AN ABSTRACT OF THE THESIS OF

Patricia K. Lebow for the degree of Doctor of Philosophy in Statistics presented on October 23, 1992.

Title: Estimation of Discriminant Analysis Error Rate for High Dimensional Data

Redacted for Privacy

Abstract approved: _____

David A. Butler

Redacted for Privacy

_____

David R. Thomas

Methodologies for data reduction, modeling, and classification of grouped response curves are explored. In particular, the thesis focuses on the analysis of a collection of highly correlated, highly dimensional response-curve data of spectral reflectance curves of wood surface features.

In the analysis, questions about the application of cross-validation estimation of discriminant function error rates for data that has been previously transformed by principal component analysis arise. Performing cross-validation requires re-calculating the principal component transformation and discriminant functions of the training sets, a very lengthy process. A more efficient approach of carrying out the cross-validation calculations, plus the alternative of estimating error rates without the re-calculation of the principal component decomposition, are studied to address questions about the cross-validation procedure.

If populations are assumed to have common covariance structures, the pooled covariance matrix can be decomposed for the principal component transformation. The leave-one-out cross-validation procedure results in a rank-one update in the pooled covariance matrix for each observation left out. Algorithms have been developed for calculating the updated eigenstructure under rank-one updates and they can be applied to the orthogonal decomposition of the pooled covariance matrix. Use of these algorithms results in much faster computation of error rates, especially when the number of variables is large.

The bias and variance of an estimator that performs leave-one-out cross-validation directly on the principal component scores (without re-computation of the principal component transformation for each observation) is also investigated.

Estimation of Discriminant Analysis Error Rate for High Dimensional Data

by

Patricia K. Lebow

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed October 23, 1992

Commencement June 1993

APPROVED:

Redacted for Privacy

Professor of Statistics in charge of major

Redacted for Privacy

Professor of Statistics in charge of major

Redacted for Privacy

Head of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented October 23, 1992

Typed by Patricia K. Lebow

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ESTIMATION OF DISCRIMINANT ANALYSIS ERROR RATE FOR HIGH DIMENSIONAL DATA

Chapter 1
INTRODUCTION

With the advent of high-speed microprocessors, high-capacity storage devices, and many other technological advances, today's instrumentation allows the rapid collection of vast amounts of data. Unfortunately, the analysis of such data is hindered by theoretical and computational considerations. Chemists are one group who have experienced such problems, and in response the formal area of chemometrics (or chemostatistics) has evolved. Statisticians have also recognized that traditional statistical methods may fail when applied to large, complicated data sets, but have made little progress in developing new analytic procedures. These issues have been addressed in the Institute of Mathematical Statistics' cross-disciplinary research review (1990) and the American Statistical Associations' *Challenges for the 90's* (Spiegelman, 1989). In addition, differing terminologies impede understanding between the fields (Kowalski and Wold, 1982).

Response-curve data are typical in chemical and other areas of research either as a single response measured over time or along some spatial axis or as an observation composed of multiple responses that have some ordinal relationship. The percentage of a certain chemical's retention in a certain substance over time would be an example of the former case, while the percentage of light absorbed at several different wavelengths by a certain substance (light absorbance curves) would be an example of the latter case. In both cases, the data tends to be highly dimensional and highly correlated, and it is not uncommon that the number of variables (responses, features) measured greatly exceeds the number of response curves observed. Depending on the questions to be investigated with the data, in the former case of repeated measurements of a single response, time series analysis may be the appropriate analytical tool. But in the latter case of multiple responses across some axis,

with which this thesis is primarily concerned, traditional multivariate statistical methods, or some variation thereof, are usually applied.

It is not unusual for the observations to fall into distinct classification groups that can be observed at the time the original response curves are gathered. With such information available, the common goals include developing a model to characterize the data to gain a better understanding of the underlying response-curve structure and to allow for data reduction, and developing classification rules for future unknown observations. This will also facilitate the understanding of the data structure in relation to the groups.

In developing a model to describe this data, reducing the dimensionality is of primary importance. Methods for data reduction and modeling are numerous, but in the applied literature we find the most common procedures employed with response-curve data (grouped or ungrouped) are principal component analysis and factor analysis. For grouped data in which discrimination is desired, this may be followed by linear discriminant analysis. The specifics of these procedures, however, tend to vary throughout the literature. This thesis reviews this practice, looks at its implications for a given data set, and offers computationally efficient alternatives.

The organization of this thesis is as follows.

In Chapter 2, the methodology and terminology of linear models, principal component analysis, and linear discriminant analysis are introduced. Additionally, the chapter reviews the literature relevant to the analysis of response-curve data when a response curve is an observation composed of multiple responses having some ordinal relationship and the curve has some meaningful classification that can be noted at the time of measurement.

Chapter 3 presents a data set typical of physical and chemical applications, a data base of spectral reflectance curves of wood surface features. This data set, a collection of grouped response curves, exhibits the characteristics of high dimensionality and high correlation in terms of the original variables measured. The initial goal in analyzing this data was to determine a model that achieves maximal data reduction while maintaining the reproducibility of the original data. A secondary goal, since the type of wood surface feature could be observed at the time of the spectral reflectance

measurement, was to develop classification procedures and to determine to the regions of the spectrum that were the most discriminating.

A review of the literature, the analysis goals, and a preliminary statistical analyses of the data set led to the application of principal component analysis followed by linear discriminant analysis. However, the problem of estimating discrimination misclassification rates (error rates) with the reduced data arises. In particular, performing the cross-validation estimation procedure for error rates requires re-calculating the principal component decomposition and discriminant functions of the training sets, a very lengthy process. The alternative is to perform the cross-validation procedure on the principal component scores without re-calculation of the principal component transformation. These statistical analyses of the spectral reflectance data are reviewed in Chapter 3.

If common covariance structures between populations are assumed, the pooled covariance matrix may be used for decomposition in principal component analysis. The leave-one-out cross-validation procedure results in a rank-one change in the pooled covariance matrix for each observation left out. Algorithms have been developed for calculating the updated eigenstructure under rank-one changes, and in Chapter 4, one particular algorithm is incorporated to calculate the orthogonal decomposition of the updated pooled covariance matrix. Use of the algorithm results in much faster computation of the estimated error rates, especially if the number of original variables is very large. The parallelism employed in the algorithm is another advantage as computers employing parallel processing become more widely available.

Chapter 5 gives a more detailed comparison of the error rate estimators for the wood data set under varying conditions, including changes in the number of variables in the spectral reflectance curve, the number of principal components kept for discrimination, the number of groups (wood surface feature types), and the number of observations (sample size).

In Chapter 6 the faster alternative of performing leave-one-out cross-validation directly on the principal components without recomputating the principal component analysis for each observation is compared to the rank-one procedure. The bias and variance of the estimators is investigated using

simulated normally distributed data that exhibits characteristics similar to the observed spectral reflectance data.

Chapter 7 summarizes the findings of this thesis and offers suggestions for further research.

Chapter 2
LITERATURE REVIEW

Usually when first analyzing a data set, the simplest statistical procedures are used unless there is an indication prior to the application of the procedures that the assumptions for their use are violated. After the simple procedures have been applied, the results may indicate that more complicated procedures are necessary for the proper evaluation of the data. Such is the case in the applied literature for the analysis of response curves.

General Methodology -- Linear Model

Consider having a population of response curves that are measurable by some means at a fixed set of points. It is possible to characterize a random curve from the population by the random vector $\mathbf{x} = [x_1\ x_2 \cdots\ x_p]'$ where x_j, j=1, 2, $\cdots$, p, are the individual measurements at the points represented by j. Unless the researcher has a particular nonlinear model in mind for describing a response curve, it is commonly assumed $\mathbf{x}$ can be represented by a multivariate linear model, not only because of the linear model's simplicity but because examination of its residuals may show the assumptions underlying its use are being violated and, hence, suggest a more complicated model.

Specifically, let the linear model for $\mathbf{x}$ be expressed as

$$\mathbf{x} = \mathbf{G}\,\boldsymbol{\gamma} + \boldsymbol{\epsilon} = \sum_{k=1}^{m} \gamma_k\,\mathbf{g}_k + \boldsymbol{\epsilon} \tag{2.1}$$

where $m \leq p$, $\boldsymbol{\gamma} = [\gamma_1\ \gamma_2 \cdots\ \gamma_m]'$ is a parametric vector, $\boldsymbol{\epsilon} = [\epsilon_1\ \epsilon_2 \cdots\ \epsilon_p]'$ is the residual vector, and $\mathbf{G} = [\ \mathbf{g}_1\ \ \mathbf{g}_2\ \cdots\ \mathbf{g}_m\ ]$ with $\mathbf{g}_k = [\ g_k(1)\ \ g_k(2) \cdots\ g_k(p)]'$. The $g_k(\cdot)$ may represent polynomial functions, trigonometric functions, or some other suitable functions that depend on the nature and structure of the response curve. In this context, the $g_k(\cdot)$ are referred to as basis functions, and the response curve is just a linear combination of these basis functions.

Examples in the literature for modeling optical response curves include Legendre polynomials (Healey and Binford, 1987), band-limited trigonometric

functions (Stiles *et al.*, 1977), unit step functions (Stiles and Wyszecki, 1962) and empirical orthogonal functions (Cohen, 1964). Moon and Spencer (1945) appear to be the first to have used polynomials for this purpose but it is unknown what type of polynomial basis functions they used as they only reported their polynomial coefficients.

Typically, a sample of n observations $\mathbf{X} = [\ \mathbf{x}_1\ \ \mathbf{x}_2 \cdots\ \mathbf{x}_n\ ]'$, where each $\mathbf{x}_i$ is a response curve, is gathered from the population of interest. This data can be used for estimating the model (2.1) with iterated least-square methods (Wold, 1966). This results in the basis functions $g_k(\cdot)$ being estimated by the empirical orthogonal functions, otherwise known as the principal components. (In this respect, with $\mathbf{G}$ being considered unknown, the model can also be viewed as a nonlinear model.)

Among the assumptions of model (2.1) are the additivity of the error and the independence of the observations (experimental units).

Preprocessing -- Data Reduction and Modeling

Methods for data reduction and modeling are numerous, but in the applied literature we find the most common procedures employed with response curve data are principal component analysis and factor analysis. Principal component analysis has been applied to spectral reflectance curves (Cohen, 1964, Maloney, 1986, Healy, 1989, and Tominaga and Wandell, 1989, Parkkinen *et al.*, 1989), radiation curves of paper (Grum and Wightman, 1960), spectral response curves of neurons in monkeys (Young, 1986), spectral densities of patches of color film (Morris and Morrissey, 1954), spectral irradiance curves of daylight (Judd *et al.*, 1964), absorbance curves (Rao, 1964, Cochran and Horne, 1977), transmittance curves (Parkkinen and Jaaskelainen, 1987), mass spectra (Rozett and Peterson, 1975, 1976, Justice and Isenhour, 1975, Hoogerbrugge *et al.*, 1983), motor torque over time (Church, 1966), and so on. Simonds (1963) appears to have laid much of the foundation for many of these applications when he described the use of principal component analysis for hypothetical photographic and optical response curves.

In the chemical literature, we are likely to find that principal component analysis has been done as part of a factor analytic model (see Rozett and Peterson, 1975 and 1976, Justice and Isenhour, 1975, Malinowski and Howery,

1980, Sharaf *et al.*, 1986), usually without distinction from the closely related principal factor analysis. Principal component analysis is actually a special case of principal factor analysis with certain constraints (see Mardia *et al.*, 1982). Malinowski and Howery claim in their book *Factor Analysis in Chemistry* that before 1980 nearly 100 chemical articles had employed factor analysis, indicating the popularity of the procedures. These are commonly referred to as preprocessing techniques implying that further data analysis is likely.

Gnanadesikan (1977) offers a nonlinear approach to principal component analysis, called generalized principal component analysis, that can be used when the data can be characterized by some nonlinear coordinate system. See Hastie and Stuetzle (1989) for a recent iterative approach to finding such a coordinate system, referred to as principal curves. Multidimensional scaling offers yet another nonlinear data reduction method by describing the data set in terms of a distance matrix which is analyzed by either a metric or nonmetric technique. The distances need not be Euclidean distances; they may represent dissimilarities or similarities between objects. With Euclidean distances, however, the classical multidimensional scaling approach essentially yields the linear method of principal component analysis. See Mardia *et al.* (1982), Gnanadesikan (1977), or Green (1978) for mathematical formulations.

General Methodology -- Principal Component Analysis

Principal component analysis offers data reduction, the ability to find linear combinations of the original variables with relatively large (or small) variability, and, in the single population case, the ability to transform correlated variables into uncorrelated ones. The extension of principal components to more than one population will be discussed later. Following the terminology of Mardia et al. (1982), assume we have a random vector $\mathbf{x} = [x_1\ x_2\ \cdots\ x_p]'$ from some population with mean $\boldsymbol{\mu}$ and covariance Σ. The principal component transformation is defined as

$$\mathbf{y} = \beta'(\mathbf{x} - \boldsymbol{\mu}) \tag{2.2}$$

where β is a pxp orthogonal matrix ($\beta\beta' = \beta'\beta = \mathbf{I}$) such that $\Sigma = \beta\Lambda\beta'$ where

$\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Define $\beta = [\beta_1 \ \beta_2 \cdots \beta_p]$, then the i-th principal component of $\mathbf{x}$ is

$$y_i = \beta_i' (\mathbf{x} - \mu)$$

or, sometimes (without mean adjustment),

$$y_i + \beta_i'\mu = \beta_i'\mathbf{x} . \tag{2.3}$$

The column vector β_i is referred to as the i-th vector of principal component loadings. In this construct, the principal components are uncorrelated, the variance of the i-th principal component is the eigenvalue λ_i, and no standardized linear combination of $\mathbf{x}$ has a variance larger than λ_1. The linear combination $\mathbf{a}'\mathbf{x}$ is a standardized linear combination when $\sum a_i^2 = 1$. Verification of these properties and additional properties may be found in Mardia et al. (1982). It should be noted that principal components are not scale invariant; the random vector $\mathbf{x}$ measured on a different scale is likely to have another principal component transformation.

The principal components of a sample are defined in a manner analogous to the principal components of a population. Define the sample data matrix from the population as

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_n]' \tag{2.4}$$

where each p-vector $\mathbf{x}_i$ is an observation and n is the number of observations from the population. Now, in the above, replace Σ by some statistic $\widehat{\Sigma}$ which is symmetric and positive semidefinite. Then just take the spectral decomposition of $\widehat{\Sigma}$, which is say $\mathbf{VDV}'$ where $\mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \cdots \mathbf{v}_p]$ is the matrix of eigenvectors and $\mathbf{D} = \text{diag}(d_1, d_2, \cdots, d_p)$, $d_1 \geq d_2 \geq \cdots \geq d_p$, is the diagonal matrix of eigenvalues.

Under the assumptions of population normality and positive definite covariance structure, the maximum likelihood estimates for the population

eigenvalues and eigenvectors are those obtained from the spectral decomposition of the unbiased sample covariance matrix, **S** (Flury, 1988). Note

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \mathbf{VDV}' = \sum_{k=1}^{p} d_k \mathbf{v}_k \mathbf{v}_k' \tag{2.5}$$

where

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{\mathrm{n}} \tag{2.6}$$

is the sample mean and n is the number of observations. The k-th sample principal component (without adjustment for the mean) can be written as

$$\mathbf{u}_k = \mathbf{X}\mathbf{v}_k$$

If, say, r population characteristic roots (eigenvalues) are the same, there are some modifications (Flury, 1988). In this situation, the eigenvectors associated with the equal eigenvalues can be chosen arbitrarily as long as they are orthogonal to each other and span an r-dimensional subspace that is orthogonal to the remaining eigenvectors (that are associated with the distinct eigenvalues).

Data reduction and modeling in principal component analysis is achieved by writing each observation as a linear combination of the empirical functions, or eigenvectors of $\hat{\Sigma}$,

$$\mathbf{x}_i = \sum_{k=1}^{q} \mathbf{v}_k u_{ki} + \mathbf{e}_i$$

where u_{ki} is the i-th element of the k-th principal component vector and $q < p$. One attempts to choose q such that the remainder, $\mathbf{e}_i$, can be considered noise.

Variations and Extensions of Principal Component Analysis

For the single population case with little or no distributional assumptions, the choice of $\hat{\Sigma}$ in principal component analysis varies throughout the literature and may depend on the actual application. In addition to the

variance-covariance matrix, the correlation matrix is a popular choice since it is the variance-covariance matrix of the standardized variables. A standardized variable is the column of observations corresponding to the variable of interest divided by the appropriate standard deviation. If variables are measured on different scales or there are large differences in the variables' variances, performing principal component analysis on the correlation matrix will remove the influence of scale. As Green (1978) points out, other sum-of-product matrices that can be decomposed include the raw sums-of-squares and cross-products, and the mean-corrected sums-of-squares and cross-products.

For several populations, principal component analysis is not as well defined as in the single-population situation. Assuming a common covariance structure for the populations simplifies matters somewhat, but there remains the decision of what estimator of Σ ($= \Sigma_1 = \cdots = \Sigma_g$) to use. The maximum likelihood estimator of Σ, $\mathbf{S} = \sum n_j \mathbf{S}_j / n$ with $n = \sum n_j$, is a reasonable choice (Krzanowski, 1984). Anderson (1984, p. 405) derives this estimator for the likelihood ratio test of equal covariance matrices. Because this estimator is biased, however, other statisticians prefer the unbiased estimator (given in 2.8) (Manly and Rayner in likelihood ratio tests, 1987, and Flury, 1988). Other linear combinations of single population estimators may be used, such as the within-group sums-of-squares and cross-products matrix or some type of pooled correlation matrix.

If modeling is of less importance than data reduction and retainment of among-group variation for discrimination, then decomposition of the total sums-of-squares and products matrix, or some variant thereof, may be a more suitable approach. Hoogerbrugge *et al.* (1983) uses the total covariance matrix in preprocessing before discrimination; Rao (1958) suggests this alternative in testing mean differences between groups; and Church (1966) also uses total variation in his analysis of response curve data. Taking single observations from identifiable groups do not allow estimation of within-group variation; in this case, the overall variation is the only alternative. For the illustrative example in Chapter 3, modeling the within-group variation is more important and the within-group variation dominates any of the between-group variations.

The application of principal component analysis when other relationships besides equality are assumed between population covariance matrices has been

investigated by several researchers. Krzanowski (1979), apparently the first to publish on this subject, mathematically compared principal component subspaces between groups to obtain a measure of similarity. Flury (1988), who has summarized much of his and others' work, constructed a hierarchy of covariance matrix relationships that includes proportionality, common eigenvector structures, and partially common eigenvector structures. Although the theoretical work for the common principal component model (ie., common eigenvector structures) appears well-founded, the results for the proportional and partially common eigenvector structures are based on approximations from the common principal component model and are still under investigation. Flury's methods are developed with the assumption that the sample covariance matrices have a Wishart distribution (as in the case of normal i.i.d. observations).

Following his order-of-covariance-matrix relationships, Flury (1988) partitions the likelihood ratio statistic to test structural associations between covariance matrices. Manly and Rayner (1987) similarly break down the likelihood ratio test according to a somewhat different hierarchy.

Graphical comparison procedures as well as formal tests for the equality of covariance matrices exist. Seber (1984) provides a good discussion. The M-test, an unbiased version of the likelihood ratio test based on multivariate normal distributions, compares generalized variances of the individual groups to the generalized variance of the pooled groups. Generalized variance is mathematically defined as the determinant of the covariance matrix. Unfortunately, the test is well-known to be nonrobust because of its sensitivity to nonnormality and kurtosis; also, the test was developed for populations with positive definite covariance matrices. See also Anderson (1984).

Seber (1984) also describes graphical procedures for examining equality of covariance matrices. These include comparisons between groups of traces or determinants of the corrected sum-of-squares matrices and comparisons between groups of the transformed distinct elements of the covariance and correlation matrices. The latter procedure becomes prohibitive as the number of variables increases.

If no covariance relationships between groups are assumed, then one proceeds as if the population covariance matrices are unequal. In such

instances, one should analyze using separate principal component models, one for each group. SIMCA, a supervised learning procedure popular for analyzing chemical data (to be discussed later), employs such disjoint principal component models. Under the assumption of equal covariance matrices, such a scheme would fail to utilize all available information and would require more data since more parameters must be estimated.

Data Reduction With Principal Component Analysis

Data reduction in principal component analysis can be achieved as outlined earlier by choosing a number of the first components, q, that is fewer than the original number of variables, p (i.e., $q < p$). This is based on the fact that the smaller components contribute less towards the total variance than do the larger ones, implying that the variation explained by the smaller components is of less importance and essentially noise. Total variance is mathematically defined as the sum of the variances of all variables or, equivalently, the trace of the covariance matrix. In principal component analysis total variance is given by the trace of the matrix that is decomposed and, depending on the choice of matrix, can have different interpretations. Data reduction should not be attempted if small components contribute substantially.

Assuming the sample eigenvalues are in decreasing order, $d_1 \geq d_2 \geq \cdots \geq d_p$, many procedures for the selection of q exist, including

(i) choosing q by Cattell's scree test, which plots the eigenvalues (or their contribution to total variation) in decreasing order. An 'elbow' in the plot, i.e. where the change in the ordered eigenvalues takes a noticeable jump, indicates the number of components, q, to keep. (Mardia *et al.*, 1982, Green, 1978, Malinowski and Howery, 1980)

(ii) choosing q so that $\sum^{q} d_i \; / \; \sum^{p} d_i > a$, where the d_i are the observed ordered eigenvalues and a defines some percentage. (Mardia *et al.*, 1982)

(iii) choosing the maximum q so that $\min_{q} \{d_q\} \geq \bar{d}$ (Kaiser's criterion), where d_q is the observed q-th ordered eigenvalue and $\bar{d}$ is the average of all the eigenvalues. (Mardia *et al.*, 1982)

(iv) choosing q so that the test of the hypothesis H: $\lambda_q = \lambda_{q+1} = \cdots = \lambda_p$ fails (known as isotropy test, Barlett's sphericity test, Barlett's test of homogeneity). This test is usually applied sequentially. (Mardia *et al.*, 1982, Flury, 1988, Green, 1978)

(v) choosing q using PRESS, predicted residual sum of squares (Wold, 1978, Krzanowski, 1987).

With the exception of Barlett's test, these tests are rather subjective. In the factor analysis literature there exist many more tests for determining the dimensionality of the factor space (see Malinowski and Howery, 1980, pp. 72-86). All are developed for the single-population situation and thus do not consider that smaller components may carry discriminating information. These above tests will be explored for a grouped-data example in Chapter 3.

The PRESS statistic is generally a measure of prediction error associated with a particular principal component model. The PRESS statistics for different models are then compared by an F-test or other ratio criteria. Wold (1976, 1978) and Krzanowski (1987) have both proposed using cross-validation in calculation of the PRESS statistics, although their application of cross-validation differs. Wold (1976) initially introduces what is called the single-cross procedure (Stone, 1974) followed by a double-cross procedure (1978). Krzanowski has suggested that only a single element of each observation be withheld for the cross-validation.

Measurement of response-curve data can be costly and principal component analysis can also be used to identify redundant variables. It is unknown how the methods of variable elimination using principal component analysis would perform for highly correlated response curve data. The method of elimination based on removing the variables most heavily weighted in the low-order components is likely to be unstable and yield results incompatible with instrumentation (since many instruments only work for equal interval sizes). Furthermore, in the presence of group structure the choice of estimator for Σ would probably affect variable exclusion, and either help or hinder discrimination.

Supervised Learning

If the data has distinguishable groupings, it is likely the researcher will want to develop discrimination rules not only to gain a better understanding of the underlying data structure, but to determine if more or fewer measurements (variables) or observations (sample size) should be made and to allocate future unknown observations. With groups known *a priori*, one of the most common and simplest methods for classification is linear discriminant analysis. The wide availability of statistical software (SAS, SPSS, Statgraphics, BMPD, etc.) and the procedure's simplicity contribute to its popularity, not necessarily its appropriateness for a particular problem.

General Methodology -- Linear Discriminant Analysis

Suppose there are g distinct groups of r-dimensional populations each with an associated probability density $f_j(\cdot)$ and it is desired to classify an unknown observation $\mathbf{x} = (x_1, x_2, \cdots, x_r)$ based on its measurements. Discriminant analysis is the development of rules separating $\mathbb{R}^r$ into disjoint regions such that any point $\mathbf{x}$ is assigned to the group k with the highest probability of occurence in the region containing $\mathbf{x}$, that is, k such that $f_k(\mathbf{x}) = \max_j f_j(\mathbf{x})$.

If the populations are known to be normally distributed with means $\boldsymbol{\mu}_j$ and common covariance Σ ($\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$), the maximum likelihood discriminant rule allocates $\mathbf{x}$ to the group which minimizes the square of the Mahalanobis distance

$$D^2 = \min_j (\mathbf{x} - \boldsymbol{\mu}_j)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) . \tag{2.7}$$

Note this requires Σ to be nonsingular; if Σ is singular then this rule must be modified by replacing the inverse of Σ with a generalized-inverse of Σ (Rao and Mitra, 1971, p. 204). This rule establishes linear boundaries separating the groups; if the populations have different covariance structures ($\Sigma_i \neq \Sigma_j$ for some $i \neq j$), then the boundaries should be quadratic since the intersecting contours of the covariance structures will be nonlinear.

Suppose the population parameters, $\boldsymbol{\mu}_j$, $j = 1, 2, \cdots, g$, and Σ, are unknown, but n_j observations are observed from each population j, $j = 1, 2, \cdots, g$. The sample maximum likelihood discriminant rule is the maximum likelihood discriminant rule with the unbiased estimators $\bar{\mathbf{x}}_j$ and

$$\mathbf{S} = \frac{\sum_{j=1}^{g} (n_j - 1)\, \mathbf{S}_j}{n - g} \tag{2.8}$$

of $\boldsymbol{\mu}_j$ and Σ, respectively, substituted into (2.7). S is called the pooled or within-group covariance matrix. For the special case $g = 2$, the sample maximum likelihood rule simplifies, classifying an observation $\mathbf{x}$ as coming from population 1 if and only if

$$\mathrm{W}(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \, \mathbf{S}^{-1} \left(\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right) \quad > \quad 0. \tag{2.9}$$

See Mardia et al. (1982).

Linear discriminant analysis provides a discriminating model of the data; that is, it gives us a representation of the data in a way such that the ratio of the between-group variation to the within-group variation is nearly maximized for each pair of groups. In fact, this is the criterion for Fisher's linear discriminant functions, a heuristic, nonparametric approach to discriminant analysis. The sample maximum likelihood discriminant rule for multinormal populations with equal covariance matrices and Fisher's linear discriminant functions will be similar if the sample group means are nearly collinear; they will be equal if the sample group means are exactly collinear (Mardia *et al.*, 1982). Prediction of the original observations (excluding the classification variable) from a discriminating model is unlikely to perform well; another method of modeling should be used for this purpose.

Alternatives to Linear Discriminant Analysis

Linear discriminant functions are the basis of Nilsson's linear learning machine (LLM), a supervised learning method of pattern recognition, used by chemists and others (Nilsson, 1965). Other supervised learning techniques common in the chemical literature include K-Nearest Neighbors (KNN),

SIMCA, and the more generalized Bayes classification analysis, of which linear discriminant analysis is a special case. Neural networks, which have evolved from Nilsson's LLM, are gaining popularity as pattern classifiers as software availability increases.

SIMCA is an acronym for many things including Soft Independent Modeling of Chemical Analogy, Statistical Isolinear Multiple Components Analysis, SIMple Classification Program, etc. See Wold and Sjostrom, 1977, and Kowalski and Wold, 1982. SIMCA was put forth by Wold (1976) when he described the suitability of separate principal component models for purposes of modelling and classification for applications in chemistry and biology. Others in pattern recognition had proposed disjoint principal component models for classification, but it was Wold who brought the idea to the chemical community. For each group a separate principal component analysis is performed; a new observation is then classified based on smallest residual fit. No relationships between groups are assumed. SIMCA is very similar to the subspace method used by Parkkinen and Jaaskelainen (1987) for their classification of transmittance curves and reflectance curves.

Data Reduction Prior to Classification

Besides prediction, other reasons that modeling should precede classification analysis are that response curve data tend to be highly correlated and that the ratio of variables to observations may be greater than one (Hoogerbrugge *et al.*, 1983). Applying linear discriminant analysis to this type of data will result in unstable discriminant function estimates since the discriminant function depends on the inverse of Σ. Rao and Mitra (1971) address this problem by suggesting that the null space of Σ be considered.

Some authors have promoted the use of more robust estimates of Σ in the linear discriminant function, usually based on the assumption that insufficient sample sizes relative to dimensionality produces poor results (and thus outliers are overly influential, see Peck and Van Ness, 1982, Randles *et al.*, 1978). Friedman's regularized discriminant analysis (1989) is also established along these lines.

Cheng *et al.* (1992) suggest a rank decomposition method for addressing the problems of singularity of the sample covariance matrix in the calculation of Fisher's discriminant functions. They compare this to generalized-inverse and perturbation methods. Biscay *et al.* (1990) introduce modified Fisher's Linear Discriminant Analysis which uses a metric that combines Euclidean distance (in principal component space) and Mahalanobis distance (in space orthogonal to the principal component space).

Kshirsagar *et al.* (1990) outline two-stage discriminant analysis, which divides variables into smaller groups such that sample size is greater than the number of variables. They calculate discriminant functions on these smaller groups, then combine them for an overall discriminant function. One problem in applying this method to spectral data would be the division of the variables into smaller groups. However, the authors do provide a good discussion of the application of principal component analysis prior to linear discriminant analysis.

Discriminant Analysis Error Rates

The maximum likelihood discriminant rule is actually a special case of the Bayes discriminant rule when prior probabilities of classification are equal. In this context, posterior probabilities of classification and misclassification associated with allocating a random observation to a particular population can be discussed. For further information about the Bayes discriminant rule see Mardia *et al.* (1982). Posterior probabilities of misclassification, also referred to as error rates, help judge the effectiveness of classification or discriminant rules. They are estimated by various techniques including resubstitution, cross-validation, and bootstrapping, among others. Commonly, the probabilities of misclassification for the populations are assumed equal and an overall estimate of the probability of misclassification is reported as a general measure of the performance of the discriminant functions.

Resubstitution estimates the probability of misclassification for a population by the proportion of observations from that population which when substituted into the sample discriminant rule are allocated to another population. The resubstitution estimate is called an apparent error rate. This method tends to be overly optimistic, that is, biased towards zero.

Cross-validation and bootstrapping are sample re-use methods; they calculate the estimates by using portions of the original sample. Cross-validation repeatedly divides the original sample into two exclusive groups. For one group, the training set, the discrimination rule is calculated, and then the observations in the second group, the evaluation set, are classified by this discrimination rule. The observations in the evaluation set then become members of the training set, while a set of observations in the training set that have yet to be classified become the new evaluation set. The discriminant rule is re-calculated on this new training set, and the members of the new evaluation set are allocated according to this rule. This process of evaluation is repeated until all observations in the sample have been classified. The estimates of the posterior probabilities of misclassification are calculated by the percentages of observations that have been misallocated. When the second group is of size one, this method is referred to as the leave-one-out method.

Subtracting from the resubstitution estimate an estimate of its bias gives the bootstrap estimate. The estimate of bias is obtained by resampling: many samples, each of size n, are taken from the original sample and for each an estimate of the bias is calculated; these estimates are averaged over all the samples to get an overall estimate.

Error rates, or probabilities of misclassification, are computed for classification rules to determine the effectiveness of a particular rule for separating groups in a data set and to compare among the different rules. The sample re-use methods of cross-validation and bootstrapping are generally acknowledged to be better estimators of error rates than resubstitution (Snapinn and Knoke, 1988); however, these procedures are computationally intensive, especially with a large number of variables and/or a large number of observations. Although bootstrapping appears to be the better estimator, cross-validation is computationally less expensive in terms of time and complexity and, thus, is often the method of choice. The bootstrap estimator is due to Efron (1979), while the leave-one-out cross-validation estimator is credited to Lachenbruch (1967) and Lachenbruch and Mickey (1968). Snapinn and Knoke (1988) provide a nice review, as well as comparison, of many misclassification error rate estimators.

When performing discrimination analysis on the principal components it should be noted that components associated with the most variation are not necessarily the best discriminators. Common forward variable selection procedures in discriminant analysis are outlined in Habbema and Hermans (1977). These include Wilk's Lambda (U statistic), the F statistic, and maximal estimated correct classification rate. Variable selection will be explored for the spectral data set with Wilk's Lambda and the minimal estimated error rate, since the U and F statistics are essentially equivalent for variable selection.

Rank-One Methods

Since the leave-one-out cross-validation estimator of the linear discrimination's error-rate is commonly used by practioners (because of its properties and wide availability), it was desired to investigate its appropriateness for data that has previously been reduced by principal component analysis. The principal component transformation, however, uses all of the data, while the cross-validation principle is to remove the influence of subsets of observations. To prevent the introduction of bias into the estimator, then, the principal component transformation should be recalculated prior to the calculation of the discriminant functions.

As noted the leave-one-out cross-validation procedure results in a rank-one update in the pooled covariance matrix. This update in the pooled covariance matrix when an observation is deleted has been noted recently by Friedman (1989) in his treatise of regularized discriminant analysis, although earlier references are likely to be found.

If all components up to a certain number are used in the discriminant analysis, the process of re-evaluation of the principal component model minus one observation may be sped up with a sequential procedure developed by H. Wold (1966) in addition to the rank-one methods by Bunch *et al.* (1978) alluded to in Chapter 1. The structure of the rank-one procedure allows the recalculation of components independent of each other; components unnecessary for discrimination do not have to be recomputed. This independence also allows for parallelism, which has been exploited by Cuppen(1981) and Dongarra and Sorensen (1987), in their treatment of the symmetric eigenvalue problem.

Surely it will outperform sequential methods in parallel processing environments.

Many researchers have taken advantage of the rank-one procedure of Bunch *et al.* In statistics, as mentioned earlier, Krzanowski (1983, 1986, 1987) and Eastment and Krzanowski (1982) have used the algorithm in the computation of the PRESS statistic for determining the number of components to keep and for variable selection in principal component analysis. In signal processing, where data covariance matrices are modified adaptively for time-varying signals, DeGroat and Roberts (1990), Schreiber (1986), and Yu (1991) have made use of the Bunch *et al.* rank-one updating procedure; only zero-mean processes were considered. And in chemistry, Hemel and van der Voet (1986) applied the rank-one procedure in their development of software of multivariate classification techniques in chemistry. Specifically, the procedure was used for the evaluation of the leave-one-out cross-validation estimators in SIMCA, but the authors do not mention how the algorithm is actually implemented.

Chapter 3
APPLICATION OF CROSS-VALIDATION ESTIMATION

The data to be analyzed in this thesis, spectral reflectance curves of Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco] veneer clear wood, were obtained as part of a project to develop better color scanning systems for the wood products industry. A wood surface's spectral reflectance is the primary element affecting its color as described in Brunner *et al.* (1990). The researchers on the project have studied the color of Douglas-fir veneer in three-dimensional color spaces, such as RGB, Yxy, Lab, etc., that are common to optical scanning systems (Brunner *et al.*, 1990, Brunner *et al.*, 1992, Maristany *et al.*, 1991, and Maristany *et al.*, 1992). To remove the influences of lighting and camera sensors of the machine vision system on the colors observed in the three-dimensional color spaces, it is necessary to study spectral reflectance curves.

Establishing a database (a large collection) of observed curves will aid in the development of more effective optical scanning systems by allowing researchers to see the affects in three-dimensional color spaces of hypothetical lights, camera sensors, and filters. In addition, modeling the curves will give a more basic understanding of the *true* color of wood.

Clear wood, that is, wood devoid of defects, such as knots, pitch streaks, pitch pockets, fungal stain, and other irregularities, usually exhibits distinct growth traits. As a tree ages the inner cells of the stem eventually die, forming heartwood, while new, living cells grow around the outer circumference of the stem, forming sapwood. The heartwood cells are chemically and physically different than the sapwood cells; the former containing extractives (various phenolic substances) and the latter containing sugars, starches and fats. The extractives give Douglas-fir heartwood a characteristic pinkish color, while the Douglas-fir sapwood is a less distinctive yellow-white.

Tree growth is fastest in the spring, producing large, relatively thin-walled cells (earlywood), while smaller, thick-walled cells are formed in the late summer (latewood). These alternating bands of thick and thin walled cells are

called growth rings, or annual rings, and are visibly evident in the cross-section of most wood species. The difference in cell wall thickness between the earlywood and latewood is thought to affect brightness (luminance) as opposed to true color (chromaticity), especially in sapwood. This would be seen as an overall shift in the spectral response curve rather than a change in shape. In most veneers the transition between earlywood and latewood is more gradual and less obvious because veneers are taken from the tangential surface. A more general discussion can be found in Panshin and de Zeeuw (1980) and a U.S. Forest Products Lab report on wood as engineering material (1974).

From a preliminary experiment, 50 response-curves (spectral reflectance curves) were obtained from each of sapwood earlywood, sapwood latewood, heartwood earlywood, and heartwood latewood of a single Douglas-fir veneer specimen. The locations of the observations on the specimen were randomized. Typical curves are illustrated in Figure 3.1(A)-(D). Theoretically, the response-curves are limited to values between zero and one, but some of the observed curves exceed one as a result of excess interface reflection within the measurement process. The lighting that produced this has subsequently been corrected but the updated data has yet to be gathered. It is anticipated the actual curves will exhibit characteristics very similar to those in this preliminary data set, so our analysis procedure should apply.

Response-curves are not constrained to be smooth, although the spectral response curves of naturally occurring materials are typically smooth, that is, continuous-looking (MacAdam, 1981). The smoothness exhibited by these curves invites the possibility of characterizing them by other means than principal components, such as low order polynomials or Fourier series. Representing spectral reflectances by finite dimensional linear models is supported by many researchers, including Maloney (1986), Buchsbaum and Gottschalk (1984), and Stiles *et al.* (1977). In their work, these researchers are interested in developing models that approximate all natural objects well, thus, they do not want to completely rely on an empirical model. In contrast, empirical orthogonal functions seem very suitable for the goals of the project from which this data arose. (In fact the use of polynomial functions and Fourier series were investigated as bases for this particular data set, but they

did not perform as well in modeling the data set as the principal component models.)

Descriptive Statistics

Although earlywood and latewood occur within heartwood and sapwood, the groups will be treated individually; nested designs are beyond the scope of this thesis. Define the data matrix for the j-th group, $j = 1, 2, 3, 4$ for sapwood earlywood, sapwood latewood, heartwood earlywood and heartwood latewood, respectively, as

$$\mathbf{X}_j = [\mathbf{x}_{1j} \quad \mathbf{x}_{2j} \quad . \quad . \quad . \quad \mathbf{x}_{n_j,j}]'$$

where n_j is the number of observations for this group (as in (2.4)). The p-dimensional vector $\mathbf{x}_{ij}$ defines the i-th response curve from the j-th group with each element of the vector corresponding to the observed percentage of spectral reflectance at a particular wavelength. In this chapter, p is 71, as the spectral reflectance, at a particular location of the piece of wood, was measured every ten nanometers over the range 400 to 1100 nm.

Several observed spectral reflectance curves for each wood type are displayed in Figures 3.1(A)-(D). Douglas-fir wood tends to be pinkish-red; this is confirmed by the general shape of the curves which show less blue light reflectance (at the lower wavelengths) and more red light reflectance (at the wavelengths 600-700 nm). The variations within each group are significant, and are thought to be large differences in brightness and not color, as color is determined by the shape of the curve (Brunner *et al.*, 1990). Researchers sometimes remove this excess variation by normalizing each observed response-curve (Tominaga and Wandell, 1990, Tominaga, 1991). However, in developing a database, it is of interest to characterize all of the variation associated with the data set. Decomposition into variance components, although beyond the scope of this thesis, would be desirable for this type of data.

The j-th group mean and group sample covariance are given by (2.6) and (2.5), respectively. Figure 3.2 shows the mean response vectors for each group. The general shapes between sapwood and heartwood are rather distinct; in the visible light range (approx. 400-700 nm) the heartwood is less reflectant than the sapwood owing to its darker color. The heartwood latewood does

appear to reflect more red and less blue than the heartwood earlywood suggesting there is a color difference between them; this would be visualized as the heartwood latewood having a more reddish appearance. The shapes of the sapwood earlywood and sapwood latewood mean curves of Figure 3.2 are less distinct from each other, with the exception of the sapwood latewood reflecting more red and near-infrared light than the earlywood.

The standard deviations of spectral reflectance at each wavelength are presented in Figure 3.3. Albeit these figures do not illustrate the relationships between spectral reflectances at different wavelengths, they show the commonality between the variance structures of the groups of wood. The standard deviations show an increase and then level off between 4 and 6 with several peaks near the same wavelengths.

Constant variance is a desirable trait as no spectral reflectance at a particular wavelength or band of wavelengths will be overly influential in the decomposition of the covariance matrix. Although this can be achieved by standardization with decomposition of the correlation matrix, a simple one-to-one transformation of the data vector $\mathbf{x}_{ij}$ would be preferred, especially since the spectral reflectance is measured on the same scale across wavelength. Figures 3.4(A) and (B) demonstrate the effects of simple transformations, such as the logarithm (3.4A) and square root (3.4B), of the response-curves (non-percentage data) on the standard deviations. In addition other transformations of the curves were explored, but in no case did the variance achieve nearly complete constancy across the spectrum.

Test for Equality of Covariance Matrices

Several tests for the equality of covariance matrices were briefly described in the literature review of Chapter 2. The M-test for equality of covariance matrices among populations, or any of its approximations, should not be used in this situation as the sample covariance matrix for each wood group is singular. This is a result of having fewer observations in each group (50) than observed variables (71).

The graphical procedure based on the traces of the mean-corrected sum-of-products matrices falling on a straight line indicates the covariances are somewhat similar. Using the actual reflectances (as opposed to the percentage

reflectance) to avoid problems of scale, the traces of the mean-corrected sum-of-products matrices are 9.3512, 11.9040, 11.3450, and 8.8330 for heartwood earlywood, heartwood latewood, sapwood earlywood, and sapwood latewood, respectively. These are ordered and plotted against quantiles of the gamma distribution with an estimated shape parameter 63.0 and scale parameter 6.1 (see Figure 3.5(A)). The graph illustrates that the ordered values fall relatively close about a straight line.

A similar graphical procedure based on the geometric mean of the positive eigenvalues of the mean-corrected sum-of-products matrices also illustrates the similarity of the covariance matrices. The geometric means of positive eigenvalues of the mean-corrected sum-of-products matrices are 1.0289, 1.1937, 0.9471, and 1.0791 for heartwood earlywood, heartwood latewood, sapwood earlywood, and sapwood latewood, respectively. In Figure 3.5(B), the estimated gamma distribution has a shape parameter of 143.4 and scale parameter of 135.0. This graph illustrates that the ordered values fall almost on a straight line giving a stronger indication of the similarity of the covariance matrices.

The preceding graphical procedures compare only a part of the covariance structures, that is, they examined only the magnitudes the of eigenvalues without looking at the directional aspects (eigenvectors) of the structures. It should also be kept in mind that only four points are being compared.

Principal Component Analysis

Although only 50 spectral reflectance curves were observed per group, it is still possible to decompose the individual group covariance matrices $\mathbf{S}_j$ of (2.4) as would be done following SIMCA. It should be noted, however, that in the context of modeling, the number of parameters being estimated far outweighs the amount of independent information available. It is probably wise not to place too much emphasis on these analyses.

Eigenstructure plots (i.e., scree plots and eigenvector plots) for each wood group are given in Figures 3.6 through 3.9. The eigenvalues and eigenvectors, although not identical for each group, reveal likenesses such that the violation of the assumption of equal covariance matrices among groups may

be unimportant. In fact, the procedures for comparing population covariance structures suggested that the eigenvalues are similar (see preceeding section). If all the information is pooled together, it may provide something more useful than if the groups were analyzed separately.

Geometrically, the eigenvectors can be compared using Krzanowski's between-groups comparison of principal component subspaces (Krzanowski, 1979). Comparing the principal component subspaces of sapwood earlywood and sapwood latewood it appears that the first four to seven eigenvectors are very similar (see Krzanowski, 1979, for discussion of similarity). This says that the major sources of variation appear to be common among the two groups. The angle between the first eigenvectors is 4.8^o (in 71-dimensional space); Figures 3.6(B) and 3.7(B) show that they do have similar characteristics. If the principal component subspaces of two dimensions are compared, they come within 2.1^o of each other. With seven-dimensional subspaces, the subspaces are within 0.3^o of each other. Higher dimensional principal component subspaces are closer, but the eigenvectors associated with the smaller variations begin to deviate from each other.

Krzanowski extends this comparison of subspaces when there are more than two groups, such as when comparing the four groups sapwood earlywood, sapwood latewood, heartwood earlywood, and heartwood latewood. In this situation, six-dimensional principal component subspaces appear to capture similar major sources of variation among the four groups. With six-dimensional subspaces the maximum angle between the first eigenvectors of the groups and the vector defined as the closest to the four subspaces is 0.4^o. The maximal deviation on the sixth eigenvector is 26.0^o.

If all of the group covariance matrices are pooled together, part of the resulting eigenstructure is depicted in Figure 3.10; the eigenstructure based on the pooled covariance matrix for the two groups of sapwood is illustrated in Figure 3.11. The spectral reflectances of the sapwood earlywood and sapwood latewood groups are much more difficult to distinguish as suggested by the mean spectral reflectance curves of Figure 3.2 and as analyzed later in this chapter. As this usually means more borderline classification probabilities, these wood types will be used to illustrate methods for classifying between two groups.

How well the principal component model based on common population covariance matrices for four groups depicts the observed spectral reflectance curves is displayed in Figures 3.12(A) and (B). For each graph, the solid curve is an observation, the dashed line represents the observation as predicted from the first principal component, the dotted line represents the observation as predicted from the first two components, and the dash-dot line represents the observation as predicted from the first three components. In these two instances, it appears that at least two components may be necessary to adequately predict the response curves. See the following section on the PRESS method for retaining components for a numerical indication of prediction.

The variables, spectral reflectances across the spectrum, are all measured on a common scale and do not show excessive variation across the spectrum. Thus, it would be common practice to just perform principal component analysis on the covariance matrix (either the pooled or the total). For curiosity's sake, however, the correlation matrix from the pooled covariance matrix was also decomposed. Notice in Figure 3.13(A) the likeness of the first three eigenvectors of the correlation matrix to zero-, first-, and second-degree polynomials.

In the remainder of this thesis, principal component analysis will be performed on the unbiased estimate of a common population covariance matrix, that is, the pooled covariance matrix of (2.8).

Retaining Components

The guidelines for choosing the number of meaningful principal components were outlined in Chapter 2. The results for the spectral reflectance data set are described below.

Cattell's scree tests for determining the number of meaningful principal components in describing all wood types and the sapwood groups are basically illustrated in Figures 3.10(A) and 3.11(A), respectively. They are repeated in Figures 3.14(A) and (B) to show the relative contributions of the ordered components to the total variation. Since the first component contains over 90% of the total variation while the remainder of the components contributions decrease geometrically, these plots indicate one component should be sufficient in explaining the variation.

Below Table 3.1 gives the number of components necessary to attain a certain percentage of the total variation along with the actual percentage retained for all groups and for the two sapwood groups.

Table 3.1. Number of Components Retained Based on Percentages.

	90%	95%	99%	99.9%
All wood	1 (92.35%)	2 (96.84%)	4 (99.28%)	8 (99.93%)
Sapwood	1 (93.54%)	2 (97.46%)	4 (99.54%)	7 (99.93%)

Under Kaiser's criterion, the number of eigenvalues that exceed the average eigenvalue is three for both the all wood situation and the sapwood situation. The average eigenvalue in the decomposition of the pooled covariance for all groups is 29.774 and in the decomposition of the pooled covariance matrix for sapwood is 29.004. Although difficult to visualize where the average eigenvalues fall because of the scale of the graphs, Figures 3.10(A) and 3.11(A) display the ordered eigenvalues.

Application of the PRESS statistic following the leave-one-out scheme as outlined in Chapter 2 results in ten components being retained when all wood groups are considered and eight components when the sapwood groups are considered alone. Let $\mathbf{x}_{ij}$ be the i-th response curve observed from the j-th group. If $\widehat{\mathbf{x}}_{ij}$ is the corresponding predicted response curve from a q-dimensional principal component model, then the PRESS statistic can be given by

$$\text{PRESS}(q) = \frac{1}{np} \sum_{j=1}^{g} \sum_{i=1}^{n_j} \| \widehat{\mathbf{x}}_{ij} - \mathbf{x}_{ij} \|^2$$

where g is the number of groups and $n = \sum_{j=1}^{g} n_j$ is the total number of observed response curves. Notice this is the mean-squared-error. Table 3.2 summarizes

these average squared deviations based on progressively larger principal component models.

Table 3.2. Summarization of PRESS(q).

	PRESS(q)	
q	All Wood	Sapwood
1	113.15	88.67
2	25.44	48.26
3	14.88	26.99
4	5.38	7.70
5	4.00	5.77
6	3.78	4.94
7	3.10	2.80
8	2.99	2.39
9	1.92	2.36
10	0.49	0.91
11	0.48	0.80
12	0.48	0.79
13	0.40	0.74
14	0.25	0.39
15	0.17	0.28

The test of sphericity based on Barlett's approximation results in about 67 components being retained for the four groups and 66 components for two groups. A problem with this test is that the geometric mean, which is used in the denominator of the test statistic, is nearly zero.

The results of the methods for choosing the number of meaningful components vary quite a bit considering the range of total variation explained by them. As is common with many analyses, the researcher is essentially left with the decision of choosing the number of components that he or she *feels*

adequately represents the data. Between two and four principal components probably are sufficient; with four components predictability is substantially improved. Several papers discussing principal component models for spectral data suggest that between five and eight dimensions are best (Maloney, 1986, Parkkinen, *et al.*, 1987), but these researchers are concerned with a much broader class of spectral reflectance curves.

The number of components retained for each of the preceeding methods is summarized in Table 3.3.

Table 3.3. Number of Components Retained.

Test	All Wood	Sapwood
Cattell's Scree	1	1
95%	2	2
99%	4	4
99.9%	8	7
Kaiser's Criterion	3	3
PRESS	10	8
Test of Sphericity	67	66

Linear Discriminant Analysis

As stated in the Literature Review, calculation of the discriminant functions on highly correlated data could result in meaningless discriminators since their calculation depends on the inverse of the pooled sample covariance matrix. Figures 3.15(A) and (B) show the discriminant function for sapwood earlywood in both the four group and two group instances. Interpretation of these plots is difficult: it could be said the functions are contrasting neighboring values or the functions look like noise (as a result of unstable estimates). Based on the magnitudes of the coefficients, the method of calculation and the relationship of spectral reflectances at neighboring wavelengths, it is likely to be the latter. In fact, the condition numbers of the

pooled sample covariance matrices for the four wood types and the two sapwood types are 6.6×10^6 and 3.5×10^7, respectively. These were calculated by taking the ratio of the largest singular value to the smallest.

If the data is so highly correlated that the pooled sample covariance matrix is near-singular or ill-conditioned, before attempting discriminant analysis the data should be reduced in some manner, either by eliminating redundant variables or transforming the set of variables. A similar situation arises when insufficient data has been gathered; however, because of the nature of this data it seems that obtaining more observations will not remove the ill-conditioning of the estimator of covariance. This was mentioned earlier, and for the situation where the variables have an ordered relationship, the principal component transformation seems to be a logical choice. After the transformation and data reduction, discriminant analysis is performed on the new data set. For example if three principal components are chosen for modeling the data, the discriminant functions can be calculated based on these components.

In determining the effectiveness of such a procedure, it is a good idea to first look at scatterplots of several of the principal components to see if pairs of components may have separation of the groups. Figures 3.16(A)-(E) show scatterplots for several pairs of the first five principal components with each point being labelled with the corresponding wood type. From these plots, it can be seen that the second component is a very good discriminator, the third provides some separation between heartwood latewood and the other wood types, and the first component does not seem to provide any discriminating variation at all. The fifth component also provides good separation between the heartwood and sapwood, while the fourth component shows some slight separation between the heartwood and sapwood groups.

The discriminant functions for the four wood types based on the first five principal components is given in Table 3.4. Note these coefficients are based on the standardized principal components to help give some indication of which components are important for discrimination for each group. From the table, the coefficients show that the second and fifth principal components appear to be helpful in differentiating heartwood from sapwood, while the third principal component seems to provide additional separation between the heartwood

earlywood and heartwood latewood, especially in conjuction with the second or fifth components. Looking back at the eigenvectors of Figures 3.10(B) and 3.10(C), this suggests that the sapwood and heartwood differ in the blue-green region of the spectrum (around 500 nm) and in the higher wavelengths of the near-infrared region (1050-1100 nm) and that they also differ in the blue and red regions, and that the heartwood earlywood and heartwood latewood differ across the spectrum especially in the contrast of the red and low near-infrared regions (700-900 nm) to blue-green and higher near-infrared regions.

Table 3.4. Discriminant Functions for All Wood ($\times 10^{-1}$).

Principal Component	SE	SL	HE	HL
1	0.076	0.079	0.072	0.076
2	0.405	0.518	0.976	1.217
3	-1.202	-1.226	-1.017	-1.639
4	-3.958	-4.208	-3.318	-3.524
5	0.994	0.114	5.984	6.148

Using the leave-one-out cross-validation estimator of error-rate, the group-estimated misclassification rates are 0.22, 0.38, 0.06, and 0.08 for sapwood earlywood, sapwood latewood, heartwood earlywood and heartwood latewood, respectively. The overall estimate of misclassification is 0.185. With the sapwood alone the estimated error-rates are 0.22 for sapwood earlywood and 0.36 for sapwood latewood, and the overall estimate of misclassification is 0.29. The overall cross-validation error rate estimate for the sapwood raw data is 0.33, that is, using the raw data without prior reduction does result in somewhat worse classification. Table 3.5 provides a breakdown of the classifications when discriminant analysis is performed for all four wood groups using the first three principal components.

Table 3.5. Wood Classifications.

Actual Wood Type	Classified Wood Type			
	Sap Early	Sap Late	Heart Early	Heart Late
Sap Early	39	11	0	0
Sap Late	18	31	1	0
Heart Early	0	0	47	3
Heart Late	0	0	4	46

One problem encountered in arriving at these estimates is the time involved, since when an observation is removed the principal component transformation as well as the discriminating functions are recalculated. On a personal computer equipped with a 80386 processor and 80387 math coprocessor running at 25 MHz the total time for the four-group case (200 observations) was 57.9 minutes, while for the two-group case (100 observations) was 22.6 minutes. The number of floating point operations required was 998 million for the former case and 387 million for the latter case (see next section for discussion of floating point operations). This is summarized as the original method in Table 3.6 for the four groups and Table 3.7 for the two sapwood groups. Obviously faster alternatives, such as those mentioned in the objectives, are desirable.

Table 3.6. Algorithm Comparisons for Four Groups.

Method	Flops (million)	Time (minutes)	Mean Absolute Error ($\times 10^{-4}$)
Original	997.9	57.9	
Short-cut	6.3	2.2	3.7
Rank-one Update	38.2	6.0	0.0

Table 3.7. Algorithm Comparisons for Two Groups.

Method	Flops (million)	Time (minutes)	Mean Absolute Error ($\times 10^{-4}$)
Original	386.6	22.6	
Short-cut	4.2	0.5	4.2
Rank-one Update	15.9	2.4	0.0

To emphasize that the order of the principal components does not necessarily correspond to discriminating variation, discriminant analysis variable selection procedures were performed on a subset of components. If the first ten components are retained and forward stepwise variable selection is performed on them, the second, third and fifth components carry the most discriminating information in terms of minimizing the estimate of overall misclassification rate (in both the four and two group cases). In fact, using the second and third components for discrimination, as opposed to the first three, reduces the overall cross-validation estimate of error rate to 0.17 and 0.26 for the four groups and two groups, respectively. Forward stepwise variable selection using Wilk's Lambda results in the second, tenth, and first components being kept for the sapwood data set and the second, fifth, and third components for the all wood data set.

Costs of Calculation

From the preceding section it is apparent that faster alternatives are a necessity in obtaining these error-rate estimates. What follows is a general discussion of the number of floating point instructions and time costs involved in the calculation of the cross-validation error-rates estimates stated above; this will allow the introduction of other, faster options.

A floating point operation (flop) is an arithematic operation such as the addition, subtraction, multiplication, or division of two floating point numbers. The number of flops required by an algorithm is one measure of its performance; other measurements, such as time, may give a truer image of a

computer implementation of the algorithm. Of course, different implementations of the algorithm on different computers can have vastly different times associated with them, whereas the measurement of floating point operations should be rather consistent. See Golub and Van Loan (1989) for further discussion.

Suppose there are g groups from which n_j observations are gathered in each group, $j = 1, \cdots, g$, and that the total number of observations is $n = \sum_{j=1}^{g} n_j$. Further suppose that each observation is a p-vector, and that q principal components are chosen for discrimination. When one observation from group i is left out, the number of flops necessary:

1. to calculate the pooled covariance matrix is $(2n + 3g - 2)p^2 + (n-1)\,4p + 3g$. (See Appendix A.)

2. to compute the eigenvalue decomposition of the pooled covariance matrix as implemented in MATLAB (1989), the software used for this study, is approximately $9p^3$.

3. to calculate the principal components is $2npq$. (If the total covariance matrix is used, as opposed to the pooled covariance matrix, then steps 2 and 3 could be combined by computing the singular value decomposition.)

4. to estimate the posterior probability for the left-out observation based on the linear discriminant function is $2q^3 + (2n + 5g - 2)q^2 + (5n + 3g - 5)q + 9g - 1$. (See Appendix A.)

This must be calculated n times, once for each observation. The number of misclassified observations in step 4 gives the leave-one-out cross-validation estimate of the discriminant function's error-rate.

Since the intention of a principal component model is usually to allow for data reduction by elimination of noise, from a practical point of view, steps 1 through 3 could be computed just once and, then, step 4 calculated n times, once for each observation. This results in what is referred to as the short-cut estimator. By assuming the principal component model has no error, however, a small amount of bias is introduced. This will be discussed in Chapter 6. (Using this short-cut procedure, the number of flops necessary in step 4 can be reduced by using Barlett's identity (Bartlett, 1951).)

For the same situation with three components, as outlined in the prior section, the number of flops required to calculate the short-cut overall misclassification estimates of 0.185 and 0.3 is 6.3 million and 4.2 million for the all wood and sapwood cases, respectively. Only 130 and 32 seconds are needed to obtain these estimates, respectively. See Tables 3.6 and 3.7.

In Tables 3.6 and 3.7 the mean absolute errors between the classification probabilities of each observation for this short-cut method and the original method are listed. Overall, the short-cut method appears to approximate the probabilities rather well. As a consequence, the overall misclassification estimates of error rate are very close to the original estimates listed in the previous section, with a slight difference when the two sapwood groups were analyzed. (These calculations were based on six digits of accuracy.)

The number of flops in step 1 can be reduced by taking advantage of the fact that the pooled covariance matrix minus one observation can be obtained by some simple calculations from the pooled covariance matrix based on all the observations. This simple perturbation of the covariance matrix, which is more specifically known as a rank-one update in the covariance matrix, combined with the idea that a complete decomposition is unnecessary in step 2 allows for a relatively fast algorithm by Bunch *et al.* (1978) to be employed. The application of this algorithm in this context is fully described in the next chapter and summarized in Tables 3.6 and 3.7 as the rank-one method.

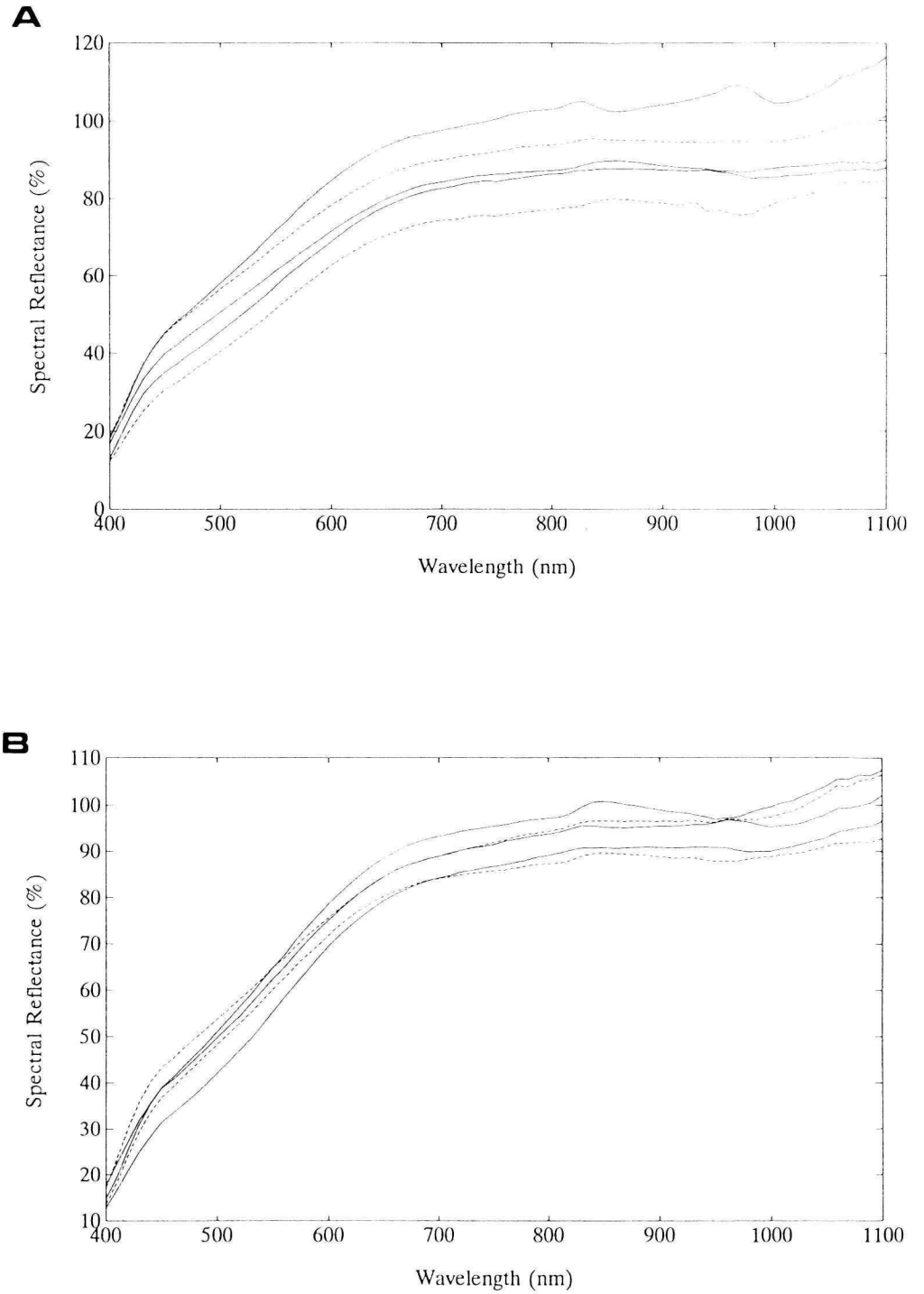
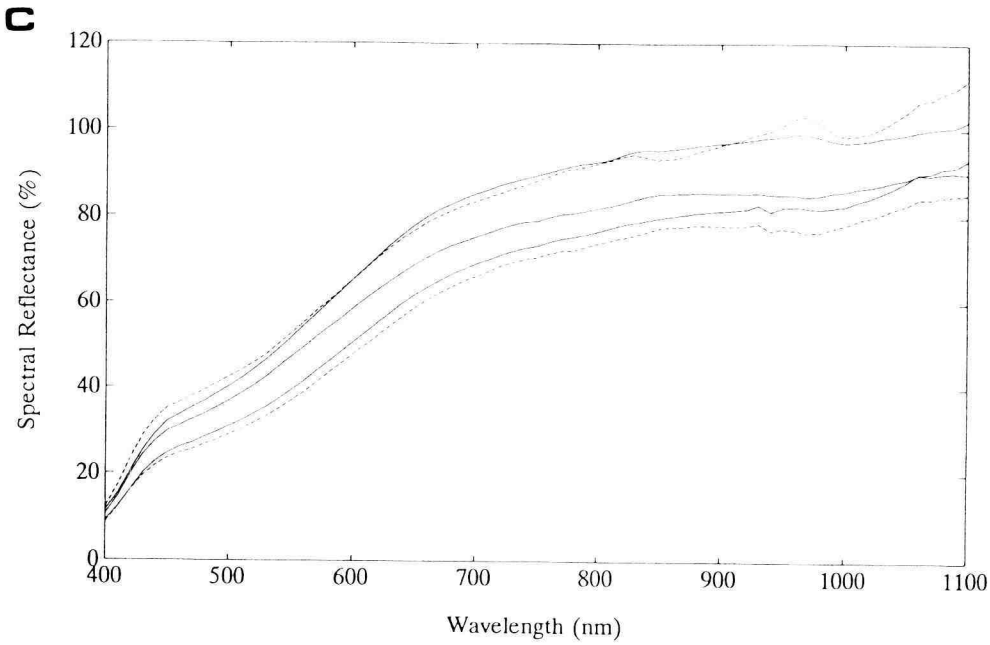
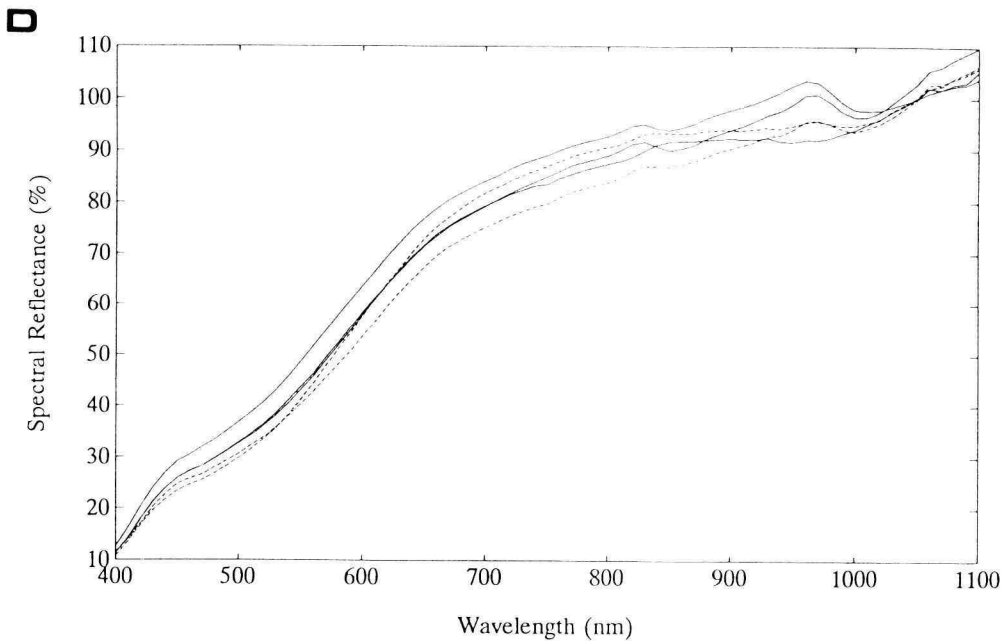


Figure 3.1. Spectral reflectance curves for five different sample sites of A) sapwood earlywood, B) sapwood latewood, C) heartwood earlywood, and D) heartwood latewood.

Figure 3.1. (continued)





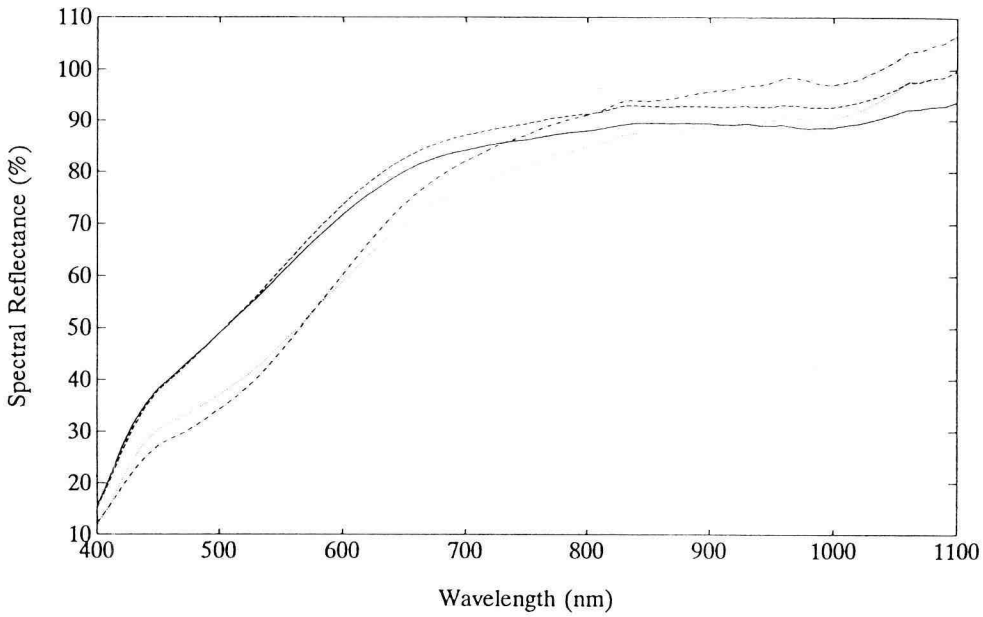


Figure 3.2. Mean percentage spectral reflectance curves by wood group. The solid, dashed, dotted, and dash-dotted lines are the mean curves for sapwood earlywood, and sapwood latewood, heartwood earlywood, heartwood latewood, respectively.

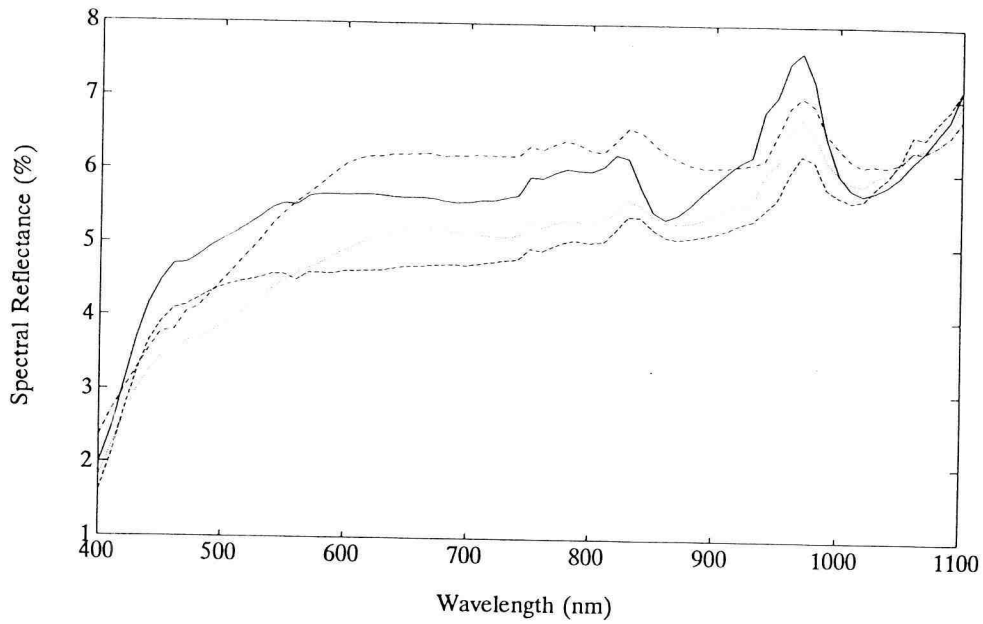


Figure 3.3. Standard deviations of spectral reflectance curves by wood group. The solid, dashed, dotted, and dash-dotted lines are the standard deviation curves for sapwood earlywood, and sapwood latewood, heartwood earlywood, heartwood latewood, respectively.

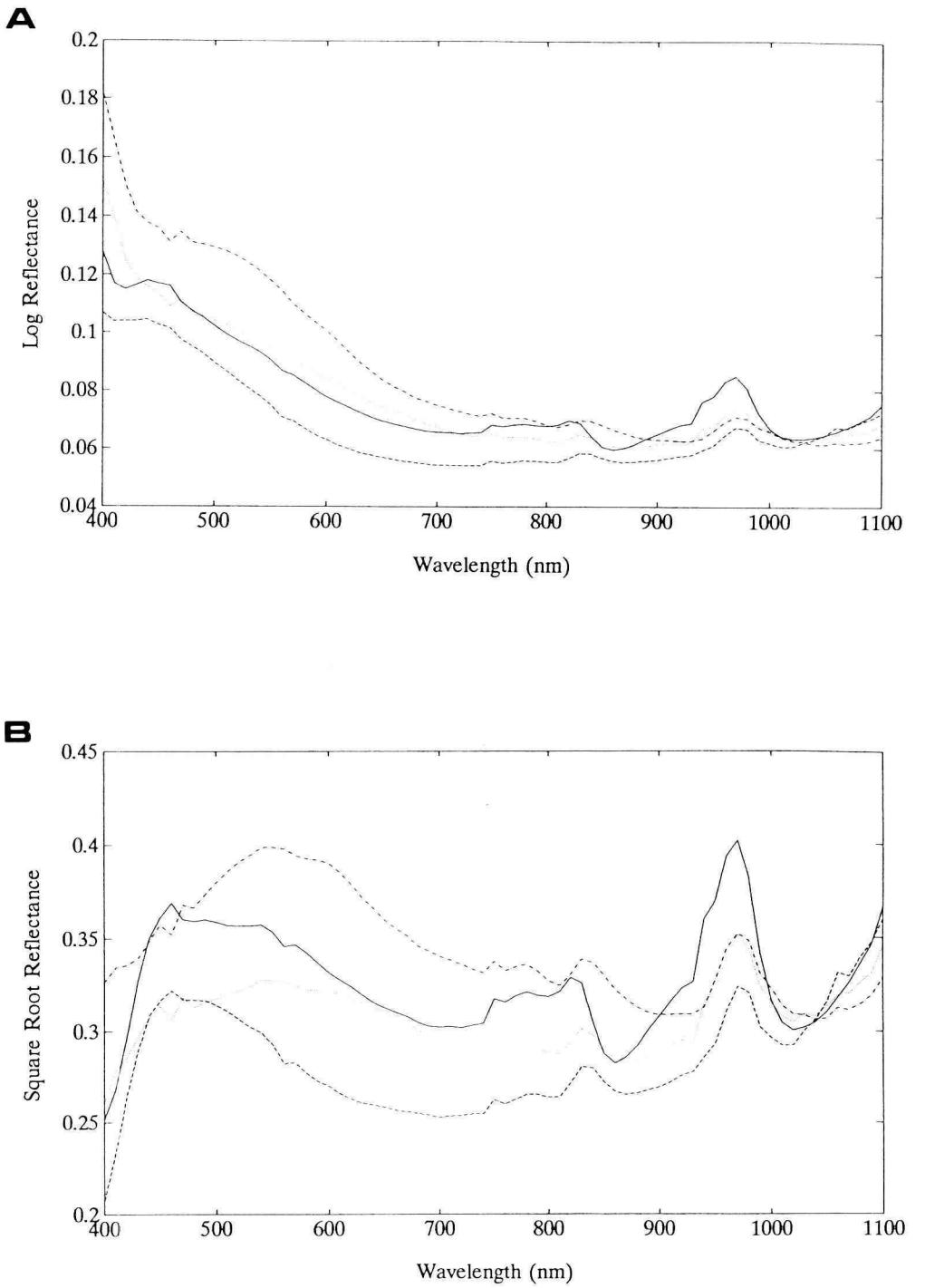


Figure 3.4. Standard deviations of transformed spectral reflectance curves by wood group. The solid, dashed, dotted, and dash-dotted lines are the standard deviations for spectral reflectance curves from heartwood earlywood, heartwood latewood, sapwood earlywood, and sapwood latewood, respectively, as transformed by A) logarithms and B) square-roots.

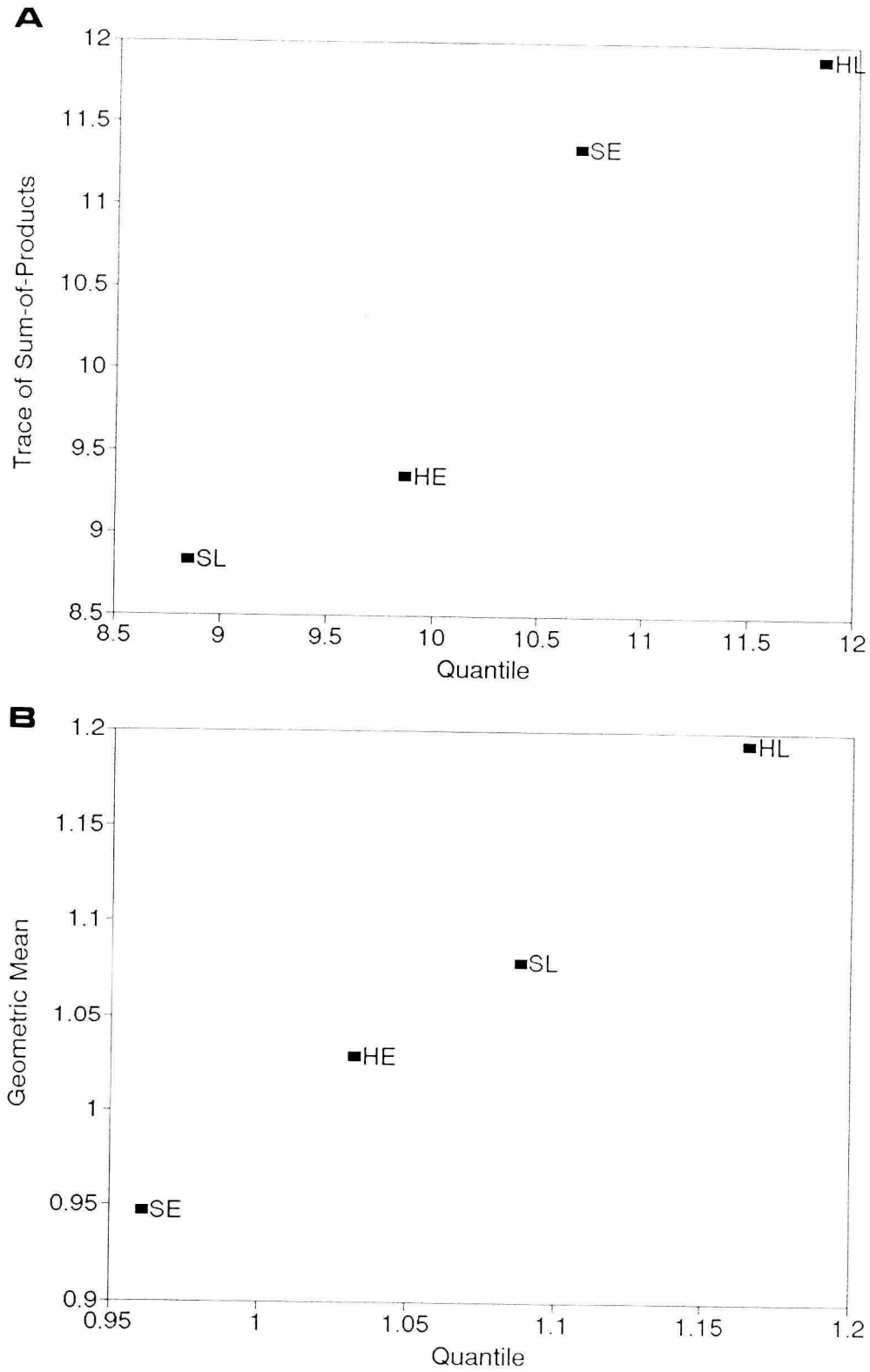


Figure 3.5. Graphical comparisons of the group covariance matrices using A) the traces of the groups' sum-of-products matrices and B) the geometric mean of the positive eigenvalues of the groups' sum-of-products matrices.

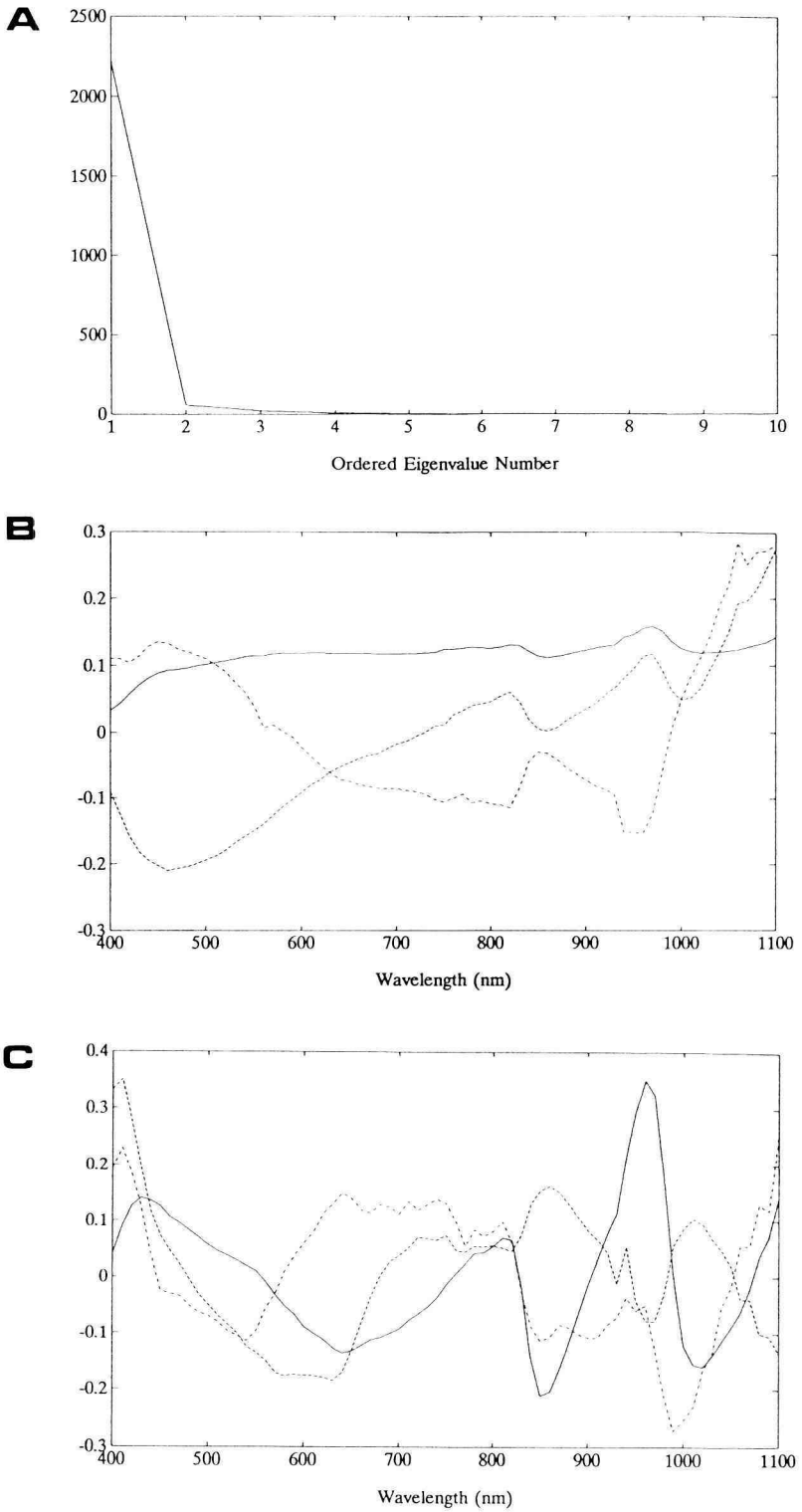


Figure 3.6. Part of the eigenstructure of the sapwood earlywood group sample covariance matrix. A) plots the first ten ordered eigenvalues, B) plots the first (solid line), second (dashed line), and third (dash-dotted line) eigenvectors, and C) plots the fourth (solid line), fifth (dashed line), and sixth (dash-dotted line) eigenvectors.

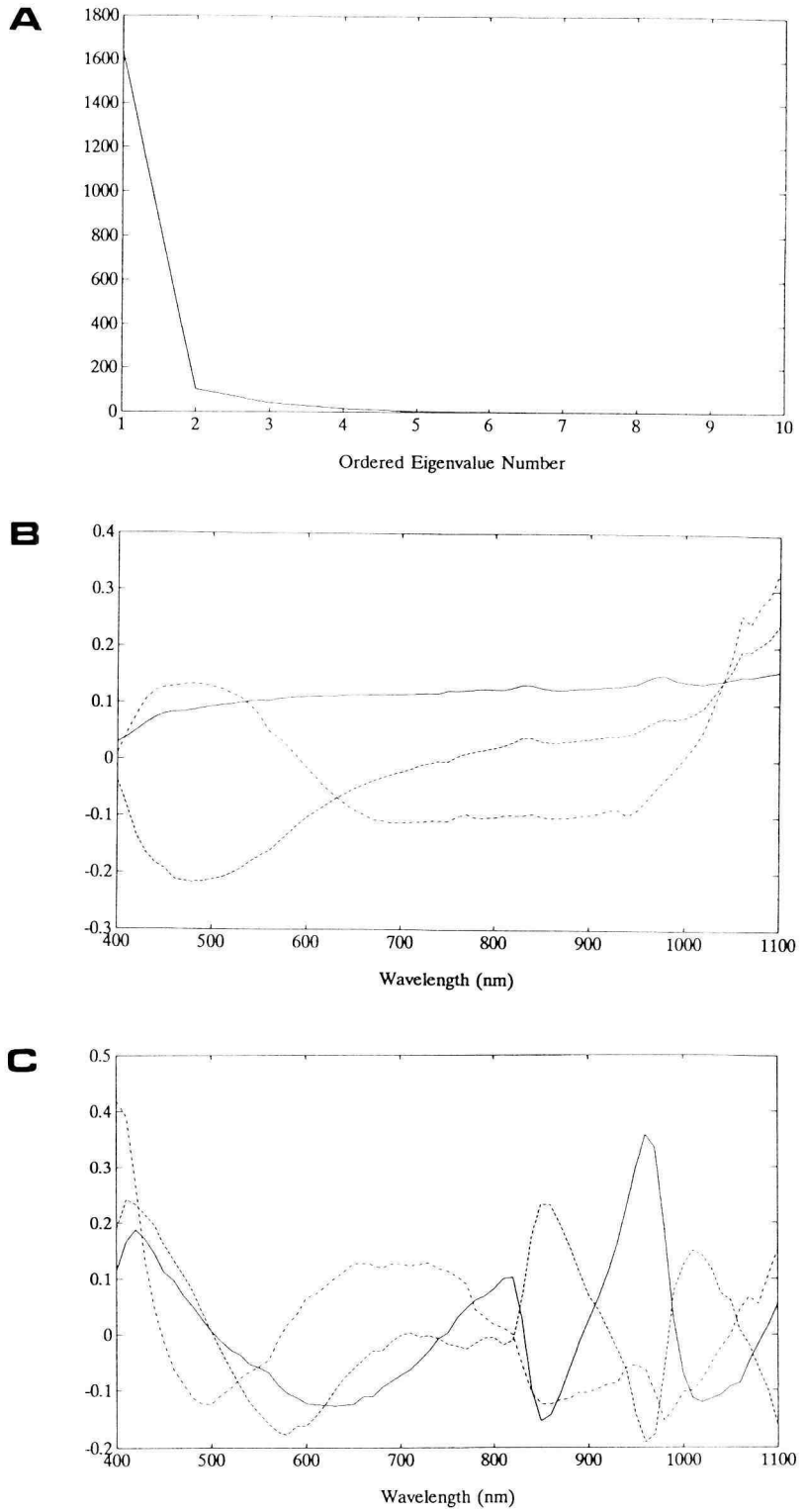


Figure 3.7. Part of the eigenstructure of the sapwood latewood group sample covariance matrix. A) plots the first ten ordered eigenvalues, B) plots the first (solid line), second (dashed line), and third (dash-dotted line) eigenvectors, and C) plots the fourth (solid line), fifth (dashed line), and sixth (dash-dotted line) eigenvectors.

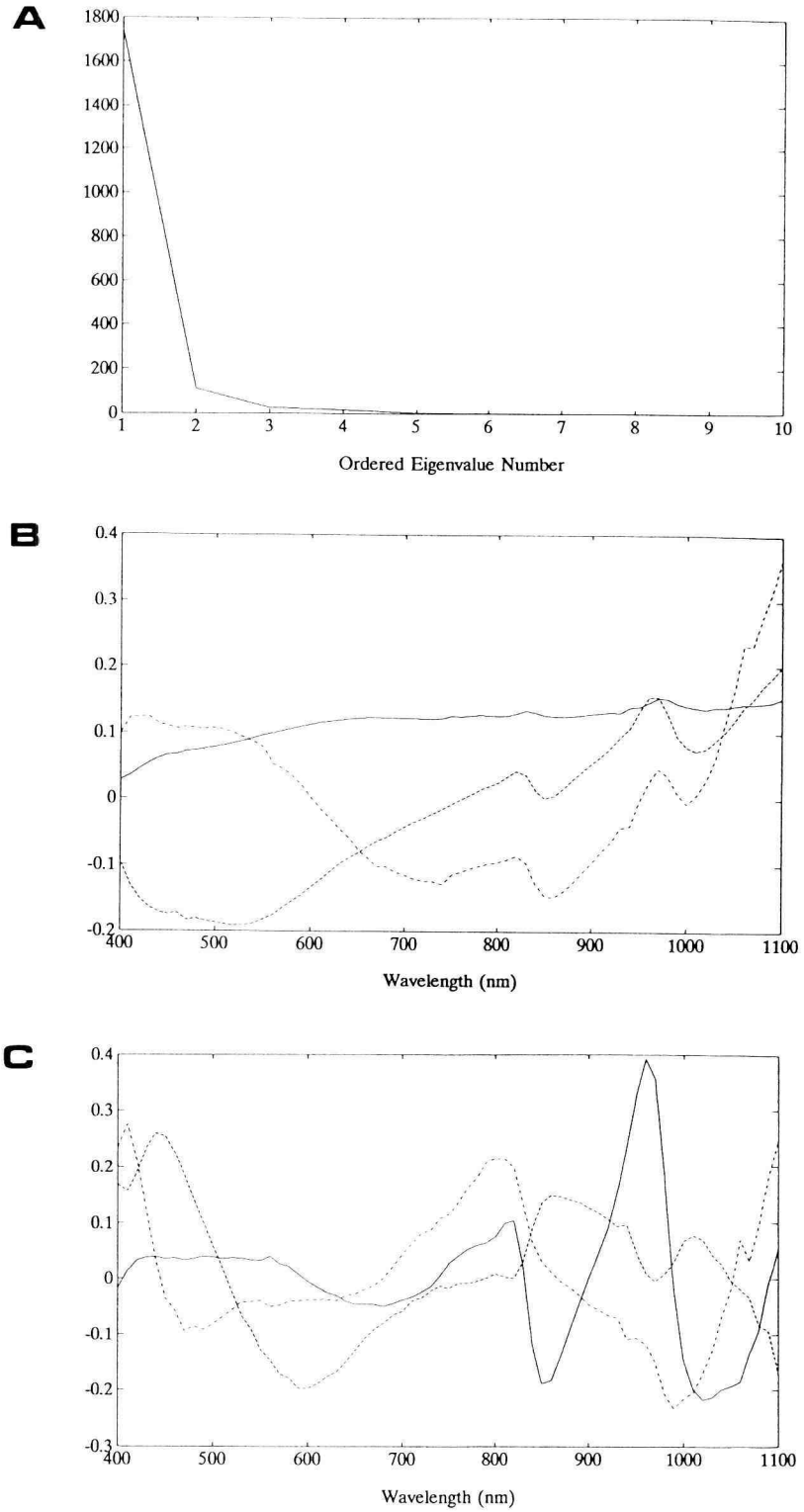


Figure 3.8. Part of the eigenstructure of the heartwood earlywood group sample covariance matrix. A) plots the first ten ordered eigenvalues, B) plots the first (solid line), second (dashed line), and third (dash-dotted line) eigenvectors, and C) plots the fourth (solid line), fifth (dashed line), and sixth (dash-dotted line) eigenvectors.

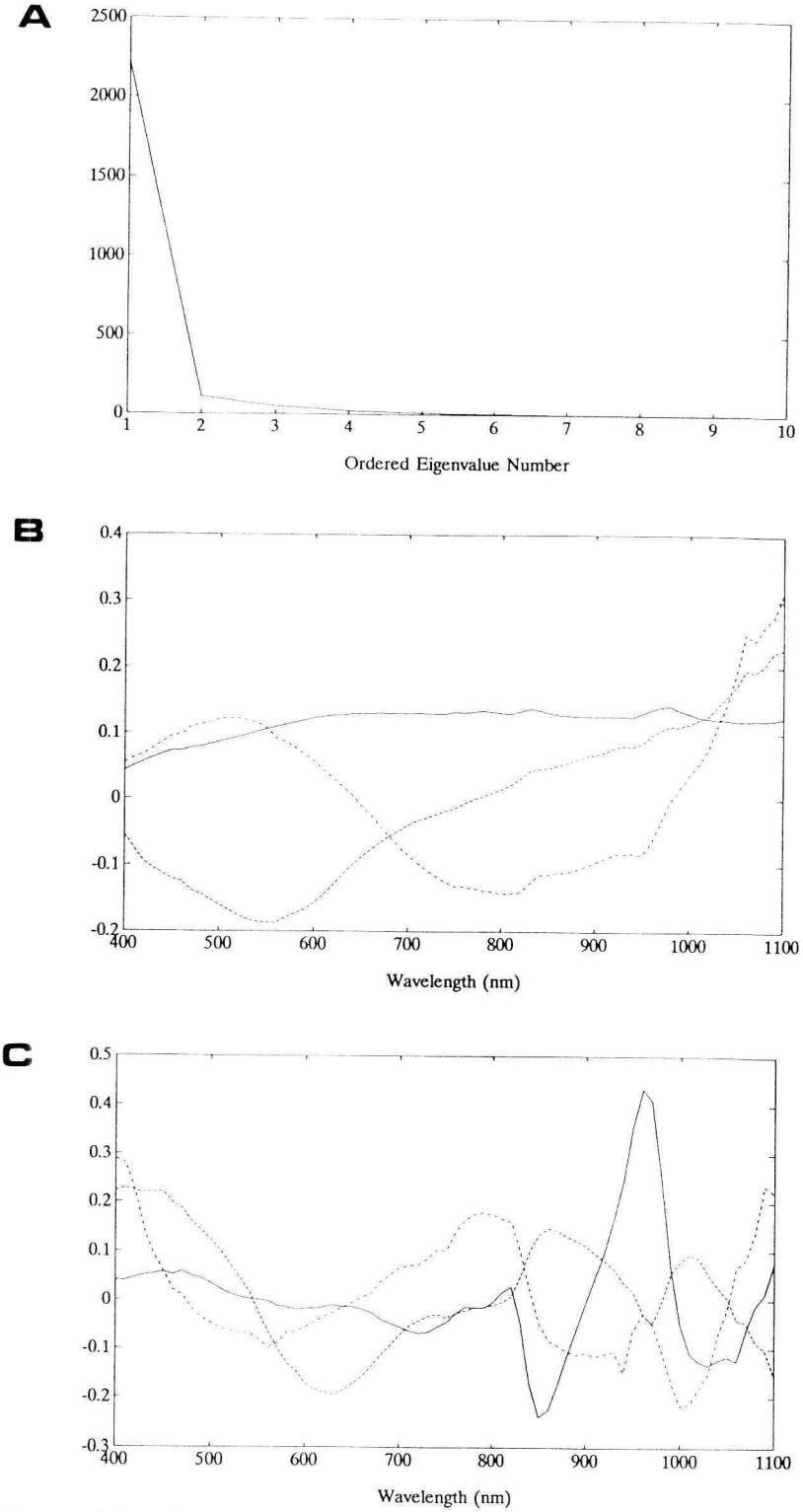


Figure 3.9. Part of the eigenstructure of the heartwood latewood group sample covariance matrix. A) plots the first ten ordered eigenvalues, B) plots the first (solid line), second (dashed line), and third (dash-dotted line) eigenvectors, and C) plots the fourth (solid line), fifth (dashed line), and sixth (dash-dotted line) eigenvectors.

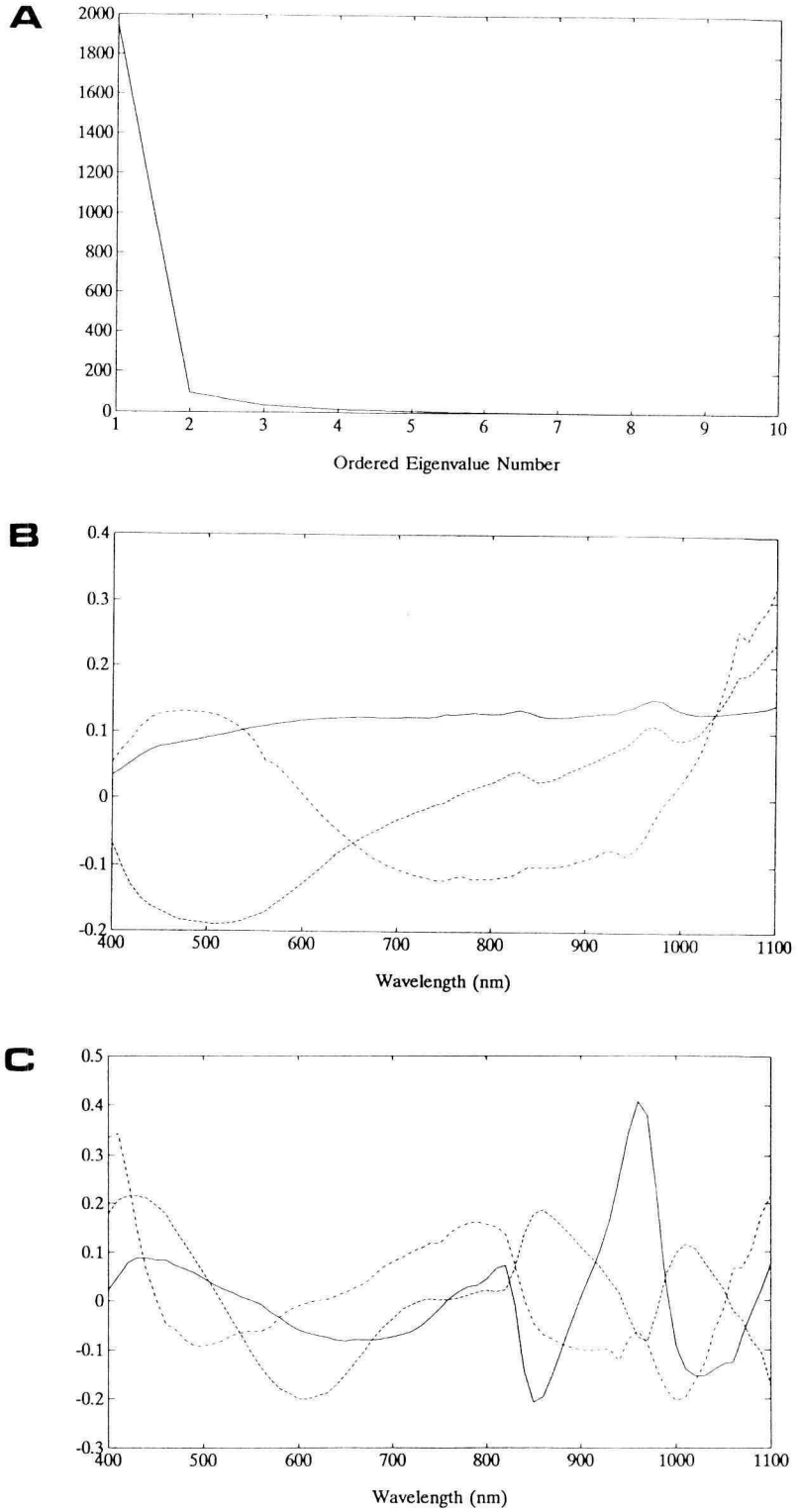


Figure 3.10. Part of the eigenstructure of the pooled covariance matrix for all four groups. A) plots the first ten ordered eigenvalues, B) plots the first (solid line), second (dashed line), and third (dash-dotted line) eigenvectors, and C) plots the fourth (solid line), fifth (dashed line), and sixth (dash-dotted line) eigenvectors.

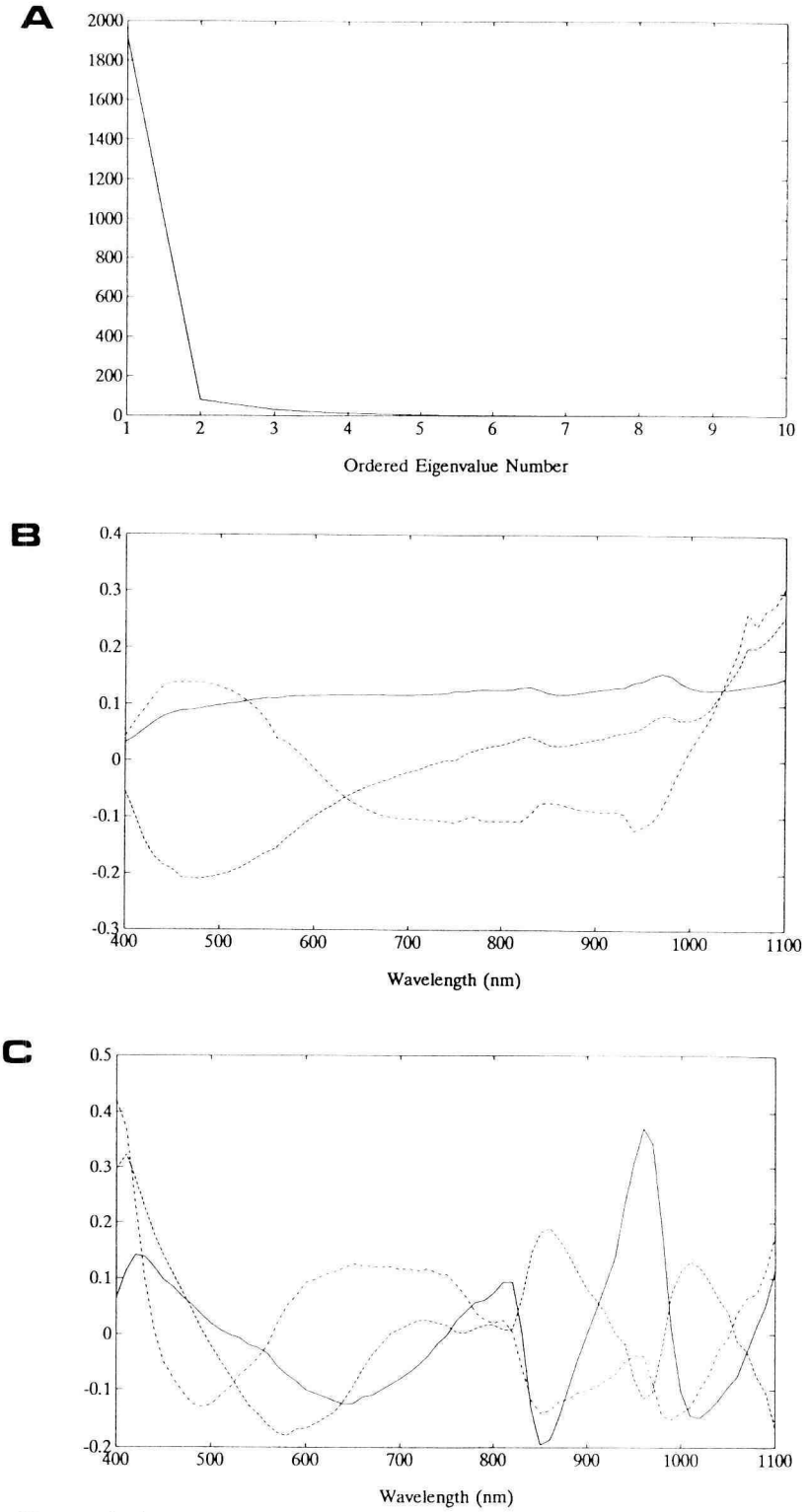


Figure 3.11. Part of the eigenstructure of the pooled sample covariance matrix for the two sapwood groups. A) plots the first ten ordered eigenvalues, B) plots the first (solid line), second (dashed line), and third (dash-dotted line) eigenvectors, and C) plots the fourth (solid line), fifth (dashed line), and sixth (dash-dotted line) eigenvectors.

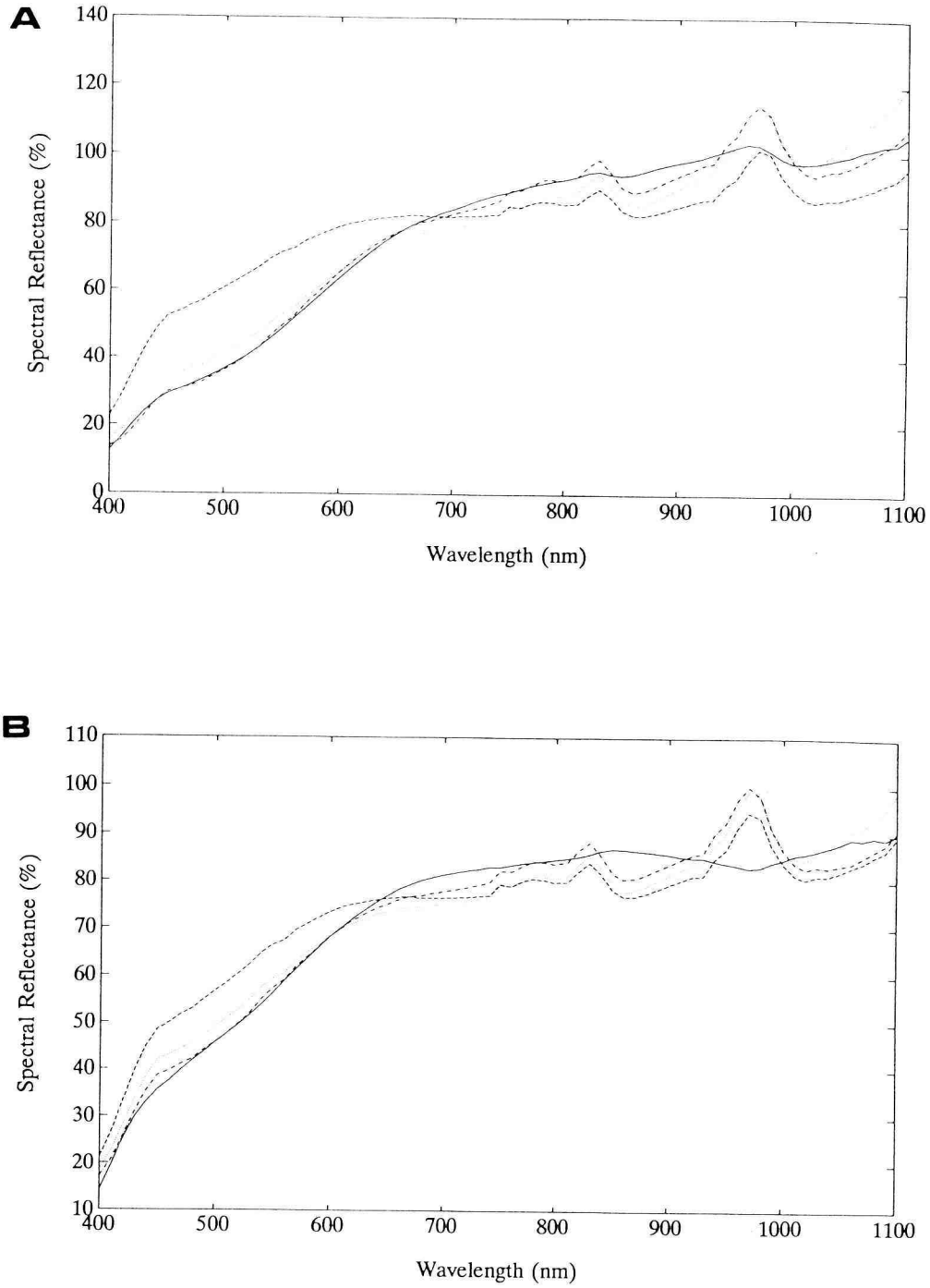


Figure 3.12. Each graph, A) and B), illustrates how well an observed reflectance curve (solid line) is predicted from subsequent principal component models. The dashed line is the predicted curve using the first principal component, the dotted line is the predicted curve using the first two principal components, and the dash-dotted line is the predicted curve using the first three principal components.

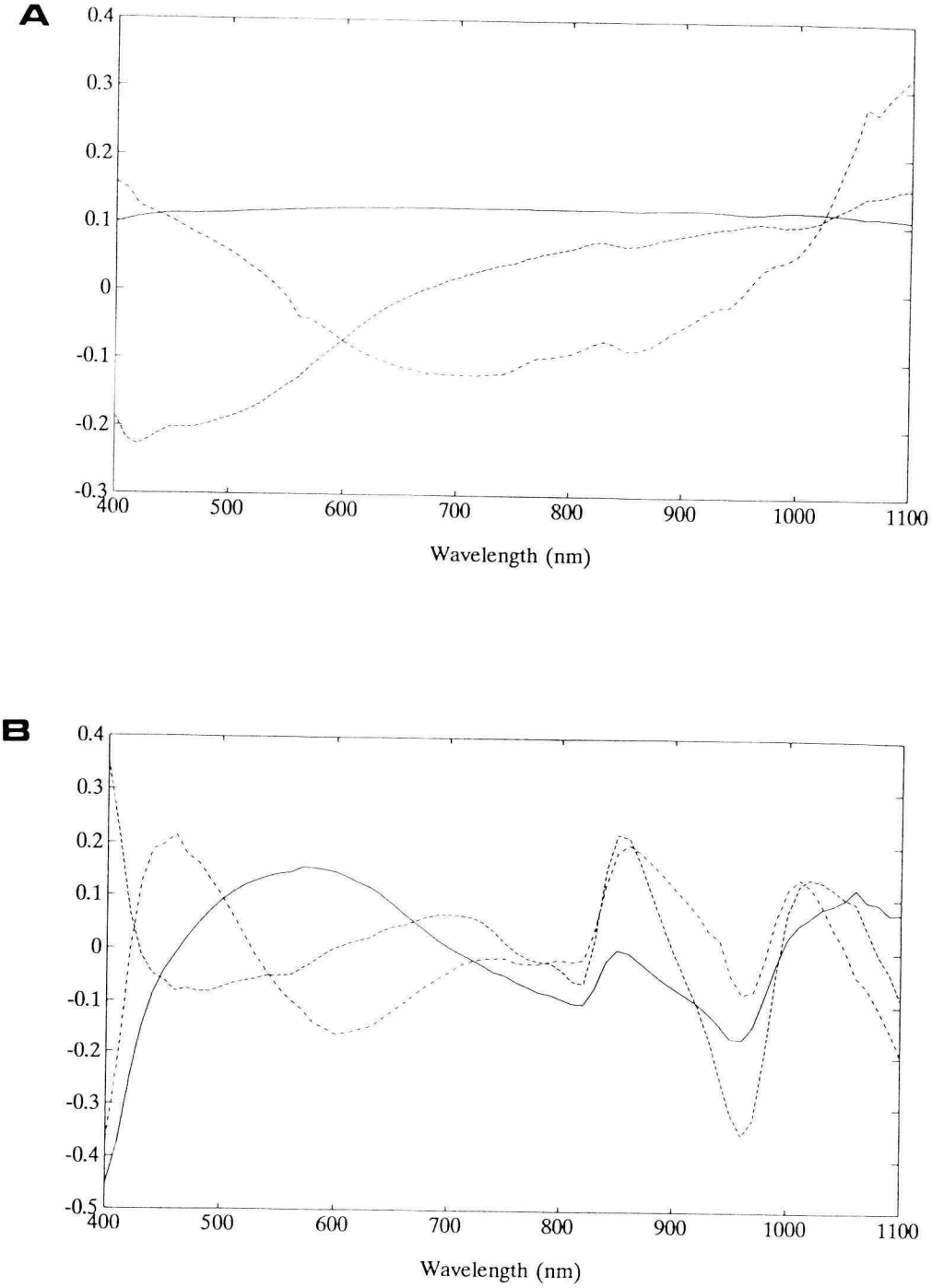


Figure 3.13. Several of the eigenvectors of the pooled correlation matrix for all four groups. A) plots the first (solid line), second (dashed line), and third (dash-dotted line) eigenvectors, and B) plots the fourth (solid line), fifth (dashed line), and sixth (dash-dotted line) eigenvectors.

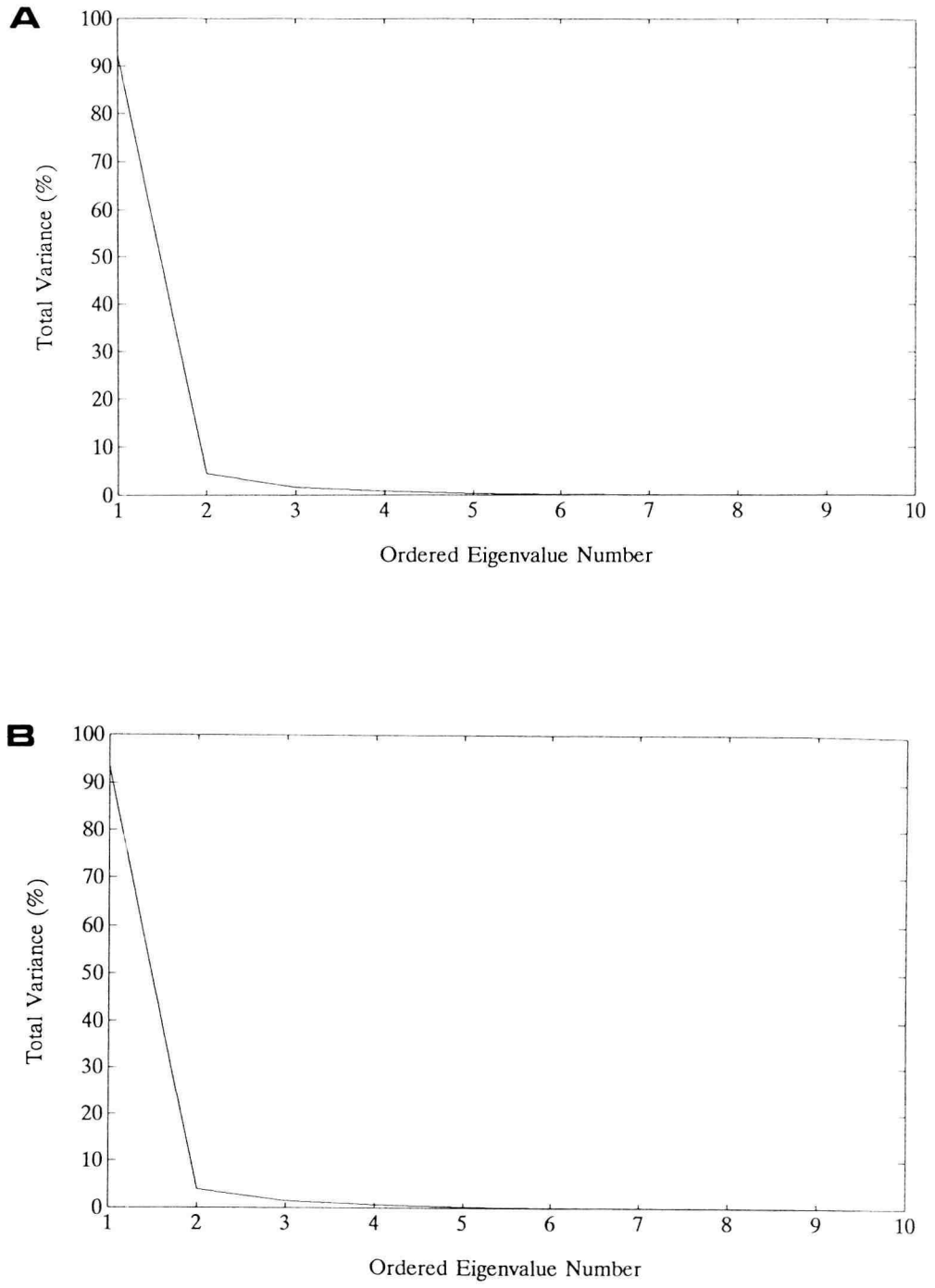


Figure 3.14. Scree plots of the first ten ordered eigenvalues from the pooled covariance matrix for A) all four wood groups and B) the two sapwood groups.

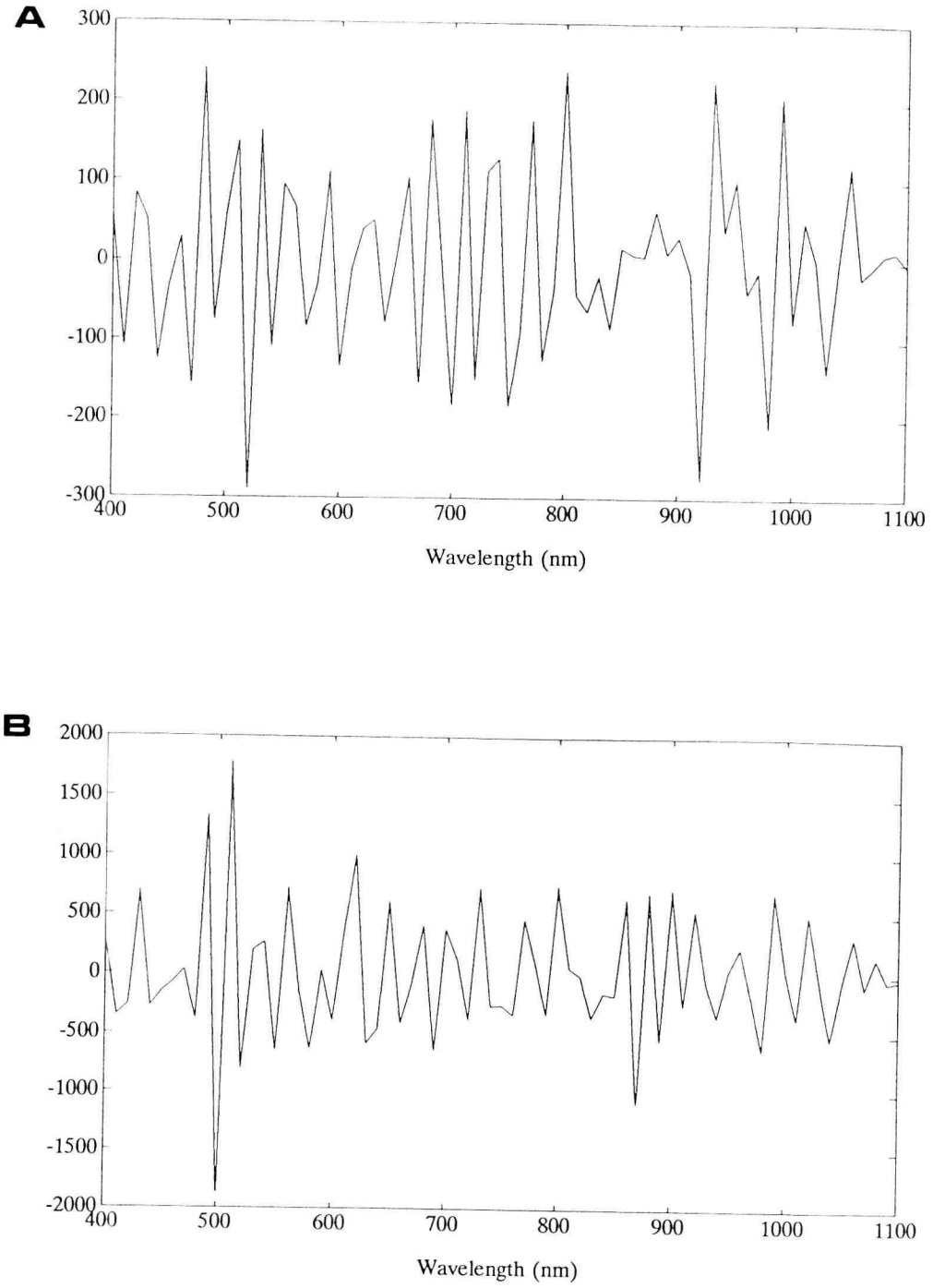


Figure 3.15. Discriminant function coefficients for sapwood earlywood based on the raw data from A) all four wood groups and B) the two sapwood groups.

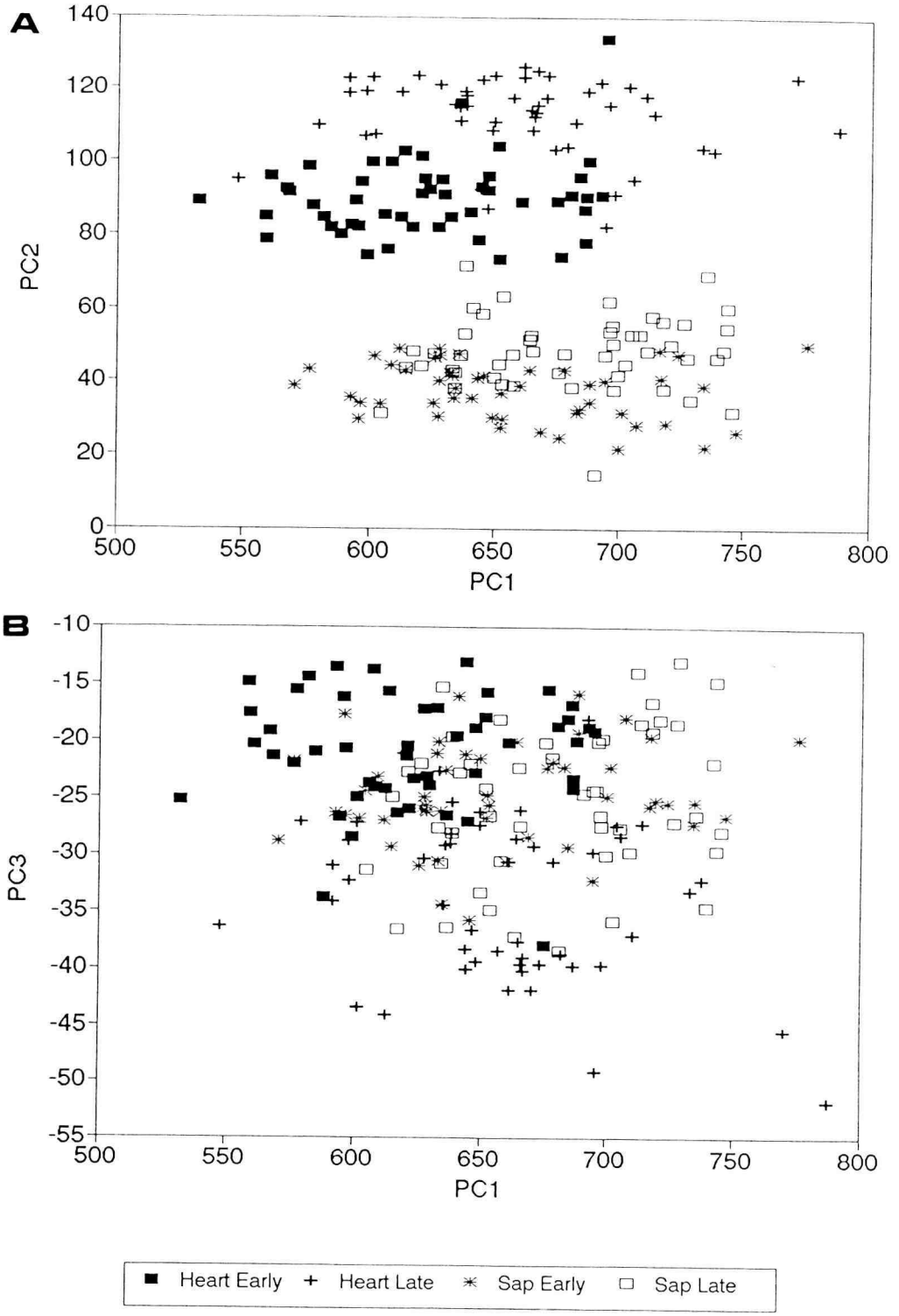


Figure 3.16. Scatterplots for several pairs of the first five principal components. A) Plot of the first principal component versus the second, B) plot of the first versus the third, C) plot of the third versus the second, D) plot of the fourth versus the second, and E) plot of the fifth versus the second.

Figure 3.16. (continued)

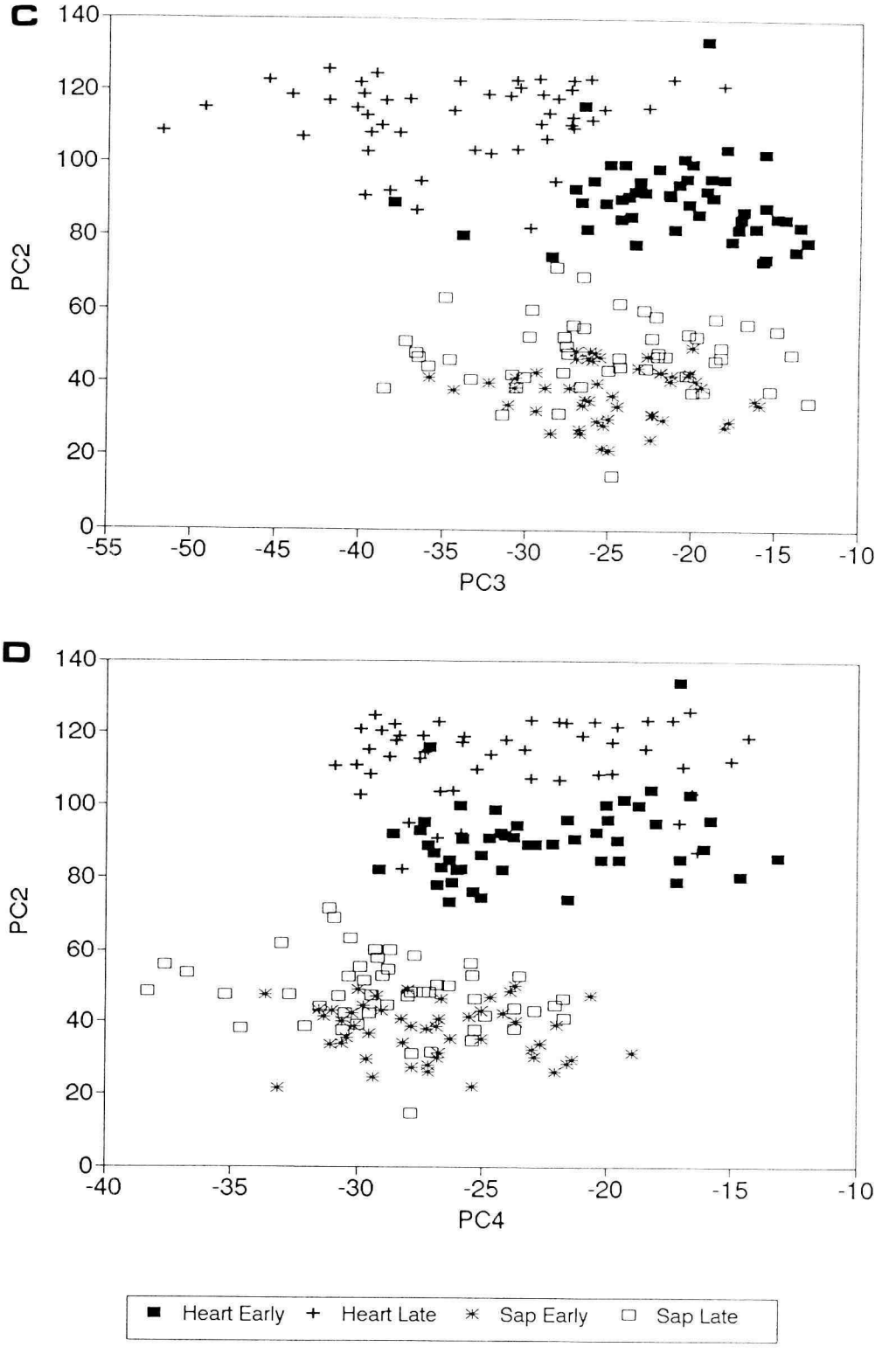
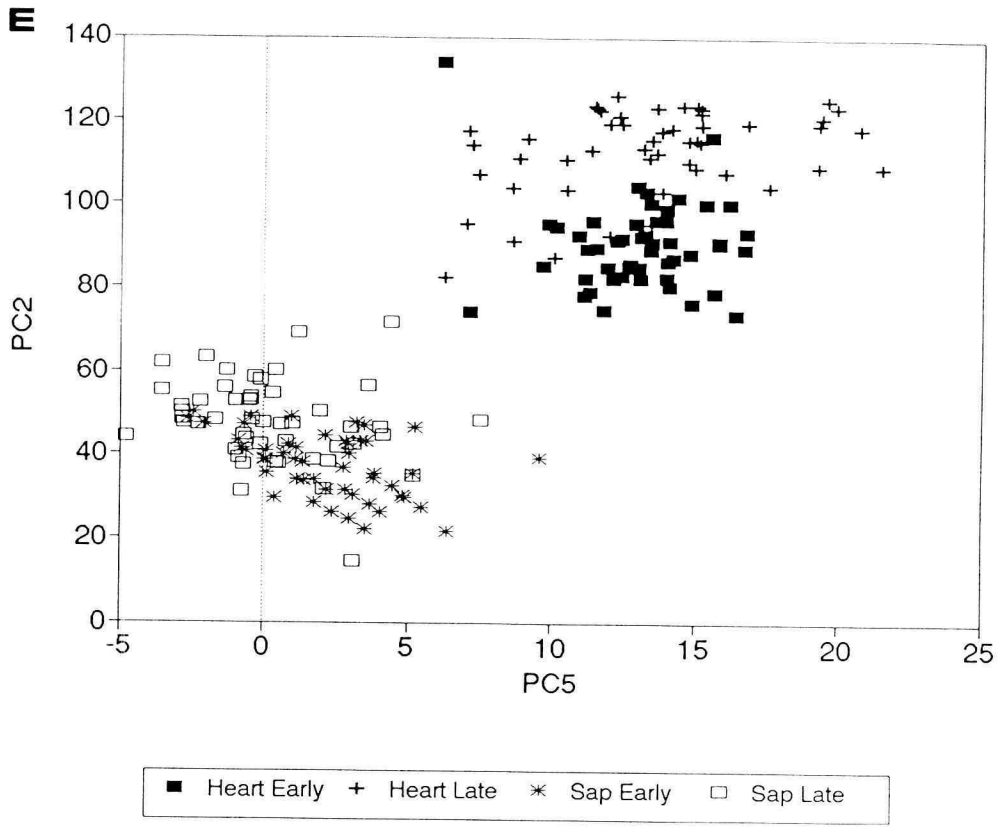


Figure 3.16. (continued)



Chapter 4
RANK-ONE MODIFICATIONS

Effect on Sample Covariance Matrix by Leaving Out One Observation

Removal of an observation from the pooled covariance matrix results in a rank-one change in the pooled covariance matrix. Friedman (1989) states this without proof in his article on regularized discriminant analysis for biased estimators of the populations' covariance matrices. It will be shown here for the unbiased estimator of common covariance and shown in decomposition form in order to apply a fast procedure for updating the eigenstructure.

Define the data matrix (as in Chapter 2) for the j-th group as

$$\mathbf{X}_j = [\mathbf{x}_{1j} \quad \mathbf{x}_{2j} \quad . \quad . \quad . \quad \mathbf{x}_{n_j,j}]'$$

where $\mathbf{x}_{ij}$ is the i-th observed p-vector from this population and n_j is the number of observations from this population. Then the j-th group mean and group sample covariance are

$$\bar{\mathbf{x}}_j = \frac{\sum_{i=1}^{n_j} \mathbf{x}_{ij}}{n_j} \quad ,$$

$$\mathbf{S}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)' \quad .$$

Without loss of generality suppose that the n_g-th observation of the g-th group is deleted. First it will be shown that a multiple of the updated group covariance matrix $\tilde{\mathbf{S}}_g$ is a multiple of the original group covariance matrix minus a rank-one matrix, that is,

$$(n_g\text{-}2)\, \tilde{\mathbf{S}}_g = (n_g\text{-}1)\, \mathbf{S}_g - \frac{n_g}{n_g - 1} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'.$$

Define

$$\tilde{\bar{\mathbf{x}}}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g - 1} \mathbf{x}_{ig} = \frac{1}{n_g - 1} (n_g \bar{\mathbf{x}}_g - \mathbf{x}_{n_g,g}) = \bar{\mathbf{x}}_g - \frac{1}{n_g - 1} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)$$

$$\tilde{\mathbf{S}}_g = \frac{1}{n_g - 2} \sum_{i=1}^{n_g - 1} (\mathbf{x}_{ig} - \tilde{\bar{\mathbf{x}}}_g)(\mathbf{x}_{ig} - \tilde{\bar{\mathbf{x}}}_g)'.$$

Then,

$$(n_g - 2)\, \tilde{\mathbf{S}}_g = \sum_{i=1}^{n_g - 1} (\mathbf{x}_{ig} - \tilde{\bar{\mathbf{x}}}_g)(\mathbf{x}_{ig} - \tilde{\bar{\mathbf{x}}}_g)'$$

$$= \sum_{i=1}^{n_g - 1} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g + \frac{1}{n_g - 1} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g))(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g + \frac{1}{n_g - 1} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g))'$$

$$= \sum_{i=1}^{n_g - 1} [(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)' + \frac{1}{(n_g - 1)^2} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)' +$$

$$\frac{1}{n_g - 1} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)' + \frac{1}{n_g - 1} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)']$$

$$= \sum_{i=1}^{n_g - 1} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)' + \frac{1}{n_g - 1} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)' +$$

$$\frac{1}{n_g - 1} [\sum_{i=1}^{n_g - 1} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)] (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)' +$$

$$\frac{1}{n_g - 1} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g) [\sum_{i=1}^{n_g - 1} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)]'$$

Rewriting $\sum_{i=1}^{n_g - 1} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g) = \sum_{i=1}^{n_g - 1} \mathbf{x}_{ig} - (n_g - 1)\, \bar{\mathbf{x}}_g$

$$= n_g \bar{\mathbf{x}}_g - \mathbf{x}_{n_g,g} - (n_g - 1)\, \bar{\mathbf{x}}_g$$

$$= -(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g). \text{ Thus,}$$

$$(n_g - 2)\, \tilde{\mathbf{S}}_g = \sum_{i=1}^{n_g - 1} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)' -$$

$$\frac{1}{n_g - 1} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'.$$

Add and subtract $(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'$ to obtain

$$= \sum_{i=1}^{n_g}(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)' - \frac{n_g}{n_g - 1}(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'$$

$$= (n_g - 1)\,\mathbf{S}_g - \frac{n_g}{n_g - 1}(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'. \qquad (4.1)$$

Assuming that there are g groups, let $n = \sum_{k=1}^{g} n_k$, the total number of observations. The pooled covariance matrix is defined as

$$\mathbf{S} = \frac{1}{n - g}\sum_{k=1}^{g}(n_k - 1)\,\mathbf{S}_k.$$

The updated pooled covariance matrix is

$$\tilde{\mathbf{S}} = \frac{1}{n-g-1}\sum_{k=1}^{g-1}(n_k - 1)\,\mathbf{S}_k + \frac{n_g - 2}{n-g-1}\tilde{\mathbf{S}}_g.$$

From (4.1) substitute in $(n_g - 2)\,\tilde{\mathbf{S}}_g$,

$$\tilde{\mathbf{S}} = \frac{1}{n-g-1}\Big[\sum_{k=1}^{g-1}(n_k - 1)\,\mathbf{S}_k + (n_g - 1)\,\mathbf{S}_g - \frac{n_g}{n_g - 1}(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'\Big]$$

$$= \frac{1}{n-g-1}\Big[\sum_{k=1}^{g}(n_k - 1)\,\mathbf{S}_k - \frac{n_g}{n_g - 1}(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'\Big]$$

$$= \frac{n-g}{n-g-1}\Big[\mathbf{S} - \frac{n_g}{(n-g)(n_g - 1)}(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'\Big] \qquad (4.2)$$

This shows the rank-one change in the pooled covariance matrix by the deletion of one observation from the data set.

It will be shown that if the spectral decomposition of $\mathbf{S}$ is $\mathbf{VBV}'$, then

$$\tilde{\mathbf{S}} = \frac{n-g}{n-g-1} \left[\mathbf{VBV}' - \frac{n_g}{(n-g)(n_g-1)} (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)'\right]$$

$$= \frac{n-g}{n-g-1} \mathbf{V} \left[\mathbf{B} - \frac{n_g}{(n-g)(n_g-1)} \mathbf{V}'(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)' \mathbf{V}\right] \mathbf{V}'$$

$$= \mathbf{V} \left[- \frac{n-g}{n-g-1} (\mathbf{D} + \rho \mathbf{Z}\mathbf{Z}')\right] \mathbf{V}'$$

where $\mathbf{D} = -\mathbf{B}$,

$$\mathbf{Z} = \mathbf{V}' (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}_g)$$

and

$$\rho = \frac{n_g}{(n-g)(n_g-1)} . \tag{4.3}$$

For the purposes of this algorithm it is computationally more efficient to keep the coefficient $-(n-g) / (n-g-1)$ on the outside of the matrix $\mathbf{D} + \rho\mathbf{Z}\mathbf{Z}'$ rather than multiplying it through. If the orthogonal decomposition of

$$\mathbf{D} + \rho\mathbf{Z}\mathbf{Z}' = \tilde{\mathbf{V}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}' , \tag{4.4}$$

then the orthogonal decomposition of $\tilde{\mathbf{S}}$ is given by

$$\tilde{\mathbf{S}} = \mathbf{V}\tilde{\mathbf{V}} \left(- \frac{n-g}{n-g-1} \tilde{\mathbf{D}} \right) \tilde{\mathbf{V}}'\mathbf{V}' \tag{4.5}$$

Algorithm for Estimating Updated Eigenstructure

Golub (1973) presents the basic eigenproblem of finding the orthogonal decomposition of a diagonal matrix with a rank-one modification as in (4.4) and suggests several alternatives for solving it. Bunch *et al.* (1978) have expanded

on one of Golub's propositions by providing explicit computations of the updated eigenvalues and eigenvectors. The eigenvalues of (4.4), $\tilde{d}_i$, $i = 1, \cdots, p$, have been shown to have precise bounds (Wilkinson, 1965),

$$d_i \leq \tilde{d}_i \leq d_{i+1} \qquad \text{for } i = 1, \cdots, p-1, \text{ and}$$

$$\tilde{d}_p \leq d_p + \rho \mathbf{Z}'\mathbf{Z} .$$

The eigenvalues $\tilde{d}_i$ are chosen to satisfy

$$\det(\mathbf{D} + \rho \mathbf{Z}\mathbf{Z}' - \lambda \mathbf{I}) = 0$$

which is equivalent to

$$\det(\mathbf{D} - \lambda \mathbf{I}) \det(\mathbf{I} + \rho(\mathbf{D} - \lambda \mathbf{I})^{-1} \mathbf{Z}\mathbf{Z}') = 0$$

when the elements ζ_i, $i = 1, \cdots, p$, of $\mathbf{Z}$ are non-zero. Since $\mathbf{D}$ is a diagonal matrix, this can be reduced, as Golub (1973) claims, to

$$\prod_{i=1}^{p} (d_i - \lambda) \left(1 + \rho \sum_{j=1}^{p} \frac{\zeta_j^2}{d_j - \lambda}\right) = 0.$$

The eigenvalues are then computed by finding the zeros of

$$\mathrm{w}(\lambda) = 1 + \rho \sum_{j=1}^{p} \frac{\zeta_j^2}{d_j - \lambda}.$$

Bunch *et al.* (1978) reformulate this problem to improve the precision of eigenvector estimates by first denoting $\tilde{d}_i = d_i + \rho \nu_i$ satisfies $\mathrm{w}(\tilde{d}_i) = 0$, $i = 1, \cdots, p$. The i-th eigenvalue is, thus, computed by solving $\mathrm{w}_i(\nu) = 0$ where

$$\mathrm{w}_i(\nu) = 1 + \psi_i(\nu) + \phi_i(\nu)$$

and

$$\psi_i(\nu) = \sum_{j=1}^{i} \frac{\zeta_j^2}{\delta_j - \nu} ,$$

$$\phi_i(\nu) = \sum_{j=i+1}^{p} \frac{\zeta_j^2}{\delta_j - \nu} ,$$

$$\delta_j = \frac{d_j - d_i}{\rho} \ .$$

Approximating functions should make use of the fact that on the interval $(0, \delta_{i+1})$ the function $-\psi_i(\cdot)$ is decreasing and convex and the function $\phi_i(\cdot)$ is increasing and convex. Rational functions are used as local approximating functions of ψ_i and ϕ_i at a point u_k, $0 < u_k < \nu_i$. A new approximation u_{k+1} of ν_i is obtained by defining the rational functions $q/(r-u)$ and $s + t/(\delta - u)$ such that

$$\begin{aligned} \frac{q}{r-u_k} &= \psi_k & s + \frac{t}{\delta - u_k} &= \phi_k \\ \frac{q}{(r-u_k)^2} &= \psi_k' & s + \frac{t}{(\delta - u_k)^2} &= \phi_k' \end{aligned} \tag{4.6}$$

where $\delta = \delta_{i+1}$, $\psi_k = \psi_i(u_k)$, $\psi_k' = \psi_i'(u_k)$, etc. Thus, the approximation u_{k+1} is found by solving

$$-\frac{q}{r-u_{k+1}} = 1 + s + \frac{t}{\delta - u_{k+1}} \ .$$

The solution to this quadratic equation is found to be

$$u_{k+1} = u_k + 2\,b \,/\, (a + \sqrt{a^2 - 4b}\,)$$

where

$$\begin{aligned} a &= (\gamma(1+\phi_k) + \psi_k^2/\psi_k')\,/\,c + \psi_k/\psi_k' \\ b &= (\gamma\omega\psi_k)\,/\,(c\psi_k') \\ c &= 1 + \phi_k - \gamma\phi_k' \\ \omega &= 1 + \phi_k + \psi_k \\ \gamma &= \delta - u_k \ . \end{aligned}$$

For the case $i = p$, this reduces to

$$u_{k+1} = u_k + \frac{1+\psi_k}{\psi_k'}\ \psi_k$$

since $\phi_i(\cdot) \equiv 0$. Bunch *et al.* (1978) show that these estimates converge quadratically to the true eigenvalues.

Initial values u_0 for ν_i, $i < p$, are calculated from

$$\frac{\zeta_i^2}{u_0} = \frac{\zeta^2{}_{i+1}}{\delta_{i+1} - u_0} + 1 + \sum_{\substack{j=1 \\ j \neq i, i+1}}^{p} \frac{\zeta_j^2}{\delta_j - \delta_{i+1}}$$

For $i = p$, take $u_0 = \mathbf{Z}'\mathbf{Z} - \sum_{i=1}^{p-1} \hat{\nu}_i = (\mathbf{x}_{n_g,g} - \bar{\mathbf{x}})'(\mathbf{x}_{n_g,g} - \bar{\mathbf{x}}) - \sum_{i=1}^{p-1} \hat{\nu}_i$.

Once the i-th eigenvalue is computed, the associated eigenvector can be calculated directly without iteration. This is based on a thereom in Bunch, *et al.* (1978). To compute the eigenvectors $\tilde{\mathbf{Q}} = \mathbf{V}\tilde{\mathbf{V}}$ of $\tilde{\mathbf{S}}$ first note they must satisfy

$$\tilde{\mathbf{S}}\tilde{\mathbf{q}}_i = -c\tilde{d}_i\tilde{\mathbf{q}}_i\ ,\quad i = 1, \cdots, p$$

where $c = (n-g)/(n-g-1)$ and $\tilde{\mathbf{q}}_i = \mathbf{V}\tilde{\mathbf{v}}_i$. Multiplying by $\mathbf{V}'$, this implies

$$\mathbf{V}'\tilde{\mathbf{S}}\tilde{\mathbf{q}}_i + c\tilde{d}_i\mathbf{V}'\tilde{\mathbf{q}}_i = 0$$

$$\Rightarrow \mathbf{V}'\ [-c\mathbf{V}(\mathbf{D} + \rho\mathbf{Z}\mathbf{Z}')\ \mathbf{V}'\ ]\ \mathbf{V}' + c\tilde{d}_i\mathbf{V}'\tilde{\mathbf{q}}_i = 0$$

$$\Rightarrow [-c(\mathbf{D} + \rho\mathbf{Z}\mathbf{Z}') + c\tilde{d}_i]\ \mathbf{V}'\tilde{\mathbf{q}}_i = 0$$

$$\Rightarrow (\mathbf{D}_i - c\rho\mathbf{Z}\mathbf{Z}')\ \tilde{\mathbf{v}}_i = 0$$

where $\mathbf{D}_i = -c\mathbf{D} + c\tilde{d}_i\mathbf{I} = c\ (-\mathbf{D} + \tilde{d}_i\mathbf{I})$. Bunch *et al.* show this is equivalent to

$$\tilde{\mathbf{v}}_i = -\theta \mathbf{D}_i^{-1}\mathbf{Z}, \quad (-\rho^{-1} + \mathbf{Z}'\mathbf{D}_i\mathbf{Z})\,\theta = 0, \quad \theta \text{ arbitrary (nonzero).}$$

$$\Rightarrow \tilde{\mathbf{q}}_i = -\theta \mathbf{V}\mathbf{D}_i^{-1}\mathbf{Z}\,.$$

To satisfy normalizing constraints $\tilde{\mathbf{q}}_i'\tilde{\mathbf{q}}_i = 1$, choose $\theta = -\left\|\mathbf{D}_i^{-1}\mathbf{Z}\right\|_2^{-1}$.

It should be noted that in the case of equal eigenvalues, the above rank-one algorithm can be accelerated further with what Bunch *et al.* call deflation. See Bunch *et al.* (1978) for details.

Results For Wood Data Set

In applying this algorithm, the leave-one-out cross-validation error rate estimates are the same as the original process of re-evaluation outlined in the last chapter. In fact, it produces the same classification probabilities for each observation as the original process to at least six decimal places (see Mean Absolute Error in Tables 3.6 and 3.7). However, since only the first three components need to be updated for each observation removed, it takes 38 million flops and six minutes to calculate the estimate of the misclassification rate for the four group case and 16 million flops and 2.4 minutes for the two group case, a 96% decrease in flops and a 90% decrease in time for both cases (see Tables 3.6 and 3.7). This increased computational efficiency does not include the advantage of the updated pooled covariance matrix (4.2) being diagonal when the original or rank-one methods are used.

Modifications to the Bunch *et al.* algorithm

DeGroat and Roberts (1990) replace Bunch's *et al.* second rational function approximation $s + t/(\delta - u)$ with $s/(t - u)$ claiming this simplification is slightly faster and less complex. This algorithm also converges quadratically to the true eigenvalues (proof in DeGroat's thesis), but does not converge monotonically as does Bunch's *et al.* (This is why DeGroat and Roberts feel their approximation is faster.)

Convergence of the Estimated Eigenstructure

The eigenvalues of the updated eigenstructure converge quadratically as mentioned above. The norm of each eigenvector error is shown through perturbation analysis to be bounded by a multiple of the error in the associated eigenvalue scaled with the minimal change in the eigenvalues. See Bunch *et al.* (1978) for details. Bounds on the posterior probability estimates, and, hence, bounds on the leave-one-out cross-validation error rate estimate can be established.

Efficiency of the Rank-One Method in Flops

The number of flops necessary for this rank-one update method is on the order of qp^2, where p and q are as described in the prior chapter. The pooled covariance matrix is calculated once taking $(2n + 3g)p^2 + 4np + 3g$ flops; the complete decomposition of this matrix requires approximately $9p^3$ flops; and for each observation, the rank-one adjustment is used to calculate the updated eigenvectors in approximately $5qp^2$ flops. The number of flops to compute the principal components and discriminant function posterior probabilities remains the same as outlined in the previous chapter and Appendix A.

Chapter 5
EXPERIMENTAL RESULTS

This chapter contains results, in particular misclassification error rates and program running times and flops, as pertaining to the wood data set. The primary focus of this chapter is to offer comparative results under different conditions to examine the effects of number of variables in the response curve, the number of components kept, the number of groups involved, and the number of response curves (observations). It should be kept in mind the results are from MATLAB implementations of the methods on a personal computer with a 80386 microprocessor (with math co-processor) running at 25 mHz.

Floating point operation (flop) counts for each of the three methods of estimating the discriminant function error rate should be similar to the formulas outlined in Chapter 3 and 4. Let p be the number of variables in the spectral reflectance curve, q be the number of principal components kept for discrimination, $n = \sum_j n_j$ be the total sample size (the number of response curves), and g be the number of groups. Then, assuming that g is typically rather small and q is also small (in comparison to p and n), the flop counts are roughly

$$[(2n + 3g)p^2 + 9p^3]n \qquad + 2n^2pq + 2n^2q^2$$

for the original method,

$$(2n + 3g)p^2 + 9p^3 \qquad + 2npq + 2n^2q^2$$

for the short-cut method, and

$$(2n + 3g)p^2 + 9p^3 + 5np^2q \quad + 2npq + 2n^2q^2$$

for the rank-one method. Thus, small changes in g and q should not have much affect on flop counts.

The experiments compare the three procedures for cross-validation estimation of the overall error rate, the original method of calculation, the method utilizing the rank-one change in the pooled covariance, and the short-cut method. The first experiment compares the three procedures based on two principal components using all the observations from the two groups sapwood

earlywood and sapwood latewood (50 observations per group -- 100 observations total). Figure 5.1(A) shows the changes in floating point operations (flops) as the number of original variables is increased while Figure 5.1(B) shows the changes in computation times. In all cases the estimate of overall error rate was 0.27 suggesting that with a large number of observations, the short-cut procedure is the best alternative.

The second experiment repeats the first experiment but uses only half of the original observations from the two groups sapwood earlywood and sapwood latewood (25 per group). See Figure 5.2(A) and (B) for the flop counts and computation times. When the number of original variables was 11 or 15, the estimate of overall error rate was 0.22 in all three procedures, while the estimate was 0.24 if the number of original variables was 19, 24, or 71. Only in the case where 36 original variables were used was there a discrepancy in the estimate: it was 0.22 for the original and rank-one methods and 0.24 for the short-cut method.

Again, repeating the first experiment this time with only one-fifth the original observations from the two sapwood groups (10 per group), the results are summarized in Figures 5.3(A) and (B). In this situation the estimate of the error rate increases to 0.35 when 11 of the original variables are used and 0.30 when 15, 19, 24, 36, and 71 original variables are used. As in the first experiment, there is no difference between estimates by the three procedures.

When the third experiment is repeated, that is on the 20 observations, but using four principal components for discrimination, there is a difference in the overall error rate estimate between the different cases. For 11 original variables the estimate is 0.30 for the three procedures, but for more than 11 original variables the estimate drops to 0.25 for the long and rank-one procedures and stays at 0.30 for the short-cut procedure. The flops and computation times are given in Figures 5.4(A) and (B). Notice in this case that although flop counts for the original method dominate for all instances, the implementation of the rank-one algorithm does not have a time advantage until the reflectance curves contain at least 24 measurements per observation. With so few observations, the sequentialness of the rank-one algorithm in a single processor environment coupled with the iterativeness of eigenvalue updating are the reasons for the longer times.

Repeating the first experiment except using the first three principal components for discrimination, there are slight increases in number of flops and computation times. See Figures 5.5(A) and (B). It is interesting that adding the third component increases the error-rate estimates to about 0.29.

The first experiment was also repeated for the sapwood groups, but only the second and third components were kept for discrimination. Figures 5.6(A) and (B) show that the original method computationally performs the same as in the first experiment, the rank-one updating method shows a slight decrease in flops and slight increase in time, while the short-cut method shows a slight increase in flops. Using the second and third components, but not the first, improve the cross-validation error-rate estimate slightly to 0.26.

The last experiment compares the procedures using all the observations from all four wood groups, sapwood earlywood, sapwood latewood, heartwood earlywood, and heartwood latewood, when the first three principal components are kept for discrimination. That is, there were 50 observations per group for a total of 200 observations. Figure 5.7(A) shows the tripling effects on the number of flops of the methods as compared when the two groups were used (100 observations total -- Figure 5.5(A)). The effects on the computation times differ among the procedures: the short-cut times are almost five times longer, the original method times are about four times longer, and the rank-one update method times are about triple. See Figure 5.7(B). In all three procedures the cross-validation error-rate is 0.165 when 11 original variables are used, 0.190 when 19 and 36 original variables are used, and 0.185 when 71 original variables are used.

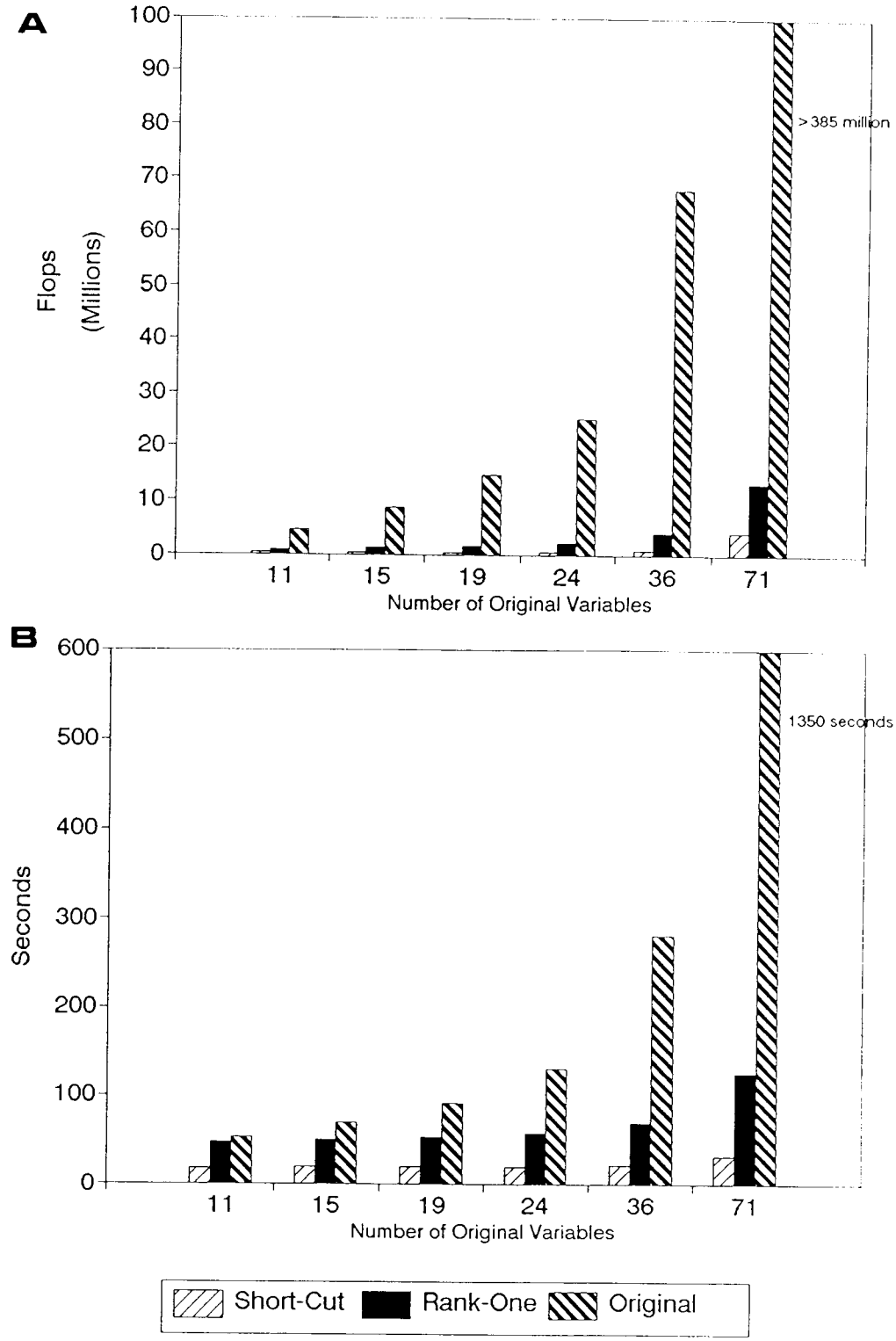


Figure 5.1. Procedure comparison graphs for calculating the discriminant function error rate based on the first two principal components for the two sapwood groups with 50 observations per group. A) is the number of floating point operations and B) is the time in seconds to compute the error rates.

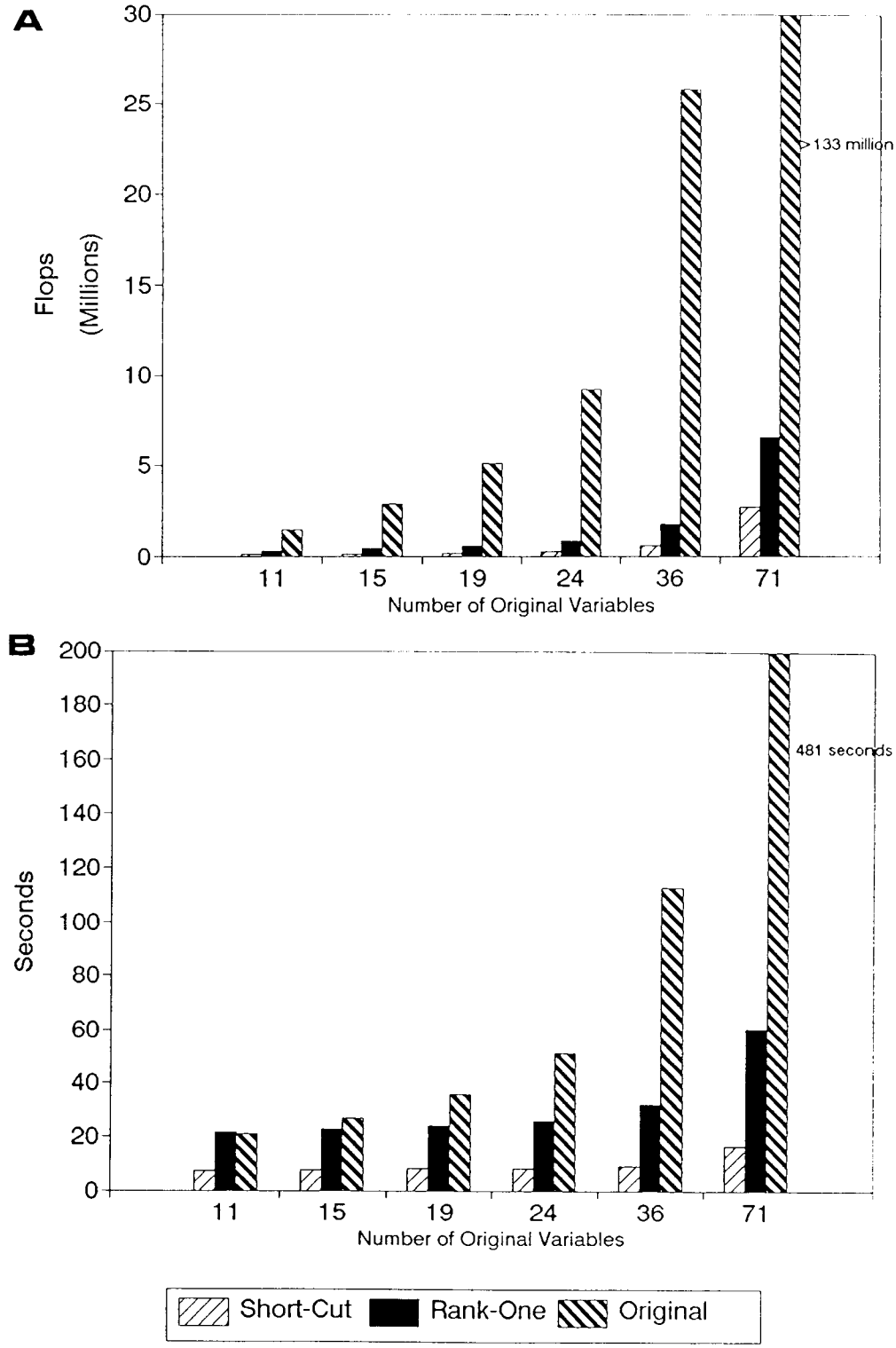


Figure 5.2. Procedure comparison graphs for calculating the discriminant function error rate based on the first two principal components for the two sapwood groups with 25 observations per group. A) is the number of floating point operations and B) is the time in seconds to compute the error rates.

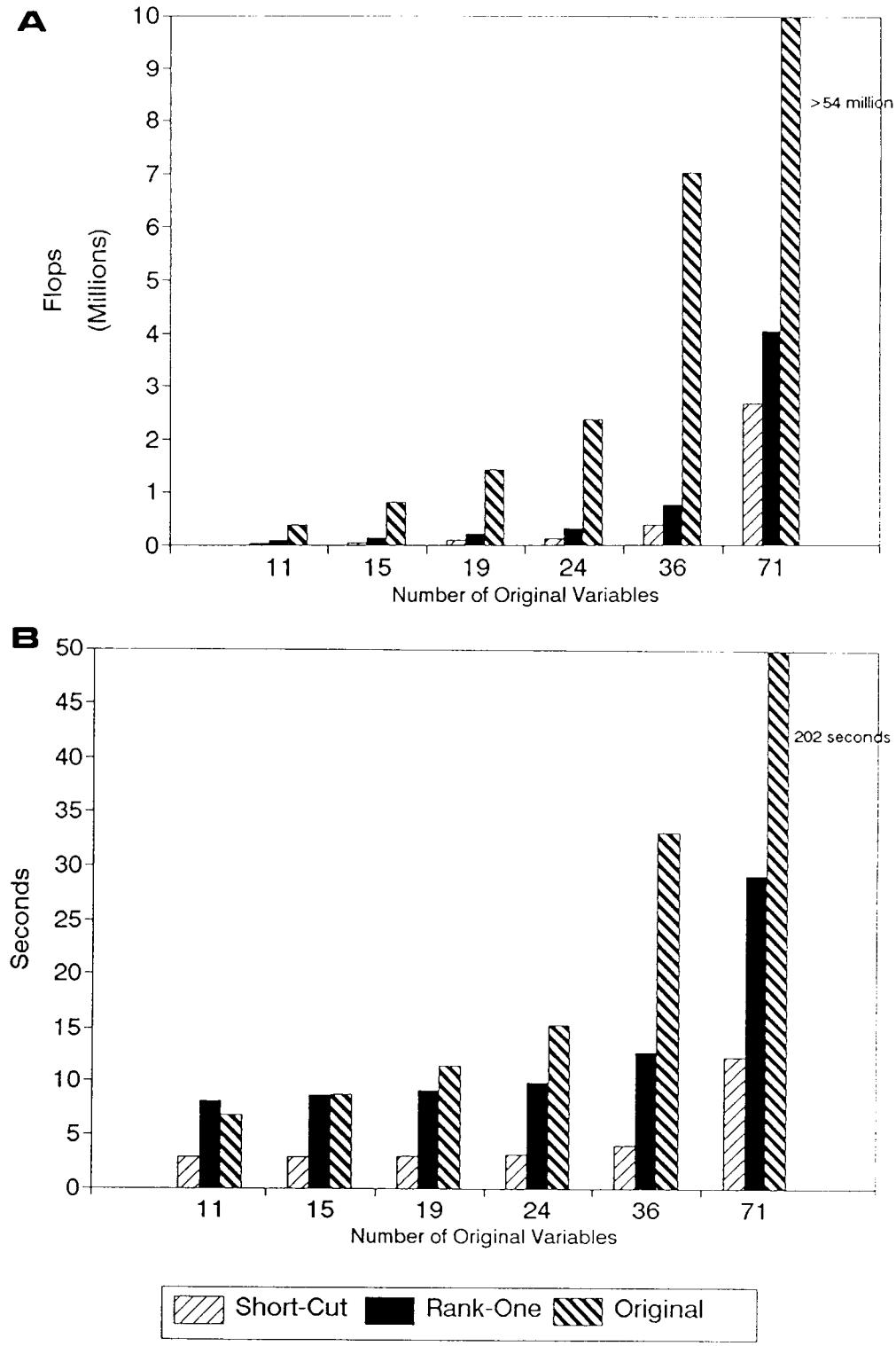


Figure 5.3. Procedure comparison graphs for calculating the discriminant function error rate based on the first two principal components for the two sapwood groups with ten observations per group. A) is the number of floating point operations and B) is the time in seconds to compute the error rates.

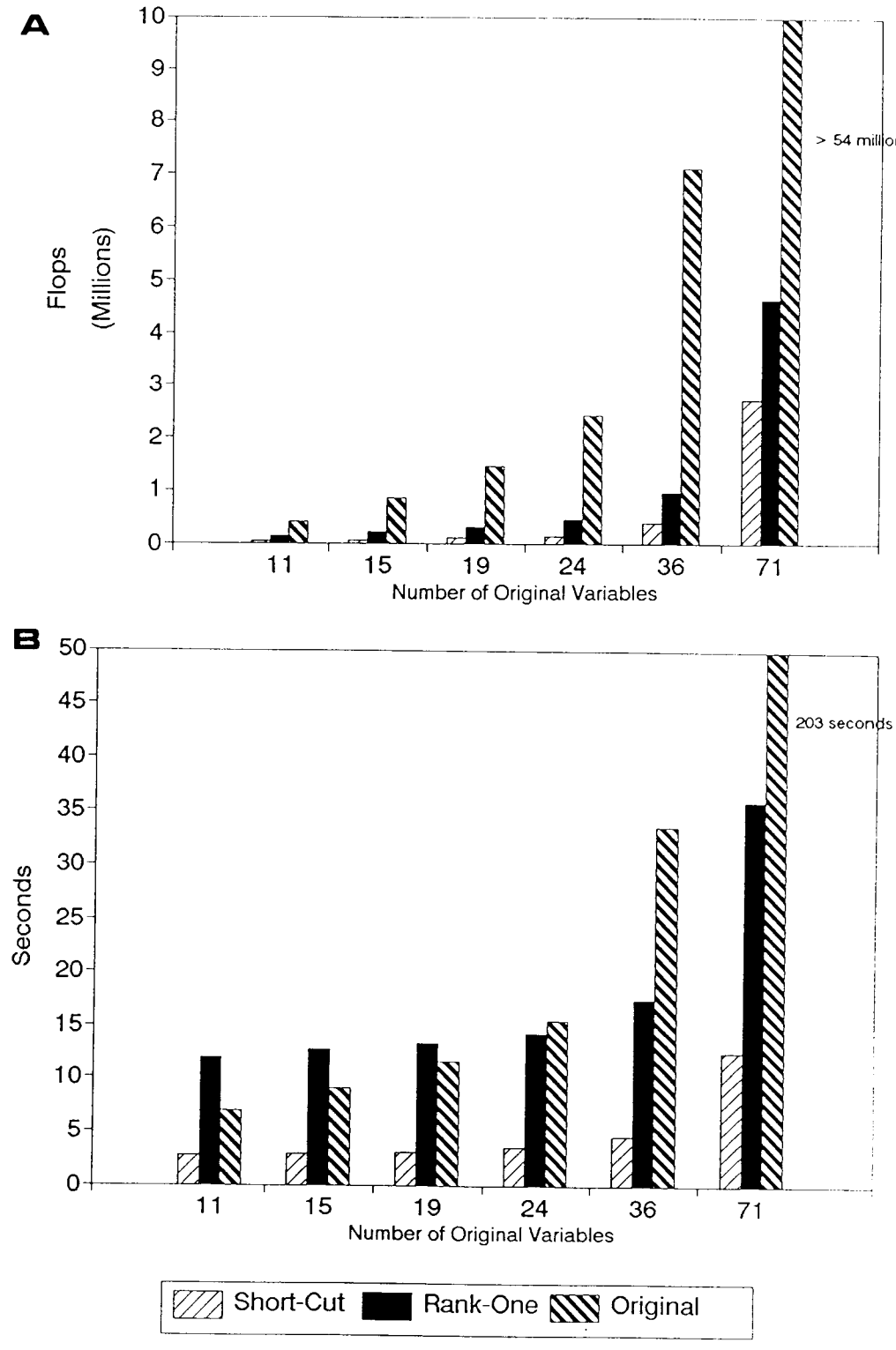


Figure 5.4. Procedure comparison graphs for calculating the discriminant function error rate based on the first four principal components for the two sapwood groups with ten observations per group. A) is the number of floating point operations and B) is the time in seconds to compute the error rates.

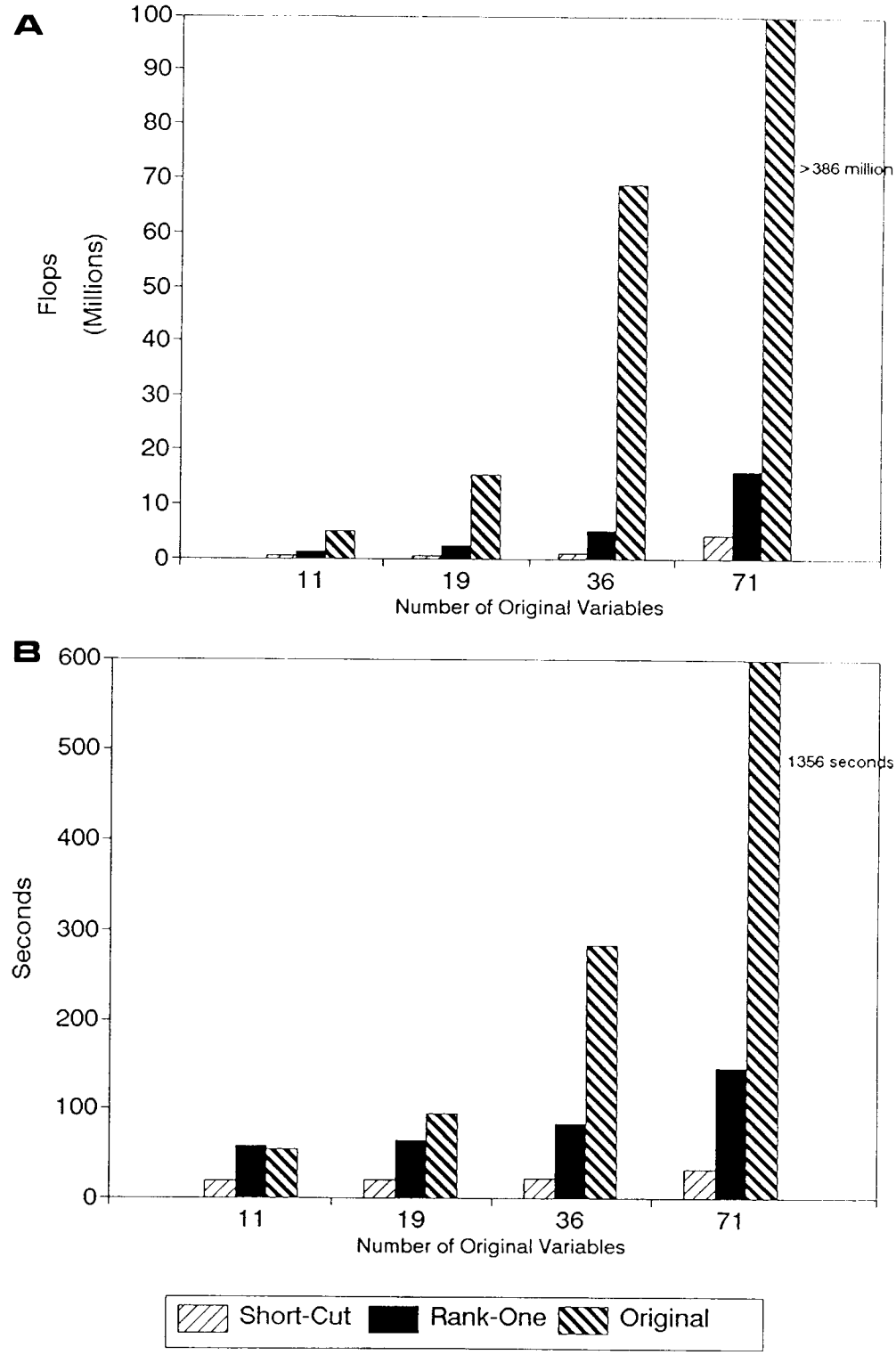


Figure 5.5. Procedure comparison graphs for calculating the discriminant function error rate based on the first three principal components for the two sapwood groups with 50 observations per group. A) is the number of floating point operations and B) is the time in seconds to compute the error rates.

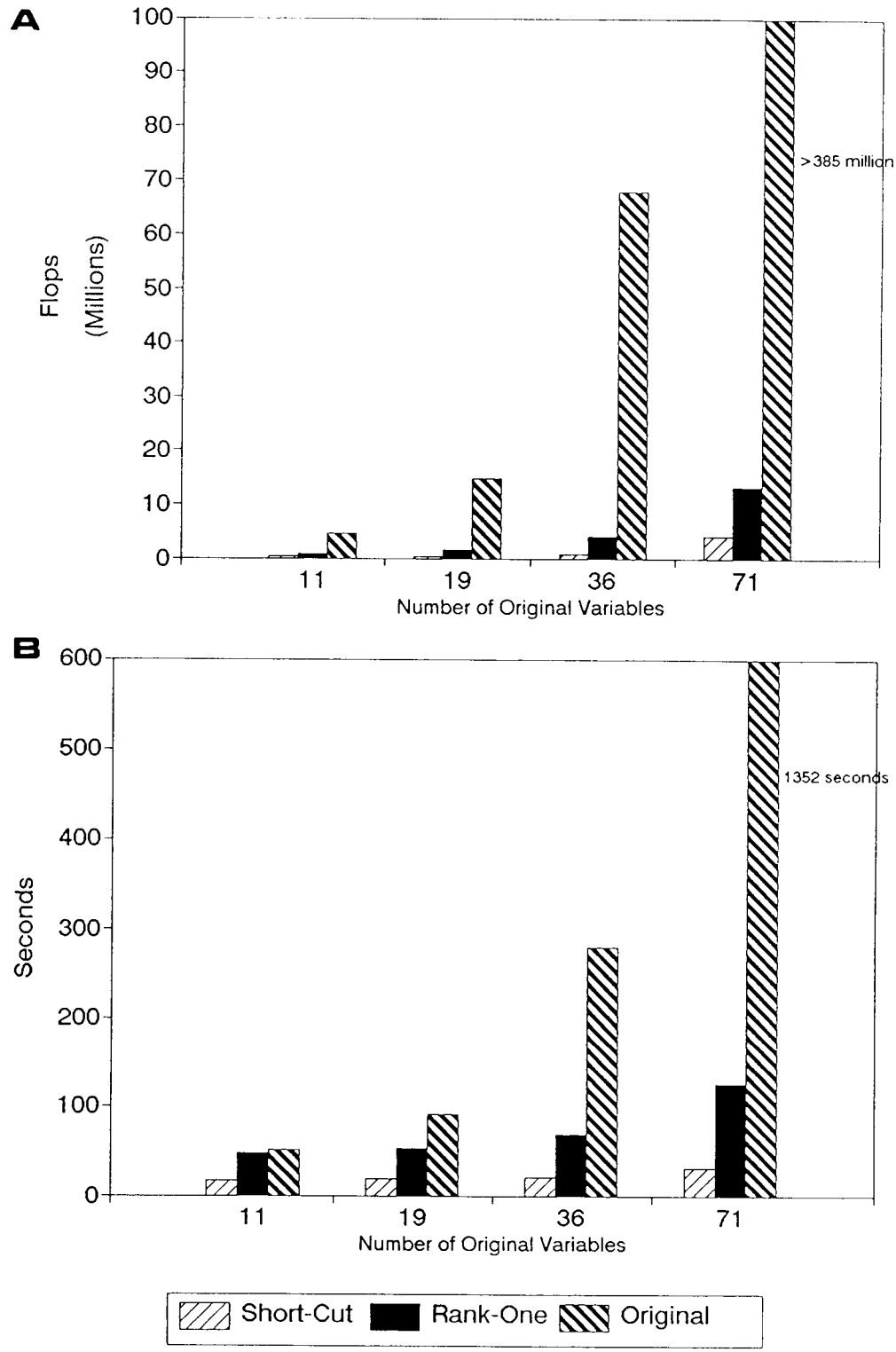


Figure 5.6. Procedure comparison graphs for calculating the discriminant function error rate based on the second and third principal components for the two sapwood groups with 50 observations per group. A) is the number of floating point operations and B) is the time in seconds to compute the error rates.

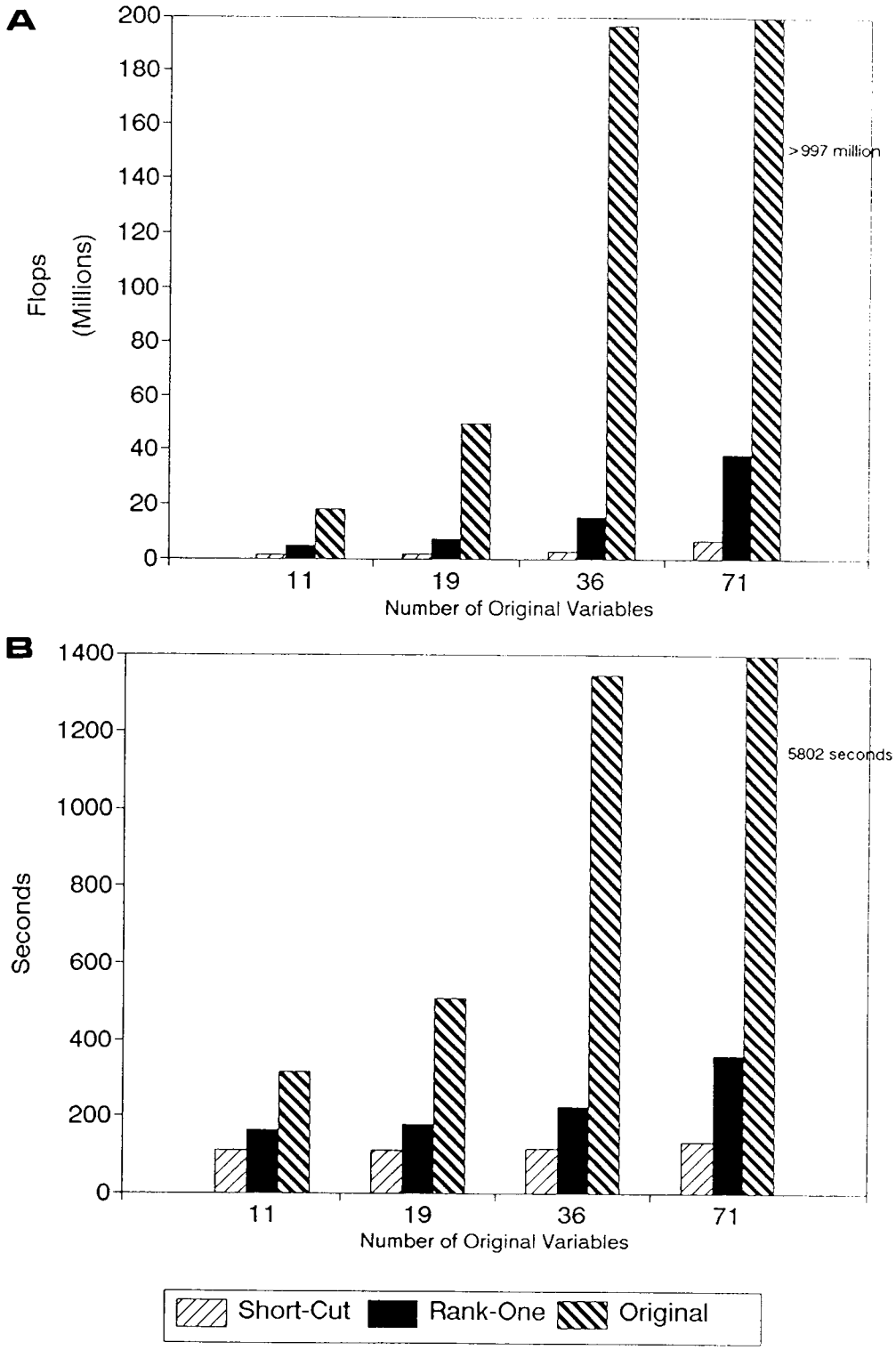


Figure 5.7. Procedure comparison graphs for calculating the discriminant function error rate based on the first three principal components for all four wood groups with 50 observations per group. A) is the number of floating point operations and B) is the time in seconds to compute the error rates.

Chapter 6
DISCUSSION OF SHORT-CUT ESTIMATOR

At the conclusion of Chapter 3 the short-cut estimator for the discriminant function's misclassification rate was illustrated for the wood data set. As seen in the last chapter, there are some slight differences in the estimates it produces versus the lengthy procedure of recalculation. However, for the wood data set, it generally performs satisfactorily considering its time savings. It can be envisioned that certain data sets could be troublesome for this estimator, especially if there are several influential outliers. Thus, a more theoretical discussion of its bias is called for.

Mean Square Error

To simplify the discussion it will be assumed that there are only two populations of interest, such as the spectral reflectances from sapwood earlywood and sapwood latewood or their dimensionally-reduced counterparts, that have equal prior probabilities. The overall short-cut cross-validation estimator of error rate will presumably give an estimate of the conditional misclassification probability, that is, assuming population normality, the probability that a random observation, $\mathbf{Y}$, from one of the populations $N_q(\gamma_1, \Psi)$ or $N_q(\gamma_2, \Psi)$ is incorrectly classified as coming from the other population based on the sample discriminant rule, $\widehat{\xi}$ ((2.9) in Chapter 2). This conditional misclassification probability has several names and is referred to as actual error rate (Snappin and Knoke, 1989, Rutter *et al.*, 1991) or conditional error rate (Snappin and Knoke, 1984, 1988, 1989).

For two normal q-variate populations, π_1 and π_2 with distributions $N_q(\gamma_1, \Psi)$ and $N_q(\gamma_2, \Psi)$, respectively, and with equal priors and equal costs of misclassification, the actual error rate is given by

$$\alpha(\widehat{\xi}) = \tfrac{1}{2}\,[1 - \alpha_2(\widehat{\xi}) + \alpha_1(\widehat{\xi})]$$

where

$$\alpha_i(\widehat{\xi}) = P\ (W(\mathbf{Y}) \leq 0 \mid \overline{\mathbf{Y}}_1\ ,\ \overline{\mathbf{Y}}_2\ ,\ \mathbf{S}_Y\ ,\ \mathbf{Y} \in \pi_i)\ .$$

$W(\cdot)$ was given as definition (2.9) in Chapter 2. Since $W(\mathbf{Y})$ is normally distributed, these probabilities are given by

$$\alpha_i(\widehat{\xi}) = \Phi\ [\ \frac{-[\gamma_i - (\overline{\mathbf{Y}}_1 + \overline{\mathbf{Y}}_2)/2]\ '\ \mathbf{S}_Y{}^{-1}\ (\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2)}{[\ (\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2)\ '\ \mathbf{S}_Y{}^{-1}\ \Psi \mathbf{S}_Y{}^{-1} (\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2)\ ]^{1/2}}\ ] \tag{6.1}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Conditional on the observed sample, this error rate is thus conditional on $\overline{\mathbf{Y}}_1$, $\overline{\mathbf{Y}}_2$ and $\mathbf{S}_Y$, the usual population means and common covariance matrix estimators. See Morrison (1976) for derivation.

Let $\widehat{\alpha}$ denote the short-cut leave-one-out estimator of actual error rate (or other estimator of error rate of interest) and $\alpha(\widehat{\xi})$ denote the actual error rate, as described above, based on the sample discriminant rule $\widehat{\xi}$ from a given training set. A common measure of performance is the mean square error (MSE)

$$\text{MSE} = E\ \{\ [\ \widehat{\alpha}\ -\ \alpha(\widehat{\xi})\ ]^{\,2}\ \}. \tag{6.2}$$

The expectation in (5.2) is taken over all possible training sets (of a particular sample size); this is why Snappin and Knoke (1984) call the MSE unconditional. The MSE measures the average closeness of $\widehat{\alpha}$ to $\alpha(\widehat{\xi})$ given the training sample for which the discriminant rule $\widehat{\xi}$ is calculated; and it is equal to (expected bias)^2 + total variance, where expected bias is the expected bias between the estimators and total variance is the variance of their difference.

Monte Carlo Sampling Results

Using Monte Carlo sampling, the MSE for the short-cut estimator can be evaluated by generating samples from known, p-multivariate normal populations with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and common covariance Σ with rank $q \leq p$, that is, $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$. Then applying a known orthogonal

transformation a system of rank q is obtained. These new variables are also normally distributed since they are linear combinations of the original variables, so (6.1) can be used in evaluating (6.2).

To investigate the bias in relation to data that has a similar structure as the wood data, the sapwood earlywood and sapwood latewood sample parameter estimates were treated as known normal population parameters, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$. The spectral decomposition of $\boldsymbol{\Sigma}$ is given by $\mathbf{VDV}'$. A random vector $\mathbf{z}$ is generated according to $N_p(\mathbf{0}, \mathbf{I})$, that is, a random, standard normal p-vector. If $\mathbf{x}_i = \mathbf{V}\mathbf{D}^{\frac{1}{2}}\mathbf{z} + \boldsymbol{\mu}_i$, where $\mathbf{D}^{\frac{1}{2}} = \text{diag}\,(\sqrt{d_1}, \cdots, \sqrt{d_p})$ and $i = 1$ or 2, then $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Let $n = n_1 + n_2$, where n_i is the number of observations from population i. In Matlab, for each sample n random, standard normal vectors were generated; then n_1 were transformed as $\mathbf{x}_1$ and n_2 were transformed as $\mathbf{x}_2$.

For each generated sample, first the necessary principal components were calculated, say q of them, and then the short-cut cross-validation estimate of discriminant analysis error rate was calculated from these q components. Based on the earlier assumptions of normality and that principal components are simply a linear transformation of the original data, the vector of q scores, $\mathbf{y}^*=\widehat{\mathbf{V}}_q{}'\mathbf{x}^*$, where $\mathbf{x}^*$ is a future observation from population i, given the generated sample is also normally distributed $N_q(\widehat{\mathbf{V}}_q{}'\boldsymbol{\mu}_i, \widehat{\mathbf{V}}_q{}'\boldsymbol{\Sigma}\widehat{\mathbf{V}}_q)$. Here $\widehat{\mathbf{V}}_q = [\widehat{\mathbf{v}}_1\ \widehat{\mathbf{v}}_2 \cdots \widehat{\mathbf{v}}_q]$ is the matrix of the first q eigenvectors from the decomposition of $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}$, the generated sample's pooled covariance matrix. To obtain the corresponding actual error rate for this generated data set $\boldsymbol{\gamma}_i = \widehat{\mathbf{V}}_q{}'\boldsymbol{\mu}_i$ and $\Psi = \widehat{\mathbf{V}}_q{}'\boldsymbol{\Sigma}\widehat{\mathbf{V}}_q$ and the estimates from the sample principal components, $\overline{\mathbf{Y}}_i$ and $\mathbf{S}_Y$, were substituted into (6.2).

The MSE was investigated by the above technique for the sapwood spectral reflectance measurements taken from wavelengths 400 to 1100 every 50 nanometers (p=15 wavelengths). Total sample sizes were n = 20, 30, 50, and 100 with equal group sizes $n_1 = n_2$ = 10, 15, 25, and 50. One thousand Monte Carlo samples (data sets) were generated for each of the sampling conditions. Figure 6.1 illustrates the behavior of the MSE as the number of principal components, q, kept varies from one to fifteen, especially the importance of larger sample sizes. On this scale, the number of components kept for discrimination has little effect for the larger group sizes.

Figures 6.2(A)-(D) show each MSE curve for group sizes 10, 15, 25 and 50, respectively, broken down into its respective parts of (expected bias)^2 and total variation. These graphs show that the major contributor to the MSE is the total variation. In fact, the variation associated with the cross-validation error rate estimator dominates. Interestingly, there is little correlation between the estimators on an individual sample basis. These graphs also show that for the smallest sample size the expected bias tends to increase with the number of components are kept for discrimination. This is to be expected as a large number of parameters are being estimated from relatively little information, especially in relation to the subspaces that are associated with smaller amounts of total variation. With 25 or 50 observations made from each population this effect disappears as these larger samples apparently contain sufficient information for more accurate estimation and the influence of individual observations is reduced.

As an illustration of the bias of the short-cut estimator when assuming normality, Figures 6.3(A)-(D) give the rank-one cross-validation, short-cut cross-validation, and expected actual error rate (true error rate) estimates from the above Monte Carlo sampling procedures for the respective group sizes. From these figures, the short-cut error rate appears more conservative than the true error rate, especially for the smallest sample size, while the rank-one cross-validation more accurately estimates the true error rate, although it also has a small amount of conservatism. The degree of conservatism disappears as sample size increases causing the short-cut estimator to become more accurate.

Based on the increasing characteristic of these error rate estimates after a subset of components is kept for discrimination, it is important not to keep too many components with a small amount of data. Even with a group size of 25, this phenomenon is still visible, but disappears when group size is increased to 50. It would interesting to incorporate variable selection into this process to assess effects on bias.

In the situation with a small number of observations, it appears that individual observations become more influential in the estimation of the higher order components which has an effect on the short-cut estimator. It is uncertain whether this increased influence is because of almost equal eigenvalues or almost equal to zero eigenvalues; in either case, when the entire

subspace is included the short-cut error rate estimate returns to the true estimate.

The above Monte Carlo sampling procedure was also carried out for the two groups, heartwood earlywood and heartwood latewood, to determine the effects on the MSE for groups with better separation and a farther distance between their means. The response of the MSE for the two groups of heartwood is given in Figure 6.4 for the group sizes 10, 15, 25, and 50. This graph illustrates that the MSE based on these two groups declines until the three principal component model, then levels off. For the group sizes of 10, the MSE begins to rise after a minimum at eight components; this is largely a result of the small amount of information available as more parameters are estimated (see Figure 6.5). Figures 6.6(A)-(D) show the bias. Only the small sample size case experiences much conservative biasing; the better separation between these groups promotes the use of the short-cut error-rate estimator.

To explore the effects of bias for response curves containing more measurements, the above sampling procedure was performed on the two sapwood groups with spectral reflectances simulated at 36 wavelengths. Figures 6.7(A)-(C) illustrate the results for the group sizes of 15, 25, and 50. For the smaller group sizes of 15 and 25, the short-cut cross-validation error rate estimator exhibits conservative bias. This bias disappears when the larger group size of 50 is considered.

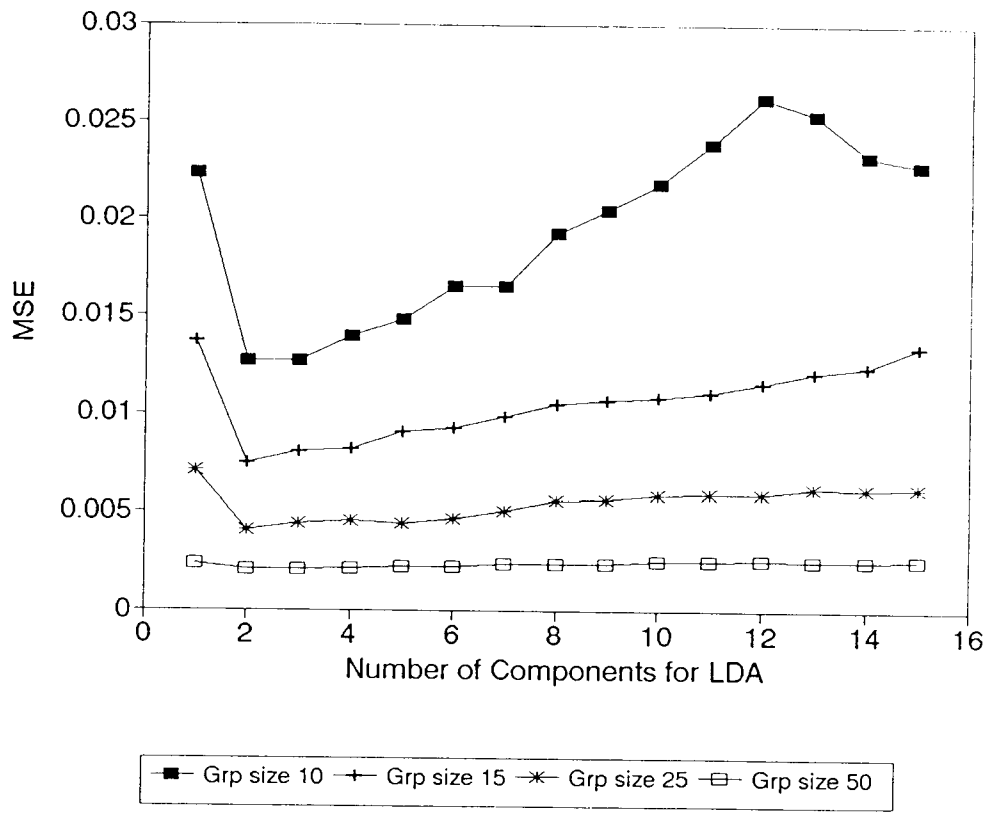


Figure 6.1. Mean square error for the short-cut cross-validation estimator of error rate based on the two groups of sapwood with 15 original measurements per response curve.

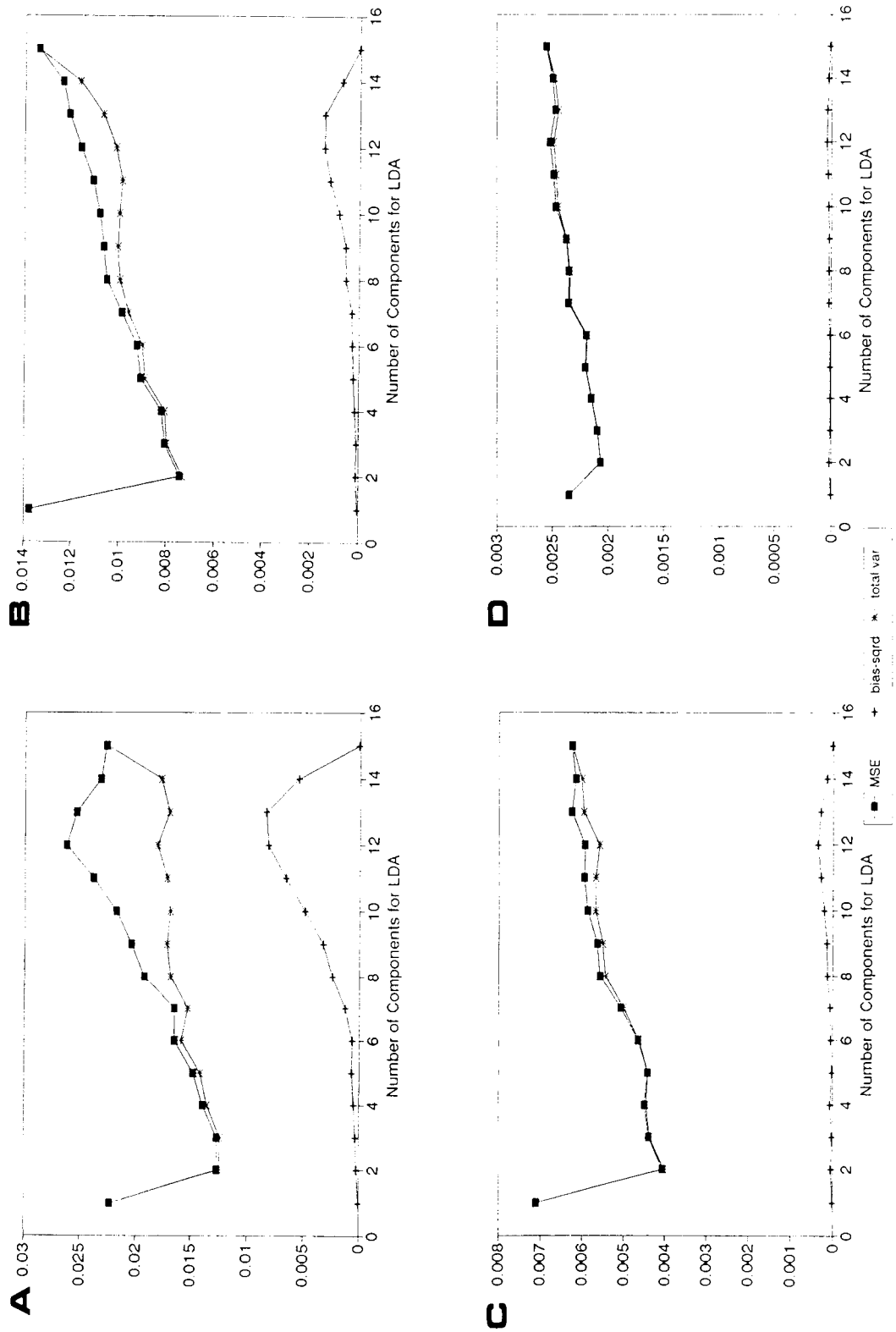


Figure 6.2. Breakdown of the mean square error of Figure 6.1 for the short-cut error rate estimator into bias-squared and total variance for the group sizes A) 10, B) 15, C) 25, and D) 50.

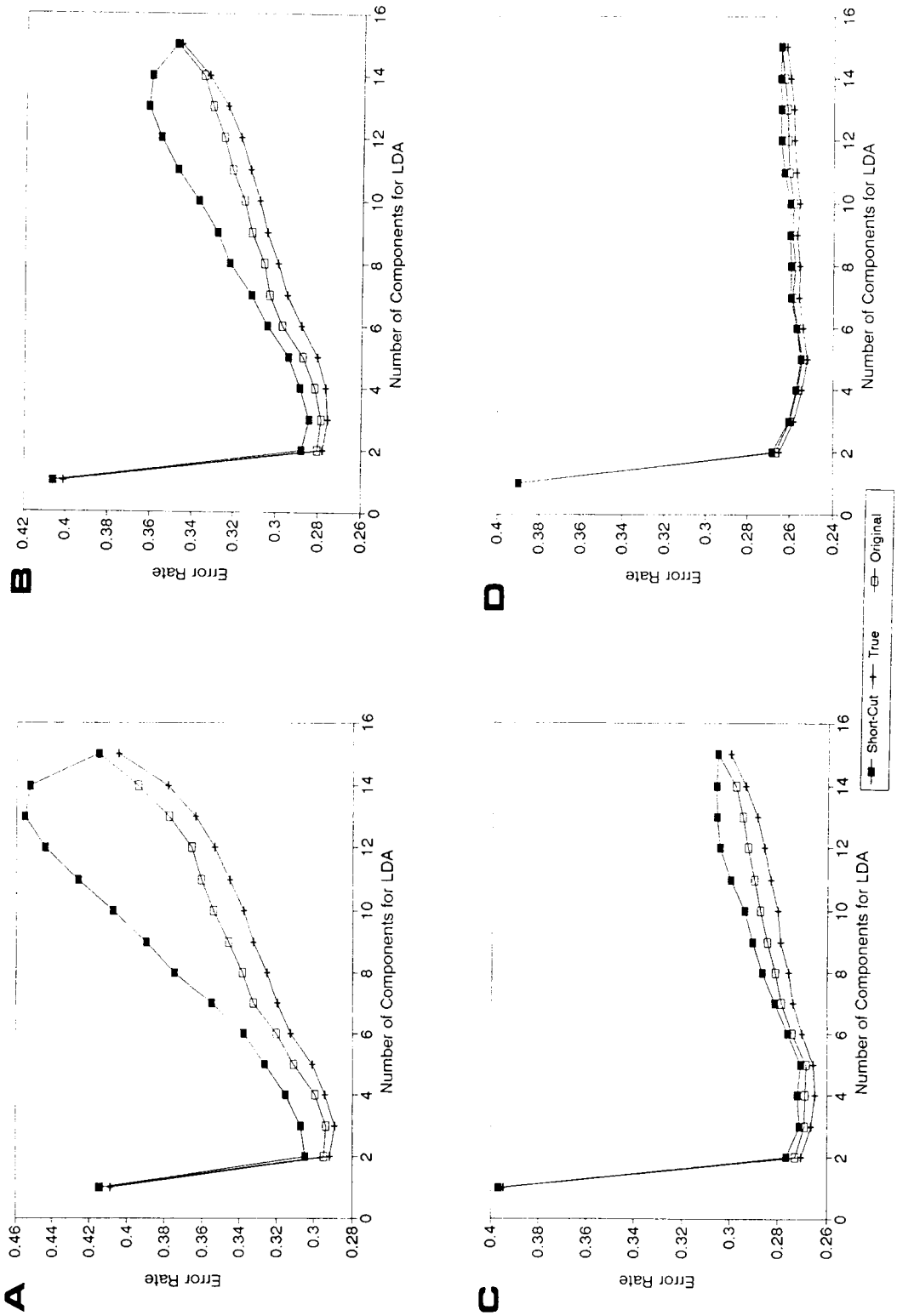


Figure 6.3. Monte Carlo results for the short-cut, true, and original error rate estimators based on the two sapwood groups for the group sizes A) 10, B) 15, C) 25, and D) 50.

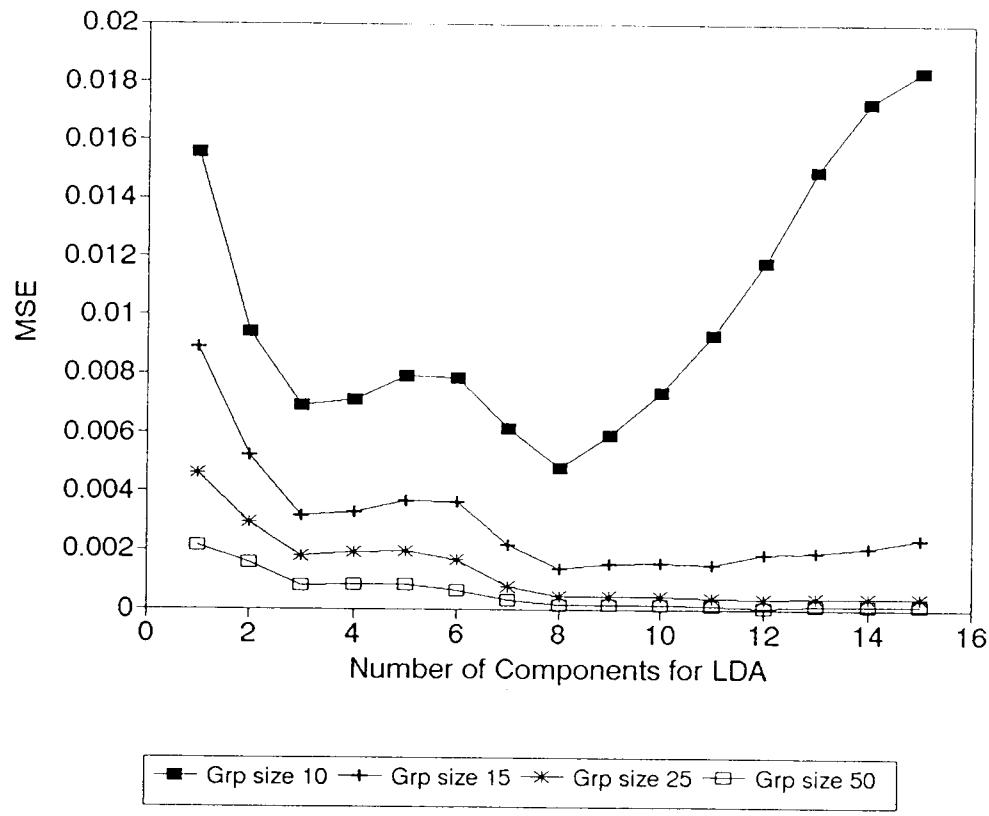


Figure 6.4. Mean square error for the short-cut cross-validation estimator of error rate based on the two groups of heartwood with 15 original measurements per response curve.

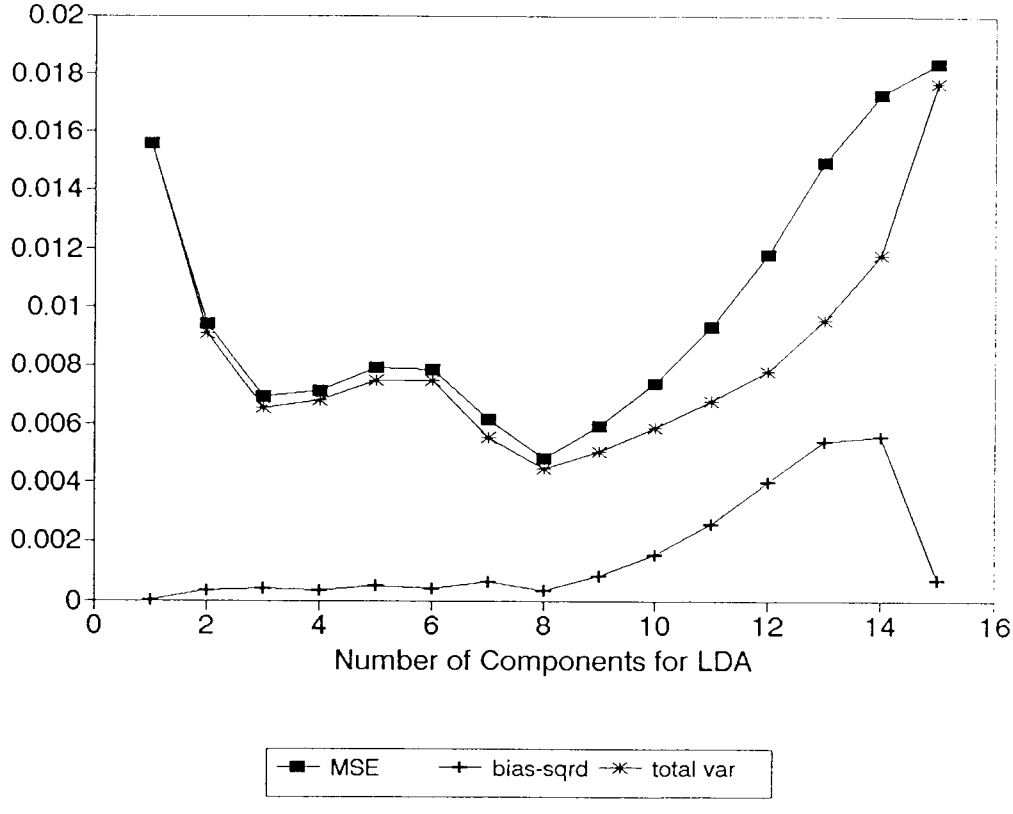


Figure 6.5. Breakdown of the mean square error of Figure 6.4 for the short-cut error rate estimator into bias-squared and total variance for the group size ten.

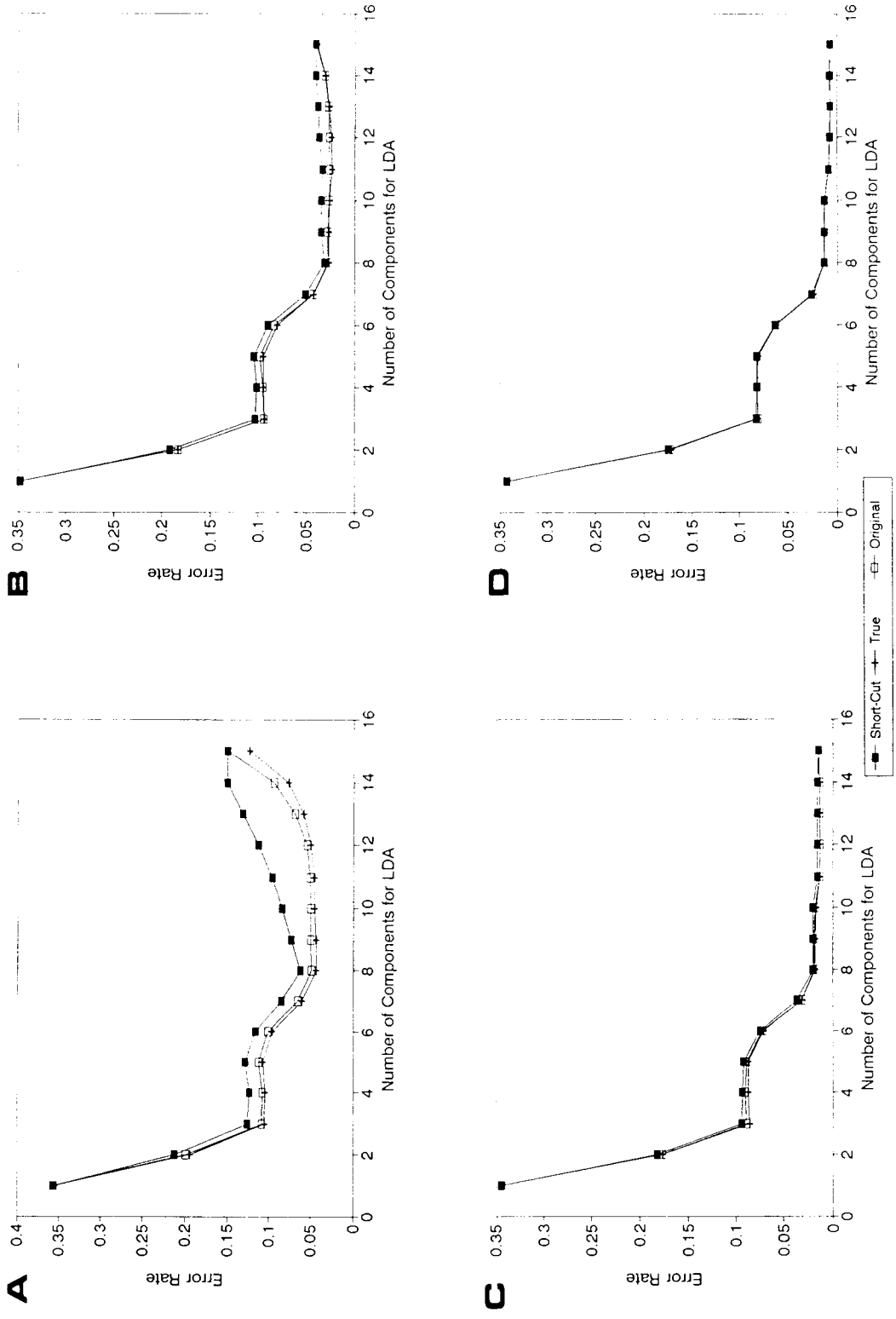


Figure 6.6. Monte Carlo results for the short-cut, true, and original error rate estimators based on the two heartwood groups for the group sizes A) 10, B) 15, C) 25, and D) 50.

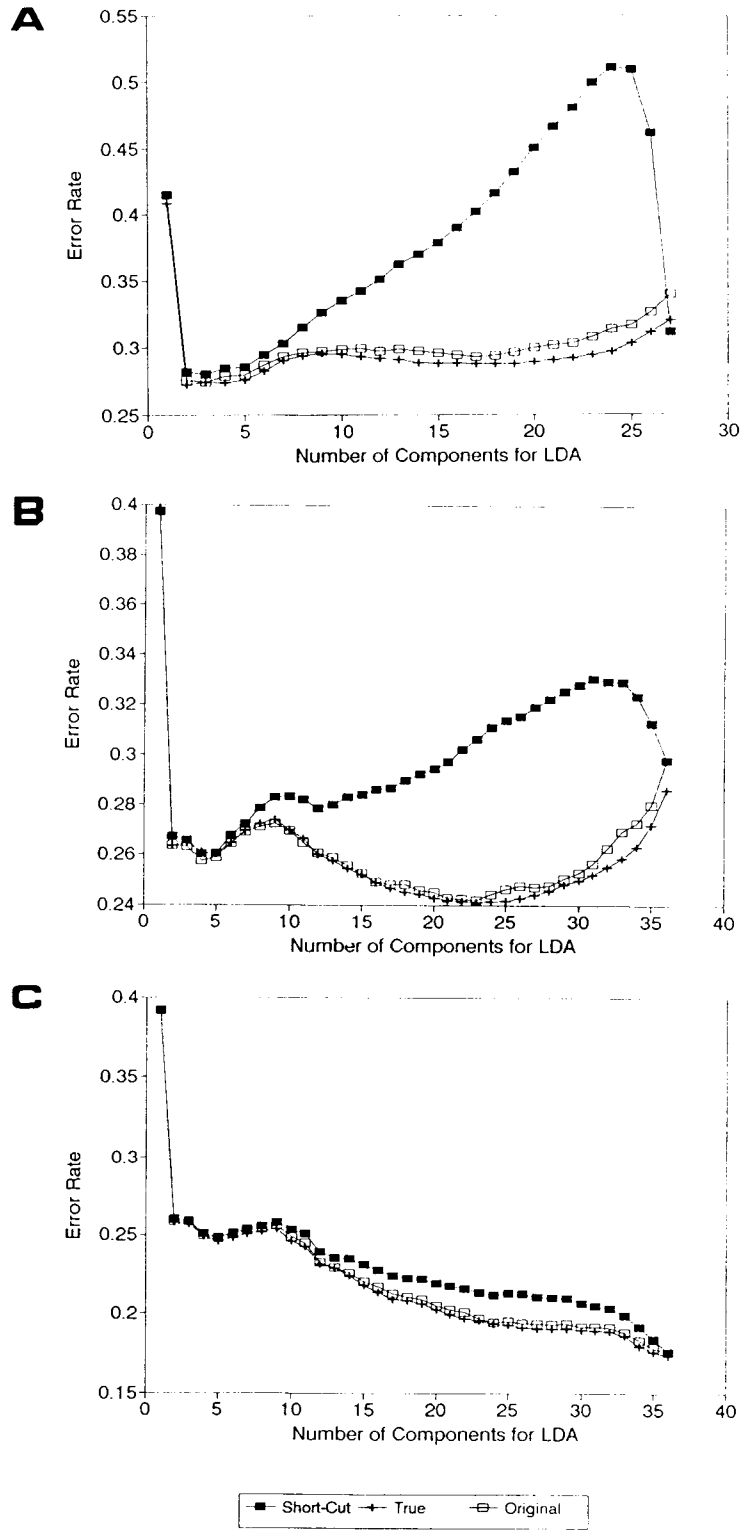


Figure 6.7. Monte Carlo results for the short-cut, true, and original error rate estimators based on the two sapwood groups with 36 original measurements per response curve for the group sizes A) 15, B) 25, and C) 50.

Chapter 7
DISCUSSION AND CONCLUSION

The goals of this thesis included reviewing methodologies for data reduction, modeling, and classification of grouped response curves, looking at the implications for a given data set, and offering computationally efficient alternatives.

A database of spectral reflectance curves of wood surface features was presented and analyzed in Chapter 3. A review of the literature, the analysis goals, and a preliminary statistical analysis of the data set led to the application of principal component analysis followed by linear discriminant analysis. For this data set, the various methods of determining the number of components to keep for modeling yielded widely different results. In addition, none of the methods are designed for the purpose of maintaining discriminating variation. Kshirsagar *et al.* (1990) provide a starting point for the two group case by suggesting that those eigenvectors that are orthogonal to the mean differences be thrown away.

In the application of linear discriminant analysis for this data set a problem of estimating discrimination misclassification rates (error rates) with the reduced data arose. In particular, performing the cross-validation estimation procedure for error rates required re-calculating the principal component decomposition and discriminant functions of the training sets, a very lengthy process. The alternative of performing the cross-validation procedure on the principal component scores without re-calculation of the principal component decomposition showed little difference in the error rate estimates for the presented data, but substantial improvements in computer time and operations.

In assuming common covariance structures between the populations, the pooled covariance matrix was used for decomposition in the principal component analysis. As described in Chapter 4, the leave-one-out cross-validation procedure results in a rank-one update in the pooled covariance matrix for each observation left out. Algorithms were developed for calculating

the updated eigenstructure under rank-one updates, and one particular algorithm was incorporated to calculate the orthogonal decomposition of the updated pooled covariance matrix. Use of the algorithm resulted in much faster computation of the estimated error rates than the lengthy process.

The error rate estimators for the wood data set under varying conditions, including changes in the number of variables in the spectral reflectance curve, the number of principal components kept for discrimination, the number of groups (wood surface feature types), and the number of observations (sample size), were compared in Chapter 5. In most instances, the faster alternative of performing leave-one-out cross-validation directly on the principal components without recomputating the principal component analysis for each observation performed satisfactory, especially considering the time savings and simplicity involved.

Although the short-cut estimator performed almost equivalently as the original estimator for this data set, it was unknown if this was happenstance or related to the distribution of the data set. Chapter 6 explored the bias and variance of the estimators using simulated normally distributed data that exhibited characteristics similar the observed spectral reflectance data. Under the assumption of normality the short-cut estimator appeared to be somewhat conservatively biased when the sample size was small in comparison to the number of original variables. However, the variance of the estimators really overshadowed any differences that may exist. This again implies that nothing is really lost by using the short-cut estimator for the wood data set.

Further exploration of this bias and variance of the estimators may prove useful, as it seems that data that is not normally distributed could cause problems. For example, data sampled from a population with a heavily-tailed distribution would be more likely to have several observations that could readily influence the principal component decomposition, which in turn might overly influence the discriminant functions.

It would also be interesting to study effects on the bias and variance of the estimators in association with variable selection procedures available for discriminant analysis. And if there exist instances where the short-cut estimator seems inappropriate, then further research could be conducted on the rank-one procedure.

BIBLIOGRAPHY

Anderson, T.W. (1984), *An Introduction to Multivariate Analysis*, John Wiley and Sons, New York.

Bartlett, M. S. (1951), "An Inverse Matrix Adjustment Arising in Discriminant Analysis," *Annals of Math. Statistics*, **22**, 107-111.

Biscay, R., Valdes, P., Pascual, R. (1990), "Modified Fisher's Linear Discriminant Function With Reduction of Dimensionality," *J. Statist. Comput. Simul.*, **36**, 1-8.

Brunner, C. C., Maristany, A. G., Butler, D. A., VanLeeuwen, D., and Funck, J. W., (1992), "An Evaluation of Color Spaces for Detecting Defects in Douglas-fir Veneer," *Industrial Metrology*, **2**(3&4), 169-184.

Brunner, C. C., Shaw, G. B., Butler, D. A., and Funck, J. W. (1990), "Using Color in Machine Vision Systems for Wood Processing," *Wood and Fiber Science*, **22**, 413-428.

Buchsbaum, G. and Gottschalk, A. (1984), "Chromaticity Coordinates of Frequency-Limited Functions," *J. Opt. Soc. Am. A*, **8**, 885-887.

Bunch, J. R. and Nielsen, C. P. (1978), "Updating the Singular Value Decomposition," *Numer. Math.*, **31**, 111-129.

Bunch, J. R., Nielsen, C. P., and Sorensen, D. C. (1978), "Rank-One Modification of the Symmetric Eigenproblem," *Numer. Math.*, **31**, 31-48.

Cheng, Y-Q., Zhuang, Y-M., and Yang, J-Y. (1992), "Optimal Fisher Discriminant Analysis Using the Rank Decomposition," *Pattern Recognition*, **25**(1), 101-111.

Church, A. Jr. (1966), "Analysis of Data When the Response is a Curve," *Technometrics*, **8**, 229-246.

Cohen, J. (1964), "Dependency of the Spectral Reflectance Curves of the Munsell Color Chips," *Psychon. Sci.*, **L**, 369-370.

Cuppen, J. J. M. (1981), "A Divide and Conquer Method for the Symmetric Eigenproblem," *Numer. Math.*, **36**, 177-195.

DeGroat, R. D. and Roberts, D. A. (1990), "Efficient, Numerically Stabilized Rank-One Eigenstructure Updating," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, **38**, 301-316.

Dongarra, J. J. and Sorensen, D. C. (1986), "Linear Algebra on High Performance Computers," *Appl. Math. and Comp.*, **20**, 57-88.

Dongarra, J. J. and Sorensen, D. C. (1986), "A Fully Parallel Algorithm for the Symmetric Eigenvalue Problem," *SIAM J. Sci. and Stat. Comp.*, **8**, S139-S154.

Eastment, H. T. and Krzanowski, W. J. (1982), "Cross-Validatory Choice of the Number of Components From a Principal Component Analysis," *Technometrics*, **24**, 73-77.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, **7**, 1-26.

Efron, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *J. Am. Stat. Assoc.*, **78**, 316-331.

Flury, B. (1988), *Common Principal Components and Related Multivariate Models*, John Wiley and Sons, New York.

Forest Products Lab, U.S.F.S (1974), *Wood Handbook: Wood as An Engineering Material*, U.S. Agriculture Handbook No. 72.

Friedman, J. H. (1989), "Regularized Discriminant Analysis," *J. Am. Stat. Assoc.*, **84**, 165-175.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, Academic Press, Inc., San Diego, CA.

Gershon, R. (1987), *The Use of Color in Computation Vision*, PhD. thesis, Univ. of Toronto, Tech. Report RBCV-TR-87-15.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley and Sons, New York.

Golub, G. H. (1973), "Some Modified Matrix Eigenvalue Problems," *SIAM Review*, **15**, 318-334.

Golub, G. H. and Van Loan, C. F. (1989), *Matrix Computations*, John Hopkins University Press, Baltimore.

Green, P. E. (1978), *Analyzing Multivariate Data*, Dryden Press, Hinsdale, Illinois.

Grum, F. and Wightman, T. (1960), "Measurement of the Contribution of Fluorescence to the Brightness of Papers Treated with Whitening Agents," *Tappi*, **43**, 400-405.

Habbema, J.D.F. and Hermans, J. (1977), "Selection of Variables in Discriminant Analysis by F-statistic and Error Rate," *Technometrics*, **19**, 487-493.

Hastie, T. and Stuetzle, W. (1989), "Principal Curves," *J. Am. Stat. Assoc.*, **84**, 502-516.

Healey, G. (1989), "Using Color for Geometry-Insensitive Segmentation," *J. Opt. Soc. Am. A*, **6**, 920-937.

Healey, G. and Binford, T. O. (1987), "The Role and Use of Color in a General Vision System," *Proc. ARPA Image Understanding Workshop*, DARPA, 599-613.

Hemel, J. B. and Van Der Voet, H. (1986), "The CLAS Program for Classification and Evaluation," *Analytica Chimica Acta*, **191**, 33-45.

Hoogerbrugge, R., Willig, S. J., Kistemaker, P. G. (1983), "Discriminant Analysis by Double Stage Principal Component Analysis," *Anal. Chem.*, **55**, 1710-1712.

Institute of Mathematical Statistics (1990), "Cross-Disciplinary Research in the Statistical Sciences," *Statistical Science*, **5**, 121-146.

Justice, J. B., Jr., and Isenhour, T. L. (1975), "Factor Analysis of Mass Spectra," *Analytical Chemistry*, **47**, 2286-2288.

Kowalski, B. R. and Wold, S. (1982), "Pattern Recognition in Chemistry," in *Handbook of Statistics 2*, P. R. Krishnaiah and L. N. Kanal, editors, North-Holland Publishing, Amsterdam.

Krzanowski, W. J. (1979), "Between-Groups Comparison of Principal Components," *J. Am. Stat. Assoc.*, **74**, 703-707.

Krzanowski, W. J. (1983), "Cross-Validatory Choice in Principal Component Analysis; Some Sampling Results," *J. Statist. Comput. Simul.*, **18**, 299-314.

Krzanowski, W. J. (1984), "Principal Component Analysis in the Presence of Group Structure," *Appl. Statist.*, **33**, 164-168.

Krzanowski, W. J. (1986), "Cross-Validation in Principal Component Analysis," in *The XIIIth Inter. Biometrics Conf. Proceedings*, Univ. Wash., Seattle, 1-15.

Krzanowski, W. J. (1987), "Cross-Validation in Principal Component Analysis," *Biometrics*, **43**, 575-584.

Kshirsagar, A. M., Kocherlakota, S., and Kocherlakota, K. (1990), "Classification Procedures Using Principal Component Analysis and Stepwise Discriminant Function," *Commun. Statist.-Theory Meth.*, **19**(1), 91-109.

Lachenbruch, P. A. (1967), "An Almost Unbiased Method of Obtaining Confidence Intervals for the Probability of Misclassification in Discriminant Analysis," *Biometrics*, **23**, 639-645.

Lachenbruch, P.A. and Mickey, M.R. (1968), "Estimation of Error Rates in Discriminant Analysis," *Technometrics*, **10**, 1-11.

Lorber, A. and Kowalski, B. (1990), "Alternatives to Cross-Validatory Estimation of the Number of Factors in Multivariate Calibration," *Appl. Spectroscopy*, **44**, 1464-1470.

MacAdam, D. L. (1985), *Color Measurement Theme and Variations*, Springer-Verlag, Berlin, Germany.

Malinowski, E. R. and Howery, D. G. (1980), *Factor Analysis in Chemistry*, John Wiley and Sons, New York.

Maloney, L. T. (1986), "Evaluation of Linear Models of Surface Spectral Reflectance With Small Numbers of Parameters," *J. Opt. Soc. Amer. A*, **3**(10), 1673-1683.

Manly, B. F. J. and Rayner, J. C. W. (1987), "The Comparison of Sample Covariance Matrices Using Likelihood Ratio Tests," *Biometrika*, **74**, 841-847.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1982), *Multivariate Analysis*, Academic Press, Orlando, Florida.

Maristany, A. G., Butler, D. A., Brunner, C. C., and Funck, J. W. (1991), "Exploiting Local Color Information for Defect Detection on Douglas-fir Veneer," *Proceedings of the Fourth International Conference of Scanning Technology in Sawmilling*, Miller Freeman Publications, San Francisco.

Maristany, A. G., Lebow, P. K., Brunner, C. C., Butler, D. A., and Funck, J. W. (1992), "Classifying Wood-Surface Features Using Dichromatic Reflection," *SPIE Conference on Optics in Agriculture and Forestry*, The International Society for Optical Engineering, Bellingham, WA.

Massart, D. L., Dijkstra, A., and Kaufman, L. (1978), *Evaluation and Optimization of Laboratory Methods and Analytical Procedures*, Elsevier Scientific Publishing, Amsterdam, Netherlands.

MATLAB User's Guide, (1989), The MathWorks, Inc., South Natick, MA.

Moon, P. and Spencer, D. E. (1945), "Polynomial Representation of Reflectance Curves," *J. Opt. Soc. Am.*, **35**(9), 597-600.

Morris, R. H. and Morrissey, J. H. (1954), "An Objective Method for Determination of Equivalent Neutral Densities of Color Film Images. II. Determination of Primary Equivalent Neutral Densities," *J. Opt. Soc. Am.*, **44**, 530-534.

Morrison, D. F. (1976), *Multivariate Statistical Methods*, McGraw-Hill, New York.

Nilsson, N. J. (1965), *Learning Machines*, McGraw-Hill, New York.

Novak, C. L. and Shafer, S. A. (1991), "Supervised Color Constancy for Machine Vision," *SPIE Vol. 1453 Human Vision, Visual Processing, and Digital Display II*, 353-368.

Panshin, A. J. and de Zeeuw, C. (1980), *Textbook of Wood Technology*, McGraw-Hill Book Co., New York.

Parkkinen, J. P. S., Hallikainen, J. and Jaaskelainen, T. (1989), "Characteristic Spectra of Munsell Colors," *J. Opt. Soc. Am. A*, **6**(2), 318-322.

Parkkinen, J. P. S. and Jaaskelainen, T. (1987), "Color Representation Using Statistical Pattern Recognition," *Applied Optics*, **26**(19), 4240-4245.
Peck, R., and Van Ness, J. (1982), "The Use of Shrinkage Estimators in Linear Discriminant Analysis," *IEEE Trans. Pat. Anal. Mach. Intell.*, **4**, 530-537.

Randles, R. H., Broffitt, J. D., Ramberg, J. S., and Hogg, R. V. (1978), "Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates," *J. Am. Stat. Assoc.*, **73**, 564-568.

Rao, C. R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhyā A.*, **26**, 329-358.

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

Rao, C. R. and Mitra (1971), *Generalized Inverse of Matrices and Its Applications*, John Wiley and Sons, New York.

Rozett, R. W. and Petersen, E. M. (1975), "Methods of Factor Analysis of Mass Spectra," *Analytical Chemistry*, **47**, 1301-1308.

Rozett, R. W. and Petersen, E. M. (1976), "Classification of Compounds by the Factor Analysis of Their Mass Spectra," *Analytical Chemistry*, **48**, 817-825.

Rutter, C., Flack, V., and Lachenbruch, P. (1991), "Bias in Error Rate Estimates in Discriminant Analysis When Stepwise Variable Selection is Employed," *Commun. Statist. - Simul.*, **20**(1), 1-22.

Seber, G.A.F. (1984), *Multivariate Observations*, John Wiley and Sons, New York.

Sharaf, M. A., Illman, D. L., and Kowalski, B. R. (1986), *Chemometrics*, John Wiley and Sons, New York.

Simonds, J. L. (1963), "Application of Characteristic Vector Analysis to Photographic and Optical Response Data," *J. Opt. Soc. Am.*, **53**, 968-974.

Snapinn, S. M. and Knoke, J. D. (1984), "Classification Error Rate Estimators Evaluated by Unconditional Mean Squared Error," *Technometrics*, **26**, 371-378.

Snapinn, S. M. and Knoke, J. D. (1988), "Bootstrapped and Smoothed Classification Error Rate Estimators," *Commun. Statist. - Simul.*, **17**, 1135-1153.

Snapinn, S. M. and Knoke, J. D. (1989), "Estimation of Error Rates in Discriminant Analysis with Selection of Variables," *Biometrics*, **45**, 289-299.

Snee, R. D. (1972), "On the Analysis of Response Curve Data," *Technometrics*, **14**, 47-62.

Spiegelman, C. H. (1989), "Chemostatistics," *Challenges for the 90's*, publication of the American Statistical Association.

Stiles, W.S., Wyszecki, G. and Ohta, N. (1977), "Counting Metameric Object-Color Stimuli Using Frequency-Limited Spectral Reflectance Functions," *J. Opt. Soc. Am.*, **67**(6), 779-784.

Tominaga, S. and Wandell, B. A. (1989), "Standard Surface-Reflectance Model and Illuminant Estimation," *J. Opt. Soc. Am. A*, **6**, 576-584.

Wilkinson, J. H. (1965), *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford.

Wold, H. (1966), "Nonlinear Estimation by Iterative Least Squares Procedures," in *Research Papers in Statistics*, F. N. David editor, John Wiley and Sons, London.

Wold, S. (1976), "Pattern Recognition by Means of Disjoint Principal Components Models," *Pattern Recognition*, **8**, 127-139.

Wold, S. (1978), "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics*, **20**, 397-405.

Wold, S. and Sjostrom, M. (1977), "SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy," in *Chemometrics: Theory and Application*, B.R. Kowalski editor, ACS Symposium Series 52, Washington, D.C.

Young, R. A. (1986), "Principal-component Analysis of Macaque Lateral Geniculate Nucleus Chromatic Data," *J. Opt. Soc. Am. A*, **3**, 1735-1742.

Yu, K.-B. (1991), "Recursive Updating the Eigenvalue Decomposition of a Covariance Matrix," *IEEE Trans. on Signal Proc.*, **39**(5), 1136-1145.

APPENDIX

APPENDIX

This appendix will show how the number of floating point operations (flops) for certain steps of the algorithms were derived. These were given in Chapter 3. Since the programs were written in MATLAB, the calculations are based, in part, on MATLAB's implementations of specific functions.

Suppose there are g groups from which n_j observations are gathered in each group, $j = 1, \cdots, g$, and that the total number of observations is $n = \sum_{j=1}^{g} n_j$. Further suppose that each observation is a p-vector, and that q principal components are chosen for discrimination. It will first be shown that the number of flops necessary to calculate the unbiased estimate of the pooled covariance matrix is $(2n + 3g)p^2 + 4np + 3g$.

For each group sample covariance matrix, $(2n_j + 1)p^2 + 4n_jp + 1$ flops are required. $4n_jp$ flops are necessary to adjust the observations for the mean, $2n_jp^2$ are necessary to 'square' the mean-corrected observation matrix, and p^2+1 are necessary to divide the sum-of-squares matrix by $n_j - 1$. See the MATLAB M-file cov.m.

The pooled covariance matrix then requires an additional $2gp^2 + 2g$ flops. Thus the total number of flops to calculate the pooled covariance matrix is $\sum_{j=1}^{g} (2n_jp^2 + 4n_jp + p^2 + 1) + 2gp^2 + 2g = (2n + 3g)p^2 + 4np + 3g$. When one observation from group j is left out, this reduces to $(2n + 3g - 2)p^2 + 4(n - 1)p + 3g$.

To estimate the posterior probability for the left-out observation based on the linear discriminant function the number of flops needed is $2q^3 + (2n + 5g - 2)q^2 + (5n + 3g - 5)q + 9g - 1$. This includes $(2n + 3g - 2)q^2 + (n - 1)\, 4q + 3g$ flops to compute the pooled covariance matrix, approximately $2q^3$ flops to compute the inverse of the pooled covariance matrix (a closer approximation is $13q^3/6$, see Golub and Van Loan, 1989), and $2gq^2 + (n + 3g - 1)q + 6g - 1$ flops to compute posterior probabilities. Note that if the squared Mahalanobis distance from the left-out observation $\mathbf{y}$ to the i-th group

mean is given by $d_i^2(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}}_i)' \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}}_i)$, then the posterior probability that $\mathbf{y}$ belongs to group i is given by

$$p(i|\mathbf{y}) = \frac{\exp\{-\frac{1}{2} d_i^2(\mathbf{y})\}}{\sum_{j=1}^{g} \exp\{-\frac{1}{2} d_j^2(\mathbf{y})\}} .$$