AgRISTARS

SR-L1-04031
JSC-16853
JAN 2 1 1981

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

Supporting Research

January 1981

MAXIMUM LIKELIHOOD CLUSTERING WITH DEPENDENT FEATURE TREES

C. B. Chittineni

(E81-10186) MAXIMUM LIKELIHOOD CLUSTERING WITH DEPENDENT FEATURE TREES (Lockheed Engineering and Management) 54 p
HC A04/MF A01 CSCL 12A G3/43
N81-29502
Unclas 00186

Lockheed Engineering and Management Services Company, Inc.
1830 NASA Road 1, Houston, Texas 77058

Lyndon B. Johnson Space Center
Houston, Texas 77058

1 Report No	2 Government Accession No	3 Recipient's Catalog No
SR-L1-04031, JSC-16853		

4 Title and Subtitle	5 Report Date
Maximum Likelihood Clustering With Dependent Feature Trees	January 1981
	6 Performing Organization Code

7 Author(s)	8 Performing Organization Report No
C. B. Chittineni Lockheed Engineering and Management Services Company, Inc.	LEMSCO-15683
9 Performing Organization Name and Address	10 Work Unit No
Lockheed Engineering and Management Services Company, Inc. 1830 NASA Road 1 Houston, Texas 77058	11 Contract or Grant No NAS 9-15800
12 Sponsoring Agency Name and Address	13 Type of Report and Period Covered Technical Report
National Aeronautics and Space Administration Lyndon B. Johnson Space Center Houston, Texas 77058 (Technical Monitor Dr. R. Heydorn)	14 Sponsoring Agency Code

15 Supplementary Notes

16 Abstract

In this report, maximum likelihood clustering for the decomposition of mixture density of the data into its normal component densities is considered. The densities are approximated with first-order dependent feature trees using criteria of mutual information and distance measures. Expressions are presented for the criteria when the densities are Gaussian. By defining different types of nodes in a general dependent feature tree, maximum likelihood equations are developed for the estimation of parameters using fixed-point iterations. The field structure of the data is also taken into account in developing maximum likelihood equations. Furthermore, experimental results from the processing of remotely sensed multispectral scanner imagery data are presented.

17 Key Words (Suggested by Author(s))	18 Distribution Statement
Clustering, distance measures, dependent feature trees, fields, fixed-point iteration schemes, link, maximum likelihood equations, mutual information, parameter estimation, types of nodes	

19 Security Classif (of this report)	20 Security Classif (of this page)	21 No of Pages	22 Price*
Unclassified	Unclassified	54	

*For sale by the National Technical Information Service, Springfield, Virginia 22161

NASA — JSC

SR-L1-04031
JSC-16853

MAXIMUM LIKELIHOOD CLUSTERING WITH DEPENDENT FEATURE TREES

Job Order 73-306

This report describes Classification activities of the Supporting Research project of the AgRISTARS program.

PREPARED BY

C. B. Chittineni

APPROVED BY

T. C. Minter, Supervisor
Techniques Development Section

J. E. Wainwright, Manager
Development and Evaluation Department

LOCKHEED ENGINEERING AND MANAGEMENT SERVICES COMPANY, INC.

Under Contract NAS 9-15800

For

Earth Resources Research Division

Space and Life Sciences Directorate

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
LYNDON B. JOHNSON SPACE CENTER
HOUSTON, TEXAS

January 1981

LEMSCO-15683

PREFACE

The techniques which are the subject of this report were developed to support the Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing program. Under Contract NAS 9-15800, Dr. C. B. Chittineni, a principal scientist for Lockheed Engineering and Management Services Company, Inc., performed this research for the Earth Resources Research Division, Space and Life Sciences Directorate, National Aeronautics and Space Administration, at the Lyndon B. Johnson Space Center.

CONTENTS

FIGURES

1. INTRODUCTION

Recently, considerable interest has been shown in developing techniques for the classification of imagery data (such as remotely sensed multispectral scanner data acquired by the Landsat series of satellites) for inventorying natural resources, monitoring crop conditions, and detecting changes in natural and manmade objects. Nonsupervised classification or clustering techniques have been found to be effective in the analysis of remotely sensed data (ref. 1). The approach of clustering for imagery data classification, in general, involves two steps: (1) partitioning the image into its inherent modes or into its homogeneous parts and (2) labeling the clusters using information from a given set of labeled patterns.

In practical applications of pattern recognition such as remote sensing, it is difficult to obtain labels for the patterns. In remote sensing imagery, an analyst-interpreter provides the labels for the picture elements (pixels) by examining imagery films and using other information (e.g., crop growth stage models and historic information). Remote sensing imagery usually has a field structure, and it is recognized that fields are easier to label than are pixels. The development of algorithms for locating fields has attracted the attention of several researchers in the recent literature (refs. 2-5).

Considerable interest has been shown in applying maximum likelihood equations for the decomposition of the mixture density of the imagery data into its normal component densities (refs. 5-9). Recently, methods have been developed (refs. 10, 11) for probabilistically labeling the modes of the data using information from a given set of labeled patterns and, also, from a given set of labeled fields.

In decomposing the mixture density of the data into its normal component densities, the parameters of the component densities and the a priori probabilities of the modes are iteratively computed using maximum likelihood equations coupled with a split and merge sequence. The updating of the parameters is usually stopped after a few iterations because of the large amount of computation.

Also, in practical problems (remote sensing imagery data of several acquisitions), a large number of parameters will be estimated. For a fixed sample size, the accuracy of estimation usually decreases (ref. 12) as the number of parameters to be estimated increases. To overcome the computational requirements and the large number of parameters to be estimated with the usual maximum likelihood clustering technique, maximum likelihood equations are obtained in this report by approximating the cluster conditional densities with first-order tree dependence (refs. 13, 14) among the features. The field structure of the data is also taken into account. Either the average mutual information between the features (ref. 13) or the probabilistic distance measures (ref. 15) can be used to construct optimal dependent feature trees for a given data type.

This paper is organized as follows. General maximum likelihood equations are presented in section 2. Section 3 concerns the problem of approximating probability density functions with dependent feature trees using the criteria of information measure and probabilistic distance measure. Expressions are derived for the criteria when the distributions of the features are Gaussian. In section 4, a general dependent feature tree and its various types of nodes are described, and expressions for the covariance between the features not connected by a single link are derived. Maximum likelihood equations for the parameters of the density functions when approximated by dependent feature trees are developed in section 5. Experimental results from the processing of remotely sensed multispectral scanner imagery data are presented in section 6. Section 7 contains the concluding remarks. Detailed derivations of maximum likelihood equations are given in appendix A. In appendix B, the field structure of the data is taken into account in developing maximum likelihood equations. An expression is derived in appendix C for the mutual information between the feature subsets when they are represented by the nodes in a dependent feature tree. Also, expressions are derived for the covariance between the feature subsets when they are connected by a path in a dependent feature tree.

2. GENERAL MAXIMUM LIKELIHOOD EQUATIONS

General maximum likelihood equations are presented in this section for the decomposition of the mixture density of the data into its component densities. It is assumed that a set $\mathcal{X} = \{X_1, \cdots, X_N\}$ of N unlabeled patterns, each of dimension n, is given. These patterns are assumed to be drawn independently from the mixture density

$$p(X|\theta) = \sum_{j=1}^{m} p(X, \omega_j, \theta_j) P(\omega_j) \tag{2-1}$$

where θ is a fixed but unknown parameter vector, θ_j is a parameter vector for the j^{th} cluster, and m is the number of modes or clusters in the data. Let $P(\omega_j)$ and $p(X|\omega_j)$ be the a priori probabilities of the modes and mode conditional densities, respectively. The likelihood of the observed pattern vectors is, by definition, the joint density

$$p(\mathcal{X}|\theta) = \prod_{k=1}^{N} p(X_k|\theta) \tag{2-2}$$

Since the logarithm is a monotonic function of its argument, taking the gradient of the logarithm of equation (2-2) with respect to θ_i results in

$$\nabla_{\theta_1} \ell = \sum_{k=1}^{N} \frac{1}{p(X_k|\theta)} \nabla_{\theta_1} \left[\sum_{j=1}^{m} p(X_k|\omega_j, \theta_j) P(\omega_j) \right] \tag{2-3}$$

where

$$\ell = \sum_{k=1}^{N} \log[p(X_k|\theta] \tag{2-4}$$

and $\nabla_{\theta_1} \ell$ is the gradient of ℓ with respect to θ_1. From the Bayes rule, the a posteriori probability can be written as

$$p(\omega_1|X_k, \theta) = \frac{p(X_k|\omega_i, \theta_i) P(\omega_i)}{p(X_k|\theta)} \tag{2-5}$$

If the elements of θ_i and θ_j are assumed to be functionally independent, using equation (2-5) in equation (2-3) yields

$$\nabla_{\theta_i} \ell = \sum_{k=1}^{N} p(\omega_1 | X_k, \theta) \nabla_{\theta_i} \left\{ \log[p(X_k | \omega_i, \theta_i) P(\omega_1)] \right\} \tag{2-6}$$

The following likelihood equation for the a priori probabilities can easily be obtained from equation (2-6) by introducing Lagrangian multipliers to take into account the probability constraints on $P(\omega_i)$.

$$P(\omega_i) = \frac{1}{N} \sum_{k=1}^{N} p(\omega_i | X_k, \theta) \tag{2-7}$$

Since θ_i is a parameter vector of the density of the i^{th} cluster, equation (2-6) can be written as

$$\nabla_{\theta_i} \ell = \sum_{k=1}^{N} p(\omega_i | X_k, \theta) \nabla_{\theta_i} \left\{ \log[p(X_k | \omega_i, \theta_i)] \right\} \tag{2-8}$$

From equation (2-8), general maximum likelihood equations for the parameters of the cluster conditional densities can be obtained.

3. APPROXIMATING PROBABILITY DENSITY FUNCTIONS WITH DEPENDENT FEATURE TREES

If the probability density function of the i^{th} class is approximated by a first-order dependent feature tree, it can be written as

$$p_i(X) = \prod_{\ell=1}^{n} p_i\left[x_{m_\ell} \middle| x_{m_{j(\ell)}}\right] \quad ; \quad 0 \leq j(\ell) < \ell \tag{3-1}$$

where x_{m_ℓ} is the m_ℓ^{th} feature of pattern vector X; $(m_1, \cdots, m_n)$ is an unknown permutation of integers 1, 2, $\cdots$, n; and $p(x_i|x_0)$, by definition, is equal to $p(x_i)$. Each variable in the above expansion may be conditioned upon, at most, one of the other variables. Figure 3-1 shows an example of a dependent feature tree.

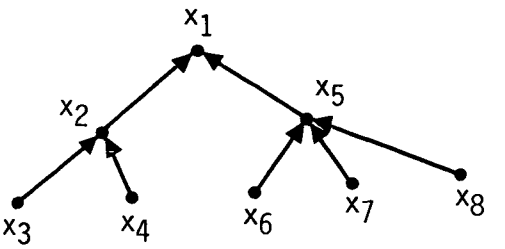

Figure 3-1.- An example of a dependent feature tree.

The component of the density in the product approximation that is represented by a single link, such as the one connecting features x_5 and x_8 in figure 3-1, is $p(x_8|x_5)$. The density that is approximated by the dependence tree of figure 3-1 can be written as

$$p(X) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_2)p(x_5|x_1)p(x_6|x_5)p(x_7|x_5)p(x_8|x_5) \tag{3-2}$$

3.1 CONSTRUCTION OF OPTIMAL DEPENDENT FEATURE TREES

This section concerns the problem of constructing dependent feature trees. The dependent feature tree, the density of which best approximates the true density, is proposed to be constructed using either the criterion of information preservation (ref. 13) or the criterion of class separability (ref. 15). An algorithm developed by Kruskal (ref. 16) provides an efficient computational procedure for constructing optimal dependent feature trees using the expressions developed in the following.

3.1.1 A CRITERION BASED ON INFORMATION MEASURE

Let $p_{1t}(X)$ be the approximate density of the 1^{th} class with the product approximation. That is,

$$p_{1t}(X) = \prod_{\ell=1}^{n} \left\{ p_i \left[x_{m_\ell} \middle| x_{m_{j(\ell)}} \right] \right\} \tag{3-3}$$

Consider the following measure of closeness between the true and approximate densities (ref. 13). That is,

$$I(p,p_t) = \sum_{i=1}^{C} P(\omega_i) \int p_1(X) \log \left[\frac{p_i(X)}{p_{it}(X)} \right] dX \tag{3-4}$$

where C is the number of classes. From equation (3-4), it is seen that $I(p,p_t) = 0$ whenever $p_i(X)$ is equal to $p_{it}(X)$ for all X and that $I(p,p_t) > 0$ if $p_1(X)$ is different from $p_{1t}(X)$ for some X. To find the product approximation for the densities or the dependent feature tree that minimizes $I(p,p_t)$, consider

$$\begin{aligned} I(p,p_t) &= - \sum_{i=1}^{C} P(\omega_i) \int p_i(X) \log[p_{it}(X)] dX \\ &\quad + \sum_{i=1}^{C} P(\omega_1) \int p_i(X) \log[p_i(X)] dX \\ &= - \sum_{\ell=1}^{n} I[x_\ell, x_{j(\ell)}] + K \end{aligned} \tag{3-5}$$

where

$$\left.\begin{aligned} I[x_\ell, x_{j(\ell)}] &= \sum_{i=1}^{C} P(\omega_i) I_1[x_\ell, x_{j(\ell)}] \\ I_1[x_\ell, x_{j(\ell)}] &= \int p_i[x_\ell, x_{j(\ell)}] \log \left\{ \frac{p_i[x_\ell, x_{j(\ell)}]}{p_i(x_\ell) p_i[x_{j(\ell)}]} \right\} dx_\ell \, dx_{j(\ell)} \end{aligned}\right\} \tag{3-6}$$

and $$K = - \sum_{\ell=1}^{n} I(x_\ell) + \sum_{i=1}^{C} P(\omega_i) \int p_i(X) \log[p_1(X)] dX$$

The quantity $I_1[x_\ell, x_{J(\ell)}]$ is the mutual information between features x_ℓ and $x_{J(\ell)}$ of class 1. From equation (3-5), Kruskal's algorithm (ref. 16) can be efficiently used to construct optimal dependent feature trees.

3.1.2 A CRITERION BASED ON PROBABILISTIC DISTANCE MEASURES

A probability density function, like any other function, can be approximated by a number of different procedures. In the sense of preserving the separability between the classes, it is proposed that a criterion based on probabilistic distance measures such as divergence be used to construct dependent feature trees. The measure of closeness between the approximate and true densities is defined as

$$J_{12} = \int p_1(X)\log\left[\frac{p_{1t}(X)}{p_{2t}(X)}\right]dX + \int p_2(X)\log\left[\frac{p_{2t}(X)}{p_{1t}(X)}\right]dX \tag{3-7}$$

From equation (3-7), it is seen that J_{12} is large whenever the ratio of $p_{1t}(X)$ to $p_{2t}(X)$ is large in the region of class 1 and the ratio of $p_{2t}(X)$ to $p_{1t}(X)$ is large in the region of class 2. By using the product approximation of equation (3-3) for the densities $p_{it}(X)$, equation (3-7) can be written as follows.

$$\begin{aligned}
J_{12} &= \sum_{i=1}^{n} \int p_1(X)\log\left\{\frac{p_1\left[x_{m_1}\middle|x_{m_{J(i)}}\right]}{p_2\left[x_{m_1}\middle|x_{m_{J(1)}}\right]}\right\}dX \\
&\quad + \sum_{i=1}^{n} \int p_2(X)\log\left\{\frac{p_2\left[x_{m_1}\middle|x_{m_{J(i)}}\right]}{p_1\left[x_{m_1}\middle|x_{m_{j(1)}}\right]}\right\}dX \\
&= \sum_{i=1}^{n} \int p_1\left[x_{m_i}, x_{m_{J(1)}}\right]\log\left\{\frac{p_1\left[x_{m_i}\middle|x_{m_{J(i)}}\right]}{p_2\left[x_{m_1}\middle|x_{m_{j(i)}}\right]}\right\}dx_{m_i}\, dx_{m_{j(i)}} \\
&\quad + \sum_{i=1}^{n} \int p_2\left[x_{m_i}, x_{m_{J(i)}}\right]\log\left\{\frac{p_2\left[x_{m_i}\middle|x_{m_{j(1)}}\right]}{p_1\left[x_{m_i}\middle|x_{m_{J(1)}}\right]}\right\}dx_{m_i}\, dx_{m_{J(i)}} \\
&= \sum_{i=1}^{n} \left\{\Delta J_{12}\left[x_{m_1}, x_{m_{J(1)}}\right]\right\} + K
\end{aligned} \tag{3-8}$$

where

$$\Delta J_{12}\left[x_{m_i}, x_{m_{j(i)}}\right] = J_{12}\left[x_{m_i}, x_{m_{j(i)}}\right] - J_{12}\left(x_{m_i}\right) - J_{12}\left[x_{m_{j(i)}}\right] \tag{3-9}$$

and

$$K = \sum_{i=1}^{n} \int p_1\left(x_{m_i}\right) \log\left[\frac{p_1\left(x_{m_i}\right)}{p_2\left(x_{m_i}\right)}\right] dx_{m_i} + \sum_{i=1}^{n} \int p_2\left(x_{m_i}\right) \log\left[\frac{p_2\left(x_{m_i}\right)}{p_1\left(x_{m_i}\right)}\right] dx_{m_i} \tag{3-10}$$

If more than two classes exist, the expected value of the measure of closeness defined over pairs of classes can be used to obtain optimal approximations for the densities (ref. 17). From equation (3-8), Kruskal's algorithm (ref. 16) can be efficiently used to construct optimal dependent feature trees.

3.2 EXPRESSIONS FOR THE CRITERIA WHEN THE DISTRIBUTIONS OF THE FEATURES ARE GAUSSIAN

Expressions are derived in this section for the mutual information and for ΔJ_{12} between the features, assuming that the distributions of the features are Gaussian. If $p_\ell(x_i)$, the density of feature x_i of the ℓ^{th} class, is Gaussian, it can be written as

$$p_\ell(x_i) = \frac{1}{\sqrt{2\pi\sigma_i(\ell)}} \exp\left\{-\frac{1}{2\sigma_i(\ell)}\left[x_i - u_i(\ell)\right]^2\right\} \tag{3-11}$$

or it is denoted as $p_\ell(x_i) \sim N[u_i(\ell), \sigma_i(\ell)]$. The joint and conditional densities of features x_i and x_j of the ℓ^{th} class can be written as follows.

$$p_\ell(x_i, x_j) = \frac{1}{2\pi\left\{\sigma_i(\ell)\sigma_j(\ell)\left[1 - \rho_{ij}^2(\ell)\right]\right\}^{1/2}} \exp\left[-\frac{1}{2} q_\ell(x_i, x_j)\right] \tag{3-12}$$

where

$$q_\ell(x_i, x_j) = \frac{1}{2\left[1 - \rho_{ij}^2(\ell)\right]}\left\{\left[\frac{x_i - u_i(\ell)}{\sqrt{\sigma_i(\ell)}}\right]^2 - 2\rho_{ij}(\ell)\left[\frac{x_i - u_i(\ell)}{\sqrt{\sigma_i(\ell)}}\right]\left[\frac{x_j - u_j(\ell)}{\sqrt{\sigma_j(\ell)}}\right] + \left[\frac{x_j - u_j(\ell)}{\sqrt{\sigma_j(\ell)}}\right]^2\right\} \tag{3-13}$$

$$\rho_{ij}^2(\ell) = \frac{\sigma_{ij}^2(\ell)}{\sigma_i(\ell)\sigma_j(\ell)} \tag{3-14}$$

and $\sigma_{ij}(\ell)$ is the covariance between features x_i and x_j of the ℓ^{th} class. From equations (3-11) and (3-12), the conditional density can be written as

$$p_\ell(x_i|x_j) = \frac{1}{\left\{2\pi\sigma_i(\ell)\left[1 - \rho_{ij}^2(\ell)\right]\right\}^{1/2}} \exp[-q_\ell(x_i|x_j)] \tag{3-15}$$

where

$$q_\ell(x_i|x_j) = \frac{1}{2\sigma_i(\ell)\left[1 - \rho_{ij}^2(\ell)\right]} \left\{[x_i - u_i(\ell)] - \rho_{ij}(\ell)\sqrt{\frac{\sigma_i(\ell)}{\sigma_j(\ell)}}\,[x_j - u_j(\ell)]\right\}^2 \tag{3-16}$$

3.2.1 AN EXPRESSION FOR THE MUTUAL INFORMATION BETWEEN FEATURES x_i AND x_j

In this section, an expression is derived for the mutual information between Gaussian-distributed features x_i and x_j of class ℓ. From equations (3-11) and (3-15), the following can easily be obtained. Consider

$$\frac{p_\ell(x_i,x_j)}{p_\ell(x_i)p_\ell(x_j)} = \frac{p_\ell(x_i|x_j)}{p_\ell(x_i)}$$

$$= \left[1 - \rho_{ij}^2(\ell)\right]^{-1/2} \exp\left(\frac{\rho_{ij}(\ell)}{\left[1 - \rho_{ij}^2(\ell)\right]}\left[\frac{x_i - u_i(\ell)}{\sqrt{\sigma_i(\ell)}}\right]\left[\frac{x_j - u_j(\ell)}{\sqrt{\sigma_j(\ell)}}\right]\right.$$

$$\left. - \frac{\rho_{ij}^2(\ell)}{2\left[1 - \rho_{ij}^2(\ell)\right]}\left\{\left[\frac{x_i - u_i(\ell)}{\sqrt{\sigma_i(\ell)}}\right]^2 + \left[\frac{x_j - u_j(\ell)}{\sqrt{\sigma_j(\ell)}}\right]^2\right\}\right) \tag{3-17}$$

From equation (3-17), the mutual information between features x_i and x_j of class ℓ can be obtained as follows.

$$I_\ell(x_i,x_j) = \int p_\ell(x_i,x_j)\log\left[\frac{p_\ell(x_i,x_j)}{p_\ell(x_i)p_\ell(x_j)}\right]dx_i\,dx_j$$

$$= -\frac{1}{2}\log\left[1 - \rho_{ij}^2(\ell)\right] \geq 0 \tag{3-18}$$

3.2.2 AN EXPRESSION FOR $\Delta J_{12}(x_i, x_j)$ WHEN FEATURES x_i AND x_j ARE NORMALLY DISTRIBUTED

From equations (3-7) and (3-9), when features x_i and x_j are normally distributed, an expression for $\Delta J_{12}(x_i, x_j)$ can be easily obtained as follows.

$$
\begin{aligned}
2\Delta J_{12}(x_i,x_j) = \Bigg\{ & \left[\frac{\sigma_j(1)\sigma_i(2) - 2\sigma_{ij}(1)\sigma_{ij}(2) + \sigma_i(1)\sigma_j(2)}{\Delta_{ij}(1)} + \frac{\sigma_j(2)\sigma_i(1) - 2\sigma_{ij}(1)\sigma_{ij}(2) + \sigma_i(2)\sigma_j(1)}{\Delta_{ij}(2)} - 2 \right] \\
& + \left[\frac{\sigma_j(1)}{\Delta_{ij}(1)} + \frac{\sigma_j(2)}{\Delta_{ij}(2)} \right] [u_i(1) - u_i(2)]^2 - \left[\frac{\sigma_{ij}(1)}{\Delta_{ij}(1)} + \frac{\sigma_{ij}(2)}{\Delta_{ij}(2)} \right] 2[u_i(1) - u_i(2)][u_j(1) - u_j(2)] \\
& + \left[\frac{\sigma_i(1)}{\Delta_{ij}(1)} + \frac{\sigma_i(2)}{\Delta_{ij}(2)} \right] [u_j(1) - u_j(2)]^2 \Bigg\} - \left\{ \left[\frac{\sigma_i(2)}{\sigma_i(1)} + \frac{\sigma_i(1)}{\sigma_i(2)} - 2 \right] + \left[\frac{1}{\sigma_i(1)} + \frac{1}{\sigma_i(2)} \right] [u_i(1) - u_i(2)]^2 \right\} \\
& - \left\{ \left[\frac{\sigma_j(2)}{\sigma_j(1)} + \frac{\sigma_j(1)}{\sigma_j(2)} - 2 \right] + \left[\frac{1}{\sigma_j(1)} + \frac{1}{\sigma_j(2)} \right] [u_j(1) - u_j(2)]^2 \right\}
\end{aligned}
\tag{3-19}
$$

where

$$
\Delta_{ij}(1) = \sigma_i(1)\sigma_j(1) - \sigma_{ij}^2(1) \tag{3-20}
$$

4. A GENERAL DEPENDENT FEATURE TREE

A general dependent feature tree is shown in figure 4-1.

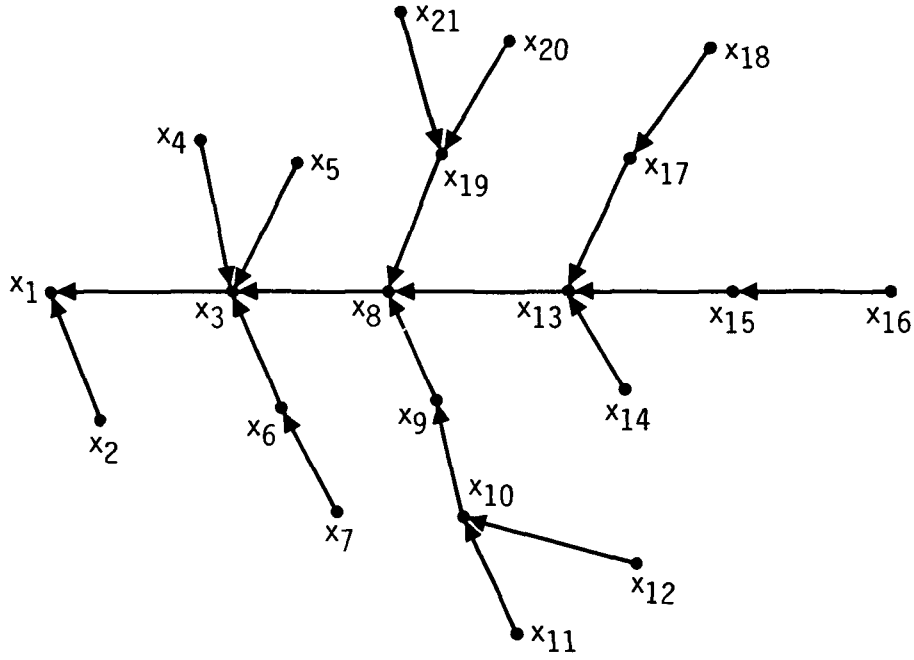

Figure 4-1.- A general dependent feature tree.

Each node of the tree represents a feature, and the feature numbers are given in figure 4-1. In approximating the probability density functions with dependent feature trees, each feature may be conditioned upon, at most, one of the other features. Node x_1 is the root node of the tree. Nodes x_2, x_4, x_5, x_7, etc., are nodes on the periphery of the tree.

4.1 DIFFERENT TYPES OF NODES

For convenience in the following analysis, the nodes of the dependent feature tree are divided into the following types.

1. Type I nodes: Except for the root node, nodes on the periphery of the tree are defined as type I nodes. For example, in figure 4-1, nodes x_2, x_4, x_5, x_7, etc., are type I nodes.

2. Type II nodes: These are nodes which are one node deep from the periphery. For example, in figure 4-1, nodes x_6, x_{10}, x_{15}, x_{17}, x_{19}, etc., are type II nodes.

3. Type III nodes: These are nodes which are at least two nodes deep from the periphery. For example, in figure 4-1, nodes x_3, x_8, x_9, x_{13}, etc., are type III nodes.

4. Type IV nodes: The root node of the tree is defined as a type IV node. The types of root nodes are divided into type IVa and type IVb. Examples of the types of root nodes are described in the following.

   a. Type IVa node: The type IVa node is the root node of a tree with a single link. As an example, node x_1 of figure 4-2 is a type IVa node.

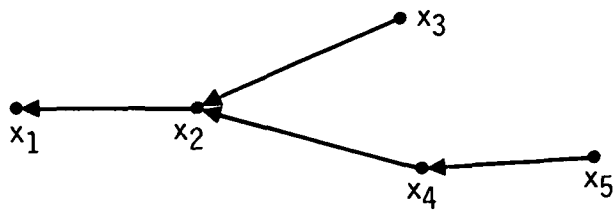

Figure 4-2.- Illustration of a type IVa node.

   b. Type IVb node: The type IVb node is the root node of a tree with two or more links. As an example, node x_1 of figure 4-3 is a type IVb node.

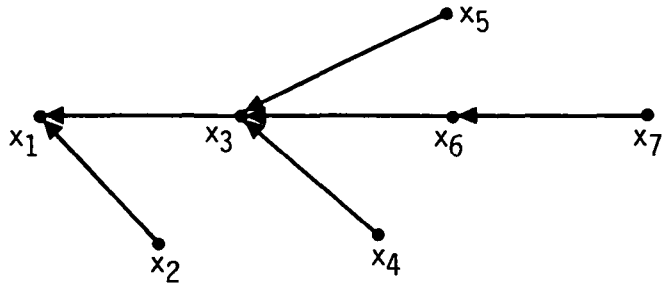

Figure 4-3.- Illustration of a type IVb node.

It is noted that the type IVb node is different from the type IVa node in that more than one node links directly with the root node of the tree.

4.2 AN EXPRESSION FOR THE COVARIANCE BETWEEN THE FEATURES CONNECTED BY A PATH IN A DEPENDENT FEATURE TREE

An expression for the covariance between the features when a path connects their representative nodes in a dependent feature tree is developed in this section. For example, features x_{11} and x_{16} are connected through features x_{10}, x_9, x_8, x_{13}, and x_{15} in the dependent feature tree of figure 4-1. For the following analysis, consider the dependent feature tree shown in figure 4-4.

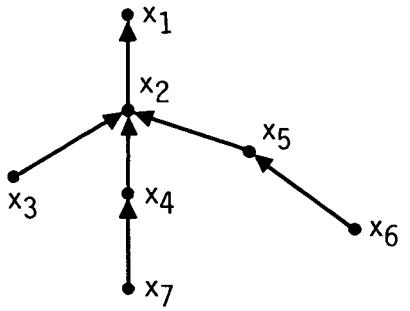

Figure 4-4.- An example dependent feature tree.

The probability density represented by the dependent feature tree of figure 4-4 can be written as follows.

$$p(X) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_2)p(x_5|x_2)p(x_6|x_5)p(x_7|x_4) \qquad (4\text{-}1)$$

In the following, an expression for the covariance between features x_6 and x_7 of figure 4-4 is derived.

$$\begin{aligned} E[(x_6 - u_6)(x_7 - u_7)] &= \int (x_6 - u_6)(x_7 - u_7)p(X)dX \\ &= \int (x_6 - u_6)p(x_2)p(x_5|x_2)p(x_6|x_5)dx_2\, dx_5\, dx_6 \\ &\quad \cdot \int p(x_4|x_2)dx_4 \int (x_7 - u_7)p(x_7|x_4)dx_7 \end{aligned} \qquad (4\text{-}2)$$

From equation (3-15), the following equations are obtained.

$$\int (x_7 - u_7)p(x_7|x_4)dx_7 = \rho_{74}\sqrt{\frac{\sigma_7}{\sigma_4}}\,(x_4 - u_4) \qquad (4\text{-}3)$$

$$\int (x_4 - u_4)p(x_4|x_2)dx_4 = \rho_{42}\sqrt{\frac{\sigma_4}{\sigma_2}}\,(x_2 - u_2) \qquad (4\text{-}4)$$

Using equations (4-3) and (4-4) in equation (4-2) yields

$$E[(x_6 - u_6)(x_7 - u_7)] = \rho_{74}\sqrt{\frac{\sigma_7}{\sigma_4}}\,\rho_{42}\sqrt{\frac{\sigma_4}{\sigma_2}}\int (x_2 - u_2)p(x_2)dx_2 \int p(x_5|x_2)dx_5 \int (x_6 - u_6)p(x_6|x_5)dx_6 \qquad (4\text{-}5)$$

Similar to equations (4-3) and (4-4), which were developed from equation (3-15), the following are obtained.

$$\left.\begin{aligned} \int (x_6 - u_6)p(x_6|x_5)dx_6 &= \rho_{65}\sqrt{\frac{\sigma_6}{\sigma_5}}\,(x_5 - u_5) \\ \int (x_5 - u_5)p(x_5|x_2)dx_5 &= \rho_{52}\sqrt{\frac{\sigma_5}{\sigma_2}}\,(x_2 - u_2) \\ \int (x_2 - u_2)(x_2 - u_2)p(x_2)dx_2 &= \sigma_2 \end{aligned}\right\} \qquad (4\text{-}6)$$

From equations (4-5) and (4-6), the covariance between features x_6 and x_7 can be obtained as follows.

$$E[(x_6 - u_6)(x_7 - u_7)] = \frac{\sigma_{74}}{\sigma_4} \cdot \frac{\sigma_{42}}{\sigma_2} \cdot \frac{\sigma_{65}}{\sigma_5} \cdot \sigma_{52} \qquad (4\text{-}7)$$

For a general case, the following theorems can easily be established.

Theorem 1: Suppose the features x_1 and x_{r+s} in a dependent feature tree are connected by a path as shown in figure 4-5. Then, the covariance between features x_1 and x_{r+s} is given by equation (4-8).

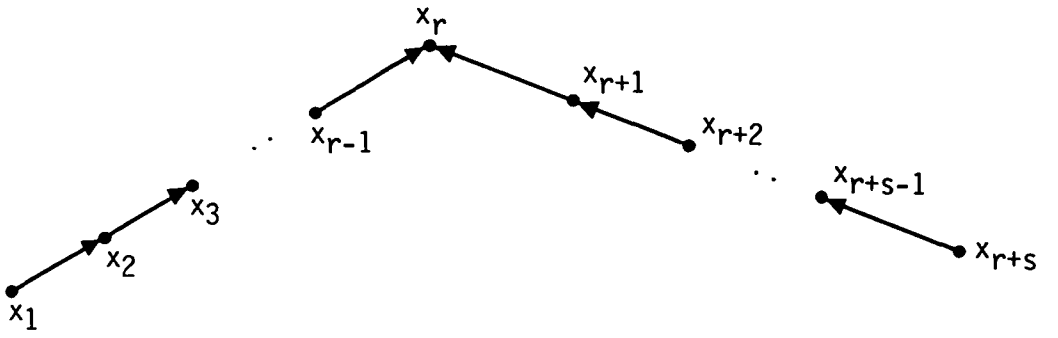

Figure 4-5.- A path between features x_1 and x_{r+s} in a dependent feature tree.

$$E[(x_1 - u_1)(x_{r+s} - u_{r+s})] = \frac{\sigma_{12}}{\sigma_2} \cdot \frac{\sigma_{23}}{\sigma_3} \cdots \frac{\sigma_{r-1,r}}{\sigma_r} \cdot \frac{\sigma_{r+s,r+s-1}}{\sigma_{r+s-1}}$$

$$\cdot \frac{\sigma_{r+s-1,r+s-2}}{\sigma_{r+s-2}} \cdots \frac{\sigma_{r+2,r+1}}{\sigma_{r+1}} \cdot \sigma_{r+1,r} \qquad (4\text{-}8)$$

Theorem 2: Suppose the features x_1 and x_r in a dependent feature tree are connected by a path as shown in figure 4-6. Then, the covariance between features x_1 and x_r is given by equation (4-9).

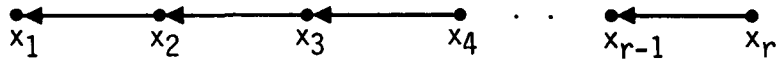

Figure 4-6.- A path between features x_1 and x_r in a dependent feature tree.

$$E[(x_1 - u_1)(x_r - u_r)] = \frac{\sigma_{12}}{\sigma_2} \cdot \frac{\sigma_{23}}{\sigma_3} \cdot \frac{\sigma_{34}}{\sigma_4} \cdots \frac{\sigma_{r-2,r-1}}{\sigma_{r-1}} \cdot \sigma_{r-1,r} \qquad (4\text{-}9)$$

5. MAXIMUM LIKELIHOOD EQUATIONS FOR THE PARAMETERS OF THE DENSITY FUNCTIONS

In this section, maximum likelihood equations are developed for estimating the parameters of the cluster conditional densities when approximated by the first-order dependent feature trees. In practice, such as in the classification of remotely sensed multispectral scanner imagery data, considerable interest has been shown in applying maximum likelihood clustering for the decomposition of the mixture density of the data into its normal component densities. The mixture density p(X) can be written as

$$p(X) = \sum_{i=1}^{m} P(\omega_i) p(X|\omega_i) \tag{5-1}$$

where m is the number of clusters, and $P(\omega_i)$ and $p(X|\omega_i)$ are the a priori probabilities of the modes and mode conditional densities, respectively. If the cluster conditional densities are Gaussian [i.e., $p(X|\omega_i) \sim N(U_i, \Sigma_i)$], by using a given set of N independent observations from the mixture density, from equation (2-6), the maximum likelihood equations for the estimates of the parameters of the mixture density can easily be shown to be the following (ref. 6).

$$\left.\begin{aligned} P(\omega_i) &= \frac{1}{N} \sum_{k=1}^{N} p(\omega_i|X_k) \\ U_i &= \frac{\sum_{k=1}^{N} X_k p(\omega_i|X_k)}{\sum_{k=1}^{N} p(\omega_i|X_k)} \\ \Sigma_i &= \frac{\sum_{k=1}^{N} (X_k - U_i)(X_k - U_i)^T p(\omega_i|X_k)}{\sum_{k=1}^{N} p(\omega_i|X_k)} \end{aligned}\right\} \tag{5-2}$$

In maximum likelihood clustering, equation (5-2) is used for updating the parameters of the densities, and this computation is coupled with a split and merge sequence. The updating is usually stopped after a few iterations because of the large amount of computation in clustering data such as imagery data.

For practical problems, the number of parameters to be estimated is large. Using equation (5-2), the number of parameters to be estimated for each mode is $\frac{n(n+3)}{2}$, where n is the dimensionality of the patterns. It is known that, with a fixed sample size, the accuracy of estimation usually decreases as the number of parameters to be estimated increases (ref. 12).

In this paper, the cluster conditional densities are approximated with first-order dependent feature trees to reduce the number of parameters to be estimated. In the product approximation for the densities discussed in the previous sections, it is noted that each feature is conditioned upon, at most, one of the other features. The number of parameters to be estimated for each mode is obtained as follows: the means n, the variances n, and the covariances (n - 1), or a total of (3n - 1), where n is the dimensionality of the patterns. When the product approximation is used for the probability densities, with an increase in the dimensionality of the patterns, the reduction in the number of parameters for each mode is as shown in figure 5-1.

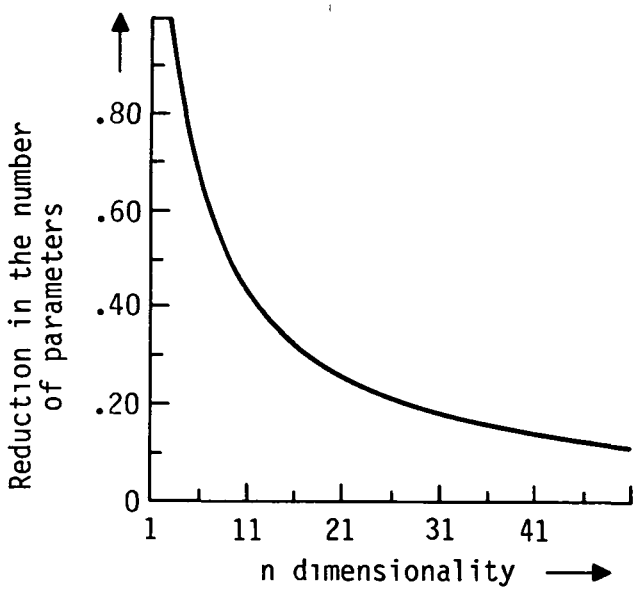

Figure 5-1.- Reduction in the number of parameters with the dimensionality.

In the following, maximum likelihood equations are developed for estimating the parameters of the cluster conditional densities when approximated with first-order dependent feature trees. It is assumed that the structure of the dependent feature tree is determined using the techniques discussed in section 3. The different types of nodes described in section 4 are considered separately.

5.1 MAXIMUM LIKELIHOOD EQUATIONS FOR THE A PRIORI PROBABILITIES, MEANS, AND VARIANCES

In this section, maximum likelihood equations similar to equation (5-2) are obtained for the a priori probabilities of the clusters and for the means and variances of features in each cluster when the cluster conditional densities are approximated with dependent feature trees. The different types of nodes discussed in section 4 are treated separately. It is assumed that a set $\mathcal{X} = \{X_1, \cdots, X_N\}$ of N unlabeled patterns, each of dimension n, drawn independently from the mixture density p(X) is given. When the cluster conditional densities are approximated by first-order dependent feature trees, the density of the i^{th} cluster can be written as

$$p(X|\omega_i) = \prod_{\ell=1}^{n} \left\{p_i[x_\ell|x_{j(\ell)}]\right\} \tag{5-3}$$

The maximum likelihood equations for the a priori probabilities of the clusters can easily be shown to be the following.

$$P(\omega_i) = \frac{1}{N} \sum_{k=1}^{N} p(\omega_i|X_k) \tag{5-4}$$

If θ_i is a parameter of the i^{th} cluster, using equation (5-3) in equation (2-8) results in

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left[\sum_{\ell=1}^{n} \frac{\partial}{\partial \theta_i}\left(\log\left\{p_i[x_\ell|x_{J(\ell)}]\right\}\right)\right] \tag{5-5}$$

In the following, it is assumed that the distributions of pattern features in each cluster are Gaussian. That is,

$$p_i(x_\ell) \sim N[u_\ell(i),\sigma_\ell(i)] \tag{5-6}$$

5.1.1 MAXIMUM LIKELIHOOD EQUATIONS FOR THE PARAMETERS OF TYPE I NODES

Consider a link in a dependent feature tree containing a type I node, as shown in figure 5-2.

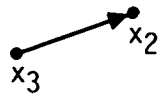

Figure 5-2.- A link in a dependent feature tree with a type I node.

Since each feature is conditioned upon, at most, one of the other features, equation (5-5) becomes

$$\frac{\partial \ell}{\partial \phi_1} = \sum_{k=1}^{N} p(\omega_1 | X_k, \theta) \frac{\partial}{\partial \phi_1} \left\{ \log[p_1(x_2, x_3)] \right\}$$

$$\text{for } \phi_1 = u_3(1)$$

$$\text{and } \phi_1 = \sigma_3(1) \qquad (5\text{-}7)$$

When $\phi_1 = u_3(1)$, from equation (5-7), the following is obtained.

$$u_3(1) = \frac{\sum_{k=1}^{N} p(\omega_i | X_k, \theta) \left\{ x_3^k - \frac{\sigma_{23}(i)}{\sigma_2(i)} \left[x_2^k - u_2(i) \right] \right\}}{\sum_{k=1}^{N} p(\omega_1 | X_k, \theta)} \qquad (5\text{-}8)$$

In equation (5-7), letting $\phi_1 = \sigma_3(1)$ and $\phi_i = \sigma_{23}(1)$ and eliminating $\left[x_3^k - u_3(1) \right]^2$ from the resulting equations yields, after simplification, an expression for the covariance between type I and type II nodes. That is,

$$\sigma_{23}(1) = \sigma_2(1) \cdot \frac{\sum_{k=1}^{N} p(\omega_1 | X_k, \theta) \left[x_2^k - u_2(i) \right] \left[x_3^k - u_3(i) \right]}{\sum_{k=1}^{N} p(\omega_1 | X_k, \theta) \left[x_2^k - u_2(i) \right]^2} \qquad (5\text{-}9)$$

Letting $\phi_1 = \sigma_3(i)$ in equation (5-7) and using equation (5-9) yields the following.

$$\sigma_3(1) = \frac{\sum_{k=1}^{N} p(\omega_1 | X_k, \theta) \left(\left[x_3^k - u_3(1) \right]^2 + \frac{\sigma_{23}(1)}{\sigma_2(i)} \left\{ \sigma_{23}(1) - \left[x_2^k - u_2(1) \right] \left[x_3^k - u_3(i) \right] \right\} \right)}{\sum_{k=1}^{N} p(\omega_1 | X_k, \theta)} \qquad (5\text{-}10)$$

5.1.2 MAXIMUM LIKELIHOOD EQUATIONS FOR THE PARAMETERS OF TYPE II NODES

A typical type II node, as defined in section 4, in a general dependent feature tree is shown in figure 5-3.

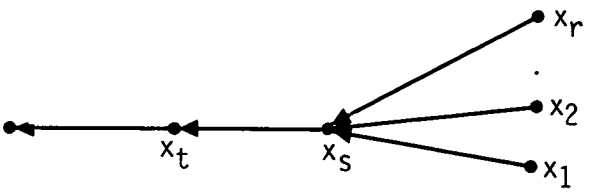

Figure 5-3.- A typical type II node in a general dependent feature tree.

In figure 5-3, node x_s is a type II node, and nodes x_1, x_2, $\cdots$, x_r are type I nodes. The terms in the product approximation of the probability density function of cluster i containing feature x_s are

$$p(X|\omega_1) = \cdots\ p(x_s|x_t)p(x_1|x_s) \cdots\ p(x_r|x_s) \cdots \tag{5-11}$$

If θ_1 is a parameter of the 1^{th} cluster, using equation (5-11) in equation (2-8) yields

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{k=1}^{N} p(\omega_1|X_k,\theta)\ \frac{\partial}{\partial \theta_i}\left[\sum_{\ell=1}^{r} p_i(x_\ell|x_s) + p_1(x_s|x_t)\right] = 0 \tag{5-12}$$

From equation (5-12), letting $\theta_1 = u_s(i)$ results in the following maximum likelihood equation for the mean of feature x_s of cluster i.

$$u_s(1) = \frac{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)x_s^k + \dfrac{1}{\left[\dfrac{r}{\sigma_s(i)} - \dfrac{\sigma_t(i)}{\Delta_{st}(i)} - \sum_{\ell=1}^{r}\dfrac{\sigma_\ell(i)}{\Delta_{s\ell}(1)}\right]}\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left\{\dfrac{\sigma_{st}(1)}{\Delta_{st}(i)}\left[x_t^k - u_t(1)\right] + \sum_{\ell=1}^{r}\dfrac{\sigma_{s\ell}(i)}{\Delta_{s\ell}(i)}\left[x_\ell^k - u_\ell(1)\right]\right\}}{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)} \tag{5-13}$$

where

$$\Delta_{s\ell}(1) = \sigma_s(i)\sigma_\ell(i) - \sigma_{s\ell}(i) \tag{5-14}$$

In equation (5-12), letting $\theta_i = \sigma_s(i), \sigma_{s1}(1), \cdots, \sigma_{sr}(i)$, and $\sigma_{st}(i)$ yields the following after simplification.

$$\sigma_s(1) = \frac{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left(r\left[x_s^k - u_s(1)\right]^2 + \sigma_s^2(1)\left\{\sum_{\ell=1}^{r}\frac{\left[x_\ell^k - u_\ell(1)\right]^2}{\Delta_{s\ell}(1)}\right\} + \frac{\sigma_s^2(i)}{\Delta_{st}(1)}\left[x_t^k - u_t(1)\right]^2\right)}{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left(r + \left[x_s^k - u_s(1)\right]^2\left[\sum_{\ell=1}^{r}\frac{\sigma_\ell(1)}{\Delta_{s\ell}(1)}\right] + \frac{\sigma_s(1)\sigma_t(1)}{\sigma_{st}(1)\Delta_{st}(1)}\left\{\left[x_s^k - u_s(1)\right]\left[x_t^k - u_t(1)\right]\right\}\right)} \tag{5-15}$$

5.1.3 MAXIMUM LIKELIHOOD EQUATIONS FOR THE PARAMETERS OF TYPE III NODES

A typical type III node in a general dependent feature tree is shown in figure 5-4.

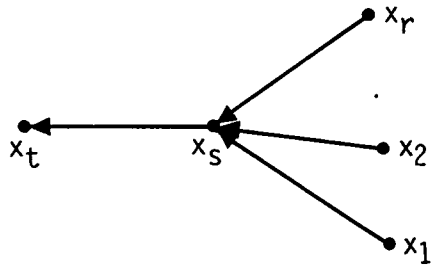

Figure 5-4.- A typical type III node in a general dependent feature tree.

In figure 5-4, x_s is a type III node; and nodes $x_1, x_2, \cdots, x_r$ and x_t may be type III nodes or other types of nodes. Proceeding as in section 5.1.2, the maximum likelihood equations for the variance and mean of feature x_s of cluster i can be shown to be the following (see appendix A).

$$\sigma_s(1) = \frac{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left(r\left[x_s^k - u_s(1)\right]^2 + \frac{\sigma_s^2(1)}{\Delta_{st}(1)}\left[x_t^k - u_t(1)\right]^2 + \sigma_s^2(1)\cdot\left\{\sum_{\ell=1}^{r}\frac{\left[x_\ell^k - u_\ell(1)\right]^2}{\Delta_{s\ell}(1)}\right\}\right)}{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left(r + \frac{\sigma_s(1)\sigma_t(1)}{\sigma_{st}(1)\Delta_{st}(1)}\left[x_s^k - u_s(1)\right]\left[x_t^k - u_t(1)\right] + \left\{\sum_{\ell=1}^{r}\frac{\sigma_s(1)\sigma_\ell(1)}{\sigma_{s\ell}(1)\Delta_{s\ell}(1)}\left[x_s^k - u_s(1)\right]\left[x_\ell^k - u_\ell(1)\right]\right\}\right)} \tag{5-16}$$

and

$$u_s(1) = \frac{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)x_s^k + \frac{1}{\left[\frac{r}{\sigma_s(1)} - \frac{\sigma_t(1)}{\Delta_{st}(1)} - \sum_{\ell=1}^{r}\frac{\sigma_\ell(1)}{\Delta_{s\ell}(1)}\right]} \cdot \sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left(\frac{\sigma_{st}(1)}{\Delta_{st}(1)}\left[x_t^k - u_t(1)\right] + \left\{\sum_{\ell=1}^{r}\frac{\sigma_{s\ell}(1)}{\Delta_{s\ell}(1)}\left[x_\ell^k - u_\ell(1)\right]\right\}\right)}{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)} \tag{5-17}$$

5.1.4 MAXIMUM LIKELIHOOD EQUATIONS FOR THE PARAMETERS OF TYPE IVA NODES

A typical type IVa node in a general dependent feature tree is shown in figure 5-5.

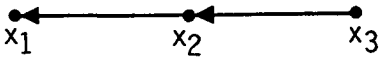

Figure 5-5.- A typical type IVa node in a general dependent feature tree.

In figure 5-5, x_1 is a root node of type IVa, and node x_2 may be of type I, type II, or type III. The maximum likelihood equations for the variance and mean of feature x_1 of cluster 1 are given in the following.

$$\sigma_1(1) = \frac{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left(\left[x_1^k - u_1(1)\right]^2 + \frac{\sigma_{21}(1)}{\sigma_2(i)}\left\{\sigma_{21}(1) - \left[x_1^k - u_1(1)\right]\left[x_2^k - u_2(1)\right]\right\}\right)}{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)} \tag{5-18}$$

and

$$u_1(i) = \frac{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left\{x_1^k - \frac{\sigma_{21}(i)}{\sigma_2(i)}\left[x_2^k - u_2(i)\right]\right\}}{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)} \tag{5-19}$$

5.1.5 MAXIMUM LIKELIHOOD EQUATIONS FOR THE PARAMETERS OF TYPE IVB NODES

A typical type IVb node in a general dependent feature tree is shown in figure 5-6.

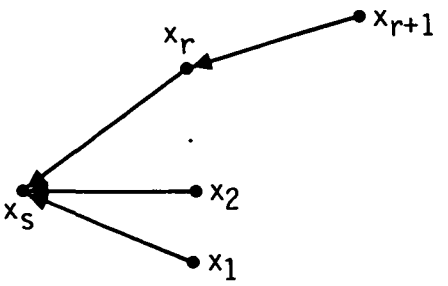

Figure 5-6.- A typical type IVb node in a general dependent feature tree.

In figure 5-6, x_s is a root node of type IVb, and nodes x_1, x_2, $\cdots$, x_r are of type I, type II, or type III. The maximum likelihood equations for the variance and mean of feature x_s of cluster i can be shown to be the following.

$$\sigma_s(i) = \frac{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left((r-1)\left[x_s^k - u_s(i)\right]^2 + \left\{\sum_{\ell=1}^{r} \frac{\sigma_s^2(i)}{\Delta_{s\ell}(i)}\left[x_\ell^k - u_\ell(i)\right]^2\right\}\right)}{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left((r-1) + \left\{\sum_{\ell=1}^{r} \frac{\sigma_s(i)\sigma_\ell(i)}{\sigma_{s\ell}(i)\Delta_{s\ell}(i)}\left[x_s^k - u_s(i)\right]\left[x_\ell^k - u_\ell(i)\right]\right\}\right)} \tag{5-20}$$

and

$$u_s(i) = \frac{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)x_s^k + \frac{1}{\left[\frac{(r-1)}{\sigma_s(i)} - \sum_{\ell=1}^{r}\frac{\sigma_\ell(i)}{\Delta_{s\ell}(i)}\right]}\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left\{\sum_{\ell=1}^{r}\frac{\sigma_{s\ell}(i)}{\Delta_{s\ell}(i)}\left[x_\ell^k - u_\ell(i)\right]\right\}}{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)} \tag{5-21}$$

5.2 MAXIMUM LIKELIHOOD EQUATIONS FOR THE COVARIANCES BETWEEN FEATURES

In this section, maximum likelihood equations are developed for the covariances between the features when the probability density functions of the clusters are approximated by first-order dependent feature trees.

5.2.1 MAXIMUM LIKELIHOOD EQUATIONS FOR THE COVARIANCE BETWEEN TYPE I AND TYPE II NODES

In this section, maximum likelihood equations for the covariance between type I and type II features are derived. A typical link connecting type I and type II nodes is shown in figure 5-7.

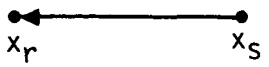

Figure 5-7.- A typical link connecting type I and type II nodes.

In figure 5-7, node x_s is of type I, and node x_r is of type II. The maximum likelihood equation for the covariance between features x_r and x_s of cluster 1 is given in the following.

$$\sigma_{rs}(1) = \sigma_r(i) \frac{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left[x_r^k - u_r(i)\right]\left[x_s^k - u_s(i)\right]}{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left[x_r^k - u_r(i)\right]^2} \tag{5-22}$$

5.2.2 MAXIMUM LIKELIHOOD EQUATIONS FOR THE COVARIANCE BETWEEN TYPE IVA AND TYPE II OR TYPE III NODES

A typical link connecting type IVa and type II or type III nodes in a general dependent feature tree is shown in figure 5-8.

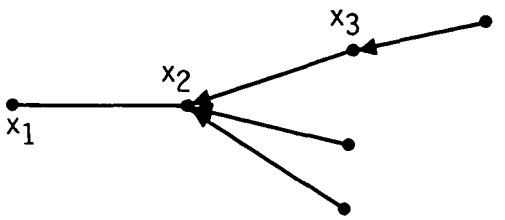

Figure 5-8.- A typical link connecting type IVa and type II or type III nodes in a general dependence tree.

In figure 5-8, node x_1 is of type IVa, and node x_2 may be of type II or type III. The maximum likelihood equation for the covariance between features x_1 and x_2 of cluster i is as follows.

$$\sigma_{12}(1) = \sigma_2(1) \cdot \frac{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left[x_1^k - u_1(i)\right]\left[x_2^k - u_2(i)\right]}{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left[x_2^k - u_2(i)\right]^2} \tag{5-23}$$

5.2.3 MAXIMUM LIKELIHOOD EQUATIONS FOR THE COVARIANCE BETWEEN TYPES OF NODES OTHER THAN THOSE CONSIDERED IN SECTIONS 5.2.1 AND 5.2.2

Let there be a link between nodes x_2 and x_3 in a general dependent feature tree whose types are other than those considered in sections 5.2.1 and 5.2.2. The maximum likelihood equation for the covariance between features x_2 and x_3 of cluster i is given in the following.

$$\sigma_{23}(i) = \frac{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left\{\sigma_2(i)\sigma_3(i) + \sigma_{23}(i)\Delta_{23}(i) + \sigma_{23}^2(i)\left[x_2^k - u_2(i)\right]\left[x_3^k - u_3(i)\right]\right\}}{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left\{\sigma_3(i)\left[x_2^k - u_2(i)\right]^2 + \sigma_2(i)\left[x_3^k - u_3(i)\right]^2\right\}} \qquad (5\text{-}24)$$

6. EXPERIMENTAL RESULTS

In this section, some results from processing remotely sensed Landsat multispectral scanner imagery data are presented. The images are of a 5- by 6-nautical-mile area called a segment. The image is divided into a rectangular array of pixels, 117 rows by 196 columns. The image is overlaid with a rectangular grid of 209 grid intersections. Two classes are considered: class 1 is wheat, and class 2 is "other." The true (ground truth) labels for the pixels at the grid intersections are acquired. The locations of the segments and the individual acquisitions used for each of the segments are listed in table 6-1. The a priori probabilities of the classes are estimated as sample estimates. Equations (3-6) and (3-18) are used to compute the weighted mutual information between the features, assuming in each class that the features are Gaussian distributed. Kruskal's algorithm (ref. 16) is used to construct optimal dependent feature trees by minimizing $I(p,p_t)$ of equation (3-5). The optimal dependent feature trees of segments 1648 and 1739 are shown in figures 6-1 and 6-2, respectively.

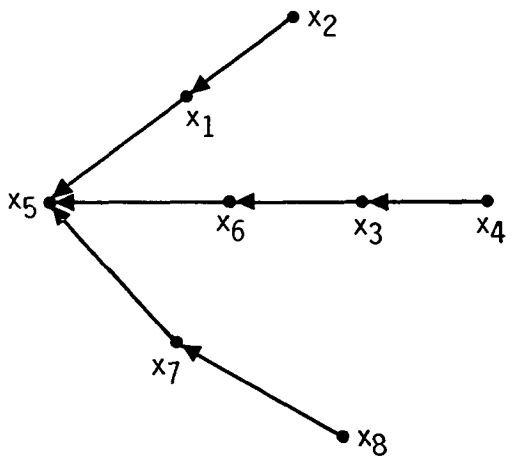

Figure 6-1.- Optimal dependent feature tree of segment 1648.

Generally, it is known that, for each acquisition, a strong dependency exists between channels 1 and 2 and between channels 3 and 4. From figures 6-1 and 6-2, it is seen that these dependencies appear in the optimal dependent feature trees.

TABLE 6-1.- CONFUSION MATRICES AND CLASSIFICATION ACCURACIES OF BAYES CLASSIFIER

Segment	Location (county, state)	Acquisition dates	Full covariance matrix		Independent features		Optimal dependent feature tree		Arbitrary dependent feature tree	
			Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
1520	Big Stone, Minnesota	77174 77156 77120	[a] $\begin{bmatrix}28 & 3\\ 3 & 70\end{bmatrix}$ [b] 0.9423	$\begin{bmatrix}17 & 14\\ 9 & 65\end{bmatrix}$ 0.7809	$\begin{bmatrix}26 & 5\\ 13 & 60\end{bmatrix}$ 0.8269	$\begin{bmatrix}23 & 8\\ 8 & 66\end{bmatrix}$ 0.8476	$\begin{bmatrix}25 & 6\\ 6 & 67\end{bmatrix}$ 0.8846	$\begin{bmatrix}23 & 8\\ 6 & 68\end{bmatrix}$ 0.8667	$\begin{bmatrix}26 & 5\\ 19 & 54\end{bmatrix}$ 0.7692	$\begin{bmatrix}24 & 7\\ 15 & 59\end{bmatrix}$ 0.7904
1604	Renville, North Dakota	77143 77125	$\begin{bmatrix}49 & 9\\ 17 & 29\end{bmatrix}$ 0.75	$\begin{bmatrix}38 & 20\\ 24 & 23\end{bmatrix}$ 0.5809	$\begin{bmatrix}45 & 13\\ 20 & 26\end{bmatrix}$ 0.6827	$\begin{bmatrix}36 & 22\\ 20 & 27\end{bmatrix}$ 0.6000	$\begin{bmatrix}47 & 11\\ 20 & 26\end{bmatrix}$ 0.7019	$\begin{bmatrix}41 & 17\\ 15 & 32\end{bmatrix}$ 0.6952	$\begin{bmatrix}45 & 13\\ 21 & 25\end{bmatrix}$ 0.6731	$\begin{bmatrix}37 & 21\\ 25 & 22\end{bmatrix}$ 0.5619
1648	Bowman, North Dakota	77179 77125	$\begin{bmatrix}30 & 11\\ 18 & 45\end{bmatrix}$ 0.7211	$\begin{bmatrix}24 & 18\\ 18 & 45\end{bmatrix}$ 0.6571	$\begin{bmatrix}33 & 8\\ 24 & 39\end{bmatrix}$ 0.6923	$\begin{bmatrix}28 & 14\\ 27 & 36\end{bmatrix}$ 0.6095	$\begin{bmatrix}36 & 5\\ 21 & 42\end{bmatrix}$ 0.7500	$\begin{bmatrix}30 & 12\\ 22 & 41\end{bmatrix}$ 0.6762	$\begin{bmatrix}28 & 13\\ 23 & 40\end{bmatrix}$ 0.6538	$\begin{bmatrix}26 & 16\\ 28 & 35\end{bmatrix}$ 0.5809
1739	Teton, Montana	77222 77168 77132 76263	$\begin{bmatrix}34 & 23\\ 4 & 63\end{bmatrix}$ 0.9327	$\begin{bmatrix}19 & 19\\ 17 & 50\end{bmatrix}$ 0.6571	$\begin{bmatrix}20 & 17\\ 12 & 55\end{bmatrix}$ 0.7212	$\begin{bmatrix}18 & 20\\ 14 & 53\end{bmatrix}$ 0.6762	$\begin{bmatrix}25 & 12\\ 12 & 55\end{bmatrix}$ 0.7692	$\begin{bmatrix}23 & 15\\ 17 & 50\end{bmatrix}$ 0.6952	$\begin{bmatrix}18 & 19\\ 13 & 54\end{bmatrix}$ 0.6923	$\begin{bmatrix}16 & 22\\ 13 & 54\end{bmatrix}$ 0.6667
1853	Ness, Kansas	77193 77067 76253	$\begin{bmatrix}27 & 4\\ 4 & 69\end{bmatrix}$ 0.9231	$\begin{bmatrix}17 & 14\\ 8 & 66\end{bmatrix}$ 0.7905	$\begin{bmatrix}24 & 7\\ 10 & 63\end{bmatrix}$ 0.8365	$\begin{bmatrix}20 & 11\\ 11 & 63\end{bmatrix}$ 0.7905	$\begin{bmatrix}24 & 7\\ 6 & 67\end{bmatrix}$ 0.875	$\begin{bmatrix}20 & 11\\ 6 & 68\end{bmatrix}$ 0.8381	$\begin{bmatrix}24 & 7\\ 10 & 63\end{bmatrix}$ 0.8365	$\begin{bmatrix}20 & 11\\ 11 & 63\end{bmatrix}$ 0.7905
Mean classification accuracy			0.8554	0.6933	0.7519	0.7047	0.7961	0.7543	0.7250	0.6781

[a] Confusion matrix.

[b] Probability of correct classification.

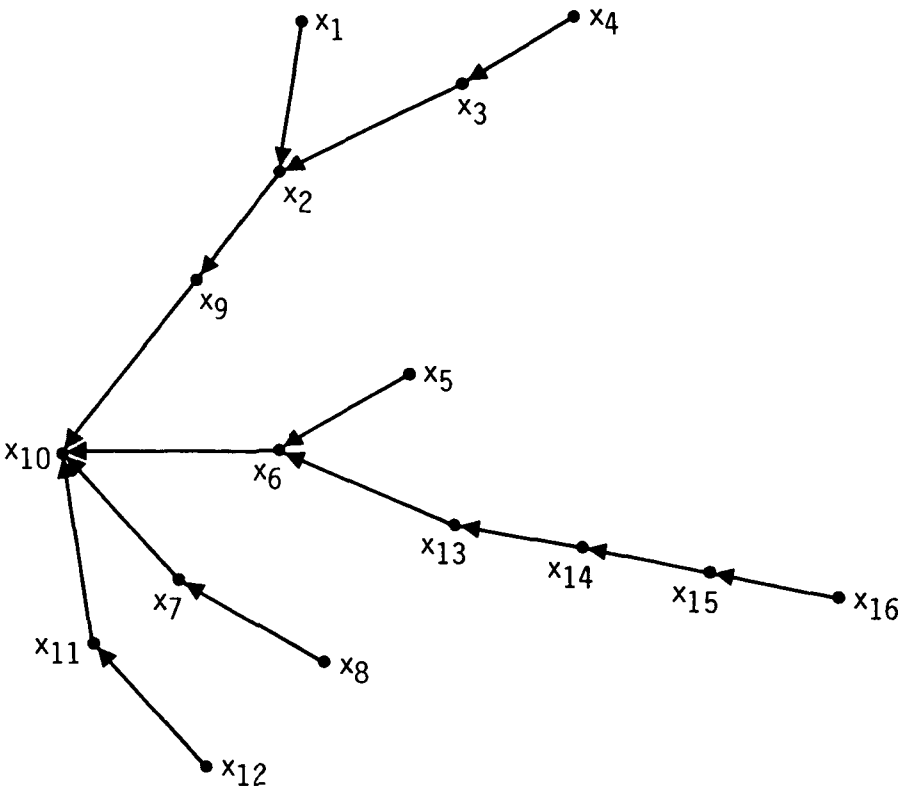

Figure 6-2.- Optimal dependent feature tree of segment 1739.

To investigate the effectiveness of the optimal dependent feature trees in classification, an experiment is performed to compare the classification accuracies of the Bayes classifier (1) when the densities are approximated with optimal dependent feature trees, (2) when no approximation is used for the densities (full covariance matrix), (3) assuming the features are independent, and (4) when the densities are approximated with arbitrary dependent feature trees. Spectral vectors of 104 labeled pixels are used as the training pattern set, and the spectral vectors of the remaining 105 labeled pixels are used as the test pattern set. The structure of the arbitrary dependent feature tree used in this experiment is shown in figure 6-3.

The computed confusion matrices and the classification accuracies on the training and test sets for each of the segments processed are listed in table 6-1. From table 6-1, it is seen that, in general, better classification accuracies

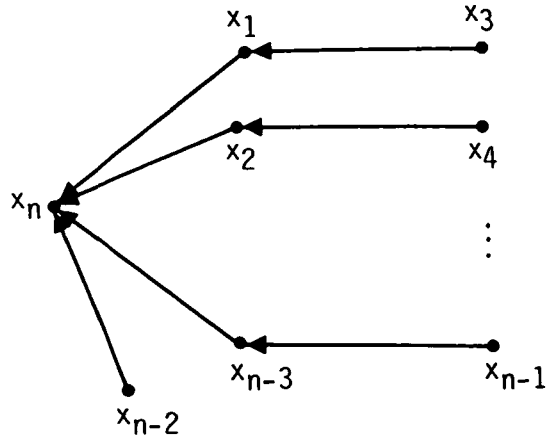

Figure 6-3.- Arbitrary dependent feature tree used in the experiment.

are obtained on the training set when the full covariance matrix is used without approximating the densities. Improved classification accuracies are obtained on the test set when the densities are approximated with optimal dependent feature trees. This might be due to the fact that a large number of parameters are estimated when the full covariance matrices are used.

One of the important objectives in the classification of remotely sensed agricultural imagery data is to estimate the proportion of the class of interest in the image. The ratio of the variance of the estimated proportion using machine classification to the variance of the estimated proportion using simple random sampling is called variance reduction factor R (ref. 1). The quantity R can be viewed as an indication of how much the machine classification improves the proportion estimation. The computed variance reduction factors for each of the segments processed are listed in table 6-2. From table 6-2, it is seen that the variance reduction factor consistently improves when the densities are approximated with dependent feature trees, compared to the other cases.

TABLE 6-2.- COMPARISON OF VARIANCE REDUCTION FACTORS

Segment	Location (county, state)	Acquisition dates	Full covariance matrix	Independent features	Optimal dependent feature tree	Arbitrary dependent feature tree
1520	Big Stone, Minnesota	77174 77156 77120	0.8194	0.6405	0.5532	0.7540
1604	Renville, North Dakota	77143 77125	0.9786	0.9637	0.8475	0.9889
1648	Bowman, North Dakota	77179 77125	0.9325	0.9509	0.8896	0.9679
1739	Teton, Montana	77222 77168 77132 76263	0.9432	0.9242	0.8724	0.9450
1853	Ness, Kansas	77193 77067 76253	0.8034	0.7701	0.6576	0.7701

7. CONCLUDING REMARKS

In the classification of imagery data, such as in the machine processing of remotely sensed multispectral scanner data, unsupervised classification techniques have been found to be effective. The application of clustering techniques for the analysis of imagery data essentially involves two steps: (1) clustering the data or partitioning the image into its inherent modes and (2) giving the probabilistic class labels to the resulting clusters. In practice, it is observed that fields are relatively easy to label when compared to pixels.

Several researchers have investigated methods for locating fields in the imagery data. Recently, considerable interest has been shown in developing techniques for probabilistically labeling the clusters using information from a given set of labeled patterns and, also, from a given set of labeled fields.

In decomposing the mixture density of the data into its normal component densities, the parameters of the component densities and the a priori probabilities of the modes are iteratively computed using maximum likelihood equations coupled with a split and merge sequence. The updating of the parameters is usually stopped after a few iterations; and for practical data, a large number of parameters must be estimated. For a fixed sample size, the accuracy of estimation usually decreases as the number of parameters to be estimated increases.

To overcome the above shortcomings, it is proposed in this paper that the densities be approximated with first-order dependent feature trees. The dependent feature trees can be constructed using criteria based on information measure and, also, based on class separability measure. Expressions are derived for the criteria when the distributions of the features are Gaussian. Expressions also are derived for the covariances between features not connected by a single link in the dependent feature tree.

Different types of nodes are defined in a general dependent feature tree. Maximum likelihood equations are derived for the parameters of the mixture

density of the data by approximating the cluster conditional densities with first-order dependent feature trees. The field structure of the data is also taken into account in the decomposition of the mixture density of the data into its normal component densities. Furthermore, experimental results from the processing of remotely sensed multispectral scanner imagery data are presented.

8. REFERENCES

1. Heydorn, R. P.; Trichel, M. C.; and Erickson, J. D.: Methods for Segment Wheat Area Estimation. Proc. LACIE Symp., NASA/JSC (Houston). JSC-16015, vol. II, July 1979, pp. 621-632.

2. Bryant, J.: On the Clustering of Multidimensional Pictorial Data. Pattern Recognition, vol. 11, no. 2, 1979, pp. 115-125.

3. Kauth, R. J.; Pentland, A. P.; and Thomas, G. S.: Blob: An Unsupervised Clustering Approach to Spatial Preprocessing of MSS Imagery. Proc. 11th Int. Symp. on Remote Sensing of Environment. Environmental Research Institute of Michigan (Ann Arbor), 1977, pp. 1309-1317.

4. Kettig, R. L.; and Landgrebe, D. A.: Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects. IEEE Trans. Geoscience Electronics, vol. GE-14, Jan. 1976, pp. 19-26.

5. Peters, B. C., Jr.: On the Consistency of the Maximum Likelihood Estimate of Normal Mixture Parameters for a Sample With Field Structure. Report no. 74, Dept. of Math., Univ. of Houston, Sept. 1979.

6. Duda, R. O.; and Hart, P. E.: Pattern Classification and Scene Analysis. John Wiley and Sons (New York), 1973.

7. Hasselblad, V.: Estimation of Parameters for a Mixture of Normal Distributions. Technometrics, vol. 8, 1966, pp. 431-446.

8. Day, N. E.: Estimating the Components of a Mixture of Normal Distributions. Biometrika, vol. 56, 1969, pp. 463-474.

9. Lennington, R. K.; and Rassbach, M. E.: Mathematical Description and Program Documentation for CLASSY: An Adaptive Maximum Likelihood Clustering Method. Lockheed Electronics Company, Inc., Tech. Memo LEC-12177, JSC-14621, NASA/JSC (Houston), Apr. 1979.

10. Chittineni, C. B.: Some Approaches to Optimal Cluster Labeling of Aerospace Imagery. Lockheed Engineering and Management Services Company, Inc., Tech. Report LEMSCO-14597, JSC-16355, NASA/JSC (Houston), May 1980.

11. Chittineni, C. B.: Probabilistic Cluster Labeling of Imagery Data. Lockheed Engineering and Management Services Company, Inc., Tech. Report LEMSCO-15358, JSC-16384, NASA/JSC (Houston), Sept. 1980.

12. Hughes, G. F.: On the Mean Accuracy of Statistical Pattern Recognizers. IEEE Trans. Information Theory, vol. IT-14, Jan. 1968, pp. 55-63.

13. Chow, C. K.; and Liu, C. N.: Approximating Discrete Probability Distributions With Dependence Trees. IEEE Trans. Information Theory, vol. IT-14, May 1968, pp. 462-467.

14. Lewis, P. M.: Approximating Probability Distributions to Reduce Storage Requirements. Information and Control, vol. 2, Sept. 1959, pp. 214-225.

15. Chittineni, C. B.: Efficient Feature Subset Selection With Probabilistic Distance Criteria. Lockheed Electronics Company, Inc., Tech. Memo LEC-13355, JSC-14870, NASA/JSC (Houston), May 1979.

16. Kruskal, J. B.: On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. Proc. American Math. Soc., vol. 7, 1956, pp. 48-50.

17. Tou, J. T.; and Gonzalez, R. C.: Pattern Recognition Principles. Addison-Wesley Publishing Co., Inc. (Mass.), 1974.

APPENDIX A

DERIVATIONS OF MAXIMUM LIKELIHOOD EQUATIONS FOR THE PARAMETERS OF A TYPE III FEATURE

APPENDIX A

DERIVATIONS OF MAXIMUM LIKELIHOOD EQUATIONS FOR THE PARAMETERS OF A TYPE III FEATURE

In this appendix, maximum likelihood equations for the parameters of a typical type III feature are derived. A typical type III node in a general dependent feature tree is illustrated in figure A-1.

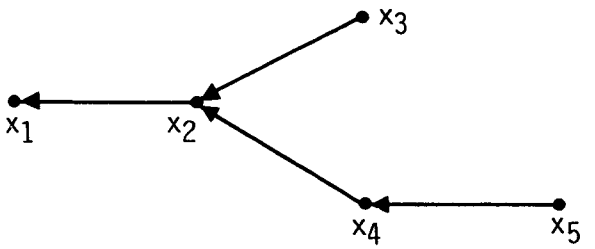

Figure A-1.- Illustration of a typical type III node in a general dependent feature tree.

In figure A-1, x_2 is a type III feature. The following is obtained from equation (5-5) by keeping only the terms that involve feature x_2 in the product approximation of the density of the i^{th} cluster.

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{k=1}^{N} p(\omega_1 | X_k, \theta) \frac{\theta}{\partial \theta_1} \left\{ \log[p_i(x_4|x_2)] + \log[p_i(x_3|x_2)] + \log[p_1(x_2|x_1)] \right\}$$

$$= \sum_{k=1}^{N} p(\omega_1 | X_k, \theta) \frac{\partial}{\partial \theta_i} \left\{ \log[p_i(x_2,x_4)] + \log[p_i(x_2,x_3)] + \log[p_1(x_1,x_2)] - 2 \log[p_i(x_2)] - \log[p_i(x_1)] \right\} \quad \text{(A-1)}$$

It is assumed that the features of the i^{th} cluster are Gaussian distributed. That is,

$$\log[p_1(x_r,x_s)] = -\log(2\pi) - \frac{1}{2}\log[\Delta_{rs}(1)] - \frac{1}{2\Delta_{rs}(i)}\left\{\sigma_s(i)[x_r - u_r(i)]^2 - 2\sigma_{rs}(i)[x_r - u_r(i)][x_s - u_s(i)] + \sigma_r(1)[x_s - u_s(i)]^2\right\} \tag{A-2}$$

where

$$\Delta_{rs}(1) = \sigma_r(i)\sigma_s(i) - \sigma_{rs}^2(i) \tag{A-3}$$

Using equation (A-2) in equation (A-1) yields

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{k=1}^{N} p(\omega_1|X_k,\theta)\frac{\partial C_k}{\partial \theta_i} \tag{A-4}$$

where

$$\begin{aligned} C_k = {} & \log[\Delta_{24}(1)] + \frac{1}{\Delta_{24}(1)}\left\{\sigma_4(1)\left[x_2^k - u_2(1)\right]^2 - 2\sigma_{24}(1)\left[x_2^k - u_2(1)\right]\left[x_4^k - u_4(1)\right] + \sigma_2(1)\left[x_4^k - u_4(1)\right]^2\right\} \\ & + \log[\Delta_{23}(1)] + \frac{1}{\Delta_{23}(i)}\left\{\sigma_3(1)\left[x_2^k - u_2(1)\right]^2 - 2\sigma_{23}(1)\left[x_2^k - u_2(1)\right]\left[x_3^k - u_3(1)\right] + \sigma_2(i)\left[x_3^k - u_3(1)\right]^2\right\} \\ & + \log[\Delta_{21}(1)] + \frac{1}{\Delta_{21}(i)}\left\{\sigma_1(1)\left[x_2^k - u_2(1)\right]^2 - 2\sigma_{21}(1)\left[x_2^k - u_2(1)\right]\left[x_1^k - u_1(1)\right] + \sigma_2(1)\left[x_1^k - u_1(1)\right]^2\right\} \\ & - 2\log[\sigma_2(1)] - \frac{2}{\sigma_2(i)}\left[x_2^k - u_2(1)\right]^2 \end{aligned} \tag{A-5}$$

Letting $\theta_i = u_2(i)$, from equation (A-5), the following is obtained.

$$\begin{aligned} \frac{\partial C_k}{\partial u_2(i)} = {} & \left\{-\frac{2\sigma_4(1)}{\Delta_{24}(1)}\left[x_2^k - u_2(i)\right] + \frac{2\sigma_{24}(1)}{\Delta_{24}(i)}\left[x_4^k - u_4(i)\right]\right\} \\ & + \left\{-\frac{2\sigma_3(i)}{\Delta_{23}(i)}\left[x_2^k - u_2(1)\right] + \frac{2\sigma_{23}(i)}{\Delta_{23}(i)}\left[x_3^k - u_3(i)\right]\right\} \\ & + \left\{-\frac{2\sigma_1(i)}{\Delta_{21}(i)}\left[x_2^k - u_2(i)\right] + \frac{2\sigma_{21}(1)}{\Delta_{21}(i)}\left[x_1^k - u_1(1)\right]\right\} \\ & + \left\{\frac{4}{\sigma_2(i)}\left[x_2^k - u_2(i)\right]\right\} \end{aligned} \tag{A-6}$$

Substituting equation (A-6) in equation (A-4) and equating the results to zero yields the following maximum likelihood equation for the mean of feature x_2 of cluster i.

$$u_2(1) = \frac{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left(x_2^k + \frac{1}{\left[\frac{2}{\sigma_2(1)} - \frac{\sigma_1(1)}{\Delta_{21}(1)} - \frac{\sigma_3(1)}{\Delta_{23}(1)} - \frac{\sigma_4(1)}{\Delta_{24}(1)}\right]} \cdot \left\{\frac{\sigma_{21}(1)}{\Delta_{21}(1)}\left[x_1^k - u_1(1)\right] + \frac{\sigma_{23}(1)}{\Delta_{23}(1)}\left[x_3^k - u_3(1)\right] + \frac{\sigma_{24}(1)}{\Delta_{24}(1)}\left[x_4^k - u_4(1)\right]\right\}\right)}{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)} \tag{A-7}$$

Letting $\theta_1 = \sigma_2(1)$ in equation (A-4) and equating the result to zero yields the following after simplification.

$$\begin{aligned}\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\Bigg(&\left\{\frac{\sigma_4(1)}{\Delta_{24}(1)} - \frac{\sigma_{24}^2(1)}{\Delta_{24}^2(1)}\left[x_4^k - u_4(1)\right]^2 - \frac{\sigma_4^2(1)}{\Delta_{24}^2(1)}\left[x_2^k - u_2(1)\right]^2 + \frac{2\sigma_4(1)\sigma_{24}(1)}{\Delta_{24}^2(1)}\left[x_2^k - u_2(1)\right]\left[x_4^k - u_4(1)\right]\right\} \\ &+ \left\{\frac{\sigma_3(1)}{\Delta_{23}(1)} - \frac{\sigma_{23}^2(1)}{\Delta_{23}^2(1)}\left[x_3^k - u_3(1)\right]^2 - \frac{\sigma_3^2(1)}{\Delta_{23}^2(1)}\left[x_2^k - u_2(1)\right]^2 + \frac{2\sigma_3(1)\sigma_{23}(1)}{\Delta_{23}^2(1)}\left[x_2^k - u_2(1)\right]\left[x_3^k - u_3(1)\right]\right\} \\ &+ \left\{\frac{\sigma_1(1)}{\Delta_{21}(1)} - \frac{\sigma_{21}^2(1)}{\Delta_{21}^2(1)}\left[x_1^k - u_1(1)\right]^2 - \frac{\sigma_1^2(1)}{\Delta_{21}^2(1)}\left[x_2^k - u_2(1)\right]^2 + \frac{2\sigma_1(1)\sigma_{21}(1)}{\Delta_{21}^2(1)}\left[x_2^k - u_2(1)\right]\left[x_1^k - u_1(1)\right]\right\} \\ &- \frac{2}{\sigma_2(1)} + \frac{2}{\sigma_2^2(1)}\left[x_2^k - u_2(1)\right]^2\Bigg) = 0\end{aligned} \tag{A-8}$$

Similarly, differentiating ℓ with respect to $\sigma_{2J}(i)$ for j = 1,3,4 and equating the resulting expression to zero yields

$$\begin{aligned}\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\Bigg\{&-\frac{2\sigma_{2J}(1)}{\Delta_{2J}(1)} - \frac{2\sigma_2(1)\sigma_J(1)}{\Delta_{2J}^2(1)}\left[x_2^k - u_2(1)\right]\left[x_J^k - u_J(1)\right] + \frac{2\sigma_{2J}(1)\sigma_J(1)}{\Delta_{2J}^2(1)}\left[x_2^k - u_2(1)\right]^2 \\ &- \frac{2\sigma_{2J}^2(1)}{\Delta_{2J}^2(1)}\left[x_2^k - u_2(1)\right]\left[x_J^k - u_J(1)\right] + \frac{2\sigma_2(1)\sigma_{2J}(1)}{\Delta_{2J}^2(1)}\left[x_J^k - u_j(1)\right]^2\Bigg\} = 0 \quad ; J = 1,3,4\end{aligned} \tag{A-9}$$

Using equation (A-9) for j = 1,3,4 in equation (A-8) yields, after simplification, the following maximum likelihood equation for $\sigma_2(i)$.

$$\sigma_2(1) = \frac{\sum_{k=1}^{N} p(\omega_i|X_k,\theta)\left(2\left[x_2^k - u_2(1)\right]^2 + \left\{\frac{\sigma_2^2(1)}{\Delta_{24}(1)}\left[x_4^k - u_4(1)\right]^2 + \frac{\sigma_2^2(1)}{\Delta_{23}(1)}\left[x_3^k - u_3(1)\right]^2 + \frac{\sigma_2^2(1)}{\Delta_{21}(1)}\left[x_1^k - u_1(1)\right]^2\right\}\right)}{\sum_{k=1}^{N} p(\omega_1|X_k,\theta)\left(2 + \left\{\frac{\sigma_2(1)\sigma_4(1)}{\sigma_{24}(1)\Delta_{24}(1)}\left[x_2^k - u_2(1)\right]\left[x_4^k - u_4(1)\right] + \frac{\sigma_2(1)\sigma_3(1)}{\sigma_{23}(1)\Delta_{23}(1)}\left[x_2^k - u_2(1)\right]\left[x_3^k - u_3(1)\right] + \frac{\sigma_2(1)\sigma_1(1)}{\sigma_{21}(1)\Delta_{21}(1)}\left[x_2^k - u_2(1)\right]\left[x_1^k - u_1(1)\right]\right\}\right)} \tag{A-10}$$

Proceeding, similar to equations (A-9) and (A-10), it can easily be shown that the maximum likelihood equations for the mean $u_s(i)$ and variance $\sigma_s(i)$ of feature x_s of cluster i of figure 5-4 are of equations (5-16) and (5-15).

APPENDIX B

MAXIMUM LIKELIHOOD EQUATIONS WITH FIELD STRUCTURE

APPENDIX B

MAXIMUM LIKELIHOOD EQUATIONS WITH FIELD STRUCTURE

In practical applications of pattern recognition, such as in the classification of remotely sensed agricultural imagery data, one of the difficult problems is to obtain labels for the training patterns. The labels for the training patterns are usually provided by an analyst-interpreter by examining imagery films and using some other information such as historic information and crop calendar models. Agricultural imagery data usually have a field-like structure, and it is observed that fields are relatively easy to label when compared to pixels. Recently, considerable interest has been shown in developing techniques for locating fields in the imagery data (ref. 2-4) and in developing methods for the probabilistic labeling (refs. 10, 11) of cluster distributions using information from a given set of labeled fields. Once the fields are located by a field-finding algorithm, the problem of fitting a mixture of Gaussian density functions to the data by taking into account the field structure of the data is considered in this appendix.

It is assumed that there are f-fields in the data. Let the j^{th} field be denoted by F_j; let it contain N_j pixels; and let X_{jk}, $k = 1,2,\cdots,N_j$, be their spectral vectors. Let m be the number of clusters in the data. Let $P(\omega_i)$ and $p(X|\omega_i)$ be the a priori probability that a field belongs to cluster ω_i and cluster conditional densities, respectively. Let $\tilde{X}_j$ be the concatenated vector of spectral vectors of the pixels in the j^{th} field. That is,

$$\tilde{X}_j = \begin{bmatrix} X_{j1} \\ X_{j2} \\ \vdots \\ X_{jN_j} \end{bmatrix} \tag{B-1}$$

It is assumed that the fields are independent. Then, the joint density of f-fields is given by

$$p(\tilde{X}_1,\cdots,\tilde{X}_f) = \prod_{j=1}^{f} p(\tilde{X}_j) \tag{B-2}$$

The mixture density $p(\tilde{X}_j)$ can be written as

$$p(\tilde{X}_j) = \sum_{\ell=1}^{m} P(\omega_\ell) p(\tilde{X}_j | \omega_\ell) \tag{B-3}$$

If it is assumed that the spectral vectors of the pixels in each field are cluster conditionally independent, then

$$p(\tilde{X}_j | \omega_\ell) = \prod_{k=1}^{N_j} p(X_{jk} | \omega_\ell) \tag{B-4}$$

Using equations (B-3) and (B-4), the joint density of equation (B-2) can be written as follows.

$$p(\tilde{X}_1, \cdots, \tilde{X}_f) = \prod_{j=1}^{f} \left\{ \sum_{\ell=1}^{m} p(\omega_\ell) \left[\prod_{k=1}^{N_j} p(X_{jk} | \omega_\ell) \right] \right\} \tag{B-5}$$

Since the logarithm is a monotonic function of its argument, taking the log of both sides of equation (B-5) and denoting it by ℓ results in

$$\ell = \sum_{j=1}^{f} \log \left\{ \sum_{r=1}^{m} P(\omega_r) \left[\prod_{k=1}^{N_j} p(X_{jk} | \omega_r) \right] \right\} \tag{B-6}$$

From equation (B-6), which is similar to equation (2-7), the maximum likelihood equation for the probability that a field belongs to a cluster can easily be obtained as the following.

$$P(\omega_r) = \frac{1}{f} \sum_{j=1}^{f} \frac{P(\omega_r) p(\tilde{X}_j | \omega_r)}{p(\tilde{X}_j)} \tag{B-7}$$

If θ_i is a parameter of the density function of the i^{th} cluster, differentiating ℓ of equation (B-6) with respect to θ_i yields the following.

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{j=1}^{f} \frac{P(\omega_1) p(\tilde{X}_j | \omega_1)}{p(\tilde{X}_j)} \sum_{k=1}^{N_j} \frac{\partial}{\partial \theta_i} \left\{ \log [p(X_{jk} | \omega_i)] \right\} \tag{B-8}$$

If the probability density functions of the clusters are Gaussian [i.e., $p(X|\omega_i) \sim N(U_1, \Sigma_i)$], from equation (B-8), the maximum likelihood equations for the mean and covariance matrix of the densities of the clusters can be shown to be the following.

$$U_1 = \frac{\sum_{j=1}^{f} N_j p(\omega_i|\tilde{X}_j)\bar{X}_{j\cdot}}{\sum_{j=1}^{f} N_j p(\omega_i|\tilde{X}_j)} \tag{B-9}$$

and

$$\Sigma_1 = \frac{\sum_{j=1}^{f} p(\omega_1|\tilde{X}_j)\left[\sum_{k=1}^{N_j} (X_{jk} - U_1)(X_{jk} - U_i)^T\right]}{\sum_{j=1}^{f} N_j p(\omega_i|\tilde{X}_j)} \tag{B-10}$$

where

$$\bar{X}_{J\cdot} = \frac{1}{N_j} \sum_{k=1}^{N_j} X_{jk} \tag{B-11}$$

If the probability density functions of the clusters are approximated with first-order feature dependence trees, maximum likelihood equations for the parameters of the densities (similar to those developed in section 5) can easily be obtained from equations (B-8) and (3-1) by taking into account the field structure of the data.

APPENDIX C

DEPENDENT FEATURE TREES WITH THE NODES REPRESENTING FEATURE SUBSETS

APPENDIX C

DEPENDENT FEATURE TREES WITH THE NODES REPRESENTING FEATURE SUBSETS

Very often it is necessary to have each node of a dependent feature tree represent a set of features instead of one feature. For example, in remote sensing, the satellite makes multiple passes over a given area and, at each acquisition, gathers several channels of data. In some instances, it is desirable to have each node of a dependent feature tree represent a set of features (e.g., the set of features corresponding to an acquisition). In this appendix, expressions are developed for the mutual information between the feature subsets and for the covariance between the feature subsets when a path connects them in a dependent feature tree. It is assumed that the features are Gaussian distributed.

Let the components of feature vectors X_i and X_j be the sets of features represented by nodes i and j, respectively. Let n_i and n_j be the dimensionality of vectors X_i and X_j, respectively. If the feature vector X_i is normally distributed, its probability density $p(X_i)$ can be represented as

$$p(X_i) \sim N(U_i, \Sigma_i) \tag{C-1}$$

where U_i is the mean vector and Σ_i is the covariance matrix.

Let $Z = \left(X_i^T, X_j^T\right)^T$. Then,

$$p(Z) \sim N(U_Z, \Sigma_Z) \tag{C-2}$$

where

$$\Sigma_Z = \begin{bmatrix} \Sigma_i & \Sigma_{ij} \\ \Sigma_{ij}^T & \Sigma_j \end{bmatrix} \tag{C-3}$$

The mutual information between feature vectors X_i and X_j can be written as

$$I(X_i,X_j) = \int p(X_i,X_j)\log\left[\frac{p(X_i,X_j)}{p(X_i)p(X_j)}\right]dX_i\ dX_j$$

$$= -\frac{1}{2}\log\left(\frac{|\Sigma_z|}{|\Sigma_i||\Sigma_j|}\right) \tag{C-4}$$

$$I(X_i,X_j) = -\frac{1}{2}\log\left(\frac{\left|\Sigma_j - \Sigma_{ij}^T\Sigma_i^{-1}\Sigma_{ij}\right|}{|\Sigma_j|}\right) \tag{C-5}$$

where $|\Sigma_z|$ is the determinant of the matrix Σ_z. Let y and v be the zero mean normal random vectors. That is, $p(y) \sim N(0,C_y)$ and $p(v) \sim N(0,C_v)$. Let $Z = (y^T,v^T)$ and $p(Z) \sim N(0,C_z)$. Let

$$C_z = \begin{bmatrix} C_y & C_{yv} \\ C_{yv}^T & C_v \end{bmatrix} \tag{C-6}$$

and

$$C_z^{-1} = \begin{bmatrix} Q_y & Q_{yv} \\ Q_{yv}^T & Q_v \end{bmatrix}$$

Consider

$$p(y|v) = \frac{p(Z)}{p(v)}$$

$$= \text{constant} \cdot \exp\left(-\frac{1}{2}A\right) \tag{C-7}$$

where

$$A = \left(y + Q_y^{-1}Q_{yv}v\right)^T Q_y \left(y + Q_y^{-1}Q_{yv}v\right) \tag{C-8}$$

Thus, the density $p(y|v)$ is Gaussian with the mean $-Q_y^{-1}Q_{yv}v$ and the covariance matrix Q_y^{-1}. Following a similar argument, it can easily be shown that, if X_i is normally distributed, $p(X_i|X_j)$ is normally distributed with the mean $\left[U_i - Q_i^{-1}Q_{ij}(X_j - U_j)\right]$ and the covariance matrix Q_i^{-1}. Now expressions for the covariance between the feature subsets, when a path connects their representative nodes in a dependent feature tree, can be derived as in section 4.2. For example, if X_4 and X_7 are Gaussian random vectors, similar to equation (4-3), the following can easily be obtained.

$$\int (X_7 - U_7)p(X_7|X_4)dX_7 = -Q_7^{-1}Q_{74}(X_4 - U_4) \qquad \text{(C-9)}$$

Thus, expressions similar to equations (4-8) and (4-9) can easily be obtained.

"Maximum Likelihood Clustering with Dependent Feature Trees"

SUPPORTING RESEARCH
TECHNICAL REPORT

SR-L1-04031
JSC-16853
January 16, 1980

DISTRIBUTION

- Distribution of this document is limited to those people whose names appear without an asterisk. Persons with an asterisk(*) beside their name will receive an abstract only (JSC Form 1424).
- To obtain a copy of this document, contact one of the following:

  F. G. Hall - (SG3)
  J. E. Wainwright - Lockheed Development & Evaluation Department (626-43) (C09)

SK/R. Erb*
J. Powers *
R. Hatch (USDA) *
W. Stephenson *
M. Helfert *
F. Barrett (USDA) *
L. Childs
G. Nixon *

SG/R. MacDonald

SG2/D. Hay
R. Musgrove
W. Weimer *
D. Frank (USDA) *
G. McKain *

SG3/F. Hall
M. Ferguson
C. Hallum
R. Heydorn
A. Houston
D. Thompson
D. Pitts
A. Feiveson
R. Juday
M. Steib
K. Henderson
J. Dietrich

Oregon State University
L. Eisgruber

GMI/R. Holmes

Data Resources, Inc.
B. Scherr

Kansas State University/
E. T. Lab
E. Kanemasu

** LOCKHEED Documents Only

SH2/R. Hill
D. Amsbury *
K. Demel *
K. Hancock *
N. Hatcher *
R. Joosten *
T. Pendleton
C. Forbes *
L. Bennett *
M. Trichel

SH3/J. Dragg
R. Eason *
R. Bizzell
D. Henninger
W. Jones *
G. Kraus *
R. McKinney *
H. Prior *
F. Ravet *
L. Wade *
V. Whitehead *
K. Baker *
J. Carney *
C. Davis
G. Gutschewski *
F. Herbert *
F. Johnson *
R. Patterson *
J. Ginns *

Tom Barnett/NOAA CEAS *
116 Federal Building
Columiba, MO 65201

LOCKHEED
C09/J. G. Baron *
M. L. Bertrand *
B. L. Carroll
P. L. Krumm *
T. C. Minter (10)
D. E. Phinney
P. C. Swanzy *
J. J. Vaccaro *
J. E. Wainwright (3)
J. L. Hawkins (2)
** Job Order File

** B09/Technical Library (5)

ERIM/R. Horvath (4)

PURDUE-LARS/M. E. Bauer (4)

PURDUE University
Dr. Paarlberg

TAMU/L. F. Guseman

UCB/C. M. Hay

IBM- Houston, TX MC16/
A. Anderson

IBM - Palo Alto, CA/Dr. Kolsky

USDA - Houston, TX
G. Boatwright

USDC - Washington, D.C./
R. Ambroziak

USDC/EDIS/CEAS - Columbia, MO
S. LeDuc
C. Sakamoto