# A NEIGHBORHOOD BASED CLASSIFIER FOR LANDSAT DATA

BY

ART OWEN and PAUL SWITZER

TECHNICAL REPORT NO. 2
NOVEMBER 1, 1982

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

# A NEIGHBORHOOD BASED CLASSIFIER FOR LANDSAT DATA

BY

ART OWEN AND PAUL SWITZER
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 2
NOVEMBER 1, 1982

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

# A NEIGHBORHOOD BASED CLASSIFIER FOR LANDSAT DATA

by

Art Owen[1]
Department of Statistics
Stanford University

## ABSTRACT

A classifier is developed which uses information from all pixels in a neighborhood to classify the pixel at the center of the neighborhood. It is not a smoother in that it tries to recognize boundaries, and it makes explicit use of the relative positions of pixels in the neighborhood. It is based on a geometric probability model for the distribution of the classes in the plane. The neighborhood based classifier is shown to outperform linear discriminant analysis on some LANDSAT data.

Running Title: Neighborhood Based Classifier

Key words: Remote Sensing, Image Analysis, LANDSAT, Discriminant Analysis, Classification

[1]This research was partially supported by National Science Foundation Grant MCS 81-09584.

# A NEIGHBORHOOD BASED CLASSIFIER FOR LANDSAT DATA

by

Art Owen
Department of Statistics
Stanford University

## 1. Introduction

This paper considers an extension of discriminant analysis motivated by a problem in remote sensing. In this problem, a portion of the surface of the earth is partitioned into a grid of small squares called pixels. A satellite records the intensity of reflected electromagnetic radiation at each of several (typically four) wavelengths, for each pixel. On the basis of the observed intensities, the pixel is assigned one of a small number of categories. The categories represent some characteristic of the surface, such as the type of vegetation growing there, or the predominant rock type.

The usual discriminant analysis described on p. 574 of Rao (1973) often performs poorly because the variability in reflected intensity for each category is large compared to the separation of the mean reflected intensities for different categories. When the categories form contiguous regions consisting of large numbers of pixels, the classification accuracy can be improved by using the observations over a neighborhood of pixels to classify the pixel in the center of the neighborhood (Switzer et al., 1981). Previously considered methodologies have mostly been smoothers; some involve replacing each observation with a weighted

average of the observations in a neighborhood of that observation, others smooth the estimated categories by replacing small isolated regions of pixels with the category that is in some sense "locally dominant".

Smoothers may be expected to improve the classification accuracy in the centers of contiguous regions, but might reasonably be expected to blur the boundaries between regions of different categories. The distinctive feature of the classifier considered here is that it attempts to recognize which pixels are surrounded by pixels of the same category and which are boundary pixels.

## 2. Notation

The neighborhoods considered here consist of a pixel $t$ and its four nearest neighbors. Compass directions can be defined on the grid of pixels and they can be used to label the other pixels in the neighborhood of $t$ as shown in Figure 1.

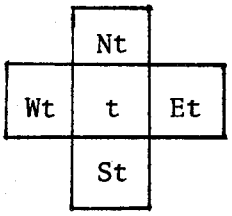


FIGURE 1

The vector of observed intensities for pixel $t$ is denoted $Z_t$ and

$$Z_t^* = (Z_t, Z_{Nt}, Z_{Et}, Z_{St}, Z_{Wt})$$

represents all the observations in the neighborhood of $t$.

There are $K$ categories: $i=1, \ldots, K$ and one category is to be assigned to each pixel. (This is an approximation in that it is likely

that some pixels would be better represented as mixtures of categories.) The category of pixel $t$ is denoted $I_t$ and

$$I_t^* = (I_t, I_{Nt}, I_{Et}, I_{St}, I_{Wt})$$

denotes the categories for the neighborhood of $t$.

The Bayes classification scheme is to choose for pixel $t$ the category $i$ that maximizes

$$\pi_i \, f(Z_t^* \mid I_t = i) = \pi_i \, \Sigma \, f(Z_t \mid I_t^*) \, P(I_t^* \mid I_t = i) \qquad (1)$$

where $f$ is the density of $Z_t^*$ given $I_t = i$, $\pi_i$ is the prior probability of category $i$, and the sum is taken over $I_t^*$ consistent with $I_t = i$.

A neighborhood based classifier is obtained by modelling the conditional density, conditional probabilities, and prior probabilities in (1). The next two sections describe such a model, and Section 5 indicates how the parameters in those models may be estimated from a training sample. (A training sample is a set of pixels for which the true category is known.)

## 3. The Conditional Density Model

The density model used was Gaussian with mean

$$E(Z_t^* \mid I_t^* = (i,j,k,\ell,m)) = (\mu_i, \mu_j, \mu_k, \mu_\ell, \mu_m)$$

where $\mu_i = E(Z_t \mid I_t = i)$ and block diagonal dispersion

$$D(Z_t^* \mid I_t^* = (i,j,k,\ell,m)) = \begin{pmatrix} S_i & & & & \\ & S_j & & & \\ & & S_k & & \\ & & & S_\ell & \\ & & & & S_m \end{pmatrix}$$

where $S_i = E([Z_t - \mu_i][Z_t - \mu_i]' \mid I_t = i)$.

With most Landsat data it would be more realistic to have a correlation structure that is positive and attenuates as the distance between pixels increases. Such a correlation could be due to cloud cover or vegetation. Modelling the correlation structure between pixels increased the computation tenfold and did not improve the accuracy of the classifier.

In the example below, the training sample was not large enough to estimate a dispersion matrix for each category, but the dispersion clearly differed from category to category by at least a scale factor. Hence the dispersion was modelled as above with the restriction that

$$S_i = \lambda_i S \qquad \lambda_i > 0 , \qquad i=1, \ldots, K .$$

## 4. The Conditional Probability Model

Rarely will there be enough training pixels to estimate $P(I_t^* \mid I_t)$ directly. Instead we model the distribution of categories in the plane and derive $P(I_t^* \mid I_t)$ from the categories' model.

In this model the plane is partitioned by straight lines into a set of convex regions. The lines are generated by a random process called the Poisson field (see Appendix I). Each region is then independently assigned a category according to a scheme which selects category i

with probability $\pi_i$. Adjacent regions can receive the same category. (This model is due to Switzer (1965) who shows that the alternation of categories along any line is a Markov process.) The category of a pixel is taken to be the category assigned to its center point.

When the categories change on a scale that is large compared to the size of a pixel, the categories will induce one of the following patterns in most of the neighborhoods. (Formally, a pattern is a partition of the set {t, Nt, Et, St, Wt}.)

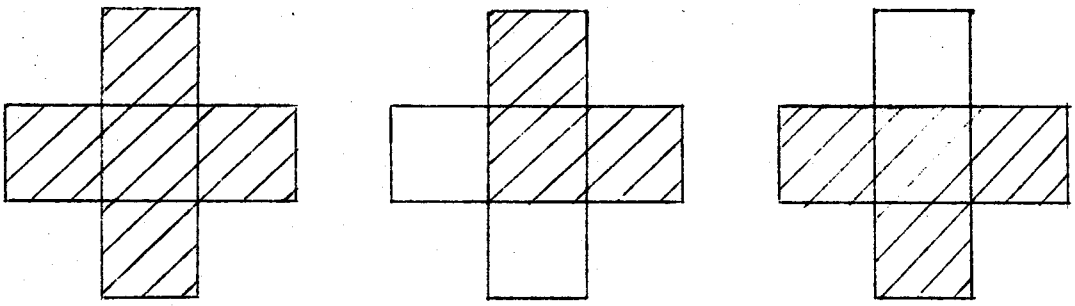

FIGURE 2

These patterns are called X, L, and T patterns, respectively. The L and T patterns can be rotated through multiples of ninety degrees, and the rotated patterns are also called L and T patterns.

Let $U_t$ denote the pattern formed by the boundary lines in the neighborhood of t. When the categories are assigned they induce a pattern in the neighborhood. This pattern will either be $U_t$ or a coarser pattern. When we need to distinguish between them, $U_t$ will be called the underlying pattern and the other pattern will be called the resultant pattern.

If the intensity parameter of the Poisson field is small (relative to the size of the pixels) then:

i) Each neighborhood has the same probability $\beta$ of intersecting one boundary line, and small probability of intersecting two or more boundaries, and irrespective of the intensity:

ii) the portion of a boundary within a neighborhood is a line segment, so that a single boundary can induce only an L or a T pattern, and

iii) the conditional probability that $U_t = L$ given that a single boundary has intersected the neighborhood is a constant $\alpha \doteq .41$ for all neighborhoods, and

iv) all four T patterns are equally probable as are all four L patterns.

The above properties together with the prior distribution provide a model for $P(I_t^* \mid I_t)$. Items i) - iv) are local properties of the process. The model should be useful on data for which these local properties are reasonable. The global properties of the boundary process do not influence the classifier. It should be noted that while $\alpha$ is a known constant (calculated in Appendix I), $\beta$ depends on the intensity parameter and will usually have to be estimated.

If $I_t^*$ and $I_t = i$ are inconsistent then

$$P(I_t^* \mid I_t = i) = 0$$

otherwise

$$P(I_t^* \mid I_t = i) = P(I_t^*)/\pi_i$$

$$= \frac{1}{\pi_i} \Sigma\, P(I_t^* \mid U_t = u)\, P(U_t = u) \qquad (2)$$

where the summation is over all possible patterns u.

From properties i) - iv) we obtain the second factor in each term of (2):

$$P(U_t = X) = 1-\beta$$

$$P(U_t = L) = \alpha\beta/4$$

$$P(U_t = T) = (1-\alpha)\ \beta/4$$

where L represents one of the four versions of that pattern and similarly for T.

The first factor in each term of (2) is obtained from the prior. Each pattern u of positive probability indicates that one or two regions created by the Poisson field intersect the neighborhood. If u indicates that only one region intersects the neighborhood then

$$P(I_t^* \mid U_t = u) = \pi_i$$

if $I_t^*$ indicates that all pixels are of category i. If u indicates that two regions intersect the neighborhood then

$$P(I_t^* \mid U_t = u) = \pi_i\pi_j$$

if $I_t^*$ indicates that the pixels in one region are of category i while the others are of category j. (It is possible that i=j.) If $I_t^*$ and u are inconsistent, then the probability is zero.

To illustrate the calculation of (2) we find $P(I_t^* = (i,i,i,i,j) \mid I_t = i)$. There is only one possible underlying pattern; call it $T_W$. Then

$$P(I_t^* = (i,i,i,i,j) \mid I_t = i)$$

$$= (1/\pi_i)\ P(I_t^* = (i,i,i,i,j) \mid U_t = T_W)\ P(U_t = T_W)$$

$$= (1/\pi_i)\ (\pi_i\pi_j)\ [\beta(1-\alpha)/4]$$

$$= \pi_j\ \beta(1-\alpha)/4\ .$$

Similarly for L patterns,

$$P(I_t^* = (i,i,i,j,j) \mid I_t = i) = \pi_j \beta\alpha/4\ .$$

Finally,

$$P(I_t^* = (i,i,i,i,i) \mid I_t = i)$$

$$= (1/\pi_i)\ [(1-\beta)\pi_i + 4(\beta\alpha\ \pi_i^2/4 + \beta(1-\alpha)\ \pi_i^2/4)]$$

$$= 1-\beta + \beta\pi_i$$

completes the specification of $P(I_t^* \mid I_t)$.

## 5. Training

To train the classifier, the true category is determined for some of the pixels and the parameters are estimated from these known categories and the observations for these pixels.

It is often most convenient to obtain the true categories for several rows of pixels; for example, a geologist might traverse the area under study, recording the rock types. We assume that this type of data is available rather than, say, a simple random sample of pixels.

The parameters $\mu_i$ and $\pi_i$ can be estimated by the usual sample estimates; $\hat{\pi}_i$ is the proportion of training pixels of category i, and $\hat{\mu}_i$ is the mean of $Z_t$ over all training pixels of category i.

The dispersion for category i was a constant $\lambda_i$ times a matrix S. The solutions to the likelihood equations for $\hat{\lambda}_i$ and $\hat{S}$ are not expressible in closed form. An iterative procedure for their calculation is described in Appendix II.

The parameter $\beta$ is estimated from the proportion of training pixels that are of the same category as both of their neighbors in the training row. The probability that a pixel t is of the same category as both its neighbors is

$$\gamma = 1-\beta + \beta \sum_i \pi_i^2 + \tfrac{1}{2}\beta(1-\alpha)\ (1 - \sum_i \pi_i^2) \ .$$

This corresponds to the three possibilities: no boundary crossed the neighborhood, one boundary crossed the neighborhood but both regions received the same category, and, a T pattern resulted for which the odd pixel was not in the training sample.

This leads to the estimate

$$\hat{\beta} = (1-\hat{\gamma})/[(1 - \Sigma\ \hat{\pi}_i^2)\ (\tfrac{1+\alpha}{2})]$$

where $\hat{\gamma}$ is the proportion of training pixels that are of the same category as both their neighbors. For simplicity this proportion is taken over those pixels in the training set that have two neighbors in the set, i.e. all but the pixels at the ends of the training rows.

## 6. Example and Conclusions

This model was applied to a 25 by 16 grid of pixels from the Yerington district of Nevada. These pixels are represented in Figure 3. The white, gray, and dark shades correspond to tertiary volcanic, altered quartz monzonite, and nonaltered quartz monzonite. These rock types are labelled 1, 2, and 3, respectively.

The training sample consisted of 100 pixels in the 4th, 8th, 12th, and 16th rows, counting from the top of the map.

A predicted map was obtained from the model for the interior 23 by 14 pixel grid. The map of predicted categories is shown in Figure 4. The map predicted by linear discriminant analysis (based on the same training sample and applied to the interior pixels) is shown in Figure 5.

To assess the accuracy of a predictive map a confusion matrix is calculated. The rows represent the true categories, the columns represent the predicted categories, and the entries represent the number of pixels of the row category that were predicted to be of the column category. The confusion matrices for the two maps are given in Figure 6.

It is apparent that category 3 is commonly misclassified as category 1. This is due to bias in the training sample. The problem is reinforced by neighborhood-based classification. Except for this problem, the neighborhood-based classifier was superior, classifying more pixels correctly and producing a smoother map.

The neighborhood-based classifier is not very successful on this data set, even though it outperforms linear discriminant analysis. It attempts to combine smoothing and edge detection, but only looks at five pixels. This data is so noisy and the underlying categories so smooth that it would make sense to use information from more than five pixels.

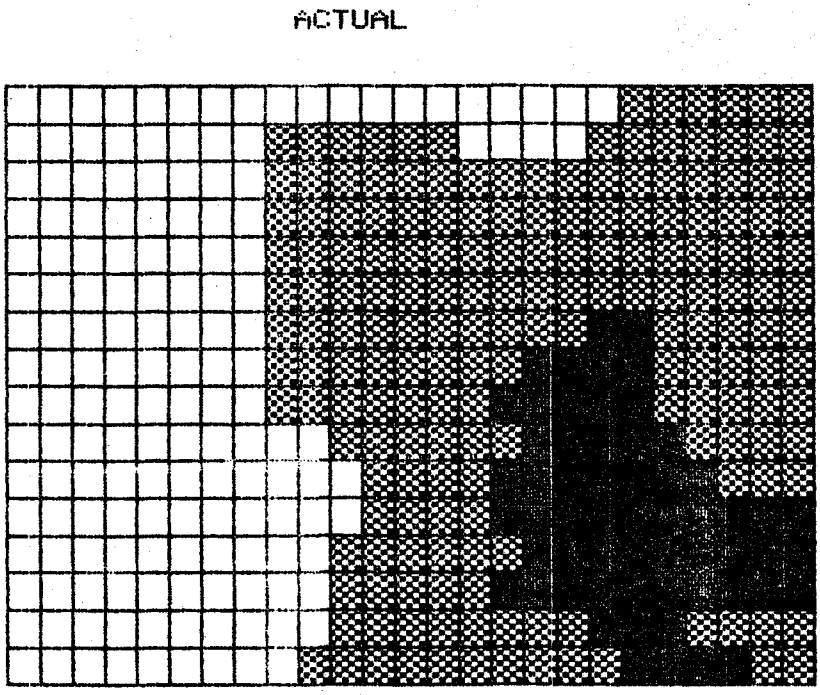


FIGURE 3

Ground Truth

PREDICTED

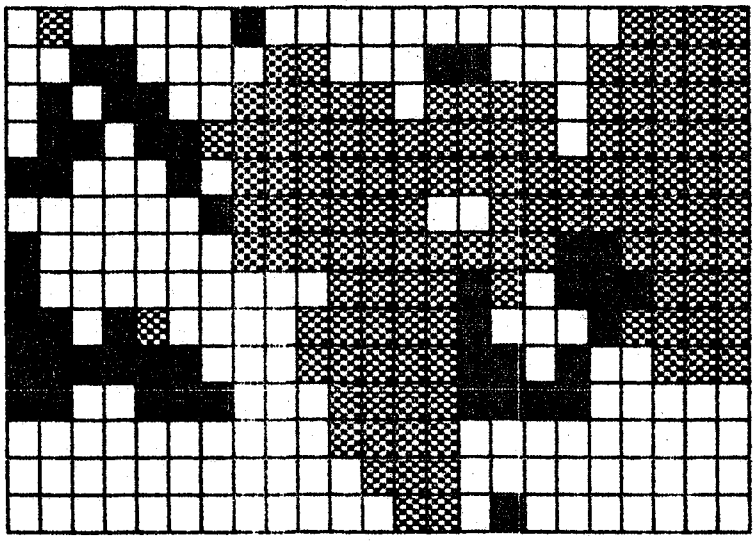

FIGURE 4

Predictions of the Neighborhood Based Classifier

PREDICTED

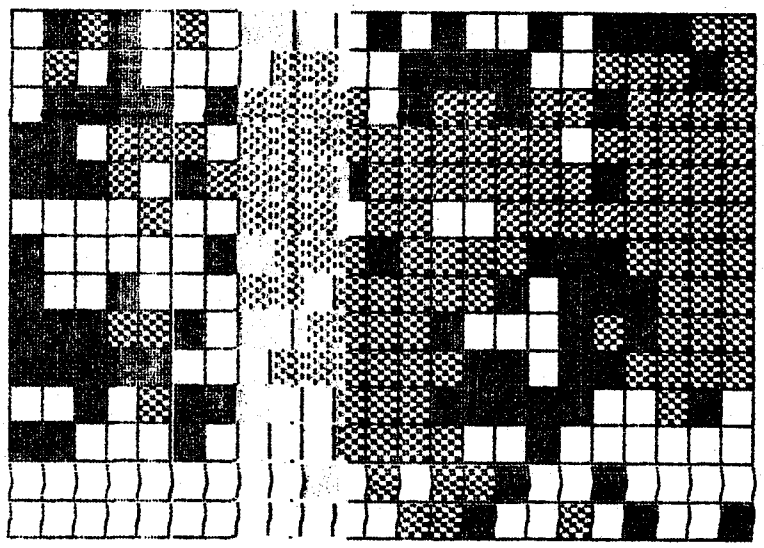

FIGURE 5

Predictions of Linear Discriminant Analysis

FIGURE 6

Confusion matrix for linear discriminant analysis:

| | | Predicted | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Actual | 1 | 52 | 13 | 27 |
| | 2 | 26 | 78 | 18 |
| | 3 | 18 | 8 | 13 |

Confusion matrix for neighborhood based classifier:

| | | Predicted | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Actual | 1 | 68 | 4 | 20 |
| | 2 | 32 | 84 | 6 |
| | 3 | 27 | 5 | 7 |

Acknowledgements

This research was supported in part also by the Natural Sciences and Engineering Research Council of Canada.

## APPENDIX I

This appendix describes the Poisson field and derives the value of $\alpha$.

One way to generate the Poisson field in a bounded region $R$ is as follows. Construct a circle with center $O$ containing $R$. Generate points $P_i$ with polar coordinates $(r_i, \theta_i)$ centered at $O$, where the $r_i$ are points of a Poisson process on the positive reals and for each $r_i$, $\theta_i$ is uniformly distributed on $(0, 2\pi]$. To each point $P_i$ associate the line $\ell_i$ through $P_i$ perpendicular to $OP_i$. The lines $\ell_i$ are the lines of the Poisson field.

For each neighborhood there are eight patterns that can be induced by a boundary. To these patterns there correspond eight disjoint regions in the plane such that points generated in one of these regions yield boundaries that induce the corresponding pattern. It follows that the number of lines that would induce a given pattern in a neighborhood has a Poisson distribution that is independent of the number of lines that would induce any other pattern.

The Poisson field is described in greater depth in Solomon (1978). In particular, Solomon shows that the number of lines of the field intersecting a line segment of length $d$ has a Poisson distribution with mean $\tau d$ for some $\tau > 0$. The process is also invariant under rotation and translation, so that the distribution of the number of boundaries that would induce a pattern is the same for all four $L$ patterns (and similarly for $T$ patterns). The distributions are also the same for all pixels.

Let $\lambda_t$ ($\lambda_\ell$) be the expected number of lines that induce one of the four T (L) patterns in a neighborhood.

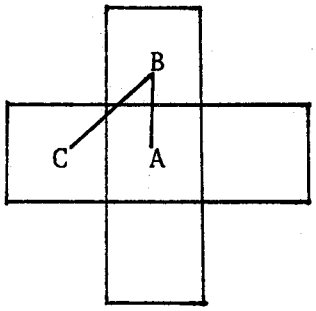


FIGURE 7

Consider the neighborhood in Figure 7. If a generated line intersects AB then it induces a T pattern or one of two possible L patterns. Conversely if any of those three patterns is induced by a line, the line must intersect AB. Taking AB to be of unit length, this implies that

$$\tau = \lambda_t + 2\lambda_\ell \ .$$

Similarly, consideration of segment BC yields

$$\sqrt{2}\,\tau = 2\lambda_t + 2\lambda_\ell$$

so that

$$\lambda_t = (\sqrt{2} - 1)\tau$$

$$\lambda_\ell = \frac{1}{2}(2 - \sqrt{2})\tau \ .$$

Let $N_t$ and $N_\ell$ be the numbers of T and L inducing lines generated in a neighborhood. They are independent Poisson variates with means $4\lambda_t$ and $4\lambda_\ell$ respectively. The conditional probability that an L pattern is induced given that exactly one boundary crosses a neighborhood is

$$\begin{aligned}
\alpha &= P(N_\ell = 1 \mid N_t + N_\ell = 1) \\
&= \frac{P(N_\ell = 1,\ N_t = 0)}{P(N_\ell = 1)\ \ N_t = 0) + P(N_\ell = 0,\ N_t = 1)} \\
&= \lambda_\ell/(\lambda_\ell + \lambda_t) \\
&= \sqrt{2} - 1 \\
&\approx .41 \ .
\end{aligned}$$

## APPENDIX II

We have a training sample with $n_i$ data vectors $Z_j$ corresponding to pixels of category $i$ for $i=1, \ldots, K$. We want maximum likelihood estimates of $\mu_i$, $\lambda_i$, and $S$ where

$$j \in i \Rightarrow Z_j \sim N(\mu_i, \lambda_i S)$$

(We identify each category with the set of items of that category.) In the above, $Z_j$ are $\nu$-vectors, $S$ is $\nu \times \nu$ positive definite, and the $\lambda_i$ are positive. We normalize by setting $\lambda_1 = 1$.

It is known that a solution to the likelihood equation exists and that such a solution is a local maximum of the likelihood (Kim, 1971). It is not known whether such a solution is unique.

A solution to the likelihood equations was found numerically. The algorithm alternated between estimating the $\lambda$'s for fixed $\mu$'s and $S$, and estimating the $\mu$'s and $S$ for fixed $\lambda$'s.

For any set of $\lambda$'s one can note that

$$j \in i \Rightarrow Z_j/\sqrt{\lambda_i} \sim N\left(\frac{\mu_i}{\sqrt{\lambda_i}}, S\right)$$

whence the likelihood is maximized at

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j \in i} Z_j$$

and

$$\hat{S}(\lambda) = \frac{1}{n} \sum_i \frac{n_i S_i}{\lambda_i}$$

where

$$n = \sum_i n_i$$

and

$$S_i = \frac{1}{n_i} \sum_{j \in i} (Z_j - \mu_i)(Z_j - \mu_i)' \ .$$

Since the estimate of $\mu_i$ does not depend on the $\lambda$'s the algorithm alternates between estimating $S$ and estimating the $\lambda$'s.

For fixed $S$ the likelihood is

$$L = \prod_i \prod_{j \in i} \frac{1}{\sqrt{2\pi}^{\nu} \ |\lambda_i S|^{\frac{1}{2}}} e^{-\frac{1}{2} \left( (\lambda_i S)^{-1} (Z_j - \mu_i)(Z_j - \mu_i)' \right)}$$

so the log likelihood is

$$\ell = \text{const} + \sum_i \sum_j \frac{-\nu}{2} \log \lambda_i - \frac{1}{2\lambda_i} \operatorname{tr}\left( S^{-1} (Z_j - \mu_i)(Z_j - \mu_i)' \right)$$

$$= \text{const} + \sum_i \left[ -\frac{n_i \nu}{2} \log \lambda_i - \frac{n_i}{2\lambda_i} \operatorname{tr}(S^{-1} S_i) \right]$$

where $S_i$ is as above.

Now for $i \neq 1$

$$\frac{\partial \ell}{\partial \lambda_i} = \frac{-n_i \nu}{2\lambda_i} + \frac{n_i \operatorname{tr}(S^{-1} S_i)}{2\lambda_i^2}$$

which vanishes at

$$\hat{\lambda}_i(S) = \frac{1}{\nu} \operatorname{tr}(S^{-1} S_i) \ .$$

In the example starting values

$$\lambda_i^{(0)} = \text{tr } S_i / \text{tr } S_1$$

were used, the convergence criterion was

$$\sum_i |\lambda_i^{(K+1)} - \lambda_i^{(K)}| < .001$$

and five iterations were required.

Round-off errors should not propagate in this algorithm since the matrices $S_i$ do not change. The author does not know of any previously published use of this M.L.E. algorithm. It is an application of the Rosenbrock method (Rosenbrock, 1960).

## REFERENCES


Kim, D. Y. (1971). "Statistical Inference for Constants of Proportionality Between Covariance Matrices," Technical Report No. 59, Department of Statistics, Stanford University.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. John Wiley & Sons, New York.

Rosenbrock, H. H. (1960). "An Automatic Method for Finding the Greatest of Least Value of a Function," Comput. J. 3, pp. 175-184.

Solomon, H. (1978). Geometric Probability. SIAM Publications, Philadelphia, PA.

Switzer, P. (1965). "A Random Set Process in the Plane with a Markov Property," Ann. Math. Statist. 36, pp. 1859-1863.

Switzer, P., Kowalki, W. S., Lyon, R. J. P. (1981). "A Prior Probability Method for Smoothing Discriminant Analysis Classification Maps," Technical Report No. 1, Department of Statistics, Stanford University.