where $a = .5[\Phi(1) - \Phi(-1)]^{-2} \simeq 1.0731$ and $b = 2a\,\Phi(1)\,\Phi(-1) \simeq .2865$.

Proof

(i) Since $E(X_i|\theta_i) = 2\theta_i - 1$ and $E(\theta_i) = 1-p$ the proof of (i) is immediate.

(ii) We have $Z_i = [\text{sgn } X_i \neq \text{sgn } X_{i+1}]$

$$= [X_i \geq 0,\ X_{i+1} < 0] + [X_i < 0,\ X_{i+1} \geq 0]\ .$$

Then,

$$EZ_i = 2P(X_i \geq 0,\ X_{i+1} < 0) \tag{2.9}$$

and the RHS (2.9) is evaluated by partitioning on (θ_i, θ_{i+1}) as follows.

Let $\quad \text{RHS}(2.9) = 2(p_1+p_2+p_3+p_4)$

where

$$p_1 = P(X_i \geq 0,\ X_{i+1} < 0,\ \theta_i = 0,\ \theta_{i+1} = 0)$$

$$p_2 = P(X_i \geq 0,\ X_{i+1} < 0,\ \theta_i = 0,\ \theta_{i+1} = 1)$$

$$p_3 = P(X_i \geq 0,\ X_{i+1} < 0,\ \theta_i = 1,\ \theta_{i+1} = 0)$$

and

$$p_4 = P(X_i \geq 0,\ X_{i+1} < 0,\ \theta_i = 1,\ \theta_{i+1} = 1).$$

p_1, p_2, p_3, p_4 can be evaluated without much difficulty. For example,

$$\begin{aligned} p_1 &= P(X_i \geq 0,\ X_{i+1} < 0 | \theta_i = 0,\ \theta_{i+1} = 0)\ P(\theta_i = 0,\ \theta_{i+1} = 0) \\ &= P(X_i \geq 0 | \theta_i = 0) P(X_{i+1} < 0 | \theta_{i+1} = 0) \delta p \\ &= (1 - \Phi(1))\Phi(1)\delta p \\ &= \Phi(-1)\Phi(1)\delta p \end{aligned}$$

and similar computations yield

$$p_2 = [\Phi(-1)]^2(1-\delta)p$$

$$p_3 = [\Phi(1)]^2 p(1-\delta)$$

$$p_4 = \Phi(1)\Phi(-1)[1-p-p(1-\delta)]\ .$$

Then

$$\text{RHS } (2.9) = 2p(1-\delta)\ [\Phi(1) - \Phi(-1)]^2 + 2\Phi(1)\Phi(-1).$$

Thus,

$$E\left[\frac{Z_i - 2\Phi(1)\Phi(-1)}{2(\Phi(1)-\Phi(-1))^2}\right] = p(1 - \delta),$$

and this completes the proof. □

Since $0 \leq p, \delta \leq 1$, it is natural to truncate U_n and V_{n-1} at 0 and 1 and use the modified estimators

$$U_n' = U_n[0 < U_n < 1] + [U_n \geq 1]$$

$$V_{n-1}' = V_{n-1}[0 < V_{n-1} < 1] + [V_{n-1} \geq 1].$$

The unknown parameters p and δ are estimated by solving the equations

$$\begin{aligned} p &= u_n' \\ p(1 - \delta) &= v_{n-1}'. \end{aligned} \tag{2.10}$$

2.4 Classification by Extended Compound Bayes Rules

We begin this section with a brief review of the Γ^k construct introduced by Gilliland and Hannan (1969) when applied to the finite–state, finite–action compound decision problem. Consider a component decision problem with states $\theta \in \Theta = \{1,2,.....,m\}$ indexing possible distributions $P_\theta \in \mathcal{P} = \{P_1,P_2,.....,P_m\}$ where P_i (i=1,2,.....,m) are distinct probability measures on $(\mathcal{X}, \mathcal{B})$. The actions $a \in \mathcal{A} = \{1,2,.....,n\}$. Let the loss function $L(\theta, a)$ be such that

$$0 \leq L(\theta,a) < \infty \quad \text{for all} \quad \theta \in \Theta \text{ and } a \in \mathcal{A}.$$

Let φ be a component decision rule. Then φ is a $\mathcal{B}$–measurable mapping into $\mathcal{A}^*$, the (n–1)–dimensional simplex of probability measures on $\mathcal{A}$. Let $R(i,\varphi)$ denote the risk of the component decision rule φ at state i. Then

$$R(i,\varphi) = \int [\sum_{j=1}^{n} L(i,j)\ \varphi_j]\ dP_i\ .$$

Let $\underline{P} = P_{\theta_1} \times P_{\theta_2} \times \ldots\ldots \times P_{\theta_N}$ and $\underline{\varphi} = (\varphi_1, \varphi_2, \ldots\ldots, \varphi_N)$ where for each α, φ_α is a $\mathcal{B}^N$–measurable mapping into $\mathcal{A}^*$. Let $R_\alpha(\underline{\theta}, \underline{\varphi})$ denote the risk in the α component decision for the compound rule $\underline{\varphi}$ and let $\underline{R}(\underline{\theta}, \underline{\varphi})$ denote the compound risk at N repetitions. Then

$$R_\alpha(\underline{\theta}, \underline{\varphi}) = \int [\sum_{j=1}^{n} L(\theta_\alpha, j)\ \varphi_{\alpha,j}] d\underline{P}$$

and

$$\underline{R}(\underline{\theta}, \underline{\varphi}) = \frac{1}{N} \sum_{\alpha=1}^{N} R_\alpha\ (\underline{\theta}, \underline{\varphi})\ .$$

For the above description of the component decision problem, we will now describe the Γ^k construct. The Γ^k decision problem has states $\underline{\theta}_k = (\theta_1, \theta_2, \ldots\ldots, \theta_k) \in \Theta^k$, observations $\underline{X}_k = (X_1, X_2, \ldots\ldots, X_k)$ distributed as $P_{\underline{\theta}_k} = P_{\theta_1} \times P_{\theta_2} \times \ldots\ldots \times P_{\theta_k} \in \mathcal{B}^k$ and θ_k is to be classified. The loss matrix L^k is $m^k \times n$ with $L^k(\underline{\theta}_k, a) = L(\theta_k, a)$.

Let $R^k(\underline{\theta}_k, \varphi)$ denote the risk of a decision rule φ in the Γ^k decision problem at state $\underline{\theta}_k$. Then

$$R^k(\underline{\theta}_k, \varphi) = E[L(\theta_k, \varphi_j(\underline{X}_k))]$$

$$= \int [\sum_{j=1}^{n} L(\theta_k, j)\ \varphi_j(\underline{x}_k)] dP_{\underline{\theta}_k}$$

$$= \int [\sum_{j=1}^{n} L(\theta_k, j)\ \varphi_j(\underline{x}_k) f_{\underline{\theta}_k}(\underline{x}_k) d\mu^k(\underline{x}_k)$$

where $f_{\underline{\theta}_k}(\underline{x}_k) = \prod_{i=1}^{k} f_{\theta_i}(x_i)$.

Letting $R^k(G,\varphi)$ denote the Bayes risk of φ versus a prior G on Θ^k, we have

$$R^k(G,\varphi) = E[R^k(\underline{\theta}_k,\varphi)]$$

$$= \sum_{\underline{\theta}_k} R^k(\underline{\theta}_k,\varphi)G_{\underline{\theta}_k}$$

$$(2.11) \qquad = \int\{\sum_{j=1}^{n} \varphi_j(\underline{x}_k)[\sum_{\underline{\theta}_k} L(\theta_k,j)f_{\underline{\theta}_k}(\underline{x}_k)G_{\underline{\theta}_k}]\}d\mu^k(\underline{x}_k) .$$

The Bayes envelope in the Γ^k decision problem is given by

$$R^k(G) = \inf_{\varphi} R^k(G,\varphi).$$

Let

$$(2.12) \qquad \Delta_j(\underline{x}_k) = \sum_{\underline{\theta}_k} L(\theta_k,j)f_{\underline{\theta}_k}(\underline{x}_k)G_{\underline{\theta}_k}.$$

From (2.11) it follows that a Γ^k Bayes rule has all of its mass placed on the j's which minimize $\Delta_j(\underline{x}_k)$.

The incorporation of asymptotic risk objectives, for the compound decision problem, more stringent than the usual standard $R^1(G_N)$ is an important development in compound decision theory. This was discussed in Section 1.3 of Chapter 1. The advantage of using these extended rules in classification problems is due to the fact that these rules benefit from empirical dependencies in the state sequence $\theta_1,\theta_2,.....,\theta_N$. Ballard and Gilliland (1978) have studied the finite risk performance of extended decision rules for the sequence version of the compound decision problem with Robbins' component.

The criterion for the classification rules is minimizing the risk and the concept of the Γ^k decision problem is useful for constructing rules for this purpose. We consider both the Γ^2 Bayes and Γ^3 Bayes rules. For the application of Γ^2 to the compound decision problem, the θ_i are grouped as $\underline{\theta}_i^2 = (\theta_{i-1}, \theta_i)$ with θ_i to be classified; for the application of Γ^3 to the compound decision problem the θ_i are grouped as $\underline{\theta}_i^3 = (\theta_{i-1}, \theta_{i+1}, \theta_i)$ with θ_i to be classified. For each $\{\theta_i\}$ and n let G_{n-1}^2 be the empirical distribution of $\underline{\theta}_2^2, \underline{\theta}_3^2, \ldots\ldots, \underline{\theta}_n^2$ and let G_{n-2}^3 be the empirical distribution of $\underline{\theta}_2^3, \underline{\theta}_3^3, \ldots\ldots, \underline{\theta}_{n-1}^3$.

For the Γ^2 decision problem, the state space is

$$\Theta^2 = \{(i, j): i, j = 0, 1\}$$

and we let $\underline{p}^2 = \{p_{ij}: i, j=0,1\}$ be a probability measure on Θ^2.

Taking $L(\theta,a) = L_1(\theta,a) = [\ \theta \neq a\ ]$ in (2.12), we obtain

$$\Delta_0(\underline{x}_2) = f_0(x_1)f_1(x_2)p_{01} + f_1(x_1)f_1(x_2)p_{11}$$

and

$$\Delta_1(\underline{x}_2) = f_0(x_1)f_0(x_2)p_{00} + f_1(x_1)f_0(x_2)p_{10}\ .$$

A Γ^2 Bayes rules versus $\underline{p}^2$ is given by

$$(2.13) \qquad \text{decide } \theta_k = \begin{matrix} 1 \\ \\ 0 \end{matrix} \quad \text{if } f_1(X_k) \sum_{i=0}^{1} p_{i1} f_i(X_{k-1}) \begin{matrix} \geq \\ \\ < \end{matrix} f_0(X_k) \sum_{i=0}^{1} p_{i0} f_i(X_{k-1}).$$

This is equivalent to

$$\text{decide } \theta_k = \begin{matrix} 1 \\ \\ 0 \end{matrix} \text{ if } X_k \begin{matrix} \geq \\ \\ < \end{matrix} \frac{1}{2} \log \frac{\sum_{i=0}^{1} p_{i0} f_i(X_{k-1})}{\sum_{i=0}^{1} p_{i1} f_i(X_{k-1})}$$

which can be written as

$$\text{(2.14)} \quad \text{decide } \theta_k = \begin{matrix} 1 \\ \\ 0 \end{matrix} \text{ if } X_k \begin{matrix} \geq \\ \\ < \end{matrix} c(p(X_{k-1}))$$

where

$$p(x) = \frac{\sum_{i=0}^{1} p_{i1} f_i(x)}{\sum_{i=0}^{1} \sum_{j=0}^{1} p_{ij} f_i(x)}$$

and

$$c(p) = \frac{1}{2} \log \left[\frac{1-p}{p}\right] .$$

The Bayes envelope in the Γ^2 decision problem is

$$R^2(p^2) = \int_{x_1} \int_{\{x_2 : x_2 \leq c(p(x_1))\}} [f_0(x_1)p_{01} + f_1(x_1)p_{11}] f_1(x_2) dx_1 dx_2$$

$$+ \int_{x_1} \int_{\{x_2 : x_2 > c(p(x_1))\}} [f_0(x_1)p_{00} + f_1(x_1)p_{10}] f_0(x_2) dx_1 dx_2$$

$$= \int_{x_1} [f_0(x_1)p_{01} + f_1(x_1)p_{11}]\Phi(c(p(x_1)) - 1)dx_1$$

$$+ \int_{x_1} [f_0(x_1)p_{00} + f_1(x_1)p_{10}][1 - \Phi(c(p(x_1)) + 1)]dx_1$$

$$= \sum_{i=0}^{1} \int f_i(x_1)[p_{i0}\{1 - \Phi(c(p(x_1)) + 1)\} + p_{i1}\Phi(c(p(x_1)) - 1)]dx_1. \tag{2.15}$$

For the Γ^3 decision problem, the state space is

$$\Theta^3 = \{(i, j, k): i, j, k = 0, 1\}$$

and we let $p^3 = \{p_{ijk}: i, j, k = 0, 1\}$ be a probability measure on Θ^3.

In this case (2.12) gives

$$\Delta_0(\underline{x}_3) = \sum_{i=0}^{1} \sum_{k=0}^{1} f_{i0k}(\underline{x}_3)p_{i0k}$$

$$= f_0(x_2) \sum_{i=0}^{1} \sum_{k=0}^{1} f_i(x_1)f_k(x_3)p_{i0k}$$

and

$$\Delta_1(\underline{x}_3) = \sum_{i=0}^{1} \sum_{k=0}^{1} f_{i1k}(\underline{x}_3)\, p_{i1k}$$

$$= f_1(x_2) \sum_{i=0}^{1} \sum_{k=0}^{1} f_i(x_1)f_k(x_k)p_{i1k}\,.$$

The Γ^3 Bayes rule versus $\underline{p}^3$ is given by

$$\text{(2.16)} \qquad \text{decide } \theta_\ell = \begin{matrix} 1 \\ 0 \end{matrix} \text{ if}$$

$$f_1(X_\ell) \sum_{i=0}^{1} \sum_{k=0}^{1} p_{i1k} f_i(X_{\ell-1}) f_k(X_{\ell+1}) \begin{matrix} \geq \\ < \end{matrix} f_0(X_\ell) \sum_{i=0}^{1} \sum_{k=0}^{1} p_{i0k} f_i(X_{\ell-1}) f_k(X_{\ell+1}) .$$

This is equivalent to

$$\text{decide } \theta_\ell = \begin{matrix} 1 \\ 0 \end{matrix} \text{ if}$$

$$X_\ell \begin{matrix} \geq \\ < \end{matrix} \frac{1}{2} \log \frac{\sum_{i=0}^{1} \sum_{k=0}^{1} p_{i0k} f_i(X_{\ell-1}) f_k(X_{\ell+1})}{\sum_{i=0}^{1} \sum_{k=0}^{1} p_{i1k} f_i(X_{\ell-1}) f_k(X_{\ell+1})}$$

which can be written as

$$\text{(2.17)} \qquad \text{decide } \theta_\ell = \begin{matrix} 1 \\ 0 \end{matrix} \text{ if } X_\ell \begin{matrix} \geq \\ < \end{matrix} c(p(X_{\ell-1}, X_{\ell+1}))$$

where

$$p(x, y) = \frac{\sum_{i=0}^{1} \sum_{k=0}^{1} p_{i1k} f_i(x) f_k(y)}{\sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{k=0}^{1} p_{ijk} f_i(x) f_k(y)} .$$

It can be shown that the Bayes envelope in the Γ^3 decision problem is

$$R^3(\underline{p}^3) = \int_{x_1}\int_{x_3}\Big[\int_{\{x_2:x_2\leq c(p(x_1,x_3))\}}\Delta_1(\underline{x}_3)dx_2 + \int_{\{x_2:x_2>c(p(x_1,x_3))\}}\Delta_0(\underline{x}_3)dx_2\Big]\,dx_1dx_3$$

$$= \sum_{i=0}^{1}\sum_{k=0}^{1}\int_{x_1}\int_{x_3} f_i(x_1)f_k(x_3)[p_{i0k}\{1-\Phi(c(p(x_1,x_3))+1)\} + p_{i1k}\,\Phi(c(p(x_1,x_3))-1)]\,dx_1dx_3\ .$$

The estimation of the Γ^2 and Γ^3 Bayes rules versus empirics in the compound problem depends on the construction of estimates for G_n^2 and G_n^3. The estimates for these empirical distributions are then plugged into (2.14) and (2.17) in place of $\underline{p}^2$ and $\underline{p}^3$ yielding extended compound Bayes rules. We follow Hannan (1957), Robbins (1964), Van Ryzin (1966) and Ballard (1974) in the construction of unbiased estimates.

Definition 2.1

$\underline{h} = (h_0, h_1)$ is said to be an unbiased estimator of the family $\mathcal{P} = \{P_0, P_1\}$ if

$$E_i h_j(X) = [\ i = j\ ] \qquad i,j = 0, 1$$

where E_i denotes expectation with respect to P_i .

From $\underline{h}$ one obtains unbiased estimators of $\mathcal{P}^k$ by taking

$$h_{\underline{i}_k}(\underline{x}_k) = h_{i_1}(x_1)h_{i_2}(x_2).\ .\ .\ .\ .h_{i_k}(x_k) \quad \text{for all } \underline{i}_k \in \Theta^k \ .$$

Such an estimator is called a product estimator [Ballard (1974)]. The unbiased estimator we use for $\{P_0, P_1\}$ is the kernel function used by Robbins (1951):

$$\underline{r}(x) = (r_0(x), r_1(x))$$

where $r_0(x) = \frac{1}{2}(1 - x)$ and $r_1(x) = \frac{1}{2}(1 + x)$.

These are used to produce estimators for the components of G_n^2 and G_n^3 as follows:

$$\hat{G}^2_{n-1;ij} = \frac{1}{n-1} \sum_{\alpha=2}^{n} r_i(X_{\alpha-1})r_j(X_\alpha) \ , \qquad n \geq 2 \ , \tag{2.18}$$

and

$$\hat{G}^3_{n-2;ijk} = \frac{1}{n-2} \sum_{\alpha=2}^{n-1} r_i(X_{\alpha-1})r_j(X_\alpha)r_k(X_{\alpha+1}) \ , \quad n \geq 2 \ . \tag{2.19}$$

These estimators are unbiased and consistent for the components of G^2_{n-1} and G^3_{n-2}. The extended compound rule ER2 is formed by substituting (2.18) for p_{ij} in (2.13) and letting $k = 2,3,.....n$. Similarly, the extended compound rule ER3 is formed by substituting (2.19) for p_{ijk} in (2.16) and letting $\ell = 2,3,.....,n-1$.

2.5 Comparison of the Methods

Even though we have suggested two types of classification rules, MAP and extended, our main interest is in the performance of the extended rules. Chapter 4 describes in more detail the simulations and the classifications obtained using various rules. As an example, we present some of the PCC comparisons for the extended rule ER3 and its competitor HEB3 defined following (1.7) in Chapter 1. Table 12 in Chapter 4 gives more detailed information on the performance of these two rules.

In 100 simulations for each of 84 different (p, δ) combinations:

A. with n = 50 components

(i) In 76 (90%) cases the ER3 has a higher PCC than the HEB3. The improvement of ER3 [=PCC of ER3 – PCC of HEB3] has a mean of 3.74 with standard deviation 2.62.

(ii) In 8 (10%) cases the HEB3 has a higher PCC than the ER3. The improvement of HEB3 has a mean of .82 with standard deviation .48.

B. with n = 200 components.

(i) In 75 (89%) cases the ER3 has a higher PCC than the HEB3. The improvement of ER3 has a mean of 2.30 with standard deviation 2.25.

(ii) In 9 (11%) cases the HEB3 has a higher PCC than the ER3. The improvement of HEB3 has a mean of .33 with standard deviation .24

It was also observed that the ER has a very high improvement over HEB when either p is very small ($< .1$) or δ is very large ($>.9$).

In implementing the empirical MAP rule (2.8), the unknown parameters (p,δ) were estimated by using Proposition 2.1 and the following convention was used in cases where either the estimates were undefined or truncation was necessary. Let u,v denote u_n , v_{n-1} respectively.

(i) If $\{u \leq 0\} \cup \{0 < u \leq .5 , v \leq 0\}$ then decide all $\theta_i = 1$.

(ii) If $\{u > .5 , v \leq 0\}$ then decide all $\theta_i = 0$.

(iii) If $\{u \geq 1 , v \geq 1\} \cup \{.5 < u < 1 , u \leq v\}$ then decide $\theta_{2i+1} = 0$ and $\theta_{2i} = 1$.

(iv) MAP rule (2.8) for the remaining (u,v).

We used the estimators following from (2.10) for substitution into the HEB3 and HEB5 rules. Hill et al. (1984) also have proposed estimators (1.5) for (p,δ) which they apparently used to implement their empirical Bayes rules. The method of estimating (p,δ) did not effect the performance of the empirical MAP rule. However the same was not true for the HEB rules. The HEB rule with j=1 (HEB3) and the parameters estimated by (1.5) had a better performance than the same rule with the parameters estimated by (2.10) whereas the HEB rule with j=2 (HEB5) and the parameters estimated by (2.10) had a better performance than the same rule with the parameters estimated by (1.5). However even with the improvement, the HEB rules remained inferior in performance to the ER3 rules.

CHAPTER 3

THE IMAGE CASE

3.1 Introduction

In this chapter we consider the following classification problem. Suppose we have a rectangular array of pixels and on pixel (i,j) we have the random observation X_{ij}. Pixel (i,j) has a true but unknown classification θ_{ij}. The $\{\theta_{ij}\}$ gives rise to a two dimensional image. The observed image $\{X_{ij}\}$ has a distribution depending on $\{\theta_{ij}\}$. Because of the spatial correlations inherent in images, $\{\theta_{ij}\}$ is considered to be a realization of a Markov random field (MRF) in some models. This can be viewed as a variation of Robbin's empirical Bayes problem, with the parameters being dependent. As in Chapter 2, tools developed for extended compound decision problem will be applied to derive decision rules based on $\underline{X}$ to reconstruct the image. Here $\underline{\theta}$ and $\underline{X}$ denote restrictions to an $N_1 \times N_2$ rectangle of indices and $n = N_1 N_2$.

In Chapter 2 the sequence $\{\theta_i\}$ was assumed to be a realization of a two-state Markov chain. This prior distribution can be considered as a special case of a more general class of Markov distributions for $\{\theta_i\}$. To see this and to set the stage for the introduction of a MRF, we rewrite (2.3) as follows:

$$p(\underline{\theta}) = p\, p_{00}^{n-1} \left[\frac{1-p}{p}\frac{p_{10}}{p_{00}}\right]^{\theta_1} \left[\frac{p_{01}}{p_{00}}\right]^{\theta_n} \left[\frac{p_{01}p_{10}}{p_{00}^2}\right]^{\sum_{i=2}^{n-1}\theta_i} \left[\frac{p_{00}p_{11}}{p_{01}p_{10}}\right]^{\sum_{i=2}^{n}\theta_{i-1}\theta_i}$$

$$(3.1) \qquad = c(n, p, \delta) \exp\Big[-\sum_{i=1}^{n}\alpha_i\,\theta_i - \beta\sum_{i=2}^{n}\theta_{i-1}\,\theta_i\Big]$$

where

$$c(n,\ p,\ \delta) \ = \ pp_{00}^{n-1} \ = \ p\delta^{n-1},$$

$$\alpha_1 \ = \ -\log\left[\frac{1-p}{p}\ \frac{p_{10}}{p_{00}}\right] \ ,$$

$$\alpha_i \ = \ -\log\left[\frac{p_{01}p_{10}}{p_{00}^2}\right] \ , \qquad\qquad 1 < i < n \ ,$$

$$\alpha_n \ = \ -\log\left[\frac{p_{01}}{p_{00}}\right],$$

and

$$\beta \ = \ -\log\left[\frac{p_{00}p_{11}}{p_{01}p_{10}}\right].$$

Let $\tilde{\underline{\theta}}_i$ denote the vector obtained by deleting the i^{th} component from $\underline{\theta}$.

$$\tilde{\underline{\theta}}_i \ = \ (\theta_1,\theta_2,.....,\theta_{i-1},\theta_{i+1},.....\theta_n)$$

If $\underline{\theta}$ is a Markov chain, then it can be shown that

$$(3.2) \qquad P(\theta_i|\tilde{\underline{\theta}}_i) \ = \ \begin{cases} P(\theta_1|\theta_2) & i \ = \ 1 \\ P(\theta_i|\theta_{i-1},\theta_{i+1}) & 1 < i < n \\ P(\theta_n|\theta_{n-1}) & i \ = \ n \end{cases}$$

and from (3.1) for $1 < i < n$

$$(3.3) \qquad P(\theta_i|\theta_{i-1},\ \theta_{i+1}) \ = \ \frac{e^{-\alpha_i\theta_i-\beta\theta_i(\theta_{i-1}+\theta_{i+1})}}{1+e^{-\alpha_i-\beta(\theta_{i-1}+\ \theta_{i+1})}} \ .$$

From (3.3) an expression for the log odds ratio for the i^{th} class is given by

$$\log \frac{P(\theta_i=1 \mid \theta_{i-1},\theta_{i+1})}{P(\theta_i=0 \mid \theta_{i-1},\theta_{i+1})} = -\alpha_i - \beta(\theta_{i-1} + \theta_{i+1}) .$$

The expressions (3.1) and (3.2) can be generalized into more complex situation than the one dimensional Markov chain considered above. We shall concentrate on discrete two dimensional random fields defined over a finite $N_1 \times N_2$ rectangular lattice of pixels. Let

$$L = \{(i, j) \mid i \leq i \leq N_1,\ 1 \leq j \leq N_2\}.$$

Definition 3.1

A collection of subsets of L,

$$\mathcal{N} = \{N(i, j) \mid (i, j) \in L\},$$

is said to be a neighborhood system on L if for all $(i, j) \in L$,

(i) $(i, j) \notin N(i, j)$

and (ii) $(k, l) \in N(i, j)$ implies $(i, j) \in N(k, l)$.

N(i, j) is called the neighborhood of pixel (i, j).

A hierarchically ordered sequence of neighborhood systems that are commonly used in image segmentation and pattern recognition are $\mathcal{N}_1$, $\mathcal{N}_2$,..... where for each $(i, j) \in L$ (except for boundary pixels) $\mathcal{N}_1$ consists of four pixels neighboring (i, j), $\mathcal{N}_2$ consists of the eight pixels neighboring (i, j) and so on. The neighborhood systems $\mathcal{N}_1$, $\mathcal{N}_2$, $\mathcal{N}_3$, $\mathcal{N}_4$, and $\mathcal{N}_5$ are shown in Figure 1.

5	4	3	4	5
4	2	1	2	4
3	1	(i,j)	1	3
4	2	1	2	4
5	4	3	4	5

Figure 1. Neighborhood systems $\mathcal{N}_n$

In Figure 1, $\mathcal{N}_n = \{N_n(i, j)\}$ where symbolicaly $N_n(i, j) = \{k: k \leq n\}$. $\mathcal{N}_1$ is called the nearest neighbor (NN) system and $\mathcal{N}_n$ is called the n^{th} order neighborhood system.

Definition 3.2

Let $\mathcal{N}$ be a neighborhood system defined on L. A random field $\underline{\theta} = \{\theta_{ij} | (i, j) \in L\}$ is a Markov random field with respect to $\mathcal{N}$ if

$$P(\theta_{ij} | \tilde{\underline{\theta}}_{ij}) = P(\theta_{ij} | \underline{\theta}^N_{ij}) \tag{3.4}$$

for each $(i, j) \in L$, where $\underline{\theta}^N_{ij}$ denotes the restriction of $\underline{\theta}$ to $N(i, j)$.

The Markov chain on the transect $\underline{\theta}$ results in this structure by taking

$$N(1) = \{2\}$$
$$N(i) = \{i-1, i+1\}, \qquad 1 < i < n,$$
$$N(n) = \{n-1\};$$

(cf. (3.2) and (3.4)).

It should be noted that any random field is a MRF with respect to the neighborhood system $N(i, j) = L$ for all $(i, j) \in L$. A MRF is not characterized by specifying the conditional probabilities (3.4) yet (3.4) is termed the local characterization of the random field. A difficulty with the above characterization of a MRF is the unavailability of the joint distribution. These difficulties are overcome by characterizing a MRF through its Gibbs distribution. To do this we will introduce additional concepts.

Definition 3.3

A clique of the pair $(L, \mathcal{N})$, denoted C, is a subset of L such that

(i) C consists of a single pixel

or (ii) $(i, j) \in C$ and $(k, l) \in C \Rightarrow (i, j) \in N(k, l)$

for $(i, j) \neq (k, l)$.

The collection of all cliques of $(L, \mathcal{N})$ will be denoted by $\mathcal{C}$.

For the Markov chain

$$\mathcal{C} = \{\{1\}, \{2\},.....,\{n\}, \{1, 2\}, \{2, 3\},.....\{n-1, n\}\}.$$

The type of cliques associated with $\mathcal{N}_1$, $\mathcal{N}_2$ are shown in Figure 2.

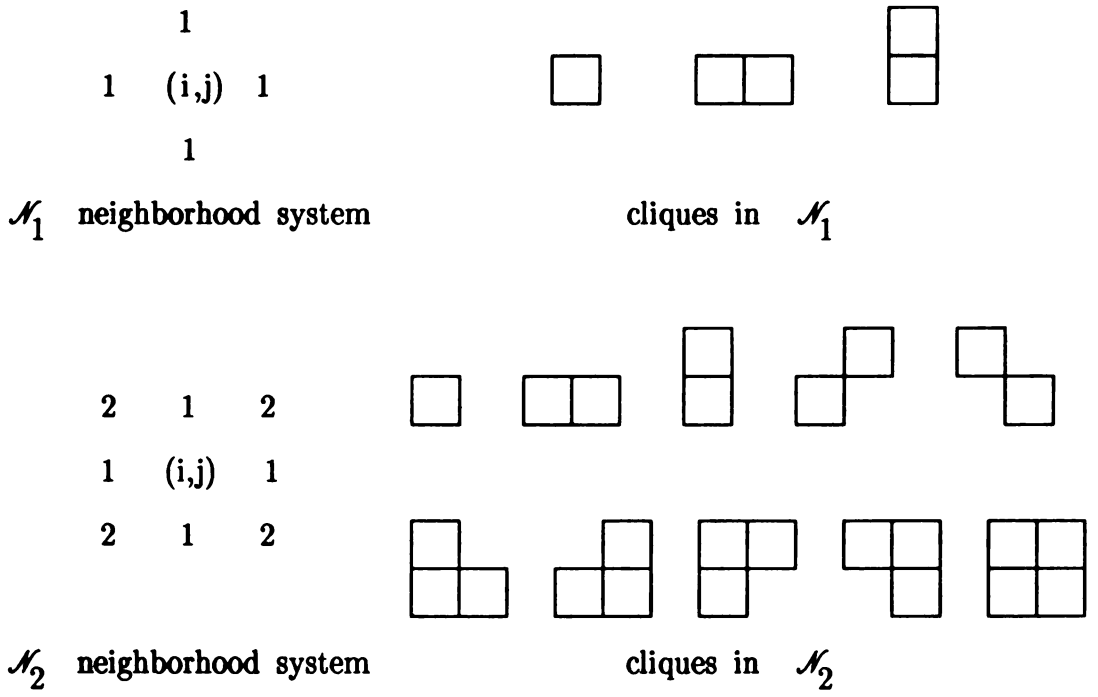


Figure 2. Neighborhood systems $\mathcal{N}_1$ and $\mathcal{N}_2$ and their associated cliques

Definition 3.4

Let $\mathcal{N}$ be a neighborhood system defined over a lattice L and $\mathcal{C}$ the collection of all cliques of $(L,\mathcal{N})$. A random field $\underline{\theta} = \{\theta_{ij} \mid (i, j) \in L\}$ has a Gibbs distribution (GD) or equivalently is a Gibbs random field (GRF) with respect to $\mathcal{N}$ if its joint distribution is of the form

$$p(\underline{\theta}) = \frac{e^{-U(\underline{\theta})}}{Z} \tag{3.5}$$

where

$$U(\underline{\theta}) = \sum_{C \in \mathcal{C}} V_C(\underline{\theta}_C)$$

$$Z = \sum_{\underline{\theta}} e^{-U(\underline{\theta})}.$$

The U, $V_C(\cdot)$ and Z in Definition 3.4 are termed, respectively, the energy function, the potential associated with clique C, and the partition function. The only condition on the otherwise fully arbitrary clique potential V_C is that it depends on the pixel values in clique C. That is, θ_{ij} and θ_{kl} may appear together in a term is $U(\cdot)$ only if (i, j) and (k, l) are neighbors.

Comparison of (3.1) and (3.5) gives

$$V_{\{i\}}(\theta_i) = \alpha_i \theta_i \quad , \qquad 1 \le i \le n \ ,$$

$$V_{\{i-1,i\}}(\theta_{i-1}, \theta_i) = \beta\theta_{i-1} \theta_i \quad , \qquad 2 \le i \le n \ .$$

The origin of GD lies in physics and statistical mechanics literature. The source of revived interest in GD, especially is the context of image segmentation and pattern recognition is due to an important result known as Hammersley–Clifford theorem. This result establishes a one–to–one correspondence between MRF and GRF. Unlike the MRF local characterization (3.4) the GD characterization (3.5) provides the joint distribution of the random field and is free from consistency problems. A detailed treatment of MRF and GRF can be found in Besag (1974, 1986), Kinderman and Snell (1980).

3.2 Introduction to Image Segmentation

The image segmentation process is an important component of scene analysis and reconstruction of noisy images. As such it has received considerable attention in computer vision and image processing literature. The general image segmentation problem can be viewed as a two dimensional classification problem. The data are collected in two spatial dimensions, specially on a rectangular lattice of points called pixels. Each pixel is assumed to have a true but unknown fixed classification.

Suppose, we have a $N_1 \times N_2$ rectangular array L of pixels and we observe $\underline{X}_{ij}$, a vector of intensities, at pixel (i, j). Let θ_{ij} denote the label of pixel (i, j) and

$$\theta_{ij} \in \Theta = \{1,2,.....,m\}.$$

The observed image is a random field $\underline{X} = \{X_{ij}\}$ defined over L. The image random field $\underline{X}$ is assumed to be a function of "scene random field" $\underline{\theta} = \{\theta_{ij}\}$ and a corruptive "noise random field" $\underline{Y}$ each defined over L. The functional relationship between the random fields is such that at each pixel the image random variable is a function of the scene random variable and the noise random variable at that pixel, that is

$$X_{ij} = f_{ij}(\underline{Y},\underline{\theta}) \qquad \text{for all } (i, j) \in L.$$

However, the interest is mainly in the case of additive noise where

$$f_{ij}(\underline{Y},\underline{\theta}) = Y_{ij} + \theta_{ij} \qquad \text{for all } (i, j) \in L.$$

The use of MRF as models for the distribution of the scene random field have been prevalent in the image segmentation literature for some time now. Geman and Geman (1984), Ripley (1986) are two, among the many papers which contain detailed references about the use of MRF in image segmentation. The MRF model has the spatial distributional properties that have proven to be useful in segmentation studies.

The objective in image segmentation problems is that, given a image realization $\underline{X}$, to determine the scene realization $\underline{\theta}$ that has given rise to $\underline{X}$. The scene realization $\underline{\theta}$ is, of course, invisible and can not be obtained deterministically from $\underline{X}$. So the problem is to obtain an estimate $\hat{\underline{\theta}} = \hat{\underline{\theta}}(\underline{X})$ of the scene, based on a realization $\underline{X}$. A Bayes decision theoretic approach requires the specification of a loss function. The two loss functions L_2 and L_1 introduced in Chapter 1 lead, respectively, to

(1) maximum a posteriori estimation

and

(2) maximization of the posterior marginal probability at each pixel.

A major reservation regarding (1) is purely computational. A brute force maximization is simply out of the question because there are m^n possible scene configurations for the maximization to be done over. Exceptions occur by design in the subclass of MRF considered. Derin et al. (1984), Geman and Geman (1984) have obtained MAP rules using a procedure from combinatorial optimization known as simulated annealing. The simulated annealing is used on a subclass of MRF's called Markov mesh random fields (MMRF), where an additional assumption that the columns of the scene constitute a Markov chain is made. Still there are computational difficulties in implementing the Bayes smoothing algorithm of Derin et al. (1984), Geman and Geman (1984). For example, for a binary image on L, at each step of the algorithm there are on the order of 2^{N_2} calculations to be performed. In order to overcome the above computational difficulties the image is processed in relatively narrow strips. The overlapping strips are chosen in such a way that the union of the middle sections of the strips give the entire image. Haslett (1985) has proposed an approximation to (2) based on maximum likelihood discriminant analysis for a subclass of MRF, called the Pickard (1977) random field (PRF). An important feature of PRF is that the columns and rows of $\underline{\theta}$ separately form a Markov chain. An easily proven fact for a Markov chain is that θ_{i-1} and θ_{i+1} are conditionally independent given θ_i. In Haslett's (1985) derivation of the classification rules, it is assumed that the four random variable $\theta_{i,j\pm1}$, $\theta_{i\pm1,j}$ are conditionally indpendent given θ_{ij}.

In the next section we introduce the empirical Bayes approach to image reconstruction which is based on estimating parameters of the MRF. In Section 3.4 we will give details of the construction of extended compound rules for the image application; these rules have favorable risk behavior that does not require distributional assumptions. We introduced the MRF because three of the tested images in our simulations were generated as a MRF.

3.3 Classification by Empirical Bayes Rules

In this section we describe the empirical Bayes rule proposed by Morris et al. (1985), Morris (1986) for the classification of a binary image, $\Theta = \{0,1\}$.

Let X_{ij} be the random observation on pixel (i, j) and θ_{ij} the true but unknown classification of the pixel (i, j), i,j = 1,2,..... . We let $\underline{\theta}$ and $\underline{X}$ be restrictions of $\{\theta_{ij}\}$ and $\{X_{ij}\}$ to the $N_1 \times N_2$ rectangular lattice L.

The specific empirical Bayes model considered by Morris et al. (1985), Morris (1986) has

(a) $X_{ij} \sim N(\cdot 5\delta(2\theta_{ij}-1), 1)$ independently, with δ a known constant, conditional on $\{\theta_{ij}\}$.

(b) $\{\theta_{ij}\}$ has a distribution G that is spatially isotropic (invariant under translations and rotations) with

$$p = P(\theta_{ij} = 0) \qquad \text{for all (i, j)}$$

and

$$\rho_{t,u} = \text{corr. } (\theta_{ij}, \theta_{i+t,j+u}) \qquad \text{for all (i, j) and (t, u).}$$

Even with the structure (a) and (b), the exact form for each marginal posterior probability $P(\theta_{ij}=1|\underline{X})$ is quite complicated. Morris et al. (1985), Morris (1986) suggest a logistic form as a good approximation to this probability, with accuracy increasing as $\delta \to 0$. Furthermore, the logistic function is approximated by a discriminant function, which predicts $P(\theta_{ij}=1|\underline{X})$ from the "ring" averages, these being averages of those data values in specified locations relative to the pixel (i, j). The ring locations $R_0, R_1, \ldots, R_5$ with center at pixel (i, j) are indicated by the integers in Figure 1 with R_0 being the singleton pixel (i, j). The four nearest neighbors marked 1 make up R_1, the next four marked 2 make up R_2, and so on.

Define the discriminant function $\lambda_{ij}(\underline{X})$ as

$$\lambda_{ij}(\underline{X}) = \log \frac{P(\theta_{ij}=1|\underline{X})}{P(\theta_{ij}=0|\underline{X})} . \tag{3.6}$$

The approximation proposed in Morris et al. (1985), Morris (1986) is based on a moving average of the form

$$\lambda_{ij}(\underline{X}) \simeq \log \frac{1-p}{p} + \sum_{t=0}^{r} \gamma_t \bar{X}_{ij,t} \tag{3.7}$$

where $\bar{X}_{ij,t}$ is the average of the observations in the ring R_t centered at pixel (i, j) and γ_t depends on $\underline{X}$ and $\underline{\ell}$. In (3.7),

$$\bar{X}_{ij,0} = X_{ij} , \quad \bar{X}_{ij,1} = \tfrac{1}{4}(X_{i,j-1} + X_{i-1,j} + X_{i,j+1} + X_{i+1,j})$$

and so on, with suitable modifications at the edges of the lattice.

Define the n×(r+1) matrix

$$Y = \begin{bmatrix} X_{11}-\bar{X} & \bar{X}_{11,1} - \bar{X}_1 & \cdots & \bar{X}_{11,r} - \bar{X}_r \\ X_{12}-\bar{X} & \bar{X}_{12,1} - \bar{X}_1 & \cdots & \bar{X}_{12,r} - \bar{X}_r \\ \vdots & \vdots & & \vdots \\ X_n-\bar{X} & \bar{X}_{n,1} - \bar{X}_1 & \cdots & \bar{X}_{n,r} - \bar{X}_r \end{bmatrix}$$

with $\bar{X}$, $\bar{X}_1,\ldots,\bar{X}_r$ the average of X_{ij}, $\bar{X}_{ij,1},\ldots,\bar{X}_{ij,r}$, respectively.

If θ_{ij} were observable, RHS(3.7) can be estimated by

$$\lambda_{ij}(\underline{x}) \simeq \log \frac{\bar{\theta}}{1-\bar{\theta}} + \frac{1}{RSS(\underline{\theta})} \sum_{t=0}^{r} b_t(\underline{\theta})\ (\bar{x}_{ij,t} - m_t(\underline{\theta})),$$

where $\bar{\theta} = \frac{1}{n} \sum_{i,j} \theta_{ij}$ and $b_t(\underline{\theta})$, $m_t(\underline{\theta})$, $RSS(\underline{\theta})$ defined as follows:

Let
$$\begin{aligned} \underline{b}(\underline{\theta}) &= (b_0(\underline{\theta}),\ b_1(\underline{\theta}),\ldots,b_r(\underline{\theta}))' \\ \underline{y} &= (\bar{x},\ \bar{x}_1,\ldots,\bar{x}_r)' \\ \underline{m}(\underline{\theta}) &= (m_0(\underline{\theta}),\ m_1(\underline{\theta}),\ldots,m_r(\underline{\theta}))' \\ C(\underline{\theta}) &= n^{-1}Y'\underline{\theta} \\ S &= n(Y'Y)^{-1}\ . \end{aligned}$$

Then

$$\begin{aligned} \underline{b}(\underline{\theta}) &= SC(\underline{\theta}) \\ \underline{m}(\underline{\theta}) &= \underline{y} + \frac{1-2\bar{\theta}}{2\bar{\theta}(1-\bar{\theta})} C(\underline{\theta}) \\ RSS(\underline{\theta}) &= \bar{\theta}(1-\bar{\theta}) - C'(\underline{\theta})SC(\underline{\theta}). \end{aligned}$$

(3.8) can be simplified by letting

$$\sum_{t=0}^{r} b_t(\underline{\theta})\ (\bar{x}_{ij,t} - m_t(\underline{\theta})) = S_1 + S_2$$

where
$$\begin{aligned} S_1 &= \sum_{t=0}^{r} b_t(\underline{\theta})\ (\bar{x}_{ij,t} - \bar{x}_t) \\ &= \underline{y}'_{(i,j)} b(\underline{\theta}), \qquad \underline{y}'_{(i,j)} \text{ is the } (i,\ j)^{th} \text{ row of } Y \\ &= n\underline{y}'_{(i,j)} (Y'Y)^{-1} C(\underline{\theta}) \end{aligned}$$

and

$$\begin{aligned} S_2 &= \sum_{t=0}^{r} b_t(\underline{\theta})\ (\bar{x}_t - m_t(\underline{\theta})) \\ &= \frac{2\bar{\theta}-1}{2\bar{\theta}(1-\bar{\theta})}\ C'(\underline{\theta}) b C(\underline{\theta}) \\ &= \frac{\bar{\theta}-.5}{\bar{\theta}(1-\bar{\theta})}\ nC'(\underline{\theta})(Y'Y)^{-1} C(\underline{\theta}). \end{aligned}$$

With the above simplifications (3.8) reduces to

$$(3.9) \quad \lambda_{ij}(\underline{x}) \simeq \log \frac{\bar{\theta}}{1-\bar{\theta}} + \frac{n}{RSS(\underline{\theta})}\left[\underline{y}'_{(i,j)} + \frac{(\bar{\theta}-.5)}{\bar{\theta}(1-\bar{\theta})}\ C'(\underline{\theta})\right](Y'Y)^{-1} C(\underline{\theta})\ .$$

Let c_t denote the sample autocovariance of elements in R_0 with those in R_t. That is

$$c_t = \frac{1}{n} \sum_{i,j} x_{ij}(\bar{x}_{ij,t} - \bar{x}_t) \qquad t = 1,2,....,r.$$

Proposition 3.1

Let $\hat{q} = .5 + \delta^{-1}\bar{X}$ and $\hat{C}(\underline{\theta}) = (\delta\hat{q}(1-\hat{q}), \delta^{-1}c_1 - n^{-2}n_1, \ldots, \delta^{-1}c_r - n^{-2}n_r)'$ where $\bar{X} = \frac{1}{n}\sum_{i,j} X_{ij}$ and n_t is the number of pixels in ring R_t. Then $\hat{q}$ and $\hat{C}(\underline{\theta})$ are unbiased estimates for θ and $C(\underline{\theta})$ respectively.

Proof: see Morris et al. (1985). □

In applications n_t is very small compared to n^2. Thus a nearly unbiased estimate of $c_t(\underline{\theta})$ is $\delta^{-1}c_t$ $(t = 1, 2, \ldots, r)$. Plugging in the estimates $\hat{q}$ and $\hat{C}(\underline{\theta})$ into (3.9), Morris et al. (1985), Morris (1986) obtain their empirical Bayes rule as

$$\text{decide } \theta_{ij} = \begin{cases} 1 & \text{if } \hat{\lambda}_{ij} \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

We investigate the risk behavior of the Morris rule based on the four nearest neighbors and based on the eight nearest neighbors. We call the rules MEB4 and MEB8.

3.4 Classification by Extended Rules

In this section tools developed for extended compound decision problem are used to obtain classification rules which enables us to reconstruct the image. These rules will be called extended rules.

Let $a_{ij} \in \mathcal{A} = \{0, 1\}$ be the action taken with respect to the pixel (i, j) and let $L(\theta_{ij}, a_{ij})$ be the loss incurred by taking action a_{ij} when the true classification of the pixel (i, j) is θ_{ij}. As in Section 2.4 assume that

$$0 \leq L(\theta, a) < \infty \qquad \theta \in \Theta \quad a \in \mathcal{A}$$

The average loss suffered over the n classifications is

$$L(\underline{\theta}, \underline{a}) = \frac{1}{n} \sum_{i,j} L(\theta_{ij}, a_{ij}(\underline{X}))$$

and the risk

$$R(\underline{\theta}, \underline{a}) = E\,[L(\underline{\theta}, \underline{a})]. \tag{3.10}$$

In order to apply the extended compound approach, that was introduced in Section 1.2, to the image case, we choose a neighborhood system for the $N_1 \times N_2$ lattice. Our computations will concern compound rules designed to achieve the asymptotic risk behavior (1.17) for a particular neighborhood structure and $k = 5$. Specifically, that neighborhood system is the nearest neighbor system $\mathcal{N}_1$ defined in Section 3.1. For simplicity of exposition, we develop our ideas in terms of this system although all results hold more generally.

It is also convenient to measure risk on the sublattice

$$\{(i, j) \mid 2 \leq i \leq N_1-1,\ 2 \leq j \leq N_2-1\}$$

since here every lattice point has a complete neighborhood $N(i, j)$ consisting of four nearest neighbors. On the sublattice, $n = (N_1-2)(N_2-2)$ classifications are made.

As in Section 1.2, we let $\underline{\theta}_{ij}$ and $\underline{X}_{ij}$ denote restrictions of $\underline{\theta}$ and $\underline{X}$ to the positions $N(i, j) \cup \{(i, j)\}$. For our neighborhood system, $k = 5$ and $\underline{\theta}^5_{ij} = (\theta_{i,j-1}, \theta_{i-1,j}, \theta_{i,j+1}, \theta_{i+1,j}, \theta_{ij})$.

We now derive the form of the Γ^5 rule that has compound risk $R^5(G^5_n)$. Let $\underline{d}$ be a compound rule such that

$$d_{ij}(\underline{X}) = d(\underline{X}_{ij}), \qquad 2 \le i \le N_1-1,\ 2 \le j \le N_2-1 , \tag{3.11}$$

where d is a Γ^5 decision rule.

Notations

$\underline{X}^k$: k–vector of observations

$\underline{\theta}^k$: k–vector of parameters

$\underline{X}^k_{ij}$: k–vector of observations with X_{ij} as k^{th} component.

$\underline{\theta}^k_{ij}$: k–vector of observations with θ_{ij} as k^{th} component.

From (3.10) the compound risk of a rule of the form (3.11) over the subarray is

$$R(\underline{\theta}, d) = E[\tfrac{1}{n} \sum_{i,j} L(\theta_{ij}, d(\underline{X}^k_{ij}))]$$

$$= \frac{1}{n} \sum_{\underline{\theta}^k \in \Theta^k} \sum_{\{(i,j):\, \theta_{ij}=\theta_k\}} E[L(\theta_k, d(\underline{X}^k_{ij})]$$

$$= \sum_{\underline{\theta}^k \in \Theta^k} \frac{1}{n} \sum_{\{(i,j):\, \theta_{ij}=\theta_k\}} \int L(\theta_k, d(\underline{x}^k)) f(\underline{x}^k | \underline{\theta}^k) d\underline{x}^k . \tag{3.12}$$

Letting $G_n^k(\underline{\theta}^k)$ denote the relative frequency of the occurrence of $\underline{\theta}^k$ in the subarray $\underline{\theta}$, (3.12) can be written as

$$R(\underline{\theta}, d) = \sum_{\underline{\theta}^k \in \Theta^k} G_n^k(\underline{\theta}^k) \int L(\theta_k, d(\underline{x}^k)) f(\underline{x}^k | \underline{\theta}^k) d\underline{x}^k$$

$$= \int \sum_{\underline{\theta}^k \in \Theta^k} G_n^k(\underline{\theta}^k) L(\theta_k, d(\underline{x}^k)) f(\underline{x}^k | \underline{\theta}^k) d\underline{x}^k. \tag{3.13}$$

Note that G_n^k is simply the empirical distribution of the $\underline{\theta}_{ij}$ across the subarray.

A Γ^k Bayes rule d^* with respect to G_n^k is obtained by minimizing the integrand in (3.13) for each $\underline{X}^k$. Thus,

$$d^*(\underline{X}_{ij}) = \text{the action a which minimizes}$$
$$\sum_{\underline{\theta}^k \in \Theta^k} G_n^k(\underline{\theta}^k) L(\theta_k, a) f(\underline{X}_{ij} | \underline{\theta}^k) . \tag{3.14}$$

Taking $L(\theta, a) = L_1(\theta, a) = [\theta \neq a]$, (3.14) reduces to

$$d^*(\underline{X}_{ij}) = \text{the action a which minimizes}$$
$$\sum_{\substack{\underline{\theta}^k \in \Theta^k \\ \theta_k \neq a}} G_n^k(\underline{\theta}^k) f(\underline{X}_{ij} | \underline{\theta}^k)$$

$$= \text{the action a which maximizes}$$
$$\sum_{\substack{\underline{\theta}^k \in \Theta^k \\ \theta_k = a}} G_n^k(\underline{\theta}^k) f(\underline{X}_{ij} | \underline{\theta}^k) . \tag{3.15}$$

Since the X_{ij}'s are assumed to be class conditionally independent

$$f(\underline{X}_{ij}|\underline{\theta}^k) = \prod_{\alpha=1}^{k} f(X_\alpha|\theta_\alpha)$$

and (3.15) becomes

$$d^*(\underline{X}_{ij}) = \text{the action a which maximizes}$$

$$\text{(3.16)} \qquad \left[\sum_{\substack{\underline{\theta}^k \in \Theta^k \\ \theta_k = a}} G_n^k(\underline{\theta}^k) \prod_{\alpha=1}^{k-1} f(X_\alpha|\theta_\alpha)\right] f(X_{ij}|a).$$

Taking $k = 5$ and

$$\underline{\theta}^5_{ij} = (\theta_{i,j-1}, \theta_{i-1,j}, \theta_{i,j+1}, \theta_{i+1,j}, \theta_{ij})$$

with θ_{ij} to be classified, (3.16) gives the Γ^5 Bayes rule as

(3.17) decide $\theta_{ij} = 1$ if

$$\left[\sum_{\substack{\underline{\theta}^5 \in \Theta^5 \\ \theta_5 = 1}} G_n^5(\underline{\theta}^5) \prod_{\alpha=1}^{4} f(X_\alpha|\theta_\alpha)\right] f(X_{ij}|1) \geq \left[\sum_{\substack{\underline{\theta}^5 \in \Theta^5 \\ \theta_5 = 0}} G_n^5(\underline{\theta}^5) \prod_{\alpha=1}^{4} f(X_\alpha|\theta_\alpha)\right] f(X_{ij}|0)$$

and

$$\theta_{ij} = 0 \qquad \text{otherwise.}$$

As in Section 2.4, we will estimate $G_n^5(\underline{\theta}^5)$ by unbiased product estimators. For example, $G_n^5(0,0,0,0,0)$ is estimated by $\hat{G}(0,0,0,0,0)$ where

$$\hat{G}(0,0,0,0,0)$$

$$= \frac{1}{(n_1-1)(n_2-1)} \sum_{i=2}^{n_1} \sum_{j=2}^{n_2} r_0(X_{i,j-1}) r_0(X_{i-1,j}) r_0(X_{i,j+1}) r_0(X_{i+1,j}) r_0(X_{ij}).$$

where $n_1 = N_1 - 1$ and $n_2 = N_2 - 1$.

We refer to the extended compound decision rule that is obtained by substituting the product estimates based on kernel r into (3.17) as ER4.

3.5 Comparison of the Methods

The performance of the extended rule (ER) of Section 3.4 was compared with the empirical Bayes rule (MEB) of Section 3.3 based on 2 different sets of neighbors, the four nearest neighbors (R_1) and the eight nearest neighbors (R_1 and R_2). The rules were tested on both deterministic and stochastic images, details of which is given in Chapter 4. In a total of 150 simulations with 6 different images, in 149(99%) cases the ER had a higher PCC than the MEB based on 4 neighbors and in 142 (95%) cases the ER had a higher PCC than the MEB based on 8 neighbors. We also considered revised classifications using the ER, which were done in the following way. After obtaining the initial classification by the ER, revised estimates of $G_n^k(\underline{\ell}^k)$ were obtained by computing the relative frequencies of $\underline{\ell}^k$ from the reconstructed image and then (3.17) was used to obtain the revised classification. The process was repeated. It was observed that in most cases there was no change in the PCC after the third revision. Furthermore all the revised classifications had a higher PCC than both of the empirical Bayes rules.

CHAPTER 4

SIMULATIONS.

4.1 Introduction.

Consider the component decision problem with $X|\theta \sim P_\theta$ where $P_\theta = N(\mu_\theta, 1)$, $\theta \in \Theta = \{0, 1\}$ and $\mu_1 > \mu_0$. The Bayes rule versus a probability measure p on $\theta = 0$ and $1 - p$ on $\theta = 1$ is

$$(4.1) \qquad \text{decide } \theta = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if } (1-p)f_1(X) \begin{matrix} \geq \\ < \end{matrix} pf_0(X)$$

which is equivalent to

$$(4.2) \qquad \text{decide } \theta = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if } X \begin{matrix} \geq \\ < \end{matrix} c(p)$$

where

$$(4.3) \qquad c(p) = \tfrac{1}{2}(\mu_0 + \mu_1) + \frac{1}{\mu_1 - \mu_0} \log\left[\frac{p}{1-p}\right].$$

The Bayes risk R(p) is given by

$$(4.4) \qquad R(p) = (1-p)\,\Phi(c(p) - \mu_1) + p[1 - \Phi(c(p) - \mu_0)].$$

We shall call the Bayes rule versus the uniform prior on Θ the simple rule (SR). Then (4.2) and (4.3) gives the simple rule as

$$(4.5) \qquad \text{decide } \theta = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if } X \begin{matrix} \geq \\ < \end{matrix} \tfrac{1}{2}(\mu_0 + \mu_1)$$

and (4.4) become

$$(4.6) \qquad R(.5) = \Phi(\tfrac{1}{2}(\mu_0 + \mu_1)) \ .$$

Table 1 below gives the Bayes risk and the theoretical value of the PCC when classification is performed with the SR, in the special case when $\mu_0 = -\mu_1 = -\mu$.

Table 1. Bayes risk and PCC for the SR

μ	Bayes risk	PCC
.25	.4013	59.87
.50	.3085	69.15
.75	.2266	77.34
1.00	.1587	84.13
1.50	.0668	93.32
1.75	.0401	95.99
2.00	.0228	97.92

The PCC of the SR is a reasonable measure against which to compare the performance of other classification rules. In our simulation studies we have taken $\mu = 1$ so that, if a decision rule is to be of any practical use, then its PCC has to be at least 84.13. The simulation programs were written by Mr. G. Heidari under the supervision of the author. The programs are in Turbo Pascal, Version 87. Standard normal deviates were generated in pairs (z_1, z_2). Using the inverse probability transform for the Raleigh distribution and a uniform (0, 1) random variable, a radius R was generated. Using the uniform $(0, 2\pi)$, an angle ω was generated and (z_1, z_2) determined by $z_1 = R \cos \omega$, $z_2 = R \sin \omega$.

4.2 Simulations on the Transect

In our simulation studies the ℓ sequence was generated as a Markov chain with the parameters (p, δ) specified by (2.1) and (2.2). A total of of 84 different combinations of (p, δ) were considered, with

$p = 0, .01, .10, .20, .30, .40, .50$

$\delta = 0, .10, .20, .30, .40, .50, .60, .70, .80, .90, .99, 1.00$.

The observable $X_k (k = 1,2,.....,n+1)$ were generated as

$$X_k = Z_k + 2\theta_k - 1$$

where Z_k is i.i.d standard normal. For each (p, δ), 100 replications were performed with $n = 50$ and $n = 200$. The simulated data were classified by the following 5 decision rules.

1. EMAP: Empirical MAP rule (2.8)
2. ER2: Γ^2 Bayes rule
3. ER3: Γ^2 Bayes rule
4. HEB3: Empirical Bayes rule of Hill et al. (1.7) with $j = 1$.
5. HEB5: Empirical Bayes rule of HIll et al. (1.7) which $j = 2$.

Tables 2 – 11 display the mean and the standard error of the PCC for the 5 decision rules. (All tables appear at the end of this Chapter.) The ER3 and HEB3 when used for classifying θ_i are (X_{i-1}, X_i, X_{i+1}) measurable. Table 12 compares the performance of ER3 versus HEB3.

4.3 Classification of Images

Due to the complexity of the simulation of the true image as a MRF only special types of patterns and MRF were considered. The patterns considered were

1. checkerboard size 25 × 25
2. road sign size 25 × 25
3. great lakes size 31 × 44.

In the case of MRF simulations, the Markovian dependence of θ_{ij} was restricted to only one side. In particular, we assume

$$P(\theta_{ij}|\text{rest}) = P(\theta_{ij}|\theta_{i,j-1}, \theta_{i-1,j}). \tag{4.7}$$

MRF of type (4.7) with additional simplifying approximations were considered by Hansen and Elliott (1982), Haslett (1985). In our simulations the MRF was generated as (4.7) by specifying values for the 4 parameters

$$p_1 = P(\theta_{ij} = 0|\theta_{i,j-1} = 0, \theta_{i-1,j} = 0)$$
$$p_2 = P(\theta_{ij} = 0|\theta_{i,j-1} = 0, \theta_{i-1,j} = 1)$$
$$p_3 = P(\theta_{ij} = 0|\theta_{i,j-1} = 1, \theta_{i-1,j} = 0)$$
$$p_4 = P(\theta_{ij} = 0|\theta_{i,j-1} = 1, \theta_{i-1,j} = 1)$$

The MRF simulated were of size 31 by 44. In each of the images the lattice was extended to size 27 by 27 or 33 by 46 is an obvious way, in order to provide the border pixels with enough neighbors. The conventions used in obtaining the image was to color the pixel black if $\theta = 1$ and white if $\theta = 0$. As in the transect case, for each pixel (i, j) the observable X_{ij} were generated as

$$X_{ij} = Z_{ij} + 2\theta_{ij} - 1$$

where Z_{ij} is i.i.d. standard normal.

The image segmentation was performed by the following 4 decision rules

1. ER4: Extended rule (3.16) with 4 neighbors
2. SR: Simple rule (4.5)
3. MEB4: MEB rule (3.7) with 4 neighbors
4. MEB8: MEB rule (3.7) with 8 neighbors

After obtaining the image by using ER4 three revisions (1RER, 2RER, 3RER) of the extended rule (3.16) were implemented. The method of revision is as follows. At the k^{th} ($k \geq 1$) revision, the data were classified using ER4 with the empirics $G_n^5(\underline{\theta}^5)$ estimated from the $(k-1)^{st}$ revision (0^{th} revision is the classification obtained by using ER4)

For each image 25 simulations were performed. Morris et al. (1985) reports their findings based on one simulation except for one image where 10 repetitions were done. Tables 13–18 give the PCC for the decision rule used on the 6 images. For each image we have reconstructed (Figures 3–14) the best and the worst (based on the PCC of ER4) classification. Figure 15 shows a reconstruction of the great lakes pattern along with the PCC and the ACC at each stage.

4.4 Conclusion

The simulations performed with regard to the transect case show that overall the Γ^3 Bayes rule performs better than the existing empirical Bayes methods proposed by Hill et al. (1984).

The simulations performed with various images, both deterministic and stochastic, show that again the extended rules perform better than the empirical Bayes rules suggested by Morris et al. (1985), Morris (1986).

In a total of 150 simulations with various images, MEB4 had a higher PCC than ER4 only once. Furthermore in 142 simulations the ER4 had a higher PCC than even MEB8. All of the revised extended rule classifications had a higher PCC than both MEB4 and MEB8. For some images, great lakes pattern and MRF with $(p_1, p_2, p_3, p_4) = (.80, .40, .50, .60)$, the simple rule performed better than both the MEB4 and MEB8 rules in all 50 simulations, whereas for ER4 only 3 simulations had a lower PCC than the simple rule. Even in those 3 instances the revisions took care of that. The revisions of the extended rules had a small improvement on the PCC and no more than 3 revisions were needed to produce stability.

The work of Geman and Geman (1984) falls within the empirical Bayes structure in that the hyperparameters are estimated from the marginal distribution of the data. It will be of interest to compare the performance of the extended rules with the method of simulated annealing proposed by Geman and Geman (1984).

Table 2. The PCC for the EMAP rule with n = 50

	p= .00	.01	.10	.20	.30	.40	.50
δ=.00	92.20	90.42	78.66	79.88	78.78	77.68	75.42
	1.27	1.58	1.50	.98	.65	1.15	2.26
.10	92.70	91.32	79.44	78.38	79.40	79.18	71.74
	1.58	1.35	1.78	1.31	.62	.80	1.83
.20	92.46	89.74	83.20	81.52	80.42	78.72	75.72
	1.37	1.27	1.23	1.01	.77	.78	1.22
.30	89.22	88.98	81.74	80.12	80.40	78.24	77.74
	1.76	1.51	1.38	1.47	.91	.88	.99
.40	87.08	89.66	81.36	77.64	80.54	79.80	79.24
	1.61	1.35	1.33	1.50	.85	.71	.83
.50	93.30	89.98	80.54	79.08	78.88	79.30	80.24
	1.29	1.88	2.12	1.25	1.06	.83	.78
.60	92.20	91.18	83.78	77.70	73.00	76.50	80.50
	1.27	1.53	1.49	1.51	1.51	1.18	.59
.70	90.00	93.12	80.82	73.36	73.62	74.02	77.24
	1.60	1.21	1.78	1.92	1.60	1.37	1.02
.80	92.20	92.44	81.82	74.02	72.00	70.74	72.48
	1.27	1.41	1.80	1.83	1.49	1.53	1.30
.90	90.00	93.66	81.94	78.22	73.76	70.86	69.62
	1.60	1.15	2.16	1.80	1.91	1.78	1.41
.99	93.04	90.08	90.08	88.22	90.64	90.06	88.64
	1.48	1.63	1.85	1.57	1.46	1.39	1.41
1.00	93.94	93.32	91.66	94.10	95.02	95.10	94.36
	1.13	1.20	1.29	1.05	1.12	1.01	1.32

Table 3. The PCC for the EMAP rule with n = 200

	p= .00	.01	.10	.20	.30	.40	.50
δ= .00	94.36	90.84	81.82	80.57	77.69	77.37	67.01
	1.27	1.48	1.03	.42	.41	.83	2.95
.10	90.48	91.33	81.85	82.50	77.43	74.07	72.59
	1.59	1.58	1.09	.48	.50	.85	1.24
.20	90.33	90.19	85.85	83.44	80.94	74.58	75.67
	1.58	1.43	.78	.78	.48	.70	.89
.30	92.67	92.28	81.10	85.26	82.97	79.61	79.00
	1.29	1.37	1.47	.46	.43	.56	.53
.40	92.26	91.46	81.09	83.62	83.95	82.50	81.74
	1.75	1.31	1.73	.70	.37	.36	.45
.50	93.91	89.77	79.34	80.24	82.18	82.33	83.17
	1.09	1.54	1.79	1.34	.56	.33	.28
.60	92.30	90.99	77.20	78.17	79.91	81.52	82.64
	1.20	1.06	2.21	1.37	.76	.43	.32
.70	91.89	93.30	79.39	72.06	70.99	75.53	79.91
	1.47	1.02	1.92	1.86	1.42	.96	.40
.80	92.43	91.56	80.40	69.52	66.23	66.57	72.10
	1.33	1.54	2.14	1.70	1.69	1.13	.94
.90	91.92	92.55	80.57	66.22	61.53	61.56	65.78
	1.29	1.19	2.18	2.16	1.69	1.16	.94
.99	92.23	91.60	87.11	81.30	78.43	72.89	70.87
	1.34	1.33	1.80	1.99	1.78	2.20	1.98
1.00	92.93	94.41	94.01	93.47	92.86	96.19	94.05
	1.25	1.11	1.17	1.09	1.12	.96	1.38

Table 4. The PCC for the HEB3 rule with n = 50

	p= .00	.01	.10	.20	.30	.40	.50
δ=.00	92.94	89.60	84.60	83.38	83.28	88.12	95.02
	1.70	2.11	1.28	.79	.61	.49	.42
.10	94.66	91.94	86.26	82.58	83.46	85.70	91.12
	1.44	1.79	1.33	.78	.51	.50	.45
.20	91.94	91.94	86.86	85.36	83.74	83.26	86.98
	1.86	1.67	1.28	.82	.56	.53	.51
.30	88.84	88.62	88.48	85.30	82.80	82.80	84.20
	2.21	2.05	1.07	.80	.63	.49	.51
.40	91.72	89.52	88.68	84.98	84.38	82.22	82.50
	1.80	2.02	1.08	.86	.49	.57	.54
.50	91.62	93.58	88.52	88.00	85.32	82.42	81.70
	1.96	1.57	1.28	.68	.55	.54	.56
.60	92.94	92.38	89.24	88.12	86.12	83.32	81.56
	1.70	1.83	1.49	.74	.64	.50	.61
.70	89.14	88.54	90.12	88.88	86.82	85.54	84.26
	2.25	2.34	1.46	.70	.86	.58	.61
.80	92.94	92.64	91.02	91.52	88.84	88.08	86.34
	1.70	1.83	1.44	.89	.67	.59	.57
.90	89.14	88.14	91.30	90.14	91.68	90.74	91.12
	2.25	2.48	1.67	1.66	.93	.58	.46
.99	92.48	93.86	92.74	91.86	92.00	93.52	92.78
	1.84	1.42	1.81	1.81	1.79	1.46	1.33
1.00	91.06	92.42	92.34	93.76	94.18	93.18	90.74
	1.92	1.85	1.80	1.65	1.61	1.60	2.08

Table 5. The PCC for the HEB3 rule with n = 200

	p= .00	.01	.10	.20	.30	.40	.50
δ= .00	94.48	94.04	88.81	86.45	85.53	88.17	95.85
	1.65	1.58	.77	.35	.28	.22	.19
.10	94.31	93.19	90.75	87.58	84.63	86.27	91.32
	1.39	1.58	.37	.30	.30	.26	.21
.20	93.39	90.64	90.90	87.64	84.73	84.27	87.23
	1.71	2.12	.37	.25	.30	.30	.23
.30	92.64	94.48	91.75	87.90	85.57	82.77	84.74
	1.91	1.60	.34	.24	.25	.28	.26
.40	94.09	94.84	92.59	88.32	85.08	83.82	84.00
	1.66	1.24	.27	.26	.29	.26	.26
.50	96.45	96.38	93.10	89.12	85.62	83.64	83.30
	1.25	1.15	.33	.24	.25	.24	.25
.60	94.80	91.07	94.21	90.39	87.53	84.89	83.90
	1.36	2.05	.20	.22	.24	.22	.29
.70	94.14	96.00	94.89	91.75	88.89	86.31	84.81
	1.50	1.18	.22	.25	.24	.28	.24
.80	93.24	97.00	94.46	92.74	90.96	88.72	87.82
	1.89	.89	.97	.23	.20	.27	.25
.90	93.63	96.38	95.93	94.75	93.68	92.25	91.17
	1.65	1.25	.52	.21	.21	.20	.22
.99	90.45	96.23	95.12	95.56	94.13	96.21	95.32
	2.25	1.07	1.35	1.24	1.24	.38	.96
1.00	92.08	96.96	94.69	95.01	92.87	96.90	96.43
	2.04	.99	1.40	1.39	1.72	1.04	1.07

Table 6. The PCC for the HEB5 rule with n = 50

	p= .00	.01	.10	.20	.30	.40	.50
δ=.00	90.90	84.44	78.66	79.84	82.48	87.82	97.46
	1.89	2.47	1.74	1.13	.65	.49	.31
.10	89.80	86.64	82.04	78.46	81.28	84.56	91.44
	2.06	2.35	1.57	.90	.55	.49	.43
.20	89.66	85.56	83.14	82.78	81.60	81.48	86.40
	1.89	2.26	1.55	.92	.62	.54	.55
.30	82.58	85.76	84.36	81.96	81.22	80.72	83.28
	2.65	2.35	1.42	.81	.65	.49	.54
.40	86.38	87.24	84.62	83.84	82.86	80.62	81.22
	2.33	2.12	1.32	.90	.53	.58	.57
.50	88.92	92.80	86.78	86.02	83.58	80.08	80.92
	2.16	1.39	1.40	.81	.62	.54	.58
.60	90.90	88.08	86.20	86.96	85.04	82.10	80.16
	1.89	2.20	1.87	.87	.74	.55	.64
.70	87.18	86.88	85.82	87.38	85.40	84.26	83.44
	2.36	2.26	2.10	.98	.84	.62	.58
.80	90.90	88.18	89.40	91.60	88.26	87.98	85.82
	1.89	2.27	1.69	.91	.79	.59	.59
.90	87.18	86.98	87.10	89.50	90.72	90.98	90.72
	2.36	2.30	2.29	1.90	1.10	.74	.64
.99	89.18	90.62	89.60	89.08	89.54	90.96	89.90
	2.12	1.97	2.27	2.09	2.01	1.86	1.89
1.00	87.52	88.76	86.84	90.04	90.16	90.60	86.42
	2.36	2.27	2.31	2.02	1.94	1.81	2.48

Table 7. The PCC for the HEB5 rule with n = 200

	p= .00	.01	.10	.20	.30	.40	.50
δ= .00	93.88	88.81	87.02	85.63	85.55	88.39	98.62
	1.65	2.12	.79	.39	.27	.25	.12
.10	90.20	90.31	87.88	86.31	84.47	86.06	91.71
	1.97	1.96	.76	.35	.30	.26	.21
.20	90.28	86.54	89.14	87.01	84.00	83.76	86.84
	2.27	2.45	.49	.27	.29	.27	.25
.30	90.67	92.93	89.95	87.11	84.79	82.03	83.98
	2.11	1.83	.70	.32	.28	.27	.25
.40	92.90	90.78	90.80	87.85	84.46	82.82	83.12
	1.95	1.70	.56	.27	.29	.30	.28
.50	91.36	93.37	92.60	88.35	85.25	83.06	82.56
	1.87	1.50	.32	.26	.27	.24	.27
.60	92.36	89.52	93.71	89.71	86.90	84.15	83.05
	1.71	1.96	.28	.23	.27	.26	.29
.70	90.40	93.41	94.95	91.55	88.55	85.64	84.02
	2.09	1.40	.21	.25	.25	.28	.28
.80	90.19	94.69	94.21	93.29	90.83	88.55	87.54
	2.07	1.41	1.03	.22	.25	.27	.26
.90	89.00	92.70	95.79	95.88	94.43	93.45	91.79
	2.19	1.75	.67	.18	.21	.19	.24
.99	87.99	91.41	94.25	95.98	92.97	95.58	95.67
	2.43	1.85	1.47	1.19	1.69	1.22	1.37
1.00	87.28	94.10	90.60	93.15	88.14	93.49	94.51
	2.58	1.64	1.96	1.76	2.10	1.78	1.42

Table 8. The PCC for the ER2 rule with n = 50

	p= .00	.01	.10	.20	.30	.40	.50
δ=.00	98.38	97.94	91.50	88.26	85.88	86.90	90.30
	.31	.27	.45	.44	.52	.53	.56
.10	98.80	98.14	92.00	87.16	85.90	85.50	87.14
	.22	.24	.40	.48	.47	.47	.59
.20	98.78	97.56	91.68	88.58	85.56	84.42	84.86
	.25	.28	.47	.36	.51	.50	.50
.30	98.46	97.82	91.64	87.84	84.90	84.14	84.14
	.22	.27	.47	.52	.49	.50	.53
.40	98.32	97.82	91.94	88.08	85.36	83.32	83.94
	.25	.26	.49	.48	.48	.50	.50
.50	98.76	98.34	92.76	89.26	86.82	83.42	83.02
	.19	.22	.49	.45	.52	.51	.54
.60	98.38	98.34	93.36	89.04	86.18	83.96	83.40
	.31	.28	.50	.47	.55	.49	.56
.70	98.78	98.36	95.00	88.74	86.76	85.56	84.80
	.20	.23	.38	.55	.61	.58	.50
.80	98.38	98.66	94.02	91.32	87.80	86.30	84.96
	.31	.25	.65	.55	.56	.57	.59
.90	98.78	98.52	96.86	93.08	91.78	88.80	87.98
	.20	.29	.38	.59	.55	.59	.49
.99	98.68	98.50	98.44	97.12	96.92	97.46	96.30
	.22	.21	.28	.40	.44	.35	.48
1.00	98.48	98.82	98.64	98.78	98.82	98.78	98.68
	.24	.20	.20	.21	.21	.21	.24

Table 9. The PCC for the ER2 rule with n = 200

	p= .00	.01	.10	.20	.30	.40	.50
δ= .00	99.47	98.71	92.52	89.01	87.42	87.59	91.62
	.07	.09	.18	.23	.30	.25	.23
.10	99.45	98.50	92.38	88.80	86.18	85.90	88.16
	.08	.11	.17	.23	.25	.23	.28
.20	99.46	98.66	92.34	88.38	85.78	84.54	85.80
	.07	.12	.17	.21	.25	.30	.29
.30	99.55	98.78	92.76	88.65	85.99	83.49	84.29
	.08	.11	.19	.21	.24	.26	.27
.40	99.56	98.32	92.81	88.62	85.63	84.41	84.41
	.06	.13	.20	.28	.30	.25	.24
.50	99.60	98.88	93.10	89.10	86.06	84.50	84.02
	.07	.11	.21	.24	.26	.23	.23
.60	99.41	98.47	93.91	89.69	87.22	85.20	84.38
	.07	.12	.21	.22	.26	.22	.26
.70	99.58	98.62	94.09	90.45	88.16	86.22	84.45
	.08	.14	.25	.25	.25	.26	.26
.80	99.59	98.93	94.76	91.00	88.99	87.10	86.58
	.06	.15	.24	.28	.26	.27	.24
.90	99.50	98.77	95.87	92.54	90.78	89.22	88.64
	.07	.17	.27	.27	.24	.23	.23
.99	99.42	99.26	98.26	96.48	95.59	94.52	93.89
	.07	.14	.27	.37	.36	.41	.35
1.00	99.54	99.56	99.51	99.46	99.43	99.53	99.45
	.07	.07	.08	.07	.08	.07	.07

Table 10. The PCC for the ER3 rule with n = 50

	p= .00	.01	.10	.20	.30	.40	.50
δ=.00	97.58	97.46	90.56	87.78	86.22	88.86	93.52
	.36	.28	.54	.48	.51	.49	.46
.10	98.36	97.42	91.60	86.40	85.64	86.62	89.80
	.28	.31	.42	.53	.49	.50	.55
.20	98.16	97.02	90.62	87.90	84.88	84.42	86.08
	.33	.33	.55	.44	.56	.50	.58
.30	97.94	97.26	91.10	87.48	84.10	83.22	83.88
	.28	.34	.50	.53	.50	.53	.52
.40	97.50	97.22	91.42	87.34	84.96	82.50	82.94
	.33	.32	.55	.52	.49	.49	.55
.50	98.34	97.86	92.66	89.00	86.12	83.10	82.04
	.23	.25	.48	.49	.51	.50	.54
.60	97.58	97.88	93.32	88.80	86.96	84.20	82.46
	.36	.29	.51	.54	.56	.53	.63
.70	98.04	97.92	94.70	89.36	87.52	85.82	84.74
	.29	.26	.42	.53	.57	.59	.58
.80	97.58	98.16	94.16	92.02	88.74	87.12	86.50
	.36	.26	.56	.53	.56	.58	.57
.90	98.04	98.02	96.62	94.00	92.28	90.20	90.22
	.29	.30	.38	.50	.56	.60	.49
.99	98.16	97.98	98.00	97.24	96.62	97.62	96.56
	.28	.29	.32	.34	.44	.32	.41
1.00	97.78	98.12	98.20	98.30	98.28	98.22	98.22
	.29	.31	.25	.24	.25	.25	.27

Table 11. The PCC for the ER3 rule with n = 200

	p= .00	.01	.10	.20	.30	.40	.50
δ= .00	99.16	98.35	92.10	89.19	88.71	90.14	95.05
	.10	.12	.20	.23	.26	.24	.20
.10	99.23	98.16	92.13	88.72	86.45	87.33	90.99
	.11	.15	.19	.23	.23	.23	.25
.20	99.08	98.42	92.09	88.08	85.48	84.96	87.52
	.11	.14	.17	.22	.26	.28	.24
.30	99.26	98.55	92.45	88.20	85.82	83.30	85.00
	.10	.12	.21	.22	.24	.25	.26
.40	99.26	98.21	92.87	88.72	85.29	84.06	84.52
	.09	.14	.20	.27	.30	.27	.24
.50	99.28	98.72	93.52	89.15	86.19	84.14	83.65
	.11	.12	.18	.25	.26	.23	.22
.60	99.10	98.25	94.23	90.43	87.99	85.58	84.11
	.10	.13	.21	.22	.26	.22	.26
.70	99.27	98.45	94.73	91.71	89.19	87.04	85.30
	.09	.12	.23	.27	.25	.27	.26
.80	99.12	98.74	95.68	92.49	90.85	89.02	88.52
	.10	.15	.20	.24	.23	.31	.26
.90	99.11	98.66	96.60	94.46	93.10	91.80	91.22
	.07	.17	.21	.22	.20	.21	.24
.99	99.10	98.98	98.40	97.23	96.95	96.30	95.99
	.10	.13	.19	.26	.27	.30	.24
1.00	99.26	99.21	99.16	99.12	99.14	99.16	99.23
	.09	.09	.11	.10	.12	.10	.10

Table 12. The performance of the ER2 rule versus the HEB3 rule

A. n=50

	p=.00	.01	.10	.20	.30	.40	.50
δ=.00	4.64	7.86	5.96	4.40	2.94	0.74	–1.50
.10	3.70	5.48	5.34	3.82	2.18	0.92	–1.32
.20	6.22	5.08	3.76	2.54	1.14	1.16	–0.90
.30	9.10	5.64	2.62	2.18	1.30	0.42	–0.32
.40	5.78	7.70	2.74	2.36	0.58	0.28	0.44
.50	6.72	4.28	4.14	1.00	0.80	0.68	0.34
.60	4.64	5.50	4.08	0.68	0.84	0.88	0.90
.70	8.90	9.38	4.58	0.48	0.70	0.28	0.48
.80	4.64	5.52	3.14	0.50	–0.10	–0.96	0.16
.90	8.90	9.88	5.32	3.86	0.60	–0.54	–0.90
.99	5.68	4.12	5.26	5.38	4.62	4.10	3.78
1.00	6.72	5.70	5.86	4.54	4.10	5.04	7.48

76/84 (90%) ER3 better: mean = 3.74 std. dev. = 2.62
8/84 (10%) HEB3 better: mean = .82 std. dev. = .48

B. n=200

	p=.00	.01	.10	.20	.30	.40	.50
δ=.00	4.68	4.31	3.29	2.74	3.18	1.97	–0.80
.10	4.92	4.97	1.38	1.14	1.82	1.06	–0.33
.20	5.69	7.78	1.19	0.44	0.75	0.69	0.29
.30	6.62	4.07	0.70	0.30	0.25	0.53	0.26
.40	5.17	3.37	0.28	0.40	0.21	0.24	0.52
.50	2.83	2.34	0.42	0.03	0.57	0.50	0.35
.60	4.30	7.18	0.02	0.04	0.46	0.69	0.51
.70	5.13	2.45	–0.16	–0.04	0.30	0.73	0.49
.80	5.88	1.74	1.22	–0.25	–0.11	0.30	0.70
.90	5.48	2.28	0.67	–0.29	–0.58	–0.45	0.05
.99	8.65	2.75	3.28	1.67	2.82	0.09	0.67
1.00	7.18	2.25	4.47	4.11	6.27	2.26	2.80

75/84(89%) ER3 better: mean = 2.30 std. dev. = 2.25
9/84(11%) HEB3 better: mean = 0.33 std. dev. = 0.24

Table 13. The PCC with the GREAT LAKES pattern

	ER4	1RER	2RER	3RER	SR	MEB4	MEB8
	93.842	94.868	94.795	94.795	83.871	71.408	73.680
	94.795	94.575	94.721	95.015	83.798	70.894	73.837
	95.015	95.528	95.528	95.455	84.824	69.941	73.900
	93.842	95.015	94.868	94.868	83.871	71.408	73.680
	95.528	96.554	96.481	96.481	84.897	71.188	74.194
	95.015	95.381	95.455	95.381	83.871	70.528	72.947
	95.308	95.601	95.601	95.601	83.284	72.067	73.607
	94.501	95.528	95.528	95.528	84.091	70.894	73.314
	95.968	96.554	96.408	96.188	85.557	71.774	73.754
	94.868	95.968	95.894	95.894	85.191	70.455	72.874
	94.648	94.648	94.648	94.648	85.264	71.188	72.874
	94.868	95.968	95.821	95.821	85.557	70.894	73.021
	95.381	95.528	95.455	95.455	85.337	72.727	74.633
	95.455	95.748	95.601	95.601	84.971	71.848	73.387
	93.915	95.894	96.114	96.041	85.484	71.334	74.707
	94.941	95.528	95.528	95.601	84.091	70.528	73.754
	95.088	95.674	95.674	95.674	83.798	71.041	73.240
	94.575	94.795	94.941	94.941	83.138	69.795	72.581
	92.815	94.721	94.941	94.941	84.531	70.455	72.727
	94.575	95.601	95.601	95.601	83.358	70.528	72.287
	94.648	95.088	95.235	95.161	84.091	70.088	72.141
	95.601	95.821	95.748	95.748	84.824	70.015	73.974
	95.308	95.821	95.601	95.601	84.604	71.334	72.067
	95.088	95.674	95.674	95.601	82.845	69.868	72.067
	94.721	94.868	94.795	94.868	84.091	71.848	74.633
mean	94.812	95.478	95.466	95.460	84.370	70.962	73.337
s.d.	.674	.542	.499	.462	.800	.756	.784

Table 14. The PCC with the CHECKERBOARD pattern

	ER4	1RER	2RER	3RER	SR	MEB4	MEB8
	90.560	91.360	91.360	91.360	82.720	88.000	89.280
	91.360	92.480	92.640	92.640	83.520	89.120	90.240
	93.600	93.440	93.440	93.440	85.760	92.320	93.440
	89.440	89.280	89.600	89.600	83.200	87.680	88.000
	90.720	90.080	90.240	90.240	84.160	89.600	89.600
	92.480	92.160	92.160	92.160	84.000	90.400	90.880
	93.760	94.080	93.760	93.600	85.440	90.880	92.320
	90.720	91.360	90.880	90.880	83.840	89.280	90.880
	93.120	93.120	93.280	93.280	84.480	90.720	92.160
	91.520	92.320	92.320	92.320	83.040	88.000	89.440
	89.600	89.760	89.920	89.920	82.560	87.840	88.880
	88.960	89.760	89.440	89.600	84.480	88.960	89.600
	91.360	91.520	91.520	91.520	84.800	89.120	89.920
	92.960	91.840	92.000	92.320	84.160	91.040	91.840
	92.000	92.800	91.840	92.000	83.520	89.440	90.240
	92.480	93.760	93.920	93.760	84.800	91.840	92.800
	91.360	92.320	92.640	92.480	85.440	89.760	90.400
	88.480	90.400	90.400	90.720	83.840	87.520	88.800
	92.160	92.000	91.840	92.000	85.920	90.080	90.560
	93.280	93.440	93.440	93.600	86.240	90.080	91.680
	92.320	93.280	93.120	93.280	84.960	89.760	91.200
	92.960	93.920	93.920	93.920	87.040	91.840	94.240
	92.000	93.600	93.920	93.760	84.800	90.720	91.840
	92.000	92.640	92.800	92.800	85.280	90.040	91.520
	93.440	93.920	93.920	93.920	85.280	90.880	92.000
mean	91.706	92.186	92.173	92.205	84.531	89.731	90.867
s.d.	1.467	1.441	1.443	1.399	1.128	1.412	1.533

Table 15. The PCC with the ROAD SIGN pattern

	ER4	1RER	2RER	3RER	SR	MEB4	MEB8
	94.880	95.680	95.520	95.680	83.360	86.400	91.360
	95.680	97.280	97.280	97.600	87.680	93.120	95.200
	94.400	95.200	95.360	95.360	85.760	90.240	93.440
	90.560	92.800	93.600	93.440	83.200	85.920	90.880
	94.560	95.680	96.160	96.160	84.000	85.600	90.560
	95.200	95.520	95.360	95.520	83.040	87.360	90.560
	92.320	96.800	96.960	96.640	86.560	91.200	94.080
	95.040	95.360	95.200	95.200	85.120	87.840	92.320
	94.240	94.880	94.720	94.720	83.520	86.880	92.800
	93.280	93.600	94.080	94.080	84.640	88.160	92.320
	93.440	94.240	95.040	95.040	82.720	84.800	90.720
	94.240	96.320	95.840	96.000	83.520	87.040	89.280
	93.440	95.680	95.680	95.520	86.720	90.720	93.440
	93.600	94.400	94.560	94.880	82.400	86.880	90.720
	92.960	94.400	94.560	94.560	83.840	88.480	92.160
	94.560	94.560	95.360	95.680	83.840	86.240	91.040
	93.280	95.200	95.520	95.520	86.400	89.440	93.760
	95.200	96.800	96.640	96.640	86.880	90.240	92.320
	95.840	96.800	96.640	96.640	85.280	89.280	92.640
	95.840	96.160	96.640	96.320	85.920	88.480	91.360
	95.520	96.000	96.000	96.000	84.640	87.200	92.000
	94.400	94.880	95.200	95.200	85.280	88.000	92.000
	94.880	95.040	95.040	95.200	84.000	86.400	91.360
	95.360	96.960	96.640	96.640	88.000	90.400	93.760
	95.200	96.000	95.840	95.840	86.560	87.520	92.000
mean	94.317	95.450	95.578	95.603	84.915	88.154	92.083
s.d.	1.239	1.099	.912	.908	1.607	2.006	1.359

Table 16. The PCC with the (.90,.50,.50,.10) MRF pattern

	ER4	1RER	2RER	3RER	SR	MEB4	MEB8
	89.223	89.883	89.809	89.736	85.191	87.537	88.270
	88.416	88.563	88.563	88.416	84.164	86.804	87.170
	89.003	89.589	89.003	88.930	83.431	86.510	86.950
	88.196	88.710	88.710	88.710	84.897	86.364	86.657
	87.023	88.050	87.977	88.123	84.018	86.510	87.023
	88.196	88.490	88.636	88.563	83.871	86.657	87.023
	87.683	87.830	87.683	87.683	83.358	86.070	85.997
	87.903	89.003	88.416	88.636	83.871	87.648	87.454
	87.683	87.830	88.050	88.123	83.651	85.777	85.777
	88.783	89.589	89.809	89.809	84.677	86.657	87.170
	87.390	87.757	87.683	87.683	84.897	86.804	87.097
	89.296	89.589	89.516	89.443	84.384	87.170	87.610
	88.856	89.296	89.150	89.223	84.677	86.510	88.050
	88.710	88.123	88.270	88.343	83.871	86.144	86.437
	87.537	87.977	87.977	87.977	83.211	84.971	86.364
	90.396	90.469	90.396	90.396	85.924	88.270	88.416
	88.196	88.563	88.343	88.343	83.871	85.411	85.997
	87.463	87.830	87.830	87.903	82.258	85.264	85.264
	88.490	88.783	88.563	88.636	84.897	87.243	87.317
	88.563	88.783	88.710	88.490	83.504	86.217	86.217
	87.463	88.270	88.343	88.416	83.504	86.437	86.730
	88.930	89.883	90.029	90.029	84.018	87.023	87.463
	88.343	88.636	88.856	88.930	84.384	87.243	87.097
	87.977	88.563	88.710	88.563	85.777	88.636	87.830
	90.249	90.396	90.469	90.323	85.337	87.390	88.343
mean	88.399	88.818	88.780	88.777	84.226	86.691	87.029
s.d.	.842	.819	.814	.784	.854	.867	.822

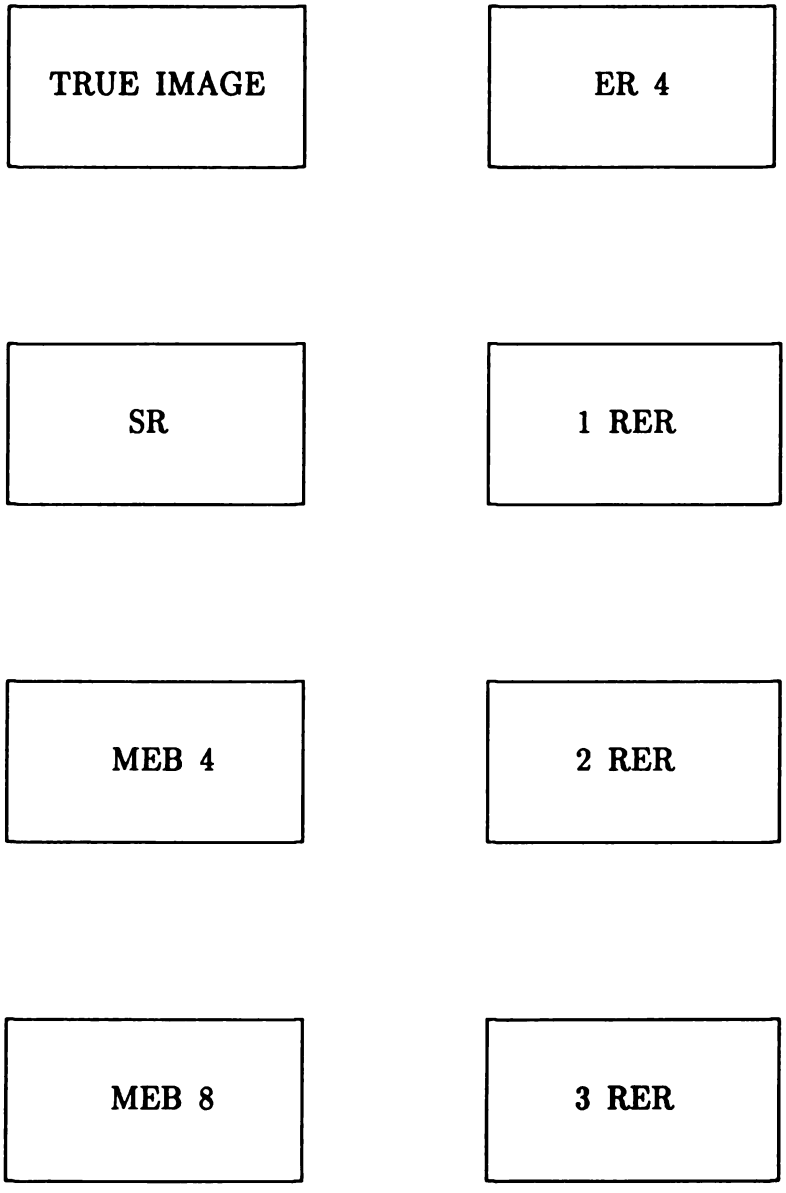
Table 17. The PCC with the (.01,.10,.01,.90) MRF pattern

	ER4	1RER	2RER	3RER	SR	MEB4	MEB8
	91.569	91.496	91.569	91.642	82.771	88.636	90.616
	91.569	93.182	92.669	92.962	83.871	89.223	90.909
	92.229	93.035	93.182	93.182	82.258	87.830	91.129
	93.035	93.182	93.035	93.109	82.918	88.415	90.396
	92.595	93.548	93.622	93.622	83.871	89.516	91.496
	92.889	93.548	93.622	93.402	84.091	89.736	91.129
	92.815	92.522	92.522	92.595	83.211	88.856	91.422
	93.768	93.475	93.548	93.255	84.897	90.543	92.522
	92.962	93.182	93.475	93.402	84.457	89.296	91.349
	92.595	92.375	92.302	92.155	84.238	88.710	90.909
	92.595	92.962	92.962	92.962	84.824	90.103	91.056
	92.082	92.522	92.669	92.595	82.331	88.416	90.689
	92.815	94.135	93.915	94.062	85.191	92.155	92.815
	93.475	93.475	93.109	93.255	83.724	89.736	92.155
	91.642	92.155	92.009	92.009	83.138	88.490	91.129
	93.109	93.109	93.328	93.182	83.798	88.930	91.569
	93.109	92.595	92.742	92.669	82.478	88.343	90.689
	93.475	94.282	94.355	94.501	84.897	90.396	92.302
	93.328	93.255	93.402	93.548	84.164	89.956	91.202
	93.182	92.889	92.522	92.522	83.798	89.003	91.276
	93.035	93.548	93.402	93.402	82.258	88.710	91.422
	93.842	94.282	94.355	94.208	85.997	89.956	91.349
	92.082	92.449	92.742	92.669	82.918	87.463	90.396
	92.522	92.742	92.669	92.742	84.091	89.223	90.689
	92.669	93.328	93.255	93.109	82.771	88.270	89.956
mean	92.760	93.091	93.079	93.070	83.718	89.196	91.223
s.d.	.633	.658	.666	.661	1.002	.997	.677

Table 18. The PCC with the (.80,.40,.50,.60) MRF pattern

	ER4	1RER	2RER	3RER	SR	MEB4	MEB8
	85.924	85.411	85.117	84.897	84.311	83.724	83.944
	82.771	83.358	83.578	83.798	81.965	81.158	81.085
	86.217	86.364	86.217	86.144	84.897	83.724	83.358
	84.091	84.238	84.018	83.944	84.018	82.625	82.771
	85.191	85.704	85.630	85.630	85.044	83.284	83.184
	84.824	85.264	84.824	85.044	83.431	82.185	82.331
	84.457	84.677	84.604	84.531	85.557	83.358	83.358
	86.730	86.730	86.730	86.730	85.484	84.604	84.897
	85.117	84.971	84.897	84.824	84.164	82.625	82.771
	84.531	86.144	86.437	86.657	85.191	83.578	83.504
	86.290	85.924	85.924	85.777	84.897	84.384	84.238
	85.411	84.751	85.117	84.824	84.018	83.798	84.018
	85.557	86.144	85.924	86.070	85.557	84.238	84.238
	84.311	85.044	84.677	85.117	84.677	83.358	83.211
	83.358	83.284	82.478	82.405	82.551	81.745	81.672
	85.191	86.437	86.290	86.364	84.604	83.211	83.358
	84.531	84.384	84.677	84.751	83.724	82.478	82.405
	85.630	85.850	85.850	85.850	84.897	83.651	83.724
	85.264	85.630	85.850	85.704	83.724	83.065	83.138
	85.337	84.824	84.824	84.677	83.944	81.745	81.745
	83.724	83.651	83.431	83.504	82.918	82.038	81.891
	84.971	85.411	85.191	85.337	84.384	83.431	83.211
	86.510	86.510	86.510	86.584	85.557	84.164	84.238
	85.484	85.997	85.777	85.850	84.311	83.504	83.138
	84.164	84.384	84.164	84.311	83.724	82.991	83.211
mean	85.023	85.243	85.149	85.173	84.302	83.147	83.144
s.d.	.972	.983	1.063	1.067	.942	.885	.915

In Figures 3 – 15 the images are arranged as shown below.



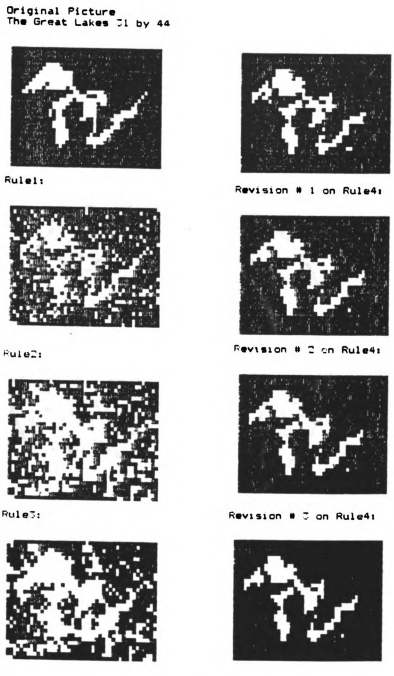


Figure 3. GREAT LAKES: The best classification

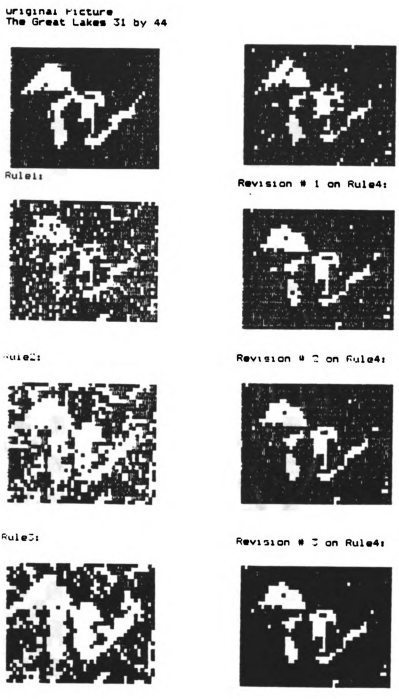


Figure 4. GREAT LAKES: The worst classification

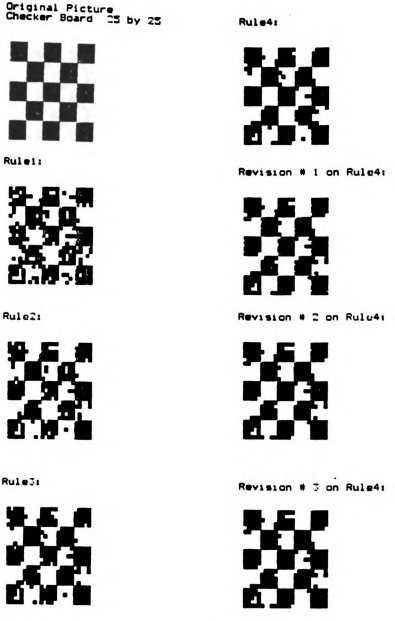


Figure 5. CHECKERBOARD: The best classification

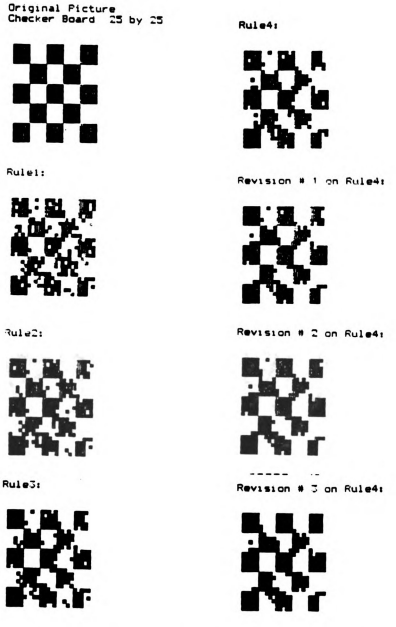


Figure 6. CHECKERBOARD: The worst classification

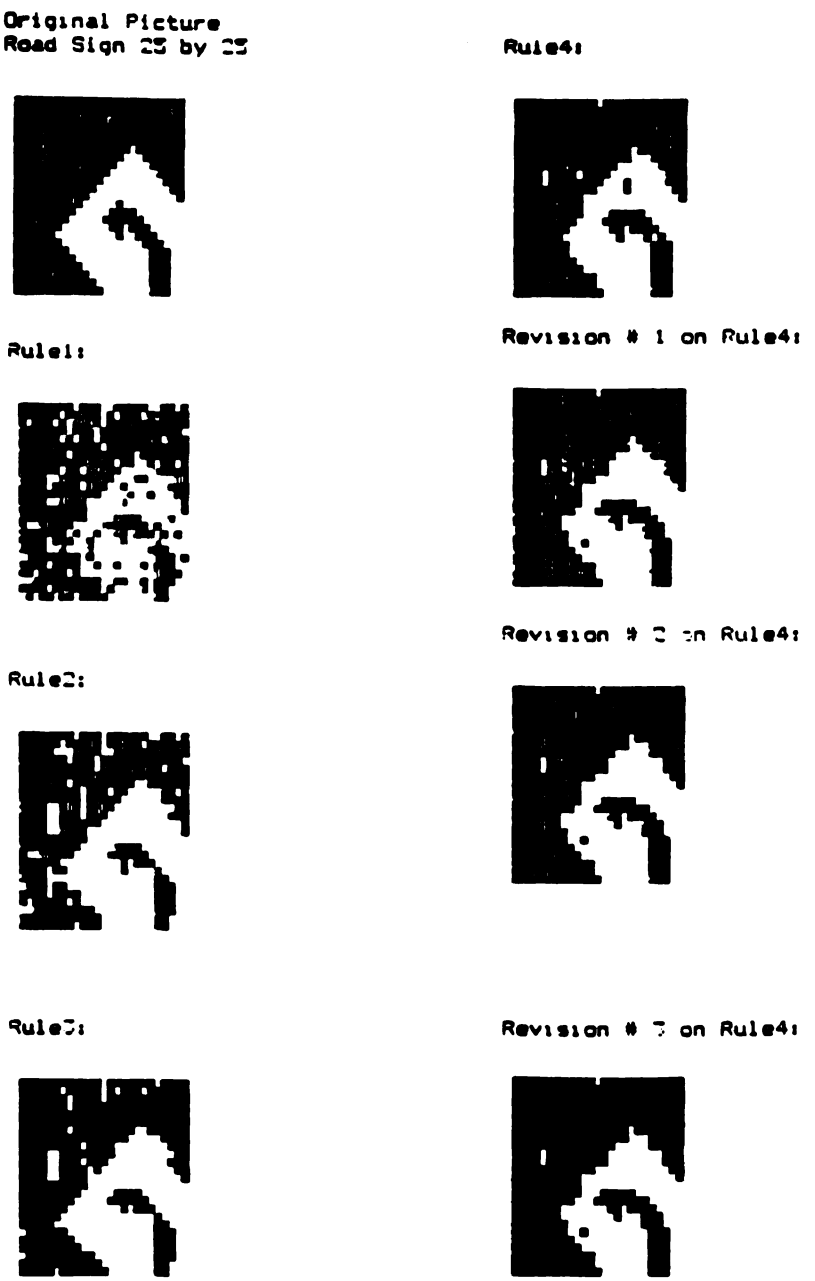


Figure 7. ROAD SIGN: The best classification

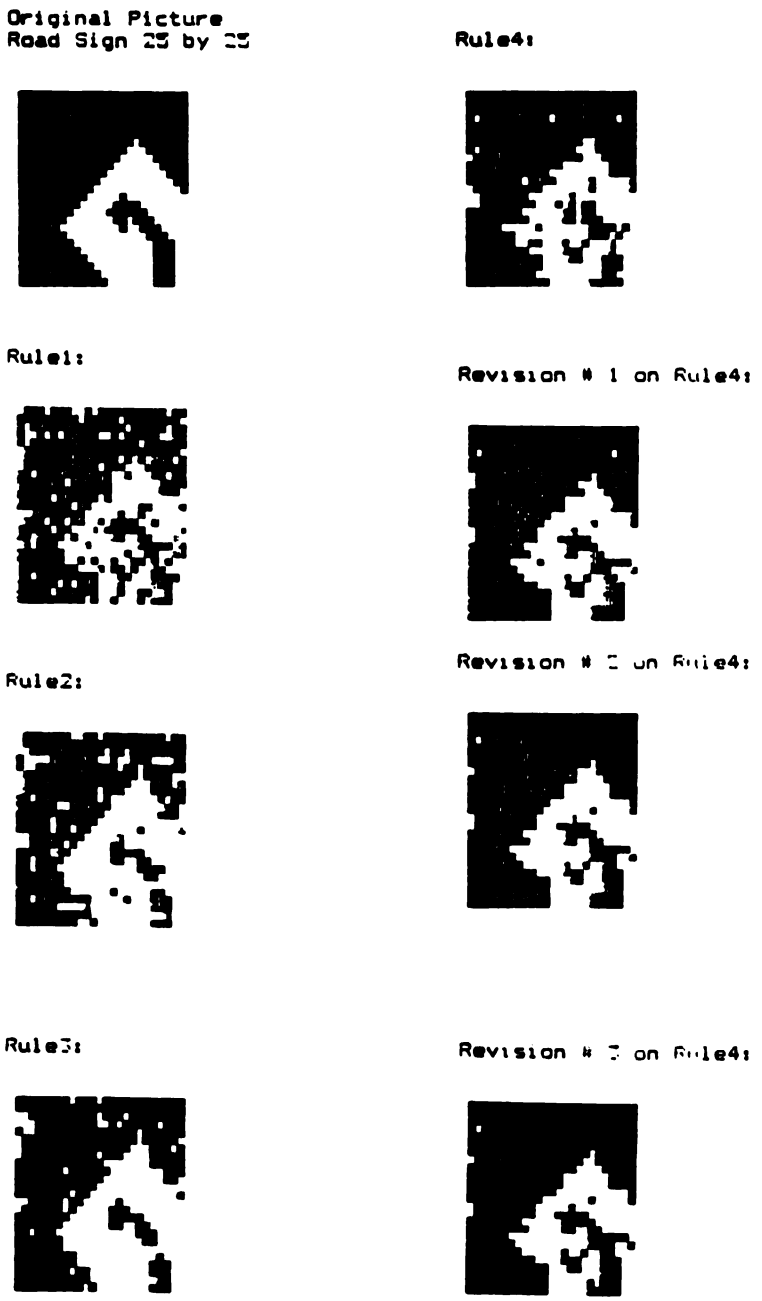


Figure 8. ROAD SIGN: The worst classification

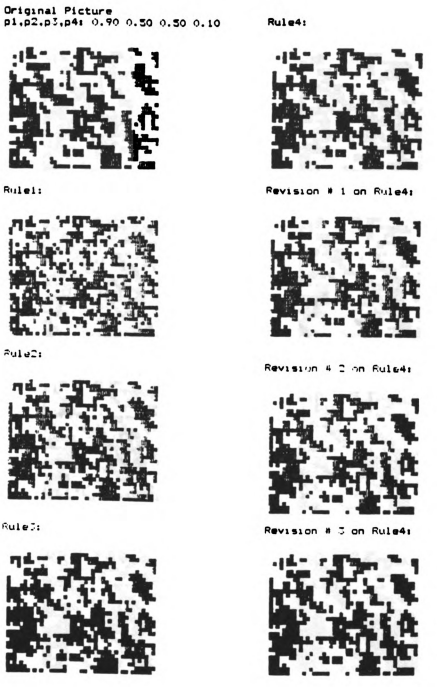


Figure 9. MRF (.90,.50,.50,.10): The best classification

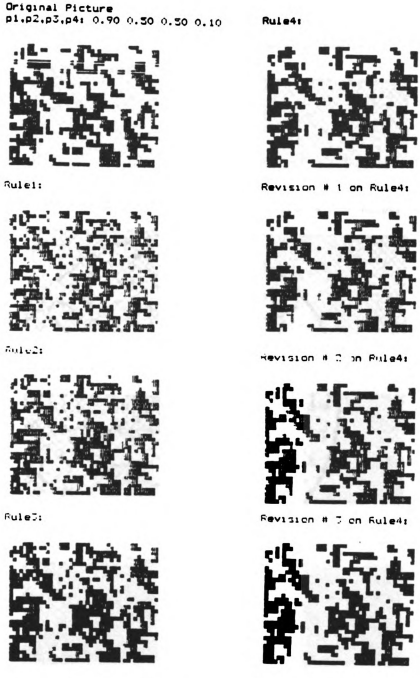


Figure 10. MRF (.90,.50,.50,.10): The worst classification

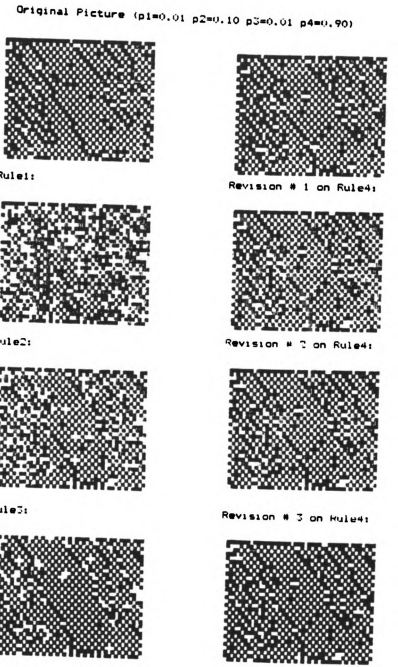


Figure 11. MRF (.01,.10,.01,.90): The best classification

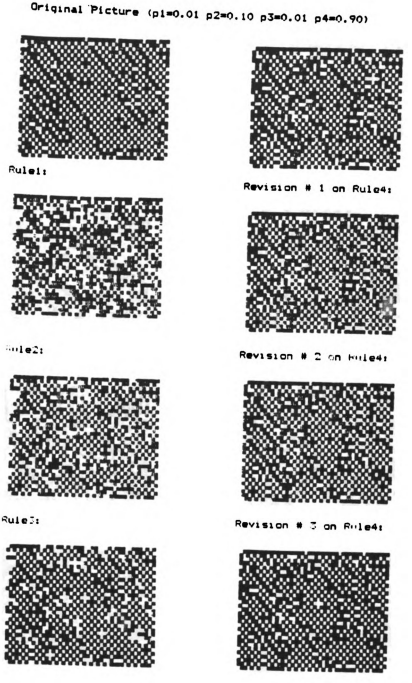


Figure 12. MRF (.01,.10,.01,.90): The worst classification

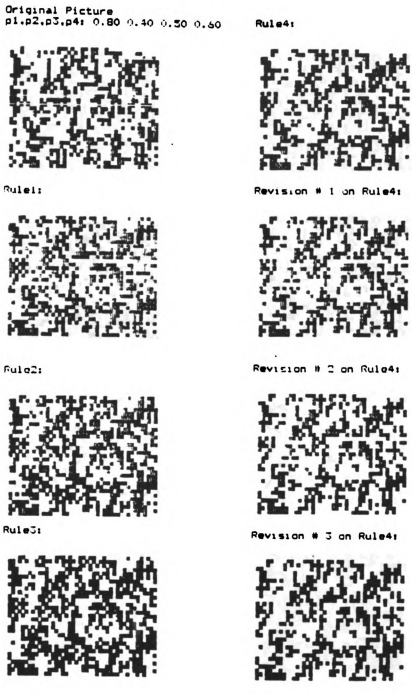


Figure 13. MRF (.80,.40,.50,.60): The best classification

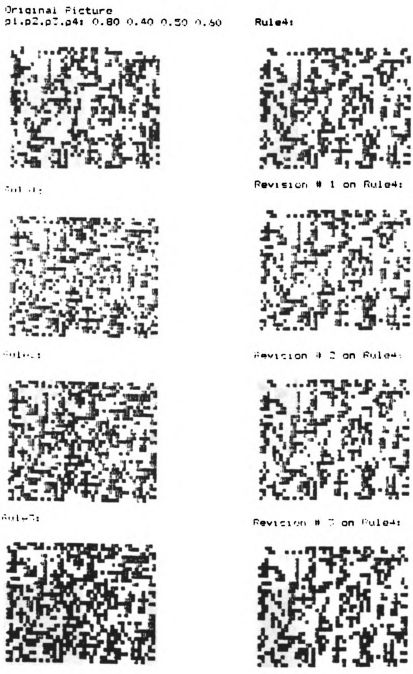


Figure 14. MRF (.80,.40,.50,.60): The worst classification

Original Picture:

Revision # 1 on the Result of Rule4 :

PCC =96.04%
PCC0 =83.41%
PCC1 =98.35%
ACA =90.80%

Revision # 2 on the Result of Rule4 :

PCC =93.50%
PCC0 =84.76%
PCC1 =83.75%
ACA =87.85%

PCC =96.19%
PCC0 =84.87%
PCC1 =98.27%
ACA =91.55%

Result of Rule4:

PCC =95.16%
PCC0 =80.57%
PCC1 =97.83%
ACA =89.20%

Revision # 3 on the Result of Rule4 :

PCC =96.11%
PCC0 =85.31%
PCC1 =98.09%
ACA =91.70%

Figure 15. GREAT LAKES CLASSIFICATION BY SR and ER4

APPENDIX

APPENDIX

DYNAMIC PROGRAMMING RECURSION

Suppose

$$A_n(\underline{\theta}^n) = \sum_{i=1}^{n} a_i\theta_i + \sum_{i=2}^{n} b_i\theta_{i-1}\theta_i \tag{A.1}$$

where $\underline{\theta}^n = (\theta_1,\theta_2,\ldots,\theta_n)$, $\theta_i \in \{0, 1\}$.

Let $$A_1^0 = 0, \quad A_i^1 = a_1 .$$

For $1 \leq i \leq n-1$, let

$$A_{i+1}^0 = \max\,(A_i^0, A_i^1)$$

$$A_{i+1}^1 = \max\,(A_i^0 + a_{i+1}, A_i^1 + a_{i+1} + b_{i+1}).$$

Then,

$$\max_{\underline{\theta}^n} A_n(\underline{\theta}^n) = \max\,(A_n^0, A_n^1).$$

Proof:

The underlying similarity of all dynamic programming processes is the creation of a set of functional equations of a particular type, called recurrence relations.

(A.1) can be written as

$$A_n(\underline{\theta}^n) = A_{n-1}(\underline{\theta}^{n-1}) + a_n\theta_n + b_n\theta_{n-1}\theta_n \tag{A.2}$$

From (A.2)

$$\max_{\underline{\theta}^n} A_n(\underline{\theta}^n) = \max\left[\max_{\underline{\theta}^{n-1}} A_{n-1}(\underline{\theta}^{n-1}),\ \max_{\underline{\theta}^{n-1}} A_{n-1}(\underline{\theta}^{n-1}) + a_n + b_n\theta_{n-1}\right]$$

$$= \max\,(A_n^0, A_n^1)$$

where

$$A_n^0 = \max_{\underline{\theta}^{n-1}} A_{n-1}(\underline{\theta}^{n-1})$$

$$A_n^1 = \max_{\underline{\theta}^{n-1}} A_{n-1}(\underline{\theta}^{n-1}) + a_n + b_n\theta_{n-1}.$$

Using (A.2) again, we get

$$A_n^0 = \max \left[\max_{\underline{\theta}^{n-2}} A_{n-2}(\underline{\theta}^{n-2}),\ \max_{\underline{\theta}^{n-2}} A_{n-2}(\underline{\theta}^{n-2}) + a_{n-1} + b_{n-1}\theta_{n-2} \right]$$

$$= \max\ (A_{n-1}^0,\ A_{n-1}^1)$$

and

$$A_n^1 = \max \left[\max_{\underline{\theta}^{n-2}} A_{n-2}(\underline{\theta}^{n-2}) + a_n,\ \max_{\underline{\theta}^{n-2}} A_{n-2}(\underline{\theta}^{n-2}) + a_{n-1} + b_{n-1}\theta_{n-2} + a_n + b_n \right]$$

$$= \max\ (A_{n-1}^0 + a_n,\ A_{n-1}^1 + a_n + b_n).$$

Proceeding as above, we obtain for $1 \le i \le n-1$

$$A_{i+1}^0 = \max(A_i^0,\ A_i^1)$$

$$A_{i+1}^1 = \max\ (A_i^0 + a_{i+1},\ A_i^1 + a_{i+1} + b_{i+1}).$$

Letting $A_1^0 = 0$ and $A_i^1 = a_1$,

we get

$$\max_{\underline{\theta}^n} A_n(\underline{\theta}^n) = \max\ (A_n^0,\ A_n^1). \qquad \square$$

Note that,

$$A_i^0 = \max_{\{\underline{\theta}^i\ ;\ \theta_i = 0\}} A_i(\underline{\theta}^i) \quad \text{and} \quad A_i^1 = \max_{\{\underline{\theta}^i\ ;\ \theta_i = 1\}} A_i(\underline{\theta}^i).$$

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Ballard, R. J. (1974). Extended rules for the sequence compound decision problem with m×n component. Ph.D. dissertation, Department of Statistics and Probability, Michigan State University, East Lansing.

2. Ballard, R. J. and Gilliland, D. (1978). On the risk performance of extended sequence compound rules for classification between N(−1, 1) and N(1, 1). *J. Stat. Comput. Simul.*, **6**, 265–280.

3. Ballard, R.J., Gilliland, D. and Hannan, J. (1975). $O(N^{-1/2})$ convergence to k–extended Bayes risk in the sequence compound decision problem with m×n components. RM–333, Department of Statistics and Probability, Michigan State University, East Lansing.

4. Bertsekas, D. P. (1976). *Dynamic Programming and Stochastic Control*, Academic Press, New York.

5. Besag, J. E. (1974). Spatial interaction and statistical analysis of lattice systems. *J. Royal Statist. Soc.*, Series B, **36**, 192–236.

6. Besag, J. E. (1986). On the statistical analysis of dirty pictures. *J. Royal Statist. Soc.*,Series B, **48**, 259–302.

7. Cooper, L. and Cooper, M. V. (1981). *Introduction to Dynamic Programming*, Pergamon Press, New York.

8. Cox, D. R. (1960). Serial sampling acceptance schemes derived from Bayes' theorem. *Technometrics*, **2**, 353–360.

9. Derin, H., Elliott, H., Christi, R. and Geman, D. (1984). Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields. *IEEE Trans. Pattern Anal. Machine Intell.*, *PAMI* **6**, 707–720.

10. Devore, J. L. (1975). Linear near neighbor classifiers for correlated categories. *Commun. Statist.*, **4**, 891–905.

11. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, *PAMI* **6**, 721–741.

12. Gilliland, D. and Hannan, J. (1969). On an extended compound decision problem. *Ann. Math. Statist.*, **40**, 1536–1541.

13. Gilliland, D., Hannan, J. and Huang, J.S. (1976). Asymptotic solutions to the two state component compound decision problem, Bayes versus diffuse priors on proportions. *Ann. Statist.*, **4**, 1101–1112.

14. Hannan, J. (1957). Approximations to Bayes risk in repeated play. *Contributions to the Theory of Games*, **3**, 97–139, *Ann. Math. Studies*, No. 39, Princeton University Press.

15. Hannan, J. and Robbins, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distribution. *Ann. Math. Statist.*, **26**, 37–51.

16. Hannan, J. and Van Ryzin, J. (1965). Rate of convergence in the compound decision problem for two completely specified distributions. *Ann. Math. Statist.*, **36**, 1743–1752.

17. Hansen, F. R. and Elliott, H. (1982). Image segmentation using simple Markov random field models. *Computer Graphics and Image Processing*, **20**, 101–132.

18. Haslett, J. (1985). Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context. *Pattern Recognition*, **18**, 287–296.

19. Hill, J. R., Hinkley, D. V., Kostal, H. and Morris, C. N. (1984). Spatial estimation from remotely sensed data via empirical Bayes models. *Proc. of the NASA Symp. on Math. Pattern Recognition and Image Analysis*, 115–136.

20. Kinderman, R. and Snell, J. L. (1980). *Markov random fields and their applications*, American Mathematical Society, Providence.

21. Morris, C. N. (1986). Binary image classification. *Proc of the* 18^{th} *Symp. on the Interface*, 71–74.

22. Morris, C. N., Hinkley, D. V. and Johnston, W. (1985). Classification in a spatially correlated environment. *Proc. of the Third Annual Symp. on Math. Pattern Recognition and Image Analysis*; also, CSS Technical Report No. 20, Center for Statistical Sciences, The University of Texas, Austin.

23. Pickard, D. K. (1977). A curious binary lattice process. *J. Appl. Prob.*, **14**, 717–731.

24. Preston, P. F. (1971). An empirical Bayes problem with a Markovian parameter. *Biometrika*, **58**, 535–543.

25. Ripley, B. D. (1986). Statistics, images and pattern recognition. *Cand. J. Statist.*, **14**, 83–111.

26. Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math Stat. Prob.*,University of California Press, 131–148.

27. Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, **35**, 1–20.

28. Scharf, L. L. and Elliott, H. (1981). Aspects of dynamic programming in signal and image processing. *IEEE Trans. Automat. Contr.*, *AC* **26**, 1018–1029.

29. Swain, P. H., Vardeman, S. B. and Tilton, J. C. (1981). Contextual classification of multispectral image data. *Pattern Recognition*, **13**, 429–441.

30. Tilton, J. C., Vardeman, S. B. and Swain, P.H. (1982). Estimation of context for statistical classification of multispectral image data. *IEEE Trans. on Geosciences and Remote Sensing*, *GE* **20**, 445–452.

31. Van Ryzin, J. (1966). The compound decision problem with m×n finite loss matrix. *Ann. Math. Statist.*, **37**, 412–424.

32. Vardeman, S. (1975). k–extended procedures in the finite state sequence compound decision problem. RM–337, Department of Statistics and Probability, Michigan State University, East Lansing.

33. Vardeman, S. (1980). Admissible solutions of k–extended finite set and sequence compound decision problems. *Journ. Multivariate Anal.*, **10**, 426–441.

34. Vardeman, S. (1981). On the small N performance of bootstrap and Bayes extended and unextended set compound rules for classification between N(–1,1) and N(1,1). *J. Statist. Comput. Simul.*, **13**, 255–271.