Dimension reduction in PCA : likelihood-based methods
by Kamolchanok Choochaow

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics
Montana State University

Abstract:
The main objective of this thesis is to develop procedures for making inferences about the eigenvalues and eigenvectors of a covariance matrix. Specifically, new procedures for examining dimension reduction in principal component analysis (PCA) are developed. The dimension reduction consists of the following two aspects: reduction in the number of components and reduction in the number of original variables. The procedures are based on a likelihood approach. Parameterizations of eigenvalues and eigenvectors are presented. The parameterizations allow arbitrary eigenvalue multiplicities. The use of the Fisher scoring algorithm for computing maximum likelihood estimates of the covariance parameters subject to multiplicity and other constraints is discussed. Asymptotic distributions of estimators of covariance parameters are derived under normality and non-normality. Likelihood ratio tests and Bartlett corrections are described. Simulation studies show the effectiveness of Bartlett corrections. The new procedures are demonstrated to give better overall results than some existing methods.

DIMENSION REDUCTION IN PCA:

LIKELIHOOD-BASED METHODS

by

Kamolchanok Choochaow

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

July 2002

APPROVAL

of a dissertation submitted by

Kamolchanok Choochaow

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Robert J. Boik Robert J. Boik (Signature) 7/22/02 Date

Approved for the Department of Mathematical Sciences

Kenneth L. Bowers Kenneth L. Bowers (Signature) 7/25/02 Date

Approved for the College of Graduate Studies

Bruce McLeod Bruce R. McLeod (Signature) 7-25-02 Date

STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to Bell & Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature Kawlebank Chambers

Date July, 22 2002

ACKNOWLEDGEMENTS

I would like to thank my adviser, Professor Robert J. Boik, for his invaluable assistance and guidance during the preparation of this thesis. I truely appreciate his encouragement and patience.

I am also grateful for John Borkowski, Jim Robinson-Cox, Steve Cherry, and Bill Quimby's advice throughout my graduate studies.

Lastly, I would like to thank my dad, mom, and husband for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The main objective of this thesis is to develop procedures for making inferences about the eigenvalues and eigenvectors of a covariance matrix. Specifically, new procedures for examining dimension reduction in principal component analysis (PCA) are developed. The dimension reduction consists of the following two aspects: reduction in the number of components and reduction in the number of original variables. The procedures are based on a likelihood approach. Parameterizations of eigenvalues and eigenvectors are presented. The parameterizations allow arbitrary eigenvalue multiplicities. The use of the Fisher scoring algorithm for computing maximum likelihood estimates of the covariance parameters subject to multiplicity and other constraints is discussed. Asymptotic distributions of estimators of covariance parameters are derived under normality and non-normality. Likelihood ratio tests and Bartlett corrections are described. Simulation studies show the effectiveness of Bartlett corrections. The new procedures are demonstrated to give better overall results than some existing methods.

CHAPTER 1

INTRODUCTION

The main objective of this thesis is to develop procedures for making inferences about the eigenvalues and eigenvectors of a covariance matrix. Specifically, new procedures for examining dimension reduction in principal component analysis (PCA) are developed. The procedures are based on a likelihood approach.

The first chapter of this thesis briefly describes principal component analysis followed by a discussion of dimensionality reduction methods. Two examples are given. In Chapter 2, parameterizations of eigenvalues and eigenvectors are presented. Two eigenvalue parameterizations are adopted under different objectives of use. The parameterizations allow arbitrary eigenvalue multiplicities. The Newton-Raphson algorithm is used to solve implicit parameter equations. The hypothesis test of redundancy is described and an eigenvector parameterization under redundancy is proposed.

Chapter 3 describes the use of the Fisher scoring algorithm for computing maximum likelihood estimates of the covariance parameters subject to multiplicity and other constraints. Asymptotic distributions of estimators of covariance parameters are derived under normality and non-normality. In Chapter 4, likelihood ratio tests and Bartlett corrections are described. The results of simulations that examine the

effectiveness of Bartlett corrections are presented in Chapter 5. Two numerical examples also are illustrated.

Principal Component Analysis

Investigators often measure or make observations on a large number of variables. There are several useful techniques to reduce the dimensionality of data without the loss of much information such as factor analysis and cluster analysis. Principal component analysis also is one such technique and is one of the most widely-used multivariate techniques.

Typically, principal component analysis is used to reduce the dimensionality of a data set, while retaining as much of the original information as possible. This is achieved by transforming the original set of variables into a smaller set of linear combinations called principal components. These new variables are uncorrelated and ordered so that the first principal component accounts for the largest proportion of the variation present in the original set of variables. The usual objective of the analysis is to determine whether the first few components account for most of the variation in the original data. If they do, then these components can be used to summarize the data with little loss of information. This dimension reduction can be useful in simplifying subsequent analyses.

Before defining principal components, some notation will be described. Matrices will be denoted by boldface upper case letters. Vectors will be denoted by boldface

lower case letters. A diagonal matrix having diagonal elements $a_1, a_2, \ldots, a_n$ is denoted by $\text{diag}(a_1, a_2, \ldots, a_n)$ and the trace of matrix $\boldsymbol{A}$ is denoted by $\text{tr}(\boldsymbol{A})$. The $p \times p$ identity matrix is denoted by $\boldsymbol{I}_p$.

Suppose that the vector of original variables $\boldsymbol{y} = (y_1 \ y_2 \ \cdots \ y_p)'$ has a positive definite covariance matrix $\boldsymbol{\Sigma}$. For mathematical convenience and without loss of generality, assume that the mean of y_i is zero for all $i = 1, 2, \ldots, p$. Because $\boldsymbol{\Sigma}$ is symmetric and positive definite, all eigenvalues of $\boldsymbol{\Sigma}$ are real and positive. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ be the ordered eigenvalues of $\boldsymbol{\Sigma}$ and let $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1 \ \boldsymbol{\gamma}_2 \ \cdots \ \boldsymbol{\gamma}_p)$ be a $p \times p$ orthogonal matrix such that

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}', \tag{1.1}$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$. That is, $\boldsymbol{\gamma}_i$ is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to the eigenvalue λ_i. The principal components of $\boldsymbol{y}$ are the entries of the p-vector $\boldsymbol{z}$, where $\boldsymbol{z}$ is the linear combination that can be written as

$$\boldsymbol{z} = \boldsymbol{\Gamma}'\boldsymbol{y}. \tag{1.2}$$

The covariance matrix of $\boldsymbol{z}$, $\boldsymbol{\Sigma}_z$, can be written as

$$\boldsymbol{\Sigma}_z = \boldsymbol{\Gamma}'\boldsymbol{\Sigma}\boldsymbol{\Gamma}.$$

Substituting $\boldsymbol{\Sigma}$ from (1.1) yields

$$\boldsymbol{\Sigma}_z = \boldsymbol{\Gamma}'(\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}')\boldsymbol{\Gamma} = \boldsymbol{\Lambda}.$$

Hence the components $z_1 = \gamma_1'\boldsymbol{y},\ z_2 = \gamma_2'\boldsymbol{y}, \ldots, z_p = \gamma_p'\boldsymbol{y}$ are uncorrelated. The variance of z_i is λ_i. The maximum value of variance of $\gamma_1'\boldsymbol{y}$ satisfying $\gamma_1'\gamma_1 = 1$ is equal to λ_1, the largest eigenvalue of $\boldsymbol{\Sigma}$. This maximum occurs when γ_1 is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to λ_1. That is,

$$\max_{l'l=1} \boldsymbol{l'\Sigma l} = \gamma_1'\boldsymbol{\Sigma}\gamma_1 = \lambda_1.$$

Then, the component z_1 has the largest variance λ_1, z_2 has the second largest variance λ_2, and so on. The total variance of the p principal components is equal to the total variance of the original variables so that

$$\sum_{i=1}^{p} \lambda_i = \mathrm{tr}(\boldsymbol{\Sigma}).$$

Consequently, the j^{th} principal component accounts for a proportion

$$t = \frac{\lambda_j}{\mathrm{tr}(\boldsymbol{\Sigma})}$$

of the total variance of the original variables.

This thesis focuses on two major aspects of dimensionality reduction:

1. Reduction in the number of components.
2. Reduction in the number of original variables.

Reduction in the Number of Components

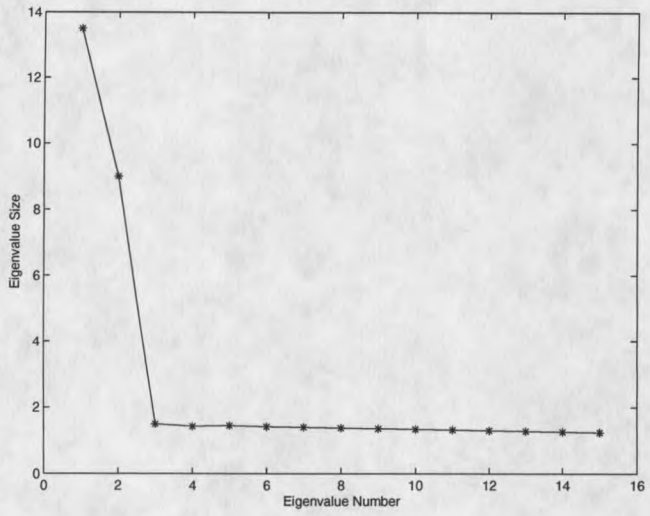
The major problem in reducing the number of components is deciding how many principal components should be retained in order to account for most of the variability

in the original data. Some rules of thumb have been suggested. Kaiser (1958) suggested a cut off point to retain the components whose sample variances are greater than the average, i.e. greater than 1 for a correlation matrix. It can be argued that if one variable is nearly independent of all others variables, it will dominate a component with variance slightly less than 1 assuming that variables have been standardized. Since this variable provides independent information from the other variables, there is no reason to discard it. Kaiser's rule is arbitrary and tends to retain too few components (Jolliffe, 1986). Rencher (1995) suggested that investigators should retain sufficient components to account for a specified percentage of the total variance. Figures between 70 and 90 percent are generally suggested.

Three methods are in common use for determining the number of components to be retained. The first method is graphical and is called a scree graph. This technique was discussed and named by Cattell (1966). The scree graph consists of a plot of the ordered eigenvalues against eigenvalue numbers. With this plot, the components to be retained correspond to the eigenvalues plotted in the steep curve above the straight line formed by the smaller eigenvalues. Thus in Figure 1, the first two components would be retained. Jolliffe (1986) suggested plotting $\log(\boldsymbol{\lambda})$ rather than $\boldsymbol{\lambda}$. This plot is known as a log-eigenvalue diagram.

The second method is testing the hypothesis that the last $m = p-k$ eigenvalues of $\boldsymbol{\Sigma}$ are equal. Anderson (1963) derived the likelihood ratio test for H_{0k}: $\lambda_{k+1} = \cdots = \lambda_p = \lambda$ against the alternative that at least two of the last m eigenvalues are different.

Figure 1. Ideal Scree Graph.



Let $y_1, y_2, \ldots, y_N$ be a random sample of size $N = n+1$ from a p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. If $\boldsymbol{y}$ can be written as $\boldsymbol{C}\boldsymbol{w} + \boldsymbol{\varepsilon}$, where $\boldsymbol{C}$ is a $p \times k$ matrix with rank k, $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_p$, and $\mathrm{var}(\boldsymbol{w}) = \boldsymbol{I}_k$, then $\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}' + \sigma^2 \boldsymbol{I}_p$. Write $\boldsymbol{C}\boldsymbol{C}'$ in diagonal form as $\boldsymbol{C}\boldsymbol{C}' = \boldsymbol{\Gamma}_1 \boldsymbol{\Lambda}_1 \boldsymbol{\Gamma}_1'$ and let $\boldsymbol{\Gamma}_2$ be an orthogonal complement to $\boldsymbol{\Gamma}_1$. That is, $\boldsymbol{I}_p = \boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1' + \boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2'$ and $(\boldsymbol{\Gamma}_1 \quad \boldsymbol{\Gamma}_2) \in \mathcal{O}(p)$, where $\mathcal{O}(p)$ denotes group of orthogonal $p \times p$ matrices. The covariance matrix of $\boldsymbol{y}$ can be rewritten as

$$\begin{aligned}
\boldsymbol{\Sigma} &= \boldsymbol{\Gamma}_1 \boldsymbol{\Lambda}_1 \boldsymbol{\Gamma}_1' + \sigma^2 \boldsymbol{I}_p \\
&= \boldsymbol{\Gamma}_1 \boldsymbol{\Lambda}_1 \boldsymbol{\Gamma}_1' + \sigma^2 (\boldsymbol{\Gamma}_1 \boldsymbol{\Gamma}_1' + \boldsymbol{\Gamma}_2 \boldsymbol{\Gamma}_2') \\
&= \boldsymbol{\Gamma}_1 (\boldsymbol{\Lambda}_1 + \sigma^2) \boldsymbol{\Gamma}_1' + \boldsymbol{\Gamma}_2 \sigma^2 \boldsymbol{\Gamma}_2'.
\end{aligned}$$

Therefore, the first k eigenvalues are $\lambda_i+\sigma^2,\ i=1,2,\ldots,k$, where $\mathbf{\Lambda}_1=\text{diag}(\lambda_i),\ i=1,2,\ldots,k$, and the last $m=p-k$ eigenvalues are σ^2. Anderson derived the test for testing the equality of the last m eigenvalues. The test statistic is

$$\mathbf{\Lambda}_k=\left\{\prod_{i=k+1}^{p} l_i/\left(\sum_{i=k+1}^{p} l_i/m\right)^m\right\}^{n/2}, \tag{1.3}$$

where $l_i,\ i=1,2,\ldots,p$, are eigenvalues of the sample covariance matrix, $\boldsymbol{S}$. If H_{0k} is true, then the limiting distribution of $-2\ln\mathbf{\Lambda}_k$ is χ^2_ν, where $\nu=\frac{1}{2}(m+2)(m-1)$. These results can be summarized as

$$T_1=n\left[m\ln\bar{l}-\sum_{i=k+1}^{p}\ln l_i\right]\sim\chi^2_\nu+O_p(n^{-\frac{1}{2}}),$$

where $\bar{l}=\sum_{i=k+1}^{p} l_i/m$.

Bartlett (1954) improved Anderson's test by using a multiplying factor. Bartlett's test is to reject H_{0k} whenever

$$T_2=\left(n-k-\frac{2m^2+m+2}{6m}\right)\left[m\ln\bar{l}-\sum_{i=k+1}^{p}\ln l_i\right]\geq\chi^2_{1-\alpha,\nu},$$

where $\chi^2_{1-\alpha,\nu}$ is the 100(1-α) percentile of the χ^2_ν distribution.

Lawley (1956) proposed a further improvement in Bartlett's multiplying factor. Lawley's test is to reject H_{0k} whenever

$$T_3=\left(n-k-\frac{2m^2+m+2}{6m}+\sum_{i=1}^{k}\frac{\bar{l}^2}{(l_i-\bar{l})}\right)\left[m\ln\bar{l}-\sum_{i=k+1}^{p}\ln l_i\right]\geq\chi^2_{1-\alpha,\nu}.$$

Acceptance of H_{0k} suggests that each of the last m components contain the same amount of information. If these eigenvalues are very small, then little information is

lost by discarding the corresponding principal components. The first k components should be retained for further analysis. A sequence of tests can be conducted starting with $k = 0$ and increasing k until the null hypothesis is accepted.

Bentler and Yuan (1996) developed a method of testing the linear trend in the last $m = p - k$ eigenvalues of the covariance matrix. The hypothesis of interest is H_{0m}: $\lambda_{k+i} = \alpha + \beta x_i,\ x_i = m - i,\ i = 1, 2, \ldots, m$. The likelihood ratio test statistic is

$$\Lambda_m = \frac{\underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}_0}{\text{Sup}}\ \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)}{\underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}}{\text{Sup}}\ \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})},$$

where $\boldsymbol{\Sigma}_0 = \boldsymbol{\Gamma}_0 \boldsymbol{\Lambda}_0 \boldsymbol{\Gamma}_0'$ and $\boldsymbol{\Lambda}_0 = \text{diag}(\lambda_1, \ldots, \lambda_k, \alpha + \beta x_1, \alpha + \beta x_2, \ldots, \alpha + \beta x_m)$. For $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}$, $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is maximized when $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$. To maximize $\ell(\bar{\boldsymbol{y}}, \boldsymbol{\Sigma}_0)$, Bentler and Yuan showed that $\hat{\lambda}_i = \frac{n}{n+1} l_i$ for $i = 1, 2, \ldots, k$ are the maximum likelihood estimators (MLEs) of $\lambda_1, \ldots, \lambda_k$ under H_{0m}. Maximum likelihood estimates of α and β were numerically obtained by solving the following equations:

$$\begin{aligned} \sum_{j=1}^{m} \frac{[n\, l_{k+j} - (n+1)(\alpha + \beta x_j)]}{(\alpha + \beta x_j)^2} &= 0, \\ \sum_{j=1}^{m} \frac{[n\, l_{k+j} - (n+1)(\alpha + \beta x_j)] x_j}{(\alpha + \beta x_j)^2} &= 0. \end{aligned}$$

The asymptotic distribution of $-2 \ln \Lambda_m$ is χ^2_ν with $\nu = m - 2$. If all x_i are equal to zero, then the hypothesis H_{0m} is equivalent to H_{0k} as considered by Anderson (1963).

Boik (2002a) proposed a method for modeling the eigen-structure of several covariance matrices simultaneously. The proposed model for the i^{th} covariance matrix

is as follows:

$$\mathbf{\Sigma}_i = \mathbf{\Gamma}_i \mathbf{\Lambda}_i \mathbf{\Gamma}_i', \ i = 1, 2, \ldots, g,$$

where the eigenvector matrices $\{\mathbf{\Gamma}_i\}$ and eigenvalue matrices $\{\mathbf{\Lambda}_i\}$ may share certain properties. The model extends common principal components (Flury, 1988) and subsumes Anderson's (1963) model, Lawley's (1956) model and Bentler and Yuan's (1996) model as special cases when g=1. Boik presented an algorithm to compute maximum likelihood estimates of covariance parameters and also gave likelihood ratio tests including Bartlett corrections. Several of his results are referred to throughout this thesis.

The third method is based on the cumulative proportion of total variance. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are the ordered eigenvalues, then the cumulative proportions of the first k eigenvalues, δ_k, and the last $m = p - k$ eigenvalues, δ_m, are as follows:

$$\delta_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \quad \text{and} \quad \delta_m = 1 - \delta_k = \frac{\sum_{i=k+1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

Let $\hat{\delta}_k = \dfrac{\sum_{i=1}^{k} l_i}{\sum_{i=1}^{p} l_i}$ and $\hat{\delta}_m = \dfrac{\sum_{i=k+1}^{p} l_i}{\sum_{i=1}^{p} l_i}$, where $l_i, \ i = 1, 2, \ldots, p,$ are defined in (1.3).

Anderson (1963) proposed a test statistic for the hypothesis,

$$\mathrm{H}_{0m}\colon \delta_m = \delta_{m,0}, \tag{1.4}$$

where $\delta_{m,0}$ is a known constant. If δ_m is small, then little information is lost by discarding the corresponding principal components. Anderson showed that if $\lambda_k >$

λ_{k+1}, then $\hat{\delta}_m$ is asymptotically distributed as

$$\sqrt{n}(\hat{\delta}_m - \delta_m) \xrightarrow{L} N\left(0, \frac{2\delta_m^2 \sum_{i=1}^{k} \lambda_i^2 + 2(1-\delta_m)^2 \sum_{i=k+1}^{p} \lambda_i^2}{(\sum_{i=1}^{p} \lambda)^2}\right). \tag{1.5}$$

The test rejects H_{0m} whenever $|z^*| \geq z_{\alpha/2}$, where

$$z^* = \frac{\hat{\delta}_m - \delta_{m,0}}{\sqrt{\hat{V}(\hat{\delta}_m)}}, \tag{1.6}$$

$z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution, and $\hat{V}(\hat{\delta}_m)$ is an estimator of the variance of $\hat{\delta}_m$ which can be computed by replacing λ_i by l_i in the variance of $\hat{\delta}_m$ in (1.5).

Sugiyama and Tong (1976) studied an approximate distribution of the cumulative proportion of the first k eigenvalues, δ_k. Sugiyama discussed that this quantity is interpreted as a measure of the amount of information in the k retained components. Assuming that the eigenvalues are distinct, he derived a perturbation expansion for the distribution of $\hat{\delta}_k$ that is accurate to $O(n^{-3/2})$.

Huang and Tseng (1992) devised the following method of selecting the number of components to be retained. Let $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i = \delta_k$ and

$$\Omega_q(\omega) = \{\boldsymbol{\lambda};\ \delta_q \geq \omega\},$$

where ω is a fixed constant and q is an integer less than p. Their decision rule is to select k components, where k is the smallest integer that satisfies

$$\hat{\delta}_k \geq c,$$

where c is a fixed constant to be determined. They discussed how to determine N and c such that the minimum probability of retaining the important components is at least some specified level P^*, $0 \leq P^* \leq 1$. That is,

$$\min_k \left\{ \inf_{\boldsymbol{\lambda} \in \Omega_k(\omega)} P_{\boldsymbol{\lambda}}(\hat{\delta}_k \geq c) \right\} \geq P^*. \tag{1.7}$$

Huang and Tseng showed that if $\omega \geq 0.5$, $k \leq \frac{p}{2}$, and N and c satisfy

$$c \leq \omega\{1 - (1-\omega)\sqrt{2p/(p-1)(N-1)}\Phi^{-1}(P^*)\},$$

where $\Phi(x)$ is the c.d.f of the standard normal distribution, then (1.7) holds.

In this thesis, a likelihood approach for making inferences about the cumulative proportion of total variance is given. The hypothesis of interest is

$$\mathrm{H}_0 \colon \frac{\boldsymbol{C}_1'\boldsymbol{\lambda}}{\mathrm{tr}(\boldsymbol{\Sigma})} = \boldsymbol{c}_0, \tag{1.8}$$

where $\boldsymbol{C}_1$ is a matrix of known constants and $\boldsymbol{c}_0$ is a vector of known constants. This hypothesis is more flexible and general than the hypothesis in (1.4). If $\boldsymbol{C}_1$ is $(\mathbf{0}_k' \quad \mathbf{1}_{p-k}')'$ and $\boldsymbol{c}_0$ is a scalar, then the hypothesis in (1.4) is subsumed as a special case of the hypothesis in (1.8). Specific goals are to obtain maximum likelihood estimates (MLEs) of covariance parameters, to construct a likelihood ratio test of the hypothesis, and to construct confidence intervals for $\frac{\boldsymbol{C}_1'\boldsymbol{\lambda}}{\mathrm{tr}(\boldsymbol{\Sigma})}$. The MLEs of parameters and a likelihood ratio test for the hypothesis in (1.8) are given using one of the parameterizations of eigenvalues presented in Chapter 2. The parameterizations allow for arbitrary eigenvalues multiplicities. A Bartlett correction is derived to improve the

likelihood ratio test. The tests are inverted to obtained confidence intervals. The likelihood ratio intervals and Bartlett-corrected intervals are compared with Anderson's intervals inverted from (1.6) by means of simulations. The results of the simulations are described in Chapter 5.

Reduction in the Number of Original Variables

A principal component is a linear combination of all of the original variables. Thus, interpretation and subsequent data analysis involve all of the variables even if some components are discarded. Not only is it useful to reduce the number of components but it also is useful to reduce the number of the original variables. If the first k principal components have zero coefficients on a set of original variables, then those variables are redundant. These redundant variables can be eliminated from the study and need not be collected in any subsequent studies. The term "redundancy" has been employed in other contexts , e.g. multivariate regression (Lazraq, 2001), discriminant analysis (Fujikoshi, 1985), canonical correlation (Van Den Wollenberg, 1977; Gleason, 1976), and growth curve analysis (Fujikoshi and Rao, 1991). The term takes a different meaning in each of these contexts. In this thesis, the principal components of $\boldsymbol{y}$ are defined as in (1.2). It follows that

$$\boldsymbol{y} = \boldsymbol{\Gamma z}.$$

The coefficient for obtaining the i^{th} variable from the j^{th} principal component is γ_{ji} and is called a loading. A redundant variable is defined to be an original variable

whose loadings on the first k principal components are zero.

Two approaches for determining the number of variables and which variables to retain will be discussed, namely the rejection approach and the hypothesis testing approach. A rejection approach is a set of cut-off rules for choosing how many and which variables to be retained. A hypothesis test is a method of inference to decide which of two complementary hypotheses is true. The rejection approach will be described first.

Beale, Kendall and Mann (1967) developed cut-off rules that specify the number of variables to be discarded. However, they focused primarily on multiple regression analysis rather than on principal component analysis. In multiple regression with p independent variables, they suggested selecting the set of r variables that maximizes the multiple correlation of the dependent variable with the r independent variables. An extension of this rule to principal components is to retain set of r variables which maximizes the minimum multiple correlation between r selected variables and any of $p - r$ discarded variables. They also mentioned another rejection approach for use in principal component analysis. With all p variables, p principal components are obtained. If p_1 eigenvalues are smaller than some number, λ_0, then the last p_1 components are inspected. Starting with the one corresponding to the smallest eigenvalue, then the next component corresponding to the second smallest eigenvalue and so forth, the variable that has the largest coefficient in the component and has not already been deleted by a previously considered component is deleted and the

number of variables is reduced from p to $p-p_1$. Another principal component analysis is done on the remaining $p-p_1$ variables. Similarly, if p_2 eigenvalues are smaller than the λ_0, p_2 variables associated with the largest coefficients in the last p_2 components are rejected. The process continues until all eigenvalues are greater than the λ_0. The number of variables is reduced from p to $r = (p - p_1 - \cdots - p_k)$. The value of r is determined by choice of λ_0. Jolliffe (1972) discussed the appropriate λ_0 based on simulation studies.

Jolliffe (1972) discussed eight rejection methods for deciding which variables to be discarded. The eight rejection methods were divided into three groups: namely multiple correlation methods (A1 and A2), principal component methods (B1, B2, B3, and B4), and cluster methods (C1 and C2). These methods are briefly described below.

- Multiple Correlation Methods
  - Method A1 is the method of Beale, Kendall and Mann (1967).
  - Method A2 is a stepwise method. In multiple regression with p independent variables, a subset of independent variables is selected , in each step, such that the variable having maximum multiple correlation with the remaining variables is deleted. The process is repeated until r variables are retained.
- Principal Component Methods
  - Method B1 also is based on Beale, Kendall and Mann (1967) using PCA.

- Method B2 is the same as Method B1 except that only one principal component analysis is done. If k variables are to be retained, then the last $p-k$ components are inspected. For each of these $p-k$ components, starting with the last component, the variable that has the largest coefficient in the component and has not already been deleted by a previously considered component is deleted.
- Method B3 uses the last $p-k$ components. The sum of squares of coefficients of each of the p variables in the last $p-k$ components are computed. The $p-k$ variables that have the largest sum of squares are rejected.
- Method B4 uses the first k components. For each of k components, starting with the first component, the variable that has the largest coefficient in the component and has not already been selected by a previously considered component is selected. Variables that are not selected are rejected.

- Cluster Methods
  - Method C1 is a single-linkage method (Seber, 1984). A measure of similarity between two clusters of variables X and Y is defined by r_{xy} such that

$$r_{xy} = \max r_{ij}, \tag{1.9}$$

where r_{ij} is the correlation coefficient between variable i and j. For each of $p(p-1)/2$ pairs of p variables, r_{xy} are computed. Two clusters having the maximum r_{xy} are combined into a single cluster. For the new set of

clusters, the process returns to calculate r_{xy} and so forth. The process continues until k clusters of variables remain. The principal component analysis is based on k variables, one chosen from each cluster. Jolliffe described 3 ways to select a variable from each cluster.

- Method C2 is an average-linkage method (Seber, 1984). It follows the same steps as Method C1 except that r_{xy} in (1.9) is replaced with

$$r_{xy} = \frac{\sum_{i \in X} \sum_{j \in Y} r_{ij}}{p_1 p_2},$$

where p_1, p_2 are the numbers of variables in X and Y, respectively.

These rejection methods were tested on artificial data. The artificial data were constructed so that certain variables were linear combination of other variables. For example, if $y_1 = y_2 + e$, where e is a random disturbance, then either y_1 or y_2 may be discarded without loss of information. It was shown that several rejection methods eliminated precisely those certain variables. In general, Method A2 is the most consistent method because it only retained good or best subsets. Method B4 retained best subsets more often than Method A2 but it retained moderate or bad sets fairly frequently.

Jolliffe (1973) discussed the five methods, A1, B2, B4, C1, and C2 which had been successfully used on artificial data in Jolliffe (1972). He applied these methods on four sets of real data. The methods were equally effective with real and artificial data, although none of rejection methods was apparently better than the others.

McCabe (1984) proposed methods to select a subset of variables called principal

variables that contains as much information as possible. McCabe gave four optimality criteria to evaluate all possible subsets. These criteria are based on the conditional covariance matrix of variables not selected, given those selected. Let Σ be partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} is the covariance matrix of the selected variables and Σ_{22} is the covariance matrix of the discarded variables. The conditional covariance matrix of the variables not selected given those selected can be written as

$$\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Let $\theta_i,\ i = 1, 2, \ldots, p-k$, be the eigenvalues of $\Sigma_{22\cdot 1}$ and $\lambda_i,\ i = 1, 2, \ldots, p$, be the eigenvalues of Σ. McCabe proposed the following four criteria:

1. $\min\ |\Sigma_{22\cdot 1}| = \min \prod_{i=1}^{p-k} \theta_i$,
2. $\min\ \mathrm{tr}(\Sigma_{22\cdot 1}) = \min \sum_{i=1}^{p-k} \theta_i$,
3. $\min\ \|\Sigma_{22\cdot 1}\|^2 = \min \sum_{i=1}^{p-k} \theta_i^2$,
4. $\max \sum_{i=1}^{k} \rho_i^2 = \max\ \mathrm{tr}(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$,

where ρ_i are canonical correlations between the variables not selected and those selected. The percentage of variance explained by a set of k principal variables is

$$P = \left(1 - \frac{\sum_{i=1}^{p-k} \theta_i}{\sum_{i=1}^{p} \lambda_i}\right) 100\%.$$

This percentage can be compared to the variation explained by the first few principal components.

Jolliffe and Cadima (1995) showed that discarding variables that have a small loading is not appropriate in various respects. They examined the effects on a single component if one or more variables are eliminated. They did not examine the effects on a subspace spanned by several components simultaneously. Jolliffe and Cadima implied that the correlation between the i^{th} variable and the j^{th} principal component, $\hat{\rho}_{ij}$, is an appropriate quantity to examine. This quantity can be written as

$$\hat{\rho}_{ij} = \hat{\gamma}_{ij} \left(\frac{l_j}{s_i^2} \right)^{1/2},$$

where $\hat{\gamma}_{ij}$ is loading for the i^{th} variable in the j^{th} component, l_j is the eigenvalue associated with that component, and s_i^2 is the variance of the i^{th} variable. They used an example to show that similar coefficients, $\hat{\gamma}_{ij}$, even very large ones, may translate into very different correlations between those variables and principal components or very different coefficients may be associated with similarly correlated variables and principal components. Jolliffe and Cadima examined the distance between the principal component and the truncated principal component. The truncated principal component is the principal component in which variables with small loading are ignored. The distance between the principal component and the truncated principal component can be expressed as

$$d = \frac{\|\boldsymbol{X}\hat{\gamma}_i - \boldsymbol{X}_k\hat{\gamma}_i^k\|}{\|\boldsymbol{X}\hat{\gamma}_i\|} = \left(1 - 2\hat{\gamma}_j^{k\prime}\hat{\gamma}_j^k + \frac{\hat{\gamma}_j^{k\prime}\boldsymbol{S}_k\hat{\gamma}_j^k}{l_j}\right)^{1/2},$$

where $\hat{\gamma}_j^k$ is the sub-vector of $\hat{\gamma}_j$ which results from retaining only the k coefficients associated with the variables that were retained, $\boldsymbol{S}_k$ is the sample covariance of the k selected variables, and $\boldsymbol{X}_k$ is the sub-matrix of a $n \times p$ column-centered data matrix, $\boldsymbol{X}$, obtained by retaining only k of its columns. The last term of d shows that selecting the k variables with the largest loadings is not guaranteed to yield the smallest distance.

The second approach for determining the number of variables is hypothesis testing. Anderson (1963) gave the asymptotic distribution of the sample eigenvalues and eigenvectors when the population eigenvalues are equal in sets. Anderson considered the hypothesis $\mathrm{H}_0\colon \gamma_1 = \gamma_0$, where γ_1 is the eigenvector of the first principal component and γ_0 is a specified vector such that $\gamma_0'\gamma_0 = 1$. If the hypothesis is true, then

$$n[l_1\gamma_0'\boldsymbol{S}^{-1}\gamma_0 + l_1^{-1}\gamma_0'\boldsymbol{S}\gamma_0 - 2]$$

has the limiting Chi-square distribution with degree of freedom $p-1$, where l_1 is the first eigenvalue of the sample covariance matrix, $\boldsymbol{S}$. Similar results can be applied for any other eigenvector corresponding to an eigenvalue with multiplicity 1.

Tyler (1981) modified and extended Anderson's results. Tyler derived asymptotic procedures for testing more general hypotheses concerning eigenvectors. Let $\boldsymbol{H}$ be a $p \times p$ matrix and $\boldsymbol{W}$ be a $p \times p$ positive definite symmetric matrix such that $\boldsymbol{WH}$ is symmetric. These conditions ensure that the eigenvalues and eigenvectors of $\boldsymbol{H}$ are real. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ be the ordered eigenvalues of $\boldsymbol{H}$ and let $\boldsymbol{A}$ be a

$p \times q$ matrix with rank q. Lastly, let w be a subset of m integers $(1 \leq m < p)$ from $\{1, 2, \ldots, p\}$ such that $\lambda_i \neq \lambda_j$ for all $i \in w,\ j \notin w$. Tyler considered the following two hypotheses:

for $q \leq m$,

H_0 : The columns of $\boldsymbol{A}$ lie in the subspace generated by the set of eigenvectors of $\boldsymbol{H}$ associated with m eigenvalues, $\lambda_i,\ i \in w$, and

for $q \geq m$,

H_0^* : The eigenvectors of $\boldsymbol{H}$ associated with the m eigenvalues, $\lambda_i,\ i \in w$, lie in the subspace generated by the columns of $\boldsymbol{A}$.

Let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p$ be the eigenvectors of $\boldsymbol{H}$ and let $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p$ be the eigenvectors of $\boldsymbol{H}_n$, where

$$\sqrt{n}\,\mathrm{vec}(\boldsymbol{H}_n - \boldsymbol{H}) \xrightarrow{L} N(\mathbf{0}, \Xi).$$

Denote the $p \times m$ matrix whose columns are $\boldsymbol{\beta}_i$ for $i \in w$ by $\boldsymbol{\beta}_w$ and denote the $p \times m$ matrix whose columns are $\mathbf{b}_i$ for $i \in w$ by $\mathbf{B}_w$. Let $d_1 \geq d_2 \geq \cdots \geq d_p$ be the ordered eigenvalues of $\boldsymbol{H}_n$, $\boldsymbol{P}_0 = \boldsymbol{\beta}_w(\boldsymbol{\beta}_w'\boldsymbol{\beta}_w)^{-1}\boldsymbol{\beta}_w'$, and $\hat{\boldsymbol{P}}_0 = \sum_{i \in w} \hat{\boldsymbol{P}}_i = \mathbf{B}_w(\mathbf{B}_w'\mathbf{B}_w)^{-1}\mathbf{B}_w'$. Then the test statistic for H_0 is

$$T_n(\boldsymbol{A}) = n\left\{\mathrm{vec}\left[(\boldsymbol{I}_p - \hat{\boldsymbol{P}}_0)\boldsymbol{A}\right]\right\}' \left[\hat{\Xi}(\boldsymbol{A})\right]^{-} \mathrm{vec}\left[(\boldsymbol{I}_p - \hat{\boldsymbol{P}}_0)\boldsymbol{A}\right], \tag{1.10}$$

where

$$\hat{\Xi}(\boldsymbol{A}) = (\boldsymbol{A}' \otimes \boldsymbol{I}_p)\boldsymbol{C}_w' \hat{\Xi}\, \boldsymbol{C}_w(\boldsymbol{A} \otimes \boldsymbol{I}_p),$$

$$\boldsymbol{C}_w = \sum_{i \in w} \sum_{j \notin w} (d_i - d_j)^{-1} \hat{\boldsymbol{P}}_i \otimes \hat{\boldsymbol{P}}_j'.$$

Tyler showed that, under H_0, $T_n(\boldsymbol{A}) \xrightarrow{L} \chi^2_{(p-m)q}$. The test is to reject H_0 if $r = \mathrm{rank}(\hat{\boldsymbol{P}}_0 \boldsymbol{A}) < q$ or if $r = q$ and $T_n(\boldsymbol{A}) > \chi^2_{1-\alpha,(p-m)q}$. The hypothesis H_0^* can be tested by using the following approach. Let $\boldsymbol{B}$ be a fixed $p \times (p-q)$ matrix with rank $p-q$ whose columns are orthogonal to $\boldsymbol{A}$, i.e. $\boldsymbol{A}'\boldsymbol{B} = \boldsymbol{0}$. The hypothesis H_0^* can be rephrased as

H_0 : The columns of $\boldsymbol{B}$ lie in the subspace generated by the set of eigenvectors of $\boldsymbol{H}'$ associated with $p-m$ eigenvalues, $\lambda_i,\ i \notin w$,

and can be treated as H_0.

The test statistic can be simplified if a random sample has been drawn from an elliptical distribution. Denote the covariance matrix by $\boldsymbol{H}$. Let $\hat{\kappa}$ be an estimator of κ, the kurtosis parameter. An estimator of the covariance matrix of $\mathrm{vec}(\boldsymbol{H}_n)$ can be written as

$$\hat{\Xi} = (1+\hat{\kappa})(\boldsymbol{I}_{p^2} + \boldsymbol{I}_{(p,p)})(\boldsymbol{H}_n \otimes \boldsymbol{H}_n) + \hat{\kappa}\,\mathrm{vec}(\boldsymbol{H}_n)[\mathrm{vec}(\boldsymbol{H}_n)]'.$$

The test statistic in (1.10) can be simplified as

$$T_n(\boldsymbol{A}) = n(1+\hat{\kappa}) \sum_{j \notin w} d_j^{-1}\, \mathrm{tr}\left\{ \boldsymbol{A}'\hat{\boldsymbol{P}}_i \boldsymbol{A} \left[\boldsymbol{A}'\boldsymbol{B}_w \boldsymbol{D}_j \boldsymbol{B}_w' \boldsymbol{A}\right]^{-1} \right\}, \qquad (1.11)$$

where $\boldsymbol{D}_j$ is an $m \times m$ diagonal matrix with entries $d_i/(d_i - d_j)^2$, $j \notin w$. For $q = m = 1$, $T_n(\boldsymbol{A})$ is asymptotically equivalent to the statistic given by Anderson (1963).

Flury (1986) discussed the test for redundancy of variables in the comparison of two covariance matrices. Let $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ be the covariance matrices of two independent p-vectors. Flury analyzed the eigenvectors of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$. Let $\{\boldsymbol{\beta}_j\}_{j=1}^p$ be a set of eigenvectors of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$ and partition $\{\boldsymbol{\beta}_j\}_{j=1}^p$ into the first $p-q$ and the last q rows as follows:

$$\{\boldsymbol{\beta}_j\}_{j=1}^p = \left\{ \begin{array}{c} \boldsymbol{\beta}_{j1} \\ \boldsymbol{\beta}_{j2} \end{array} \right\}_{j=1}^p .$$

The hypothesis of interest is

$$\mathrm{H}_0(p,q)\colon \boldsymbol{\beta}_{j2} = \mathbf{0} \quad \text{for all } j \in w. \tag{1.12}$$

This hypothesis can be formulated in the form H_0^* of Tyler (1981), by putting $\boldsymbol{A} = \left(\boldsymbol{I}_{p-q} \quad \mathbf{0}'\right)'$.

Then by using

$$\boldsymbol{B} = \left(\begin{array}{c} \mathbf{0} \\ \boldsymbol{I}_q \end{array} \right), \tag{1.13}$$

the hypothesis can be rephrased as

$\mathrm{H}_0(p,q)$: The columns of $\boldsymbol{B}$ lie in the subspace generated by the eigenvector $\boldsymbol{\alpha}_i$, $(i \notin w)$ of $\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1}$.

The test statistic in (1.10) can be applied. Let $\{\mathbf{b}_j\}_{j=1}^{p}$ be a set of eigenvectors of $\boldsymbol{S}_1^{-1}\boldsymbol{S}_2$ and partition $\{\mathbf{b}_j\}_{j=1}^{p}$ into the first $p-q$ and the last q rows as follows:

$$\{\mathbf{b}_j\}_{j=1}^{p} = \left\{ \begin{array}{c} \mathbf{b}_{j1} \\ \mathbf{b}_{j2} \end{array} \right\}_{j=1}^{p}.$$

The test statistic is

$$R(p,q) = n\sum_{j\in w}\mathbf{b}'_{j2}\left(\sum_{i\notin w}\frac{k_1 d_j^2 + k_2 d_i d_j}{(d_i-d_j)^2}\mathbf{b}_{i2}\mathbf{b}'_{i2}\right)^{-1}\mathbf{b}_{j2},$$

where $k_i = \lim_{n_i\to\infty}\frac{n}{n_i}$ for $i = 1,2$ and $d_j, j = 1,2,\ldots,p$, are the eigenvalues of $\boldsymbol{S}_2\boldsymbol{S}_1^{-1}$. Under $\mathrm{H}_0(p,q)$, $R(p,q) \xrightarrow{L} \chi^2_{mq}$.

Fujikoshi (1989) reviewed the problem of testing the hypothesis of redundancy in various multivariate situations including principal component analysis. Let $\{\boldsymbol{\gamma}_j\}_{j=1}^{p}$ be a set of eigenvectors of $\boldsymbol{\Sigma}$ and partition $\{\boldsymbol{\gamma}_j\}_{j=1}^{p}$ into the first $p-q$ and the last q rows as follow:

$$\{\boldsymbol{\gamma}_j\}_{j=1}^{p} = \left\{ \begin{array}{c} \boldsymbol{\gamma}_{j1} \\ \boldsymbol{\gamma}_{j2} \end{array} \right\}_{j=1}^{p}.$$

The hypothesis of redundancy of the last q variables in the first k components can be written as $\mathrm{H}_0(k,q)$: $\boldsymbol{\gamma}_{j2} = \mathbf{0}$ for all $j \in w$, $w = \{1,2,\ldots,k\}$, which is equivalent to

$\mathrm{H}_0(k,q)$: The columns of $\boldsymbol{B}$ lie in the subspace generated by the eigenvectors

$$\boldsymbol{\gamma}_i, (i \notin w) \text{ of } \boldsymbol{\Sigma}, \tag{1.14}$$

where $\boldsymbol{B}$ is defined in (1.13). Fujikoshi used the test statistic in (1.11) to test the hypothesis. Let l_i and $\hat{\boldsymbol{\gamma}}_i$, $i = 1,2,\ldots,p$, be the eigenvalues and eigenvectors of $\boldsymbol{S}$,

respectively. The test statistic can be rewritten as

$$T_n(\boldsymbol{B}) = n \sum_{j \in w} \hat{\gamma}'_{j2} \left(\sum_{i \notin w} \frac{l_i l_j}{(l_i - l_j)^2} \hat{\gamma}_{i2} \hat{\gamma}'_{i2} \right)^{-1} \hat{\gamma}_{j2}, \tag{1.15}$$

where under $\mathrm{H}_0(k, q)$, $T_n(\boldsymbol{B}) \xrightarrow{L} \chi^2_{kq}$.

Schott (1991) also employed the test statistic in (1.11) to test the hypothesis that in each of the first k components have zero loadings on the last q original variables. This hypothesis is equivalent to (1.14). The test is to reject $\mathrm{H}_0(k, q)$ if $r = \mathrm{rank}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{B}) \leq q$, or if $r = q$ and

$$T_{k,q} = n \sum_{j \notin w} l_j^{-1} \operatorname{tr} \left\{ \boldsymbol{B}' \hat{\gamma}_j \hat{\gamma}'_j \boldsymbol{B} [\boldsymbol{B}' \boldsymbol{\Gamma} \boldsymbol{D}_j \boldsymbol{\Gamma}' \boldsymbol{B}]^{-1} \right\} \geq \chi^2_{1-\alpha,kq}, \tag{1.16}$$

where $\boldsymbol{B}$ is defined in (1.13) and $\boldsymbol{D}_j$ is an $m \times m$ diagonal matrix with entries $l_i/(l_i - l_j)^2$, $j \notin w$, $w = \{1, 2, \ldots, k\}$. The statistic $T_{k,q}$ in (1.16) is a matrix expression for $T_n(\boldsymbol{B})$ in (1.15). A Bartlett adjustment was proposed based on the idea that if a test statistic T has a mean which can be expressed as

$$\mathrm{E}(T) = a \left\{ 1 + \frac{c}{n} + O(n^{-3/2}) \right\}, \tag{1.17}$$

where a is the asymptotic mean and c is a constant, then the mean of adjusted statistic,

$$T^* = \left(1 - \frac{c}{n} \right) T, \tag{1.18}$$

approaches a. It follows from $T_{k,q} \xrightarrow{L} \chi^2_{kq}$ that the asymptotic mean of $T_{k,q}$ in (1.16) is $a = kq$. An expression for c can be found in Schott (1991). Schott compared the unadjusted statistic, $T_{k,q}$, and adjusted statistic, T^*, under several conditions using

simulated data. He showed that, in general, the Bartlett adjusted statistic performs better than $T_{k,q}$ with respect to Type I error.

Dümbgen (1995) proposed a likelihood ratio test for principal components of a matrix $\mathbf{\Sigma}$. Let $\mathbf{\Omega}$ be the set of all symmetric matrices in $\mathbb{R}^{p\times p}$ and $\mathcal{O}(p)$ be the set of all orthogonal matrices in $\mathbb{R}^{p\times p}$. For $\mathbf{\Sigma} \in \mathbf{\Omega}$, let $\mathbf{\Gamma} \in \mathcal{O}(p)$ such that $\mathbf{\Gamma}'\mathbf{\Sigma}\mathbf{\Gamma} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Let $m = (m_1, m_2, \ldots, m_{\bar{a}})$ be a partition of {1,2,...,p} into $\bar{a} > 1$ sets. The hypothesis of interest is

$$\mathrm{H}_0 \colon \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}',$$

where $\mathbf{\Gamma} = \oplus_{i=1}^{\bar{a}} \mathbf{\Gamma}_{ii}$, $\mathbf{\Lambda} = \oplus_{i=1}^{\bar{a}} \mathbf{\Lambda}_{ii}$, $\mathbf{\Lambda}_{ii} = \text{diag}(\lambda_i)$ for $i \in m_i$, $\mathbf{\Gamma}_{ii}$ is an $m_i \times m_i$ matrix, $\mathbf{\Gamma}_{ii} \in \mathcal{O}(m_i)$, $\sum_{i=1}^{\bar{a}} m_i = p$, and $\oplus$ stands for the direct sum. Let $l_1 \geq l_2 \geq \cdots \geq l_p$ be the eigenvalues of $\boldsymbol{S}$ and define

$$\boldsymbol{\mu} = \arg \min_{\boldsymbol{\nu} \in \boldsymbol{\lambda}(\mathbf{\Omega})} \|\boldsymbol{\nu} - \boldsymbol{\lambda}_m(\boldsymbol{S})\|^2,$$

where $\boldsymbol{\lambda}(\boldsymbol{S})$ is the vector of the ordered eigenvalues of $\boldsymbol{S}$, and

$$\boldsymbol{\lambda}_m(\boldsymbol{S}) = \begin{pmatrix} \boldsymbol{\lambda}(\boldsymbol{S}_{ij})_{i,j\in m_1} \\ \boldsymbol{\lambda}(\boldsymbol{S}_{ij})_{i,j\in m_2} \\ \vdots \\ \boldsymbol{\lambda}(\boldsymbol{S}_{ij})_{i,j\in m_{\bar{a}}} \end{pmatrix}.$$

The exact computation of $\boldsymbol{\mu}$ is described in Dümbgen (1995). The likelihood ratio test is to reject H_0 whenever $t_k(\boldsymbol{S}) \leq c(\alpha)$, where

$$t_k(\boldsymbol{S}) = \sum_{i=1}^{p} \ln\left(\frac{\mu_i}{l_i}\right).$$

The test statistic requires a simulation to obtain critical values.

In this thesis, a likelihood approach for making inferences about the covariance parameters is proposed and is applied to redundancy test. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ be the diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}^*$ be a set of the corresponding eigenvectors. The subspace spanned by the columns of a matrix, $\boldsymbol{T}$, is denoted by $\mathcal{R}(\boldsymbol{T})$. Let $\boldsymbol{A}$ be a known $p \times q$ semi-orthogonal matrix with rank q, and $\boldsymbol{M}^*$ be a $p \times m$ semi-orthogonal matrix such that $\mathcal{R}(\boldsymbol{M}^*)$ spans the subspace generated by m specific columns of $\boldsymbol{I}_p$. The hypotheses of interest are

$$\text{H}_0 \quad : \quad \boldsymbol{A} \in \mathcal{R}(\boldsymbol{\Gamma}^*\boldsymbol{M}^*) \quad \text{for } q \leq m, \text{ and} \tag{1.19}$$

$$\text{H}_0^* \quad : \quad \boldsymbol{\Gamma}^*\boldsymbol{M}^* \in \mathcal{R}(\boldsymbol{A}) \quad \text{for } q > m. \tag{1.20}$$

It follows from (1.20), see Tyler (1981), that

$$\boldsymbol{\Gamma}^*\boldsymbol{M}^* \in \mathcal{R}(\boldsymbol{A}) \Longleftrightarrow \boldsymbol{A}^c \in \mathcal{R}(\boldsymbol{\Gamma}^*\boldsymbol{M}^{*c}),$$

where $\boldsymbol{A}^c$ and $\boldsymbol{M}^{*c}$ are orthogonal complement to $\boldsymbol{A}$ and $\boldsymbol{M}^*$, respectively. That is, $\boldsymbol{A}^c\boldsymbol{A}^{c\prime} = \boldsymbol{I}_p - \boldsymbol{A}\boldsymbol{A}'$ and $\boldsymbol{M}^{*c}\boldsymbol{M}^{*c\prime} = \boldsymbol{I}_p - \boldsymbol{M}^*\boldsymbol{M}^{*\prime}$. Accordingly, the hypothesis H_0^* in (1.20) can be rewritten and treated as the hypothesis H_0 in (1.19). Only hypothesis H_0 in (1.19) is employed in the remainder of this thesis.

To match the theoretical setup in the later chapters and without loss of generality, $\boldsymbol{M}^*$ can always be equated to the first m columns of $\boldsymbol{I}_p$. This simplification is possible because the columns of $\boldsymbol{\Gamma}^*$ and the rows of $\boldsymbol{M}^*$ can be permuted to satisfy

$$\mathcal{R}(\boldsymbol{M}^*) = \mathcal{R}\begin{pmatrix} \boldsymbol{I}_m \\ \boldsymbol{0} \end{pmatrix}.$$

If $\mathcal{R}(\boldsymbol{M}^*) \neq \mathcal{R}\begin{pmatrix} \boldsymbol{I}_m \\ \boldsymbol{0} \end{pmatrix}$, then replace $\boldsymbol{\Gamma}^*\boldsymbol{M}^*$ by $\boldsymbol{\Gamma}\boldsymbol{M}$, where $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^*\boldsymbol{P}$ and $\boldsymbol{M} = \boldsymbol{P}'\boldsymbol{M}^*$, where $\boldsymbol{P}$ is a $p \times p$ permutation matrix. Choose $\boldsymbol{P}$ so that

$$\mathcal{R}(\boldsymbol{P}'\boldsymbol{M}^*) = \mathcal{R}\begin{pmatrix} \boldsymbol{I}_m \\ \boldsymbol{0} \end{pmatrix}.$$

In the remainder of this thesis, hypothesis (1.19) will be written as

$$\mathrm{H}_0\colon \boldsymbol{A} \in \mathcal{R}(\boldsymbol{\Gamma}\boldsymbol{M}), \text{ where } \mathcal{R}(\boldsymbol{M}) = \mathcal{R}\begin{pmatrix} \boldsymbol{I}_m \\ \boldsymbol{0} \end{pmatrix}. \tag{1.21}$$

This hypothesis is general and the hypothesis of redundancy is subsumed as a special case. In Chapter 3, an algorithm for computing MLEs of the covariance parameters satisfying multiplicity and other constraints is proposed under H_0 in (1.21). A likelihood ratio test and a Bartlett correction for the test of (1.21) are also developed in Chapter 4.

The hypothesis of redundancy can be written as (1.21) in the following manner. First, the eigenvalues will be reordered so that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p$. This order is used throughout the remainder of the thesis. The rational for reordering the eigenvalues is described in Chapter 2. Let $\boldsymbol{\Gamma}$ be a matrix whose columns are the corresponding eigenvectors. Then, by equating $\boldsymbol{A}$ to $\boldsymbol{A} = (\boldsymbol{I}_q \quad \boldsymbol{0}')'$, the hypothesis in (1.21) matches with the hypothesis in (1.14). This hypothesis states that each of the first k components has zero loadings on the last q original variables. The redundancy hypothesis is identical to that tested by Fujikoshi (1989) and Schott (1991) but it differs from the variable selection approaches of Jolliffe (1972) and McCabe (1984). To understand

the distinction, first note that it follows from (1.2) that $\boldsymbol{y} = \boldsymbol{\Gamma z}$. Partition $\boldsymbol{y}$ as

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix},$$

where $\boldsymbol{y}_1$ is $(p-q) \times 1$ and $\boldsymbol{y}_2$ is $q \times 1$. Partition $\boldsymbol{z}$ as

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{pmatrix},$$

where $\boldsymbol{z}_1$ is $k \times 1$ and $\boldsymbol{z}_2$ is $(p-k) \times 1$. The last q variables are redundant if their loadings on the first k principal components are zero. That is, the last q variables are redundant if

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \mathbf{0} & \boldsymbol{\Gamma}_{22} \end{pmatrix}. \tag{1.22}$$

It follows from (1.2) that

$$\text{var}\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = \text{var}\begin{pmatrix} \boldsymbol{I}_p \\ \boldsymbol{\Gamma}' \end{pmatrix} \boldsymbol{y} = \begin{pmatrix} \boldsymbol{\Gamma\Lambda\Gamma}' & \boldsymbol{\Gamma\Lambda} \\ \boldsymbol{\Lambda\Gamma}' & \boldsymbol{\Lambda} \end{pmatrix},$$

where $\boldsymbol{\Gamma}$ is defined in (1.22). The conditional covariance of $\boldsymbol{y}_2$ given $\boldsymbol{z}_2$ can be written as

$$\text{var}(\boldsymbol{y}_2|\boldsymbol{z}_2) \;=\; \text{var}(\boldsymbol{y}_2) - \text{cov}(\boldsymbol{y}_2, \boldsymbol{z}_2)\{\text{var}(\boldsymbol{z}_2)\}^{-1}\text{cov}(\boldsymbol{z}_2, \boldsymbol{y}_2).$$

If the last q variables are redundant, then

$$\text{var}(\boldsymbol{y}_2|\boldsymbol{z}_2) = \boldsymbol{\Gamma}_{22}\boldsymbol{\Lambda}_2\boldsymbol{\Gamma}'_{22} - \boldsymbol{\Gamma}_{22}\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Lambda}_2\boldsymbol{\Gamma}'_{22} = \mathbf{0}.$$

That is, if $\boldsymbol{y}_2$ is redundant with respect to $\boldsymbol{z}_1$ then $\text{var}(\boldsymbol{y}_2|\boldsymbol{z}_2) = \mathbf{0}$.

On the other hand, Jolliffe (1972) and McCabe (1984) reduced the number of

variables in such a way that the variables can be discarded if all linear functions of $\boldsymbol{y}$ can be approximately reproduced as linear combination of $\boldsymbol{y}_1$. It can be shown that satisfaction of the constraint (1.22) does not guarantee that any of McCabe criteria (see page 17) are optimized. For example, if $\boldsymbol{\Gamma}$ satisfies (1.22), then the conditional covariance of $\boldsymbol{y}_2$ given $\boldsymbol{y}_1$ can be written as

$$\boldsymbol{\Sigma}_{22\cdot 1} = \left(\boldsymbol{\Gamma}_{22}\boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Gamma}_{22}'\right)^{-1},$$

which can have norm as large as $\left\|\left(\boldsymbol{\Gamma}_{22}\boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Gamma}_{22}'\right)^{-1}\right\|^2 = \mathrm{tr}(\boldsymbol{\Lambda}_2^2) = (p-k)\lambda_2^2$ if all eigenvalues in $\boldsymbol{\Lambda}_2$ are equal to λ_2. Conversely, $\|\boldsymbol{\Sigma}_{22\cdot 1}\|^2$ can be small even though redundancy is not satisfied. For example, suppose that $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_1 + \boldsymbol{\varepsilon} \end{pmatrix}$, where $\mathrm{var}(\boldsymbol{y}_1) = \boldsymbol{\Sigma}_{11}$, $\mathrm{var}(\boldsymbol{\varepsilon}) = a\boldsymbol{\Sigma}_{11}$, and a is a scalar. The covariance matrix of $\boldsymbol{y}$ can be written as

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 1+a \end{pmatrix} \otimes \boldsymbol{\Sigma}_{11}.$$

Let $\boldsymbol{\Gamma}_{11}$ be the matrix of eigenvectors of $\boldsymbol{\Sigma}_{11}$. Then, the eigenvector matrix of $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$, can be expressed as

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & 1 \\ \frac{a}{2} + \sqrt{1+\frac{a^2}{4}} & \frac{a}{2} - \sqrt{1+\frac{a^2}{4}} \end{pmatrix} \otimes \boldsymbol{\Gamma}_{11}.$$

The conditional covariance of $\boldsymbol{y}_2$ given $\boldsymbol{y}_1$ can be written as

$$\boldsymbol{\Sigma}_{22\cdot 1} = (1+a)\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11} = a\boldsymbol{\Sigma}_{11}.$$

McCabe criteria can be satisfied by choosing a to be a small number. If a is close to zero, then $\boldsymbol{\Gamma}$ approaches

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \boldsymbol{\Gamma}_{11}$$

which does not satisfy the redundancy constraint in (1.22). Accordingly, the criteria of Jolliffe and McCabe were not constructed to reduce the number of variables with respect to redundancy.

The likelihood ratio test for redundancy is compared with Schott's tests in (1.16) and (1.18) using simulated data. A simulation study of the effectiveness of the Bartlett correction is described in Chapter 5.

To illustrate the dimensionality reduction problem, two examples are described below.

Example 1

A large number of variables were observed in a study of the quality of pictures produced by a photographic process. The data were originally presented by Jackson and Morris (1957) and were discussed by Schott (1991). The procedure for a check on the process was as follows: A film strip was given a graded series of exposures to white light and was processed. Optical densities were measured through red, green, and blue filters at the high-density portion of the characteristic curve (shadow areas), at the middle-tone portion of the curve (average picture density) and at the toe portion of the curve (highlights, whites, etc.). There were $p = 9$ measurements: three density levels and three colors at each level. The sample covariance matrix based on $N = 109$ was given in Jackson and Morris (1957). The eigenvalues and cumulative proportions of total variance are given in Table 1. The eigenvectors of the first two