

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ESTIMATING PARAMETERS IN NEIGHBOURHOOD
BASED CLASSIFIERS FOR REMOTELY SENSED DATA,
USING UNCLASSIFIED VECTORS

by

Nils Lid Hjort

Norwegian Computing Center

and

Stanford University

TECHNICAL REPORT NO. 12

OCTOBER 1985

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT MCS 81-09584

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA



**ESTIMATING PARAMETERS IN NEIGHBOURHOOD
BASED CLASSIFIERS FOR REMOTELY SENSED DATA,
USING UNCLASSIFIED VECTORS**

**by
Nils Lid Hjort
Norwegian Computing Center
and
Stanford University**

**TECHNICAL REPORT NO. 12
OCTOBER 1985**

**PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT MCS 81-09584**

**DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA**

1. INTRODUCTION

A scene of picture elements (pixels) is considered. Observed in pixel i is a d -dimensional vector X_i , the components of which are spectral measurements from different channels. X_i may also have components bearing for example topographic information. K classes are defined. The X 's from class k are distributed according to a density $f_k(x)$, usually assumed to be the multi-normal. Further involved in most classification rules are the so-called a priori probabilities $\pi(1), \dots, \pi(K)$.

Hjort & Mohn (1984) review several classification procedures, both contextual and the more traditional non-contextual. (By a contextual classification rule is meant one where the assignment of class membership for a particular pixel is based on observed vectors also from neighbouring pixels.) Nearly every supervised classification procedure involves a set of statistical parameters, and estimates of these are needed to perform the classification in practice. Besides being instrumental in the classification of a scene these estimates have some interest in their own right in that they may provide further insight in the data. The parameter estimates are usually obtained from samples of 'ground truth', traditionally in the form of a training set for each class, possibly in combination with past experience.

The purpose of the present paper is to show that parameter estimates actually may be updated for the given scene, utilising also the observed vectors in unclassified pixels, i.e. whose classes are unknown. The updated estimates will be more reliable than the initial ones, in that much more information is drawn upon when constructing the former, and should also result in better classification accuracy.

What makes the updating approach possible is the fact that a vector X from a pixel randomly chosen from the set of unclassified pixels (i.e. not belonging to the sampled ground truth) has density

$$f(x) = \pi(1)f_1(x) + \dots + \pi(K)f_K(x), \quad (1.1)$$

i.e. a mixture of the K class densities with the a priori probabilities as mixture proportions. Thus it is not a contradiction in terms when Section 2 provides methods giving estimates of a priori probabilities for the given scene, based on known (or fixed estimates of) class densities.

Remark. Obviously there may be many pixels on the scene that are outliers, i.e. do not belong to any of the K predefined classes. It is assumed in (1.1) that the pixel in question really belongs to one of the classes. A separate procedure committed to the detection of all outliers is presented in Hjort (1985a). In the rest of the present paper we assume that the outliers have been taken care of, so that (1.1) applies to the remaining pixels.

Note that the a priori probabilities are of particular importance in remote sensing analysis, not only because they enter into discriminant rules, but because they are intimately tied to the areas of the K classes in the scene under study. Ideally, the $w(k)$ that should be inserted in classification algorithms is the one obtained as the frequency of class k in an imaginary very large scene of which the given scene (or rather the rest of the given scene, after removing training and outlier pixels) is a random part; $k=1, \dots, K$.

An assumption often made in remote sensing analysis is that the class density $f_k(x)$ is well described as a multinormal distribution, i.e.

$$f_k(x) = N_d(\mu_k, \Gamma_k)(x), \quad k=1, \dots, K, \quad (1.2)$$

where μ_k is the mean vector and Γ_k is the covariance matrix. If a training set for each class is available, then estimates $\hat{\mu}_k, \hat{\Gamma}_k$ may be obtained in the usual way, see (3.13), (3.14). Section 3 proposes a method which results in simultaneously updated parameter estimates

$$\pi_k^*, \mu_k^*, \Gamma_k^*; \quad k=1, \dots, K.$$

This procedure may be followed in any application of the 'best quadratic rule' of discriminant analysis where a (perhaps large) number of objects are to be classified based on experience from training sets. The updating of the covariance matrix estimate may be of particular importance when applied to the classification of remotely sensed data, for the following reason: the training set is often chosen by the user as a reasonably homogeneous class of pixels lying safely within the class concerned, i.e. the training set may be too homogeneous to be quite representative. The result is a too optimistic estimate $\hat{\Gamma}_k$, say which in its turn may introduce bias in classification. $\hat{\Gamma}_k^*$ intends to repair such possible defects in $\hat{\Gamma}_k$.

Also considered in Section 3 is an updating procedure for the case where the K covariance matrices are assumed equal, i.e.

$$f_k(x) = N_d(\mu_k, \Gamma)(x), \quad k=1, \dots, K. \quad (1.3)$$

The potential of the 'updating approach' is further illustrated in Section 4 where it is applied to the estimation of transition probabilities

$$\pi(m|k) = \Pr\{C_j = m | C_i = k\} \quad (1.4)$$

involved in a more complicated statistical model. In (1.4) C_i and C_j are used to denote the true classes of pixels i and j , which furthermore are assumed to be immediate neighbours. The transition probabilities are of separate interest since they reveal characteristic information about the structure of the landscape under study. They are also involved in several contextual classification procedures, cf. Haslett (1983), Hjort & Mohn (1984), Hjort, Mohn & Storvik (1985). The traditional way to obtain estimates of $\pi(m|k)$, $m, k=1, \dots, K$, will need random sampling of ground truth of pairs (at least) of pixels; the traditional training sets, one for each class, will not suffice, meaning trouble for the type of user that defines 'ground truth' by looking at a screen. The estimation procedures of Section 4 do not require more than the traditional training sets.

2. ESTIMATING A PRIORI PROBABILITIES ON THE GIVEN SCENE

Assume that the class densities $f_1, \dots, f_K$ are known, or that reliable estimates have been obtained, for example on the basis of training sets and/or past experience. $f_k(x)$ could typically be $N_d(\hat{\mu}_k, \hat{\Sigma}_k)(x)$ with $\hat{\mu}_k, \hat{\Sigma}_k$ obtained from training sets as in (3.13), (3.14). We wish to point out, however, that the f_k 's may be completely arbitrary as far as the present section is concerned, they might for example have been obtained via nonparametric density estimation.

The vectors X_i from pixels outside the training sets follow the mixture density

$$f(x) = \sum_{k=1}^K \pi(k) f_k(x) \quad (2.1)$$

(but see the remark following (1.1)). Vectors from pixels close to each other may be positively correlated, cf. Hjort & Mohn (1984). Consider M pixels lying far enough apart to make independence between their vectors a reasonable assumption; they may for example constitute a lattice. Denote the vectors $X_1, \dots, X_M$. By construction, these have simultaneous likelihood

$$\prod_{i=1}^M f(X_i) = \prod_{i=1}^M \left\{ \sum_{k=1}^K \pi(k) f_k(X_i) \right\}. \quad (2.2)$$

Now several procedures are possible for the estimation of $\pi(1), \dots, \pi(K)$ on the basis of $X_1, \dots, X_M$. We review first the maximum likelihood method, which has been applied to the present problem before in statistical literature, and then present a new method.

2.1 Maximum likelihood estimators

The maximum likelihood principle urges us to find the parameter set $\hat{\pi}(1), \dots, \hat{\pi}(K), \sum_{k=1}^K \hat{\pi}(k) = 1$, that maximises (2.2). Consider its logarithm

$$H(\pi(1), \dots, \pi(K-1)) = \sum_{i=1}^M \log \left\{ \sum_{k=1}^K \pi(k) f_k(X_i) \right\}, \quad (2.3)$$

where $\pi(K) = 1 - \sum_{k=1}^{K-1} \pi(k)$. Then

$$\begin{aligned} \frac{\partial H}{\partial \pi(k)} &= \sum_{i=1}^M [f_k(x_i) - f_K(x_i)] / f(x_i) \\ &= \frac{1}{\pi(k)} \sum_{i=1}^M P(k|X_i) - \frac{1}{\pi(K)} \sum_{i=1}^M P(K|X_i), \end{aligned}$$

where

$$P(k|x) = \pi(k) f_k(x) / \sum_{m=1}^K \pi(m) f_m(x) \quad (2.4)$$

is the a posteriori probability of class k given that vector $X = x$ is observed in a pixel. It follows that a maximum point for H has to fulfill the equations

$$\pi(k) = \frac{1}{M} \sum_{i=1}^M P(k|X_i), \quad k = 1, \dots, K. \quad (2.5)$$

These are indeed natural equations since $\sum_{i=1}^M P(k|X_i)$ is the expected number of pixels from class k among the M , given the observations

$X_1, \dots, X_M$.

The log-likelihood function H is fortunately concave, so that it has a unique maximum point, i.e. the equations (2.5) have only one solution. H is concave since it is easily shown that the matrix with elements

$$-\frac{\partial^2 H}{\partial \pi(k) \partial \pi(m)} = \sum_{i=1}^M [f_k(X_i) - f_K(X_i)] [f_m(X_i) - f_K(X_i)] / f(X_i)^2$$

is positive definite.

The $\pi(k)$'s enter into the r.h.s. of (2.5), and the equations cannot be solved explicitly or recursively. The solutions $\hat{\pi}(1), \dots, \hat{\pi}(K)$ have to be arrived at by an iterative computational procedure. The Newton-Raphson technique may be applied, but the easiest and completely satisfactory

method is to start out with initial values, say $\pi^{(0)}(k) = \frac{1}{K}$, $k = 1, \dots, K$, and then use (2.5) as iteration equations, i.e.

$$\pi^{(t+1)}(k) = \frac{1}{M} \sum_{i=1}^M P^{(t)}(k|X_i), \quad k = 1, \dots, K, \quad (2.6)$$

$t = 1, 2, \dots$, where $P^{(t)}(k|x) = \pi^{(t)}(k)f_k(x) / \sum_{m=1}^K \pi^{(t)}(m)f_m(x)$. This may be seen to be an application of the so-called EM algorithm, cf. Dempster, Laird & Rubin (1977). The general properties of this algorithm (cf. Wu, 1983) ensure that the sequence $\pi^{(t)}(k)$, $t = 1, 2, \dots$ will converge to the maximum likelihood estimator $\hat{\pi}(k)$ regardless of the chosen initial values; $k = 1, \dots, K$.

The estimators $\hat{\pi}(k)$ above were obtained on the basis of observed vectors $X_1, \dots, X_M$ from pixels chosen far enough from each other to make independence between their vectors a reasonable assumption. The set of pixels involved may now be replaced by a new set, producing new estimates for the $\pi(k)$'s, etc. Eventually an average should be made to define the final estimates.

To be specific, assume that J sets of distant pixels are to be used in the estimation process above, set no. j having $M(j)$ pixels; $j = 1, \dots, J$. Then set no. j and its vectors $X_{j,i}$, $i = 1, \dots, M(j)$, produce estimates $\pi_j(k)$ fulfilling equations

$$\pi_j(k) = \frac{1}{M(j)} \sum_{i=1}^{M(j)} P_j(k|X_{j,i}),$$

where

$$P_j(k|x) = \pi_j(k)f_k(x) / \sum_{m=1}^K \pi_j(m)f_m(x), \quad k = 1, \dots, K.$$

The proposed average estimates are

$$\hat{\pi}(k) = \frac{1}{J} \sum_{j=1}^J \frac{M(j)}{M} \hat{\pi}_j(k) = \frac{1}{M} \sum_{j=1}^J \sum_{i=1}^{M(j)} \hat{P}_j(k|X_{j,i}), \quad k = 1, \dots, K, \quad (2.7)$$

where $M = \sum_{j=1}^J M(j)$ is the total number of pixels involved.

When the $M(j)$'s are reasonably large the estimates obtained from the different sets of pixels will tend to be stable. A reasonable approximation then leads to

$$\pi(k) = \frac{1}{M} \sum_{i=1}^M P(k|X_i), \quad k = 1, \dots, K \quad (2.8)$$

again, where $P(k|X_i) = \pi(k)f_k(X_i) / \sum_{m=1}^K \pi(m)f_m(X_i)$, and $X_1, \dots, X_M$ are all the vectors involved in the J sets of pixels. In this way we are led to the solution we get if we assume (2.2) to be the likelihood for $X_1, \dots, X_M$ even if many pixels are close to each other. Accordingly, we do not need to be bothered with the choice of separate lattices etc., as far as the estimation of $\pi(k)$'s are concerned, if we accept the approximation that led to (2.8). However, such lattices may be needed anyway for other purposes, cf. Section 4, and it is felt that the estimates (2.7) are a bit more proper than those based on (2.8).

2.2 Bayes estimators

The Bayes estimation principle requires in the present problem the specification of an a priori distribution for $(\pi(1), \dots, \pi(K))$. Assume that this distribution is given by a density

$$g(\pi) = g(\pi(1), \dots, \pi(K-1)) \quad (2.9)$$

for $\pi = (\pi(1), \dots, \pi(K-1))$ in the simplex

$$B = \{\pi | \pi(k) > 0, \quad k = 1, \dots, K-1, \quad \sum_{k=1}^{K-1} \pi(k) \leq 1\}. \quad (2.10)$$

Then π given the data $X_1, \dots, X_M$ has a posteriori density

$$g(\pi)L_M(\pi)d\pi / \int_B g(\pi)L_M(\pi)d\pi,$$

where

$$L_M(\pi) = L_M(\pi, X_1, \dots, X_M) = \prod_{i=1}^M \left\{ \sum_{k=1}^K \pi(k) f_k(X_i) \right\} \quad (2.11)$$

is the likelihood of the observed data, given the unknown parameters.

The Bayes estimators are the a posteriori expectations

$$\tilde{\pi}(k) = E(\pi(k) | X_1, \dots, X_M) = N_M(k) / D_M(k), \quad (2.12)$$

where

$$N_M(k) = \int_B \pi(k) g(\pi) L_M(\pi) d\pi, \quad (2.13)$$

$$D_M(k) = \int_B g(\pi) L_M(\pi) d\pi, \quad (2.14)$$

$k = 1, \dots, K$.

It is difficult to obtain the Bayes estimates in practice. The likelihood $L_M(\pi)$ may be expanded in a large sum of terms of the type

$$\pi(j_1) \dots \pi(j_M) f_{j_1}(X_1) \dots f_{j_M}(X_M), \quad (2.15)$$

where $1 \leq j_i \leq K$, $i = 1, \dots, M$. Consequently, both the numerator and denominator of $\tilde{\pi}(k)$ may in theory be written as a sum of K^M terms, each of which is explicitly computable from moments

$$A(a_1, \dots, a_K) = E \pi(1)^{a_1} \dots \pi(K)^{a_K} = \int_B \prod_{k=1}^K \pi(k)^{a_k} g(\pi) d\pi \quad (2.16)$$

(writing $\pi(K) = 1 - \prod_{k=1}^{K-1} \pi(k)$) of the a priori distribution. Such a procedure might well take a computer several years unless M is small, however, so that alternative procedures have to be devised.

A possible method consists of obtaining $N_M(k)$ and $D_M(k)$ by numerical integration. This would also pose implementational difficulties since typically $L_M(\pi)$ would be a product of a large number of possibly small terms.

We will instead outline another method. When expanding $L_M(\pi)$ the terms (2.15) may be sorted and collected together in a way that gives

$$L_M(\pi) = \sum_{C(M)} \pi(1)^{a_1} \dots \pi(K)^{a_K} S_M(a_1, \dots, a_K) \quad (2.17)$$

where $C(M)$ is the set of $(a_1, \dots, a_K)$ having $a_k \geq 0$, $k = 1, \dots, K$, and $\sum_{k=1}^K a_k = M$. $S_M(a_1, \dots, a_K)$ is the sum of all terms $f_{j_1}(x_1) \dots f_{j_M}(x_M)$ that have exactly a_k f_k -factors, $k = 1, \dots, K$. $S_M(a_1, \dots, a_K)$ has $M! / (\prod_{k=1}^K a_k!)$ terms, whereas the number of terms in the expansion (2.17) may be seen to be

$$\#C(M) = \sum_{a_1=0}^M \sum_{a_2=0}^{M-a_1} \dots \sum_{a_{K-1}=0}^{M-a_1-\dots-a_{K-2}} 1 = \frac{1}{(K-1)!} (M+1) \dots (M+K), \quad (2.18)$$

utilising the identity

$$\sum_{j=1}^n j(j+1) \dots (j+b-1) = \frac{1}{b+1} n(n+1) \dots (n+b),$$

which can be proved by induction.

By (2.13), (2.14), (2.16) and (2.17)

$$N_M(k) = \sum_{C(M)} A(a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_K) S_M(a_1, \dots, a_K),$$

$$D_M(k) = \sum_{C(M)} A(a_1, \dots, a_k, \dots, a_K) S_M(a_1, \dots, a_K),$$

giving us a method to obtain $\tilde{\pi}(k)$, provided the moments $A(a_1, \dots, a_K)$ are easy to compute (as they are w.r.t. a Dirichlet a priori density, for example), provided $\# C(M)$ in (2.18) is a manageable number (notice that it is dramatically less than K^M in the typical remote sensing application), and provided the terms $S_M(a_1, \dots, a_K)$ can be easily computed.

It can now be seen that

$$S_M(a_1, \dots, a_K) = \sum_{k=1}^K f_k(X_M) S_{M-1}(a_1, \dots, a_{k-1}, \dots, a_K), \quad (2.19)$$

which holds even if some a_k 's are zero, if $S_M(a_1, \dots, a_K)$ is defined as zero when some a_k equals -1 . This makes it possible to obtain the S_M 's in a recursive way. At intermediate step m the full set of $(m+1) \dots (m+K)/(K-1)!$ S_m 's will need to be stored.

Suppose for example that the uniform prior $g_0(\pi) \equiv (K-1)!$, $\pi \in B$ of (2.10), is chosen, which is one way of formalising 'total ignorance a priori' about $\pi(1), \dots, \pi(K)$. Then

$$E \prod_{k=1}^K \pi(k)^{a_k} = \left(\prod_{k=1}^K a_k! \right) / K \dots (K+M-1),$$

from known properties of Dirichlet distributions. Hence in this case the Bayes estimates can be computed as

$$\tilde{\pi}(k) = \frac{1}{K+M} \frac{\int_{C(M)} a_1! \dots a_K! (a_k+1) S_M(a_1, \dots, a_k, \dots, a_K)}{\int_{C(M)} a_1! \dots a_K! S_M(a_1, \dots, a_k, \dots, a_K)}. \quad (2.20)$$

Discussion. The Bayes estimates (2.12) are generally harder to compute than the maximum likelihood estimates (2.5) - (2.8). They may have some statistical advantages for moderate sample sizes, however, as a priori knowledge often exists in a form that could be approximated with an appropriate Dirichlet density. The computations needed in for example (2.20) increase considerably with the number of classes, whereas the maximum likelihood procedure is essentially unbothered.

One may prove that Bayes estimators and maximum likelihood estimators are asymptotically equivalent, i.e.

$$\sqrt{M} (\hat{\pi}(k) - \tilde{\pi}(k)) \xrightarrow{P} 0$$

as $M \rightarrow \infty$; $k = 1, \dots, K$, regardless of a priori distribution (2.9). This follows from a multidimensional extension of Theorem 6.7.2 in Lehmann (1983) in the case where the assumed model (2.1) is absolutely correct, and may be shown to be true even in the more realistic case where the model is considered to be an approximation only. It may also be proved that both $\hat{\pi}(k)$ and $\tilde{\pi}(k)$ are asymptotically optimal.

Thus we may recommend the maximum likelihood procedure on the grounds of computational simplicity and statistical objectivity. The Bayes estimators were included in the paper partly because of general theoretical interest. Bayes estimators may constitute good alternatives to maximum likelihood estimators in the more general mixture distribution problem where the class densities have unknown parameters, in which case there are severe problems with the latter approach.

On the whole, however, Subsection 2.2 may be considered a digression from the main theme of the present paper.

3. ESTIMATING A PRIORI PROBABILITIES AND UPDATING CLASS DESCRIPTIONS SIMULTANEOUSLY

Section 2 was concerned with the model (2.1) in the ideal case where the class densities $f_1, \dots, f_K$ were known. Assume now that there are unknown parameters in them, say

$$f_k(x) = f(x, \theta_k), \quad (3.1)$$

where θ_k is a p -dimensional parameter characterising class k ; $k = 1, \dots, K$. Hence vectors from pixels whose class labels are unknown follow the mixture distribution

$$f(x) = \sum_{k=1}^K \pi(k) f(x, \theta_k). \quad (3.2)$$

Let as in Section 2 $X_1, \dots, X_M$ be data from a set of distant pixels, so that their simultaneous likelihood may be taken to be

$$L_{\text{rest}} = \prod_{i=1}^M \left(\sum_{k=1}^K \pi(k) f(x_i, \theta_k) \right). \quad (3.3)$$

It is now possible to construct estimators for $\pi(k)$, θ_k , $k=1, \dots, K$ simultaneously, based on $X_1, \dots, X_M$ only. Indeed, the maximum likelihood approach may be invoked, and results in natural likelihood equations for the estimates. These equations may have several solutions, however, and another difficulty is that the likelihood (3.3) may have unpleasant singularities at edges of the parameter space. Nevertheless the described approach will be successful if consistent estimates $\hat{\theta}_{k,0}$ from the training sets are available and if the needed iterative computational procedure have $\pi(k) = \frac{1}{K}$, $\theta_k = \hat{\theta}_{k,0}$, $k = 1, \dots, K$ as initial values.

Better yet is to utilise the training data in combination with the 'new' vectors $X_1, \dots, X_M$. Assume that a training set of the form

$$Z = \{X_j^{(k)}; j = 1, \dots, n_k, k = 1, \dots, K\} \quad (3.4)$$

is available, where $x_1^{(k)}, \dots, x_{n_k}^{(k)}$ are from pixels in class k . Resorting to the trick of choosing only a subset of distant pixels for each class, if necessary, we may assume that $(Z, X_1, \dots, X_M)$ has total likelihood

$$L_{\text{full}} = L_{\text{training}} L_{\text{rest}}$$

$$= \prod_{k=1}^K \prod_{j=1}^{n_k} f(x_j^{(k)}, \theta_k) \prod_{i=1}^M \left(\sum_{k=1}^K \pi(k) f_k(X_i) \right). \quad (3.5)$$

The likelihood equations become

$$\frac{\partial \log L_{\text{full}}}{\partial \pi(k)} = \frac{1}{\pi(k)} \sum_{i=1}^M P(k|X_i) - \frac{1}{\pi(K)} \sum_{i=1}^M P(K|X_i), \quad k = 1, \dots, K-1,$$

$$\frac{\partial \log L_{\text{full}}}{\partial \theta_{k,l}} = \sum_{j=1}^{n_k} \frac{\partial \log f(x_j^{(k)}, \theta_k)}{\partial \theta_{k,l}} + \sum_{i=1}^M P(k|X_i) \frac{\partial \log f(X_i, \theta_k)}{\partial \theta_{k,l}},$$

$l = 1, \dots, p, \quad k = 1, \dots, K$, where

$$P(k|x) = \frac{\pi(k) f(x, \theta_k)}{\sum_{m=1}^K \pi(m) f(x, \theta_m)}. \quad (3.6)$$

It follows that any local maximum $(\pi(1)^*, \dots, \pi(K)^*, \theta_1^*, \dots, \theta_K^*)$ of L_{full} must satisfy

$$\pi(k) = \frac{1}{M} \sum_{i=1}^M P(k|X_i), \quad k = 1, \dots, K, \quad (3.7)$$

$$\frac{\partial \log L_{\text{full}}}{\partial \theta_{k,l}} = 0, \quad l = 1, \dots, p, \quad k = 1, \dots, K. \quad (3.8)$$

In the presence of L_{training} the full likelihood will most often have a global maximum (whereas L_{rest} alone may be unbounded). One may prove that the sequence of estimators obtained by iterating equations (3.7) - (3.8)

in an EM-manner with starting values $\pi(k) = \frac{1}{K}$, $\theta_k = \hat{\theta}_{k,0}$, $k = 1, \dots, K$, where $\hat{\theta}_{k,0}$ is the maximum likelihood estimator for θ_k based on $x_1^{(k)}, \dots, x_{n_k}^{(k)}$, is consistent. The asymptotic framework implicitly referred to here is one where both the n_k 's and M grow towards infinity. The practical interpretation, as far as a typical remote sensing analysis application is concerned, is that the start values above will suffice to find the correct bump on the likelihood surface, even for moderately sized training sets.

Let us apply the program above to the important case where

$$f_k(x) = N_d(\mu_k, \Gamma_k)(x) = (2\pi)^{-d/2} |\Gamma_k|^{-1/2} \exp\left\{-\frac{1}{2} (x-\mu_k)' \Gamma_k^{-1} (x-\mu_k)\right\},$$

$k = 1, \dots, K$. Then some manipulations entail

$$\frac{\partial \log f_k(x)}{\partial \mu_{k,a}} = \Lambda_{k,(a)}(x-\mu_k),$$

$$\frac{\partial \log f_k(x)}{\partial \lambda_{k,ab}} = \left(1 - \frac{1}{2} \delta_{ab}\right) \left[\sigma_{k,ab} - (x_a - \mu_{k,a})(x_b - \mu_{k,b})\right],$$

$a, b = 1, \dots, d$, where $\Lambda_k = \Gamma_k^{-1} = \{\lambda_{k,ab}\} = \{\sigma_{k,ab}\}^{-1}$ and $\Lambda_{k,(a)}$ is the a 'th row in Λ_k . Hence the estimates we seek must obey

$$\Lambda_{k,(a)} \left\{ \sum_{j=1}^{n_k} (x_j^{(k)} - \mu_k) + \sum_{i=1}^M P(k|X_i) (X_i - \mu_k) \right\} = 0,$$

$$\begin{aligned} \sum_{j=1}^{n_k} \{\sigma_{k,ab} - (x_{j,a}^{(k)} - \mu_{k,a})(x_{j,b}^{(k)} - \mu_{k,b})\} \\ + \sum_{i=1}^M P(k|X_i) \{\sigma_{k,ab} - (X_{i,a} - \mu_{k,a})(X_{i,b} - \mu_{k,b})\} = 0, \end{aligned}$$

for $k = 1, \dots, K$ and $a, b = 1, \dots, d$. The equations simplify to

$$\mu_k = \frac{\sum_{j=1}^{n_k} X_j^{(k)} + \sum_{i=1}^M P(k|X_i) X_i}{n_k + \sum_{i=1}^M P(k|X_i)}, \quad (3.10)$$

$$\Gamma_k = \frac{\sum_{j=1}^{n_k} (X_j^{(k)} - \mu_k)(X_j^{(k)} - \mu_k)' + \sum_{i=1}^M P(k|X_i) (X_i - \mu_k)(X_i - \mu_k)'}{n_k + \sum_{i=1}^M P(k|X_i)}, \quad (3.11)$$

$k = 1, \dots, K$, where

$$P(k|x) = \frac{\pi(k) N_d(\mu_k, \Gamma_k)(x)}{\sum_{m=1}^K \pi(m) N_d(\mu_m, \Gamma_m)(x)}. \quad (3.12)$$

These equations may be considered as natural generalisations of known ones for the case where data from only the mixture distribution are available, cf. Duda & Hart (1973, p.200), to the case where also training data are present.

The maximum likelihood estimators from the training sets alone are

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} X_j^{(k)}, \quad (3.13)$$

$$\hat{\Gamma}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} (X_j^{(k)} - \hat{\mu}_k)(X_j^{(k)} - \hat{\mu}_k)', \quad (3.14)$$

$k = 1, \dots, K$, see for example Anderson (1958, p.47). We are now in a position to define the estimators $\pi(k)^*$, μ_k^* , Γ_k^* , $k = 1, \dots, K$. Compute successive estimates $\pi_{(t)}(k)$, $\mu_{k,(t)}$, $\Gamma_{k,(t)}$, $k = 1, \dots, K$, using the iteration equations

$$\pi_{(t+1)}(k) = \frac{1}{M} \sum_{i=1}^M P_{(t)}(k|X_i), \quad (3.15)$$

$$\mu_{k,(t+1)} = \frac{n_k \hat{\mu}_k + \sum_{i=1}^M P_{(t)}(k|X_i) X_i}{n_k + \sum_{i=1}^M P_{(t)}(k|X_i)}, \quad (3.16)$$

$$\begin{aligned} \Gamma_{k,(t+1)} = & \frac{n_k \hat{\Gamma}_k + n_k (\mu_{k,(t)} - \hat{\mu}_k) (\mu_{k,(t)} - \hat{\mu}_k)' +}{n_k + \sum_{i=1}^M P_{(t)}(k|X_i)} + \\ & + \frac{\sum_{i=1}^M P_{(t)}(k|X_i) (X_i - \mu_{k,(t)}) (X_i - \mu_{k,(t)})'}{n_k + \sum_{i=1}^M P_{(t)}(k|X_i)} \end{aligned} \quad (3.17)$$

$t = 0, 1, 2, \dots$, with initial values

$$\pi_{(0)}(k) = \frac{1}{K}, \quad \mu_{k,(0)} = \hat{\mu}_k, \quad \Gamma_{k,(0)} = \hat{\Gamma}_k, \quad k = 1, \dots, K. \quad (3.18)$$

Here $P_{(t)}(k|X_i)$ is as in (3.12), but with t -th generation parameter values. The iteratively defined sequence of points

$$\{\pi_{(t)}(k), \mu_{k,(t)}, \Gamma_{k,(t)}; k = 1, \dots, K\}$$

in $K(1+d+d(d+1)/2)$ -dimensional space are ensured convergence as $t \rightarrow \infty$ by the general theory of EM algorithms, cf. Wu (1983). The limiting values are our maximum likelihood estimates $\pi(k)^*$, μ_k^* , Γ_k^* ; $k = 1, \dots, K$.

Note that all $K(1+d+d(d+1)/2)$ points in the t -th generation need to be computed before the $(t+1)$ -st generation points can be obtained. Observe further that fewer iterations will be needed to obtain the maximum likelihood estimates, probably, if $\pi_{(0)}(k) = \hat{\pi}(k)$ is used in (3.18), where $\hat{\pi}(k)$ is found as in Section 2.

Next consider the case where the covariance matrices are taken to be equal, i.e.

$$f_k(x) = N_d(\mu_k, \Gamma)(x), \quad k = 1, \dots, K. \quad (3.19)$$

Then the program above may be applied with a few technical changes. The essential equations for the maximum likelihood estimators in the previous case were (3.7), (3.10), (3.11). After elaborations similar to those presented there one arrives in the present case at (3.7), (3.10) as before, whereas the K equations in (3.11) are replaced by

$$\Gamma = \frac{\sum_{k=1}^K \sum_{j=1}^{n_k} (X_j^{(k)} - \mu_k)(X_j^{(k)} - \mu_k)' + \sum_{i=1}^M \sum_{k=1}^M P(k|X_i)(X_i - \mu_k)(X_i - \mu_k)'}{N + \sum_{i=1}^M \sum_{k=1}^K P(k|X_i)} \quad (3.20)$$

where $N = \sum_{k=1}^K n_k$. An iterative computational procedure may now be devised, resulting in a sequence of parameter points $(\pi_{(t)}(k), \mu_{k,(t)}, \Gamma_{(t)}), t = 0, 1, 2, \dots$, converging to estimates $\pi(k)^{**}, \mu_k^{**}, \Gamma^{**}$. Start out with

$$\pi_{(0)}(k) = \frac{1}{K}, \mu_{k,(0)} = \hat{\mu}_k, \Gamma_{(0)} = \sum_{k=1}^K \frac{n_k}{N} \hat{\Gamma}_k = \hat{\Gamma}; \quad (3.21)$$

variations exist, for example $\pi_{(0)}(k) = \hat{\pi}(k)$ obtained from $X_1, \dots, X_M$ as described in Section 2. Then compute successively

$$\pi_{(t+1)}(k) = \frac{1}{M} \sum_{i=1}^M P_{(t)}(k|X_i), \quad (3.22)$$

$$\mu_{k,(t+1)} = \frac{n_k \hat{\mu}_k + \sum_{i=1}^M P_{(t)}(k|X_i) X_i}{n_k + \sum_{i=1}^M P_{(t)}(k|X_i)}, \quad (3.23)$$

$$\Gamma_{(t+1)} = \frac{N \hat{\Gamma} + \sum_{k=1}^K n_k (\mu_{k,(t)} - \hat{\mu}_k)(\mu_{k,(t)} - \hat{\mu}_k)' + \sum_{i=1}^M \sum_{k=1}^K P_{(t)}(k|X_i)(X_i - \mu_{k,(t)})(X_i - \mu_{k,(t)})'}{N + \sum_{i=1}^M \sum_{k=1}^K P_{(t)}(k|X_i)} \quad (3.24)$$

$t = 0, 1, 2, \dots$ until a convergence criterion is met. Here $P_{(t)}(k|X_i)$ is evaluated according to t -th generation parameter values.

Remark. This section has provided methods to obtain updated class description parameter estimates for the given scene, along with estimates of prior probabilities. The resulting estimators combine information from training sets (the supervised part) with information from the rest of the scene (the unsupervised part) in an optimal way. In applications one may witness the gradual adjustment of the original class descriptions $\hat{\mu}_k, \hat{\Gamma}_k$ to the full-scene estimates μ_k^*, Γ_k^* . This will give the experimenter information on how representative his training set really was. Better yet in this respect might be to compute maximum likelihood estimates $\tilde{\mu}_k, \tilde{\Gamma}_k$ for the unsupervised part alone, which is possible by methods similar to those presented in the present section, and compare them with the maximum likelihood estimates $\hat{\mu}_k, \hat{\Gamma}_k$ for the supervised part alone.

4. ESTIMATING NEIGHBOUR TRANSITION PROBABILITIES

Consider the probabilities

$$\pi(m|k) = \Pr\{C_j = m | C_i = k\}; \quad k, m = 1, \dots, K, \quad (4.1)$$

where C_i and C_j denote the true classes of pixels i and j , and where j is one of the four immediate neighbours to i . (We assume stationarity and isotropy). These are termed neighbour transition probabilities. Their interpretation is not obvious. A possibility is to tie them to the corresponding (unobserved) relative frequencies on the given scene. Such a definition, which would have to be made more precise due to certain ambiguities related to multiple counts of a pixel and its neighbours, would cause the $\pi(m|k)$'s to change if the scene is only slightly redefined, however.

Another possibility is to relate the neighbour transition probabilities to a statistical model for the behaviour of classes in a given region of land. We will subscribe to this second viewpoint in this section, since it will enable us to devise estimation procedures for the $\pi(m|k)$'s and related quantities from general statistical principles. Thus we consider the actual distribution of classes in a scene to be a realisation of a stochastic process.

4.1 A multiplicative Markov random field model

Consider a pixel cross with centre pixel i and four neighbours, and denote their classes $C_i, C_{iN}, C_{iE}, C_{iS}, C_{iW}$. A simple statistical model considered by Haslett (1983), Hjort & Mohn (1984), and Hjort, Mohn & Storvik (1985), for the purpose of deriving contextual classification procedures, has

$$\begin{aligned} g(a,b,c,d|k) &= \Pr\{C_{iN} = a, C_{iE} = b, C_{iS} = c, C_{iW} = d | C_i = k\} \\ &= \pi(a|k)\pi(b|k)\pi(c|k)\pi(d|k) \end{aligned} \quad (4.2)$$

for each of the K^4 possible configurations (a,b,c,d) , for each given class k . Thus $\pi(k)\pi(a|k)\pi(b|k)\pi(c|k)\pi(d|k)$ is the probability of seeing class k in the centre of a given pixel cross combined with neighbour classes (a,b,c,d)

in the appropriate north, east, south, west positions.

To arrive at a statistical model for the observed vectors in the pixel cross,

$$\Delta_i = \{X_i, X_{iN}, X_{iE}, X_{iS}, X_{iW}\}, \quad (4.3)$$

it is also necessary to specify the conditional density of $X_{iN}, X_{iE}, X_{iS}, X_{iW}$ given classes k, a, b, c, d , and given X_i in the centre pixel; say

$$h(X_{iN}, X_{iE}, X_{iS}, X_{iW} | k, a, b, c, d, X_i), \quad (4.4)$$

cf. Hjort & Mohn (1984). Then the likelihood of (4.3) becomes

$$\begin{aligned} f(\Delta_i) &= \sum_{k=1}^K \sum_{a,b,c,d} \pi(k) g(a,b,c,d|k) f(\Delta_i | k, a, b, c, d) \\ &= \sum_{k=1}^K \pi(k) f_k(X_i) \sum_{a,b,c,d} g(a,b,c,d|k) h(X_{iN}, X_{iE}, X_{iS}, X_{iW} | k, a, b, c, d, X_i) \end{aligned} \quad (4.5)$$

in general terms, and the posterior probability that the class label is k for the centre pixel, given the data Δ_i from its pixel cross, becomes

$$P(k|\Delta_i) = \frac{1}{f(\Delta_i)} \pi(k) f_k(X_i) \sum_{a,b,c,d} g(a,b,c,d|k) h(X_{iN}, X_{iE}, X_{iS}, X_{iW} | k, a, b, c, d, X_i). \quad (4.6)$$

The reader is referred to Hjort & Mohn (1984) for applications of this general formula to the construction of contextual classifiers. The present aim is to devise estimation procedures for the $\pi(m|k)$'s involved in (4.2).

4.1.A. Conditional independence

A simplifying but not always adequate assumption often made in the literature is that of conditional independence between spectral vectors given the class labels, which implies that the h -factor in (4.4) reduces to

$$f_a(X_{iN}) f_b(X_{iE}) f_c(X_{iS}) f_d(X_{iW}).$$

Hence

$$f(\Delta_i) = \prod_{k=1}^K \pi(k) f_k(X_i) T_k(X_{iN}) \dots T_k(X_{iW}) \quad (4.7)$$

and

$$P(k|\Delta_i) = \frac{1}{f(\Delta_i)} \pi(k) f_k(X_i) T_k(X_{iN}) \dots T_k(X_{iW}), \quad (4.8)$$

where

$$T_k(x) = \prod_{m=1}^K \pi(m|k) f_m(x). \quad (4.9)$$

Assume now that M pixel crosses, indexed $i = 1, \dots, M$, are chosen sufficiently distant from each other to make independence between $\Delta_1, \dots, \Delta_M$ a reasonable assumption. Then $\prod_{i=1}^M f(\Delta_i)$ is the total observed likelihood, and we may try to obtain maximum likelihood estimates for the neighbour transition probabilities.

The typical remote sensing analysis application will involve separate homogeneous training sets, one for each class, cf. (3.4). Then knowledge about the class densities in the form of class description estimates may be obtained with reasonable precision, and likewise reliable estimates of the $\pi(k)$'s may be computed by the methods of previous sections. Thus it is not unreasonable to consider the $\pi(m|k)$'s as the 'principal unknowns' and take $\pi(k)$'s and f_k 's as known, when considering the likelihood $\prod_{i=1}^M f(\Delta_i)$.

We are, therefore, to maximise $\prod_{i=1}^M \log f(\Delta_i)$ w.r.t. the K^2 transition probabilities. These are constrained by

$$\prod_{m=1}^K \pi(m|k) = 1; \quad k = 1, \dots, K. \quad (4.10)$$

Upon writing $\pi(K|k) = 1 - \sum_{m=1}^{K-1} \pi(m|k)$ we get

$$\begin{aligned} \frac{\partial \log f(\Delta_i)}{\partial \pi(m|k)} &= \frac{1}{f(\Delta_i)} \pi(k) f_k(X_i) \\ &\cdot \left[\{f_m(X_{iN}) - f_K(X_{iN})\} T_k(X_{iE}) T_k(X_{iS}) T_k(X_{iW}) \right. \\ &\quad \left. + \dots + T_k(X_{iN}) T_k(X_{iE}) T_k(X_{iS}) \{f_m(X_{iW}) - f_K(X_{iW})\} \right] \\ &= P(k|\Delta_i) \left[\frac{U(m|k, X_{iN})}{\pi(m|k)} - \frac{U(K|k, X_{iN})}{\pi(K|k)} \right. \\ &\quad \left. + \dots + \frac{U(m|k, X_{iW})}{\pi(m|k)} - \frac{U(K|k, X_{iW})}{\pi(K|k)} \right], \quad (4.11) \end{aligned}$$

where

$$U(m|k, x) = \pi(m|k) f_m(x) / T_k(x); \quad m, k = 1, \dots, K.$$

Define

$$V(m|k) = \sum_{i=1}^M P(k|\Delta_i) \frac{1}{4} \left[U(m|k, X_{iN}) + \dots + U(m|k, X_{iW}) \right]. \quad (4.13)$$

The equations for the maximum likelihood estimates $\hat{\pi}(m|k)$ become

$$V(m|k)/\pi(m|k) = V(K|k)/\pi(K|k), \quad m = 1, \dots, K,$$

which in view of (4.10) can be rewritten

$$\pi(m|k) = V(m|k) / \sum_{m=1}^K V(m|k) = V(m|k) / \sum_{i=1}^M P(k|\Delta_i), \quad (4.14)$$

$m, k = 1, \dots, K.$

This is not an explicit solution in that the full set of $\pi(m|k)$'s are involved on each of the right hand sides of (4.14). It is however possible to find the solutions $\hat{\pi}(m|k)$ as the limits arising from an iterative computational procedure, having (4.14) as the defining iterative equations. Good starting values $\pi_{(0)}(m|k)$ may be essential.

4.1.B General case

Subsection 4.1.A provided an estimation procedure for the $\pi(m|k)$'s when the multiplicative model (4.2) was tied to the conditional independence hypothesis, so that (4.7) - (4.8) were in force. As this hypothesis is dubious when applied to high resolution satellites, cf. Hjort & Mohn (1984), the general case should also be considered.

Assume that a particular model for the distribution of vectors given the classes has been proposed and fitted to the data, with needed parameter estimates obtained from the supervised part of the scene. This leads to a specific

$$h(X_{iN}, X_{iE}, X_{iS}, X_{iW} | k, a, b, c, d, X_i) = h(\Delta_i | k, a, b, c, d),$$

say, in formulae (4.5), (4.6). Such an approach is taken in Hjort, Mohn & Storvik (1985).

A direct maximum likelihood approach applied to $\prod_{i=1}^M f(\Delta_i)$ gives a very cumbersome set of equations for the neighbour transition probabilities $\pi(m|k)$. A practical and fully satisfactory solution is obtained by writing

$$g(a, b, c, d | k) = \pi_N(a|k) \pi_E(b|k) \pi_S(c|k) \pi_W(d|k) \quad (4.15)$$

and treating the sets $\{\pi_N(a|k)\}, \dots, \{\pi_W(d|k)\}$ as being pro forma different, and average

$$\pi(m|k) = \frac{1}{4} \left[\pi_N(m|k) + \dots + \pi_W(m|k) \right]$$

afterwards.

Introduce

$$U_N(m|k, \Delta_i) = \frac{\sum_{b,c,d} \pi_N(m|k) \pi_E(b|k) \pi_S(c|k) \pi_W(d|k) h(\Delta_i|m, a, b, c, d)}{\sum_{a,b,c,d} \pi_N(a|k) \pi_E(b|k) \pi_S(c|k) \pi_W(d|k) h(\Delta_i|k, a, b, c, d)} \quad (4.16)$$

and analogously $U_E(m|k, \Delta_i)$, $U_S(m|k, \Delta_i)$, $U_W(m|k, \Delta_i)$; $k, m = 1, \dots, K$. Note that $U_N(m|k, \Delta_i)$ reduces to $\pi_N(m|k) f_m(X_{iN})/T_k(X_{iN}) = U(m|k, X_{iN})$ of (4.12) under conditional independence.

Then it is seen from (4.5) and (4.6) that

$$\begin{aligned} \frac{\partial \log f(\Delta_i)}{\partial \pi_N(m|k)} &= \frac{1}{f(\Delta_i)} \pi(k) f_k(X_i) \\ &\quad \left[\sum_{b,c,d} \pi_E(b|k) \pi_S(c|k) \pi_W(d|k) h(\Delta_i|k, m, b, c, d) \right. \\ &\quad \left. - \sum_{b,c,d} \pi_E(b|k) \pi_S(c|k) \pi_W(d|k) h(\Delta_i|k, K, b, c, d) \right] \\ &= P(k|\Delta_i) \left[\frac{U_N(m|k, \Delta_i)}{\pi_N(m|k)} - \frac{U_N(K|k, \Delta_i)}{\pi_N(K|k)} \right] \end{aligned}$$

for $m = 1, \dots, K-1$, writing $\pi_N(K|k) = 1 - \sum_{m=1}^{K-1} \pi_N(m|k)$.

Define

$$V_N(m|k) = \sum_{i=1}^M P(k|\Delta_i) U_N(m|k, \Delta_i) \quad (4.17)$$

and similar expressions for $V_E(m|k)$, $V_S(m|k)$, $V_W(m|k)$. The likelihood equations

$$\sum_{i=1}^M \frac{\partial \log f(\Delta_i)}{\partial \pi_N(m|k)} = 0, \quad m = 1, \dots, K-1,$$

simplify to

$$\pi_N(m|k) = V_N(m|k) / \sum_{m=1}^K V_N(m|k) = V_N(m|k) / \sum_{i=1}^M P(k|\Delta_i), \quad (4.18)$$

$m = 1, \dots, K$.

Thus (4.14) is arrived at once more, if we let

$$V(m|k) = \sum_{i=1}^M P(k|\Delta_i) \frac{1}{4} \left[U_N(m|k, \Delta_i) + \dots + U_W(m|k, \Delta_i) \right], \quad (4.19)$$

which is the proper generalisation of (4.13).

An iterative computational process may now be performed, using (4.14) as the iteration equations. Start values may be

$\pi(m|k) = (1-\epsilon) \delta_{mk} + \frac{\epsilon}{K-1} (1-\delta_{mk})$ for some ϵ in $(0,1)$, i.e. a matrix with $1-\epsilon$ on the diagonal and $\frac{\epsilon}{K-1}$ outside. The values obtained after a convergence criterion is met are our estimates $\hat{\pi}(m|k)$; $m, k = 1, \dots, K$.

This procedure may in general demand efforts in terms of implementational work and CPU time. An important example is considered by Hjort, Mohn & Storvik (1985; their 'Specification 3'), wherein the $U_N(m|k, \Delta_i)$'s and the $V(m|k)$'s simplify drastically, so that estimation of the $\pi(m|k)$'s can be carried out in that case.

Remark 4.1. One may want to estimate the a priori probabilities $\pi(k)$ along with the $\pi(m|k)$'s, based on M pixel crosses of unclassified pixels. This can indeed be done, as it follows from (4.5) and (4.6) that

$$\frac{\partial \log f(\Delta_i)}{\partial \pi(k)} = \frac{P(k|\Delta_i)}{\pi(k)} - \frac{P(K|\Delta_i)}{\pi(K)}, \quad k = 1, \dots, K-1,$$

using $\pi(K) = 1 - \sum_{k=1}^{K-1} \pi(k)$ again. Hence the iterative procedure that will lead to maximum likelihood estimates $\pi(k)^*$, $\pi(m|k)^*$ simultaneously works by using

$$\pi(k) = \frac{1}{M} \sum_{i=1}^M P(k|\Delta_i) ,$$

$$\pi(m|k) = V(m|k) / \sum_{i=1}^M P(k|\Delta_i) ,$$

$k, m = 1, \dots, K$, where $V(m|k)$ is given by (4.19).

It may also be possible to obtain estimates for parameters involved in the h-part of the model (4.5), along with $\pi(m|k)$'s and $\pi(k)$'s, by the same approach.

Remark 4.2. It should perhaps be stressed that the estimators for $\pi(m|k)$'s constructed above are obtained by the fitting of a statistical model to data, cf. in particular (4.2). Of course a statistical model indexed by a small number of parameters can only hope to approximate the complex reality with which remote sensing analysis is concerned. It follows that the solutions $\hat{\pi}(m|k)$ cannot in general be taken as estimates of the true neighbour transition probabilities (defined in terms of relative frequencies of neighbour configuration occurrences on the given scene). This explains why $\hat{\pi}(m|k)$'s obtained by the presented methods do not necessarily agree with estimates obtained by sampling the ground truth, and also why two different h-models may lead to apparently different sets of estimates for $\pi(m|k)$'s.

This remark also points out that it would be useful to construct and apply goodness of fit tests for the employed statistical models.

4.2. Estimating parameters in a simple geometric probability model

Hjort & Mohn (1984), Hjort (1985b) consider a simple geometric probability model for the behaviour of classes in a scene. The model involves unknown parameters p, q, r , being non-negative and summing to one, and in terms of the g-function defined in (4.2) postulates

$$g(k, k, k, k | k) = p + (q+r) \pi(k),$$

$$g(k, k, m, m | k) = g(m, k, k, m | k) = g(m, m, k, k | k) = g(k, m, m, k | k) = q\pi(m)/4,$$

$$g(k, k, m, k | k) = g(k, k, k, m | k) = g(m, k, k, k | k) = g(k, m, k, k | k) = r\pi(m)/4; \quad (4.20)$$

$m \neq k$. The model is a modest generalisation of one proposed by Owen & Switzer (1982), who use

$$p = 1 - \beta, \quad q = (\sqrt{2} - 1)\beta, \quad r = (2 - \sqrt{2})\beta \quad (4.21)$$

as a consequence of assuming a certain Poisson field model.

Under the hypothesis of conditional independence one can derive from (4.5) and (4.6) that

$$f(\Delta_i) = \sum_{k=1}^K \pi(k) f_k(X_i) R_k(\Delta_i), \quad (4.22)$$

$$P_i(k | \Delta_i) = \frac{1}{f(\Delta_i)} \pi(k) f_k(X_i) R_k(\Delta_i), \quad (4.23)$$

where

$$R_k(\Delta_i) = p A_k(\Delta_i) + q B_k(\Delta_i) + r C_k(\Delta_i), \quad (4.24)$$

$$A_k(\Delta_i) = f_k(X_{iN}) f_k(X_{iE}) f_k(X_{iS}) f_k(X_{iW}),$$

$$B_k(\Delta_i) = \frac{1}{4} [f(X_{iN}) f_k(X_{iE}) f_k(X_{iS}) f_k(X_{iW})$$

$$+ f_k(X_{iN}) f(X_{iE}) f_k(X_{iS}) f_k(X_{iW})$$

$$+ f_k(X_{iN}) f_k(X_{iE}) f(X_{iS}) f_k(X_{iW})$$

$$+ f_k(X_{iN}) f_k(X_{iE}) f_k(X_{iS}) f(X_{iW})],$$

$$\begin{aligned}
C_k(\Delta_i) = & \frac{1}{4} [f_k(X_{iN}) f_k(X_{iE}) \bar{f}(X_{iS}, X_{iW}) \\
& + f_k(X_{iE}) f_k(X_{iS}) \bar{f}(X_{iW}, X_{iN}) \\
& + f_k(X_{iS}) f_k(X_{iW}) \bar{f}(X_{iN}, X_{iE}) \\
& + f_k(X_{iW}) f_k(X_{iN}) \bar{f}(X_{iE}, X_{iS})],
\end{aligned}$$

and where

$$\begin{aligned}
f(x) &= \sum_{k=1}^K \pi(k) f_k(x), \\
\bar{f}(x, y) &= \sum_{k=1}^K \pi(k) f_k(x) f_k(y).
\end{aligned}$$

We may now pursue the maximum likelihood program to obtain estimates $\hat{p}, \hat{q}, \hat{r}$ based on observations $\Delta_1, \dots, \Delta_M$ from M pixel crosses in the unsupervised part of the scene. From

$$\begin{aligned}
\frac{\partial \log f(\Delta_i)}{\partial p} &= \frac{1}{f(\Delta_i)} \sum_{k=1}^K \pi(k) f_k(X_i) \left[A_k(\Delta_i) - C_k(\Delta_i) \right], \\
\frac{\partial \log f(\Delta_i)}{\partial q} &= \frac{1}{f(\Delta_i)} \sum_{k=1}^K \pi(k) f_k(X_i) \left[B_k(\Delta_i) - C_k(\Delta_i) \right],
\end{aligned}$$

which are obtained upon writing $r = 1 - p - q$, the likelihood equations become

$$\begin{aligned}
\sum_{i=1}^M \sum_{k=1}^K P(k|\Delta_i) \frac{A_k(\Delta_i)}{R_k(\Delta_i)} &= \sum_{i=1}^M \sum_{k=1}^K P(k|\Delta_i) \frac{C_k(\Delta_i)}{R_k(\Delta_i)}, \\
\sum_{i=1}^M \sum_{k=1}^K P(k|\Delta_i) \frac{B_k(\Delta_i)}{R_k(\Delta_i)} &= \sum_{i=1}^M \sum_{k=1}^K P(k|\Delta_i) \frac{C_k(\Delta_i)}{R_k(\Delta_i)}.
\end{aligned}$$

The solutions $\hat{p}, \hat{q}$ and $\hat{r} = 1 - \hat{p} - \hat{q}$ to these equations may be obtained by iterating the equations

$$\begin{aligned} p &= \frac{1}{m} \sum_{i=1}^M \sum_{k=1}^K P(k|\Delta_i) \alpha_k(\Delta_i) , \\ q &= \frac{1}{M} \sum_{i=1}^M \sum_{k=1}^K P(k|\Delta_i) \beta_k(\Delta_i) , \\ r &= \frac{1}{M} \sum_{i=1}^M \sum_{k=1}^K P(k|\Delta_i) \gamma_k(\Delta_i) , \end{aligned} \quad (4.25)$$

where

$$\begin{aligned} \alpha_k(\Delta_i) &= p A_k(\Delta_i)/R_k(\Delta_i) , \\ \beta_k(\Delta_i) &= q B_k(\Delta_i)/R_k(\Delta_i) , \\ \gamma_k(\Delta_i) &= r C_k(\Delta_i)/R_k(\Delta_i) . \end{aligned} \quad (4.26)$$

One may similarly consider the problem of finding an estimate $\hat{\beta}$ for the parameter β in Owen & Switzer's model (4.21). The equation for $\hat{\beta}$ becomes

$$\sum_{i=1}^M \sum_{k=1}^K P_k(\Delta_i) \left[-A_k(\Delta_i) + (\sqrt{2} - 1) B_k(\Delta_i) + (2 - \sqrt{2}) C_k(\Delta_i) \right] / R_k(\Delta_i) = 0. \quad (4.27)$$

Variations exist. One may take p, q, r to be class-dependent, or one might consider other sampling schemes than that of obtaining M distant pixel crosses.

It seems reasonable to sample pixel crosses when the aim is to obtain parameter estimates needed in classifiers that use crosses as context, although for example a sample of pairs of pixels, or observations along a long row of pixels, often will suffice. The methods presented in this section can easily be adjusted.

Remark 4.3. The present section has provided methods to obtain estimates of parameters involved in statistical models for the distribution of classes, based on a sample of M distant crosses of pixels. This set of pixel crosses should next be moved to obtain new estimates, and so on, and eventually an average should be made to define the final estimates, cf. Section 2.

REFERENCES

- Anderson, T.W. (1958). Introduction to multivariate analysis. Wiley.
- Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977).
Maximum likelihood from incomplete data via the EM algorithm
(with discussion). J. Roy. Statist. Ser. B, 39, 1-38.
- Duda, R.O. & Hart, P.E. (1973). Pattern Classification and Scene Analysis. Wiley.
- Haslett, J. (1983). Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context.
Dep. of statistics, Trinity College, Dublin.
- Hjort, N.L. (1985a). Detecting outliers. In Sæbø et al.
- Hjort, N.L. (1985b). Neighbourhood-based classification of remotely sensed data based on geometric probability models. In Sæbø et al.
- Hjort, N.L. & Mohn, E. (1984). A comparison of some contextual methods in remote sensing classification. In Proceedings of the Eighteenth International Symposium on Remote Sensing of Environment. CNES, Paris.
- Hjort, N.L., Mohn, E. & Storvik, G. (1985). Contextual classification of remotely sensed data based on an autocorrelation model. In Sæbø et al.
- Lehmann, E. (1983). Theory of point estimation. Wiley.
- Owen, A. & Switzer, P. (1982). A neighbourhood based classifier for LANDSAT data. Techn. Report. Dept. of Statistics, Stanford University.
- Sæbø, H.V., Bråten, K.H., Hjort, N.L., Llewellyn, B. & Mohn, E. (1985)
"Contextual classification of remotely sensed data: statistical methods and development of a system". Project Report, Norwegian Computing Center, Oslo.
- Wu, C.F.J. (1983). On the convergence properties of the EM algorithm.
Ann. Statist. 11, 95-103.