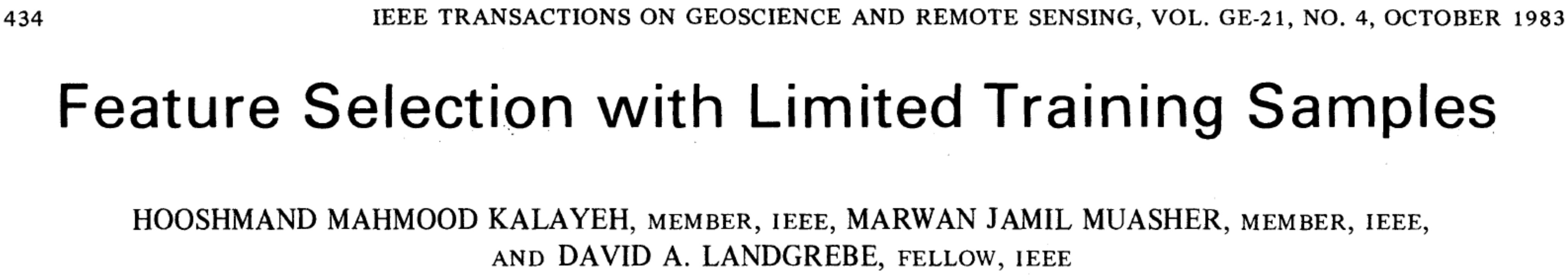


434 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. GE-21, NO. 4, OCTOBER 1983

Feature Selection with Limited Training Samples

HOOSHMAND MAHMOOD KALAYEH, MEMBER, IEEE, MARWAN JAMIL MUASHER, MEMBER, IEEE, AND DAVID A. LANDGREBE, FELLOW, IEEE

Presentado por: Alexandre Savio

Grupo de Inteligencia Computacional (GIC)
Universidad del País Vasco (UPV/EHU)

EUSKO JAURLARITZA
GOBIERNO VASCO

Summary

What?

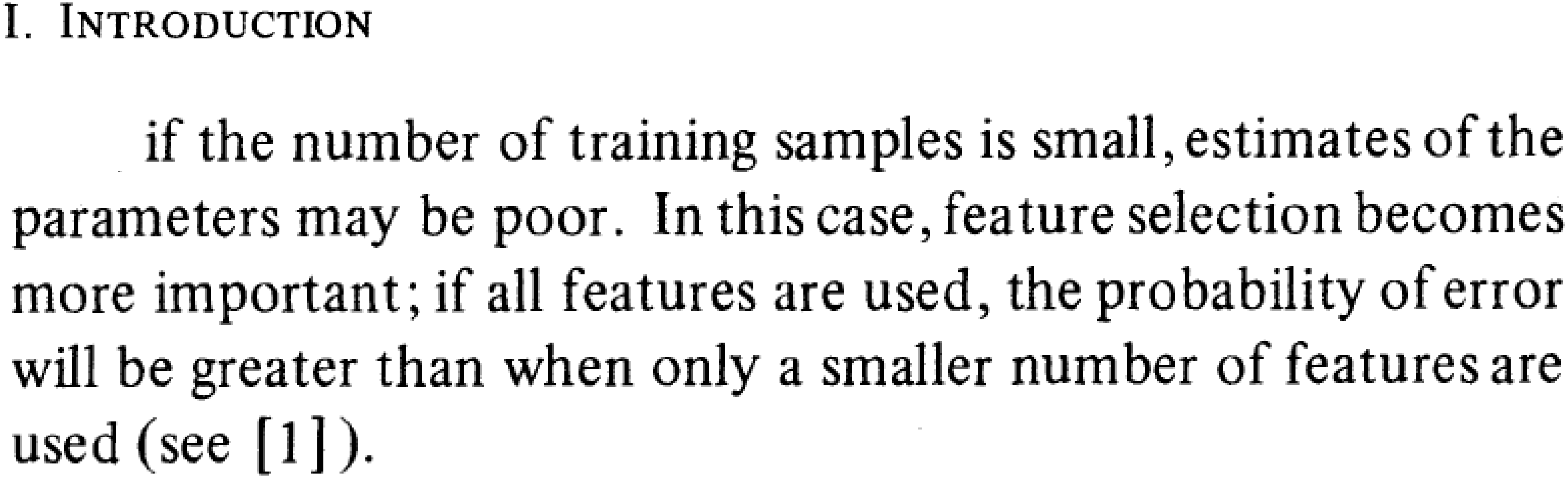
- A criterion to measure the quality of estimates of the parameters of multivariate normal distributions.

Definition

- Definition of the criterion

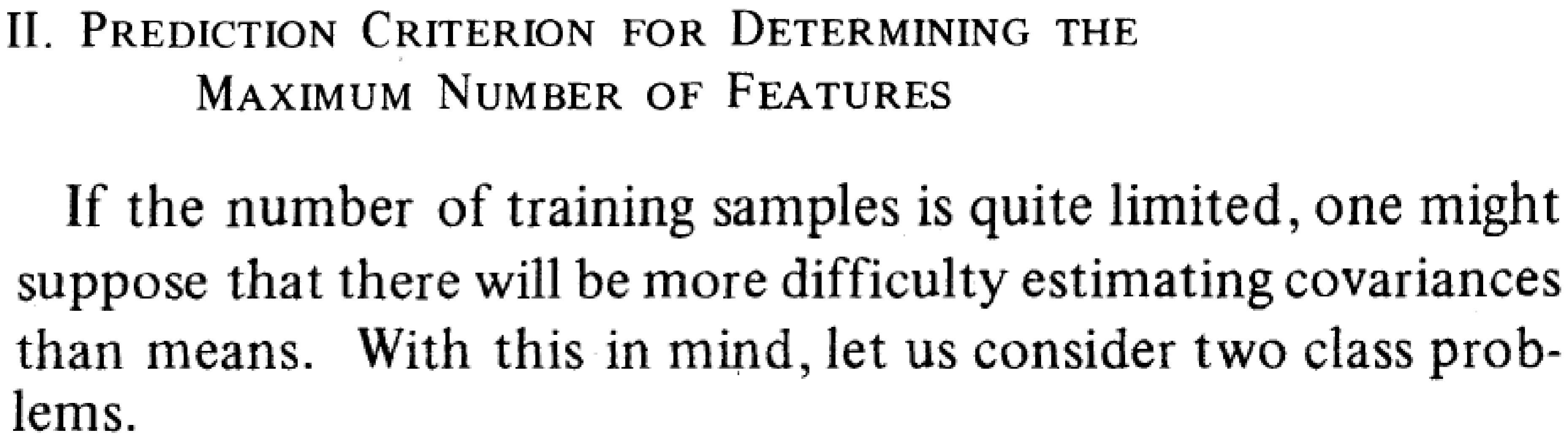
Experiments

- Experimental results
  - An aircraft dataset
  - LADSAT set



I. Introduction

if the number of training samples is small, estimates of the parameters may be poor. In this case, feature selection becomes more important; if all features are used, the probability of error will be greater than when only a smaller number of features are used (see [1]).



II. Prediction Criterion for Determining the Maximum Number of Features

If the number of training samples is quite limited, one might suppose that there will be more difficulty estimating covariances than means. With this in mind, let us consider two class problems.

II. Prediction Criterion for Determining the Maximum Number of Features

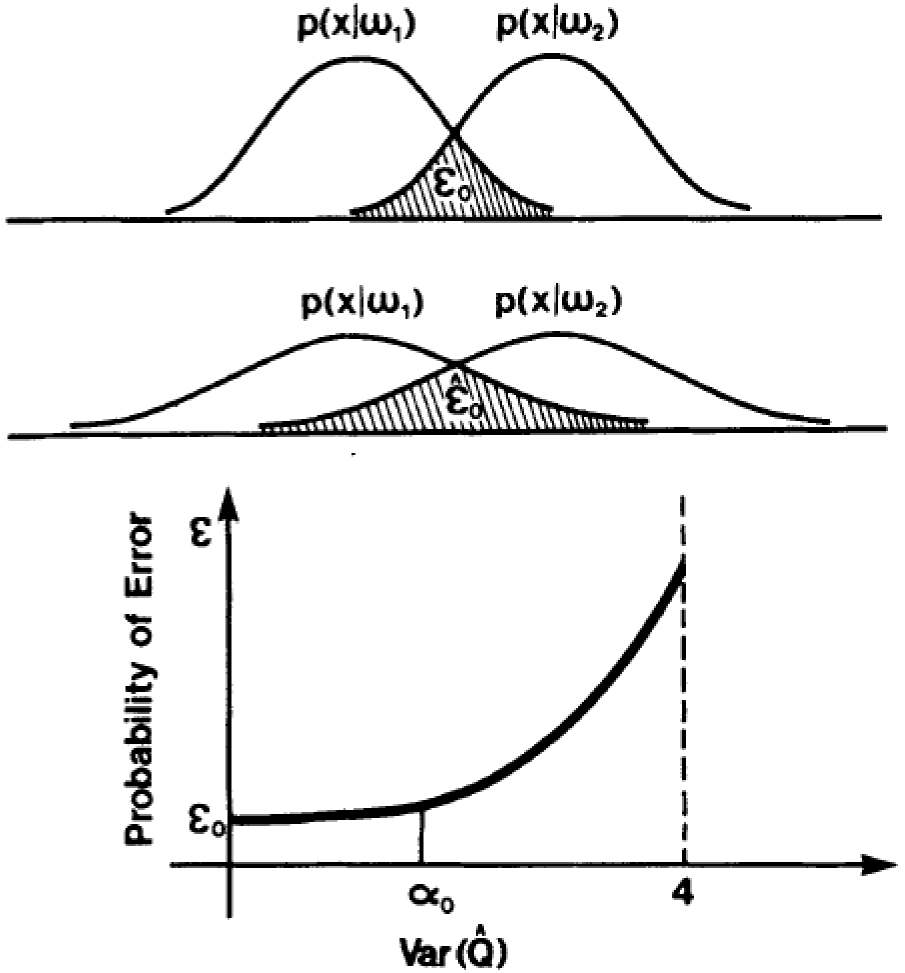


Fig. 1. Explanation of degradation in accuracy.

II. Prediction Criterion for Determining the Maximum Number of Features

Let Σ_1 and Σ_2 be estimates of the covariance matrices based on n_1 and n_2 samples of class ω_1 and class ω_2, respectively. Then let $\hat{I}$, $\hat{\Lambda}$ be the estimates of covariance matrices of ω_1 and ω_2 after applying simultaneous diagonalization transformations where $\hat{I}$ is an estimate of the identity matrix and $\hat{\Lambda}$ is a diagonal matrix (see [2]). Let α_{ii} and λ_{ii} be the diagonal elements of $\hat{I}$ and $\hat{\Lambda}$, respectively. Then let

$$\hat{Q}_1 = \sum_{i=1}^{q} \hat{\alpha}_{ii} \tag{1}$$

$$\hat{Q}_2 = \sum_{i=1}^{q} \frac{\hat{\lambda}_{ii}}{\lambda_{ii}} \tag{2}$$

$$\hat{Q} = \hat{Q}_1 + \hat{Q}_2 \tag{3}$$

where q is the number of features used.

II. Prediction Criterion for Determining the Maximum Number of Features

where q is the number of features used. In the new space, the features are independent and so are their variances in class ω_1 and class ω_2. Furthermore, we are assuming the elements of the covariance matrices of two classes in the new space are independent. Consequently, it can be said that $\hat{Q}_1$ and $\hat{Q}_2$ are two independent random variables; then we can write

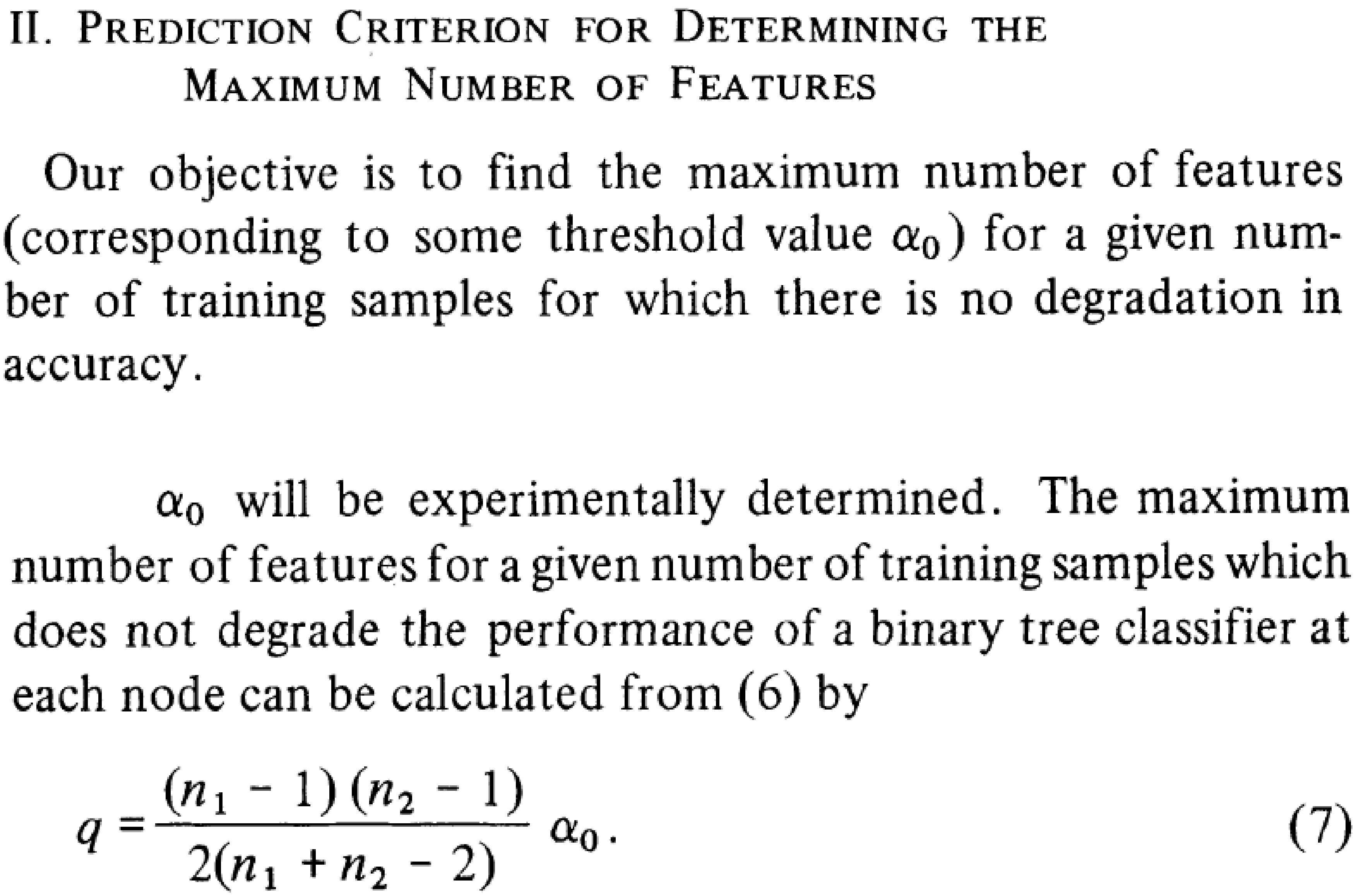
$$\operatorname{var}(\hat{Q}) = \operatorname{var}(\hat{Q}_1 + \hat{Q}_2) = \operatorname{var}(\hat{Q}_1) + \operatorname{var}(\hat{Q}_2). \tag{4}$$

In [3], it is shown that

$$\operatorname{var}(\hat{Q}_1) = 2\,q/(n_i - 1). \tag{5}$$

We will choose $\operatorname{var}(Q_1 + Q_2)$ as our prediction criterion to determine the maximum number of features for which there is no degradation in accuracy. Then we have

$$\operatorname{var}(\hat{Q}) = 2\,q/(n_1 - 1) + 2\,q/(n_2 - 1). \tag{6}$$

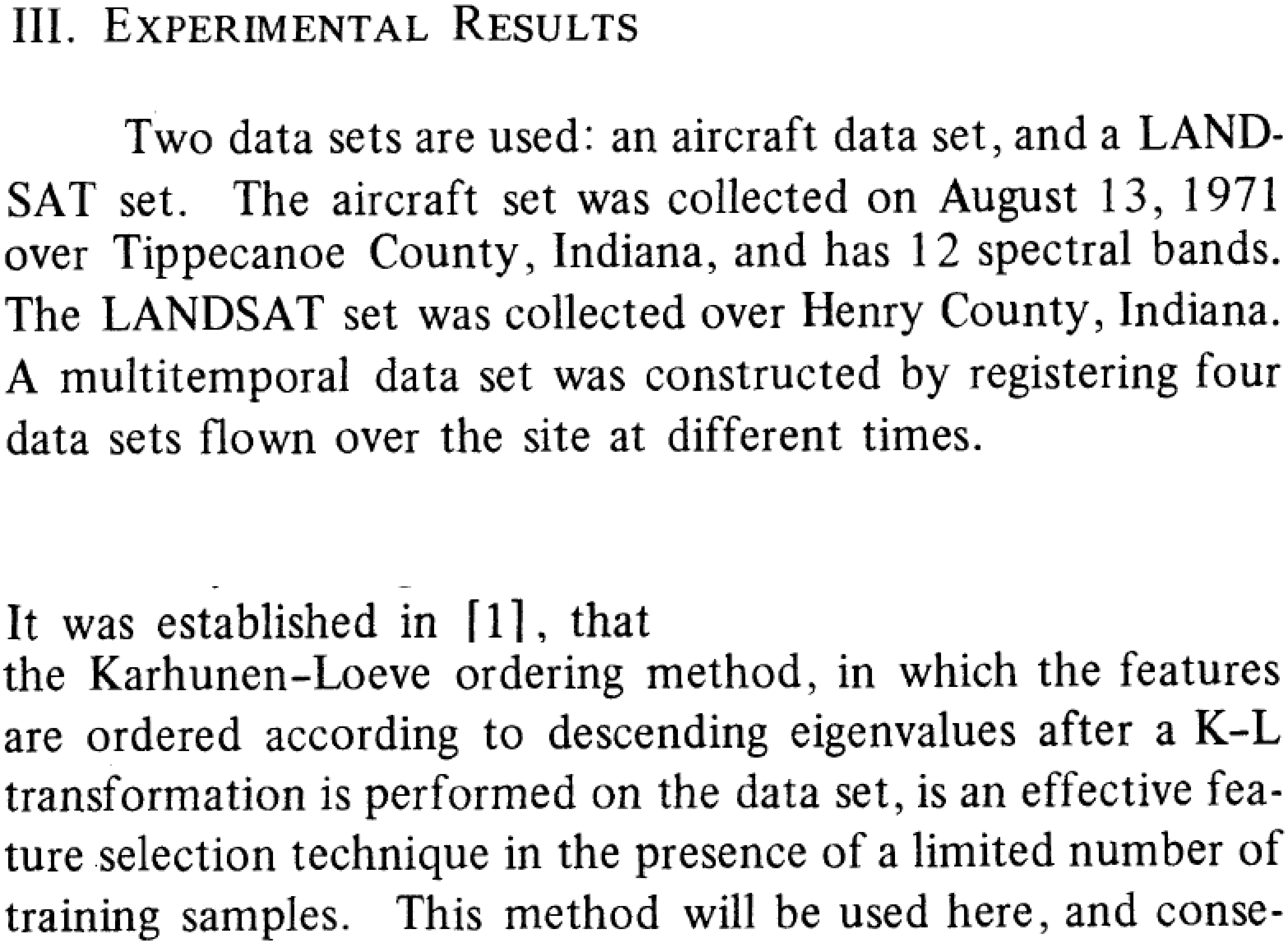


II. Prediction Criterion for Determining the Maximum Number of Features

Our objective is to find the maximum number of features (corresponding to some threshold value α_0) for a given number of training samples for which there is no degradation in accuracy.

α_0 will be experimentally determined. The maximum number of features for a given number of training samples which does not degrade the performance of a binary tree classifier at each node can be calculated from (6) by

$$q = \frac{(n_1 - 1)(n_2 - 1)}{2(n_1 + n_2 - 2)} \alpha_0 . \tag{7}$$

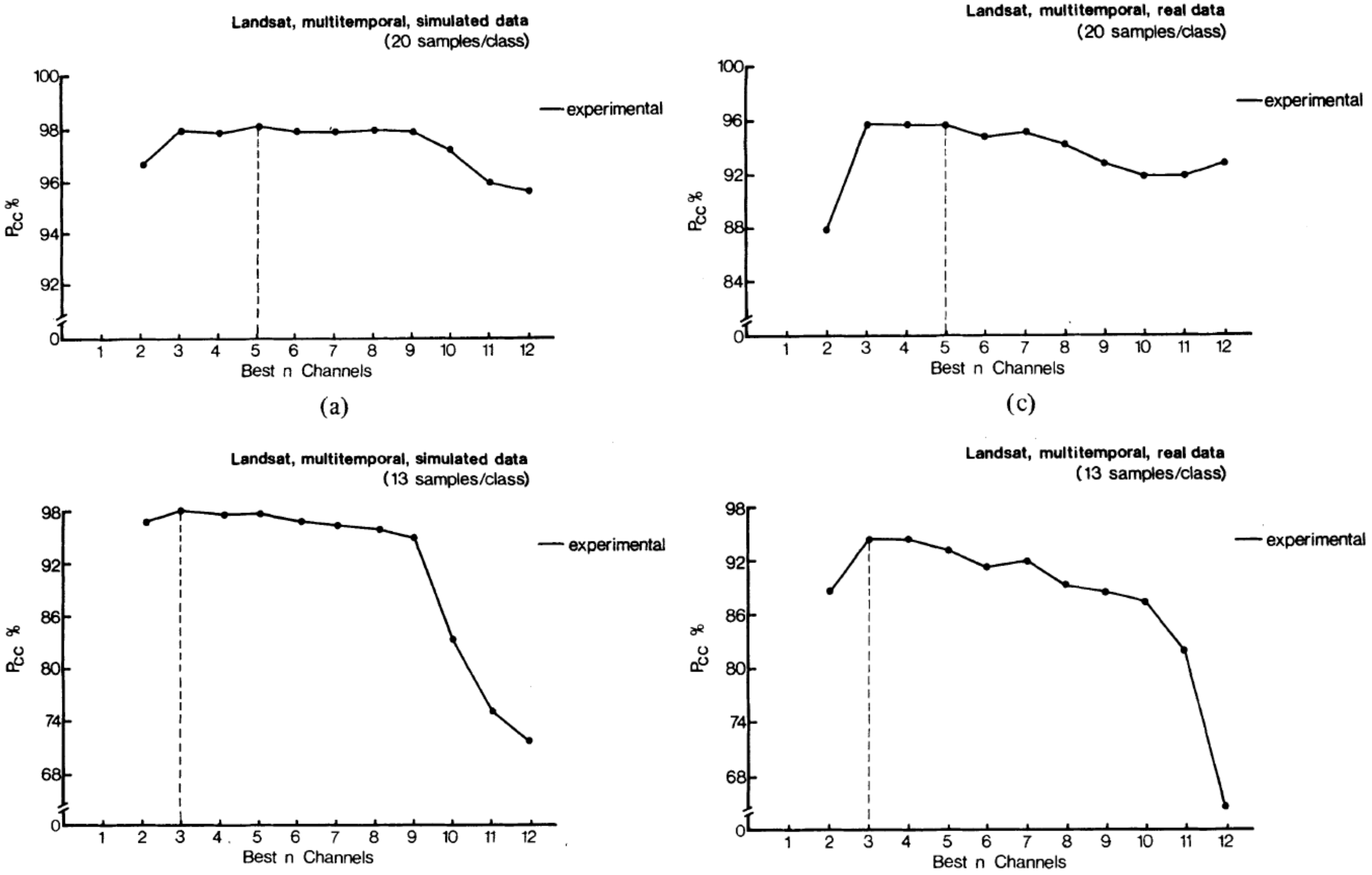


III. Experimental Results

Two data sets are used: an aircraft data set, and a LANDSAT set. The aircraft set was collected on August 13, 1971 over Tippecanoe County, Indiana, and has 12 spectral bands. The LANDSAT set was collected over Henry County, Indiana. A multitemporal data set was constructed by registering four data sets flown over the site at different times.

It was established in [1], that
the Karhunen–Loeve ordering method, in which the features are ordered according to descending eigenvalues after a K–L transformation is performed on the data set, is an effective feature selection technique in the presence of a limited number of training samples. This method will be used here, and conse-

III. Experimental Results



Landsat, multitemporal, simulated data
(20 samples/class)

(a)

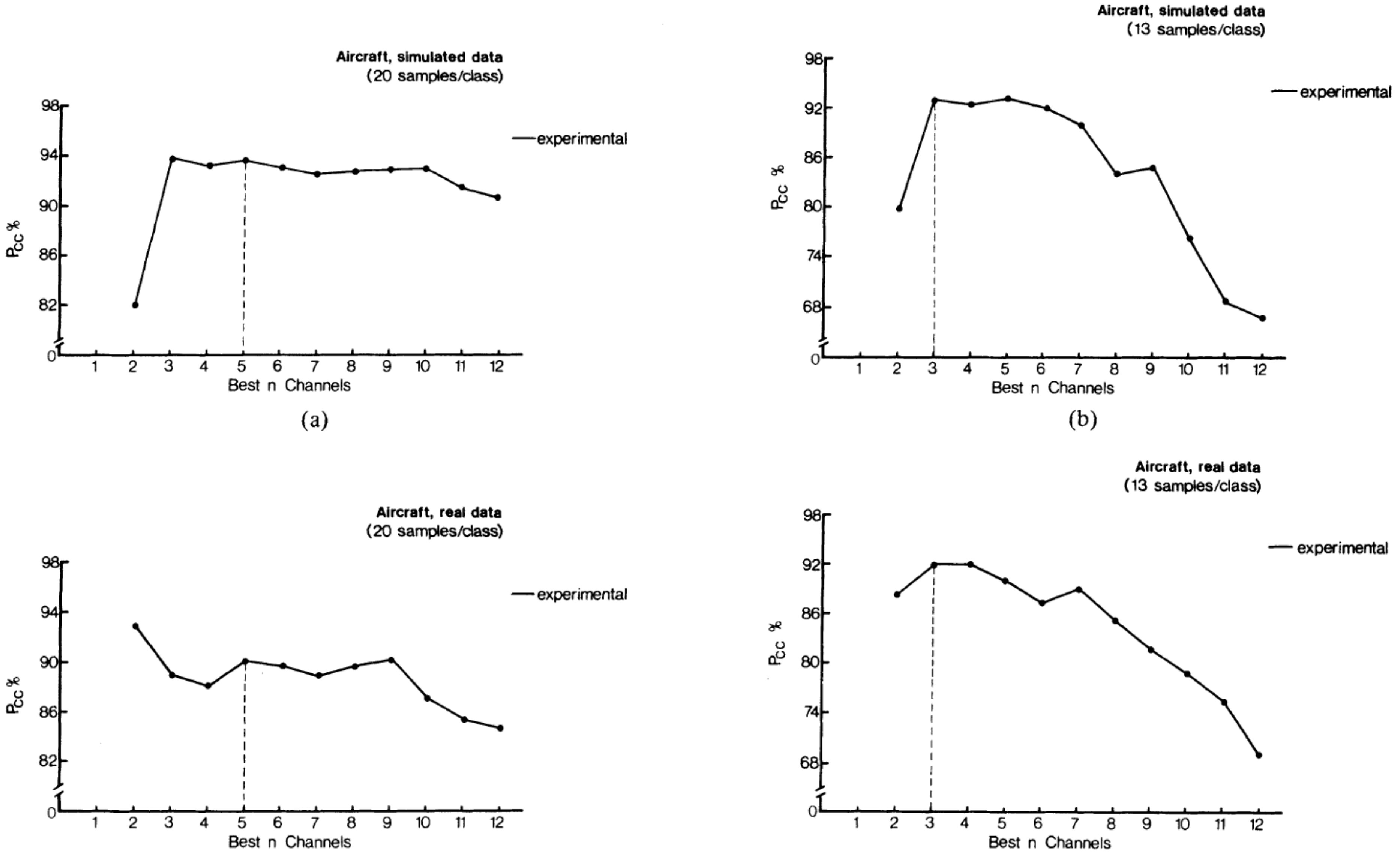
Landsat, multitemporal, real data
(20 samples/class)

(c)

Landsat, multitemporal, simulated data
(13 samples/class)

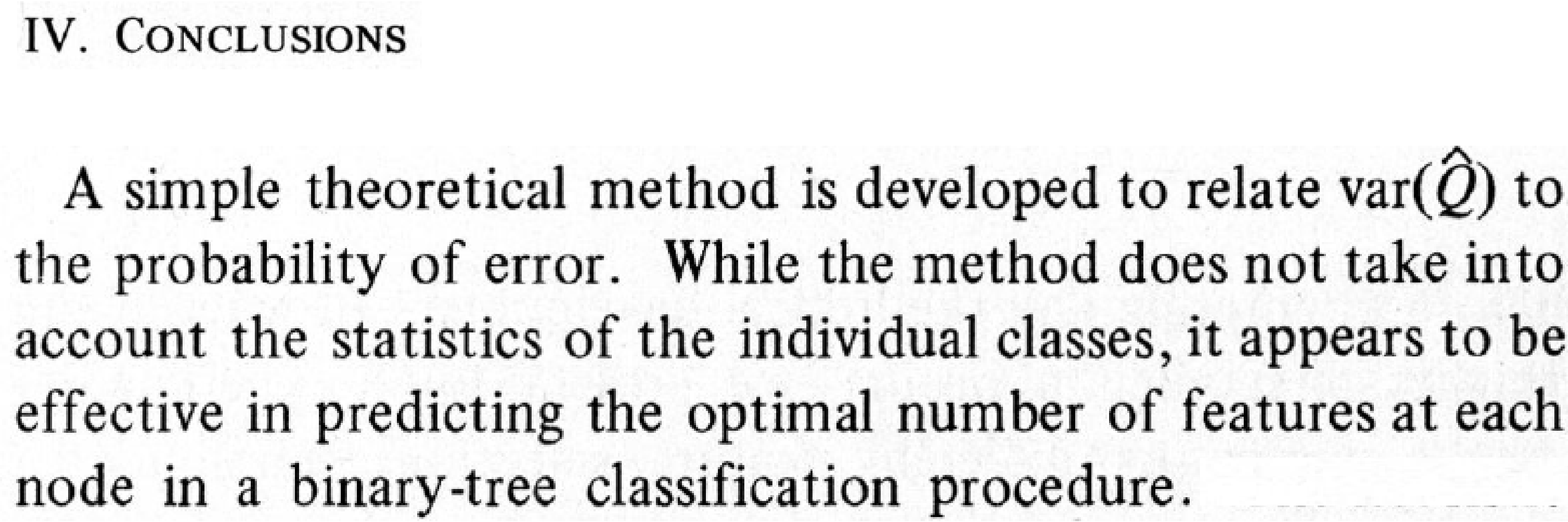
Landsat, multitemporal, real data
(13 samples/class)

III. Experimental Results



(a)

(b)



IV. Conclusions

A simple theoretical method is developed to relate var($\hat{Q}$) to the probability of error. While the method does not take into account the statistics of the individual classes, it appears to be effective in predicting the optimal number of features at each node in a binary-tree classification procedure.